Summer 8-6-2015

Network Structure, Network Flows and the Phenomenon of Influence in Online Social Networks: An Exploratory Empirical Study of Twitter Conversations about YouTube Product Categories

Nitin Venkat Mayande
Portland State University

Network Structure, Network Flows and the Phenomenon of Influence in Online Social Networks: An Exploratory Empirical Study of Twitter Conversations about YouTube Product Categories

by

Nitin Venkat Mayande

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Technology Management

Dissertation Committee:
Charles M. Weber, Chair
Timothy R. Anderson
Charla Mathwick
Antonie J. Jetter

Portland State University
2015

Abstract

Traditional marketing models are swiftly being upended by the advent of online social networks. Yet, practicing firms that are engaging with online social networks neither have a reliable theory nor sufficient practical experience to make sense of the phenomenon. Extant theory in particular is based on observations of the real world, and may thus not apply to online social networks. Practicing firms may consequently be misallocating a large amount of resources, simply because they do not know how the online social networks with which they interact are organized.

The purpose of this dissertation is to investigate how online social networks that are in stark contrast to real-world social networks behave and how they get organized. In particular, I explore how network structure and information flow within the network impact each other, and how they affect the phenomenon of influence in online social networks. I have collected retrospective data from Twitter conversations about six YouTube product categories (Music, Entertainment, Comedy, Science, Howto and Sports) in continuous time for a period of three months. Measures of network structure (Scale Free Metric, Assortativity and Small World Metric), network flows (Total Paths, Total Shortest Paths, Graph Diameter, Average Path Length, and Average Geodesic Length) and influence (Eigenvector Centrality/Centralization) were computed from the data. Experimental measures such as power law distributions of paths, shortest paths and nodal eigenvector centrality were introduced to account for node-level structure.

Factor analysis and regression analysis were used to analyze the data and generate results.

The research conducted in this dissertation has yielded three significant findings.

1. Network structure impacts network information flow, and conversely; network flow and network structure impact the network phenomenon of influence. However, the impact of network structure and network flow on influence could not be identified in all instances, suggesting that it cannot be taken for granted.

2. The nature of influence within a social network cannot be understood just by analyzing undirected or directed networks. The behavioral traits of individuals within the network can be deduced by analyzing how information is propagated throughout the network and how it is consumed.

3. An increase or decrease in the scale of a network leads to the observation of different organizational processes, which are most likely driven by very different social phenomena. Social theories that were developed from observing real-world networks of a relatively small scale (hundreds or thousands of people) consequently do not necessarily apply to online social networks, which can exhibit significantly larger scale (tens of thousands or millions of people).

The primary contribution of this dissertation is an enhanced understanding of how online social networks, which exhibit contrasting characteristics to social networks that have been observed in the real world, behave and how they get organized. The empirical findings of this dissertation may allow practicing managers that engage with online social networks to allocate resources more effectively, especially in marketing. The primary limitations of this research are the inability to identify the causes of change within networks, glean demographic information and generalize across contexts. These limitations can all be overcome by follow-on studies of networks that operate in different contexts. In particular, further study of a variety of online social networks that operate on different social networking platforms would determine the extent to which the findings of this dissertation are generalizable to other online social networks. Conclusions drawn from an aggregation of these studies could serve as the foundation of a more broadly-based theory of online social networks.

Dedication

Every challenging work needs self-effort and guidance in equal parts especially from those who are very close to our heart. I dedicate my humble efforts to my parents, Dr. V. M. Mayande and Mrs. Vijaya V. Mayande, who are my guiding light and source of inspiration; Nutan Mayande, my elder sister, for always being there for me and Sonal Patil, my wife for her unconditional love and support.

I also dedicate this dissertation to all my teachers who have inspired me over the years and kept my curiosity alive, especially Dr. Charles M. Weber. I am grateful for his unwavering support and encouragement over the years.

Acknowledgements

Over the years I have received support and encouragement from a great number of individuals. Dr. Charles M. Weber has been a mentor, colleague, and friend. His guidance has made this a thoughtful and rewarding journey. I would like to thank my dissertation committee of Dr. Timothy Anderson, Dr. Charla Mathwick and Dr. Antonie Jetter for their support over the past years as I moved from an idea to a completed study. I express my special gratitude to Dr. Antonie Jetter for guiding me through the comprehensive exam process and for always being there whenever I needed her support and guidance.

I would like to thank Matt Hixson for being my partner in crime and for all the valuable conversations and ideas over the years, my Tellagence team (especially Trevor and Jeff) for being supportive of my efforts and a constant source of encouragement.

Special thanks to Ann-Marie Lamb for being a good friend and support system. I would also like to thank Bill Pearson from Intel Corporation for lighting the fire and providing opportunities to explore social networks. Finally, thanks to my colleagues and friends at ETM Department who made my time spent in school memorable.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Research Problem

1.1.1 Online Social Networks

Online social networks are aggregations that emerge from the Internet when people carry on public discussions (Preece, 2000, Rheingold, 1993, Schoberth and Schrott, 2001). They have enabled organizations to leverage the network value of business ecosystems (Afsarmanesh and Camarinha-Matos, 2005) in activities such as marketing, customer service and product innovation (Bressler and Grantham, 2000). Online social networks are at the core of many successful business models, and they are used to coordinate business and information exchanges (Feller et al., 2008).

People all around the world are utilizing online social networks at an astonishing rate. It is estimated that there will be around 2.13 billion social network users around the globe in 2016, up from 1.4 billion in 2012[1]. Due to their rapidly growing popularity, online social networks are having a major and growing impact on consumer behavior. A study conducted by Interactive Advertising Bureau (IAB) and Pricewaterhouse Coopers US (PwC US) concludes that, “Consumers are turning to interactive media in droves to look for the latest information, to connect with their social networks, and simply to be

[1] http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Accessed on 5/13/2015

entertained."[2] Many of the online conversations concern products and services (Chakrabarti and Berthon, 2012), implying that the commercial impact of online social networks can no longer be ignored.

Marketers not only need to pay attention to these conversations (Chakrabarti and Berthon, 2012); they must also try to become a part of these conversations, in order to shape them. When the conversations are positive, they can lead to free advertising and better brand recognition (Longart, 2010). However, when the conversations are negative, they can do irreparable financial damage (Ayres, 2009, Khammash and Griffiths, 2011). Online conversations can therefore make or break a product or a service.

Today's marketers are responding to the increasing importance of online social networks by spending billions of dollars in digital marketing. According to Proctor and Gamble's chief executive A. G. Lafley, " … digital spending on things like online ads and social media ranges from 25% to 35% of *the company's* marketing budget and is currently near the top of that range in the U.S., its biggest market." [3] If these investments in online ads and social media do not yield demonstrable improvements in sales, then large sums of money will have been misallocated. The stakes in marketing via online social networks could therefore not be higher.

---

[2] Randall Rothenberg, President and CEO, IAB1; http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-060313 . Accessed on 04/01/2014

[3] http://online.wsj.com/article/SB10001424127887323681904578641993173406444.html Accessed on 04/01/2014

1.1.2 Social Network Analytics

Many companies are reallocating their marketing resources to specifically target users of Facebook and Twitter, [4] two of the most popular social networking platforms of the mid-2010s, where the majority of online conversations about products and services take place. [5] Yet even on Facebook and Twitter, companies tend to use traditional approaches to marketing, which rely on broadcasting information that is passively consumed. [6] However, advertising via social media requires users of online social networks to deliberately spread the information they receive through word of mouth (Hodas and Lerman, 2014), an approach that is demonstrably more efficient and effective than broadcasting information.[7] It is thus not surprising that traditional methods of marketing on the Internet have produced disappointing outcomes in online social networks (Edward, 2012, Rusli and Eavis, 2012, Terlep et al., 2012). This implies that traditional Internet marketing paradigms and processes are being upended by

---

[4] http://adage.com/article/digital/ad-age-reader-survey-twitter-facebook-youtube/293923/ Accessed on 05/13/2015

[5] http://www.sas.com/resources/whitepaper/wp_23348.pdf Accessed on 05/13/2015

[6] http://www.theguardian.com/technology/2013/aug/12/engage-dont-broadcast-the-need-for-authenticity-in-social-media Accessed on 05/13/2015

[7] The Nielsen agency conducted a global survey of trust in advertising, in which it polled more than 29,000 Internet respondents in 58 countries to measure consumer sentiment on 19 forms of paid, earned and owned advertising formats. Not surprisingly, this study concluded that word-of-mouth formats, such as recommendations from family and friends and consumer opinions posted online, prompted the highest levels of self-reported action among 84 percent and 70 percent of respondents, respectively. http://www.nielsen.com/us/en/press-room/2013/nielsen--earned-advertising-remains-most-credible-among-consumer.html Accessed on 04/01/2014

swiftly evolving social platforms and technology (Deighton, 2012), and that billions of dollars in marketing resources have been misallocated. [8]

With increased spending on social media, businesses are feeling the pressure to gain new insights into customer behavior. They need to know who the online influencers are and how they exert their influence (Lindsay et al.,2014). [9](Lindsay et al.)They require analytics to transform enormous volumes of data into actionable strategies (Halavais, 2015). According to a report by the research firm Gartner, companies spent a total of $76 million on social media analytics in 2011. This number is expected to increase by almost $1 billion every year to reach over $4 billion by 2016.[10]

Success in marketing though online social media critically depends upon understanding the virtual community that may have a potential interest in your product or service and by identifying the key influencers that will spread your marketing message (Lindsay et al.,2014). However, due to the fluid nature of social media, this is easier said than done. As Jure Klepic, social media innovator, states in Huffington Post: “A topic may be trending one day, someone may be popular the next or themes may

---

[8] http://www.businessinsider.com/priceline-ceo-facebook-and-twitter-are-useless-for-ads-2014-4 Accessed on 05/13/2015

[9] http://www.forbes.com/sites/kylewong/2014/09/10/the-explosive-growth-of-influencer-marketing-and-what-it-means-for-you/ Accessed on 05/13/2015

[10] http://www.forbes.com/sites/louiscolumbus/2012/10/15/using-search-analytics-to-see-into-gartners-232b-big-data-forecast/ Accessed on 04/01/2014

change almost hourly. By the time a marketer develops a response, the social universe has moved on." [11]

Many firms that engage in social media analytics (e.g., Klout, Kred, PeerIndex, and Traackr) have tried to overcome these challenges by finding the individuals that have the most friends and followers or generate the most output. [12] This approach has not been particularly successful (Cha et al., 2010). Evidently, those who have the most connections or generate the most activity online are not the true influencers in social media (Cha et al., 2010, Wu et al., 2011), and whatever influence they have is ephemeral (Wu et al., 2011, Romero et al., 2011). Instead, people appear to consume information from people they know and from people they trust,[13] just as they do in the real world (Rogers, 2003).

1.1.3 Network Flows, Network Structure and Network Phenomena

Many of the approaches that practitioners of social network analytics have deployed are grounded in theory that was developed almost entirely from observing social networks in the real world (e.g., (Bailey, 1990, Luhmann, 1986, Miller, 1978, Parson, 1951). For example, practitioners track the deliberate propagation of information, through word of mouth, from one user to another (Granovetter, 1973,

[11] http://www.huffingtonpost.com/jure-klepic/discover-the-next-advance_b_3991536.html Accessed on 04/01/2014
[12] http://blog.crazyegg.com/2013/06/04/dont-like-klout/ Accessed on 05/13/2015
[13] http://www.nielsen.com/us/en/press-room/2013/nielsen--earned-advertising-remains-most-credible-among-consumer.html Accessed on 04/01/2014

Tichy et al., 1979, Rogers, 2003). This method of information transfer is henceforth referred to as *network flows* in this dissertation.

Social scientists have long understood the importance of network flows in spreading information (Granovetter, 1973) and in diffusion of innovations (Rogers, 2003) in real-world social networks. All network flows in the real world take place between the seeker of information and the source of information, and all network flows transpire within existing social relationships (Bristor, 1989, Duhan et al., 1997, Money et al., 1998). Individuals in a strong relationship tend to interact more frequently and exchange more information, compared to those in a weak relationship (Brown and Reingen, 1987).

In real-world social networks, interactions only happen between people who have social relationships (Burt, 1987). Thus an individual's relationship network and his/her interaction network are considered to be *one and the same* (Burt, 1987). Therefore, the structure of an individual's relationship network or the structure of his/her interaction network is henceforth defined as *network structure* in this dissertation.

In extant theory on social networks, network structure defines the boundaries of communities (Bailey, 1990, Luhmann, 1986, Miller, 1978, Parson, 1951). For example, in living systems theory (Miller, 1978), a system is defined as a set of interacting units and the relationships among them. The boundaries of these interacting units are determined by the processes through which these units get organized. These units are organized hierarchically. For example, two or more people and their relationships comprise a group; communities consist of two or more groups and two or more communities comprise a society. There are comparatively few barriers to information transfer within units than there are between the units. Therefore, the boundaries between units (e.g., groups, communities, societies) constrain network flows between the units.

Within communities, network structure guides the network flows (Bailey, 1990, Luhmann, 1986, Parson, 1951) and network flows give rise to network phenomena such as social influence (Cartwright, 1965, March, 1955, Simon, 1957), social capital (Bourdieu, 1986, Burt, 1992, Burt, 2005, Coleman, 1988, Putnam, 1995), social behavior (Allen, 1977, Burt, 1976, Granovetter, 1973) and economic benefit (Allen, 1977, Bourdieu, 1986, Burt, 1976, Burt, 1992, Cartwright, 1965, Coleman, 1988, Granovetter, 1973). The network phenomenon of interest in this thesis is *social influence*. Henceforth, any reference to social phenomena or network phenomena implies social influence, unless specifically stated otherwise.

Social influence in real-world networks occurs when an actor adapts his/her behavior to the behavior of other actors in the community (March, 1955, Simon, 1957, Cartwright, 1965). A precondition for social influence is the availability of information, through network flows, about the other actors (Leenders, 1995). The scope of the network flows within all real-world networks is constrained by factors such as connectivity (the number of actors to which an individual is connected) (Allen, 1977, Burt, 1976, Burt, 1992, Granovetter, 1973) and physical distance between the actors in the network (Allen, 1977). Therefore, an individual influence in a real world network depends upon the individual's connectivity, his/her access to an individual with high connectivity or a combination of both.

1.1.4 Social Networks: Real-World versus Online

Online social networks differ from real world social networks in a variety of ways. First and foremost, online social networks tend to be larger than the social networks that have been studied in the real world. Known real-world social networks tend to consist of hundreds or thousands of people (e.g., Granovetter, 1973, Tichy et al., 1979, Burt, 1987, Rogers, 2003); online network may contain hundreds of thousands or millions (Mislove et al., 2007, Dodds et al., 2011, Moon et al., 2011). Networks of such different scale could thus behave differently; some social processes may transpire in very large but not in comparatively small processes, and conversely. Social theories that were developed from observing real-world networks may thus not necessarily apply to online social networks.

Secondly, the ability to conduct searches in online social networks (Watts et al., 2002, Adamic and Adar, 2005) makes the network structure and the network flows, which result from the interaction that follows that search, highly dynamic (Dodds et al., 2003). Real world constraints such as connectedness and distance may consequently not have any significant impact on the behavior online social networks (Borgatti and Cross, 2003, Borgatti, 2005). Instead, the online social networks may be most affected by topological organization of network structure (e.g., "scale free" (Barabási and Albert, 1999), "assortativity" (Newman, 2002) and "small world" (Watts and Strogatz, 1998)) or by various attributes of network flows (e.g., paths, geodesics) (Borgatti, 2005), which extant theory of social networks does not really consider (Borgatti and Cross, 2003) and prior empirical studies have not explored.[14]

As a consequence, network flows in online social networks cannot all be attributed to social relationships (Pei et al., 2014). We do know from observation of practicing firms (Wiertz et al., 2010) that online social networks are an emergent phenomenon (in the sense of (Sandelands and Drazin, 1989, Drazin and Sandelands, 1992)), and that network flows can be generated by ad hoc interactions. For example, the DARPA Network Challenge successfully tested the ability of online social networks to mobilize massive ad hoc teams to solve problems (Greenemeier, 2009), suggesting that an individual's online social network and his/her online interaction network are *not* one and the same thing. We also know from observing hashtag communities that people in online social networks may interact virtually with people with whom they share a common interest. The observation of hashtag communities also tells us that online social networks and network flows can be ephemeral (Weng et al., 2012). They can

[14] Neither do studies of phenomena that are somewhat related to social networks, such as business ecosystems (Iansiti & Levien, 2004; Adner, 2006; Moore, 2006) or open source software development (von Hippel, E. & von Krogh, G., 2003; Shah, 2005; West and Lakhani, 2008).

disappear on short notice, as the common interest of the community dissipates (Weng et al., 2012).

The above observations suggest that bonding between people in online social networks may be very different from what it is in the real world. In the real world, social relationships are required for a social network to form and function. This is not necessarily true in the virtual world. As a consequence, conversations may be more structured in the real world than they are online. Theories of social networks that assume strong bonds cause or enable network phenomena may therefore not apply to online social networks.

Table 1: Real-World versus Online Networks

Real World Networks	Online Networks
Limited scale (e.g., 3 to 1000 members)	Unlimited scale (up to millions of members)
Non-emergent (Static network structure) (Burt, 1987, Moffitt, 2001)	Emergent (Dynamic network structure) (Centola, 2010, Chomutare et al., 2014, Sasidharan et al., 2011, Wiertz et al., 2010)
Networks flows transpire within social relationships. (Bristor, 1989, Duhan et al., 1997, Money et al., 1998)	Networks flows do not need social relationships. (Watts et al., 2002, Adamic and Adar, 2005, Pei et al., 2014)
Connected network and interactive network are the same. (Bristor, 1989, Duhan et al., 1997, Money et al., 1998, Brown and Reingen, 1987)	Connected network and interactive network differ significantly. (Dodds et al., 2003, Wilson et al., 2012)

1.1.5 Toward a Theory of (Online) Social Networks

Table 1 summarizes the attributes of scientifically observed real-world social networks and contrasts them with attributes that have been observed in online social networks. Table 1 clearly illustrates what has been stated above—extant theory, which is based upon observation of the real world, cannot be relied upon to explain the nature and behavior of online social networks effectively. The observed differences between online and real-world social networks are simply too vast. Even a comprehensive theory of online social networks is difficult to frame, because the degree to which many of the abovementioned attributes of online social networks occur may be platform specific or network specific. Such a theory would have to be platform independent, scalable and take directionality of network flows into consideration. An overarching theory of social networks that covers real-world and online social networks is even more difficult to build. It would have to explain how all social networks, real-world or online, behave.

It goes beyond the scope of this dissertation to develop an empirically grounded theoretical framework for all social networks or even all online networks. However, this dissertation can make a significant contribution to theory by empirically investigating online social networks that exhibit the greatest contrast to real-world social networks, which are relatively well understood. Follow-on studies (perhaps conducted by other researchers) can subsequently investigate other social networks, which exhibit less of a contrast with those that occur in the real world. A comprehensive, empirically grounded theory of social networks—real-world and online—could potentially be developed once all these empirical studies have been performed.

Online social networks that are in stark contrast to those that have been observed in the real world would have to exhibit the following characteristics. They would have to be very large, emergent, dynamic and potentially ephemeral. They would have to contain network flows that do not rely on social relationships. Characteristics that are associated with network structure, such as "scalefreeness," "assortativity" and "smallworldness" would have to be demonstrably observable, and the phenomenon of influence would have to be readily identifiable. Furthermore, the existence of relationships between network flows, network structure and network phenomenon would have to be demonstrated as a prerequisite to gaining an understanding of how these networks get organized.

Due to the emergent and dynamic nature of online social networks, the relationship between network structure, network flows and the resulting network phenomenon in these networks is not very well understood. Recent research on network structure (Centola, 2010, Chomutare et al., 2014, Sasidharan et al., 2011), network flow (Hodas and Lerman, 2014, Burt et al., 2013, Aral and Walker, 2011, Dellarocas et al., 2013) and network phenomena (Aral and Walker, 2012, Pei et al., 2014, Khammash and Griffiths, 2011, Muchnik et al., 2013a, Muchnik et al., 2013b) focuses on these individual categories. However, studies that characterize the mechanisms through which network structure, network flow and the network phenomenon collectively emerge and operate are woefully lacking (Aral et al., 2013). We cannot even identify the loci of influence within an online social network reliably. Thus we are unable to explain how and why online social networks respond to a marketing message. To date, we do not know how online social networks form, how they get organized and how they evolve. As practitioners concede (Li and Bernoff, 2008), firms that are considering engaging in online social networks have neither a reliable theory nor sufficient practical experience to manage these networks effectively. Even companies that are very adroit at marketing via online social networks have experienced unintended consequences when they attempted to direct and control social networks (Wiertz et al., 2010). Using online social networks deliberately to gain competitive advantage may consequently turn out to be challenging. Nonetheless, the social and economic impact of online social networks on the modern world is increasing rapidly. The case for further academic study of the nature of online social networks is therefore compelling and urgent.

1.2 Purpose, Scope and Setting of Dissertation Research

The purpose of this dissertation is to investigate how online social networks that are in stark contrast to real-world social networks behave and how they get organized. To achieve this purpose, I conduct an exploratory empirical study that investigates how network structure and network flows in these networks impact each other and how they impact the network phenomenon of influence *in the aggregate*. Therefore, inferences about individual influencers cannot be drawn. The essential management question being addressed in this research is: "How does the relationship between network structure, network flows and the network phenomenon of influence affect the course of action that marketers should take when they engage with an online social network?" This dissertation will consequently not investigate other network phenomena such as governance, social capital, task complexities and interdependencies.

Twitter conversations constitute an ideal setting for this study because they exhibit the abovementioned characteristics of online social networks that contrast sharply with social networks that occur in the real world. Many of the lessons learned from these conversations may, however, be applicable to other social networking platforms, as well as to the real world itself. Furthermore, it is important for marketers to understand network structure, network flows and the impact that network flows and network structure have on the network phenomenon (influence) in a Twitter network. The results of the study proposed in this dissertation could consequently allow companies to optimize their marketing resources on Twitter. However, future further studies of network flows, network structure and network phenomena on other platforms could potentially verify that the findings of this dissertation are generalizable to other platforms.

1.3 Dissertation Outline

This dissertation consists of an introduction, a literature review that leads to a conceptual framework, a set of testable hypotheses, a discussion of research methods, a chapter that presents the results of the proposed study, and a chapter that draws conclusions from these results. The final chapter reviews the study's contributions and limitations. It also discusses theoretical and practical implications of the study and makes suggestions for further research.

1.3.1 Chapter 1 – Introduction

Chapter 1 familiarizes the reader with the dissertation topic. The first section describes the research problem; the second introduces the purpose and scope of this dissertation. Both sections argue that the study which this dissertation proposes should be performed. The third section presents an outline of the dissertation.

1.3.2 Chapter 2 – Literature Search

Chapter 2 reviews the literature that pertains to the research that is proposed in this dissertation. The first section summarizes the literature on network structure. It discusses previous attempts to explain holistic models of society. The second section presents prior insights into how information flows within a social network and into how these 'network flows' lead to a variety of observable phenomena within the social network. The third section bridges the gap between the literature on network structure and the literature on network flows. It also identifies the primary gap that this dissertation intends to address. Sections four and five respectively discuss the characteristics of network structure and network flows. Section six, brings forth the literature regarding the phenomenon of influence in a social network and how this influence is measured. Finally, section seven summarizes the research gaps that have been identified in the literature review, and the research questions that have been formulated based on the research gaps.

1.3.3 Chapter 3 – Research Framework, Scope and Hypothesis

Chapter 3 proposes a novel research framework that intends to overcome the shortcomings of extant theory. This research framework determines the scope of this dissertation. Chapter 3 subsequently identifies research hypotheses, which are based on the proposed theoretical framework. These hypotheses focus on the degree to which social network structure and information flow impact the network phenomenon of influence and each other.

1.3.4 Chapter 4 – Research Methods

Chapter 4 describes the research methods that will be used in my dissertation. This description includes discussions of the unit of analysis; the setting of the study; variables and measures; data collection; validity and reliability; and the data analyses that have been deployed in the study.

1.3.5 Chapter 5 – Analysis and Results

Chapter 5 of my dissertation details the results of the proposed study, as well as all statistical analyses.

1.3.6 Chapter 6 – Conclusions and Discussion

Chapter 6 draws conclusions from the results presented in chapter 5.

1.3.7 Chapter 7 – Contributions and Limitations

The final chapter identifies some of the study's limitations. It also reviews the study's contributions, discusses theoretical and practical implications of the study and makes suggestions for further research.

2. Literature Review

The management question that motivates this dissertation is: "How does the relationship between network structure, network flows and the network phenomenon of influence affect the course of action that marketers should take when they engage with an online social network?" As noted earlier, this dissertation covers network structure, network flows and network phenomenon of influence within social networks. A social network will be viewed from a graph theoretic point of view in terms of nodes and ties. A node represents an actor within a network and a tie represents a relationship between actors.

In the review of the academic literature that follows, I look at the prior research that has been done, and based on this prior research I identify gaps in knowledge that warrant further scientific study. From these gaps, I shall generate research questions for my dissertation. The major contributions of this dissertation will close the gaps in knowledge that I identify in this chapter, and address the research questions that they generate.

The following issues, which are addressed in section 2.1 and 2.2, are of particular interest to practicing technology managers:

1. What is the role of network structure and network flow in topological organization of a social network? (Section 2.1)

2. What role does network flow play in a social phenomenon within a social network? (Section 2.2)

My focus on network structure and network flows within a social network raises the following issues, which are addressed in sections 2.3, 2.4, 2.5 and 2.6:

3. How do network structures and network flows come together in a social network? (Section 2.3)
4. How do network constraints shape social theories? (Section 2.3)
5. How do online and real world social networks differ? (Section 2.4)
6. What types of structures can networks form? (Section 2.5)
7. What are the characteristics of network flows? (Section 2.6)

My proposed research also raises some broadly based issues pertaining to the phenomenon of influence within a network, which is addressed in sections 2.7:

8. How is the phenomenon of influence within a social network defined and measured? (Section 2.7)

In the following sections, I discuss each of the abovementioned issues one by one, and I identify the literature stream in which these issues have been discussed.

2.1 Topological Organization of Social Network

To understand the topological organization of social networks, I look at theories that take a broad, integrated view of social networks. These theories attempt to explain the topological organization of social systems by using analogies from the biological sciences, the physical sciences and systems science. They cover social phenomena pertaining to network structure (groups, societies, organizations, countries, etc...) and network flows (processes like communication, collaboration, reproduction, coordination, control, etc...), as well as the constraints that impact network structure and network flows (geographic distance/boundaries, land availability, etc...). They also attempt to build a unified theory of social systems that encompasses all the social phenomena that can be observed in a real world network.

An overview of the theories to be reviewed in this section is exhibited in Table 2.

Table 2: Theories of Social Organization--Literature Overview

Authors	Theory	Focus	Unit of Analysis	Processess
T. Parson (1951)	Theory of Social System	Analysis of social processes in relation to the structure of social systems and their variability.	Society	mechanisms of socialization, social roles, deviant behaviour, control.
N. Luhman (1986)	Autopoietic Theory	Social systems use self-reference as their mode for autopoietic reproduction to retain structure and functioning.	Society	communication
J.G. Miller (1978)	Living Systems Theory	Social systems work and deals with the notion of emergence and interaction. A system is defined as a set of interacting units with relationships among them.	Groups, Organizations, Communities, Society	Interacting units in a system; input-throughput-output processes
K.D. Bailey (1990)	Social Entropy Theory (SET)	SET uses the social system's internal entropy level as an indicator of system state. The chief goal of self-steering from the standpoint of SET is to keep system entropy levels from getting too high.	Society	Entropy of a system; global, mutable, and immutable variables.

The intent of this review is not to compare, contrast or assess the impact of these theories. Instead, I summarize these theories briefly, and I subsequently engage in a discussion that brings out their underlying commonalities. From these I develop a conceptual model that encompasses them all.

2.1.1 The Theory of Social Systems (Parson, 1951)

The theory of social systems was initially proposed by Talcott Parson in 1951. The author advocated a functionalist approach and hypothesized that all social systems perform the following basic functions:

1. *Adaptation:* acquiring sufficient resources
2. *Goal Attainment:* setting and achieving goals
3. *Integration:* maintaining coordination amongst sub-systems
4. *Latency:* creating, preserving and propagating systems distinct culture and values.

Parsons states that a social system comprises one of the three aspects of structuring a completely concrete system of social action (Parson, 1951). The other two are the "personality system" of the individual actors and the "cultural system," which is built into the individuals' actions (Parson, 1951). According to Parsons "a social system consists in a plurality of individual actors interacting with each other in a situation which has at least a physical or environmental aspect, actors who are motivated in terms of a tendency to the 'optimization of gratification' and whose relation to their situations, including each other, is defined and mediated in terms of a system of culturally structured and shared symbols" (Parson, 1951). He defines cultural systems as "symbolic element of the cultural tradition, ideas or beliefs, expressive symbols or value patterns so far as they are treated as situational objects by ego and are not 'internalized' as a constitutive elements of the structure of his personality" (Parson, 1951). These signs and symbol acquire a common meaning and serve as media of communication between actors. In order to define personality systems, Parsons states that 'action' is a process in an actor-situation system that has motivational significance to the individual actor because orientation of the action has a bearing on the attainment of gratification. The orientation of the action depends on the actor's personality structures, which are a function of the relation of the actor to his situation and the history of that relation (Parson, 1951). Parsons emphasizes that it is not theoretically possible to reduce any of the systems to a combination of other two. The fundamental building blocks of the theory of social systems, personality systems, and cultural systems are the same but the ways in which the conceptual material is built into theoretical structures are not the same.

Parsons approach of "structural functionalism" has been highly influential among sociologist trying to understand the shift from preindustrial societies to industrial societies, in particular complex relationships between different parts of society and the impact of social institution on individual behavior (Robertson, 1992). However, Parson's work was criticized for the absence of conflict and dysfunction (Mills, 2000, Gouldner), (Wrong, 1961). Despite these perceived flaws, Parson's theories of structural functionalism were credited with providing stimulus to the field of sociology (Turner, 1985, Merton, 1973).

In the theory of social systems, individuals do not act as the fundamental units of society. Instead, society is based the actions out of which personality systems and cultural systems are built. Therefore, the theory of social systems does not treat the personality systems and the cultural systems independently. Instead, it is concerned with how these components of the social system affect the overall structure of the social system and how it functions. The theory analyzes social processes in relation to the structure of social systems and their variability. It describes the mechanisms of socialization, patterns of orientation in social roles, tendencies toward deviant behavior and mechanisms of social control.

2.1.2 Autopoietic Theory (Luhmann, 1986)

Autopoietic theory has its origins in biological systems (Maturana and Varela, 1980). In this theory, Maturana and Varela define living systems as systems that use self-reference to reproduce. Every unit of offspring possesses a copy of its parents' genes. Throughout the interactions and transformations that the offspring encounter in their lifetimes, they continuously regenerate the network processes that have produced them. As a consequence, they retain a structure, which is similar to that of their parents, and they perform functions, which are similar to those that their parents performed.

Niklas Luhman extended autopoietic theory to social systems and suggested that social systems use communication as their mechanism for autopoietic reproduction (Luhmann, 1986). Communications are not living units; they are not conscious units; and they are not actions. A unit of communication consists of a synthesis of three components: information, utterance and understanding (including misunderstanding).

In essence, every actor within the social system has to make three choices: 1) whether to accept or reject information; 2) understand (or perhaps misunderstand) the information; and 3) and propagate it to other actors. The synthesis that results in communication is produced by the network in which the communication takes place; it is not derived from some kind of inherent power of consciousness or from the inherent quality of the information. In addition, the synthesis of information, utterance and understanding cannot be preprogrammed by language. It has to be recreated from situation to situation by referring to previous communications and to the possibility of further communications. In every situation, communication is restricted by the actual event, requiring self-reference. Furthermore, information, utterance and understanding cannot reside independently in a system; they are inherently co-created.

Autopoietic theory is based upon the following properties of communication:

- Communication is atomic. The elementary, indecomposable units of the system are communications of minimal size. However, this minimal size is context specific—it cannot be determined independently of the system.
- An elementary unit of communication has a minimal meaning, which still can be negated. This minimal meaning is necessary for reference in further communication.
- The social system also includes further communication or the prospect of further communication. Further communication can very well separate pieces of information, utterances and understandings and discuss them separately, but this still would presuppose their synthesis in previous communication.

- Communication includes understanding as a necessary part of the unity of its operation. It does not include the acceptance of its content. It is not the function of communication to produce a consensus as the favored state of mind.
- Communication always results in an open situation of either acceptance or rejection. It reproduces situations with a specified and enforced choice. Such situations are not possible without communication; they do not occur as natural happenings. Only communication itself is able to reach a point at which the meaning of the communication is either accepted or rejected. This bifurcation results in a reduction of complexity and, by this very fact, an enforcement of selection. Automatically, the selection of further communication is either an acceptance or rejection of previous communication or a visible avoidance or an adjournment of the issue.
- Whatever its content and intention, communication reacts within the framework of enforced choice. To take one course is not to take the other. This highly artificial condition structures the self-reference of the system; it makes it unavoidable to take other communications of the same system into account, and every communication renews the same condition within a varied context.

If a social system were set up to produce consensus, it would soon come to an end. It would never produce and reproduce to form a society. In fact, however, social systems are designed to reproduce themselves by submitting themselves to self-reproduced selectivity. Only this arrangement makes the evolution of social systems possible.

Autopoietic theory has been further reviewed by many researchers in the field of organizational theory (Mingers, 2003) and information systems (e.g., Baca. et. al., 2010, Malekovic and Schatten, 2008). For example, Mingers (2003), who evaluated Luhmann's theory from an organizational perspective states: “Social systems are networks of communication that produce further communication and only communication” (Mingers, 2003), pp. 104-105). Therefore, they are autopoietic. In addition, information systems are a critical subsystem of both social systems and organizations (Brumec, 1997), which raises the issue whether information systems autopoietic as well (Bača et al., 2007; Maleković and Schatten, 2008). Information systems can be viewed as a set of relations between communicative events that reproduce new communicative events based on previous (stored) communication. The organization of such systems consists of the relations between communicative events described through their semantics (meaning) and the means that are used to produce communication (Maleković and Schatten). According to Baca et al.(2007), Autopoiesis in the context of information systems denotes the ability of an information system to continuously adapt to the needs of its current users and also to keep all the characteristics that make it unique and recognizable as an information system (Bača et al., 2007). This tends to be an attribute of organizational and social systems.

2.1.3 Living Systems Theory (Miller, 1978)

Living systems theory is a general theory about how living systems work. It deals with the notion of emergence and interaction. A system is defined as a set of interacting units and the relationships among them. Miller’s model of living systems constitutes a hierarchy that consists of the following eight levels:

- Cells: the basic building block of life
- Organs: the principle components are cells, organized in simple, multi-cellular systems.
- Organisms: there are three kinds of organisms: fungi, plants and animals. Each has distinctive cells, tissues and body plans and carries out life processes differently.
- Groups: these contain two or more organisms and their relationships.
- Organizations: these involve one of more groups with their own control systems for doing work.
- Communities: these include individual persons and groups, as well as groups which are formed and are responsible for governing or providing services to them.
- Societies: these are loose associations of communities, with systematic relationships between and among them.
- Supranational systems: organizations of societies with a supra-ordinate system of influence and control.

The properties (behavior) of a system as a whole emerge from the interaction between the components that comprise the system. Regardless of their complexity, they each depend upon the same essential twenty subsystems that perform specific processes, in order to survive and to continue the propagation of their species or types beyond a single generation. The twenty subsystems and the processes of all living systems are arranged by input-throughput-output processes. Some of these processes deal with material and energy for the metabolic processes of the system. Other subsystems process information for the coordination, guidance and control of the system. Some subsystems and their processes are concerned with both. They are as follows:

Subsystems/processes that take place in the Systems Input Stage

- *Input transducer:* brings information into the system
- *Ingestor:* brings material-energy into the system

Subsystems/processes which take place in the Systems Throughput Stage

A. Information processes:

- *Internal transducer:* receives and converts information brought into system channel
- *Net:* distributes information throughout the system
- *Decoder:* prepares information for use by the system
- *Timer:* maintains the appropriate spatial/temporal relationships
- *Associator:* maintains appropriate relationships between information sources
- *Memory:* stores information for system use
- *Decider:* makes decisions about various system operations
- *Encoder:* converts information to needed and usable form

B. Material-Energy processes:

- *Reproducer:* with information, carries on reproductive function
- *Boundary:* with information, protects system from outside influences
- *Distributor:* distributes material-energy for use throughout the system

- *Converter:* converts material-energy into suitable form for use by the system
- *Producer:* synthesizes material-energy for use within the system
- *Storage:* stores material-energy used by the system
- *Motor:* handles mobility of various parts of the system
- *Supporter:* provides physical support to the system

Subsystems/processes which take place in the Systems Output Stage

- *Output transducer:* handles information output of the system
- *Extruder:* handles material-energy discharged by the system

Living Systems Theory has been used to explain the behavior of some large industrial corporations (Duncan, 1972); in general analyses of organizations (Lichtman and Hunt, 1971, Reese, 1972, Noell, 1974); for explaining the pathologies of organizations (Cummings and DeCotiis, 1973); and in studies of accounting (Swanson and Miller, 1989), and management accounting (Weekes, 1984). Other studies assess the effectiveness of a hospital (Merker and Lusher, 1987) and a metropolitan transportation utility (Bryant and Merker, 1987). The largest application of Living systems theory has been a study of the performance of 41 US Army battalions (Ruscoe et al., 1985). All these studies revealed important relationships between characteristics of matter-energy, information processing and organizational effectiveness.

2.1.4 Social Entropy Theory (Bailey, 1990)

Social Entropy Theory (SET) uses the system's internal entropy level as an indicator of system state, where entropy is a measure of system disorder. Entropy can show up in the system as various indicators of system disorder, such as faulty communication, errors, inadequate supply levels, lack of energy, resources, or even clutter. If entropy gets too high, the functionality of the system is impaired or even threatened. From the standpoint of SET, entropy can best be properly managed by a self-steering process, where the chief goal of self-steering is to keep system entropy levels from getting too high.

SET presents six structural dimensions that are salient for all social systems. These are, respectively: population size (P), information (I), level of living of the social system (L), organization (O), technology (T), and spatial area or territory (S). In conjunction, these dimensions are known by the acronyms PILOTS or IPLOTS. Energy has been assumed in this model as being present in the territory or spatial area (S), but that has not been clearly specified. As energy plays an extremely important role in self-steering, it is helpful at this point to add energy (E) specifically to the model to attain EIPLOTS.

SET facilitates the goal of analyzing self-steering through its distinction between characteristics or variables that are global, mutable or immutable. Global variables are macro-variables that are defined only for the society as a whole; they cannot be defined for individuals. These include such variables as total wealth of the nation (L), the social-class structure (O), the occupational division of labor (O), the total land area of the territory (S), etc.

The polar opposites of the global variables are the immutable variables, which are micro-variables that describe the characteristics of individuals. Immutables are properties that are only defined for individual persons and cannot be defined for the society as a whole. Immutable variables are generally present from birth, and are thus similar to “ascribed” variables. Immutables generally cannot be changed (or at least not without extreme difficulty). Examples of common immutables are an individual’s birth date, skin color, height, eye color, sex, etc.

It is clear that global variables are highly relevant for the process of self-steering, as they provide a context which facilitates or constrains the steering process. A fortuitous set of global characteristics can make self-steering quite easy. In contrast, an unfortunate array of globals can make self-steering very difficult. It is less clear how immutables affect self-steering, but they certainly do. Aside from such activities as voting, or participating in various holiday festivities or rituals, the self-steering of a social system is not generally accomplished by all members of the society, but only a subset of the population. These individuals are selected by a variety of means, but often their selection is not random. Rather, persons who steer societies (either individually or collectively) tend to be represented non-randomly on key immutable variables such as race, sex, age, etc.

Between the globals and immutables in SET are the mutable variables. These intermediate variables are true micro-macro links, as they can serve either as individual or as societal characteristics. The mutables at the individual level are similar to "achieved" variables. These are the individual counterparts of the EIPLOTS dimensions. For example, in addition to his or her immutable variables such as age or sex, each individual has mutable characteristics such as his or her educational level (I), income (L), real estate ownership (S), and access to a computer (T). These mutables, which are exhibited by all persons in the society, may be aggregated to form mutable distributions, such as the average income of the society (L), the average educational level of the society (I), etc. Notice that these distributions are not globals, but they are aggregated macro properties of society, and serve to link the individual to the society. Depending on their specific levels in a given society, the mutable distributions can also serve to either facilitate or hinder the process or self-steering in the social system.

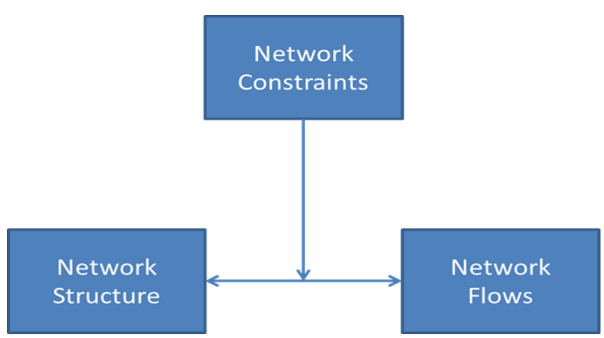
Swanson, Bailey, and Miller (1997) discuss a progression of entropy-related measures in systems ranging from physical through biological to social, with an emphasis on social systems (Swanson et al., 1997). This progression is discussed in the context of Living Systems Theory, as developed by Miller (Miller, 1978), and integrates that theory with Social Entropy Theory (Miller, 1978), as developed by Bailey (Bailey, 1990), and Macro Accounting Theory as developed by Swanson (Swanson, 1993). This integration is important for at least two reasons. The first reason is that the domains of the theories being integrated are contained progressively each in the other. The very broad domain of Living Systems Theory concerns all living systems existing in space-time and thus contains the domain of the more narrowly focused Social Entropy Theory, which in turn contains the domain of Macro Accounting Theory (which concerns economic systems within social systems).

2.1.5 Commonalities among Theories of Topological Organization of Social Networks

The theories discussed above explain the organization of network structure (groups, societies, organizations, countries, etc...) through network flows (processes like communication, collaboration, reproduction, coordination, control, etc...), as well as the constraints that impact network structure and network flow (geographic distance/boundaries, land availability etc...). For example, in the theory of social systems, Parson deals with the analysis of social processes in relation to the structure of social systems and their variability. He states that all social systems perform certain basic functions (adaptation, goal attainment, integration and latency) (Parson, 1951). In autopoietic theory, Luhmann talks about how human systems use communication as a medium to structure themselves through the process of self-reference (Luhmann, 1986). In living systems theory, Miller states a general theory about how living systems work and how they organize themselves through emergence and interaction (Miller, 1978). In social entropy theory, Bailey uses internal entropy as an indicator of the state of a system (Bailey, 1990) and its organization. Though these theories have their origins in different branches of science and constitute different elaborations of the organization of networks, they all state that some kind of information is transferred within the network which guides the topological organization of the network structure through which information flows in the real world.

The conceptual model emerging from the literature in this section is shown below. Figure1 provides a conceptual framework for the theories of topological organization. It depicts a linkage between network structure and network flows. The relationship between network structure and network flows is subject to constraints on the network. Theories of social organization are thus well suited to explain organizations as a whole. However, they do not treat individual network phenomena that are observed within social systems, such as trust and reciprocities. Theories of social organization are thus inherently incomplete.

Figure 1: Topological Organization of Network - Conceptual Model



2.2 Network Flow in Social Phenomena

I start this section by defining social phenomena. Social phenomena include "all behavior that influences or is influenced by organisms sufficiently alive to respond to one another" (John, 1925). Theories of social phenomena show that social phenomena within a social network are caused by network flows. These theories, in contrast to theories of social organization, do not address the organization of social network instead they try to explain a specific social phenomenon within the broader context of social networks that occur in the real world. An overview of theories to be explored is shown below in Table 3. Theories chosen in this section specifically focus on the role of network flows in social phenomena. These theories are reviewed with the intent of showing the underlying commonalities, from which I derive a conceptual model that encompasses them all. I do not to compare, contrast or assess the impact of these theories. The commonalities and the conceptual model that arises will be discussed after briefly summarizing the theories below:

Table 3: Theories of Social Phenomena--Literature Overview

Authors	Theory	Focus	Unit of Analysis	Process
Rogers(2003)	Diffusion of Innovations	Adoption of innovative products through word-of mouth.	Society	Diffusion of information
Granovetter (1973)	The Strength of Weak Ties	Weak ties are a source of novel information	Society	Access to novel information
Burt (1976 & 1992)	Structural Holes	Social capital is created in a network in which people can broker connections between disconnected network segments	Society	Brokerage
Coleman (1988)	Closure Theory of Social Capital	Social capital is created by a network in which people can control functioning of the network to achieve goals.	Society	Control and co-ordination
Watts and Strogatz (1998)	Small World Theory	Adding even a small number of random ties to a heavily clustered network could radically reduce distances among nodes. The reason was that many of these random ties would be between clusters, which formed bridges.	Society	Ease of access
Putnam (1995)	Social Capital	Social capital as a feature of social organization, such as trust, norms and networks that can improve efficiency of society by facilitating coordinated action.	Society, Country	Efficiency
Bourdieu (1977)	Social Capital	Resource that resulted from social structure	Society	Role, position
Allen (1977)	na	Communication tends to increase as a function of spatial proximity in an organizational setting	Organization	Efficiency
Powell (1990)	na	Reciprocal patterns of communication and exchange are alternatives to hierarchically or market-based governance structures.	Organization	Efficiency, Co-ordination
Uzzi (1997)	na	Embeddedness in an intra-firm network promotes economies of time, integrative agreements, Pareto improvements in allocative efficiency, and complex adaptation	Organization	Economic benefit, Efficiency
Podolony (1994)	na	Organizations overcome problems of market uncertainty by adopting a principle of exclusivity in selecting exchange partners	Organization	Economic benefit, Trust

2.2.1 Diffusion of Innovations (Rogers, 2003)

In his book *Diffusion of Innovations*, Everett Rogers describes the process of adoption of new innovations. He emphasizes the role of interpersonal communication in the adoption of innovations. According to Rogers, diffusion is “the process in which an innovation is communicated through certain channels over time among the members of a social system” (p. 5), the key components in this definition being innovation, communication channel, time and social system.

For Rogers, “diffusion is a very social process that involves interpersonal communication relationships” (p. 19). He defines communication as “a process in which participants create and share information with one another, in order to reach a mutual understanding” (p. 5). This communication occurs through channels between sources. Rogers defined a source as “an individual or an institution that originates the message and an interpersonal channel consists of two-way communication between two or more individuals through which the message gets to the receiver” (p. 204). These interpersonal channels are powerful enough to create or change strong attitudes held by an individual.

Rogers (2003) defined the social system as “a set of interrelated units engaged in joint problem solving to accomplish a common goal” (p. 23). Since diffusion of innovations takes place in the social system, it is influenced by the social structure of the social system. For Rogers (2003), structure is “the patterned arrangements of the units in a system” (p. 24). He further claimed that the nature of the social system affects individuals’ innovativeness, which is the main criterion for categorizing adopters into innovators, early adopters, early majority, late majority and laggards (p. 22).

Although Rogers’s theory has influenced innovation studies in various fields over the last several decades, subsequent empirical research challenges the notion of an idealized, linear 'technology push-market pull' dichotomy first proposed in his work (Baskerville and Pries-Heje, 2001, Dosi, 1982). In later work, even Rogers broke away from the linear orientation of his original project. The author suggests that his original framework might be augmented through the use of complex adaptive systems, resulting in a hybrid framework to explain the diffusion of innovations (Rogers et al., 2005).

2.2.2 The Strength of Weak Ties (Granovetter, 1973)

Granovetter asserts that acquaintances (weak ties) are less likely to be socially involved with one another than close friends (strong ties). Thus the set of people made up of any individual and his or her acquaintances comprises a low-density network (one in which many of the possible relational lines are absent), whereas the set consisting of the same individual and his or her close friends will be densely knit (many of the possible lines are present).

The overall social structural picture suggested by this argument can be seen by considering the situation of some arbitrarily selected individual. This individual will have a collection of close friends, most of which are in touch with one another, *i.e.*, a densely knit clump of social structure. Moreover, the individual will have a collection of acquaintances, few of whom know one another. Each of these acquaintances, however, is likely to have close friends in his own right and therefore to be enmeshed in a closely knit clump of social structure, but one different from the individual's. The weak tie between the individual and his acquaintance, therefore, becomes not merely a trivial acquaintance tie but rather a crucial bridge between the two densely knit clumps of close friends. To the extent that the assertion of the previous paragraph is correct, these clumps would not, in fact, be connected to one another at all were it not for the existence of weak ties. Thus, individuals with few weak ties will be deprived of information from distant parts of the social system and will be confined to the provincial news and views of their close friends. This deprivation will not only insulate them from the latest ideas and fashions but may put them in a disadvantaged position in the labor market, where advancement can depend on knowing about appropriate job openings at just the right time.

2.2.3 Structural Holes (Burt, 1976, Burt, 1992)

Burt, through his structural holes argument, suggests that social capital is created by a network in which people can broker connections between disconnected network segments. He views society as a network in which people or groups of people can exchange all types of goods and ideas in order to achieve their goals. Some of these people or groups of people achieve better returns in lieu of their efforts than others do. For example, some people earn a better remuneration, some become more important and some lead more important projects. The human capital explanation of this inequity is that people who do better are more able people, more intelligent, more articulate, more attractive or more skilled. Social capital is a contextual complement of human capital, suggesting that people who are better connected should be more successful. Thus, holding a specific position in the network structure is associated with a certain level of social capital.

Burt defines structural holes as weaker connections between two groups in a social structure. These holes in the structure create competitive advantage for the people who have relationship that span these holes. This does not mean that the people in each group are unaware of the existence of the other group. Instead, the people in each group are more focused on their own activities and do not participate in the activities of the other group. Thus, structural holes are an opportunity to broker and control the flow of information across groups.

2.2.4 Closure Theory of Social Capital (Coleman, 1988)

Coleman's network closure argument suggests that networks in which everybody is connected to everybody and no one can escape notice of the other (in other words dense networks) are the source of social capital. He defines social capital as a resource for action within a social structure (Coleman, 1988). Network closure does two things for people in a network. First, it affects access to information. Second, network closure facilitates collective sanctions, and fear of sanctions for behavior that is out of the norm fosters conformity. It also reinforces trust between those who already conform.

Coleman's study of high school students (Coleman, 1988) illustrates his argument. He argues that closure explains why some students are more likely to drop out of school. When adults in a child's life are more connected to each other the closure argument predicts that norms, trust and consensus on sanctions are more likely among adults. This suggests adults can more effectively enforce their interest in the child completing his or her education. Coleman presents three bits of evidence, which show that children living in closed networks are less likely to drop out from school. They are as follows:

1. Children living in a family of two parents with few children are less likely to drop out of school (two parents living together can collaborate more effectively to supervise a child's education than two parents living apart).
2. Children who have lived in the same neighborhood are less likely to drop out of school (parents, teachers and are more likely to know each other and collaborate on a child's education than parents who have moved in a new neighborhood).
3. Children in religious schools (e.g., Catholic school) are less likely to drop out of school (parents, teachers and parents of other students are more likely to know each other and collaborate in the child's education).

2.2.5 Small World Theory

Another well-known area of network theorizing is small world theory. In the 1950s and 60s, a stream of mathematical research sought to explain coincidences of mutual acquaintanceship (Rapoport and Horvath, 1961, Sola Pool and Kochen, 1978–1979). The basic thrust of the research was to show that societies were probably much more close-knit than popularly believed. A field experiment by Milgram (Milgram, 1967, Travers and Milgram, 1969) supported this theory, finding that the paths that link any two random Americans were incredibly short. Restarting this stream of research twenty years later, Watts and Strogatz (Watts and Strogatz, 1998) asked how human networks could have such short average distances, given that human networks were so clustered, a property which was known to lengthen network distances (Rapoport and Horvath, 1961). The answer, Watts and Strogatz showed, was simple: adding even a small number of random ties to a heavily clustered network could radically reduce distances among nodes. The reason was that many of these random ties would be between clusters, which formed bridges.

2.2.6. Other Theories of Social Phenomena

There are many more theories of social phenomena. For example, Putnam (Putnam, 1995) described social capital as feature of social organization, such as trust, norms and networks that can improve efficiency of society by facilitating coordinated action. Bourdieu (Bourdieu, 1986) defines social capital as "the aggregate of the actual or potential resources which are linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition." Allen (Allen, 1977) found that communication tends to increase as a function of spatial proximity in an organizational setting. Powell (Powell, 1990) found that network forms of organization with reciprocal patterns of communication and exchange are alternatives to hierarchically or market-based governance structures. They are more suited to describing companies involved in an intricate latticework of collaborative ventures with other firms over extended periods of time. Uzzi (Uzzi, 1997) found that embeddedness in an intra-firm network promotes economies of time, integrative agreements, Pareto improvements in allocative efficiency, and complex adaptation. However, embeddedness also insulates firms within a network from information that exists beyond their network, making the firm vulnerable to exogenous shocks that can derail the firm's economic performance. Podolony (Podolny, 1993) proposes that organizations overcome problems of market uncertainty by adopting a principle of exclusivity in selecting exchange partners. His research suggests that organizations that operate in an environment of high market uncertainty tend to engage in exchange relations with organizations with whom they have transacted in the past or organizations with similar status.

2.2.7 Commonalities among Theories of Social Phenomena

The theories of social phenomena described above identify a social phenomenon within a network and explain the phenomenon within the broader context of a social network that exists in the real world. These theories do not attempt to explain the organization of the social network. In all instances, the social phenomena under observation within a network structure are caused by network flow. For example, Rogers talks about the importance of interpersonal communication within a social system for diffusion of innovation (Rogers, 2003). Granovetter (Granovetter, 1973) suggests that weak ties are the sources of new information that flows into the network from the outside. In his structural holes theory, Burt talks about competitive advantage being derived by creating network flows between two different cliques. This suggests competitive advantage is obtained from being on the fringe of a network (Burt, 1976). In contradiction to Burt, Coleman talks about the advantage of being in the middle of network flows within a clique and the risks of being on the fringe of a network (Coleman, 1988). Small world theory shows that creating random ties within a heavily clustered network reduces the distance between the people in the network. This improves network flow, which results in better communication between the members of the network (Watts and Strogatz, 1998). In summary, theories of social phenomena are different elaborations of the impact of network flow on social phenomena.

Figure 2: Theories of Social Phenomena - Conceptual Model

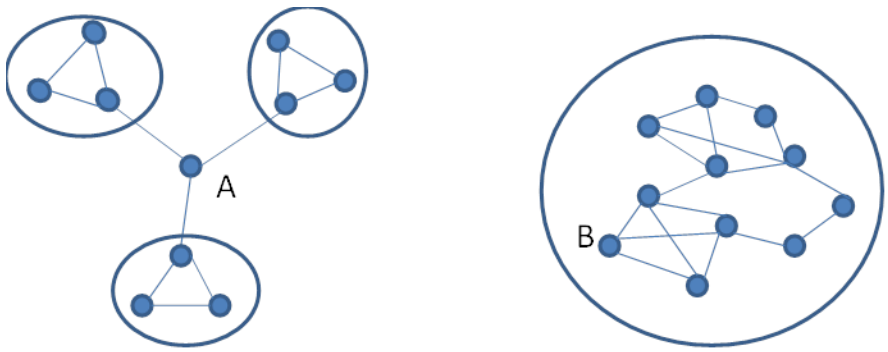
Figure 2 illustrates the conceptual model that underlies all theories of social phenomena. A network phenomenon is derived from network flows. In other words, the paths that information takes as it spreads throughout a network and the distance between the sources and the recipients of information give rise to observable network phenomena in real world networks. However, theories of social phenomena do not treat structural factors. Thus they will have difficulty explaining the organization of a network and its impact on the network's overall performance.

2.3 Integrated Network Theory and Perspective on Network Constraints

Sections 2.1 and 2.2 illustrate that network constraints play an important role in directing network flows. In autopoietic theory, Luhmann talks about communication acting as a constraint on the process of self-reference, and thereby being a constraint on network organization (Luhmann, 1986). In social entropy theory, Bailey uses internal entropy as an indicator of the state of a system (Bailey, 1990) and as a constraint on its organization. Similarly, in theories of social phenomena, Allen found communication to be a function of spatial proximity (Allen, 1977). Rogers found that in interpersonal channels, the communication may have a characteristic of homophily, that is, “the degree to which two or more individuals who interact are similar in certain attributes, such as beliefs, education, socioeconomic, status, and the like,” but the diffusion of innovations requires at least some degree of heterophily, which is “the degree to which two or more individuals who interact are different in certain attributes” (Rogers, 2003). Thus homophily and heterophily can act as constraints on network flow.

To better understand the role of network constraints, I look at Atkin’s seminal work in which he referred to network structure and network flow as backcloth and traffic. The backcloth consists of an underlying infrastructure that enables and constrains the traffic, and the traffic consists of what flows through the network, such as information (Atkin, 1974). According to Borgatti and Foster, most of the differences between theories of topological organization of networks and theories of social phenomena are elaborations of the same theory (Borgatti and Foster, 2003). They look at the network constraints from a contextual perceptive of their research.

Figure 3: Networks with Different Structures but the Same Number of Nodes and Ties



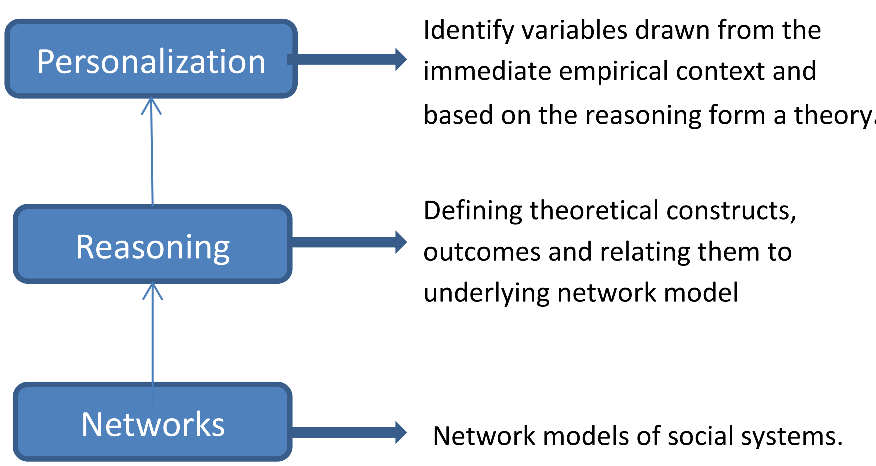
To illustrate this point, Borgatti and Halgin (Borgatti and Halgin, 2011) provide the example in Figure 3. The authors suggest that Burt's theory may look different from Granovetter's, but the differences are largely in language and focus. In Burt's language, A in figure 3 has more structural holes than B, which means A has more non-redundant ties. In Granovetter's language, A has more bridges than B. But whether we call them non-redundant ties or bridges, the concept is the same, and so are the consequences: more novel information. Where Granovetter and Burt differ is that Granovetter further argues that a tie's strength determines whether it will serve as a bridge. Burt does not disagree and even provides empirical evidence that bridging ties are weaker in that they are more subject to decay (Burt, 1992, Burt, 2005). However, Burt sees tie strength as a mere "correlate" of the underlying principle, which is non-redundancy (Burt, 1992). Thus, the difference between these theories comes down to either preferring the distal cause (strength of ties), as Granovetter does, or the proximal cause (bridging ties), as

Burt does. The former yields an appealingly ironic and counterintuitive story line, while the latter "captures the causal agent directly and thus provides a stronger foundation for theory" (Burt, 1992).

Similarly, Burt (2005) points out that the conflict between Burt's structural holes theory (Burt, 1992) and Coleman's closure theory (Coleman, 1988) is more apparent than real, as both assume that ties constrain relationships in a network (Burt, 2005). The difference is simply that in Coleman's educational setting, constraint is good, and in Burt's corporate setting, constraint is typically bad. It is really only the orientation of the social capital concept that creates contradiction.

Based on the above commonalities, Borgatti and Kidwell (Borgatti and Kidwell, 2011) proposed a three layer model to explain the social theory building process as follows:

Figure 4 : Social Theory Building Process



The bottom layer consists of a very simple model of how social systems work, which is essentially that they are networks through which information (or any resource) flows from node to node along network paths consisting of ties that are interlocked through shared endpoints. Therefore the bottom layer is characterized by fundamental network properties such as centrality and centralization (Wasserman and Faust, 1994). Centrality and centralization are explained in section 4.3.4.1 and section 4.3.4.2.

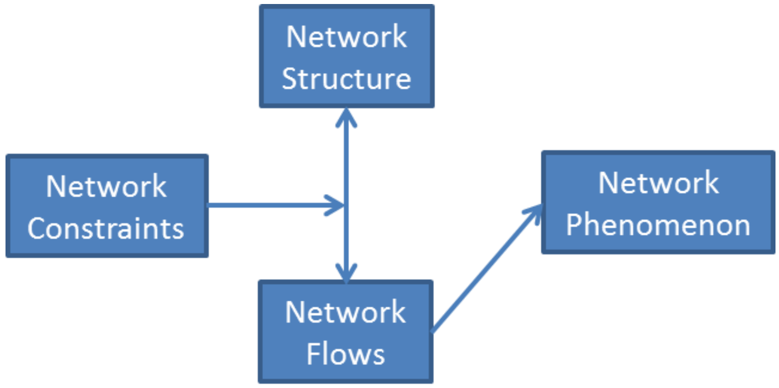
Scholars impose paradigmatic constraints upon the fundamental attributes of the network, in order to provide a theoretical explanation of the underlying phenomena. They define theoretical constructs and outcomes, from which they derive theorems about the underlying network structure and network flows using their particular line of reasoning. Theory, at this intermediate level (middle layer in fig 4.), consists of relating fundamental network properties (such as betweenness centrality) to outcomes in the same conceptual universe (such as frequency and time of first arrival of something flowing through the network). These outcomes may thus be influenced by the paradigmatic constraints that have been imposed by the scholar.

The top layer of figure 4 provides an empirical context to the theories that emerge from the middle layer. It can be viewed as a “personalization” of the theory, because the empirical context under which the scholar has formulated his/her theory may vary. For example, Granovetter and Burt both look at non-redundant ties. Granovetter (Granovetter, 1973) investigates social networks that pertain to finding a job. In that context, he focuses on the strength of ties and how they act as an antecedent to novel network flows. Burt (1992), on the other hand, studies social capital in a corporate setting. By focusing on structural holes he is able to connect information flows to personal creativity and the production of value (Burt, 1992).

The most important point conveyed by the model in figure 4 is that the information flowing through a network provides a conceptual universe, within which we can impose conceptual constraints like connectedness and relate them to other properties like the probability of receiving information. Theoretical constructs that pertain to a particular conceptual universe are thus true only within the contextual model of that universe; they may be false in a different context (Borgatti and Kidwell, 2011). These constructs are derivations of the particular model under consideration, yet, as theories of network phenomena show, they are widely misperceived to be unconnected to the theory (Borgatti and Kidwell, 2011). In addition, theoretical constructs that pertain to a particular conceptual universe cannot be considered generic measures or generic techniques like regression, which can be divorced from an underlying model of how things work (Borgatti, 2005).

Figure 5 below illustrates the integrated conceptual model of network organization and network phenomena, which has emerged from the literature so far. It depicts a linkage between network structure and network flows. The relationship between network structure and network flows is subject to constraints on the network. Network flows cause the network phenomenon.

Figure 5: Integrated Conceptual Model



2.4 Differences between Real-World and Online Social Networks

All cases that have been described until now occur in the real world. Network flows in the real world take place between the seeker of information and the source of information, and all network flows transpire within existing social relationships (Bristor, 1989, Duhan et al., 1997, Money et al., 1998). Individuals in a strong relationship tend to interact more frequently and exchange more information, compared to those in a weak relationship (Brown and Reingen, 1987). Because interactions only happen between people who have social relationships, an individual's relationship network and his/her interaction network were considered to be one and the same (Burt, 1987). In general, researchers in these studies observe a stable network structure in which the correlations among friends could be higher than those among strangers. Hence, the task is to determine whether the difference in correlation is due to social interaction or something else (Moffitt, 2001).

To my knowledge, *no study has been undertaken, which suggests that extant social theories developed for the real world networks can be applied to social networks that are formed online.* Consequently, it cannot be said that real world constraints such as connectedness and distance have any significant impact on the behavior online social networks (Borgatti and Cross, 2003). Instead, online social networks may be most affected by the topological organization of network structure (e.g., “scale free” (Barabási and Albert, 1999), “assortativity” (Newman, 2002) and “small world” (Watts and Strogatz, 1998) or by various attributes of network flows (e.g., paths, geodesics) (Borgatti, 2005), which extant theory of social networks does not really consider (Borgatti and Cross, 2003). These topics, which have been mentioned in section 1.1, are covered in sections 2.5 and 2.6.

As mentioned in section 1.1, online social networks are different from real world social networks. We know from observation of practicing firms (Wiertz et al., 2010), that online social networks are an emergent phenomenon (in the sense of (Drazin and Sandelands, 1992, Sandelands and Drazin, 1989). Unlike real world social networks, not all network flows generated in an online social network can be attributed to social relationships (Pei et al., 2014). People in online social networks may interact virtually with people with whom they share common interest. However, this does not necessarily mean that they are connected with each other. For example, in hashtag communities on Twitter converse on a particular topic. This does not mean that they are “friends” with or “followers” of each other (Weng et al., 2012). Also, the ability to conduct a search on online social networks (Watts et al., 2002, Adamic and Adar, 2005) makes the network structure and the network flows, which result from the interaction that follows that search, highly dynamic (Dodds et al., 2003).

The nascent body of research on online social networks treats network structure (Centola, 2010, Chomutare et al., 2014, Sasidharan et al., 2011), network flow (Hodas and Lerman, 2014, Burt et al., 2013, Aral and Walker, 2011, Dellarocas et al., 2013, Hodas and Lerman, 2012) and network phenomena (Aral and Walker, 2012, Pei et al., 2014, Khammash and Griffiths, 2011, Muchnik et al., 2013a, Muchnik et al., 2013b) separately. Studies that characterize the mechanisms through which network structure, network flow and network phenomena collectively emerge and operate are woefully lacking (Aral et al., 2013). We cannot even identify the loci of influence within a social network reliably.

Several studies that analyze interactions between users of online social networks have been published to date. Flicker data was used to study user interaction about photos that have been posted (Cha et al., 2009, Valafar et al.). Twitter data has been used to study how information diffuses online (Cha et al., 2010, Kwak et al., 2010). Facebook data has been used to study the time-varying dynamics of user interactions (Viswanath et al., 2009). The general consensus of this growing body of research is that a network interaction graph represents relationships that are meaningful online, whereas a graph of all social connections does not (Wilson et al., 2012). Only a fraction of all connections represent active connections, as interactions are not evenly distributed across a user's connected network. In an interaction graph, a link between two actors in an interaction network exists, only if they have interacted, irrespective of whether they are connected or not (Wilson et al., 2012). This means that interactions between actors that are not socially connected can occur. In addition, interaction graphs demonstrate significantly different properties from connected graphs. For example, interaction graphs exhibit larger graph diameters and lower clustering coefficients than connected graphs (Wilson et al., 2012).[15] It has also been observed that the selection of influential nodes and their effective range of influence change when interactivity is taken into account (Chen et al., 2009).

In summary, online social networks and real world social networks differ from each other in following ways:

1. Social networks online are significantly larger than the real world social networks.
2. Real world social networks are non-emergent whereas online social networks are emergent.
3. Network structures in real world social networks are static whereas online social networks have dynamic network structure.
4. Networks flows generated in real world social networks transpire within social relationships. Therefore, their connected network and their interactive network are the same. By contrast, in online social networks the connected network and the interactive network differ significantly.

[15] These properties of networks are defined in section 4.4 of this dissertation.

This section has established that, to date, no study has shown that social network theories from the real world directly apply to online social networks, and inherent differences between real-world and online social networks have been identified. In addition, studies that characterize the mechanisms through which network structure, network flow and network phenomena collectively emerge and operate in online social networks are woefully lacking. The ***primary research gap*** of this dissertation can thus be stated as follows: *no useful behavioral theory of online social networks, which integrates network structure, network flow and network phenomena, exists.* The behavior of online social networks has not really been characterized, and it definitely cannot be predicted.

2.5 Network Structure Topologies

Network structures have been widely studied in various disciplines of science (Pastor-Satorras and Vespignani, 2004, Westerberg and Wennergren, 2003, Keeling, 2005, Watts and Strogatz, 1998). Biological networks (Keeling, 2005), neural networks (Hopfield and Herz, 1995) and the World Wide Web (Pastor-Satorras and Vespignani, 2004) constitute examples of network structures that have been studied. The availability of large databases has allowed the study of the topology of interactions in variety of systems as diverse as communication systems to biological systems (Pastor-Satorras and Vespignani, 2004, Westerberg and Wennergren, 2003, Keeling, 2005). The main outcome of this activity has been to reveal that, despite the inherent differences, most of the real networks are characterized by the same topological properties, such as relatively small characteristic path lengths and high clustering coefficients (Watts and Strogatz, 1998)[16]. All these features make real networks radically different from regular lattices and random graphs, the standard models studied in mathematical graph theory (Watts, 1999).

The most important topological properties of networks discussed in the literature are scale free and small world properties of networks (Klemm and Eguiluz, 2002) which are described below.

2.5.1 Scale-free Networks

Networks that grow by attaching new nodes to existing nodes (by adding one tie only) form trees. They have no cycles (Figure 6 (a)) If, in this process, new nodes attach preferentially to existing nodes with a large number of ties, then the result is a scale-free network (Albert and Barabasi, 2000). Scale-free networks are distinguished by two characteristics. First, they are highly clustered (Barabási and Bonabeau, 2003); if two nodes share a common neighbor, it is likely the two are themselves adjacent. Second, the node degrees are distributed according to a power law (Barabási and Albert, 1999).

In Scale Free networks, the distribution of different network parameters acts in an exponential fashion (Figure 6(b)). The most interesting of these parameters is the Out Degree (Goh et al., 2002)—it measures the distribution of connections from each node

[16] These network properties are defined in section 4.4.

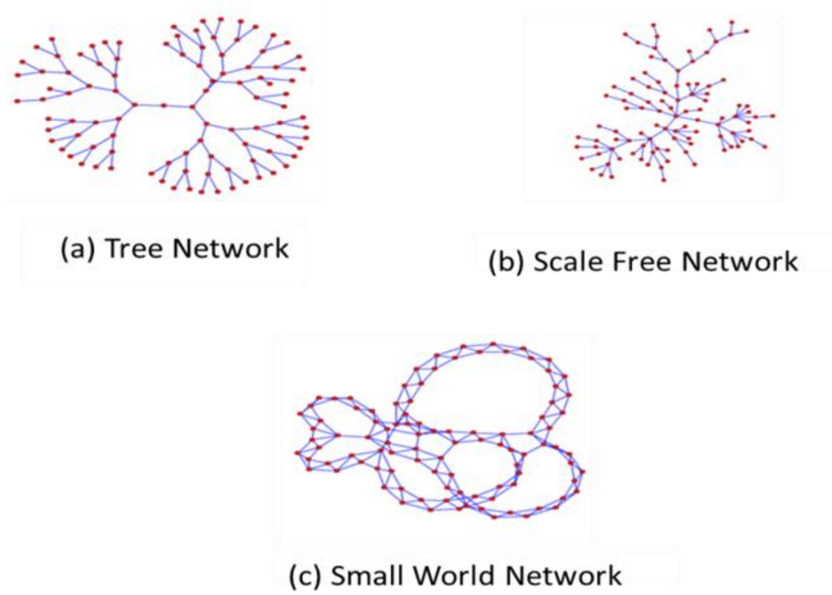
outward. In Scale Free networks this distribution of connections is highly uneven. Some of the members are connected to a lesser degree and some of the members are connected to greater degree, which is how they hold a senior position in the network (Goh et al., 2002). Networks of this type are relatively resilient, but are not at all immune to attack. In other words, a random removal of network members (a crash) will not hurt its stability, but a directed removal of key points will cause the network to collapse quickly (Doyle et al., 2005). Finally, in Scale Free networks, the distribution of density or congestion is constant and not dependent on the exponential coefficient of the distribution of the number of connections (Jeong, 2003).

2.5.2 Small World Networks

A Small World network is a network in which most nodes are not neighbors of each other but most nodes can be reached by other nodes in the networks by hopping over a few nodes (Watts and Strogatz, 1998). The small-world phenomenon is not merely a curiosity of social networks or an artefact of an idealized model (Milgram, 1967, Kochen, 1989). It is probably generic for many large, sparse networks found in nature (Kretzschmar and Morris, 1996). These networks form when long distance connections are added at random to regular networks (Figure 6(c)) (Watts and Strogatz, 1998). They are characterized by low path lengths between nodes and by high clustering coefficients (CC) (Watts and Strogatz, 1998).

The clustering coefficient (CC) is the extent to which the nodes in the graph tend to create a unified group with many internal connections but few connections leading out of the group (Watts and Strogatz, 1998). The clustering coefficient (CC) can be seen as a measurement of the nodes' isolation. The Characteristic Path Length (CPL) is a measurement of the average distance needed to pass from node to node (Watts and Strogatz, 1998). A network can be considered a Small World network when its CPL is similar to the CPL of a random network of the same length, but its CC is much larger (at least by a single order of magnitude) when compared to a random network (Watts and Strogatz, 1998). In other words, in Small World networks, we expect to find a large unified group (Herman, 2003).

Figure 6: Common structures in networks. (a) A tree has branches, usually from a root node, and no cycles (loops). (b) A scale free network has a negative exponential distribution of ties per node. (c) A small world has a regular structure of local connections with some randomly placed long-range connections. (Source: (Paperin et al., 2008)



2.5.3 Summary of Section

Table 4: Network structure characteristics

Types of Network	Characteristics
Scale-free Networks	High Clustering Co-efficient, Power Law Degree Distribution
Small-world Networks	High Clustering Co-efficient, Shorter Characteristic Path Length

Table 4 shows the defining characteristics of scale-free and small-world networks. The above literature clearly states that different social networks share very similar topological characteristics, mainly small world and scale free characteristics (Klemm and Eguiluz, 2002), which are very different from the regular lattice structures (Watts, 1999), or random structures (Watts and Strogatz, 1998) studied in graph theory. Description of these characteristics of network topology is provided in the literature that has been referenced in this section. Methods for measuring these topological characteristics will be discussed in detail in the variables and measures section (section 4.3).

2.6 Network Flows (Borgatti, 2005)

Borgatti (Borgatti, 2005) argues that the various flow types can be distinguished by two properties, the routes through which the traffic flows and the method by which the flows are propagated. Routes are important because, for example, in some flow processes it is desirable for traffic to flow over the shortest possible routes, as in a package delivery system, whereas in other flows the traffic meanders aimlessly, as in gossip passing through a communication network. Methods of propagation, too, differ among networks. For example, the propagation of an e-mail chain letter, which gets sent simultaneously to a list of e-mail addresses, is quite different than that of a traditional, paper-based chain letter, which is sent to one person at a time.

Based on the above explanation, Borgatti (Borgatti, 2005) classified routes into 4 types:

- *Paths:* A path is a sequence of distinct nodes, with each node in the sequence being a neighbor of the preceding node. If one travels from the first node in the path to the last by following ties, then the number of ties that are traveled is the path's length. Each node in a path can only be visited once; each tie can be travelled only once.
- *Geodesics:* There might be multiple paths of varying lengths from one node to another, and a shortest path amongst such paths is called a geodesic.
- *Trails:* A trail is like a path, except that nodes can be visited more than once. However, ties cannot be travelled more than once.
- *Walks:* A walk is the most general type of route, where it is permissible both for nodes to be visited more than once and for ties to be traveled more than once.

Methods of Propagation can be classified into 3 types:

- *Parallel Duplication Propagation:* Propagation occurs by replicating what is at one node to multiple neighbors of the node simultaneously. An example of this process is forwarding email to everybody on the mailing list simultaneously.
- *Serial Duplication Propagation:* Propagation occurs by replicating what is at one node to multiple neighbors of the node one at a time. An example of this process

is gossip network amongst friends. A communicator might pass the gossip to a friend, and then to another, and then to another.

- *Transfer:* Propagation of this type, allows the traffic to be in only a single location at any point in time. An object being passed from node to node, for example a package delivery system where the package exists in only one place at a time.

Based on the classification of routes and method of propagation, Borgatti proposed the following typology for the flow process:

Table 5: Flow Process Based on Route Classification and Method of Propagation (Borgatti, 2005)

	Parallel duplication	**Serial duplication**	**Transfer**
Geodesics	<No process>	Mitotic reproduction	Package delivery
Paths	Internet server	Viral infection	Mooch
Trails	E-mail broadcast	Gossip	Used goods
Walks	Attitude influencing	Emotional support	Money exchange

The examples of flow process from Table 5 are explained below:

- Internet Server: Information on a server can be accessed by multiple peripheral computers at once. For example, in a star network, every computer has a unique path to access the server, which is independent of other paths. Therefore, multiple computers can access the server simultaneously. The path may not necessarily be the shortest path.
- E-mail Broadcast: A message is forwarded from one person to several of his contacts, often by sending one message to all of them simultaneously. It is possible that one of the people on the mailing list might have received the same

message from one of his other contacts. It highly unlikely that he receives the same broadcasted message from the same person again.

- Attitude Influencing: Attitude influencing is an influence process in which individuals effect changes in each other's beliefs or attitudes through interaction. For example, a speaker may persuade many people at the same time about his/her fashion beliefs and continue to influence the same people about the same thing over time.
- Mitotic Reproduction: In this type of reproduction, a cell distributes exact copies of genetic material so the daughter nuclei are genetically identical to each other and identical to the mother nucleus from which they came. The daughter nuclei in turn produce further identical clones. The clones, once fully formed, bifurcate from the parent thereby taking the shortest possible route. This is a phenomenon in which the information spreads through shortest paths.
- Viral Infection: Consider the case of an infection to which the host becomes immune. The infection spreads from person to person by duplication, like gossip, but does not re-infect anyone who already has had it because they have become immune. By contrast, in case of gossip, repeated exposure to a message may cause the recipient to believe it. This (viral infection) is a phenomenon in which information spreads through multiple paths, not just the shortest paths.
- Gossip: Imagine a juicy, very private, story moving through the informal network of employees within an organization. The story is confidential, which does not impede its flow, but means it is typically told behind closed doors to just one person at a time. It spreads by replication rather than transference. Gossip normally does not pass the same link twice (i.e., I do not tell the same person the same story), but can pass the same node multiple times. Thus, it traces trails through the network rather than walks.
- Emotional support: A person dealing with cancer receives emotional support when other people say or do things that help him or her to feel better. For some,

words of encouragement, hope, and optimism are felt to be emotionally supportive. These words of encouragement can come from same people over a period of time; therefore the same information can travel through same nodes and links.

- Package Delivery: A package, to be delivered by a package delivery service, can only be at one place at a time. Its route is designed to be the shortest one possible, in order to reduce the package's delivery time.
- Mooching: Consider a free loading friend, who stays with you as long as he/she is supported and moves on to other people once the support stops, never to revisit again. The node and the links are visited only once.
- Used Goods: A book can only be in one place at a time. As it goes from person A to person B to person C, etc., it could easily return to a person earlier in the chain, simply because person G has no idea that person B had previously received it.
- Money Exchange: Consider a specific dollar bill that moves through the economy, changing hands with each economic transaction. The dollar bill is indivisible and can only be in one place at a time. It could easily move from A to B, B back to A, A to B again, then B to C, and so on. From a graph-theoretic point of view, the bill traverses the network via walks rather than trails.

Borgatti (2005) mapped the best known centrality measures to the flow types as shown in Table 6 below:

Table 6: Flow Process and Major Centrality Measures (Borgatti, 2005)

	Parallel duplication	Serial duplication	Transfer
Geodesics		Freeman closeness	Freeman closeness, Freeman betweenness
Paths	Freeman closeness, Freeman degree	No metrics defined	No metrics defined
Trails	Freeman closeness, Freeman degree	No metrics defined	No metrics defined
Walks	Freeman closeness, Freeman degree, Bonacich eigenvector	No metrics defined	No metrics defined

The definition of centrality and centrality measures are discussed in section 2.7.3.2.

The literature in this section classifies the network flow based on the routes that the network flow takes and the method of propagation of information. It illustrates some of the prominent work that can be categorized based on the routes and propagation methods (Borgatti, 2005). Borgatti noted that most of the sociologically interesting processes are not covered by the existing centrality measures. The examples from above illustrate that transfer follows Markov processes, in which the probability distribution of next step within the process depends only on the current state of the network and not on its previous steps (Norris, 1998). Transfer consequently only relates to the exchange of goods. By contrast, in a parallel duplication process and in a serial duplication process a copy of the information exchanged is maintained at the source, who decides whom to whom he/she will pass on the information. This decision can be based on where the information came from. Therefore, parallel and serial duplication processes are the only ones that are applicable to information propagation in social networks.

2.7 Social Influence within a Network

Social influence occurs when an actor adapts his behavior to the behaviors of other actors in the social system (March, 1955, Cartwright, 1965, Simon, 1957). A precondition for social influence to occur is the availability of information about the behavior of other actors (Leenders, 1995). The sociology literature contains many different theories of social influence (Homans, 1950, Homans, 1974, Festinger et al., 1950, Lazarsfeld et al., 1944, Tajfel, 1972, Linton, 1936, Merton, 1957, Nadel, 1957, Burt, 1987). Most of these state that the attitudes and opinions of people significant to the person influences the way in which a person comes to view a situation (Leenders, 1995). The opinions of others are seen as an appropriate standard against which an actor evaluates his own opinion. In other words, when forming his own opinion, an actor uses other actors as his frame of reference and takes their opinions into account (Leenders, 2002). This idea of frame of reference has been narrowed down to two processes, namely communication and comparison (Leenders, 2002).

2.7.1 Communication

Communication refers to social influence through direct contact between actors. The more frequent and vivid the communication between actors, the more likely actors will adopt each other's ideas and beliefs. The work of Homans (Homans, 1950, Homans, 1974) provides a theoretical foundation for influence through communication. Classical early empirical work was performed by Festinger et al., 1950; Festinger and Kelly, 1951; Lazarsfeld et al., 1944; and Berelson et al., 1954. Lazarsfeld et al. (1944), for instance, argued that people rely on personal contacts to help them select relevant arguments in political affairs. An actor trusts the judgment and evaluation of those who are respected around him. Berelson and colleagues (Berelson et al., 1954) show that political preferences of friends and coworkers strongly determine an actor's preference and that these preferences alter the strength of conviction with which actor's vote preference is held. Baerveldt and Snijders (Baerveldt and Snijders, 1994), who studied the impact of network effects on cultural behavior, have found that petty crime offenses among pupils to be correlated with the number of offenses committed by their friends.

2.7.2 Comparison

In the process of comparison, an actor compares him/herself to others that are considered similar in relevant respects (Tajfel, 1972). Comparisons are fundamental to the traditional view of social structure as a system of statuses interlocked by role relations (Linton, 1936, Merton, 1957, Nadel, 1957). Comparison models were developed during the 1970s explicitly as a vehicle for describing the structure of role relations defining social status across multiple networks. Burt (1987) argues that a comparison is triggered if actors are in competition with one another. By comparison, actors evaluate their relative adequacy. Role playing and imitation are similar to comparison (Burt, 1987).

2.7.3 Measuring Social Influence

Communication and comparison constitute the two most common approaches to measuring the degree of social influence within a network. They tend to be based on four observable phenomena: structural cohesion (Wasserman and Faust, 1994), equivalence (Wasserman and Faust, 1994), centrality (Freeman, 1977) and centralization (Freeman, 1977).

2.7.3.1 Structural Cohesion and Equivalence

The proximity of actors in a social network is associated with occurrence of influence between two actors (Burt and Doreian, 1982, Erickson, 1988, Friedkin, 1983). Two ways of measuring social proximity, structural cohesion and equivalence, have provided contrasting approaches to studying social influence (Marsden and Friedkin, 1993). Structural cohesion determines an actor's influence based on number of actors to which he/she is connected and the strength of the paths between these actors (Wasserman and Faust, 1994). The most restrictive definition of structural cohesion is simple adjacency where two actors are proximate if and only if they are directly tied in a network (Wasserman and Faust, 1994). This is very similar to the process of communication (Mokken, 1979, Seidman and Foster, 1978).

The equivalence approach defines influence in terms of actor's similarity of profiles of network relationships (Wasserman and Faust, 1994). For example, in a binary network, a structurally equivalent pair is indistinguishable when they exhibit exactly the same set of present and absent relations with an identical set of third actors. In effect, one equivalent actor can substitute for another because the two relational patterns are impossible to tell apart. The most restrictive case defines two actors as proximate when they have identical relationships with others in the network (Wasserman and Faust, 1994). This is very similar to comparison (Lorrain and White, 1971).

2.7.3.2 Centrality and Centralization

The next generation of researchers in the field of networks dedicated their efforts to developing metrics for social networks that were inclusive of both structural cohesion and equivalence. Freeman (1977) proposed centrality metrics as a measure of how influential (central) is a particular actor in a social network (Freeman, 1977) and proposed centralization (Freeman, 1979) as a way to measure how centralized a network is. Therefore, centrality is a property of an actor, whereas centralization is a property of a network. A network is considered to be highly centralized if one or few nodes are more connected as compared to other nodes. Similarly, a network is less centralized if all the nodes have more or less similar number of connections in a network.

In his seminal work, Bavelas (1950) investigated formal properties of centrality. He suggested that a particular node or nodes in a group which lies on the communication paths of other nodes and connects them hold a more central position in a social network (Bavelas, 1950). A similar point of view was also expressed by other researchers (Shimbel, 1953, Shaw, 1954). Freeman (Freeman, 1979, Freeman, 1977) argued that, to measure the centralization of a network, the centrality measures should take into consideration the difference between the most central nodes and all other nodes in the network. He went on to propose three different measure of centrality, whose relative efficacy depended upon what the researcher in measuring.

- *Degree Centrality* measures the communication activity of a node. This is a simple count of number of neighbors a node has, with whom it is directly connected.
- *Betweenness Centrality* measures the control a node can exert on the communication process in a network. This measure counts the number of shortest paths between any two nodes in a network, passing through a particular node. The node that has highest number of shortest paths passing through it exerts a better control on the communication process, in the sense that it can force the other nodes to take longer paths, which are sub-optimal.
- *Closeness Centrality* measures the efficiency of a node’s communication process. The distance between two nodes in a network is the shortest path connecting the two nodes. This measure counts the sum distances from a node to all other nodes in a network. The smaller the sum is, the more central the node.

Bonacich (Bonacich, 1972) proposed eigenvector centrality to measure the influence of one particular node on the other nodes in a network. The eigenvector centrality of a particular node is high, if it influences just one other node, who subsequently influences many other nodes (who themselves influence still more nodes). The first node in this network of nodes then regarded as highly influential.

Bonacich (Bonacich, 2007) states that eigenvectors have advantages over graph-theoretic centrality measures like degree, betweenness and closeness when it comes to measuring the influence of a node in a network. Degree, betweenness, and closeness centralities are defined only for classically simple graphs, those with strictly binary relations between nodes. Eigenvector centrality is designed to be distinctively different from mere degree centrality. Degree, betweenness, and closeness measures are especially sensitive to situations in which a high degree position is connected to many low degree positions or a low degree position is connected to a few high degree positions. By contrast, eigenvector centrality can be used with graphs that allow for variations in the degree to which status is transmitted from position to position. For example, a nodes degree, betweenness, and closeness centralities values are high when the node connects to more nodes without consideration for status of the connecting nodes in the network. However, the eigenvector centrality of a node tends be higher, if the node connects to another node with higher eigenvector centrality as opposed to lower eigenvector centrality.

2.7.3.3 Summary of Section 2.7.3

From the above literature it can be safely said that eigenvector centrality is the best measure so far in measuring the influence of an actor in a network. It not only takes proximity based on structural cohesion and equivalence into consideration. It also considers the status of actors based on to whom they are connected within a network.

2.8 Research Gaps and Research Questions

I restate the primary research gap.

<u>Primary Research Gap</u>: Currently, no useful behavioral theory of online social networks, which integrates network structure, network flow and network phenomena, exists.

The primary research gap breaks down into the following two subordinate research gaps, which have been mentioned in section 2.4.1:

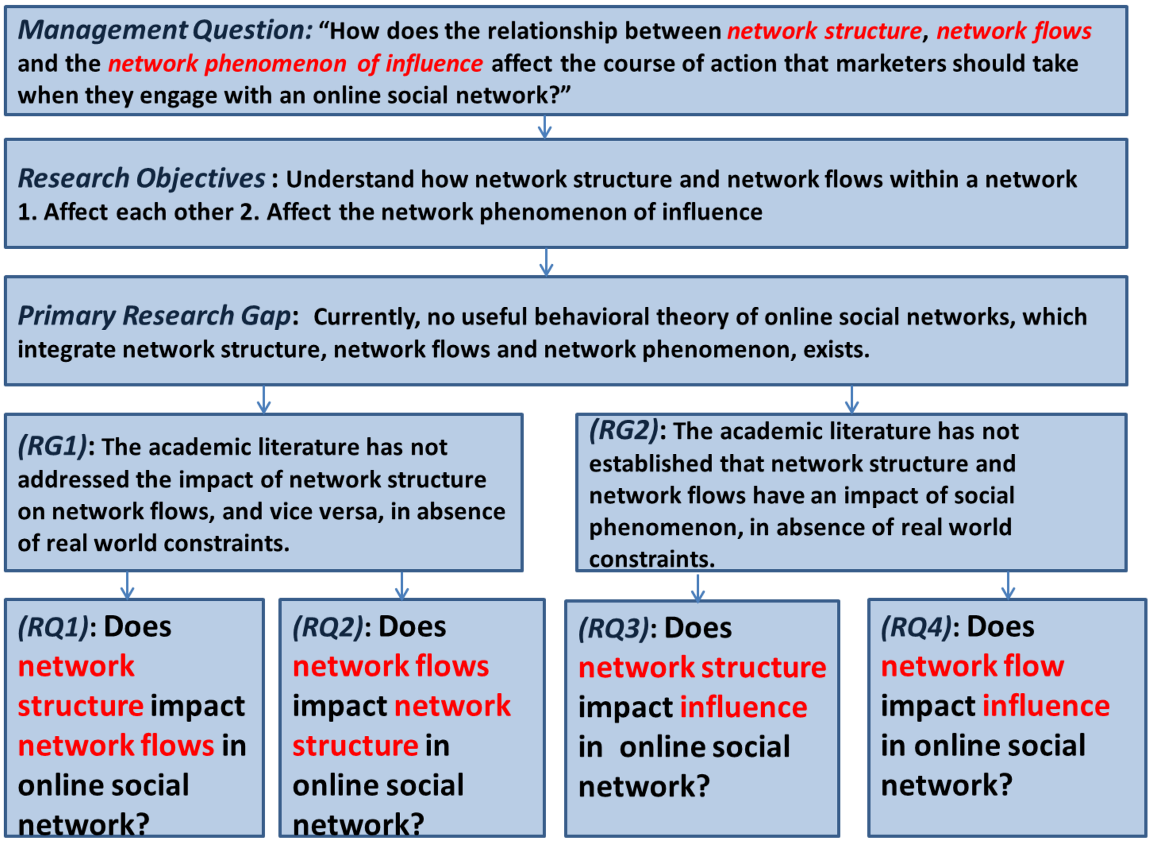
- *Research Gap 1 (RG1)*: Like their counterparts in the real world, online social networks have properties such as network structure and network flows. However, the academic literature has not addressed the impact of network structure on network flows, and vice versa, in absence of real world constraints.
- *Research Gap 2 (RG2):* The academic literature has not established that network structure and network flows have an impact on social phenomena such as influence, in absence of real world constraints.

Research Questions: In order to address the above research gaps, I ask the following research questions:

- Research Question 1 (RQ1): Does network structure impact network flows in a social network that primarily exists online?
- Research Question 2 (RQ2): Does network flow impact network structure in a social network that primarily exists online?
- Research Question 3 (RQ3): Does network structure impact influence within an online social network?
- Research Question 4 (RQ4): Does network flow impact influence within an online social network?

Addressing these research questions will hopefully allow me to achieve my research objective, which has been stated as follows (in section 1.2): to investigate how an online social network's structural organization and the network flows within the network impact each other and network phenomenon of social influence within network. Figure 7 below, illustrates the relationship between my management question, my research objective, the gaps in the existing literature, and my research questions.

Figure 7: The Relationship between Management Question, Research Objective, Research Gaps and Research Questions.

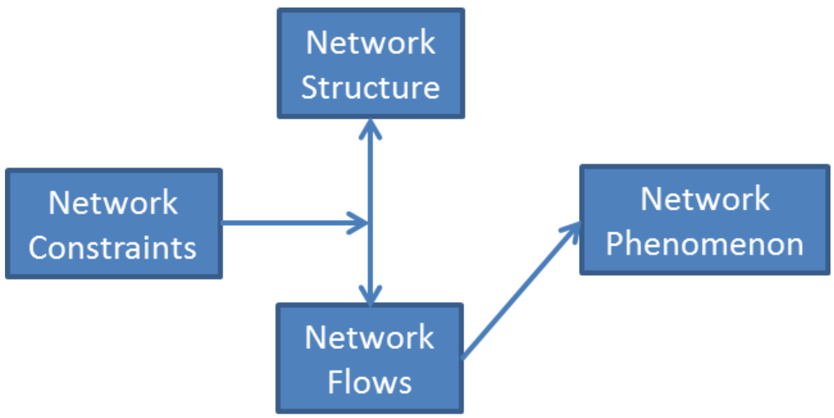


3. Research Framework, Scope and Hypothesis

3.1 Research Framework

The conceptual frame work that has evolved from the literature is shown below (and in figure 5).

Figure 8: Integrated Conceptual Model



As stated in the literature search above, this framework essentially talks about the networks that are formed in the real world, which have constraints like geography, physical distance and connections (section 2.1, section 2.2, and section 2.3). Within this model, literature has shown that network structure and network flows impact each other and network constraints mediate the level of impact between network structure and network flows (section 2.1.5 and section 2.2.6). The flows that emerge are the ones that shape the network phenomena that happen with in a network (section 2.3.1).

The gaps in the literature, shown above (section 2.8), make it clear that social networks that are formed virtually do not have the real world constraints. Thus it is not clear how network structure and network flows impact each other or, for that matter, whether they have an impact on each other at all. Given this change, it also cannot be assumed that only network flows have an impact on network phenomena; network structure could influence network phenomena as well. It also cannot be assumed that network flows are the only factor to have an impact on network phenomena. The literature search in chapter 2 has identified the impact of network structure on the network phenomena as a gap.

In order to address these gaps in the literature, I propose the following framework, which is in line with the research questions asked in section 2.8.

Figure 9: Experimental Framework for Research

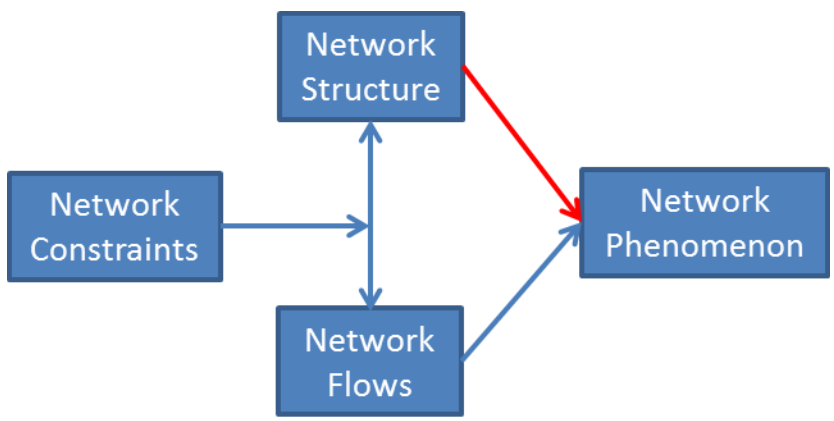


Figure 9 depicts an experimental framework, which incorporates the conceptual model in Figure 8. In addition, it is useful for exploring the impact of network structure on network phenomena, a shortcoming of the conceptual framework depicted in Figure 8. The experimental framework from Figure 9 will therefore be utilized to address research questions RQ1 through RQ4. Hypotheses that pertain to research questions RQ1 through RQ4 will be formed by using this framework.

3.2 Research Scope

3.2.1 Serial Propagation

In section 2.6, I discussed three types of network propagation: parallel, serial and transfer (Borgatti, 2005). Of these, only parallel and serial propagation are applicable to social networks (section 2.6.1). In parallel propagation, one message can be passed from one node to many nodes simultaneously, whereas in serial propagation a message is passed from one node to one node at a time (Borgatti, 2005). Parallel and serial propagation can thus be respectively associated with broadcast communication (Katz and Lazarsfeld, 1955, Kotler, 1994) and word-of-mouth (Roger, 1983, Granovetter, 1973) communication.

In broadcast communication, which has been covered extensively in prior work (Kotler, 1994, Stewart and Ward, 1994, Rice, 1992, Rubin, 1984), information is propagated in parallel, *i.e.* to multiple people at once. Thus broadcast communication tends to be one sided. Advertisements printed in a newspaper, radio shows and television advertisement (Katz and Lazarsfeld, 1955, Kotler, 1994) are examples of parallel propagation.

By contrast, word-of-mouth communication transpires through serial flows. Information is passed from one person to another, one at a time, through a process of interaction (Roger, 1983, Granovetter, 1973). This form of communication has been recognized as "the world's most effective, yet least understood marketing strategy" (Misner, 1999) because the Internet provides companies with more word-of-mouth marketing opportunities than ever. In addition, word-of-mouth communication is significantly cheaper than many forms of broadcast communication, such as, for example, tossing away millions of dollars on Superbowl ads (Whitman, 2006). As online social networks are virtual social aggregations in which information flows happen due to people interacting with each other by word of mouth (Knoke and Kuklinski, 1982, Wellman, 1983, Rheingold, 1993), *I limit the scope of this dissertation to serial flows*.

3.2.2 Paths and Geodesics

In section 2.6, I discussed four types of routes through which propagation happens: geodesics, paths, trails and walks. This dissertation will *not deal with trails and walks* for two reasons. First, calculating trails and walks can be very expensive in terms of time and compute power, especially in highly connected datasets (Kashima et al., 2003, Gartner, 2002). Second, the impact of an actor's ability to exert influence over other actors can be studied adequately by considering paths and geodesics. *Paths* are used because the phenomena under study are serially based -- one actor will interact with only one other actor at one time. Therefore, paths will be used as proxy for *information spread process* (explained in 4.3.3.2.2). *Geodesics* are the shortest paths between any two specific actors within the network. Therefore, geodesics will be used as proxy for *speed of information spread* (explained in 4.3.3.2.2). Influence could be a stronger function of paths than geodesics, or vice versa, or not correlated to either. To date, no study has determined which of these possibilities is correct.

3.2.3 Directionality

In graph theory, networks are classified as directional or non-directional.

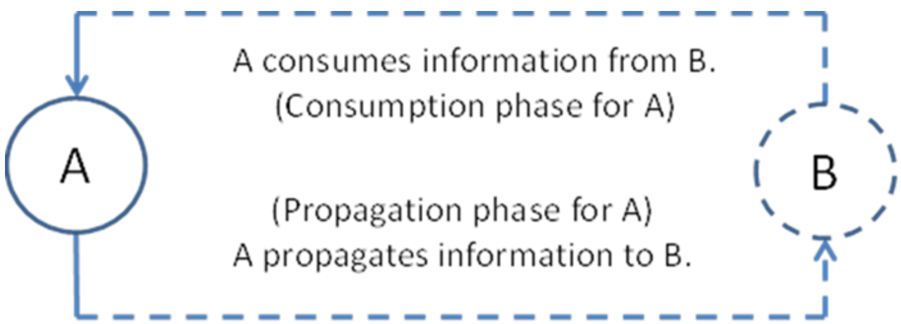
1. *Non-Directional Network*: This is a type of network in which all relations are symmetrical. If an actor A interacts with actor B, it is assumed that actor B also interacts with actor A (Wasserman and Faust, 1994).
2. *Directional Network*: This is a type of network in which relations are not symmetrical. If an actor A interacts with actor B, it is not assumed that actor B interacts with actor A (Wasserman and Faust, 1994).

In most extant analyses of social networks relationships have been treated as reciprocal (e.g., (Burt, 1976, Burt and Doreian, 1982, Granovetter, 1973). However, directionality has been a factor in some studies (e.g. (Allen, 1977, Roger, 1983) because relationships are not necessarily reciprocal. Thus directionality is taken into consideration in this dissertation.

In order to further understand the role of interaction, I partition the interaction process into a consumption phase and a propagation phase, as shown below in Figure

10. In figure 10, A consumes information from B in the consumption phase, whereas A propagates information to B in the propagation phase.

Figure 10: Information consumption and propagation flow



The consumption and propagation phases only impact directional networks, as they are non-symmetrical. In non-directional networks, the consumption phase and the propagation phase are equivalent. Therefore, I will consider the consumption and propagation phases in directional networks separately to understand the impact of each phase on the influence of nodes and then collectively to understand their combined impact.

3.3 Research Hypotheses

The empirical study that has been proposed for this dissertation intends to determine whether the structure of a network formed due to a virtual social aggregation impacts the network flows within that network; whether the network flows associated with such a network impact the network's structure; and whether network structure and network flows affect the ability of an actor within a network to exert influence over other actors within the same network. This objective is achieved by addressing research gaps RG1 and RG2, as well as by answering research questions RQ1 though RQ4. Directionality enters the hypotheses for reasons explained in section 3.2.3.

The following hypotheses address the research question RQ1:

- *Hypothesis 1 (HP1)*: The structural characteristics of a social network impact its network flows.
    - *Hypothesis 1a (HP1a)*: The structural characteristics of a non-directional social network impact its network flows.
    - *Hypothesis 1b (HP1b)*: The structural characteristics of a directional social network impact its network flows.
    - *Hypothesis 1c (HP1c)*: The structural characteristics of a directional social network impact its network flows in the consumption phase.
    - *Hypothesis 1d (HP1d)*: The structural characteristics of a directional social network impact its network flows in the propagation phase.

The following hypotheses address the research question RQ2:

- *Hypothesis 2 (HP2)*: Network flows impact the structural characteristics of a social network.
    - *Hypothesis 2a (HP2a):* Network flows impact the structural characteristics of a non-directional social network.

- *Hypothesis 2b (HP2b)*: Network flows impact the structural characteristics of a directional social network.
- *Hypothesis 2c (HP2c)*: Network flows impact the structural characteristics of a directional social network in the consumption phase.
- *Hypothesis 2d (HP2d)*: Network flows impact the structural characteristics of a directional social network in the propagation phase.

The following hypotheses address the research question RQ3:

- *Hypothesis 3 (HP3)*: Network structure impacts influence within an online social network.
    - *Hypothesis 3a (HP3a)*: Network structure impacts influence within an online social network in a non-directional social network.
    - *Hypothesis 3b (HP3b)*: Network structure impacts influence within an online social network in a directional social network.
    - *Hypothesis 3c (HP3c)*: Network structure impacts influence within an online social network in a directional social network during the consumption phase.
    - *Hypothesis 3d (HP3d)*: Network structure impacts influence within an online social network in a directional social network during the propagation phase.

The following hypotheses address the research question RQ4:

- *Hypothesis 4 (HP4)*: Network flow impacts influence within an online social network.
  - *Hypothesis 4a (HP4a)*: Network flow impacts influence within an online social network in a non-directional social network.
  - *Hypothesis 4b (HP4b)*: Network flow impacts influence within an online social network in a directional social network.
  - *Hypothesis 4c (HP4c)*: Network flow impacts influence within an online social network in a directional social network during the consumption phase.
  - *Hypothesis 4d (HP4d)*: Network flow impacts influence within an online social network in a directional social network during the propagation phase.

4. Research Methods

In this chapter, I discuss issues related to research design, including the unit of analysis and the choice of research setting. I subsequently explain my approach to collecting data for the research I have conducted, and how to measure the variables described in chapter2. I also discuss the validity and reliability of the measures. At the end of this chapter, I describe my approach to data analysis.

4.1 Research Design

I am looking at social networks from the point of view of product categories. I would thus like to know whether the patterns in a social network (structure, information flows and loci of influence) vary as a function of product category. I am interested in scale in particular, because I would like to find out whether the social networks that discuss products categories in which content is consumed at high volumes behave differently from social networks that discuss product categories in which content is consumed at relatively low volumes. Therefore, scale becomes a control variable in my research design and the theoretical criterion for case selection.

This is a population study. Due to modern data extraction capabilities on the Internet, I can study whole populations. Studying the population in its entirety not only eliminates the sample selection bias; it also ensures that the results observed are valid and generalizable to the entire population under study. This is especially important in studies that involve networks, as selecting only a sample instead of the population can break a network into multiple small networks, leading to faulty results. Furthermore, my data collection method (see section 4.3) allows me to extract large amount of data from which statistically significant conclusions can be drawn. Quantitative analyses of network phenomena (influence), the impact of network attributes (network structure and network flow) on network phenomena, and the impact of network attributes on each other consequently become feasible.

In my study, I use the case study research method to establish my experimental setting ((Yin, 1984), as cited by (Eisenhardt, 1989), p. 534). A product category that is discussed by a social network is considered a case. The social network that discusses the product category is my unit of analysis. The product category in each case will be sufficiently mature, so as to avoid any bias associated with startup effects. Conversely, the product category should not be in rapid decline, so as to avoid any bias that pertains to rapid decay of the social network under study.

In general, case study research tends to deploy inductive reasoning and qualitative methods (Eisenhardt, 1989, Yin, 1994). However, when guiding propositions have been established, the case study research method can be used to confirm or reject these propositions (Yin, 1994) through deductive reasoning. In addition, quantitative methods have been used to identify common sequences of events in large samples (e.g., (Abbott, 1990)).

I have established specific hypotheses in section 3.3, which I would like to test under a particular set of circumstances that may change as events unfold. It is thus appropriate for me to conduct deductive research in which I confirm the existence of phenomena that have been proposed a priori.

4.1.1 Research Setting

A mentioned in section 1.2, Twitter conversations were chosen as a setting for this study because they exhibit the characteristics of online social networks, which contrast sharply with social networks that occur in the real world. In addition, Twitter is the only social media platform that can capture changes in the context and content of online conversations at the rate at which they actually occur. Furthermore, all data on Twitter are available in the public domain. Finally, Twitter is popular enough for it to cover a sufficient number of conversations to enable a comprehensive analysis of the product categories under study. Twitter gets almost 190 million[17] unique visits every month, which makes it the eighth most popular website in the world. Over 1 billion tweets[18] are generated on Twitter every 5 days.

Twitter is a micro-blogging platform (Zhao and Rosson, 2009) founded in 2006. Microblogs are short comments usually delivered to a network of associates (Huberman et al., 2008). Microblogging is also referred to as micro-sharing, micro-updating, or Tweeting (Huberman et al., 2008). Tweeting directly impacts word of mouth

[17] http://preview.alexa.com/siteinfo/youtube.com (accessed on 04/09/2014)
[18] http://www.statisticbrain.com/twitter-statistics/ (accessed on 04/09/2014)

communication because it allows people to share thoughts almost anywhere (i.e., while driving, getting coffee, or sitting at their computer) to almost anyone "connected" (e.g. Web, cell phone, IM, email) on a scale that has not been seen in the past (Honeycutt and Herring, 2009). While the shortness of the microblog keeps people from writing long thoughts, it is precisely the micro part that differentiates microblogs from other word-of-mouth media, including full blogs, web pages, and online reviews (Ramage et al., 2010). A standard microblog is approximately the length of a typical newspaper headline and subheading (Milstein et al., 2008) which makes it easy to both produce and consume. (In Twitter's case, a tweet is limited to 140 characters.) Tweets commonly ask for or share information, news, opinions, complaints, or details about daily activities. Tweets may include hyperlinks to news stories, blogs, pictures, videos, etc. Tweets show up in the stream of those following the poster of the tweet; most posts are also publically available.

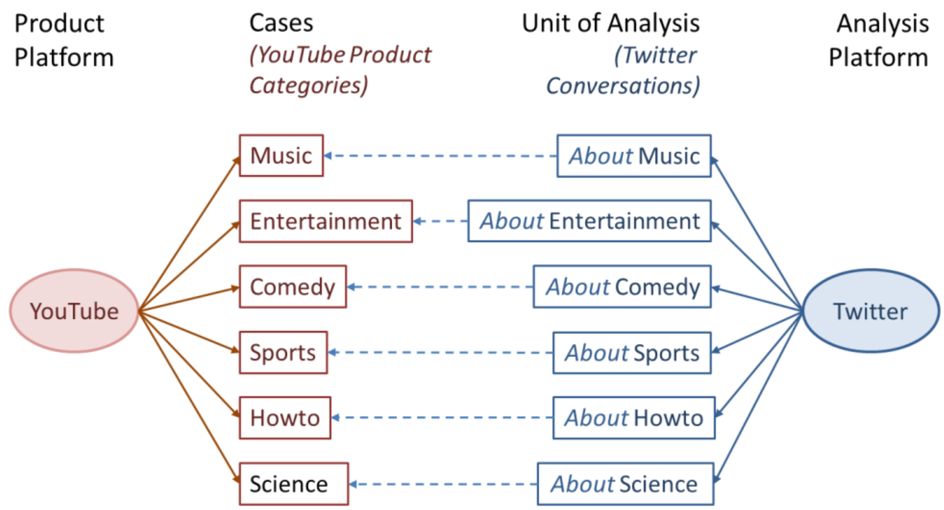
4.1.2 Case Selection

Given that Twitter is the research setting, Twitter communities become the unit of analysis. A Twitter community formed around a specified YouTube product category forms a case, for which all hypotheses will be tested. Case selection in this study (like in many others) depends upon theoretical sampling and replication logic (Yin, 1994, Leonard-Barton, 1990). The key criterion for theoretical sampling is scale, primarily because community behavior may vary as a function of community size.

4.1.2.1 Replication Logic

Replication logic manifests itself by selecting two product categories from each level of distribution volume. All hypotheses will be tested in more than one case. I will use the same input variables, moderating variables, control variables and output variables. However, I do not necessarily expect to get similar results from replication because social networks function autonomously. As explained in section 2.3, the relationship between the properties is only true within the contextual model; it may be false outside the contextual model (Borgatti and Kidwell, 2011). However, replication of cases "requires that the phenomenon being studied be defined by some characteristics common to all the research situations" ((Yin, 1984), as cited by Leonard-Barton, 1990, p. 251). Thus, all cases in my research come from a common delivery platform—YouTube.

Figure 11: Case Selection Process



4.1.2.2 YouTube

YouTube product categories were chosen to identify Twitter communities, as shown in fig.11. It is assumed that more popular product categories on YouTube will generate bigger communities on Twitter. This assumption will be tested during the analysis in chapter 5.

The success of a product category delivered on YouTube depends on its "popularity" or distribution volume, which is generally measured by the total number of views per unit time (Xu et al., 2008). Theoretical sampling (Yin, 1984, Eisenhardt, 1989, Leonard-Barton, 1990) in my study consequently consists of choosing product classes that either have very high or relatively low distribution volumes, as well as some product classes of intermediate scale.

YouTube was chosen as a delivery platform for this research because some of its product categories are an order of magnitude more popular than others. I consequently expect that the largest Twitter community in my sample will be much bigger than the smallest. Music, comedy, entertainment and sports have been identified as categories of interest on YouTube in the academic literature (Thelwall et al., 2012, Xu et al., 2008) as well as in industry reports. [19] "Music" has been rated to be the most popular category as it comprises of almost 31% of all videos. "Entertainment" has been slated to be the second most popular category with 14.59% of all videos. Music and

[19] http://www.sysomos.com/reports/youtube/#categories (accessed on 04/09/2014)

Entertainment have consequently been chosen as cases in the “large” volume category. “Comedy” and “Sports” categories are in the middle range of popularity with each category comprising of almost 6% of all videos. They will serve as cases in the “medium” category. I also intend to analyze “Howto” and “Science” categories, as they lie on the lower end of popularity, comparatively, with each category comprising of only 2.5% to 3% of overall videos.

4.2 Data Collection

I have conducted a retrospective study, and the data for this retrospective study was collected in continuous time. When data are recorded in a continuous time, the number and sequence of events and the duration between them can all be calculated. The main advantage of this approach lies in the greater detail and precision of information (Blossfeld and Rohwer, 1995). It also reduces time required to collect data, and it enhances the chances of recognizing the overall patterns (Leonard-Barton, 1990).

Data on the conversations about the chosen product categories was collected on Twitter. Twitter data is easily available through application programming interfaces (API’s) from which the networks forming within a context can be easily deduced. For the sake of simplicity, I use keyword search as a means of finding contextual network (Jansen et al., 2009). Both, Twitter platform as data source and keyword search as data filter, have been used respectively in previous studies (Williams et al., 2013, Teevan et al., 2011, Jansen et al., 2009).

Tweets have a very unique character. In contrast to any other message, they are limited to 140 characters (Ramage et al., 2010). Every person or entity (like alias, company, etc…) is identified by its Twitter handle. Every Twitter handle can tweet. A Twitter handle can direct a tweet towards another Twitter handle by "@ mentioning" them. The recipient Twitter handle can either forward the message to its network by retweeting "RT @" the sender's message, or reply to the sender by "@ mentioning" the sender's Twitter handle. The recipient can choose to do neither. Tweets are time stamped and publicly displayed on the Twitter platform.

Twitter generates almost than 1 billion tweets every 5 days. Therefore, in order to reach the relevant audience, it is important to weed out noise, which is classified into two categories:

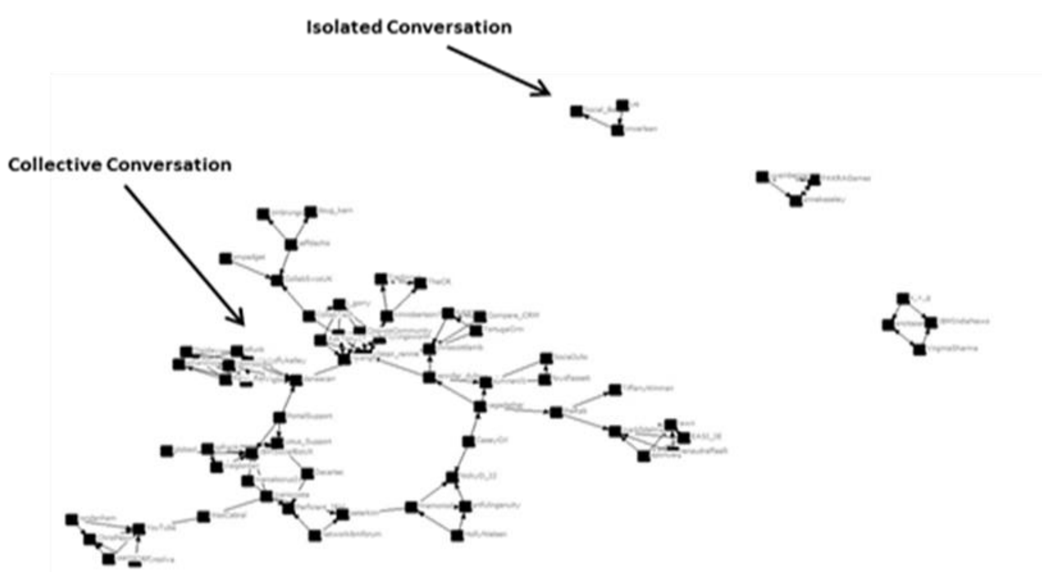
1. *Contextual Noise:* People have multiple topics of interest which may vary from the work that they do, their hobbies, their likes and dislikes, lifestyle choices, etc. Hence, they tweet about these multiple topics of interest. In order to identify a relevant social network, the context of conversations that is relevant to the business objectives (marketing, brand perception, customer support, etc…) needs to be identified. The remaining conversations fall under contextual noise. Contextual noise is very subjective and depends upon the business objective. Reducing contextual noise is achieved by using keyword searches.

2. *Broadcast Noise:* After identifying the context, a social network forming within that context can be identified. In order to identify these networks, it is necessary to identify the relationships people form within the network. Relationships in this case are formed when people interact with each other. In this case, we

consider two actions that form relationships when they are tweeting somebody (@ mentioning) or retweeting somebody (RT @). The tweets that do not evoke any response, i.e., nobody interacts (@mentions or RT @), are considered broadcast noise.

The removal of broadcast noise provides people engaged in the contextual conversation. Within the contextual conversation only the largest network of people (community) engaged in a collective conversation everyday will be considered for analysis. The distinction between the collective conversation and isolated conversations is shown in Figure 12 below. A large group of people are engaged in a collective conversation, whereas small isolated groups converse on the side in isolated conversations.

Figure 12: Collective vs. Isolated Conversations



The rate of participation in the largest network does not impact the size of network, but it does impact the volume of tweets associated with the largest network. Therefore, while considering the total number of people participating in the largest network, only the Twitter user names that participate on a particular day will be counted for that day. Even if the participants tweet more than once, they will still only be counted once as the 'daily unique'. But while considering the total number of tweets, only the tweets associated with the largest network will be counted for analysis. Same process will be followed while measuring number of people participating on daily basis and tweet volumes on a daily basis associated with overall topic, broadcast and engaged activity within the overall topic.

It is noteworthy to mention that the rate of interaction between two people may be seen as strength of their relationship, thereby defining strong ties and weak ties within a network. The changing values of the rate of interaction over a period of time can be used to define the dynamics of the relationship, i.e., are the relationships getting stronger or weaker. The impact of the rate of participation is out of scope for the thesis on hand. However, I identify impacts of rate of participation on network structure, network flow and network phenomenon as an area for future research.

This data collection process will be used to obtain data for the topics mentioned above. Data has been gathered for a period of three months, from Dec31st, 2013 to March 31st, 2014. Metadata for all the chosen topics will consist of 'Total_Tweets', 'Broadcast_Tweets', 'Engaged_Tweets', 'Community_Tweets', ' Total_People', 'Broadcast_People', 'Engaged_People' and 'Largest_Community'. Definitions for the Metadata are shown in Table 7 below.

Table 7: Definitions of Metadata

Meta Data	Definitions
Total_Tweets	Cumulative sum of daily volume of tweets associated with the topic
Broadcast_Tweets	Cumulative sum of daily volume of tweets associated with all the broadcast activity in a topic
Engaged_Tweets	Cumulative sum of daily volume of tweets associated with all the engaged activity in a topic
Community_Tweets	Cumulative sum of daily volume of tweets associated with the largest network engaged in collective conversation within a topic
Total_People	Cumulative sum of daily unique people associated with the topic
Broadcast_Tweet	Cumulative sum of daily unique people associated with all broadcast activity in a topic
Engaged_People	Cumulative sum of daily unique people associated with all engaged activity in a topic
Largest_Community	Cumulative sum of daily unique people associated with the largest network engaged in collective conversation within a topic

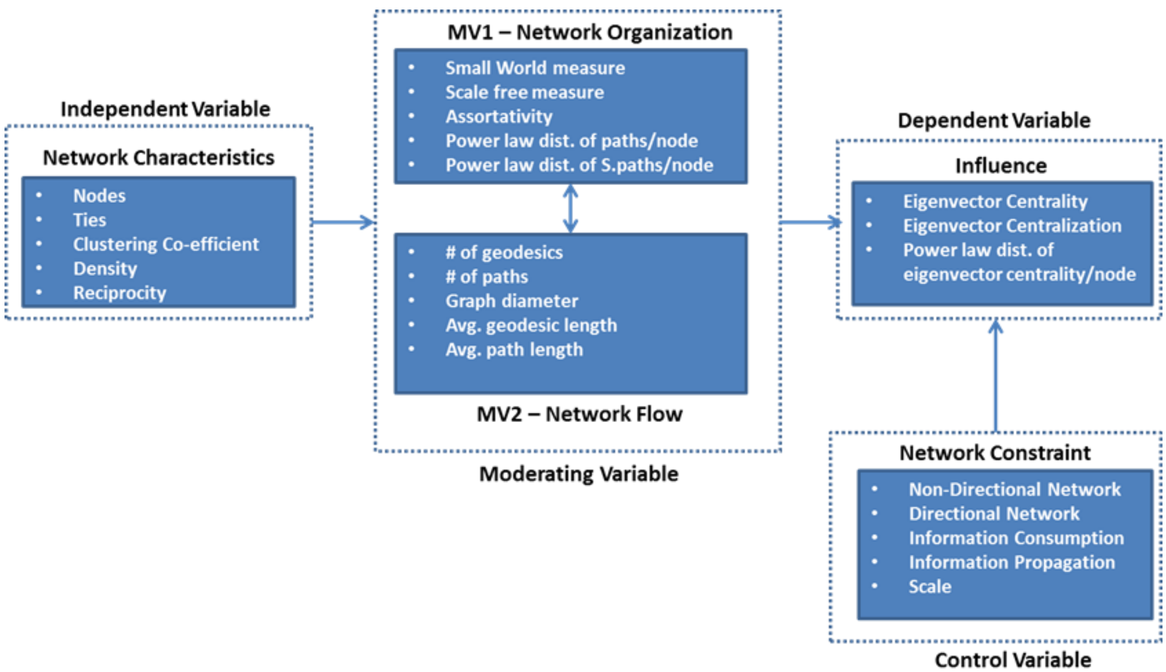
The time period of data collection was chosen at random. A period of three months of data was chosen to control for any monthly periodicity in the data (Gonçalves and Ramasco, 2008, Meiss et al., 2008). The data has been analyzed in daily intervals, in order to capture tweet volatility patterns caused due to daily routine (Dodds et al., 2011, Frank et al., 2013). For example, Twitter users in Tokyo tweet a lot less during working hours.[20] The 24 hour started in accordance with Greenwich Mean Time (GMT). The details of the analysis, the variables and the measures are described in the next section.

[20] http://gigaom.com/2012/06/04/twitter-shows-when-we-tweet-and-explains-why-its-search-sucks/ accessed on 4/27/2014

4.3 Variables and Measures

Figure 13 shows the research framework along with the measures and variables that will be used for this research. This frame work consists of four types of variables: 1) *independent variables* that will used to measure the level of activity within a network; 2) *moderating variables* that measure the network structure and network flow; 3) *dependent variables* that measure the influence of an actor within a network; and 4) *control variables*, which impact the dependent variable.

Figure 13: Research Framework with Variables



The details about what these measures mean and how to measure them is discussed in the following sections.

4.3.1 Mathematical Preliminaries (Wasserman and Faust, 1994)

The social network is the unit of analysis for this research; it will be viewed as a graph. In this section, I will explain the mathematical preliminaries from graph theory that is required to understand the variables and measures generated in subsequent sections.

Let G be a network such that

$$G = (V, E) \ldots\ldots\ldots\ldots (1)$$

where

V – is a finite and non-empty set of nodes

Therefore,

$$V = \{1, 2, 3, \ldots\ldots N\}$$

and

E – is a finite and non-empty set of ties

Therefore,

The tie (i,j) ∈ E is incident with nodes i and j.

(i,j) ∈ E is a link, if i ≠j……………………………….(2)

(i,j) ∈ E is a loop, if i = j………………………………(3)

If two nodes are incident with the same tie, then they are adjacent. Adjacent nodes are called neighbors. Defining the N×N adjacency matrix A = (a_{ij}) by setting a_{ij} equal to 1 if (i, j) ∈ E, and 0 if not. Therefore, the adjacency matrix is a matrix representation of a graph displaying connectivity of the graph. The rows and columns of the graph are labeled by the nodes. If there is a tie between two nodes, then the tie is indicated in the matrix as 1; otherwise the link takes the value of 0. This is also the first order adjacency matrix, i.e., it defines nodes that are connected directly. The first order adjacency matrix does not define relations that are not direct. In order to do so, a higher order of adjacency matrices are required, which can be achieved by the multiplying the first order adjacency matrix with itself. For example, to identify nodes that have just one node between them, a first order adjacency is multiplied with itself. The resultant matrix is called the second order adjacency matrix.

Similarly, Let "A" be an N×N adjacency matrix; then a degree matrix "D" is a second order adjacency matrix in which all the elements except the diagonal elements are non-zero.

Then the second adjacency matrix is given by

$$A^2\ (\alpha_{I,j}) = A^1 \times A^1\ \ldots\ldots\ldots\ldots\ldots\ldots.(4)$$

Hence, an N×N degree matrix (D) is given by

$$D = A^2 \ldots\ldots\ldots\ldots\ldots\ldots..(5)$$

where

$$\alpha_{i,j} = \alpha_{i,j} \ldots\ldots\ldots\ldots\ldots\ldots..(6)$$

iff i=j and (i,j) ∈ A^2

and

$$\alpha_{i,j} = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(7)$$

iff i ≠j and (i,j) ∈ A^2

4.3.2 Measuring Independent Variables

The independent variables in this research are the number of nodes, number of ties, clustering co-efficient, network density and reciprocity. Nodes and ties have been defined in section 4.3.1. In this section, I will define clustering co-efficient, network density and reciprocity and elucidate how these variables are measured.

4.3.2.1 Clustering Coefficient

Clustering is a typical property of acquaintance networks, where two individuals with a common friend are likely to know each other (Wasserman and Faust, 1994). The clustering coefficient was described by Watts and Strogatz, in context of social networking, as the degree to which the nodes in the graph cluster together (Watts and

Strogatz, 1998). Newman et al. also described clustering coefficient to be same as the transitivity of a graph and defined it as follows (Newman et al., 2002)

$$\mathrm{C}\,(\mathrm{G}) = \frac{3\ \times\ \Delta(\mathrm{G})}{\tau(\mathrm{G})} \ldots\ldots\ldots\ldots(8)$$

where C(G) – clustering coefficient of the graph,

$\Delta(\mathrm{G})$ – total number of triangles in the graph, and

$\tau(\mathrm{G})$ - total number of connected triples in the graph.

Calculating the total number of triangles:

Let A^3- third order adjacency matrix of a graph.

The diagonal elements of A^3 contain elements that start from node i and after passing through 2 other nodes ends at the same node i. This can happen only if it is triangle. The diagonal element counts each triangle 3 times. Example: triangle ijk is counted i to j to k and i to k to j. Thus every triangle is counted 6 times.

Therefore,

$$\Delta(\mathrm{G}) = \frac{1}{6}\sum_{(\mathrm{i,j})=1}^{\mathrm{n}} \mathrm{A^3}_{\mathrm{i,j}} \ \ldots\ldots\ldots\ldots.(9)$$

where i = j

Calculating total number of connected triples:

Let A^2 - Second order adjacency matrix of a graph

The elements of A^2 contain elements that start from node i and after passing through 2 other nodes ends at the same node i or any other node in the network j. These are called connected triples. Thus every connected triple is counted 4 times.

Therefore,

$$\tau(G) = \frac{1}{4} \sum_{(i,j)=1}^{n} A^2{}_{i,j} \quad \ldots\ldots\ldots\ldots.(10)$$

where i ≠ j

4.3.2.2 Density

Graph density measures the fullness of a graph. It is a measure which looks at all the ties in the graph and compares it to the all the possible ties in a graph (Wasserman and Faust, 1994).

Therefore,

$$\text{Density (D)} = \frac{\text{Total number of ties in a graph(E)}}{\text{All possible ties in a graph } (E_T)} \quad \ldots\ldots\ldots(11)$$

4.3.2.3 Reciprocity

Reciprocity is an important characteristic of directed networks which helps quantify tendency of node pairs to form mutual connections with each other (Newman et al., 2002). Reciprocity is a ratio of bi-directional ties in the network to non-bi-directional ties in the network.

Therefore,

$$\text{Reciprocity} = \frac{\text{Total number of bi} - \text{directional ties in a graph(E)}}{\text{Total number of non bi} - \text{directional ties in a graph(E)}} \ldots\ldots(12)$$

4.3.3 Measuring Moderating Variables

4.3.3.1 Network Structure (MV1)

4.3.3.1.1 The Small World Measure

A network G with n nodes and m ties is a small-world network (Watts and Strogatz, 1998), if it has a similar path length but a greater clustering of nodes than an equivalent random graph with the same number of nodes n and same number of ties m. A random graph is constructed by uniquely assigning each tie to a node pair with uniform probability (Bollobás, 1984).

A key concept in defining small-world networks is that of 'clustering' (C (G)) which measures the extent to which the neighbors of a node are also interconnected.

This concept has been defined in section 4.3.2.1. The other key concept that pertains to network structure is *path length*, which has been defined as the minimum number of ties that must be traversed to get from one node to the other. By extension, the *minimum path length* between two nodes is the minimum number of ties that must be traversed to get from one node to the other (Fronczak et al., 2004). The mean value of the minimum path length over all node pairs will be denoted by L_g.

More formally, let L_g be the mean path length of graph G and C_g its clustering coefficient. Let L_{gr} and C_{gr} be the corresponding mean path length and clustering co-efficient for a random graph. Then a network is said to be a small world network if SM is greater than 1,

where

$$\mathrm{SM} = \frac{\mathrm{C_{sm}}}{\mathrm{L_{sm}}} \ldots\ldots\ldots..(13)$$

such that SM>1

where

$$\mathrm{C_{sm}} = \frac{\mathrm{C_g}}{\mathrm{C_{gr}}} \ldots\ldots\ldots..(14)$$

such that $C_g >> C_{gr}$

and where

$$\mathrm{L_{sm}} = \frac{\mathrm{L_g}}{\mathrm{L_{gr}}} \ldots\ldots\ldots.(15)$$

such that $L_g \geq L_{gr}$

4.3.3.1.2. Scale Free Measure

In order to understand the scale free structure of the network and quantify the level of scalefreeness displayed by the network, Li et al. proposed the S-metric (Li et al., 2005), which is defined as follows:

$$s(g) = \sum_{(i,j)\in G} d_i \ d_j \ldots\ldots(16)$$

where

d_i , d_j denote the degree of node i and node j.

The value of s(g) depends explicitly on the graph and not the process through which it is constructed. The s(g) metric measures the extent to which the graph G has a hub like structure as the value of s(g) is maximized when nodes with high degrees are connected to each other. Similarly, s(g) takes a lower value when the nodes with high degree are connected to nodes with low degree. Therefore, when value of s(g) is high, the graph is scale free, and the value of s(g) is low, the graph is scale rich.

We can compute s(g) with respect to any “background” set G of graphs. Moreover, for any background set, there exists a graph whose connectivity maximizes the s-metric and is referred to as “s_{max} graph”. The s_{max} graphs for different background sets are of interest since they are essentially unique and also have the most "hub-like" core structure. Therefore, s_{max} value can be used for normalizing s(g) value between 0 and 1 as shown below.

$$S = \frac{s(g)}{s_{max}} \ldots\ldots\ldots\ldots\ldots(17)$$

This also means s_{max} has to be generated for every degree sequence that results from each trial. Constructing the s_{max} element among these graphs can be achieved trivially, by applying the following two-stage process. First, for each vertex i, if d_i is even, then attach d_i /2 self-loops; if d_i is odd, then attach (d_i − 1)/2 self-loops, leaving one available "stub". Second, for all remaining vertices with "stubs", connect them in pairs according to decreasing values of d_i. Obviously, the resulting graph is not unique, as the s_{max} element (indeed, two vertices with the same degree could replace their self-loops with connections between one another). Nonetheless, this construction does maximize s(g), and in the case when d_i is even for all i ∈ V, one achieves an s_{max} graph with

$$s_{max} = \sum_{i=1}^{n} \frac{d_i}{2} \cdot d_i^{\,2} \ldots\ldots\ldots\ldots(18)$$

In the case where some d_i are odd, the s_{max} graph will have a value of s(g) that is somewhat less than s_{max}, and will depend on the specific degree sequence. Thus, the value of $s_{\max}$ represents an idealized upper bound among unconstrained graphs, but it can only be realized in the case when all vertex degrees are even.

Scale Free (SF) graphs are defined as graphs with scaling or power law degree distributions. They are generated by a stochastic construction mechanism that is based on incremental growth (i.e. nodes are added one at a time) and preferential attachment (i.e. nodes are more likely to attach to nodes that already have many connections). The main properties of SF graphs that appear in the existing literature can be summarized as:

1. SF networks have scaling (power law) degree distribution (Barabási and Albert, 1999).
2. SF networks can be generated by a variety of random processes, foremost among which is preferential attachment. (Albert and Barabasi, 2000).
3. SF networks have highly connected "hubs" which "hold the network together" and give the "robust yet fragile" feature of error tolerance, but attack vulnerability (Albert et al., 2000, Alderson and Willinger, 2005).
4. SF networks are generic in the sense of being preserved under random degree preserving rewiring (Doyle et al., 2005).
5. SF networks are self-similar (Itzkovitz et al., 2005).
6. SF networks are universal in the sense of not depending on domain-specific details (Itzkovitz et al., 2005).

4.3.3.1.3. Assortativity

The measure r(g) of assortativity in networks was introduced by Newman (2002), who describes assortative mixing (for r > 0) as “a preference for high-degree vertices to attach to other high-degree vertices” and disassortative mixing (for r < 0) as the converse, where “high-degree vertices attach to low-degree ones.” (Newman, 2002) Assortativity has been developed in the context of an ensemble of graphs, but Newman provides a sample estimate of assortativity of any given graph g. Using our notation, Newman’s formula can be written as:

$$r(g) = \frac{\left[\sum_{(i.j)\in E} d_i d_j\right] - \frac{\left[\sum_{i\in V} \frac{1}{2} d_i^{\,2}\right]^2}{l}}{\left[\sum_{i\in V} \frac{1}{2} d_i^{\,3}\right] - \frac{\left[\sum_{i\in V} \frac{1}{2} d_i^{\,2}\right]^2}{l}} \dots\dots\dots\dots\dots\dots..(19)$$

where,

l - number of ties in the graph.

Conceptually, r(g) and s(g) have the same aim, but with different and largely incomparable normalizations, both of which are interesting. The first term in the numerator is the same as s(g). The first term in the denominator is same as s_{max}. The second term in both numerator and denominator can be interpreted as the “center” or zero assortativity case. Thus, the perfectly assortative graph can be viewed as the s_{max} graph (within a particular background set G), and the assortativity of graphs is measured relative to the s_{max} graph, with appropriate centering. Therefore, the assortative measure is linearly related to the scale free measure. The assortative measure helps in understanding the connection preference within a graph, whereas the scale free measure helps in understanding the formations of hubs.

4.3.3.1.4. Power Law Distribution of Total Paths per Node and Total Shortest Paths per Node

Social networks have been characterized by power law distribution of connections (Castellano et al., 2009, Muchnik et al., 2013b, Barabási and Albert, 1999, Barabási and Bonabeau, 2003). This means that a node in a network that is most connected has at least twice as many connections than the node that is second most connected. Mathematically it is expressed as

$$P(x) \propto x^{-\alpha} \ldots \ldots \ldots \ldots (20)$$

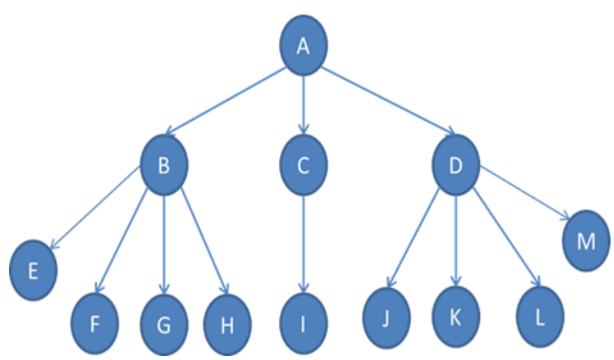
where P(x) is the probability distribution

α is the scaling parameter

Usually, many empirical measures cluster around a typical value, for example average height of an American. Even the largest deviations, which are exceptionally rare, are still only about a factor of two from the mean in either direction and hence the distribution can be well-characterized by quoting just its mean and standard deviation. However, not all distributions fit that pattern and those that do not fit are considered defective due to presence of data outliers (Clauset et al., 2009). In this case though, the data not fitting the standard pattern of mean and standard deviation leads to interesting characteristics like scale free structure in a network, as seen in section 4.3.3.1.2.

The scale free measure is a network level metric, that assess power law distribution of node connectivity (connections per node in the network). However, the scale free measure throws no light on the distribution of paths or the shortest paths amongst the nodes in a network, which explain the process of information spread and information speed (explained in 4.3.3.2.2). It is possible for a node to have low connectivity and still be responsible for large number of paths and shortest paths as shown in fig.14.

Figure 14: Power Law Distribution of Paths in a Network



Consider the directed network shown above in fig.14, A is connected to B, C, D. B is connected to E, F, G, H. C is connected to I. D is connected to J, K, L, M. B and D are more connected than A but A forms more number of paths than either B or D. Paths formed by A are A-B, A-C, A-D, A-B-E, A-B-F, A-B-G, A-B-H, A-C-I, A-D-J, A-D-K, A-D-L, A-D-M. Paths formed by B are B-E, B-F, B-G, and B-H. Paths formed by D are D-J, D-K, D-L, and D-M. Therefore, distribution of connectivity vs paths differs in a network.

In order to account for this structural characteristic of the network, I consider power law distribution of paths per node and power law distribution of shortest paths per node. These measures are experimental, as I have not come across any literature that states about the distribution paths and the shortest paths amongst nodes across networks and its impact. The relationship between the scale free metric of a network and the power law distribution of paths per node and power law distribution of shortest paths per node will be tested during the analysis in chapter 5. Also, in order to understand the relationship between power law distribution of paths per node and power law distribution of shortest paths per node with influence of node, power law distributions of eigenvector centrality (section 4.3.4.3) will be tested.

4.3.3.2 Network Flow (MV2)

The adjacency matrix provides information about the number of paths that exist in a graph (Wasserman and Faust, 1994). The order of the adjacency matrix conveys information about number of paths that exist in a graph with a particular path length (Wasserman and Faust, 1994). The path length is defined as number of nodes travelled to reach the destination node from the source node (Fronczak et al., 2004). The order, at which the calculation needs to stop, since all nodes are accessible to each other in the graph, is dictated by the diameter of the graph. A diameter of a graph is the longest shortest path required to connect any two nodes (Wasserman and Faust, 1994). In this section, I show the basics behind calculating metrics that pertain to network flow. These include the graph diameter (Shimbel, 1953), the number of geodesics, the number of paths, the length of the average geodesics and the average path length (Fronczak et al., 2004).

4.3.3.2.1 Shimbel Matrix (graph diameter) (Shimbel, 1953)

A Shimbel Matrix is a simple adaptation of the Adjacency Matrix. It holds the shortest path between nodes of a network, which is either lesser than or equal to the diameter of the graph. The Shimbel Matrix is constructed as shown in Figure 15:

Figure 15: First Order Shimbel Matrix

A^1 =

	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	0	0
C	1	1	0	1	1
D	1	0	1	0	0
E	0	0	1	0	0

D^1 =

	A	B	C	D	E
A	0	1	1	1	
B	1	0	1		
C	1	1	0	1	1
D	1		1	0	
E			1		0

The First Order Shimbel Matrix is constructed from the First Order Adjacency Matrix, where all direct links are kept. The number 1 in the Shimbel Matrix indicates that the shortest path is of 1st order (path length is 1). The diagonal elements are assigned the valued of 0 as shortest distance between a node and itself is 0. The cells which have a value of 0 in the A^1 matrix are left blank as the shortest path between those nodes is yet to occur.

Figure 16: Second Order Shimbel Matrix

A^2 =

	A	B	C	D	E
A	3	1	2	1	1
B	1	2	1	2	1
C	2	1	4	1	0
D	1	2	1	2	1
E	1	1	0	1	1

D^2 =

	A	B	C	D	E
A	0	1	1	1	2
B	1	0	1	2	2
C	1	1	0	1	1
D	1	2	1	0	2
E	2	2	1	2	0

The Second Order Shimbel Matrix is built from the empty cells of the First Order Shimbel Matrix and the Second Order Adjacency Matrix. A value of 2 is assigned to the empty cells of D^2 for which the corresponding value in A^2 are greater than 0. The number 2 in the Shimbel Matrix indicates that the shortest path of 2nd order (path length is 2). Since, all the cells in D^2 are occupied, we have identified that the highest path length of shortest paths in the graph is 2. Therefore, the diameter of the graph is 2.

4.3.3.2.2 Metrics for Network Flow (Fronczak et al., 2004)

In order to identify all the shortest paths, the Adjacency Matrix and the Shimbel Matrix need to be compared, as shown below:

Figure 17: Identifying the Shortest Paths

A^1

	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	0	0
C	1	1	0	1	1
D	1	0	1	0	0
E	0	0	1	0	0

D^1

	A	B	C	D	E
A	0	1	1	1	
B	1	0	1		
C	1	1	0	1	1
D	1		1	0	
E			1		0

P^1

	A	B	C	D	E
A	0	1	1	1	
B	1	0	1		
C	1	1	0	1	1
D	1		1	0	
E			1		0

A^2

	A	B	C	D	E
A	3	1	2	1	1
B	1	2	1	2	1
C	2	1	4	1	0
D	1	2	1	2	1
E	1	1	0	1	1

D^2

	A	B	C	D	E
A	0	1	1	1	2
B	1	0	1	2	2
C	1	1	0	1	1
D	1	2	1	0	2
E	2	2	1	2	0

P^2

	A	B	C	D	E	Sum
A	0	1	1	1	1	4
B	1	0	1	2	1	5
C	1	1	0	1	1	4
D	1	2	1	0	1	5
E	1	1	1	1	0	4
						22

The Shimbel Matrix gives the shortest path orders, and the Adjacency Matrix gives the total number of paths with specific path lengths. In P^1 we identify the shortest path of 1st order and in P^2 we fill the empty cells of P^1 with the shortest path of 2nd order. For example, consider the cell (A,E) in the P^2 matrix. A^2 indicates that the cell (A,E) has 1 path, D^2 indicates that the path from AE is of 2nd order. Therefore (A, E) in P^2 takes the value of 1, indicating that there exists one shortest path, which is of the second order between (A, E). The overall sum gives us the total number of shortest paths (geodesics) in the graph, which in this case are 22. The elements of P^2 convey the total number of shortest paths (geodesics) that exist between any two nodes. The shortest paths in P^2 show the speed with which all the nodes in the network can be reached. This defines the process of *speed of information spread*.

By adding all the values in A^1 and A^2 gives us the total number of paths in the graph, as shown below:

Figure 18: Total Paths

A^1

	A	B	C	D	E	Sum1
A	0	1	1	1	0	3
B	1	0	1	0	0	2
C	1	1	0	1	1	4
D	1	0	1	0	0	2
E	0	0	1	0	0	1

A^2

	A	B	C	D	E	Sum2
A	0	1	2	1	1	4
B	1	0	1	2	1	5
C	2	1	0	1	0	4
D	1	2	1	0	1	5
E	1	1	0	1	0	3

	Sum 1	Sum2	Total
A	3	4	7
B	2	5	7
C	4	4	8
D	2	5	7
E	1	3	4
Total number of paths			33

Therefore, the total number of paths in the graph is 33. The total paths show in how many ways the information from one particular node can reach the any other node in the network. This defines the *information spread process*.

The Average Geodesic Length (AGL) and the Average Path Length (APL) can be easily calculated by respectively dividing the total number of geodesics and the total number of paths by the number of nodes in the graph.

- AGL = 22/5 = 4.4……………………(21)
- APL =32/5 = 6.6……………………(22)

Average geodesic length and average path length are measures of ease of accessibility of nodes with a network through the speed and spread process.

4.3.4 Measuring the Dependent Variables

In section 2.6, the literature on eigenvector centrality and its role as the metric of influence (dependent variable), was discussed at length. It was also noted that centrality is an attribute of an actor, and centralization is a network. In the following section, I will describe how eigenvector centrality and eigenvector centralization are measured.

4.3.4.1. Measuring Eigenvector Centrality (Freeman, 1979)

Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Eigenvector centrality acknowledges that not all connections are equal. In general, connections to people who are themselves influential will lend a person more influence than connections to less influential people.

Denote the centrality of vertex "i" by " x_i"; then we can allow for this effect by making x_i, proportional to the average of the centralities of i's network neighbors:

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} A_{ij}\, x_j \dots \dots . . (23)$$

where λ denotes a constant

Defining the vector of centralities **x** = (x_1, x_2,…), we can rewrite this equation in matrix form as

$$\lambda\, x \ = \ A.x \dots \dots \dots . (24)$$

We see that **x** is an eigenvector of the adjacency matrix with eigenvalue ʎ. If we wish the centralities to be non-negative, then ʎ must be the largest eigenvalue of the adjacency matrix and **x** the corresponding eigenvector.

The eigenvector centrality defined in this way accords each vertex a centrality that depends both on the number and the quality of its connections: having a large number of connections still counts for something, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts.

As explained in section 3.2.2 and section 4.3.3.2.2, I am measuring two different processes within the network, information spread and speed of information spread. Total paths and Total Shortest Paths are respectively used as proxies for information spread and speed of information spread. *Therefore, the correlation coefficient of eigenvector centrality with total paths and total shortest paths will be used as a measure of influence with respect to information spread and speed of information spread processes.*

4.3.4.2. Measuring Eigenvector Centralization (Freeman, 1979)

Freeman (1979) also showed that the eigenvector centrality measures can be used to calculate network centralization as follows:

1. Compute the eigenvector centrality for each node in the network to determine the largest value.
2. Subtract each node's centrality from the largest centrality value within the network and sum the difference.
3. In a highly centralized network the sum of difference will be large; the sum of difference will be small in a less centralized network.
4. The measure is normalized by dividing the sum-of-difference value of the network under investigation with the largest possible value for the sum-of-difference in a network of equal size. This normalizes the value to a number between 0 and 1.

4.3.4.3. Power law distribution of Eigenvector Centrality per Node

In order to assess the effect of power law distribution of total paths per node and total shortest paths per node (section 4.3.3.1.4) on the influence of nodes within a network, I generate a power law distribution of eigenvector centrality, which will be compared with the network structure and network flow variables in analysis.

4.3.5 Control Variables

A large volume of research has been devoted to the development of algorithmic methods to analyze social networks (Danon et al., 2005). Nearly all of these methods have one thing in common: they are intended for the analysis of undirected network data. The common approach to analyzing social networks in directed networks has been simply to ignore the tie directions and apply algorithms designed for undirected networks (Leicht and Newman, 2008). This works reasonably well in some cases, although in others it does not. Even in the cases where it works, however, it is clear that in discarding the directions of ties a good deal of information about network structure is lost. This information could, at least in principle, allow us to make more accurate determinations about the nature of the social networks under study (Leicht and Newman, 2008). Therefore, I will consider the undirected and directed versions of the network. The directed network will be further analyzed considering only the information consumption patterns and information propagation patterns, as explained in section 3.2.2. These form the network constraints that will be used as control variables. Every network to be analyzed will be analyzed four times in the following forms:

1. Non-Directional Network
2. Directional Network
3. Information Consumption
4. Information Propagation

5. Scale

As explained in section 4.1, I would like to find out whether the social networks that discuss product categories in which content is consumed at high volumes behave differently from social networks that discuss product categories in which content is consumed at relatively low volumes. Therefore scale becomes a control variable in my research design.

In the following sections, I will discuss these issues in detail.

4.4 Validity and Reliability

In this study, I will conduct empirical research for hypotheses testing. The variables and measures used in this research are discussed in this section. It is important in hypothesis testing that type one error (α- error) and type two error (β- error) are eliminated. The α- error and the β- error are defined as follows: (Montgomery, 2008)

1. α- error: The study results lead to the rejection of null hypothesis even though it's true.
2. β- error: The study results lead to the acceptance of null hypothesis even though it's false.

Also, in order for the research to be viable, it is important to show the validity and reliability of the research design. The criteria that determine the validity of a research design are defined as follows:

1. *Construct Validity*: The degree to which both the independent and dependent variables accurately reflect or measure the constructs of interest. (Judd et al., 1991)

2. *Criterion Validity*: The extent to which one measure estimates or predicts the values of another measure or quality. (Cooper and Emory, 1995)

3. *Reliability*: The degree to which hypotheses are homogeneous and reflect the same underlying constructs. (Cronbach, 1990)

4.4.1. Eliminating α- error and β- error

In principle, one can compute any network measure for any network that is built on the basis of empirical data. Many conclusions can be drawn based on these network measures. Unfortunately, one cannot be confident that the network measure that has been computed is a true reflection of the network's structural features or a random variation. In order to overcome this predicament, Erdös and Rényi (1959) proposed comparing the network and the network measures of the network in question to the network and the network measures of a randomly generated network with same number of nodes and ties such that the every tie is chosen with equal probability (Erdös and Rényi, 1959).

This method is very similar to testing a mean using z-test. In a z-test, a sample of data is taken where the value of each data point is considered a random value. One single mean is calculated for the sample and every random value is compared with this mean. The mean in question here is the expected value. If the random value is different from the expected value, then one rejects the hypothesis that the true mean is equal to expected mean. In case of a network the random values are the ties. A single network is observed and the network measures are calculated for this network. These network measures are compared with the expected network measures of a network in which every tie has an equal probability. If the expected and the random network measures are different, then one can conclude that random network has different characteristics than the expected network. In case the network measure is non-existent for the network in question, then the network is considered to be random.

This method also prevents one from drawing the wrong conclusions because of a lack of reference point. For example, let's consider a network A with a clustering coefficient of 0.25 and a network B with a clustering coefficient of 0.5. It is easy to conclude that network B is highly clustered as compared to network A because 0.5 is greater than 0.25. But it may be that network A may have a higher clustering coefficient than one would expect in a random network, whereas network B has a clustering coefficient that is same as that of a random network. As a result, one can consider the cluster coefficient of network A as a true network feature and the clustering coefficient of network B as same as one might observe in a random network.

Kejzar et al. used such networks as the basis for modeling the dynamics of acquaintanceships (Kejzar et al., 2008). Donninger used this approach to derive the distribution of degree centralization (Donninger, 1986). Anderson et al. used it to simulate the distribution of degree and betweenness centralization (Anderson et al., 1999).

Therefore, in order to do meaningful comparisons of a network measure and eliminate α- error and β- error, I generate an Erdös-Rényi (E-R) random network which has the same number of nodes, the same number of ties and the same density. The networks that have similar clustering coefficient as a random network will be considered to have formed due to random process.

4.4.2. Construct Validity

The construct validity represents the degree to which both the independent and dependent variables accurately reflect or measure the constructs of interest (Judd et al., 1991). In hypothesis testing both the independent variables and the dependent variables should be decided prior to doing the analysis, so as to know how to measure them (Judd et al., 1991). Researchers can choose to use existing measurement scales, conduct exploratory preliminary studies, make theoretical considerations, or draw on experiences from practice (Judd et al., 1991). Using existing scales has been recommended, as it has added advantage of being able to compare results with previous studies in the same field (Judd et al., 1991). This research will use preexisting scale of measurement for all variables, as described in section 4.3. This study will also undertake factor analysis to assess construct validity (Cooper and Emory, 1995).

4.4.3. Criterion Validity

A criterion is a measure used to determine the accuracy of a decision. In psychometrics, criterion validity is a measure of how well a variable or a set of variables predicts the outcomes based on data from other variables (Murphy and Davidshofer, 1991, Pennington, 2003). Criterion validity measures the degree to which the predictor is adequate in capturing the relevant aspects of the criterion (Cooper and Emory, 1995). The correlation between the predictor and a measure of the outcome (or the criterion) provides an overall measure of the accuracy of predictions. The correlation between the predictor scores and criterion scores can be considered as a measure of the validity of decision (Murphy and Davidshofer, 1991). To confirm the criterion-related validity, a researcher can use correlation (Cooper and Emory, 1995). Therefore, this study will undertake correlational analysis to measure criterion validity.

4.4.4. Reliability

Cronbach's alpha will be used as a measure of internal consistency and by implication as a measure of reliability. Cronbach's alpha can be described as the number of test items and the average inter-correlation among the items (Cronbach, 1990). A commonly accepted rule of thumb for describing internal consistency using Cronbach's alpha is as follows:

Table 8: Cronbach Scale for Internal Consistency

Cronbach's alpha	**Internal consistency**
$\alpha \geq 0.9$	Excellent (High-Stakes testing)
$0.7 \leq \alpha < 0.9$	Good (Low-Stakes testing)
$0.6 \leq \alpha < 0.7$	Acceptable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

Hair et al. (Hair et al., 1998) and Field (Field, 2005) suggested the value of Cronbach's alpha should be higher than .70 for a reliable scale. The threshold value may decrease to .60 in exploratory research (Hair et al., 1998). Though a "high" value of alpha is often used as evidence that the items measure an underlying (or latent) construct, it does not imply that the measure is one-dimensional (Cortina, 1993). Therefore, factor analysis will be used to measure the dimensionality of variables.

4.5 Data Analysis

The key objective of this dissertation is to understand the impact of network structure and network flows on each other and their impact on the network phenomenon of influence. In previous sections, I have shown the model, the variables and the measures that will be used. In section 4.4, I have provided justification to ensure validity and reliability. In this section, I shall discuss the data analysis methods to be used.

4.5.1 Correlation Analysis

To assess the degree of interdependence between variables, this study will consider both correlation coefficient and statistical significance. Pearson's correlation coefficient is the most common measure of effect size. It is controlled to lie between -1 and 1 (Field, 2005). The effect size provides an objective measure between variables. The correlation matrix will be extremely useful for getting an idea of the relationships between dependent variables and independent variables. In this study, I will use two-tailed tests for statistical significance analysis of Pearson's correlation coefficient because the direction of moderating variables may have and positive or negative impact on the dependent variable. One -tailed tests are used when there is a specific direction to the hypothesis being tested, and two-tailed tests should be used when a relationship is expected, but the direction of the relationship is not predicted (Field, 2005).

4.5.2 Exploratory Factor Analysis

The goal of exploratory factor analysis is to find the smallest number of interpretable factors that can adequately explain the correlations among a set of variables (Field, 2005, p. 619). Items that are grouped together are presumed to be measuring the same underlying construct (Kerlinger and Lee, 1964). Exploratory factor analysis is a useful tool for understanding the dimensionality of a set of variables and also for isolating variables that do not represent the dimensions (Field, 2005, pp.622). It is extremely helpful during pilot work in the development of a set of items as all loadings are free to vary. This analysis will be conducted using varimax rotation as the rotation procedure. A Scree test (Cattell, 1965) will then be conducted to produce a more interpretable solution. Factors need to explain at least 80% of cumulative variance. The factors will be examined and given a descriptive title that represented the characteristics of the constructs.

4.5.3 Regression Analysis

This study intends to use linear regression analysis, because regression analysis helps determine the relative impact of the independent variables on the dependent variable. To provide a statistical test of the model's ability to predict the dependent variables, the value R square and the adjusted R square will be used (Field, 2005, pp.179).

R-squared is a statistical measure of how close the data is to the fitted regression line. It is also known as the coefficient of determination. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} \dots \dots \dots \dots (25)$$

R-squared is always between 0 and 100%. 0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits.

The adjusted R square can be used to assess how well the model is able to predict the outcome in a different sample. Field mentions cross-validation is a way to assess the accuracy of a model across different samples (Field, 2005 p. 171). In regression, the value of adjusted R square should be very close to the value of R square.

Table 9 below presents abbreviations of all the variables. In following sections an 'x' is added in place of 'ud', 'd', 'in' or 'out' to any of the variables mention below to indicate the applicability of the variable to all (undirected, directed, consumption and propagation) networks.

Table 9: Variables Definitions[21]

Variable	Explanation
Nodes	Number of nodes in a network
Edges_ud	Number of ties in a undirected network
Edges_d	Number of ties in a directed network
Reciprocity	Reciprocal ties in a directed network
Den_ud	Density of an undirected network
Den_d	Density of a directed network
CC_ud	Clustering coefficient of an undirected network
CC_d	Clustering coefficient of a directed network
GD_ud	Graph diameter of an undirected network
Tpaths_ud	Total paths in an undirected network
TSpaths_ud	Total shortest paths in an undirected network
AvgPL_ud	Average path length in an undirected network
AvgGL_ud	Average geodesic length in an undirected network
GD_d	Graph diameter of a directed network
Tpaths_d	Total paths in a directed network
TSpaths_d	Total shortest paths in a directed network
AvgPL_d	Average path length in a directed network
AvgGL_d	Average geodesic length in a directed network
S_ud	Scale free metric for an undirected network
S_d	Scale free metric for a directed network
S_con	Scale free metric for a consumption network
S_pro	Scale free metric for a propagation network
R_ud	Assortativity of an undirected network
R_d	Assortativity of a directed network
R_con	Assortativity of a consumption network
R_pro	Assortativity of a propagation network
SMSP_ud	Small world metric for an undirected network
SMSP_d	Small world metric for a directed network
PL_TpudN	Power law distribution of total paths per node in an undirected network
PL_TpdN	Power law distribution of total paths per node in a directed network
PL_TpinN	Power law distribution of total incoming paths per node in a consumption network
PL_TpoutN	Power law distribution of total outgoing paths per node in a

[21] The terms "edges" and "ties" are used interchangeably in graph theory. The word "ties" is preferred in this dissertation. However, the word edges appears in some aspects of statistical analysis. For example, in Table 9 and henceforth the variables Edges_ud and Edges_d refer to ties.

	propagation network
PL_TSpudN	Power law distribution of total shortest paths per node in an undirected network
PL_TSpdN	Power law distribution of total shortest paths per node in a directed network
PL_TSpinN	Power law distribution of total shortest incoming paths per node in a consumption network
PL_TSpoutN	Power law distribution of total shortest outgoing paths per node in a propagation network
ECud	Eigenvector centralization in an undirected network
ECd	Eigenvector centralization in a directed network
ECin	Eigenvector centralization in a consumption network
ECout	Eigenvector centralization in a propagation network
PL_EVCudN	Power law distribution of eigenvector centrality per node in an undirected network
PL_EVCdN	Power law distribution of directed-eigenvector centrality per node in a directed network
PL_EVCinN	Power law distribution of in-eigenvector centrality per node in a consumption network
PL_EVCoutN	Power law distribution of out-eigenvector centrality per node in a propagation network
EVCud_TpudN	Correlation coefficient of eigenvector centrality vs. total paths per node in an undirected network
EVCd_TpdN	Correlation coefficient of directed-eigenvector centrality vs. total directed paths per node in a directed network
EVCin_TpinN	Correlation coefficient of in-eigenvector centrality vs. total incoming paths per node in a consumption network
EVCout_TpoutN	Correlation coefficient of out-eigenvector centrality vs. total outgoing paths per node in a propagation network
EVCud_TSpudN	Correlation coefficient of eigenvector centrality vs. total shortest paths per node in an undirected network
EVCd_TSpdN	Correlation coefficient of directed-eigenvector centrality vs. total shortest directed paths per node in a directed network
EVCin_TSpinN	Correlation coefficient of in-eigenvector centrality vs. total shortest incoming paths per node in a consumption network
EVCout_TSpoutN	Correlation coefficient of out-eigenvector centrality vs. total shortest outgoing paths per node in a propagation network
CCudran	Clustering coefficient of an undirected random network (E-R network)
CCdran	Clustering coefficient of a directed random network (E-R network)

5. Analysis and Results

Data was collected on Twitter for the six product categories described in section 4.1.2.2; using the data collection process described in section 4.2. In this chapter, the results of the study conducted for this dissertation will be explained. This chapter is divided into 6 sections. In section 5.1, I start by providing an overview of metadata (described in section 4.2). I start the analysis process by testing the assumption made in section 4.1.2.2: the more popular product categories on YouTube will generate bigger communities on Twitter. In Section 5.2, I provide an overview of the daily patterns seen in the independent variables, moderating variables and dependent variables. Section 5.3, discusses whether the networks formed are random or not, in order to eliminate α-error and β- error, as explained in section 4.4.1. In Section 5.4, findings pertaining to factor analysis and correlation analysis are provided. In Section 5.5, findings pertaining to regression analysis that address RQ1, RQ2, RQ3 and RQ4 are provided.

The detailed overview, descriptive statistics of independent variables, moderating variables and dependent variables for all six product categories for undirected, directed, consumption and propagation networks are provided in case reports shared in Appendix A. The case reports also contain statistical analysis for the chosen product categories (correlation analysis, factor analysis and regression analysis) for the undirected, the directed, the consumption and the propagation networks.

5.1 Overview of Metadata

Table 10: Metadata Overview

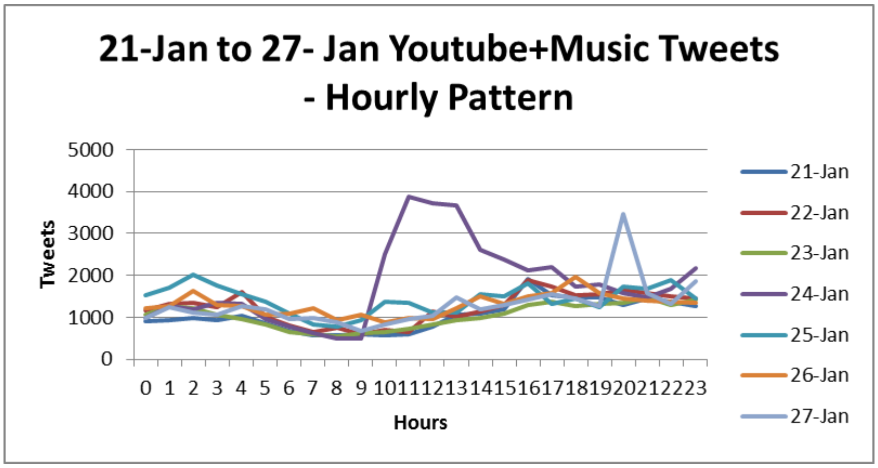
Cases	Product Category	Popularity	% of All Videos	Total_ Tweets	Broadcast _Tweets	Engaged_ Tweets	Community _Tweets	Total_ People	Broadcast _People	Engaged_ People	Largest_ Community
1	Music	High	30.7	3,097,847	713,824	2,384,023	1,586,149	2,586,586	898,282	1,688,304	1,456,770
2	Entertainment	High	14.59	44,984	10,762	34,222	16,365	45,236	16,670	28,566	15,822
3	Comedy	Medium	5.2	94,111	33,350	60,761	25,624	83,175	37,456	45,719	24,555
4	Sports	Medium	6	129,182	67,476	61,706	32,778	77,617	25,776	51,841	29,998
5	Howto	Low	3.1	10,856	3,213	7,643	4,299	10,557	4,082	6,475	4,203
6	Science	Low	2.86	49,332	13,462	35,870	22,598	52,785	20,157	32,628	21,277

As shown in Table 10, the six chosen cases were categorized based on their popularity. They were binned into three categories: high, medium and low. (Definitions for the Metadata are shown in Table 8 (section 4.2).) Table 10 shows cumulative numbers for Total Tweets, Broadcast Tweets, Engaged Tweets, Community Tweets and also Total People, Broadcast People, Engaged People and Largest Community over a period of 91 days (31/12/2013 to 31/03/2014). Total Tweets and Total People show the total number of tweets collected and the total daily unique people involved in these tweets respectively. Broadcast Tweets and Broadcast People respectively show the total tweets that were categorized as broadcast and total daily unique people involved in the broadcast activity. The definition of broadcast is provided in section 4.2. Engaged Tweets and Engaged People respectively show total tweets in which a conversation was happening and total number of daily unique people involved in conversations. Finally, Community Tweets and Largest Community show all tweets and people engaged in collective conversations. Distinction between collective conversation and isolated conversation is described in section 4.2. Their daily values are shown in Appendix A.

As per assumption from section 4.1.2, products categorized as high were supposed to generate communities that were bigger in size, both in terms of number of tweets and people involved, than products that were categorized medium or small. As seen from the Table 10 above, this assumption does not hold true. The “Entertainment” category, which was categorized as “high” based on YouTube popularity, generated 43,377 total tweets whereas the “Comedy” category, which was categorized as medium, generated 94,111 total tweets over the same period of time. Similarly, the “Sports” category generated more tweets than the “Entertainment” category. This trend can also be seen for the community sizes, of “Comedy” and “Sports”, both in terms of number of tweets and people involved. “Comedy” and “Sports” had a larger number of community participants as compared to “Entertainment” community.

The tweets collected show a daily pattern of tweeting. For example, Figure 19 below shows an hourly pattern for data collected in “Music” category between 21st Jan, 2014 to 27th Jan 2014.

Figure 19: Hourly Patterns of Tweets between 21st Jan, 2014 to 27th Jan, 2014 in Music Category



From Figure 19 it can be clearly seen that the tweets have a recurring pattern on a 24-hour basis with the exception of a few bumps on Jan. 24th and Jan.27th. These bumps are associated with the following events:

- 24th Seoul Music Awards - 22nd Jan, 2014
- 59th Filmfare Awards - 24th Jan, 2014
- 56th Annual Grammy Awards- 26th Jan, 2014

The impact (with delay) of the Seoul Music Awards, held on Jan. 22nd, can be seen on the tweet volume of Jan. 24th. The 24 hour pattern is consistent with previous large-scale stuidies undertaken on Twitter (Frank et al., 2013, Dodds et al., 2011). The 24-hour cycle started in accordance with Greenwich Mean Time (GMT) in this study. Future research can be conducted to identify the impact of changing the start times of 24 hour cycle on the results.

5.2 Overview of Variables

In this section, I discuss the important daily patterns seen in the independent variables, moderating variables and dependent variables of this study for all six product categories. The detailed patterns of all variables are provided in the Case Overview sections of Appendix A.

5.2.1 Independent Variables

Independent variables in this study consist of the number of nodes (Nodes), number of ties in the undirected network and directed network (Edges_ud, Edges_d), Reciprocity, Density Undirected (Den_ud), Density Directed (Den_d), Clustering Coefficient Undirected (CC_ud) and Clustering Coefficient Directed (CC_d). Definitions and explanations of each variable are provided in section 4.3. Detailed patterns of each variable for all six product categories are shown in Appendix A (A.1.3, A.2.3, A.3.3, A.4.3, A.5.3, A.6.3)

For all product categories, the number of directed ties (Edges_d) in the network and the total number of nodes (Nodes) follow the same pattern. The number of nodes (Nodes) forming the largest community increases in tandem with the number of ties (Edges_ud, Edges_d) in the community. The numbers of undirected ties (Edges_ud) is greater than the number of directed ties (Edges_d), because in an undirected network every directed tie is considered to be symmetric. Hence every tie is counted twice, except for the ties that are already symmetric in the directed network.

The reciprocity (Reciprocity) is 100% for any undirected network, as all ties are considered to be symmetric. In case of a directed network, the product categories “Howto” and “Science” seldom form networks that are reciprocal. The product categories of “Entertainment”, “Comedy” and “Sports” form networks that are intermittently reciprocal. “Music” is the only product category that forms a reciprocal network every day for the duration of analysis.

For all product categories, the undirected networks are denser than the directed networks (CC_ud >CC_d). This is not surprising, since all the non-symmetric ties in a directed network are counted twice in the corresponding undirected network. Product categories that form larger networks (e.g. Music) seem to be less dense than product categories that form smaller networks (e.g. Howto). This is true for both directed and undirected networks.

The directed networks of all product categories other than the “Music” category seldom show Clustering Coefficients (CC_d) above 0. “Music” is the only directed network that shows Clustering Coefficient (CC_d) above 0 on a daily basis. The directed network shows a higher Clustering Coefficient than the undirected network in “Music” product category. “Howto” is the only product category whose undirected networks seldom show Clustering Coefficients (CC_ud) above 0.

5.2.2 Moderating Variables

5.2.2.1 Network Flow Variables (MV2)

Network flow variables in this study consist of the Total Number of Paths in an undirected network (Tpaths_ud); the Total Number of Shortest Paths in an undirected network (TSpaths_ud); Total Paths in a directed network (Tpaths_d); Total Shortest Paths in a directed network (TSpaths_d); Average Path Length and Average Geodesic Length for both directed and undirected networks (AvgPL_ud, AvgGL_ud, AvgPL_d, AvgGL_d); and the Graph Diameter for both directed and undirected networks (GD_ud, GD_d). Definitions and explanations of each of these variables have been provided in section 4.3. Detailed patterns of each variable for all six product categories are shown in Appendix A (A.1.5, A.2.5, A.3.5, A.4.5, A.5.5, A.6.5).

For all product categories, Total Paths in the undirected networks (Tpaths_ud) is greater than Total Shortest Paths in the undirected network (TSpaths_ud), Total Paths in the directed network (Tpaths_d) and Total Shortest Paths in the directed networks (TSpaths_d). As the size of the network formed increases, the difference between the Total Paths in the undirected network (Tpaths_ud) and the Total Shortest Paths in the undirected network (TSpaths_ud) increases by orders of magnitude. The difference between the Total Paths in the directed network (Tpaths_d) and the Total Shortest Paths in the directed network (TSpaths_d) is a lot less than the difference between the Total Paths in the undirected network (Tpaths_ud) and the Total Shortest Paths in the undirected network (TSpaths_ud). As the size of the network formed decreases, difference between the Total Paths in directed network (Tpaths_d) and the Total Shortest Paths in directed (TSpaths_d) network is almost negligible.

Similar trends as the Total Paths (Tpaths_x) and the Total Shortest Paths (TSpaths_x) are seen with respect to the Average Path Lengths (AvgPL_x) and the Average Geodesic Lengths (AvgGL_x). As the size of the networks increases, the difference between Average Path Length in the undirected network (AvgPL_ud) and the Average Geodesic Length in the undirected network (AvgGL_ud) increases by orders of magnitude. The difference between the Average Path Lengths in the directed network (AvgPL_d) and the Average Geodesic Lengths in the directed network (AvgGL_d) is lot less than the difference between the Average Path Lengths and the Average Geodesic lengths in the undirected network (AvgPL_ud, AvgGL_ud). As the size of the network formed decreases, difference between the Average Path Lengths in the directed network (AvgPL_d) and the Average Geodesic Lengths (AvgGL_d) in the directed network is almost negligible.

The Graph Diameter of the undirected network (GD_ud) is greater than the Graph Diameter of the directed network (GD_d). The magnitude of the Graph Diameter in both the directed and the undirected networks increases as the size of the network increases. It is also noteworthy that in all cases the Graph Diameter of the undirected (GD_ud) and the Average Path Length of the directed networks (AvgGL_d) track pretty closely.

5.2.2.2 Network Structure Variables (MV1)

The network structure variables in this study consist of the Scale Free metric (S_x), the Assortativity (R_x), the Small World metric (SMSP_x), the Total Number of Paths and the Shortest Paths Power Law Distribution per Node (PL_TpxN, PL_TSpxN). Definitions and explanations of each of these variables have been provided in section 4.3. Detailed patterns of each variable for all six product categories are shown in Appendix A (A.1.4, A.2.4, A.3.4, A.4.4, A.5.4, A.6.4).

For all product categories, the Scale Free metric for the directed and the undirected networks (S_ud, S_d) follow a similar pattern. The consumption network and the propagation network (S_con, S_pro) follow very different patterns. In the "Music", "Sports" and "Howto" categories, the propagation network is more Scale Free than the consumption network (S_con<S_pro). This trend is consistent over the duration of analysis for "Music" category whereas for the "Sports" and the "Howto" categories the trend is intermittent. In the "Entertainment", "Comedy" and "Science" categories consumption network is more Scale Free than the propagation network (S_con>S_pro). This trend is consistent over the duration of analysis for the "Comedy" category whereas for the "Entertainment" and the "Science" categories the trend is intermittent.

For all product categories, the undirected networks are more Disassortative than the directed networks, the consumption networks or the propagation networks (R_ud < (R_d, R_con, R_pro)). In the "Music" category, the consumption network is more

Disassortative than the propagation network (R_con<R_pro). In the “Entertainment” category, the propagation network is Disassortative (R_con>R_pro), whereas the consumption network toggles between being Assortative and Disassortative. In the “Comedy” category, the consumption network is always Disassortative, whereas the propagation is relatively less Disassortative and sometimes toggles to being Assortative. In the “Sports” category, both the consumption and the propagation network toggle between being Assortative and Disassortative. In the “Howto” category, the consumption network is more often Assortative than the propagation network. Both the consumption network and the propagation network are consistently Disassortative. In the “Science” category, the propagation network is consistently more Assortative than the consumption network.

The Small World measures for the consumption and propagation networks are the same as the ones for the directed network. The directed networks show stronger Small World behavior than the undirected networks in the “Music” category (SMSP_d >SMSP_ud). In the “Entertainment”, “Music”, “Howto” and “Science” categories, the directed and the undirected networks don’t show any significant Small World behavior. In the “Sports” category, the undirected network shows intermittent Small World behavior whereas the directed networks don’t show any Small World behavior.

In all categories, the distribution of undirected paths per node (PL_TpudN) shows a better power law behavior than the distribution of shortest paths per node (PL_TSpudN). As the scale of the network reduces the power law behavior of directed paths per node (PL_TpdN), incoming paths per node (PL_TpinN) and the outgoing paths per node (PL_TpoutN) becomes erratic.

5.2.3 Dependent Variables

The dependent variables in this study consist of the Eigenvector Centralization (ECx), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCxN), Clustering Coefficients of Eigenvector Centrality vs. Total Paths per Node (EVCx_TpxN) and Clustering Coefficients of Eigenvector Centrality vs. Total Paths per Node (EVCx_TSpxN). Definitions and explanations of each of these variables have been provided in section 4.3. Detailed patterns of each variable for all the six product categories are shown in Appendix A (A.1.6, A.2.6, A.3.6, A.4.6, A.5.6, A.6.6).

In all product categories, the undirected networks show better Eigenvector Centralization (ECud) than the directed (ECd) networks, the consumption (ECin) networks or the propagation (ECout) networks. The consumption and propagation networks exhibit same level of Eigenvector Centralization (ECin=ECout). The directed networks have the least Eigenvector Centralization.

The distribution of the Eigenvector Centrality across the nodes for the “Music” category exhibits a similar power law pattern (PL_EVCxN) for all networks (undirected, directed, consumption and propagation) for the whole period of time under investigation. In all other product categories, only the Eigenvector Centrality values of the undirected network are consistently distributed in a power law distribution pattern

(PL_EVCudN) for the whole period of time under investigation. In the directed, the consumption and the propagation network the distribution of Eigenvector Centrality follows a power law distribution pattern (PL_EVCinN, PL_EVCoutN) only for a portion of the period of time under investigation.

In all product categories, there is a significant correlation between the Eigenvector Centrality of a node with respect to the number of paths from a node in undirected network (EVCud_TpudN). There is no significant correlation between Eigenvector Centrality of a node with respect to shortest paths from a node in undirected network (EVCud_TSpudN). In the propagation network, for all product categories, there is no significant correlation either between the Eigenvector Centrality of a node with respect to the number of paths (EVCout_TpoutN) or between the Eigenvector Centrality of a node with respect to the number of shortest paths (EVCout_TSpoutN). In the directed and the consumption networks, for all product categories, the correlation coefficients of both Eigenvector Centrality of a node with respect to the number of paths (EVCd_TpdN, EVCin_TpinN) and Eigenvector Centrality of a node with respect to the number of shortest paths (EVCd_TSpdN, EVCin_TSpinN), increases significantly as the scale of the network reduce.

5.3 Random vs. Non-Random Networks

The product categories being analyzed are extremely dynamic and show high levels of variability from day to day, as shown in Table 11 below.

Table 11: Maximum and Minimum Daily Values

Product Category	Max.Total Tweets	Max. Total People	Min.Total Tweets	Min. Total People	Largest Community Tweets	Largest Community People	Smallest Community Tweets	Smallest Community People
Music	62,380	59,666	19,700	18,333	48,720	47,630	10,830	10,324
Entertainment	1,771	2,263	207	243	1,113	1,812	35	35
Comedy	2,178	1,968	508	526	832	833	131	130
Sports	8,562	8,624	333	360	7,881	7,882	108	108
Howto	2,448	1,279	37	42	1,370	1,213	4	5
Science	1,757	1,708	277	300	634	461	130	130

For example, in the "Music" category, the maximum of the total number of daily tweets and the maximum of the total number of daily unique people observed on a single day (the daily uniques) are 62,380 and 59,666, respectively. Similarly, the minimum of the total number of daily tweets and the minimum of the number of daily uniques are 19,700 and 18,333, respectively. The size of the largest community on a particular day and the largest number of community tweets on that day also seem to follow the trend of total people and total tweets. The largest number of daily community tweets and the largest number of daily unique people are 48,720 and 47,630, respectively. Similarly, the smallest number of daily community tweets and the smallest number of daily unique people are 10,830 and 10,324, respectively. The daily values are shown in Appendix A.

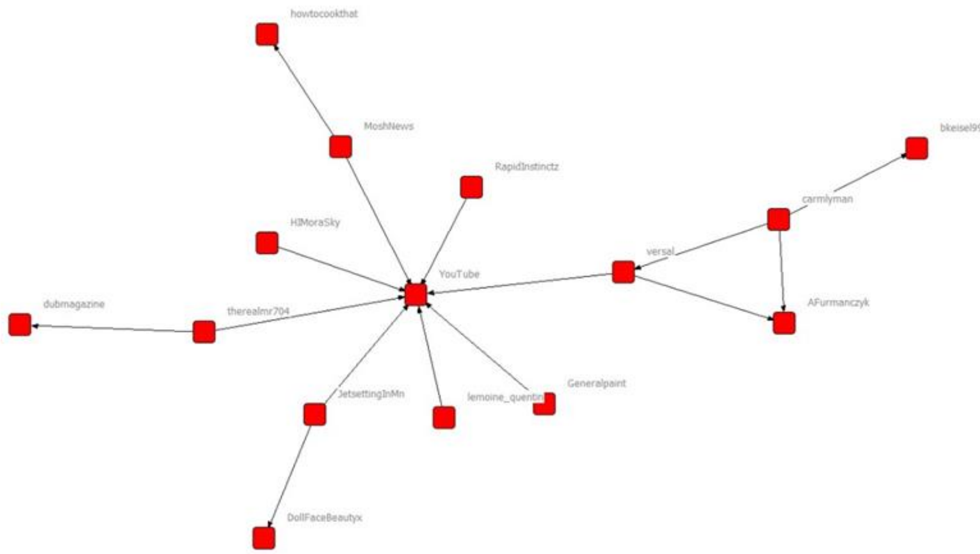
In order to understand which community networks on a given day are formed randomly and in order to eliminate α- error and β- error(as explained in section 4.4.1), I compare the Clustering Coefficients of both undirected and directed networks (CC_ud and CC_d) with their corresponding random (Erdös-Renyi, E-R) networks (CCudran and

CCdran). If the Clustering Coefficients of the undirected and directed (CC_ud and CC_d) "Music" networks are equal to those of the E-R random network (CCudran and CCdran) , then the directed and undirected networks are considered to be random, if they are not equal, then they are not random. If the Clustering Coefficient (CC_ud and CC_d) for the observed network is zero, then they are considered random as the network forms a star network.

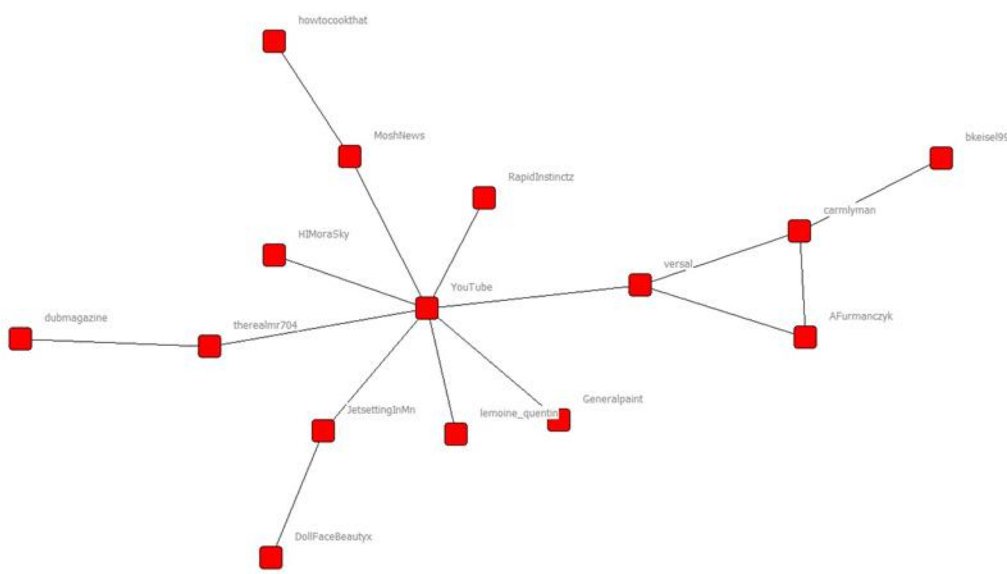
To further elucidate, consider the "Howto" community formed on January 6th, 2014 (details shown in Appendix A). The network has 15 nodes and 15 directed ties (self-ties are ignored). For the undirected version of the network all ties are considered symmetric, therefore there are 30 ties as shown in Figure 20 below.

In the directed version of the network, there is no transitivity i.e., the information only moves from a node to the connected node in a single direction. The information does not go beyond the connected node. There are no reciprocal relationships or instances where two different nodes connected to a node exchanging information with each other. The Clustering Coefficient (CC_d) of the network is 0. Therefore, the directed network is a random network.

Figure 20: “Howto” Network Jan 6th, 2014 (a) Directed (b) Undirected



(a)

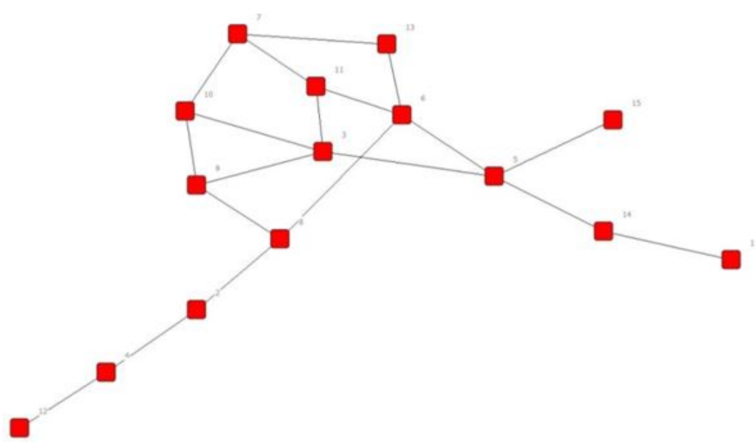


(b)

In the undirected network, in which all relationships are bi-directional, the information flows beyond just the connected node. There are reciprocal relationships and instances where two different nodes connected to a node, exchange information with each other. The Clustering Coefficient (CC_ud) of the undirected network (in fig. 19(b)) is 0.07894. The undirected network still needs to be compared to an equivalent E-R random undirected network (CCudran) to ascertain if it's random or not.

The equivalent E-R random undirected network is shown in Figure 21 below. As can be seen in Figure 21, there are instances where two different nodes connected to a node, exchange information with each other. The Clustering Coefficient of the undirected network (CCudran) is 0.133333. Comparing the Clustering Coefficients of the undirected network (CC_ud) and its equivalent random undirected network (CCudran), it is clear that the undirected network is not a random network.

Figure 21: Equivalent E-R Random Undirected Network



To understand the daily status of the networks formed for the six chosen product categories, the Clustering Coefficients of the daily undirected (CC_ud, in blue) and the daily directed networks (CC_d, in blue) were compared to the Clustering Coefficients of their respective random undirected and directed networks (CCudran and CCdran, both in red).

Figure 22: Clustering Coefficient of undirected network vs equivalent random undirected network (a) Music (b) Entertainment (c) Comedy (d) Sports (e) Howto (f) Science

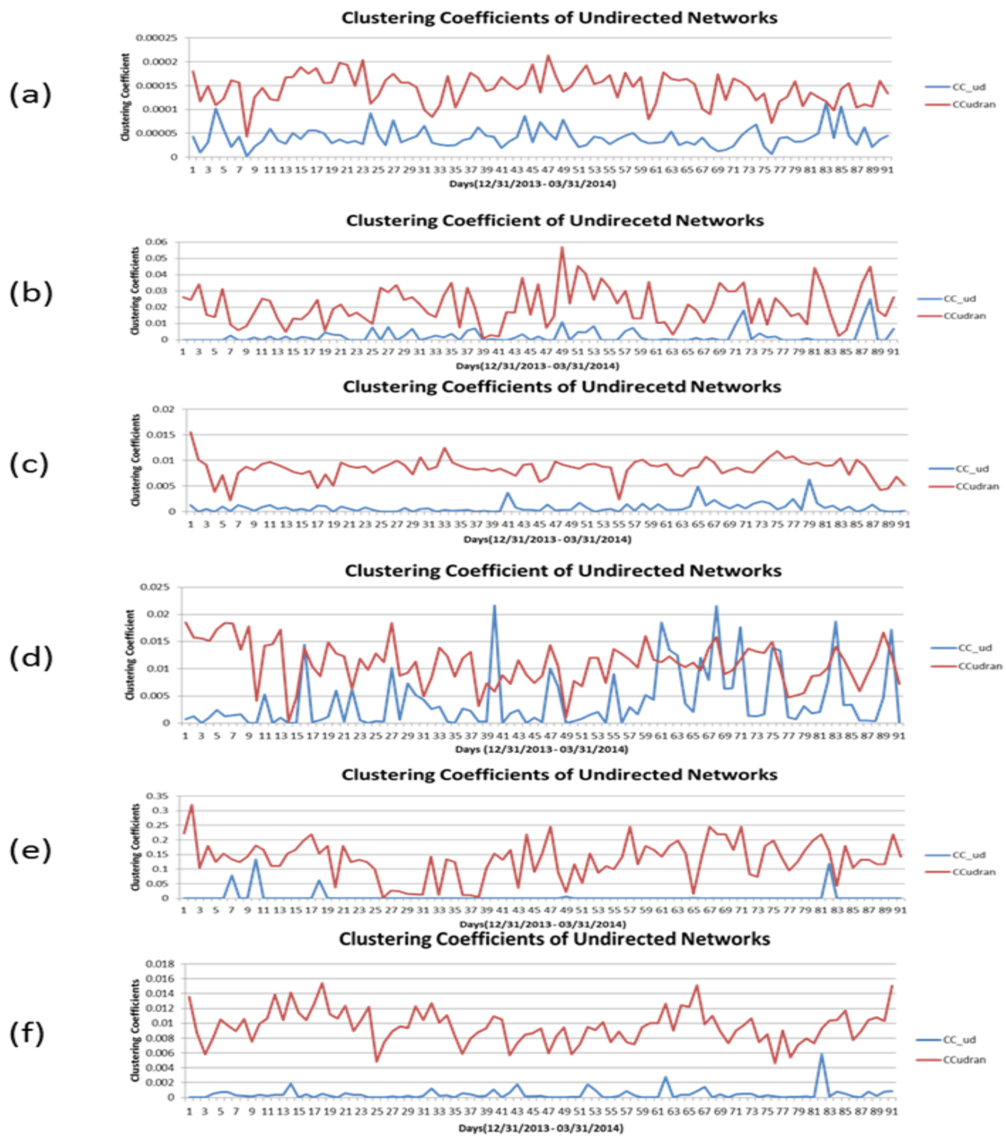


Figure 23: : Clustering Coefficient of directed network vs equivalent random directed network (a) Music (b) Entertainment (c) Comedy (d) Sports (e) Howto (f) Science

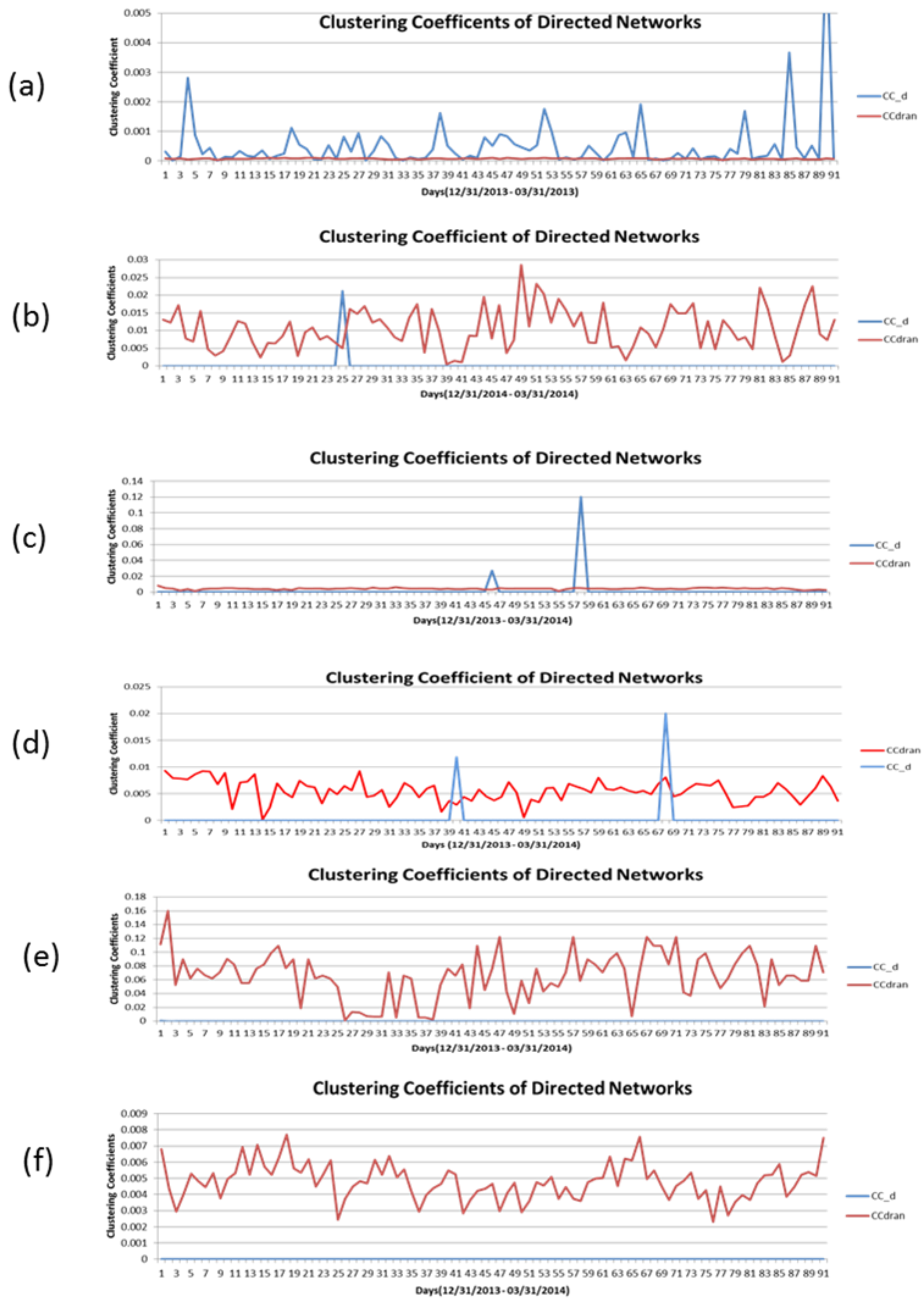


Figure 22 compares the Clustering Coefficients of undirected networks (CC_ud) to their equivalent random undirected networks (CCudran). Except for Figure 22 (e), the “Howto” undirected network, all the other undirected networks are non-random. For the “Howto” undirected networks the Clustering Coefficient (CC_ud) is zero on most of the days. For rest of the networks the Clustering Coefficient (CC_ud) displays a very distinct pattern from that of a random undirected network (CCudran).

Figure 23 shows the comparison Clustering Coefficients of directed networks (CC_d) with their equivalent random directed networks (CCdran). Except for Figure 23 (a), the “Music” directed network, all the other directed networks are random. For “Music” undirected networks the Clustering Coefficient (CC_d) displays a very distinct pattern from that of a random directed network (CCdran). For rest of the networks the Clustering Coefficient (CC_d) is zero on most of the days.

The consumption and the propagation networks emanate from the directed network. Hence, they follow the same pattern as the directed network. The results for all the networks are summarized in the Table 12 below.

Table 12: Random or Not Random Status of Six Product Category Networks

Cases	Product Category	Undirected	Directed	Consumption	Propagation
1	Music	Not Random	Not Random	Not Random	Not Random
2	Entertainment	Not Random	Random	Random	Random
3	Comedy	Not Random	Random	Random	Random
4	Sports	Not Random	Random	Random	Random
5	Howto	Random	Random	Random	Random
6	Science	Not Random	Random	Random	Random

From Table 12 above, it can be seen that for the "Howto" product category both the undirected and the directed network are random. Therefore, it will be removed from further analysis. For the "Music" product category both the directed and the undirected networks are not-random. For all the other product categories the undirected networks are not-random, while the directed networks are random.

As seen in section 2.2 of literature research, most theories of social phenomena talk about the impact of network flow on the social phenomenon. For example, Rogers talks about the importance of interpersonal communication within a social system for diffusion of innovation (Rogers, 2003). Granovetter suggests that weak ties are the sources of new information that flows into the network from the outside (Granovetter, 1973). However, for the most part, during the analysis they consider the network to be undirected. Though they allude to the existence of directionality, they do not consider them explicitly in their analysis process. Therefore, in this study, for all the product categories that have non-random undirected networks, directed networks will also be considered for analysis, even if they are random. Hence, all the undirected networks are non-random networks. In directed networks, only the networks pertaining to Music are non-random, whereas the rest are random networks.

5.4 Factor Analysis and Correlations

In this section, I present the results of factor analysis and correlation analysis undertaken for this study. The detailed factor analysis and the correlation analysis of all variables are shown in Appendix A (A.1.7.1.2, A.1.7.2.2, A.1.7.3.2, A.1.7.4.2, A.2.7.1.2, A.2.7.2.2, A.2.7.3.2, A.2.7.4.2, A.3.7.1.2, A.3.7.2.2, A.3.7.3.2, A.3.7.4.2, A.4.7.1.2, A.4.7.2.2, A.4.7.3.2, A.4.7.4.2, A.6.7.1.2, A.6.7.2.2, A.6.7.3.2, A.6.7.4.2) and Appendix A (A.1.7.1.1, A.1.7.2.1, A.1.7.3.1, A.1.7.4.1, A.2.7.1.1, A.2.7.2.1, A.2.7.3.1, A.2.7.4.1, A.3.7.1.1, A.3.7.2.1, A.3.7.3.1, A.3.7.4.1, A.4.7.1.1, A.4.7.2.1, A.4.7.3.1, A.4.7.4.1, A.6.7.1.1, A.6.7.2.1, A.6.7.3.1, A.6.7.4.1) respectively. The goal of the factor analysis in this study is to understand if the variables described as the independent variables, the moderating variables and the dependent variables for the undirected network, directed network, consumption network and the propagation network of the selected product categories measure the same constructs. Thus the factors formed are indicative of latent processes happening within the networks under consideration.

Exploratory factor analysis was used in this study to generate factors that explain the shared variability in the variables. One of the main problems in application of exploratory factor analysis is deciding how many factors to retain. In general, the best known and most utilized method is the one proposed by Kaiser, which suggests that only factors that have eigenvalues greater than one should be retained for interpretation (Kaiser, 1960). Fabrigar, et al. (1999) point out three problems with Kaiser's rule (Fabrigar et al., 1999). They are:

1. This method was proposed for principal component analysis (PCA) and not for exploratory factor analysis.
2. This rule leads to arbitrary selection of factors. It does not make to sense to regard a factor with eigenvalue of 1.01 as a valid factor and disregard a factor with eigenvalue of 0.99 as an invalid factor.
3. This rule tends to overestimate the number of factors in some cases and underestimate the number of factors in other cases.

The other popular method is scree test (Cattell, 1966). This method involves visual exploration of a graphical representation of eigenvalues, in which the eigenvalues are linked with a line and presented in a descending order. The point at which the line levels off is the point that divides the major factors from the trivial factors.

Because of the deficiencies of the Kaiser rule and the subjectivity of the scree test, I do not use these methods for factor extraction in this study.

A third method, suggested in the literature is to retain the number of factors that account for certain percentage of variance extracted. The majority of the literature suggests that 75% – 90% of the variance should be accounted for (George, 1989). This method seems more suitable for the exploratory research being undertaken in this study. In practice, for the purpose of this study, the factors that emerge from the factor analysis when 80% of the variance is explained are more meaningful and interpretable. Therefore, factors generated from the exploratory factor analysis are set to account for at least 80% of the variance. Varimax rotation as the rotation procedure is used to make the factors more interpretable.

The Kaiser-Mayer-Olkin (KMO) measure of sampling adequacy was used to test the proportion of variance in the variables that might be caused by the underlying factors. The results of factor analysis were considered, only if the KMO value was greater than 0.5. Bartlett's test of sphericity was used to test if there are correlations in the data that are appropriate for the factor analysis. The results of the factor analysis were considered, only if the significance (p-value) of Bartlett's test of sphericity were less than 0.05.

To ensure the reliability of the factors formed, only factors with Cronbach's alpha value greater than 0.6 are considered. To verify the criteria related validity of the factors, I use correlation analysis (detailed correlation analysis is shown in Appendix A). The results of the factor analysis are shown in Table 13 below (detailed factor analysis is shown in Appendix A). (Note: Cronbach's alpha in Table 13 below is not used to compare the factors. The values of Cronbach's alpha are only being shown to ascertain that the factors formed as a result of factor analysis are reliable.)

Table 13: Factors formed along with their Cronbach alpha values

NETWORK TYPE	VARIABLES	FACTORS (CRONBACH'S ALPHA)				
		Music	Entertainment	Comedy	Sports	Science
Undirected	Independent	Size(0.994)	Size (0.999)	Size (0.995)	Size (0.998)	Size (0.997)
	Network Structure (MV1)	NA	NA	NA	NA	NA
	Network Flow (MV2)	Spread(0.989)	Spread and Speed(0.937)	Spread and Speed(0.937)	Spread and Speed (0.965)	Spread and Speed (0.912)
	Dependent	NA	NA	NA	NA	NA
Directed	Independent	Size (0.994)	Size (0.999)	Size (0.995)	Size	NA
	Network Structure (MV1)	NA	Structure and Distribution (0.747)	Structure and Distribution (0.761)	Structure and Distribution (0.704)	Structure and Distribution (0.872)
	Network Flow (MV2)	Spread(0.92), Speed(0.73)	Spread and Speed (0.923)	Spread and Speed (0.991), Spread and Speed Boundry(0.813)	Spread and Speed (0.875)	Spread and Speed (0.941), Spread and Speed Boundry(0.863)
	Dependent	Influence(0.779)	Influence(0.975)	Influence (1.00)	Influence (0.978)	NA
Consumption	Independent	Size(0.994)	Size (0.999)	Size (0.995)	Size (0.997)	Size (0.977)
	Network Structure (MV1)	NA	Structure(0.666), Distribution(0.928)	Distribution(0.893)	Distribution (0.875)	Distribution(0.933)
	Network Flow (MV2)	Spread(0.92), Speed(0.73)	Spread and Speed (0.923)	Spread and Speed (0.991), Spread and Speed Boundry (0.916)	Spread and Speed (0.875)	Spread and Speed (0.905)
	Dependent	NA	Influence (0.98)	Influence (0.880)	Influence (0.729)	Influence (1.00)
Propagation	Independent	Size(0.994)	Size (0.999)	Size (0.995)	Size (0.997)	Size (0.977)
	Network Structure (MV1)	NA	Distribution(0.648)	Distribution(0.714)	NA	Structure and Distribution (0.635), Distribution (0.816)
	Network Flow (MV2)	Spread(0.92), Speed(0.73)	Spread and Speed(0.923)	Spread and Speed (0.991), Spread and Speed Boundry(0.813)	Spread and Speed (0.875)	Spread and Speed (0.905)
	Dependent	Influence(0.817)	Influence(0.999)	Influence (0.995)	NA	Influence (0.812)

5.4.1 Factors from Independent Variables

From Table13, it can be seen that the independent variables form a single factor labelled "Size" across all product categories and all network types (undirected, directed, consumption and propagation), except in the directed network of Science product category. The factor “Size” comprises of variables the total number of nodes and the total number of ties in a network. The total number of nodes (Nodes) and ties (Edges_ud, Edges_d) define the “Size” of a network. In all the cases where “Size” is a factor, the total number of nodes (Nodes) and the total number of ties (Edges_ud, Edges_d) in a network have strong factor loadings. The details of factor analysis for all independent variables are shown in Appendix A (A.1.7.1.2.1, A.1.7.2.2.1, A.1.7.3.2.1, A.1.7.4.2.1, A.2.7.1.2.1, A.2.7.2.2.1, A.2.7.3.2.1, A.2.7.4.2.1, A.3.7.1.2.1, A.3.7.2.2.1, A.3.7.3.2.1, A.3.7.4.2.1, A.4.7.1.2.1, A.4.7.2.2.1, A.4.7.3.2.1, A.4.7.4.2.1, A.6.7.1.2.1, A.6.7.2.2.1, A.6.7.3.2.1, A.6.7.4.2.1)

5.4.2 Factors from Network Flow (MV2) Variables

The network flow variables also form factors across all product categories and all network types (undirected, directed, consumption and propagation). The network flow variables form four distinct factors “Spread”, “Speed”, “Spread and Speed “and “Spread and Speed Boundary”.

The factor “Spread” consists of the following variables: the Graph Diameter (GD_x) of the network, the Total Paths (Tpaths_x) in the network and the Average Path Length (AvgPL_x). As explained in section 4.3.3.2.2, the Total Paths (Tpaths_x) in the network and the Average Path Length (AvgPL_x) is representative of the process of spreading the information in the network. In all the cases where “Spread” is a factor, the

Graph Diameter of the network (GD_x), the Total Paths in the network (Tpaths_x) and the Average Path Length (AvgPL_x) have strong factor loadings.

The factor "Speed" consists of the following variables: the Total Shortest Paths (TSpaths_x) in the network and the Average Geodesic Length (AvgGL_x). As explained in section 4.3.3.2.2, the Total Shortest Paths (TSpaths_x) in the network and the Average Geodesic Length (AvgGL_x) is representative of the process of Speed of information spread in the network. In all the cases where "Speed" is a factor, the Total Shortest Paths (TSpaths_x) in the network and the Average Geodesic Length (AvgGL_x) have strong factor loadings.

The factor "Spread and Speed" consists of the following variables: the Graph Diameter of the network (GD_x), the Total Paths in the network (Tpaths_x), the Average Path Length (AvgPL_x), the Total Shortest Paths (TSpaths_x) in the network and the Average Geodesic Length (AvgGL_x). As explained in section 4.3.3.2.2, the Total Shortest Paths in the network (TSpaths_x) and the Average Geodesic Length (AvgGL_x) are representative of the process of speed of information spread in the network and the Total Paths in the network (Tpaths_x) and the Average Path Length (AvgPL_x) are representative of the process of spreading information in the network. In most of the cases where "Spread and Speed" is a factor, variables the Graph Diameter of the network (GD_x), the Total Paths in the network (Tpaths_x), the Average Path Length (AvgPL_x) the Total Shortest Paths (TSpaths_x) in the network and the Average Geodesic

Length (AvgGL_x) have strong factor loadings. In some cases Total Shortest Paths loads independently.

The factor "Spread and Speed Boundary" consists of the following variables: the Graph Diameter of the network (GD_x), the Average Path Length (AvgPL_x) and the Average Geodesic Length (AvgGL_x). As explained in section 4.3.3.2.2, the Average Geodesic Length (AvgGL_x) is representative of the boundary of the process of Speed of information spread in the network and the Average Path Length (AvgPL_x) is representative of boundary of the process of information spread in the network. In all the cases where "Spread and Speed Boundary" is a factor, Graph Diameter of the network (GD_x), the Average Path Length (AvgPL_x) and the Average Geodesic Length (AvgGL_x) have strong factor loadings. In some cases one of the variables loads independently.

The details of factor analysis for all network flow variables are shown in Appendix A (A.1.7.1.2.3, A.1.7.2.2.3, A.1.7.3.2.3, A.1.7.4.2.3, A.2.7.1.2.3, A.2.7.2.2.3, A.2.7.3.2.3, A.2.7.4.2.3, A.3.7.1.2.3, A.3.7.2.2.3, A.3.7.3.2.3, A.3.7.4.2.3, A.4.7.1.2.3, A.4.7.2.2.3, A.4.7.3.2.3, A.4.7.4.2.3, A.6.7.1.2.3, A.6.7.2.2.3, A.6.7.3.2.3, A.6.7.4.2.3)

5.4.3 Factors from Network Structure (MV1) Variables

The network structure variables form factors only in the "Entertainment", "Comedy", "Sports" and "Science" product categories. Within these product categories the network structure variables form factors only in the directed, the consumption and the propagation networks. The network structure variables form three distinct factors "Structure", "Distribution "and "Structure and Distribution".

The factor "Structure" consists of the following variables: the Scale Free metric (S_x) and the Assortativity (R_x). As explained in section 4.3.3.1.2, the Scale Free metric (S_x) and the Assortativity (R_x) explain the presence of hubs and the patterns of connectivity in the network. In all the cases where "Structure" is a factor, the Scale Free metric (S_x) and the Assortativity (R_x) have strong factor loadings.

The factor "Distribution" consists of the following variables: the Power Law Distribution of Total Paths per Node (PL_TpxN) and the Power Law Distribution of Total Shortest Paths per Node (PL_TSpxN). As explained in section 4.3.3.1.4, variables the Power Law Distribution of Total Paths per Node (PL_TpxN) and the Power Law Distribution of Total Shortest Paths per Node (PL_TSpxN) explain the distribution of paths and shortest paths with respect to the nodes in the network. In all the cases where "Distribution" is a factor, the Power Law Distribution of Total Paths per Node (PL_TpxN) and the Power Law Distribution of Total Shortest Paths per Node (PL_TSpxN) have strong factor loadings.

The factor "Structure and Distribution" consists of the following variables: the Scale Free metric (S_x), the Assortativity (R_x), the Power Law Distribution of Total Paths per Node (PL_TpxN) and the Power Law Distribution of Total Shortest Paths per Node (PL_TSpxN). As explained in section 4.3.3.1.2, the Scale Free metric (S_x) and the

Assortativity (R_x) explain the presence of hubs and the patterns of connectivity in the network. As explained in section 4.3.3.1.4, variables the Power Law Distribution of Total Paths per Node (PL_TpxN) and the Power Law Distribution of Total Shortest Paths per Node (PL_TSpxN) explain the distribution of paths and shortest paths with respect to the nodes in the network. In all the cases where “Structure and Distribution” is a factor, the Power Law Distribution of Total Paths per Node (PL_TpxN), the Power Law Distribution of Total Shortest Paths per Node (PL_TSpxN), the Scale Free metric (S_x) and the Assortativity (R_x) have strong factor loadings.

The details of factor analysis for all network structure variables are shown in Appendix A (A.1.7.1.2.2, A.1.7.2.2.2, A.1.7.3.2.2, A.1.7.4.2.2, A.2.7.1.2.2, A.2.7.2.2.2, A.2.7.3.2.2, A.2.7.4.2.2, A.3.7.1.2.2, A.3.7.2.2.2, A.3.7.3.2.2, A.3.7.4.2.2, A.4.7.1.2.2, A.4.7.2.2.2, A.4.7.3.2.2, A.4.7.4.2.2, A.6.7.1.2.2, A.6.7.2.2.2, A.6.7.3.2.2, A.6.7.4.2.2)

5.4.4 Factors from Dependent Variables

The dependent variables form a single factor labelled “Influence” across all product categories. Within these product categories the network structure variables form factors only in the directed, the consumption and the propagation networks. The factor “Influence” consists of the following variables: the Correlation Coefficient of Eigenvector Centrality with respect to Total Paths (EVCx_TpxN) and the Correlation Coefficient of Eigenvector Centrality with respect to Total Shortest Paths (EVCx_TSpxN). As described in section 4.3.4.1, the Correlation Coefficient of Eigenvector Centrality with respect to Total Paths (EVCx_TpxN) is used as a measure of influence with respect to information spread and the Correlation Coefficient of Eigenvector Centrality with respect to Total Shortest Paths (EVCx_TSpxN) is used as a measure of influence with respect to speed of information spread. In all the cases where "Influence" is a factor, the Correlation Coefficient of Eigenvector Centrality with respect to Total Paths and (EVCx_TpxN) the Correlation Coefficient of Eigenvector Centrality with respect to Total Shortest Paths (EVCx_TSpxN) have strong factor loadings.

The details of factor analysis for all network flow variables are shown in Appendix A (A.1.7.1.2.4, A.1.7.2.2.4, A.1.7.3.2.4, A.1.7.4.2.4, A.2.7.1.2.4, A.2.7.2.2.4, A.2.7.3.2.4, A.2.7.4.2.4, A.3.7.1.2.4, A.3.7.2.2.4, A.3.7.3.2.4, A.3.7.4.2.4, A.4.7.1.2.4, A.4.7.2.2.4, A.4.7.3.2.4, A.4.7.4.2.4, A.6.7.1.2.4, A.6.7.2.2.4, A.6.7.3.2.4, A.6.7.4.2.4).

The correlation matrices presented in Appendix A confirm the criteria-related validity. For example, in the undirected network of Entertainment product category, the variables of factor "Speed" (Average Geodesic Length (AvgGL_ud) and Total Shortest Paths (TSpaths_ud)) correlate significantly with the variables of factor "Size" (the total number of nodes (Nodes) and the total number of ties (Edges_ud)), as shown in Table 14 below.

Table 14: Correlation between Variables of Factors "Speed and "Size", in the Undirected Network of "Entertainment" Category

Correlations

		Nodes	Edges_ud	TSpaths_ud	AvgGL_ud
Nodes	Pearson Correlation	1	.999**	.840**	.265*
	Sig. (2-tailed)		.000	.000	.011
	N	91	91	91	91
Edges_ud	Pearson Correlation	.999**	1	.853**	.288**
	Sig. (2-tailed)	.000		.000	.006
	N	91	91	91	91
TSpaths_ud	Pearson Correlation	.840**	.853**	1	.535**
	Sig. (2-tailed)	.000	.000		.000
	N	91	91	91	91
AvgGL_ud	Pearson Correlation	.265*	.288**	.535**	1
	Sig. (2-tailed)	.011	.006	.000	
	N	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

This shows that the factors, Size and Speed, are capable of adequately capturing the relevant aspects of each other in the undirected network of Entertainment product category.

5.5 Regression Analysis

This study uses multiple linear regression analysis to determine the relative impact of the predictor variables on the dependent variables. Multiple regression is an extension of simple regression in which an outcome is predicted by a linear combination of two or more predictor variables (Field, 2005, p. 738). The results of this regression analysis are provided in this section.

I use two approaches to regression in this study. In the first approach, I use all the predictors (independent variables and moderating variables (network flow variables and network structure variables)) to show their impact on each of the dependent variables (network phenomenon variables). In this approach, the regression model does not include interactions between the independent variables and the moderating variables. In the second approach, in order to address the research questions (RQ1, RQ2, RQ3 and RQ4); I use the following four regression models:

1. Network structure variables as predictors to show their impact on each of the network flow variables.
2. Network flow variables as predictors to show their impact on each of the network structure variables.
3. Network structure variables as predictors to show their impact on each of the dependent variables (as defined in section 5.2.3).
4. Network flow variables as predictors to show their impact on each of the dependent variables (as defined in section 5.2.3).

The stepwise regression function from IBM's SPSS version 22 (64 bit) is used in both approaches. The stepwise regression function uses both forward and backward regression models to find the best predictors.[22]

---

[22] The stepwise method calculates the contribution of each predictor on the outcome by comparing the significance value or the t-test of each predictor against a removal criterion. If a predictor meets the removal criterion or does not improve the prediction power of the model, then it is removed from the analysis. Then the model re-assesses the remaining predictors. Source: SPSS Online help (http://10.10.10.245:6908/help/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fspss%2Fbase%2Fovervw_auto_0.htm)

In order to ensure that the regression models are not suffering from multicollinearity, "Tolerance" and "VIF" (variable inflation factor) from the "Collinearity Statistics" section of SPSS results are considered. The Tolerance value is an indicator of the variance of the predictor variable (independent variable) shared with some other predictor variable (independent variable) in the regression model (Neter et al., 1996, Allison, 1999). "VIF" is the reciprocal of "Tolerance" (Neter et al., 1996, Allison, 1999).

Various recommendations for acceptable levels of Tolerance and VIF have been published in the literature. Most commonly, a value of 0.1 has been recommended as the minimum level of Tolerance (O'brien, 2007, Fidell and Tabachnick, 2003). However, a recommended minimum value as high as 0.2 has also been suggested (Menard, 2002). Similarly, a value of 10 has been commonly recommended as the maximum level of VIF (Kennedy, 2003, Marquaridt, 1970, Neter et al., 1996). A recommended maximum VIF value of 5 (Rogerson, 2010) and even 4 (Pan and Jackson, 2008) can be found in the literature. The lowest suggested value of VIF found in literature was 2.5 (Brown et al., 2007, Coumarbatch et al., 2010). As this is exploratory research, I side on the edge of caution and use conservative values -- 0.2 as the minimum level of Tolerance and 2.5 as the maximum value VIF. Therefore, if a regression model has a "Tolerance" value of less than 0.2, then the regression model is suffering from multicollinearity. Similarly, if a regression model has a "VIF" value greater than 2.5, then the regression model is suffering from multicollinearity. In both these cases, the regression model is rejected.

Cooks test is undertaken to identify the influential outliers in the data that may be skewing the regression (Cook, 1977, Cook, 1979). Any data point that has a Cooks distance greater than 1 in the regression model is considered influential (Cook, 1977, Cook, 1979). In this situation, the influential data point is deleted and the regression is undertaken again without the influential data point.

Regressions were performed on the product categories of “Music”, “Entertainment”, “Comedy”, “Sports” and “Science”. In every product category the undirected networks, the directed networks, the consumption networks and the propagation networks were considered separately, in order to address the research questions put forth in section 2.8.

Overall, there are 16 hypotheses in this desertion, which are described in section 3.3. Each product category is tested for these 16 hypotheses. These 16 hypotheses involve 72 regressions for each product category.

Table 15: Total Regressions per Product Category

	Undirected Network	Directed Network	Directed Network Consumption Phase	Directed Network Propagation Phase	
Network Structure to Network Flows (HP1)	5 Regressions	5 Regressions	5 Regressions	5 Regressions	20
Network Flows to Network Structure (HP2)	5 Regressions	5 Regressions	5 Regressions	5 Regressions	20
Network Structure to Network Phenomenon (HP3)	4 Regressions	4 Regressions	4 Regressions	4 Regressions	16
Network Flows to Network Phenomenon (HP4)	4 Regressions	4 Regressions	4 Regressions	4 Regressions	16
					72

In order to reduce the family-wise error rate that results from multiple comparisons of data, a Bonferroni adjustment is undertaken. Due to this adjustment, the statistical significance level (p-value) for each test will be lowered to 0.000694 (Dunn, 1959, Dunn, 1961).

The details of the regression analysis are provided in Appendix A. In section 5.5.1, the results of the first approach to regression are presented, in which the independent variables and moderating variables (network flow variables and network structure variables) are used as predictors. In section 5.5.2, I present results that address the research question RQ1: Does network structure impact network flows in a social network that primarily exists online (hypothesis HP1a, HP1b, HP1c and HP1d)? In section 5.5.3, I present results that address the research question RQ2: Does network flow impact network structure in a social network that primarily exists online (hypothesis HP2a, HP2b, HP2c and HP2d)? In section 5.5.4, I present results that address the research question RQ3: Does network structure impact influence within an online social network (hypothesis HP3a, HP3b, HP3c and HP3d)? In section 5.5.5, I present results that address the research question RQ4: Does network flow impact influence within an online social network (hypothesis HP4a, HP4b, HP4c and HP4d)? In every case, the impact of the predictor variables on the dependent variables is considered identified, if at least one predictor variable impacts at least one dependent variable in a statistically significant fashion. The code "NA" in subsequent tables means that no significant impact was found between the predictor variables and the dependent variable.

5.5.1 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

In this section, I present results of the first approach to regression mentioned above. The data for product categories of “Music”, “Entertainment”, “Comedy”, “Sports” and “Science” is considered in undirected, directed, consumption and propagation phases. The significance (p-value) of 0.000694 is considered after the Bonferroni adjustment. The details of the regression analysis are provided in Appendix A (A.1.7.1.3.5, A.1.7.2.3.5, A.1.7.3.3.5, A.1.7.4.3.5, A.2.7.1.3.5, A.2.7.2.3.5, A.2.7.3.3.5, A.2.7.4.3.5, A.3.7.1.3.5, A.3.7.2.3.5, A.3.7.3.3.5, A.3.7.4.3.5, A.4.7.1.3.5, A.4.7.2.3.5, A.4.7.3.3.5, A.4.7.4.3.5, A.6.7.1.3.5, A.6.7.2.3.5, A.6.7.3.3.5, A.6.7.4.3.5).

Table 16: Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon in Undirected Networks

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud, (6),GD_ud (7) Tpaths_ud (8), TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud, (11) Nodes, (12) Edges_ud, (13) Den_ud, (14) CC_ud

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Music	(0.157/0.000)[12,4,2]	(0.045/0.024)[10]	(0.046/0.024)[7]	NA
Entertainment	(0.120/0.001)[14,5]	(0.041/0.000)[3,14]	(0.597/0.000)[1,4,6]	(0.076/0.005)[10]
Comedy	(0.157/0.000)[12,4,2]	(0.045/0.024)[10]	(0.046/0.024)[7]	NA
Sports	(0.181/0.000)[14,3]	(0.032/0.049)[4]	(0.476/0.000)[1,8,6]	(0.631/0.000)[10,4,1,7]
Science	(0.192/0.000)[12,4]	(0.709/0.000)[7,10,1]	(0.595/0.000)[1,6]	(0.458/0.000)[10,4,12]

Table 16 shows that in every case the independent and the moderating variables collectively have a significant impact on the network phenomenon variables. The number in the square brackets identifies the predictor that impacts the dependent variable. For example, in the “Music” case, Edges_ud, R_ud and PL_TSpudN impact ECud. In the “Music” and “Comedy” cases, the independent and the moderating variables have no impact on EVCud_TSpudN. The cells marked in orange indicate that, although the independent and moderating variables have some impact on the network phenomenon variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694).

Table 17: Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon in Directed Networks

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5) SMSP_d, (6)GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]			
	ECd	PL_EVCdN	EVCud_TpdN	EVCud_TSpdN
Music	(0.090/0.002) [9]	(0.303/0.000) [1,7,14,1	(0.060/0.011) [15]	(0.061/0.010) [2]
Entertainment	(0.362/0.000) [7,15]	(0.456/0.000) [6,15]	(0.239/0.000) [2,6]	(0.235/0.000) [2,6]
Comedy	(0.246/0.000) [3,4,15	(0.546/0.000) [1,7,15]	(0.132/0.001) [12,15	(0.140/0.000) [12,15]
Sports	(0.229/0.000) [10]	(0.077/0.005) [2]	(0.108/0.001) [2]	(0.561/0.000) [6,15]
Science	(0.411/0.000) [3,12,1	(0.609/0.000) [3,11,15	(0.231/0.000) [6,10]	(0.233/0.000) [6,10]

Table 17 shows that in every case the independent and the moderating variables collectively have a significant impact on the network phenomenon variables. In the case of “Science”, it can be seen that the independent and the moderating variables collectively impact all the network phenomenon variables. The cells marked in orange indicate that, although the independent and moderating variables have some impact on the network phenomenon variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694).

Table 18: Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon in Consumption Networks

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5) SMSP_d, (6) GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Music	(0.199/0.000)[4,15,1]	(0.064/0.009)[14]	(0.274/0.000)[6,7]	(0.144/0.001)[4,6,7]
Entertainment	(0.325/0.000)[7,15]	(0.381/0.000)[1,14,15]	(0.092/0.005)[3,11]	(0.149/0.000)[3,8]
Comedy	(0.409/0.000)[4,12,15]	(0.325/0.000)[4,15]	(0.188/0.000)[12,15]	(0.200/0.000)[12,15]
Sports	(0.441/0.000)[15,8,2,14]	(0.306/0.000)[2,15]	(0.144/0.000)[13]	(0.169/0.000)[13]
Science	(0.308/0.000)[8,15]	NA	(0.262/0.000)[4,10]	(0.262/0.000)[4,10]

Table 18 shows that in every case the independent and the moderating variables collectively have a significant impact on the network phenomenon variables. In the case of “Science”, the independent and the moderating variables have no impact on PL_EVCinN. The cells marked in orange indicate that, although the independent and moderating variables have some impact on the network phenomenon variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694).

Table 19 : Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon in Propagation Networks

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d, (6),GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Music	(0.328/0.000)[15]	(0.487/0.000)[8,14,15]	(0.316/0.000)[4,9,11]	(0.358/0.000)[4,9,11]
Entertainment	(0.325/0.000)[7,15]	(0.413/0.000)[2,15]	(0.466/0.000)[4,13]	(0.495/0.000)[1,4,13]
Comedy	(0.328/0.000)[15]	(0.487/0.000)[8,14,15]	(0.316/0.000)[4,9,11]	(0.358/0.000)[4,9,11]
Sports	(0.462/0.000)[1,8,15]	(0.411/0.000)[1,6,15]	(0.256/0.000)[8]	(0.298/0.000)[8]
Science	(0.308/0.000)[7,15]	(0.577/0.000)[8,15]	(0.065/0.009)[15]	NA

Table 19 shows that in every case the independent and the moderating variables collectively have a significant impact on the network phenomenon variables. In case of “Science” the independent and the moderating variables have no impact on EVCout_TSpoutN. The cells marked in orange indicate that, although the independent and moderating variables have some impact on the network phenomenon variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694).

5.5.2 Impact of Network Structure on Network Flows

In this section, I present results that address research question RQ1: Does network structure impact network flows in a social network that primarily exists online? In order to do so, I address hypothesis HP1a, HP1b, HP1c and HP1d for the product categories of “Music”, “Entertainment”, “Comedy”, “Sports” and “Science”. The details of the regression analysis are provided in Appendix A (A.1.7.1.3.1, A.1.7.2.3.1, A.1.7.3.3.1, A.1.7.4.3.1, A.2.7.1.3.1, A.2.7.2.3.1, A.2.7.3.3.1, A.2.7.4.3.1, A.3.7.1.3.1, A.3.7.2.3.1, A.3.7.3.3.1, A.3.7.4.3.1, A.4.7.1.3.1, A.4.7.2.3.1, A.4.7.3.3.1, A.4.7.4.3.1, A.6.7.1.3.1, A.6.7.2.3.1, A.6.7.3.3.1, A.6.7.4.3.1)

Hypothesis 1a (HP1a): The structural characteristics of a non-directional social network impact its network flows.

Table 20: Regression Analysis (Network Structure – Network Flows) Undirected Networks

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5) SMSP_ud

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud
Music	NA	(0.093/0.005)[3,4]	(0.254/0.000)[3,5,2]	NA	(0.287/0.000)[1,2,4,5]
Entertainment	(0.477/0.000)[3,4]	(0.523/0.000)[1,3,4]	(0.501/0.000)[2,3,4]	(0.477/0.000)[3,4]	(0.406/0.000)[3,4]
Comedy	(0.410/0.000)[4]	(0.396/0.000)[3,4]	(0.634/0.000)[3,4,5]	(0.407/0.000)[4]	(0.521/0.000)[3,4]
Sports	(0.522/0.000)[3,4]	(0.534/0.000)[3,4,5]	(0.462/0.000)[1,2,3,5]	(0.524/0.000)[3,4]	(0.337/0.000)[1,4]
Science	(0.307/0.000)[3,4]	(0.379/0.000)[3,4]	(0.537/0.000)[3,4]	(0.309/0.000)[3,4]	(0.412/0.000)[2,4]

Table 20 shows that the network structure has a significant impact on the network flow variables in the undirected networks for all the product categories. In case of “Music”, the network structure variables do not impact the GD_ud and AvgPL_ud. The cells marked in orange indicate that, although the predictors have some impact on the network flow variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 1a is confirmed for all cases in the undirected networks.

Hypothesis 1b (HP1b): The structural characteristics of a directional social network impact its network flows.

Table 21: Regression Analysis (Network Structure – Network Flows) Directed Networks

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5) SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Music	(0.127/0.000)[4]	(0.052/0.017)[1]	(0.227/0.000)[1,5]	(0.156/0.000)[1,4]	(0.079/0.010)[4,5]
Entertainment	(0.252/0.000)[1,3]	(0.306/0.000)[3]	(0.351/0.000)[3]	(0.234/0.000)[3,5]	(0.333/0.000)[1,3,5]
Comedy	(0.102/0.001)[4]	(0.398/0.000)[3,4]	(0.426/0.000)[3,4]	NA	(0.037/0.039)[4]
Sports	(0.312/0.000)[1,5]	(0.336/0.000)[3,5]	(0.303/0.001)[3,5]	(0.426/0.000)[1,5]	(0.256/0.000)[2,3,5]
Science	(0.139/0.000)[1]	(0.190/0.000)[3]	(0.193/0.000)[3]	NA	NA

Table 21 shows that the network structure has a significant impact on the network flow variables in the directed networks for all product categories. In case of “Science”, the network structure variables do not impact the AvgPL_d and AvgGL_d. In case of “Comedy”, the network structure variables do not impact the AvgPL_d. The cells marked in orange indicate that, although the predictors have some impact on the network flow variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694.Therefore, hypothesis 1b is confirmed for all cases in the directed network.

Hypothesis 1c (HP1c): The structural characteristics of a directional social network impact its network flows in the consumption phase.

Table 22: Regression Analysis (Network Structure – Network Flows) Consumption Networks

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5) SMSP_d

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Music	(0.204/0.000) [1,4]	(0.411/0.000) [1,2,4]	(0.595/0.000) [1,2,3,5]	(0.407/0.000) [1,4]	(0.076/0.005) [4]
Entertainment	(0/137/0.000) [2]	(0.115/0.002) [4,5]	(0.116/0.002) [4,5]	(0.223/0.000) [4,5]	(0.373/0.000) [4,5]
Comedy	(0.210/0.000) [3,4]	(0.095/0.000) [1,3]	NA	0,230/0.000) [3]	(0.281/0.000) [1,3,4]
Sports	(0.287/0.000) [1,5]	(0.267/0.000) [4,5]	(0.162/0.000) [4,5]	(0.450/0.000) [1,5]	(0.327/0.000) [4,5]
Science	(0.285/0.000) [4]	(0.558/0.000) [2,3,4]	(0.543/0.000) [2,3,4]	(0.484/0.000) [2,3,4]	(0.456/0.000) [2,3,4]

Table 22 shows that the network structure has a significant impact on the network flow variables in the consumption networks for all the product categories. In case of "Comedy", the network structure variables do not impact TSpaths_d. The cells marked in orange indicate that, although the predictors have some impact on the network flow variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 1c is confirmed for all cases in consumption network.

Hypothesis 1d (HP1d): The structural characteristics of a directional social network impact its network flows in the propagation phase.

Table 23: Regression Analysis (Network Structure – Network Flows) Propagation Networks

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5) SMSP_d

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Music	(0.234/0.000) [4]	(0.034/0.045) [4]	(0.035/0.043) [3]	(0.240/0.000) [4]	(0.416/0.000) [3,4]
Entertainment	(0.227/0.000) [1,4]	(0.107/0.003) [2,5]	(0.143/0.001) [2,3,5]	(0.120/0.001) [2,5]	(0.197/0.000) [2,5]
Comedy	(0.240/0.000) [4]	(0.034/0.045) [4]	(0.035/0.043) [3]	(0.240/0.000) [4]	(0.416/0.000) [3,4]
Sports	(0.369/0.000) [4]	(0.205/0.000) [5]	(0.075/0.005) [5]	(0.449/0.000) [4,5]	(0.342/0.000) [4,5]
Science	(0.054/0.015) [4]	(0.113/0.001) [4]	(0.109/0.001) [4]	(0.377/0.000) [4]	(0.361/0.00) [4]

Table 23 shows that network structure has a significant impact on the network flow variables in the propagation networks for all the product categories. The cells marked in orange indicate that, although the predictors have some impact on the network flow variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 1d is confirmed for all cases in propagation network.

5.5.3 Impact of Network Flows on Network Structure

In this section, I present results that address the research question RQ2: Does network flow impact network structure in a social network that primarily exists online? In order to do so I address hypothesis HP2a, HP2b, HP2c and HP2d for the product categories of “Music”, “Entertainment”, “Comedy”, “Sports” and “Science”. The details of the regression analysis are provided in Appendix A (A.1.7.1.3.2, A.1.7.2.3.2, A.1.7.3.3.2, A.1.7.4.3.2, A.2.7.1.3.2, A.2.7.2.3.2, A.2.7.3.3.2, A.2.7.4.3.2, A.3.7.1.3.2, A.3.7.2.3.2, A.3.7.3.3.2, A.3.7.4.3.2, A.4.7.1.3.2, A.4.7.2.3.2, A.4.7.3.3.2, A.4.7.4.3.2, A.6.7.1.3.2, A.6.7.2.3.2, A.6.7.3.3.2, A.6.7.4.3.2)

Hypothesis 2a (HP2a): Network flows impact the structural characteristics of a non-directional social network.

Table 24: Regression Analysis (Network Flows – Network Structure) Undirected Networks

Predictors: (6) GD_ud, (7) Tpaths_ud, (8) TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpudN	PL_TSpudN	S_ud	R_ud	SMSP_ud
Music	(0.040/0.032)[7]	(0.0105/0.001)[10]	NA	(0.167/0.000)[10,8]	NA
Entertainment	(0.212/0.000)[6,7]	(0.207/0.000)[7]	(0.559/0.000)[8,10]	(0.496/0.000)[6,8]	NA
Comedy	(0.138/0.000)[6]	NA	(0.694/0.000)[8,10]	(0.612/0.000)[8,10]	(0.052/0.017)[7]
Sports	(0..083/0.000)[6,9]	(0.146/0.000)[10]	(0.572/0.000)[8,10]	(0.474/0.000)[9]	(0.120/0.000)[9]
Science	NA	(0.136/0.000)[8]	(0.693/0.000)[7,8,10]	(0.392/0.000)[9]	NA

Table 24 shows that the network flow has a significant impact on the network structure variables in the undirected networks for all the product categories. In case of “Music”, the network flow variables do not impact S_ud and SMSP_ud. In “Entertainment”, the network flow variables do not impact SMSP_ud. In “Comedy”, the network flow variables do not impact PL_TSpudN. In “Science”, the network flow variables have no impact on PL_TpudN. The cells marked in orange indicate that, although the predictors have some impact on the network structure variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 2a is confirmed for all cases in undirected network.

Hypothesis 2b (HP2b): Network flows impact the structural characteristics of a directional social network.

Table 25: Regression Analysis (Network Flows – Network Structure) Directed Networks

Predictors: (6) GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpdN	PL_TSpdN	S_d	R_d	SMSP_d
Music	(0.103/0.003)[8,10]	(0.163/0.000)[8,10]	(0.142/0.000)[8]	(0.226/0.000)[6,8]	(0.090/0.002)[8]
Entertainment	(0.253/0.000)[6,8]	(0.251/0.000)[6,7]	(0.351/0.000)[7]	(0.326/0.000)[6,8]	NA
Comedy	(0.208/0.000)[6,7]	(0.211/0.000)[6,8]	(0.529/0.000)[8,10]	(0.204/0.000)[6,8,10]	NA
Sports	(0.350/0.000)[6,7]	(0.281/0.000)[6,7]	(0.472/0.000)[8,9]	(0.386/0.000)[6,8]	NA
Science	(0.139/0.000)[6]	(0.111/0.001)[6]	(0.193/0.000)[8]	(0.149/0.000)[6,8]	NA

Table 25 shows that the network flow has a significant impact on the network structure variables in the directed networks for all the product categories. In cases of "Entertainment", "Comedy", "Sports" and "Science" network flow variables have no impact on SMSP_d. The cells marked in orange indicate that, although the predictors have some impact on the network structure variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 2b is confirmed for all cases in directed network.

Hypothesis 2c (HP2c): Network flows impact the structural characteristics of a directional social network in the consumption phase.

Table 26: Regression Analysis (Network Flows – Network Structure) Consumption Networks

Predictors: (6) GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpinN	PL_TSpinN	S_con	R_con	SMSP_d
Music	(0.270/0.000)[7]	(0.265/0.000)[8]	(0.360/0.000)[8,10]	(0.300/0.000)[7,9]	(0.087/0.003)[8]
Entertainment	(0.181/0.000)[6,7]	(0.205/0.000)[6,7]	(0.345/0.000)[6,10]	(0.287/0.000)[10]	NA
Comedy	NA	NA	(0.157/0.000)[6]	(0.100/0.001)[6]	NA
Sports	(0.218/0.000)[6,7]	(0.216/0.000)[6,7]	(0.049/0.020)[10]	(0.102/0.000)[8]	NA
Science	(0.037/0.038)[10]	(0.086/0.003)[10]	(0.220/0.000)[9]	(0.475/0.000)[6,7]	NA

Table 26 shows that the network flow has a significant impact on the network structural variables in consumption networks for all the product categories. In cases of "Entertainment", "Comedy", "Sports" and "Science" network flow variables have no impact on SMSP_d. Network flow variables in "Comedy" do not impact PL_TpinN and PL_TSpinN. The cells marked in orange indicate that, although the predictors have some impact on the network structure variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 2c is confirmed for all cases in consumption network.

Hypothesis 2d (HP2d): Network flows impact the structural characteristics of a directional social network in the propagation phase.

Table 27: Regression Analysis (Network Flows – Network Structure) Propagation Networks

Predictors :(6) GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]				
	PL_TpoutN	PL_TSpoutN	S_pro	R_pro	SMSP_d
Music	(0.050/0.019)[6]	(0.058/0.013)[6]	(0.392/0.000)[7,10]	(0.386/0.000)[6,7,10]	NA
Entertainment	(0.214/0.000)[6,7]	(0.171/0.000)[6]	(0.204/0.000)[6,8]	(0.315/0.000)[6,8]	NA
Comedy	(0.050/0.019)[6]	(0.058/0.013)[6]	(0.114/0.001)[6]	(0.386/0.000)[6,7,10]	NA
Sports	(0.057/0.013)[6]	(0.034/0.044)[6]	(0.115/0.001)[6]	(0.354/0.000)[6,8]	NA
Science	(0.046/0.023)[6]	(0.041/0.031)[6]	(0.297/0.000)[9]	(0.377/0.000)[9]	NA

Table 27 shows that the network flow has a significant impact on the network structure variables in the propagation networks for all the product categories. . In cases of “Music”, “Entertainment”, “Comedy”, “Sports” and “Science” network flow variables have no impact on SMSP_d. The cells marked in orange indicate that, although the predictors have some impact on the network structure variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 2d is confirmed for all cases in propagation network.

5.5.4 Impact of Network Structure on Network Phenomenon

In this section, I present results that address the research question RQ3: Does network structure impact influence within an online social network? In order to do so, I address hypothesis HP3a, HP3b, HP3c and HP3d for the product categories of “Music”, “Entertainment”, “Comedy”, “Sports” and “Science”. The details of the regression analysis are provided in Appendix A (A.1.7.1.3.3, A.1.7.2.3.3, A.1.7.3.3.3, A.1.7.4.3.3, A.2.7.1.3.3, A.2.7.2.3.3, A.2.7.3.3.3, A.2.7.4.3.3, A.3.7.1.3.3, A.3.7.2.3.3, A.3.7.3.3.3, A.3.7.4.3.3, A.4.7.1.3.3, A.4.7.2.3.3, A.4.7.3.3.3, A.4.7.4.3.3, A.6.7.1.3.3, A.6.7.2.3.3, A.6.7.3.3.3, A.6.7.4.3.3).

Hypothesis 3a (HP3a): Network structure impacts influence within an online social network in a non-directional social network.

Table 28: Regression Analysis (Network Structure – Influence) Undirected Networks

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5) SMSP_ud

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Music	(0.079/0.010)[2,5]	NA	NA	NA
Entertainmen	(0.091/0.005)[4,5]	(0.402/0.000)[4,5]	(0.0578/0.000)[1,4]	NA
Comedy	(0.105/0.003)[1,4]	(0.082/0.004)[4]	(0.640/0.000)[1]	NA
Sports	(0.133/0.000)[5]	(0.032/0.049)[4]	(0.435/0.000)[1,4]	(0.045/0.025)[2]
Science	(0.042/0.028)[3]	(0.160/0.000)[3,4]	(0.531/0.000)[1,4]	(0.060/0.000)[4]

Table 28 shows that the network structure has a significant impact on the network phenomenon variables in the undirected networks for all the product categories except “Music”. The network structure variables do not impact PL_EVCudN, EVCud_TpudN and EVCud_TSpudN in the “Music” category. Though the network structure variables have some impact on the ECud in the “Music” category, their impact is not considered as the significance (p-value) is higher than the significance (p-value) after the Bonferroni adjustment (0.000694). . In cases of “Entertainment” and

“Comedy” the network structure variables have no impact on EVCud_TSpudN. The cells marked in orange indicate that, although the predictors have some impact on the network phenomenon variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 3a is confirmed for all cases in undirected network except for the case of “Music”.

Hypothesis 3b (HP3b): Network structure impacts influence within an online social network in a directional social network.

Table 29: Regression Analysis (Network Structure – Influence) Directed Networks

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5) SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Music	(0.040/0.032)[4]	(0.140/0.001)[1,2,5]	NA	(0.061/0.010)[2]
Entertainment	(0.059/0.012)[5]	(0.089/0.002)[5]	0.157/0.000)[2,3]	(0.123/0.000)[2]
Comedy	(0.155/0.000)[3,4]	(0.086/0.003)[1]	NA	NA
Sports	NA	(0.211/0.000)[1,5]	(0.077/0.005)[2]	(0.108/0.001)[2]
Science	NA	(0.058/0.012)[1]	(0.056/0.014)[2]	(0.056/0.014)[2]

Table 29 shows that the network structure has a significant impact on the network phenomenon variables in the directed networks for all the product categories except “Music” and “Science”. In “Entertainment” category the network structure variables impact EVCd_TpdN and EVCd_TSpdN. In “Comedy” category the network structure variables impact only ECd. In the “Sports” category the network structure variables impact only PL_EVCdN. Therefore, hypothesis 3b is confirmed for all cases in the directed network except for the product categories of "Music" and "Science".

Hypothesis 3c (HP3c): Network structure impacts influence within an online social network in a directional social network during the consumption phase.

Table 30: Regression Analysis (Network Structure – Influence) Consumption Networks

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5) SMSP_d

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]			
	ECin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Music	(0.085/0.003) [4]	NA	(0.234/0.000) [1,2,5]	NA
Entertainment	NA	(0.046/0.024) [1]	(0.055/0.014) [3]	(0.094/0.002) [3]
Comedy	(0.124/0.000) [4]	(0.145/0.000) [3]	NA	NA
Sports	(0.117/0.002) [2,3]	(0.074/0.005) [2]	NA	NA
Science	NA	NA	(0.167/0.000) [4]	(0.168/0.000) [4]

Table 30 shows that the network structure has a significant impact on the network phenomenon variables in the consumption networks for all product categories except "Entertainment" and "Sports". In the "Music" category the network structure variables impact only EVCin_TpinN. In the "Comedy" category the network structure variables impact ECin and PL_EVCinN. In the "Science" category the network structure variables impact EVCin_TpinN and EVCin_TSpinN. Therefore, hypothesis 3c is confirmed only for "Music", "Comedy" and "Science" categories.

Hypothesis 3d (HP3d): Network structure impacts influence within an online social network in a directional social network during the propagation phase.

Table 31: Regression Analysis (Network Structure – Influence) Propagation Networks

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5) SMSP_d

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Music	NA	(0.034/0.44) [4]	(0.076/0.005) [4]	(0.080/0.004) [4]
Entertainment	NA	(0.128/0.001) [4,5]	(0.422/0.000) [2,4]	(0.432/0.000) [2,4]
Comedy	NA	(0.034/0.044) [4]	(0.076/0.005) [4]	(0.080/0.004) [4]
Sports	(0.138/0.000) [1]	(0.292/0.000) [1,4]	(0.037/0.038) [3]	(0.046/0.023) [3]
Science	(0.126/0.000) [4]	(0.043/0.027) [1]	NA	NA

Table 31 shows that the network structure has a significant impact on the network phenomenon variables in the propagation networks for all product categories except “Music” and “Comedy”. In “Entertainment” category the network structure variables impact EVCout_TpoutN and EVCout_TSpoutN. In “Sports” category the network structure variables impact ECout and Pl_EVCoutN. In “Science” category the network structure variables impact only ECout. Therefore, hypothesis 3d is confirmed only for “Entertainment”, “Sports” and “Science” categories.

5.5.5 Impact of Network Flows on Network Phenomenon

In this section, I present results that address the research question RQ4: Does network flow impact influence within an online social network? In order to do so, I address hypothesis HP4a, HP4b, HP4c and HP4d for the product categories of “Music”, "Entertainment", "Comedy", "Sports" and "Science". The details of the regression analysis are provided in Appendix A (A.1.7.1.3.4, A.1.7.2.3.4, A.1.7.3.3.4, A.1.7.4.3.4, A.2.7.1.3.4, A.2.7.2.3.4, A.2.7.3.3.4, A.2.7.4.3.4, A.3.7.1.3.4, A.3.7.2.3.4, A.3.7.3.3.4, A.3.7.4.3.4, A.4.7.1.3.4, A.4.7.2.3.4, A.4.7.3.3.4, A.4.7.4.3.4, A.6.7.1.3.4, A.6.7.2.3.4, A.6.7.3.3.4, A.6.7.4.3.4).

Hypothesis 4a (HP4a): Network flow impacts influence within an online social network in a non-directional social network.

Table 32: Regression Analysis (Network Flow – Influence) Undirected Networks

Predictors: (6) GD_ud, (7) Tpaths_ud, (8) TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Music	(0.062/0.010)[8]	(0.045/0.024)[10]	(0.045/0.025)[7]	NA
Entertainmen	(0.056/0.013)[8]	(0.184/0.000)[8]	(0.282/0.000)[6]	(0.076/0.005)[10]
Comedy	(0.076/0.005)[10]	(0.033/0.048)[8]	(0.097/0.000)[6]	NA
Sports	(0.054/0.015)[7]	NA	(0.167/0.000)[6]	(0.0539/0.000)[9,10]
Science	(0.033/0.048)[8]	(0.740/0.000)[6,8,10]	(0.106/0.001)[6]	(0.380/0.000)[8.10]

Table 32 shows that network flow has a significant impact on the network phenomenon variables in the undirected networks for all product categories except "Music". In "Entertainment" category the network flow variables impact PL_EVCudN and EVCud_TpudN. In "Comedy" category the network flow variables impact EVCud_TpudN. In "Sports" category the network flow variables impact EVCud_TpudN and EVCud_TSpudN. In "Science" category the network flow variables impact PL_EVCudN and EVCud_TSpudN. Therefore, hypothesis 4a is confirmed only for "Entertainment", "Comedy"," Sports" and "Science" categories.

Hypothesis 4b (HP4b): Network flow impacts influence within an online social network in a directional social network.

Table 33: Regression Analysis (Network Flow – Influence) Directed Networks

Predictors: (6) GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Music	(0.090/0.002) [9]	(0.051/0.018) [10]	NA	(0.098/0.004) [8,9]
Entertainment	(0.300/0.000) [7]	(0.149/0.000) [6]	(0.135/0.000) [6]	(0.135/0.000) [6]
Comedy	NA	(0.095/0.002) [9,10]	NA	NA
Sports	(0.229/0.000) [10]	(0.413/0.000) [6]	NA	NA
Science	(0.071/0.006) [9]	(0.146/0.000) [7]	(0.231/0.000) [6,10]	(0.233/0.000) [6,10]

Table 33 shows that the network flow has a significant impact on the network phenomenon variables in the directed networks for all product categories except "Music" and "Comedy". In "Sports" category the network flow variables impact ECd and PL_EVCdN. In "Science" category the network flow variables impact PL_EVCdN, EVCd_TpdN and EVCd_TSpdN. In "Entertainment" category the network flow variables impact ECd, PL_EVCdN, EVCd_TpdN and EVCd_TSpdN. Therefore, hypothesis 4b is confirmed only for "Entertainment"," Sports" and "Science" categories.

Hypothesis 4c (HP4c): Network flow impacts influence within an online social network in a directional social network during the consumption phase.

Table 34: Regression Analysis (Network Flow – Influence) Consumption Networks

Predictors: (6) GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Music	(0.034/0.044)[8]	NA	(0.274/0.000)[6,7]	(0.101/0.003)[6,8]
Entertainment	(0.268/0.000)[7]	NA	(0.070/0.007)[8]	(0.128/0.000)[8]
Comedy	(0.064/0.009)[9]	(0.047/0.022)[9]	(0.066/0.008)[8]	(0.071/0.006)[8]
Sports	(0.245/0.000)[10]	(0.139/0.000)[6]	(0.112/0.002)[8,9]	(0.104/0.001)[8]
Science	(0.097/0.002)[7]	(0.044/0.025)[9]	(0.205/0.000)[10]	(0.205/0.000)[10]

Table 34 shows that network flow has a significant impact on the network phenomenon variables in the consumption networks for all product categories except “Comedy”. In the “Music” and the “Entertainment” categories, the network flow variables do not impact PL_EVCinN. The cells marked in orange indicate that, although the predictors have some impact on the network phenomenon variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694. Therefore, hypothesis 4c is confirmed only for the “Music”, “Entertainment”,” Sports” and “Science” categories.

Hypothesis 4d (HP4d): Network flow impacts influence within an online social network in a directional social network during the propagation phase.

Table 35: Regression Analysis (Network Flow – Influence) Propagation Networks

Predictors: (6) GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d

	Dependent Variable (Adjusted R Square/ Significance) [Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Music	(0.064/0.009) [9]	(0.142/0.000) [6,7]	(0.058/0.013) [7]	(0.119/0.001) [7]
Entertainment	(0.268/0.000) [7]	(0.136/0.000) [6]	(0.226/0.000) [6,8]	(0.217/0.000) [6,8]
Comedy	(0.064/0.009) [10]	(0.142/0.000) [6,7]	(0.058/0.013) [7]	(0.119/0.001) [7]
Sports	(0.245/0.000) [10]	(0.221/0.000) [6]	(0.256/0.000) [8]	(0.298/0.000) [8]
Science	(0.097/0.002) [7]	(0.155/0.000) [7]	(0.066/0.008) [7]	(0.066/0.008) [7]

Table 35 shows that the network flow has a significant impact on the network phenomenon variables in the propagation networks for all product categories. For all the product categories the network flow variables impact PL_EVCoutN. The cells marked in orange indicate that, although the predictors have some impact on the network phenomenon variables, their impact is not considered, as their significance (p-value) is higher than the significance (p-value) of the Bonferroni-adjusted value (0.000694). In all other instances in the above table, the statistical significance (p-value) is less than or equal to 0.000694.Therefore, hypothesis 4d is confirmed for all cases in propagation network.

5.5.6 Summary of Hypothesis Testing

Table 36: Summary of Results of Hypothesis

Hypothesis	Music	Entertainment	Comedy	Sports	Science
Hypothesis 1a(HP1a)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 1b(HP1b)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 1c(HP1c)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 1d(HP1d)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 2a(HP2a)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 2b(HP2b)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 2c(HP2c)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 2d(HP2d)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 3a(HP3a)	Unconfirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 3b(HP3b)	Unconfirmed	Confirmed	Confirmed	Confirmed	Unconfirmed
Hypothesis 3c(HP3c)	Confirmed	Unconfirmed	Confirmed	Unconfirmed	Confirmed
Hypothesis 3d(HP3d)	Unconfirmed	Confirmed	Unconfirmed	Confirmed	Confirmed
Hypothesis 4a(HP4a)	Unconfirmed	Confirmed	Confirmed	Confirmed	Confirmed
Hypothesis 4b(HP4b)	Unconfirmed	Confirmed	Unconfirmed	Confirmed	Confirmed
Hypothesis 4c(HP4c)	Confirmed	Confirmed	Unconfirmed	Confirmed	Confirmed
Hypothesis 4d(HP4d)	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed

Table 36 summarizes the results of all hypotheses. Cells in green indicate that the hypotheses are confirmed, and the cells in red indicate that the hypotheses are unconfirmed. Overall, out of 80 hypotheses, 11 are unconfirmed and 69 are confirmed.

6. Conclusion and Discussion

In this chapter, I draw conclusions from the results presented in chapter 5. In some cases, the conclusions identify the need for further research. Section 6.1, discusses the implications identified from the metadata overview. In section 6.2, I present conclusions pertaining to network structure, network flows and the network phenomenon of interest—influence. In section 6.3, I discuss the implications of considering the consumption and propagation networks. In section 6.4, I discuss implications of scale. In section 6.5, I provide my conclusions regarding Eigenvector Centrality (EVC) as a measure of influence. In section 6.6, I present my conclusions regarding the experimental metrics (originally proposed in section 4.3.3.1.4 and 4.3.4.3), which pertain to Power Law Distribution of Paths per Node (PL_TpxN), Power Law Distribution of Shortest Paths per Node (PL_TSpxN) and Power Law Distribution of Eigenvector Centrality (PL_EVCxN). I summarize my conclusions in section 6.7.

6.1 Conclusions from Metadata Overview

As suggested in section 4.1.2.2, products categorized as high (in terms of popularity) were supposed to generate communities that were bigger in size, both in terms of number of tweets and people involved, than products that were categorized medium or small. This assumption does not hold true. Therefore, *a positive correlation between the popularity of a product category and the size of the conversation that the product category generates cannot be assumed.* “Entertainment”, which was categorized as “high” based on YouTube popularity, generated 43,377 total tweets whereas the “Comedy” category which was categorized as medium generated 94,111 total tweets over the same period of time. Similarly, the “Sports” category generated more tweets than the “Entertainment” category. This trend can also be seen for the community sizes of “Comedy” and “Sports”, both in terms of the number of tweets and the number of people involved. When compared to “Entertainment” community, “Comedy” and “Sports” had larger number of community participants.

Twitter communities were generated based on the presence of the word “YouTube” and the product category names in a tweet. A product category on YouTube, for example “Entertainment”, might encompass various types of videos that do not fall under the conversations on Twitter in which the word “Entertainment” is used. For example, videos of movie trailers might be grouped under “Entertainment” category on YouTube but people talking about the movie trailers on Twitter might not use the word “Entertainment” in their tweet. This may partly be due to the limitations put forth by the platform itself (140 character limit on Twitter). However, this might not be the case for Music category. People engaged in conversations on Twitter about “Music” may use the word “Music” in all of their conversations. As a result, the “Music” conversation might generate one large cohesive community while entertainment may spawn multiple communities on Twitter. Therefore, further research is required to understand how community definitions translate across various platforms. I identify this as an area for future research.

6.2 Network Structure, Network Flow and Network Phenomenon

The essential management question that motivated this research is: “How does the relationship between network structure, network flows and the loci of influence affect the course of action that marketers should take when they engage with an online social network?” In particular, how do network structure and network flows impact each other, and how do they impact the phenomenon of influence? In order to address the management question, a literature research was conducted in chapter 2 to identify the current state of knowledge. Gaps in the state of knowledge and the questions arising from the gaps were also presented in chapter2. In chapter3, the scope of the research was discussed and an experimental framework was provided (Figure 9). Figure 24 below is an extension of Figure 9. It illustrates the conclusions of this dissertation that pertain to network structure network, network flows and the network phenomenon of influence.

Figure 24: Validation of Research Framework

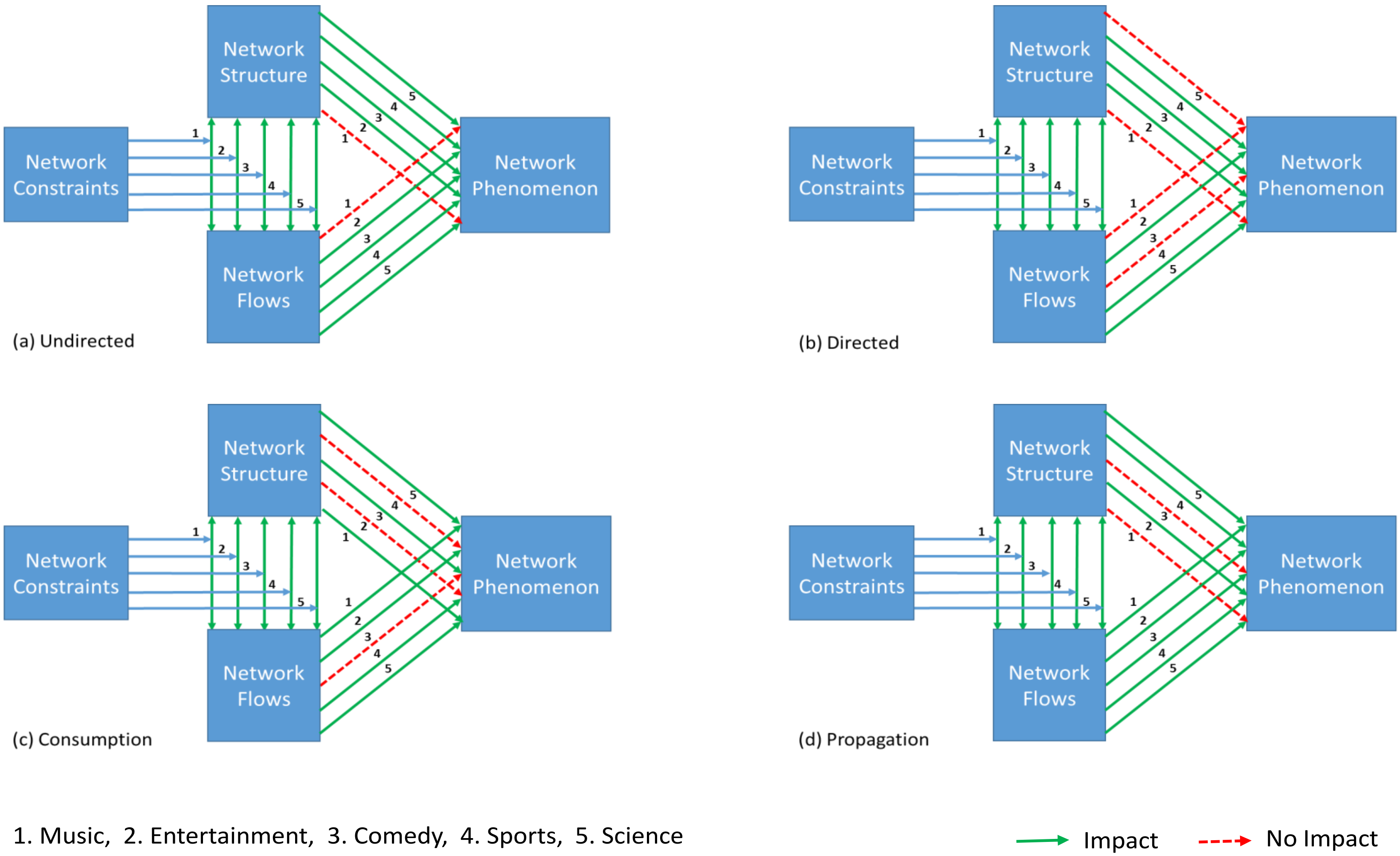


Figure 24 shows that the theoretical framework presented in Figure 9 has been validated for the five Twitter conversations that correspond to five YouTube product lines one to one. As described in section 5.5, at least one predictor variable needed to impact at least one dependent variable to confirm a hypothesis. Under these conditions, all hypotheses from section 3.3 have been confirmed in more than one case under study, but not in all cases under study. In all five cases under study, network structure has an impact on network flow, and conversely. It has also been shown that network structure and network flows impact the network phenomenon of influence, but not in all instances. In some instances, only network structure impacts influence, in others only network flows impact influence. In yet other instances, both network structure and network flows impact network influence.

The ramifications of these findings are perhaps best illustrated by an enhanced scrutiny of the "Music" case. Figure 24c shows that both network structure and network flows impact influence in the consumption phase. This suggests that someone who consumes music through YouTube is influenced by his/her network of people with whom they share a common interest on Twitter, i.e. people seem to care from whom the information comes. They also care about the content of the propagated information. Figure 24d shows that network flows but not network structure impacts influence in the propagation phase. This implies that people in the "Music" network care about the information that the community propagates, but they do not care about how the community is structured. (They may not even be aware of the community's structure.) It has been observed section 5.2.2.2 that, in the “Music” category, the consumption network is more Scale Free and more Disassortative than the propagation network over the whole time period under study. This suggests that music consumers get their information from a variety of sources and that they tend connect to people who are perceived to be more popular. These details about consumer behavior cannot be perceived in Figure 24a and 24b. This suggests that directionality needs to be studied to obtain an enhanced understanding of consumer behavior, and that studies of directed networks need to differentiate between propagation and the consumption phase.

The above observations are quite significant. They indicate that *the impact of the network structure on the network flow or the impact of network flow on network structure or the impact of network flow and network structure on the network phenomenon cannot be taken for granted*. As stated in section 2.3, the information flowing through a network provides a conceptual universe, within which we can impose conceptual constraints like connectedness and relate them to other properties like the probability of receiving information. Theoretical constructs that pertain to a particular conceptual universe are thus true only within the contextual model of that universe; they may be false in a different context (Borgatti and Kidwell, 2011). These constructs are derivations of the particular model under consideration, yet, as theories of network phenomena show, they are widely misperceived to be unconnected to the theory (Borgatti and Kidwell, 2011). In addition, theoretical constructs that pertain to a particular conceptual universe cannot be considered generic measures or generic techniques like regression, which can be divorced from an underlying model of how things work (Borgatti, 2005).

6.3 Consumption and Propagation Networks

The research in this dissertation shows that *within a directed network, the consumption and propagation networks can behave very differently from each other.* In order to elaborate, I show the Scale Free metric and the Assortativity for undirected, directed, consumption and propagation networks in the Music category.

Figure 25: Music Scale Free Metric--(a) Undirected network; (b) Directed network; (c) Consumption network; (d) Propagation network

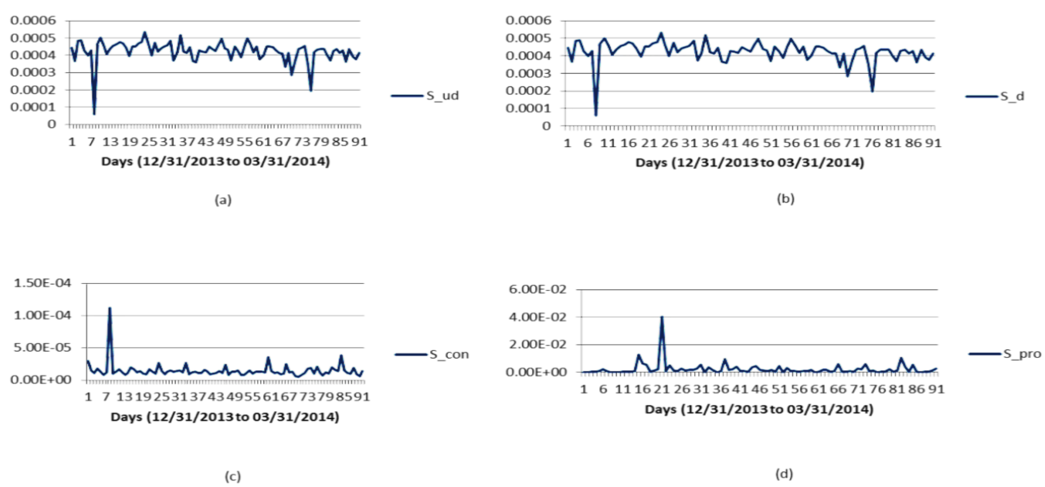


Figure 25 shows the Scale Free metric for the undirected, directed, consumption and propagation networks. The Scale Free metrics for the undirected network and the directed network are similar, but the Scale Free metric for the consumption and propagation networks are very different. The propagation network is more Scale Free than the consumption network by more than two orders of magnitude. The values of the Scale Free metrics range between 0 and 1. When the values are closer to 1, it means that the networks are more Scale Free. None of the networks are highly Scale Free in nature. This means that these networks have hubs in them. However, there is not just one hub that is the center of the community. The nodes have a uniform connectivity pattern.

Figure 26: Music Assortativity (a) Undirected Network (b) Directed Network (c) Consumption Network (d) Propagation Network

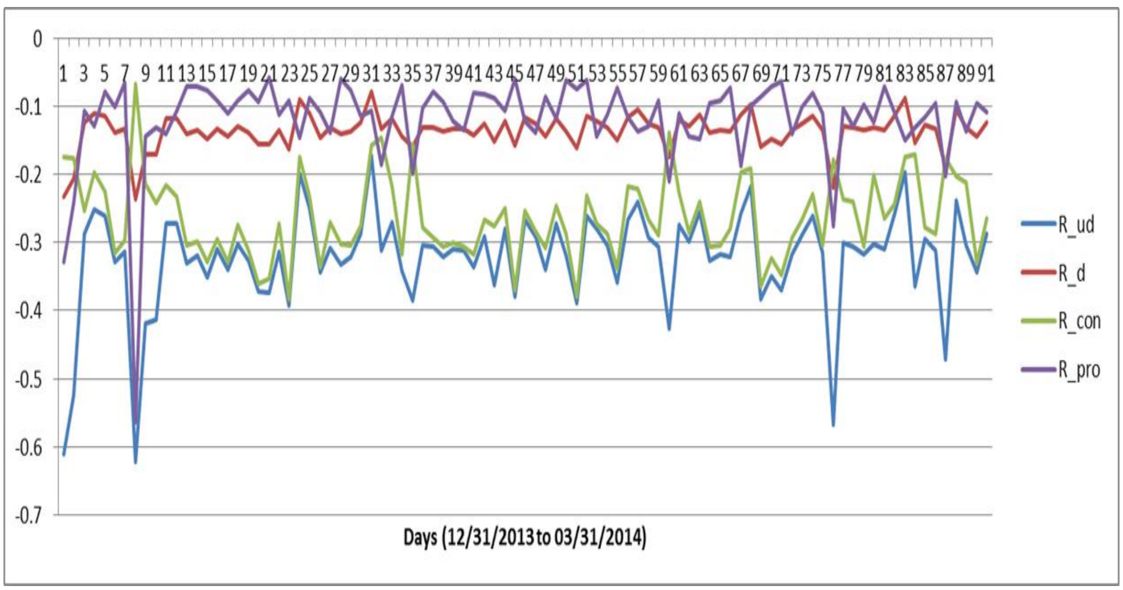


Figure 26 shows the Assortativity for the undirected, directed, consumption and propagation networks of “Music” conversations. The value of Assortativity ranges between -1 and +1. When the values are closer to -1, it means that the networks are

Disassortative. The undirected network is more Disassortative than the directed network. Among the directed networks, the consumption network is more Disassortative than the propagation network. Disassortative means that the nodes in the network connect to nodes that are very similar to themselves in connectivity pattern. This is true more so in the undirected network and in the consumption network than it is in the directed network and the propagation network. This implies that Disassortativity of consumption contributes more to the Disassortativity of the directed network than the Disassortativity of the propagation does.

The Scale Free metric and Assortativity indicated that the consumption and propagation processes happening within a network are very different, and that they cannot be deduced by just analyzing the undirected or directed network. A person who might be influential in the consumption process may not be influential in propagation process. Also, by considering the consumption and the propagation network, it is possible to deduce behavioral traits of a person in the network, which may vary greatly from person to person. For example, which people have a greater propensity to act as hubs (Scale Free); to whom do they listen; and to whom do they talk? Do some people only listen to people who have similar assortment of connections as they do, but only talk to people who have very different assortment of connections? Similar trends in Scalefreeness and Assortativity have been observed in the other YouTube categories under investigation (see in appendices A.2.4, A.3.4, A.4.4, A.5.4, and A.6.4).

6.4 Impact of Scale

Figure 27 and Figure 28 respectively show the number of nodes (Nodes) and the number of (Edges_ud) in the undirected network formed for the “Music”, “Entertainment”, “Comedy”, “Sports” and “Science” categories. Figure 29 shows Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud), Average Path Length (AvgPL_ud), Average Geodesic Length (AvgGL_ud) and Graph Diameters (GD_ud) in the undirected network formed for “Music”, “Entertainment”, “Comedy”, “Sports” and “Science” categories. From Figure 27 and Figure 28 we can see that the “Music” category networks are orders of magnitude larger than the networks formed under any other category. Figure 29 shows that the undirected networks of the “Music” category have Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud), Average Path Length (AvgPL_ud), Average Geodesic Length (AvgGL_ud) and Graph Diameters (GD_ud) that are orders of magnitude higher than the undirected networks of any other product category under observation. This provides an opportunity to study what impact scale has on the processes within the networks.

Figure 27: Nodes in Undirected Networks

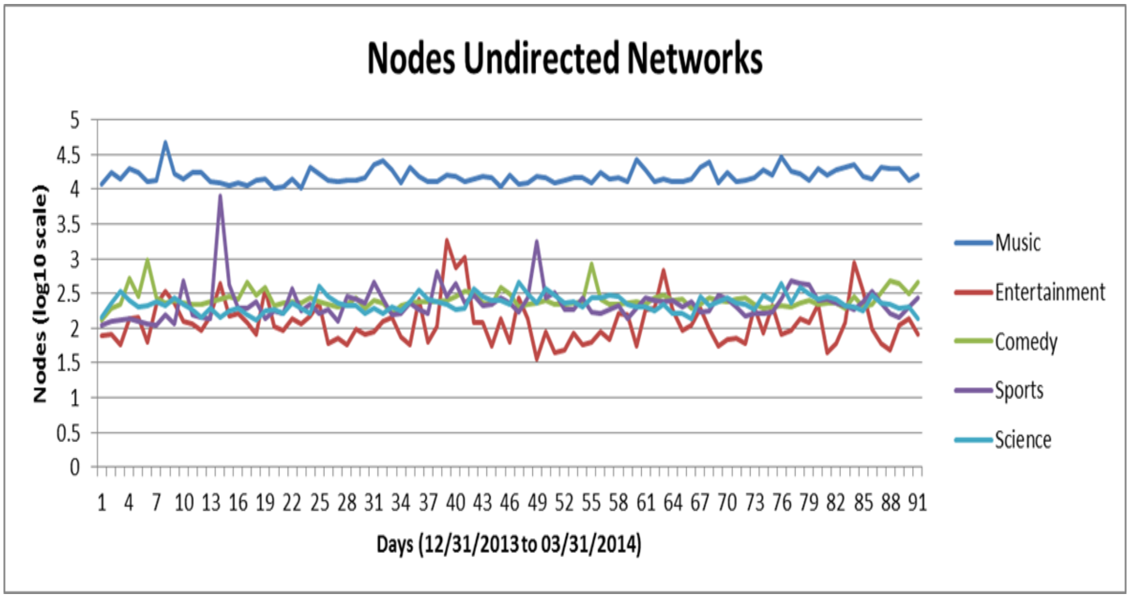


Figure 28: Ties in Undirected Networks

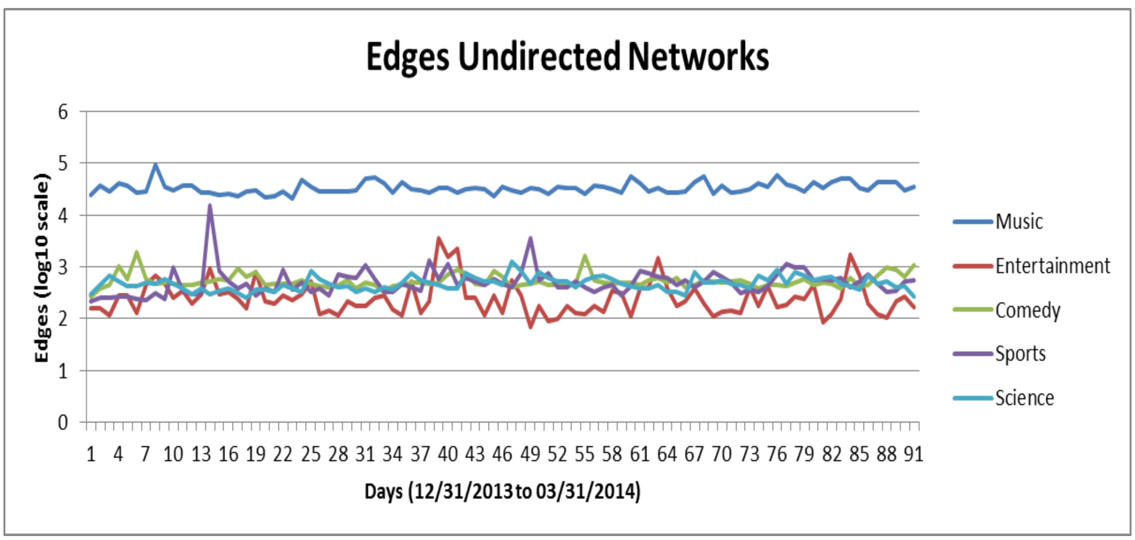
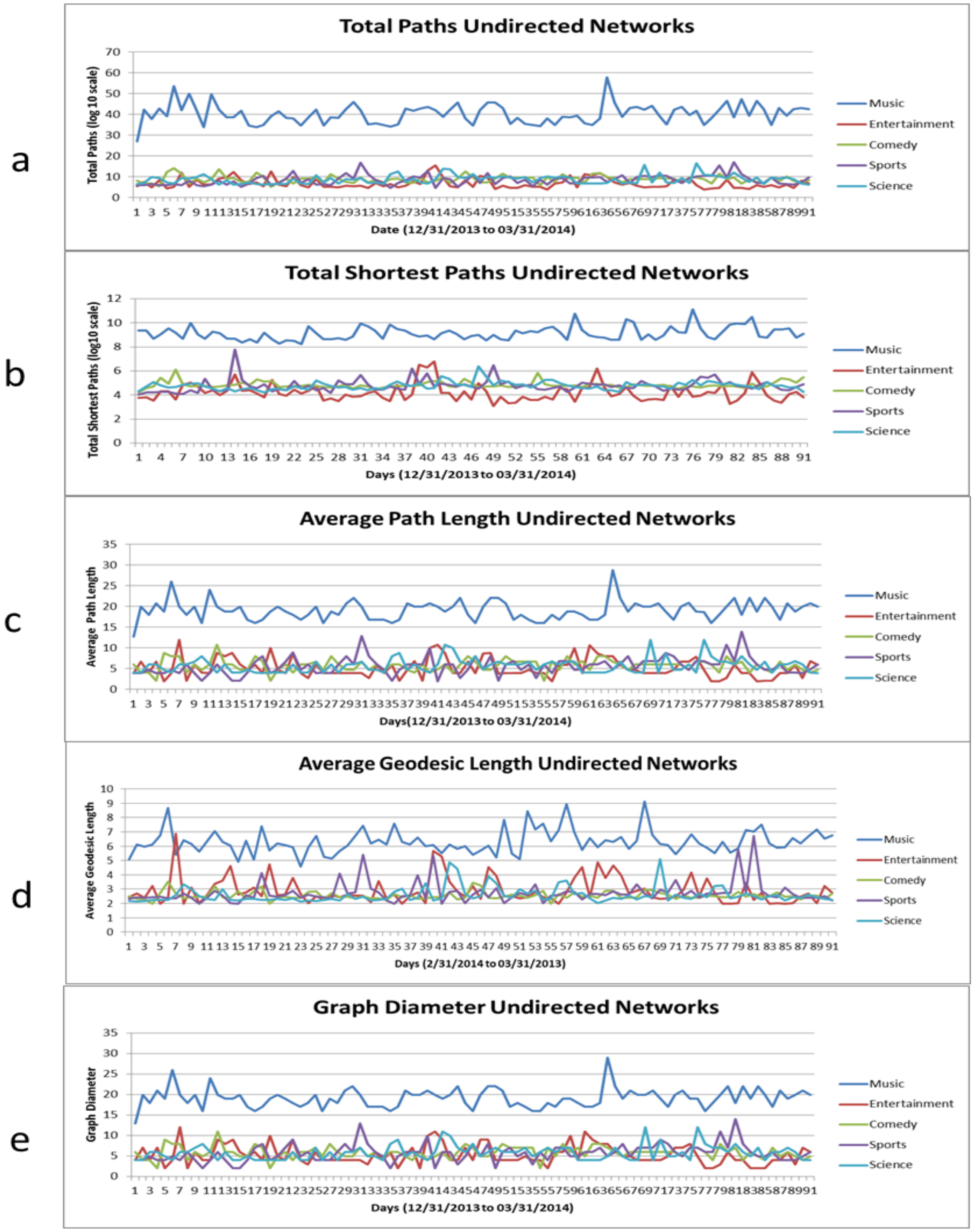


Figure 29: Undirected networks (a) Total Paths (b) Total Shortest Paths (c) Average Path Length (d) Average Geodesic Length (e) Graph Diameter



As stated in section 5.4, a factor analysis was conducted in this study with the express goal of identifying processes happening with the networks (undirected, directed, consumption and propagation). In order to show the impact of scale, I compare the changes in the factors formed by the network flow variables (Spread, Spread and Speed) with the factor formed by independent variables (Size) for the undirected network of “Music”, “Entertainment”, “Comedy”, “Sports” and “Science” categories.

As shown in section 5.4.1, the factor “Size” consists of the following variables for all categories: the total Number of nodes (Nodes) and the total number of ties (Edges_ud) in the network. As shown in section 5.4.2, the factor “Spread” consists of the following variables: the Graph Diameters (GD_ud), the Total Paths (Tpaths_ud) and the Average Path Length (AvgPL_ud). The factor “Speed” consists of the following variables: Total Shortest Paths (TSpaths_ud), and Average Geodesic Length (AvgGL_ud). The factor “Spread and Speed” consists of the following variables: Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud), Average Path Length (AvgPL_ud), Average Geodesic Length (AvgGL_ud) and Graph Diameters (GD_ud).

Table 37: Factors “Size”, “Spread” and “Spread and Speed” Along with their Cronbach Alpha Values for Undirected Networks

NETWORK TYPE	VARIABLES	FACTORS (CRONBACH'S ALPHA)				
		Music	Entertainment	Comedy	Sports	Science
Undirected	Independent	Size(0.994)	Size (0.999)	Size (0.995)	Size (0.998)	Size (0.997)
	Network Flow (MV2)	Spread(0.989)	Spread and Speed(0.937)	Spread and Speed(0.937)	Spread and Speed (0.965)	Spread and Speed (0.912)

Table37 above shows the factors formed by the network flow variables (Spread, Spread and Speed) and the factor formed by independent variables (Size) for the undirected network of "Music", "Entertainment", "Comedy", "Sports" and "Science" categories. The factor "Size" has been formed from the variable Nodes and the variable Edges_ud across all categories with significant values of Cronbach's alpha. However, different factors form from the network flow variables in different categories. In the "Music" category, "Spread" is the only significant factor. The variables that form the factor "Speed" either form independent factors or they form a factor with insignificant Cronbach's alpha (<0.60) (see Appendix A.1.7.1.2.3). As the scale of the network reduces (Figures 27 and 28), the scale of the variables that form the factors "Speed" and "Spread" also reduces (Figure 29). With the reduction of scale, the variables that measure the factors "Spread" and "Speed" load together to form a single factor labelled "Spread and Speed". This is mainly because, as the scale of the networks reduces, the difference in the magnitude of the variables measuring the processes of "Speed" and "Spread" becomes insignificant. As seen in appendices A.1.5, A.2.5, A.3.5, A.4.5 and A.6.5, the differences between the values of Total Paths (Tpaths_ud) and Total Shortest Paths (TSpaths_ud) in categories of "Entertainment", "Comedy", "Science" and "Sports" are insignificant, when compared to the differences between the values of Total Paths (Tpaths_ud) and Total Shortest Paths (TSpaths_ud) in the "Music" category. Similar trends can be seen in the case of the Average Path Length (AvgPL_ud), and the Average

Geodesic Length (AvgGL_ud) for all categories in appendices A.1.5, A.2.5, A.3.5, A.4.5, and A.6.5.

These observations imply that *the scale of networks has a significant impact on the processes that transpire within the networks. An increase or decrease in the scale of a social network gives rise to different types of processes within that network. These processes are indicative of the presence of very different social mechanisms.* This suggests that *social theories that were developed from observing real-world networks of a relatively smaller scale (hundreds or thousands of people) do not necessarily apply to online social networks of a significantly larger scale (tens of thousands or millions of people).*

6.5 Eigenvector Centrality as a Measure of Influence

Eigenvector Centrality(EVC) (Bonacich, 1972, Bonacich, 2007) has been proposed as a measure of influence in online social networks based on arguments from literature that have been made in section 2.7.3. As suggested in section 4.3.4.1, the Correlation Coefficient of Eigenvector Centrality with Total Paths and Total Shortest Paths (EVCx_TpxN and EVCx_TpxN) has been used as a measure of influence with respect to information spread and speed of information spread processes. In this section, I discuss the efficacy of using Eigenvector Centrality (EVC) as a measure of influence.

Figure 30 below depicts the correlation coefficients between Eigenvector Centrality with respect to Total Paths from a node (EVCud_TpudN) and Total Shortest

Paths from a node (EVCud_TSpudN) for the undirected network of the “Music” category. All correlations exhibit a p-value below 0.05.

Figure 30 : Music Undirected Networks Correlation Coefficients between Eigenvector Centrality vs. Total Paths and Total Shortest Paths

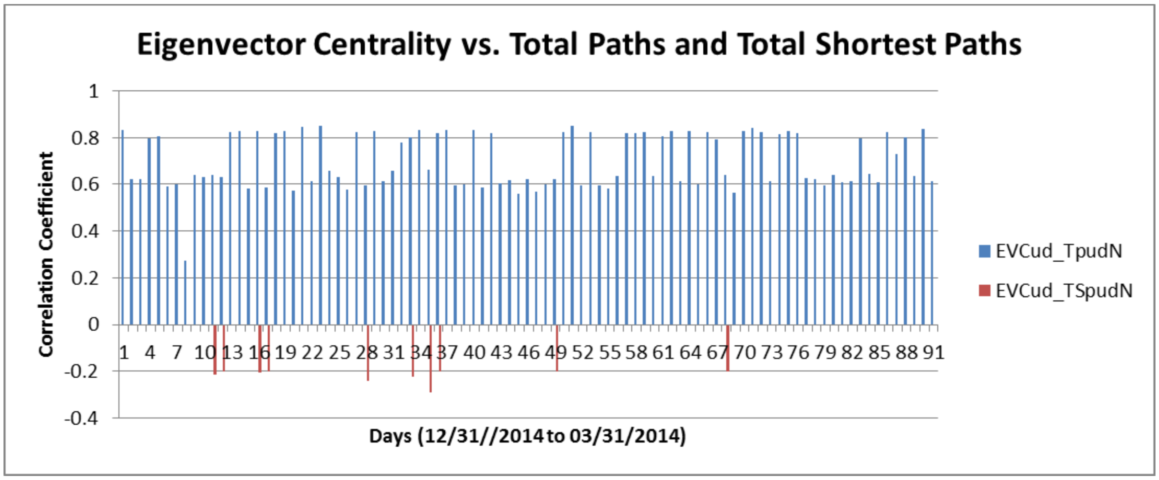
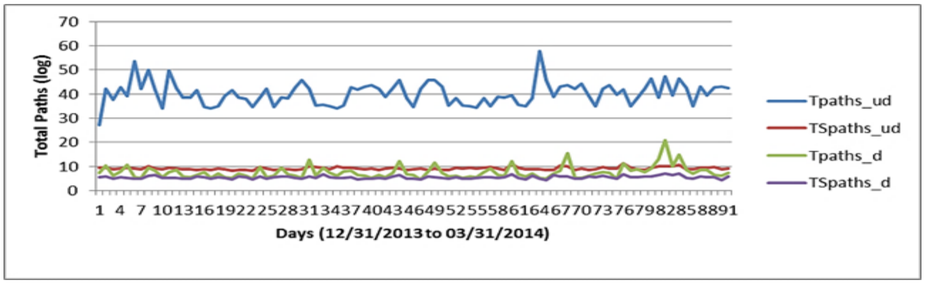


Figure 30 shows that there is a significant correlation between the Eigenvector Centrality of a node and the total number of paths from a node in the undirected network (EVCud_TpudN). There is no significant correlation between Eigenvector Centrality of a node and total number of shortest paths from a node in undirected network (EVCud_TSpudN). As seen in section 6.3, total number of paths was used as a proxy for the “Spread” process and the total number of shortest paths was used as a proxy for the “Speed” process. Based on this, it can be said that Eigenvector Centrality (EVC) is a good measure of influence in undirected networks when it comes to “Spread” process, but is not a very good measure of influence for “Speed” process. A similar trend can be seen in all undirected networks of all product categories under consideration in Appendix A (A.1.6.3, A.2.6.3, A.3.6.3, A.4.6.3, A.5.6.3, A.6.6.3)

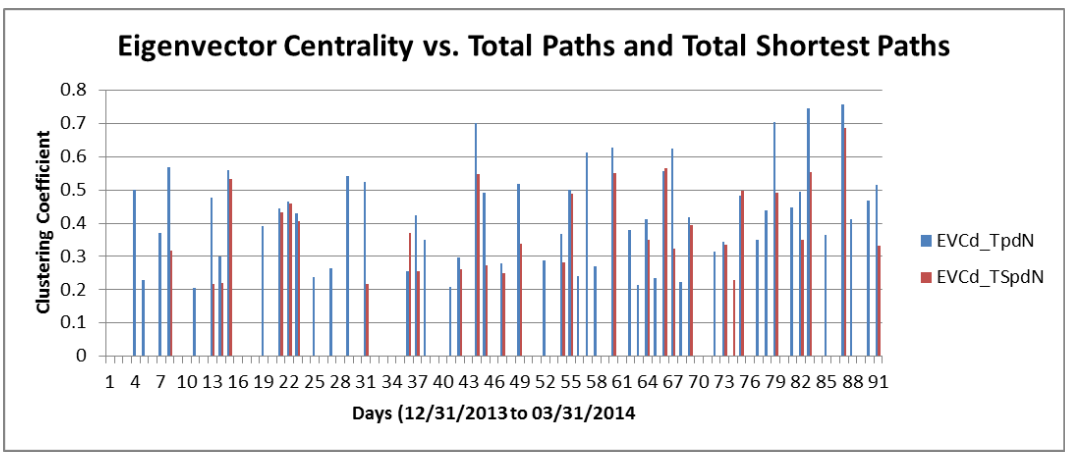
As explained in section 6.3, scale also has a significant impact on the “Spread” and “Speed” processes, which even unify at reduced scale. Reduction in scale in not just evident across categories, but also when undirected and directed networks within a category are considered. The reduction in scale across categories is due to the reduction in the number of nodes and the number of ties, as shown in Figure 27 and Figure 28 (section 6.4). The reduction in scale within the undirected and directed networks within a category is only due to the number of ties. As explained in section 5.2.1, this is mainly because in an undirected network every directed tie is considered to be symmetric. Hence every tie is counted twice, except for the ties that are already symmetric in the directed network. The impact of this reduction in scale within networks can be seen on the total paths and total shortest paths formed with respect to the undirected and directed networks. For example, as shown in Figure 31 below, in the “Music” category the difference between the Total Paths and Total Shortest Paths is significantly lower for the directed network than it is for the undirected network. The number of Total Paths and the number of Total Shortest Paths map very closely with each other in the directed network. This trend can be seen for all product categories in Appendix A (A.1.5, A.2.5, A.3.5, A.4.5, A.5.5 and A.6.5)

Figure 31: Music Category Total Paths and Total Shortest in Undirected and Directed Network



The impact of this reduction of scale in the directed network on the directed eigenvector centrality in the Music category can be seen in Figure 32 below. All correlations exhibit a p-value below 0.05.

Figure 32 : Music Directed Networks Correlation Coefficients between Eigenvector Centrality vs. Total Paths and Total Shortest Paths



In Figure 32 above, the eigenvector centrality correlates significantly more often with the total paths per node than with total shortest paths per node (EVCd_TpdN > EVCd_TSpdN) (there are more blue lines and red lines in Figure 32). On the days on which the Eigenvector Centrality correlates with both Total Paths per Node

(EVCd_TpdN) and Total Shortest Paths per Node (EVCd_TSpdN), the difference between Total Paths and Total Shortest Paths in the directed network is negligible (see Figure 32). This trend can be seen in Appendix A (A.1.1.5, A.2.1.5, A.3.1.5, A.4.1.5, A.5.1.5 and A.6.1.5) for all categories under consideration. Therefore, it can be confidently said that *Eigenvector Centrality (EVC) is a measure of influence only with respect to Total Paths per Node (Spread) but not for Total Shortest Paths per Node (Speed)*. Further research needs to be undertaken to identify metrics of measuring influence for various processes.

6.6 Experimental Metrics

The experimental metrics Power Law Distribution of Total Paths per Node (PL_TpxN), Power Law Distribution of Shortest Paths per Node (PL_TSpxN), and Power Law Distribution of Eigenvector Centrality (PL_EVCxN), were proposed in section 4.3.3.1.4 and 4.3.4.3. Table 38 below shows the output of regression for all product categories and all network types (undirected, directed, consumption and propagation). Table38 (a) shows if the network flow variables impact the Power Law Distribution of Total Paths per Node (PL_TpxN). Table38 (b) shows if the network flow variables impact the Power Law Distribution of Shortest Paths per Node (PL_TSpxN). In case an impact exists, the value in the table is represented by “Y”, else it is represented by “N”. Table38 (c) shows whether the network flow variables (NF) and the network structure variables (NS) impact the Power Law Distribution of Eigenvector Centrality (PL_EVCxN). “N” represents a lack of impact.

Table 38: Impact of Network Flow and Network Structure Variables on Power Law Distribution (a) Impact of Network Flow Variables on Power Law Distribution of Total Paths per Node, (b) Impact of Network Flow Variables on Power Law Distribution of Total Shortest Paths per Node, (c) Impact of Network Flow Variables (NF) and Network Structure Variables (NS on Power Law Distribution of Eigenvector Centrality per Node.

		Music	Entertainment	Comedy	Sports	Science
a	PL_TpudN	N	Y	Y	Y	Y
	PL_TpdN	Y	Y	Y	Y	Y
	PL_TpinN	Y	Y	Y	Y	N
	PL_TpoutN	Y	Y	N	N	N
b		Music	Entertainment	Comedy	Sports	Science
	PL_TSpudN	Y	Y	N	Y	Y
	PL_TSpdN	Y	Y	Y	Y	Y
	PL_TSpinN	Y	Y	Y	Y	Y
	PL_TSpoutN	Y	Y	N	N	N
c		Music	Entertainment	Comedy	Sports	Science
	PL_EVCudN	N	NS, NF	N	N	NS, NF
	PL_EVCdN	NS	NS, NF	NS, NF	NS, NF	NF
	PL_EVCinN	N	N	NS, NF	NS, NF	NF
	PL_EVCoutN	NF	NS, NF	NS	NS, NF	NS, NF

From Table 38 above, it can be seen that network flow variables have a significant impact on of Power Law Distribution of Total Paths per Node (PL_TpxN), Power Law Distribution of Shortest Paths per Node (PL_TSpxN). However, there is nothing in the analysis that shows the cause of the impact. Similarly, the network flow variables (NF) and network structure variables (NS) have a significant impact on Power Law Distribution of Eigenvector Centrality (PL_EVCxN). But there is nothing in the analysis that shows the cause of impact. Therefore, I conclude that more experimentation needs to be undertaken (as part of future research) to understand the cause of the various impacts shown in Table 37.

6.7 Summary of Conclusions

In this section, I summarize and restate the conclusions of my dissertation.

Conclusion 1: The size and degree of activity of online communities that discuss product lines are not necessarily correlated to the popularity of the product lines that they discuss.

Conclusion 2: The impact of network structure on network flow, the impact of network flow on network structure and the impact of network flow and network structure on the network phenomenon do exist, but their impact cannot be taken for granted.

Conclusion 3: The nature of influence within a social network cannot be understood by just analyzing the undirected or directed network. A person who might be influential in the consumption process may not be influential in propagation process, or conversely. Also, by considering the consumption *and* the propagation network, it is possible to deduce behavioral traits of a person in the network.

Conclusion 4: The scale of a network has a significant impact on the processes that transpire within the network. An increase or decrease in the scale of the network gives rise to different types of processes within a social network that are indicative of the presence of very different social mechanisms. Social theories that were developed from observing real-world networks of a relatively small scale (hundreds or thousands of people) consequently do not necessarily apply to online social networks, which can exhibit significantly larger scale (tens of thousands or millions of people).

Conclusion 5: Eigenvector Centrality (EVC) is a measure of influence only with respect to Total Paths per Node (Spread) but not for Total Shortest Paths per Node (Speed).

Conclusion 6: The introduction of new experimental metrics warrants further research.

7. Contributions and Limitations

People all around the world are utilizing online social networks at an astonishing rate, and today's marketers are responding to the increasing importance of online social networks by spending billions of dollars in digital marketing. With increased spending on social media, businesses are feeling the pressure to gain new insights into customer behavior. Success in marketing though online social media apparently critically depends upon understanding the social network that may have a potential interest in your product or service and by identifying the key attributes about the influencers that will spread your marketing message (Lindsay et al.,2014). Yet, this is easier said than done, because to date nobody really understands how online social networks get organized. Enhancing this understanding has been the primary focus of this dissertation.

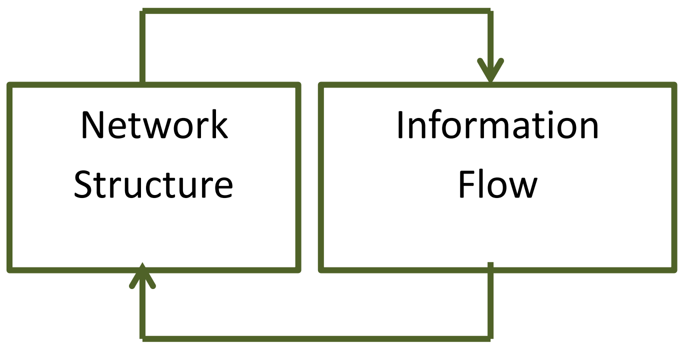
7.1 Academic Contributions

This dissertation makes contributions to various academic research streams within the fields of technology management including organizational theory, marketing and social network theory.

7.1.1 Organizational Theory and Technology Management

The primary theoretical contribution of this dissertation has resulted from addressing the stated research questions (section 2.8) and testing the hypotheses that have been derived therefrom (section 3.3). The results of the empirical portion of this dissertation suggest that network structure consistently impacts network flows, and network flows consistently impact network structure (see Figure 33). If this finding can be confirmed in other online contexts, then a fundamental property of online social networks may have been identified in this dissertation.

Figure 33: Impact of Network Structure on Network Flow and Vice Versa.



Specifically, confirming hypotheses 1 and 2 has provided concrete evidence that confirms the Theory of Structuration (Giddens, 1984, Orlikowski, 2000) in online social networks. This theory has been proven in a variety of technology management contexts (DeSanctis and Poole, 1994), (Orlikowski, 2000, Pozzebon and Pinsonneault, 2005, Walsham and Han, 1990) as well as in organization science (Van de Ven and Poole, 2005, Barley and Tolbert, 1997) and business strategy (Jarzabkowski, 2004, Biazzo, 2009). However, until now it had not been validated in online social networks. Once again, further study of online networks is warranted to establish whether structuration constitutes a broadly-based attribute of online social networks.

7.1.2 Marketing

Success in marketing on social networks depends upon identifying people who can influence the purchasing behavior of others (Brown and Hayes, 2008, Weiss, 2013, Kirby, 2012, Murphy and Schram, 2014). As of today, the measurement of influence in social networks has been based either the level of connectedness and/or the level of participation within the social network (Aral and Walker, 2011, Aral and Walker, 2012, Chomutare et al., 2014, Sasidharan et al., 2011). However, these measures of influence do not describe or predict this network phenomenon very well, and studies that characterize influence and the mechanisms that impact this network phenomenon are woefully lacking (Aral et al., 2013).

This dissertation makes an academic contribution by providing an empirically tested framework that can provide insights into the mechanisms (network flows and network structure) that impact the network phenomenon of influence. Confirming hypothesis 3 and 4 clearly shows that influence is impacted by network structure in some cases; network flows in others and by both network structure and network flows, in yet others. Further research needs to be undertaken to understand why network flow, network structure or both network structure and network flows impact phenomenon in only some cases and not in others. These findings need to be tested on various social network platforms, in order to understand whether they are broadly applicable.

This study is also the first of its kind, to the best of my knowledge, which looks at the impact of consumption and propagation of information on the network phenomenon of influence in social networks (conclusion 3). This study was able to demonstrate that network structure, network flows and their impact on influence vary signifcantly between these two modes of directionality. As a consequence, theories of online social networks, and perhaps theories social networks in general, will henceforth have to take propagation and consumption into consideration.

This dissertation also shows the impact of scale on the processes that transpire within the network (conclusion 4). An increase or decrease in the scale of the network gives rise to different types of processes within a social network that are indicative of the presence of very different social mechanisms. This observation casts severe doubt on whether extant theories of social networks, which are derived from observations of comparatively smaller social networks from the real world, apply to online social networks.

7.1.3 Social Network Theory

In earlier theories of social networks (e.g., Freeman, 1977, Freeman, 1979), measures of influence were based on connectivity within a network (section 2.7). More recent theories (e.g., Bonacich, 2007) introduced the quality of connectivity to measure influence using mesures such as Eigenvector Centrality. This lead to a better identification of the status of an individual within the network.

This dissertation points out the limitations of Eigenvector Centrality as a measure of phenomenon of influence within the social network (conclusion 5). There is a significant correlation between the Eigenvector Centrality of a node and the total number of paths from that node in the undirected network (EVCud_TpudN). There is no significant correlation between the Eigenvector Centrality of a node and the total number of shortest paths from a node in undirected network (EVCud_TSpudN). As seen in section 6.3, total number of paths was used as a proxy for the “Spread” process and the total number of shortest paths was used as a proxy for the “Speed” process. Based on this, it can be said that Eigenvector Centrality (EVC) is a good measure of influence in undirected networks when it comes to the “Spread” process, but is not a very good measure of influence for the “Speed” process. This finding casts severe doubt on theories of social networks that use EVC as a metric of influence for processes in which the speed of information propagation is considered important (e.g., Brown and Hayes, 2008, Weiss, 2013).

7.2 Contributions to Practitioners

A marketing organization might maintain a database of customers and prospective customers that are segmented according to various characteristics, and target different marketing activities to different segments. The organization may choose to invest more resources in certain segments, cross-sell to some groups, up-sell to others, and focus on reducing the cost of serving others. In such situations, the company is the main actor, addressing passive customers, whose ability to respond to the company's efforts is essentially captured in their purchasing behavior.

With the rise of social networking on a vast scale, the customer is no longer limited to a passive role in his or her relationship with a company. In addition to having more information about competitive products, customers can easily express and distribute their opinions to large audiences. Companies are likely to find it increasingly difficult to manage the messages that customers receive about their products/services. These developments are potentially detrimental to companies. If customers spread negative messages about a company, they might seriously damage its reputation.

However, the emergence of social media also offers companies opportunities to listen to and engage with their customers, and potentially to encourage them to become advocates for their products. The challenge for companies is to identify and take advantage of such opportunities, and to avoid the pitfalls they entail. The models and insights to be generated from this dissertation serve as a foundation for practicing marketing professionals, which allows them to understand the social mechanisms in the social networks they intend to target. This helps marketers make decisions regarding where to spend their resources, so that they can engage the right stakeholders and convert them into advocates. The study in this dissertation is also the first of its kind that has been undertaken to understand the impact of consumption and propagation of information within social networks, to the best of my knowledge. This will help managers understand the communication patterns within the social network, allowing them to allocate and optimize their resources accordingly.

7.3 Limitations

This study has looked at the impact of change within a social network. Specifically, it has investigated how changes in network's structural characteristics and network flows affect each other, as well as their impact on the phenomenon of influence. I identify the following limitations that pertain to this research:

1. This study does not look at the causes of change within a social network's structure or information flow. For example, how governance mechanisms, task complexities or emergent roles relate to formation of network structure, information flow or influence of an individual within a social network are not covered in this dissertation.

2. Though social networks like Twitter and YouTube provide access to a wide variety of participants, their real identity cannot be confirmed. This makes it difficult to glean demographic information like age, sex, etc.

These limitations can be overcome by follow-on research that transpires in different contexts. Further research (by others) will determine which of the lessons learned from this dissertation can be generalized to other kinds of networks (e.g., other social networks on Twitter, online social networks on other platforms, trading, e-commerce, etc…). Conclusions drawn from an aggregation of these studies could serve as the foundation of a more broadly-based theory of online social networks.

References

Abbott, A. 1990. A Primer on Sequence Methods. *Organization Science,* 1**,** 375-392.

Adamic, L. & Adar, E. 2005. How to search a social network. *Social Networks,* 27**,** 187-203.

Afsarmanesh, H. & Camarinha-Matos, L. M. 2005. A Framework for Management of Virtual Organization Breeding Environments. Valencia, Spain: *In Collaborative Networks and their Breeding Environments, (PRO-VE'05)*.

Albert, R. & Barabasi, A. L. 2000. Topology of Evolving Networks: Local Events and Universality. *Physical Review Letters,* 85**,** 5234-5237.

Albert, R., Jeong, H. & Barabasi, A. L. 2000. Error and Attack Tolerance of Complex Networks. *Nature,* 406**,** 378-382.

Alderson, D. & Willinger, W. 2005. A Contrasting Look at Self-Organization in the Internet and Next-Generation Communication Networks. *Communications Magazine, IEEE,* 43**,** 94-100.

Allen, T. J. 1977. *Managing the Flow of Technology,* Boston, MIT Press.

Allison, P. D. 1999. *Multiple regression: A primer*, Pine Forge Press.

Anderson, B. S., Butts, C. & Carley, K. 1999. The Interaction of Size and Density with Graph-Level Indices. *Social Networks,* 21**,** 239-267.

Aral, S., Dellarocas, C. & Godes, D. 2013. Introduction to the special issue-social media and business transformation: A framework for research. *Information Systems Research,* 24**,** 3-13.

Aral, S. & Walker, D. 2011. Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks. *Management Science,* 57**,** 1623-1639.

Aral, S. & Walker, D. 2012. Identifying Influential and Susceptible Members of Social Networks. *Science,* 337**,** 337-341.

Atkin, R. H. 1974. *Mathematical Structure in Human Affairs,* London, Heinemann.

Ayres, C. 2009. Revenge is best served cold – on YouTube. *The Times*.

Baerveldt, C. & Snijders, T. 1994. Influences On and From the Segmentation of Networks - Hypothesis and Tests. *Social Networks,* 16**,** 213-232.

Bailey, K. D. 1990. Social Entropy Theory: An overview. *Systemic Practice and Action Research,* 3**,** 365-382.

Barabási, A.-L. & Albert, R. 1999. Emergence of Scaling in Random Networks. *Science,* 286**,** 509-512.

Barabási, A.-L. & Bonabeau, E. 2003. Scale-Free Networks. *Scientific American,* 288**,** 60 - 69.

Barley, S. R. & Tolbert, P. S. 1997. Institutionalization and structuration: Studying the links between action and institution. *Organization studies,* 18**,** 93-117.

Baskerville, R. & Pries-Heje, J. 2001. A multiple-theory analysis of a diffusion of information technology case. *Information Systems Journal,* 11**,** 181-212.

Bavelas, A. 1950. Communication Patterns in Task-Oriented Groups. *Journal of the Acoustical Society of America,* 22**,** 725 - 730.

Bača, M., Schatten, M. & Deranja, D. 2007. Autopoietic information systems in modern organizations. *Organizacija,* 40.

Berelson, B. R., Lazarsfeld, P. F. & McPhee, W. N. 1954. *Voting: A Study of Opinion in a Presidential Campaign,* Chicago, University of Chicago Press.

Biazzo, S. 2009. Flexibility, structuration, and simultaneity in new product development. *Journal of Product Innovation Management,* 26**,** 336-353.

Blossfeld, H.-P. & Rohwer, G. 1995. *Techniques of event history modeling: New approaches to causal analysis (2nd ed.)*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Bollobás, B. 1984. The Evolution of Random Graphs. *Transactions of the American Mathematical Society,* 286**,** 257 -274.

Bonacich, P. 1972. FACTORING AND WEIGHTING APPROACHES TO STATUS SCORES AND CLIQUE IDENTIFICATION. *Journal of Mathematical Sociology,* 2**,** 113-120.

Bonacich, P. 2007. Some unique properties of eigenvector centrality. *Social Networks,* 29**,** 555-564.

Borgatti, S. P. 2005. Centrality and Network Flow. *Social Networks,* 27**,** 55-71.

Borgatti, S. P. & Cross, R. 2003. A Relational View of Information Seeking and Learning in Social Networks. *Management Science,* 49**,** 432-445.

Borgatti, S. P. & Foster, P. C. 2003. The network paradigm in organizational research: A review and typology. *Journal of Management,* 29**,** 991-1013.

Borgatti, S. P. & Halgin, D. S. 2011. On Network Theory. *Organization Science,* 22**,** 1168-1181.

Borgatti, S. P. & Kidwell, V. 2011. *Network Theorizing*, Sage Publications.

Bourdieu, P. 1986. *The Forms of Capital,* New York, Greenwood.

Bressler, S. E. & Grantham, C. 2000. *Communities of Commerce: Building Internet Business Communities to Accelerate Growth,Minimize Risk,and Increase Customer Loyalty,* New York, NY, McGraw-Hill, Inc.

Bristor, J. M. 1989. *Enhanced Explanations of Word of Mouth Communications : the Power of Relationships*, Research and Publications, School of Business Administration, University of Western Ontario.

Brown, D. & Hayes, N. 2008. *Influencer Marketing: Who really influences your customers?*, Routledge.

Brown, D. W., Giles, W. H. & Greenlund, K. J. 2007. Blood pressure parameters and risk of fatal stroke, NHANES II mortality study. *American journal of hypertension,* 20**,** 338-341.

Brown, J. J. & Reingen, P. H. 1987. Social Ties and Word-of-Mouth Referral Behavior. *Journal of Consumer Research,* 14**,** 350-362.

Brumec, J. 1997. A contribution to IS general taxonomy. *Journal of Information and Organizational Sciences,* 21**,** 1-14.

Bryant, D. & Merker, S. L. 1987. A living systems process analysis of a public transit system. *Behavioral Science,* 32**,** 293-303.

Burt, R. S. 1976. Positions in Networks. *Social Forces,* 55**,** 93-122.

Burt, R. S. 1987. Social Contagion and Innovation - Cohesion versus Structural Equivalence. *American Journal of Sociology,* 92**,** 1287-1335.

Burt, R. S. 1992. *Structural Holes: The Social Structure of Competition,* Cambridge, Mass., Harvard University Press.

Burt, R. S. 2005. *Brokerage and Closure: An Introduction to Social Capital,* Oxford -New York, Oxford University Press.

Burt, R. S. & Doreian, P. 1982. Testing a Structural Model of Perception - Conformity and Deviance with Respect to Journal Norms in Elite Sociological Methodology. *Quality & Quantity,* 16**,** 109-150.

Burt, R. S., Kilduff, M. & Tasselli, S. 2013. Social Network Analysis: Foundations and Frontiers on Advantage. *Annual Review of Psychology,* 64**,** 527-547.

Cartwright, D. 1965. *Influence, Leadership, Control,* Chicago, Rand McNally.

Castellano, C., Fortunato, S. & Loreto, V. 2009. Statistical physics of social dynamics. *Reviews of Modern Physics,* 81**,** 591-646.

Cattell, R. B. 1965. Factor Analysis: An Introduction to Essential II. The Role of Factor Analysis in Research. *Biometrics,* 21**,** 405 - 435.

Cattell, R. B. 1966. The scree test for the number of factors. *Multivariate behavioral research,* 1**,** 245-276.

Centola, D. 2010. The Spread of Behavior in an Online Social Network Experiment. *Science,* 329**,** 1194-1197.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, P. K. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM,* 10**,** 10-17.

Cha, M., Mislove, A. & Gummadi, K. P. A measurement-driven analysis of information propagation in the flickr social network. *Proceedings of the 18th international conference on World wide web, 2009. ACM*, 721-730.

Chakrabarti, R. & Berthon, P. 2012. Gift giving and social emotions: experience as content. *Journal of Public Affairs,* 12**,** 154-161.

Chen, W., Wang, Y. & Yang, S. *2009.* Efficient influence maximization in social networks. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009. ACM,* 199-208.

Chomutare, T., Xu, A. & Iyengar, M. S. 2014. Social Network Analysis to Delineate Interaction Patterns That Predict Weight Loss Performance. I*EEE 27th International Symposium on Computer-Based Medical Systems (CBMS)*, 27-29 May 2014 2014. 271-276.

Clauset, A., Shalizi, C. & Newman, M. 2009. Power-Law Distributions in Empirical Data. *SIAM Review,* 51**,** 661-703.

Coleman, J. 1988. Social Capital in the Creation of Human Capital. *American Journal of Sociology,* 94**,** S95-S120.

Cook, R. D. 1977. Detection of influential observation in linear regression. *Technometrics*, 15-18.
Cook, R. D. 1979. Influential observations in linear regression. *Journal of the American Statistical Association,* 74, 169-174.
Cooper, D. R. & Emory, C. W. 1995. *Business Research Methods, Chicago: Richard D. Irwin*, Inc.
Cortina, J. M. 1993. What is Coefficient Alpha - An Examination of Theory and Applications. *Journal of Applied Psychology,* 78, 98-104.
Coumarbatch, J., Robinson, L., Thomas, R. & Bridge, P. D. 2010. Strategies for identifying students at risk for USMLE step 1 failure. *Fam Med,* 42, 105-10.
Cronbach, L. J. 1990. *Essentials of Psychological Testing (5th ed.),* Boston, Addison-Wesley.
Cummings, L. L. & DeCotiis, T. A. 1973. Organizational correlates of perceived stress in a professional organization. *Public Personnel Management*.
Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. 2005. Comparing Community Structure Identification. *Journal of Statistical Mechanics: Theory and Experiment,* 2005, P09008-P09008.
Deighton, J. 2012. What Are Marketers' Top Concerns? (MSI Research Priorities 2012-14). Jun 25,2012 ed.: Marketing Science Institute.
Dellarocas, C., Katona, Z. & Rand, W. 2013. Media, Aggregators, and the Link Economy: Strategic Hyperlink Formation in Content Networks. *Management Science,* 59, 2360-2379.
DeSanctis, G. & Poole, M. S. 1994. Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization science,* 5, 121-147.
Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. & Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one,* 6, e26752.
Dodds, P. S., Muhamad, R. & Watts, D. J. 2003. An Experimental Study of Search in Global Social Networks. *Science,* 301, 827-829.
Donninger, C. 1986. The Distribution of Centrality in Social Networks. *Social Networks,* 8, 191-203.
Dosi, G. 1982. Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research policy,* 11, 147-162.
Doyle, J. C., Alderson, D. L., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R. & Willinger, W. 2005. The "Robust yet Fragile" Nature of the Internet. *Proceedings of the National Academy of Sciences of the United States of America,* 102, 14497-14502.
Drazin, R. & Sandelands, L. 1992. Autogenesis - A Perspective on the Process of Organizing. *Organization Science,* 3, 230-249.

Duhan, D. F., Johnson, S. D., Wilcox, J. B. & Harrell, G. D. 1997. Influences on Consumer Use of Word-of-Mouth Recommendation Sources. *Journal of the Academy of Marketing Science,* 25**,** 283-295.

Duncan, D. M. 1972. James G. Miller's Living Systems Theory: Issues for Management Thought and Practice A REVIEW ESSAY. *Academy of Management Journal,* 15**,** 513-523.

Dunn, O. J. 1959. Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics,* 30**,** 192-197.

Dunn, O. J. 1961. Multiple Comparisons among Means. *Journal of the American Statistical Association,* 56**,** 52-64.

Edward, J. 2012. DATA: Google Totally Blows Away Facebook On Ad Performance. *Business Insider.* MAY 15, 2012, 6:00 AM ed.

Eisenhardt, K. M. 1989. Building Theories from Case-Study Research. *Academy of Management Review,* 14**,** 532-550.

Erdös, P. & Rényi, A. 1959. On Random Graphs I. *Publicationes Mathematicae (Debrecen),* 6**,** 290-297.

Erickson, B. H. 1988. The Relational Basis of Attitudes. *In:* Welman, B. & Berkowitz, S. D. (eds.) *Social Structures: A Network Approach.* New York: Cambridge University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. 1999. Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods,* 4.

Feller, J., Finnegan, P., Fitzgerald, B. & Hayes, J. 2008. From Peer Production to Productization: A Study of Socially Enabled Business Exchanges in Open Source Service Networks. *Information Systems Research,* 19**,** 475-493.

Festinger, L., Schachter, S. & Back, K. 1950. Social Pressures of an Informal Groups: A Study of Human Factors of Housing. New York.: Harper.

Fidell, L. S. & Tabachnick, B. G. 2003. Preparatory data analysis. *Handbook of psychology*.

Field, A. P. 2005. *Discovering Statistics using SPSS*, SAGE publications Ltd.

Frank, M. R., Mitchell, L., Dodds, P. S. & Danforth, C. M. 2013. Happiness and the Patterns of Life: A Study of Geolocated Tweets. *Sci. Rep.,* 3.

Freeman, L. 1977. A Set of Measures of Centrality Based on Betweenness. *Sociometry,* 40**,** 35-41.

Freeman, L. C. 1979. Centrality in Social Networks Conceptual Clarification. *Social Networks,* 1**,** 215-239.

Friedkin, N. E. 1983. Horizons of Observability and Limits of Informal Control in Organizations. *Social Forces,* 62**,** 54-77.

Fronczak, A., Fronczak, P. & Hołyst, J. 2004. Average Path Length in Random Networks. *Physical Review E,* 70**,** 056110.

Gartner, T. 2002. Exponential and Geometric Kernels for Graphs. *Proceedings of Neural Information Processing Systems Workshop on Unreal Data: Principles of Modeling Nonvectorial Data*, 2002.

George, H. D. 1989. *Principal components analysis*, Sage.

Giddens, A. 1984. *The Constitution of Society : Outline of the Theory of Structuration*, Polity Press.

Goh, K. I., Oh, E., Jeong, H., Kahng, B. & Kim, D. 2002. Classification of Scale-Free Networks. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 12583-12588.

Gonçalves, B. & Ramasco, J. J. 2008. Human dynamics revealed through Web analytics. *Physical Review E,* 78**,** 026123.

Gouldner, A. W. The coming crisis of western sociology. 1973. 188-198.

Granovetter, M. S. 1973. The Strength of Weak Ties. *American Journal of Sociology,* 78**,** 1360-1380.

Greenemeier, L. 2009. Inflated Expectations: Crowd-Sourcing Comes of Age in the DARPA Network Challenge. *Scientific American*.

Hair, J., Tatham, R., Anderson, R. & Black, W. 1998. *Multivariate Data Analysis (5th Edition)*, Prentice Hall.

Herman, J. 2003. The New Science of Networks. *Business Communications Review,* 33**,** 22-23.

Hodas, N. O. & Lerman, K. 2012. How Visibility and Divided Attention Constrain Social Contagion. Privacy, Security, Risk and Trust (PASSAT), *2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 3-5 Sept. 2012. 249-257.

Hodas, N. O. & Lerman, K. 2014. The Simple Rules of Social Contagion. *Sci. Rep.,* 4.

Homans, G. 1950. *The Human Group,* New York, Harcourt Brace.

Homans, G. 1974. *Social Behavior: Its Elementary Forms,* New York, Harcourt, Brace, Jovanovich.

Honeycutt, C. & Herring, S. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. *Hawaii International Conference on System Sciences,* 0**,** 1-10.

Hopfield, J. J. & Herz, A. V. M. 1995. Rapid Local Synchronization of Action-Potentials - Toward Computation with Coupled Integrate-and-Fire Neurons. *Proceedings of the National Academy of Sciences of the United States of America,* 92**,** 6655-6662.

Huberman, B. A., Romero, D. M. & Wu, F. 2008. Social Networks that Matter: Twitter Under the Microscope. *CoRR,* abs/0812.1045.

Itzkovitz, S., Levitt, R., Kashtan, N., Milo, R., Itzkovitz, M. & Alon, U. 2005. Coarse-Graining and Self-Dissimilarity of Complex Networks. *Physical Review E,* 71.

Jansen, B., Zhang, M., Sobel, K. & Chowdury, A. 2009. Twitter Power: Tweets as Electronic Word of Mouth. *J. Am. Soc. Inf. Sci. Technol.,* 60**,** 2169-2188.

Jarzabkowski, P. 2004. Strategy as practice: recursiveness, adaptation, and practices-in-use. *Organization studies,* 25**,** 529-560.
Jeong, H. 2003. Complex Scale-Free Networks. *Physica A: Statistical Mechanics and its Applications,* 321**,** 226-237.
Judd, C. M., Smith, E. R. & Kidde, L. H. 1991. *Research Methods in Social Relations (6th ed.),* Fort Worth, TX, Holt, Rinehart and Winston.
Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educational and psychological measurement*.
Kashima, H., Tsuda, K. & Inokuchi, A. 2003. Marginalized Kernels between Labeled Graphs. *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, 2003. 321-328.
Katz, E. & Lazarsfeld, P. F. 1955. *Personal Influence,* Glencoe,IL, The Free Press.
Keeling, M. 2005. The Implications of Network Structure for Epidemic Dynamics. *Theoretical Population Biology,* 67**,** 1-8.
Kejzar, N., Nikoloski, Z. & Batagelj, V. 2008. Probabilistic Inductive Classes of Graphs. *Journal of Mathematical Sociology,* 32**,** 85-109.
Kennedy, P. 2003. *A guide to econometrics*, MIT press.
Kerlinger, F. N. & Lee, H. B. 1964. *Foundations of Behavioral Research: Educational and Psychological Inquiry*, Holt, Rinehart and Winston New York.
Khammash, M. & Griffiths, G. H. 2011. 'Arrivederci CIAO.com, Buongiorno Bing.com'— Electronic word-of-mouth (eWOM), antecedences and consequences. *International Journal of Information Management,* 31**,** 82-87.
Kirby, J. 2012. Viral marketing. *Connected marketing***,** 87-106.
Klemm, K. & Eguiluz, V. M. 2002. Growing Scale-Free Networks with Small-World Behavior. *Physical Review E,* 65.
Knoke, D. & Kuklinski, J. H. 1982. *Network Analysis,* Beverly Hills, CA, Sage.
Kochen, M. 1989. *The Small World: A Volume of Recent Research Advances Commemorating Ithiel de Sola Pool, Stanley Milgram, Theodore Newcomb.,* Norwood, NJ, Ablex Publishing Corp.
Kotler, P. 1994. *Marketing Management: Analysis, Planning,Implementation and Control, 8th ed.,* Englewood Cliffs, NJ, Prentice-Hall.
Kretzschmar, M. & Morris, M. 1996. Measures of Concurrency in Networks and the Spread of Infectious Disease. *Mathematical Biosciences,* 133**,** 165-195.
Kwak, H., Lee, C., Park, H. & Moon, S. 2010. What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web, 2010, ACM*, 591-600.
Lazarsfeld, P., Berelson, B. & Gaudet, H. 1944. *The people's choice; how the voter makes up his mind in a presidential campaign*, Duell Sloan and Pearce.
Leenders, R. T. A. J. 1995. *Structure and Influence: Statistical Models for the Dynamics of Actor Attributes, Network Structure and Their Interdependence,* Amsterdam, Thela Thesis Publishers.

Leenders, R. T. A. J. 2002. Modeling Social Influence through Network Autocorrelation: Constructing the Weight Matrix. *Social Networks,* 24**,** 21-47.

Leicht, E. A. & Newman, M. E. J. 2008. Community Structure in Directed Networks. *Physical Review Letters,* 100**,** 118703.

Leonard-Barton, D. 1990. A Dual Methodology for Case Studies: Synergistic use of a Longitudinal Single Site with Replicated Multiple Sites. *Organization Science,* 1**,** 248-266.

Li, C. & Bernoff, J. 2008. *Groundswell: Winning in a World Transformed by Social Technologies,* Boston, Harvard Business Press.

Li, L., Alderson, D., Doyle, J. C. & Willinger, W. 2005. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. *Internet Mathematics,* 2**,** 431-523.

Lichtman, C. M. & Hunt, R. G. 1971. Personality and organization theory: a review of some conceptual literature. *Psychological bulletin,* 76**,** 271.

Lindsay, A., Kaykas-Wolff, J. & Mathwick, C. 2010. Key Influencers: Locating, Measuring, and Creating Programs to Influence Social Media Influencers. Proceedings of the 2010 Academy of Marketing Science (AMS) Annual Conference, p 1.

Linton, R. 1936. *The Study of Man*, Appleton Century Crofts, Inc.

Longart, P. 2010. What drives word-of-mouth in restaurants? *International Journal of Contemporary Hospitality Management,* 22**,** 121-128.

Lorrain, F. & White, H. 1971. Structural Equivalence of Individuals in Social Networks. *The Journal of Mathematical Sociology,* 1**,** 49-80.

Luhmann, N. 1986. The Autopoiesis of Social Systems. *In:* F. Geyer, a. J. V. d. Z. (ed.) *Sociocybernetic Paradoxes: Observation, Control and Evolution of Self-Steering System.* London: Sage.

Maleković, M. & Schatten, M. 2008. Leadership in Team Based Knowledge Management-An Autopoietic Information System's Perspective. *Hrvatska znanstvena bibliografija i MZOS-Svibor.* 2008

March, J. G. 1955. An Introduction to the Theory and Measurement of Influence. *American Political Science Review,* 49**,** 431-451.

Marquaridt, D. W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics,* 12**,** 591-612.

Marsden, P. & Friedkin, N. 1993. Network Studies of Social Influence. *Sociological Methods Research,* 22**,** 127-151.

Meiss, M. R., Menczer, F., Fortunato, S., Flammini, A. & Vespignani, A. 2008. Ranking web sites with real user traffic. *Proceedings of the 2008 International Conference on Web Search and Data Mining.* Palo Alto, California, USA: ACM.

Menard, S. 2002. *Applied logistic regression analysis*, Sage.

Merker, S. L. & Lusher, C. 1987. A living systems process analysis of an urban hospital. *Behavioral science,* 32**,** 304-314.

Merton, R. K. 1957. *Social Theory and Social Structure,* New York, The Free Press.

Merton, R. K. 1973. *The sociology of science: Theoretical and empirical investigations*, University of Chicago press.
Milgram, S. 1967. The Small World Problem. *Psychology Today,* 2**,** 60-67.
Miller, J. G. 1978. *Living Systems,* New York, McGraw-Hill.
Mills, C. W. 2000. *The sociological imagination*, Oxford University Press.
Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B. & Magoulas, R. 2008. *Twitter and the micro-messaging revolution: Communication, connections, and immediacy—140 characters at a time,* Sebastopol, CA, O'Reilly Media.
Mingers, J. 2003. *Observing organizations: An evaluation of Luhmann's organization theory*, Copenhagen Business School Press.
Misner, I. R. 1999. *The World's Best Known Marketing Secret: Building Your Business with Word-of-Mouth Marketing, 2nd ed.,* Austin, TX, Bard Press.
Moffitt, R. A. 2001. Policy interventions, low-level equilibria, and social interactions. *Social dynamics,* 4**,** 45-82.
Mokken, R. J. 1979. Cliques, Clubs and Clans. *Quality & Quantity,* 13**,** 161-173.
Money, R. B., Gilly, M. C. & Graham, J. L. 1998. Explorations of National Culture and Word-of Mouth Referral Behavior in the Purchase of Industrial Services in the United States and Japan. *Journal of Marketing,* Vol. 62**,** p76.
Montgomery, D. 2008. *Design and Analysis of Experiments*, Wiley.
Muchnik, L., Aral, S. & Taylor, S. J. 2013a. Social Influence Bias: A Randomized Experiment. *Science,* 341**,** 647-651.
Muchnik, L., Pei, S., Parra, L. C., Reis, S. D. S., Andrade Jr, J. S., Havlin, S. & Makse, H. A. 2013b. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Sci. Rep.,* 3.
Murphy, K. R. & Davidshofer, C. O. 1991. Psychological Testing: Principles and Applications. Prentice-Hall, Inc.
Murphy, T. & Schram, R. 2014. What is it worth? The value chasm between brand and influencers. *Journal of Brand Strategy,* 3**,** 31-40.
Nadel, S. F. 1957. The Theory of Social Structure. London: Cohen& West.
Neter, J., Kutner, M. H., Nachtsheim, C. J. & Wasserman, W. 1996. *Applied linear statistical models*, Irwin Chicago.
Newman, M. E. J. 2002. Assortative Mixing in Networks. *Physical Review Letters,* 89**,** 208701.
Newman, M. E. J., Watts, D. J. & Strogatz, S. H. 2002. Random Graph Models of Social Networks. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 2566-2572.
Noell, J. J. 1974. On the administrative sector of social systems: an analysis of the size and complexity of government bureaucracies in the American States. *Social Forces,* 52**,** 549-558.
Norris, J. R. 1998. *Markov Chains*, Cambridge University Press.
Orlikowski, W. J. 2000. Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science,* 11**,** 404-428.

O'brien, R. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity,* 41**,** 673-690.

Pan, Y. & Jackson, R. T. 2008. Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males. *Epidemiology and infection,* 136**,** 421-431.

Paperin, G., Green, D. G. & Leishman, T. G. 2008. Dual Phase Evolution and Self-organisation in Networks. *Simulated Evolution and Learning, Proceedings,* 5361**,** 575-584.

Parson, T. 1951. *The Social System,* New York, The Free Press.

Pastor-Satorras, R. & Vespignani, A. 2004. *Evolution and Structure of Internet: A Statistical Physics Approach*, Cambridge University Press.

Pei, S., Muchnik, L., Andrade, J. J. S., Zheng, Z. & Makse, H. A. 2014. Searching for superspreaders of information in real-world social media. *Sci. Rep.,* 4.

Pennington, D. C. 2003. *Essential Personality*, A Hodder Arnold Publication.

Podolny, J. M. 1993. A Status-Based Model of Market Competition. *American Journal of Sociology,* 98**,** 829-872.

Powell, W. W. 1990. Neither Market nor Hierarchy - Network Forms of Organization. *Research in Organizational Behavior,* 12**,** 295-336.

Pozzebon, M. & Pinsonneault, A. 2005. Challenges in conducting empirical work using structuration theory: Learning from IT research. *Organization studies,* 26**,** 1353-1376.

Preece, J. 2000. *Online Communities: Designing Usability and Supporting Sociability*, John Wiley & Sons.

Putnam, R. 1995. Bowling Alone: America's Declining Social Capital. *Journal of Democracy,* 6.

Ramage, D., Dumais, S. & Liebling, D. Characterizing Microblogs with Topic Models. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010. AAAI.

Rapoport, A. & Horvath, W. J. 1961. A Study of a Large Sociogram. *Behavioral Science,* 6**,** 279-291.

Reese, W. G. 1972. An essay on administration. *American Journal of Psychiatry,* 128**,** 1249-1256.

Rheingold, H. 1993. *The Virtual Community: Homesteading on the Electronic Frontier,* New York, Addison-Wesley.

Rice, R. 1992. Task Analyzability, Use of New Media, and Effectiveness: A Multi-Site Exploration of Media Richness. *Organization Science,* 3**,** 475-500.

Robertson, R. 1992. *Globalization: Social theory and global culture*, Sage.

Roger, E. M. 1983. *Diffusion of Innovations ,*New York, Free Press.

Rogers, E. M. 2003. *Diffusion of innovations,* New York, Free Press.

Rogers, E. M., Medina, U. E., Rivera, M. A. & Wiley, C. J. 2005. Complex adaptive systems and the diffusion of innovations. *The Innovation Journal: The Public Sector Innovation Journal,* 10**,** 1-26.

Rogerson, P. A. 2010. *Statistical methods for geography: a student's guide*, Sage Publications.
Rubin, A. M. 1984. Ritualized and Instrumental Television Viewing. *Journal of Communication,* 34**,** 67-77.
Ruscoe, G. C., Fell, R. L., Hunt, K. T., Merker, S. L., Peter, L. R., Cary, J. S., Miller, J. G., Loo, B. G., Reed, R. W. & Sturm, M. I. 1985. The application of living systems theory to 41 US Army Battalions. *Behavioral science,* 30**,** 7-50.
Rusli, E. M. & Eavis, P. 2012. Facebook Raises $16 Billion in I.P.O. *The New York Times*.
Sandelands, L. & Drazin, R. 1989. On The Language of Organization Theory. *Organization Studies,* 10**,** 457-478.
Sasidharan, S., Santhanam, R., Brass, D. J. & Sambamurthy, V. 2011. The Effects of Social Network Structure on Enterprise Systems Success: A Longitudinal Multilevel Analysis. *Information Systems Research,* 23**,** 658-678.
Schoberth, T. & Schrott, G. 2001. Virtual Communities. *Wirtschaftsinformatik,* 43**,** 517-519.
Seidman, S. & Foster, B. 1978. A Graph-Theoretic Generalization of the Clique Concept. *Journal of Mathematical Sociology,* 6**,** 139-154.
Shaw, M. E. 1954. Group Structure and the Behavior of Individuals in Small Groups. *The Journal of Psychology,* 38**,** 139-149.
Shimbel, A. 1953. Structural Parameters of Communication Networks. *Bull Math Biophys,* 15**,** 501-507.
Simon, H. A. 1957. *Models of Man,* New York, Wiley.
Sola Pool, I. d. & Kochen, M. 1978–1979. Contacts and Influence. *Social Networks,* 1**,** 5-51.
Stewart, D. W. & Ward, S. 1994. Media Effects on Advertising. *In:* Bryand, J. & Zillman, D. (eds.) *Media Effects, Advances in Theory and Research.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Swanson, G. A. 1993. *Macro accounting and modern money supplies*, Quorum Books.
Swanson, G. A., Bailey, K. D. & Miller, J. G. 1997. Entropy, social entropy and money: A living systems theory perspective. *Systems Research and Behavioral Science,* 14**,** 45-65.
Swanson, G. A. & Miller, J. G. 1989. *Measurement and interpretation in accounting: a living systems theory approach*, Quorum Books.
Tajfel, H. 1972. La Catégorisation Sociale. *In:* Serge, M. (ed.) *Introduction á la Psychologie Sociale.* Larouse, Paris.
Teevan, J., Ramage, D. & Morris, M. R. #TwitterSearch: A Comparison of Microblog Search and Web Search. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011 Hong Kong, China. 35-44.
Terlep, S., Vranica, S. & Raice, S. 2012. GM Says Facebook Ads Don't Pay Off. May 16, 2012 ed.: The Wall Street Journal.

Thelwall, M., Sud, P. & Vis, F. 2012. Commenting on YouTube Videos: From Guatemalan Rock to El Big Bang. *Journal of the American Society for Information Science and Technology,* 63**,** 616-629.

Travers, J. & Milgram, S. 1969. An Experimental Study of the Small World Problem. *Sociometry,* 32**,** 425-443.

Turner, S. P. 1985. " Theoretical Logic in Sociology", Vol. 4:" The Modern Reconstruction of Classical Thought" by Jeffrey C. Alexander.

Uzzi, B. 1997. Towards a Network Perspective on Organizational Decline. *The International Journal of Sociology and Social Policy,* 17**,** 111-155.

Valafar, M., Rejaie, R. & Willinger, W. Beyond friendship graphs: a study of user interactions in Flickr. 2009. ACM, 25-30.

Van de Ven, A. H. & Poole, M. S. 2005. Alternative approaches for studying organizational change. *Organization studies,* 26**,** 1377-1404.

Viswanath, B., Mislove, A., Cha, M. & Gummadi, K. P. 2009. On the evolution of user interaction in facebook. *Proceedings of the 2nd ACM workshop on Online social networks*, 2009. ACM, 37-42.

Walsham, G. & Han, C. K. 1990. *Structuration theory and information systems research*, Management Studies Group, Department of Engineering, University of Cambridge.

Wasserman, S. & Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D. J. 1999. *Small-worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, NJ (USA).

Watts, D. J., Dodds, P. S. & Newman, M. E. J. 2002. Identity and Search in Social Networks. *Science,* 296**,** 1302-1305.

Watts, D. J. & Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature,* 393**,** 440-442.

Weekes, W. H. 1984. *A general systems approach to management accounting*, Intersystems Publ.

Weiss, R. 2013. Influencer marketing. How word-of-mouth marketing can strengthen your organization's brand. *Marketing health services,* 34**,** 16-17.

Wellman, B. 1983. Network Analysis: Some Basic Principles. *In:* Collins, R. (ed.) *Sociological Theory.* San Francisco, CA: Jossey-Bass.

Weng, L., Flammini, A., Vespignani, A. & Menczer, F. 2012. Competition among memes in a world with limited attention. *Sci. Rep.,* 2.

Westerberg, L. & Wennergren, U. 2003. Predicting the Spatial Distribution of a Population in a Heterogeneous Landscape. *Ecological Modelling,* 166**,** 53-65.

Whitman, J. 2006. Lessons Learned: New Dot-Coms Seek Attention Cheaply. *Wall Street Journal*.

Wiertz, C., Mathwick, C., De Ruyter, K. & Dellaert, B. 2010. A Balancing Act: Exploring Governance in a Virtual P3 Community. *Advances in Consumer Research,* 37**,** 672 - 673.

Williams, S. A., Terras, M. M. & Warwick, C. 2013. What do People Study When They Study Twitter? Classifying Twitter Related Academic Papers. *Journal of Documentation,* 69**,** 384-410.

Wilson, C., Sala, A., Puttaswamy, K. P. N. & Zhao, B. Y. 2012. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web (TWEB),* 6**,** 17.

Wrong, D. H. 1961. The oversocialized conception of man in modern sociology. *American sociological review***,** 183-193.

Xu, C., Dale, C. & Jiangchuan, L. 2008. Statistics and Social Network of YouTube Videos. . I*WQoS 2008. 16th International Workshop on Quality of Service*, 2008, 2-4 June 2008. 229-238.

Yin, R. K. 1984. Case Study Research: Design and Methods (1st ed.). Beverly Hills, CA, Sage Publishing.

Yin, R. K. 1994. Case Study Research: Design and Methods (2nd ed.). Beverly Hills, CA, Sage Publishing.

Zhao, D. & Rosson, M. 2009. How and Why People Twitter: The Role that Micro-Blogging Plays in Informal Communication at Work. *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, 2009 Sanibel Island, Florida, USA. ACM, 243-252.

Appendix A: Case Reports

A.1 Case 1--Music

A.1.1 Case Overview

Data for keyword “YouTube + music” was collected over a period of 91 days (31/12/2013 to 31/03/2014). As shown in table 10, overall 3,097,847 tweets were collected, out of which 713,824 were broadcast tweets and 2,384,023 were engaged tweets respectively. Out of 2,384,023 engaged tweets only 1,586,149 tweets formed the largest community. Similarly, 2,586,586 daily unique people tweeted overall, out of which 898,282 daily unique people were engaged in broadcast activity, whereas 1,688,304 daily unique people were engaged in conversations. Out of 1,688,304 daily unique people only 1,456,770 daily unique people formed the largest community. Data for the largest community was analyzed at a daily interval. The overall trends for the music data are shown below in figure 1 and figure 2.

Figure 1: Overall Tweets

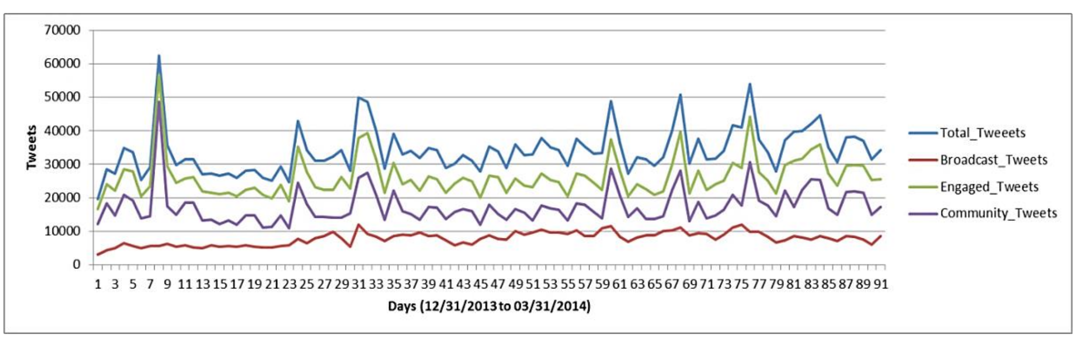


Figure 2: Overall People

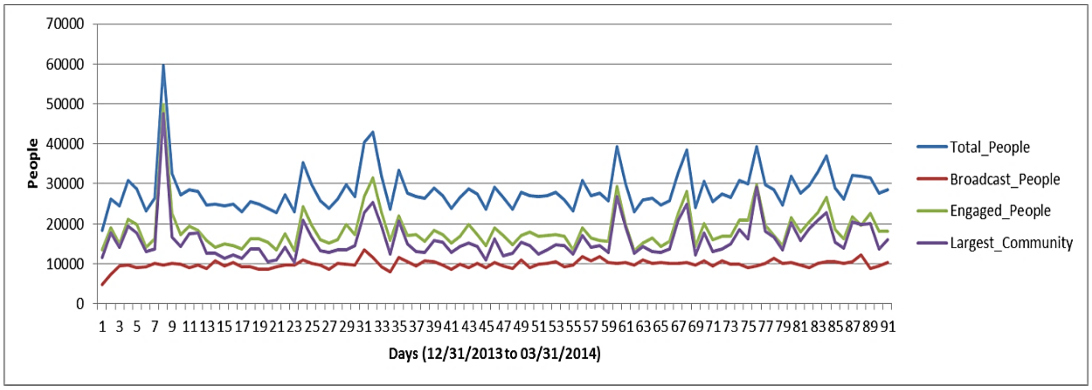


Figure 1 and figure 2 show that both the total tweets and total people involved are very dynamic, and their magnitude changes on a daily basis. The maximum of the total number of daily tweets and the maximum of the total number of daily unique people observed on a single day (the daily uniques) are 62,380 and 59,666, respectively. Similarly, the minimum of the total number of daily tweets and the minimum of the number daily unique are 19,700 and 18,333, respectively. The size of the largest community on a particular day and the largest number of community tweets on that day also seem to follow the trend of total people and total tweets. The largest number of daily community tweets and the largest number of daily unique people are 48,720 and 47,630, respectively. Similarly, the smallest number of daily community tweets and the smallest number of daily unique people are 10,830 and 10,324, respectively. As the total number of daily unique people tweeting increases, so does the size of the community. Most of the engaged people are engaged in the collective conversation forming the largest community.

A.1.2 Random or Not Random

As explained in section 4.4.1, in order to eliminate α-errors and β-errors, I compare the Clustering Coefficients of both undirected and directed networks with their corresponding random (Erdös-Rényi, E-R) networks. If the Clustering Coefficients of the undirected and directed (CC_ud, CC_d) music networks are equal to those of the E-R random network (CCudran, CCdran), then the directed and undirected networks are considered to be random, if they are not equal, then they are not random.

Figure 3: Comparison of Clustering Coefficients of Undirected Music Network with E-R Networks

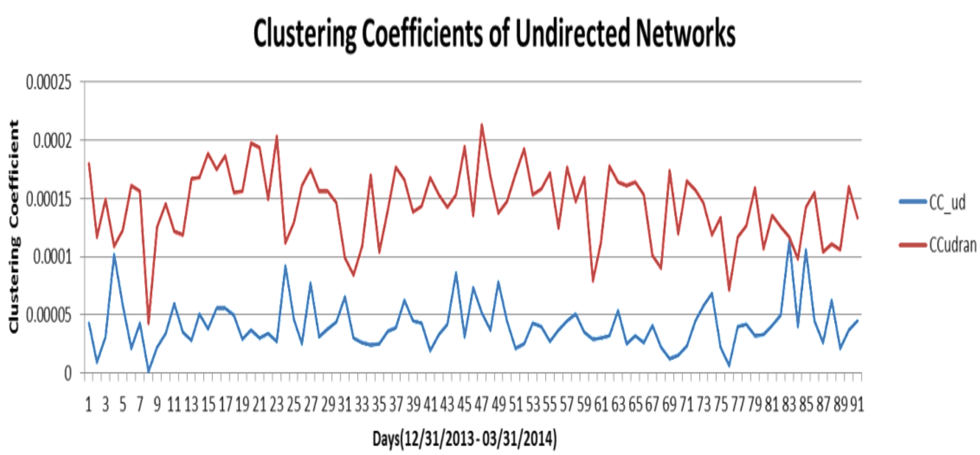
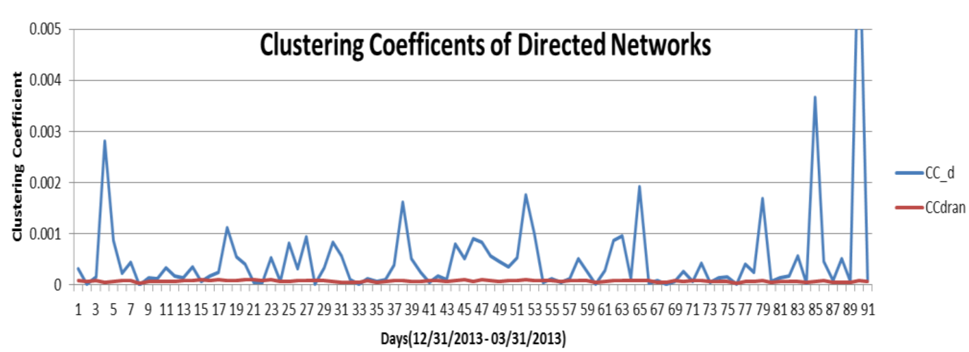


Figure 4: Comparison of Clustering Coefficients of Directed Music Network with E-R Networks



As seen in figure 3 and figure 4 Clustering Coefficients of both, directed and undirected networks (CC_ud, CC_d) follow very different pattern from their corresponding E-R networks (CCudran, CCdran). Therefore, both these networks are considered to be non-random networks, and the variables computed are a true reflection of network's features.

A.1.3 Independent Variables

The values of the independent variables for both the undirected and the directed music network are shown in figure 5 below.

Figure 5: Independent Variables--(a) Nodes and Edges (Undirected and Directed networks), (b) Reciprocity (Directed Networks), (c) Density (Undirected and Directed Networks), (d) Clustering Coefficient Undirected Network, (e) Clustering Coefficient Directed Network.

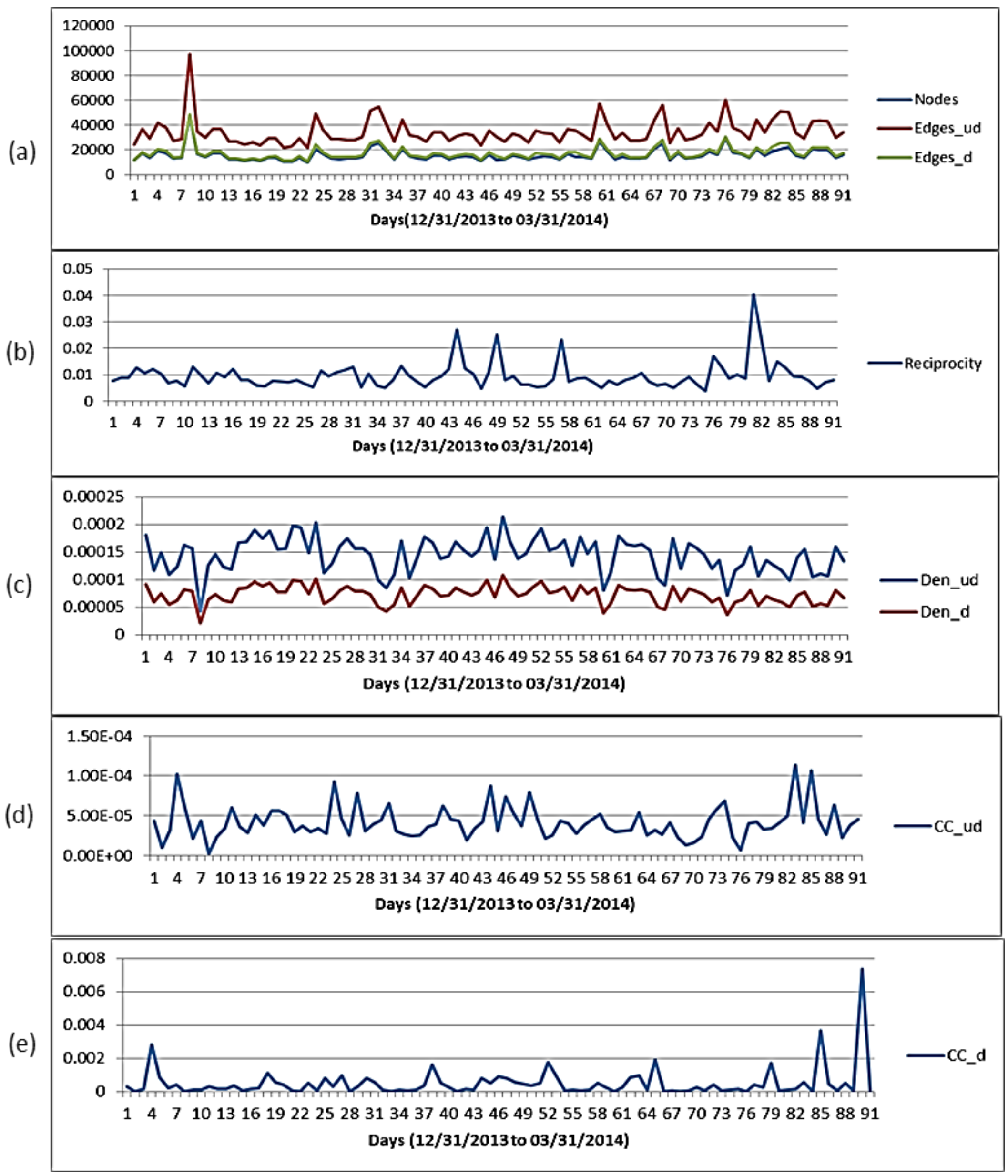
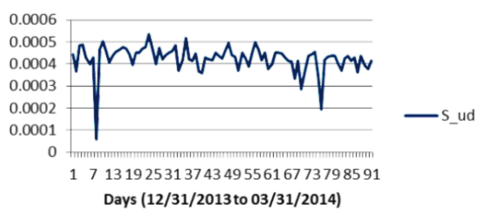


Figure 5 (a) shows that the number of directed ties (Edges_d) in the network and the total number of nodes (Nodes) overlap with each other. The numbers of undirected ties (Edges_ud) is greater than the number of directed ties (Edges_d), because in an undirected network every directed tie is considered to be symmetric. Therefore it is counted twice, except for the ones that are symmetric in a directed network. Reciprocity in figure 5 (b) indicates the presence of symmetric ties in a directed network (in an undirected network 100% are symmetric). The value of 0.01 is equal to 1% of all the ties. Figure 5(c) shows the difference between the densities of the undirected (Den_ud) and the directed (Den_d) music networks. The undirected network is denser than the directed network (Den_ud > Den_d). Figure 5 (d) and figure 5 (e) show that the directed networks have higher Clustering Coefficients than the undirected networks (CC_d > CC_ud).

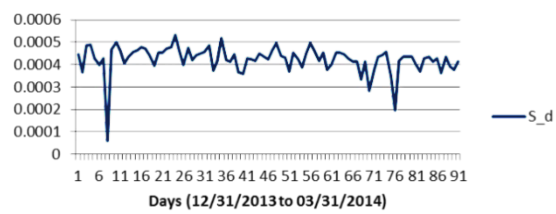
A.1.4 Network Structure Variables (MV1)

A.1.4.1 The Scale Free Metric

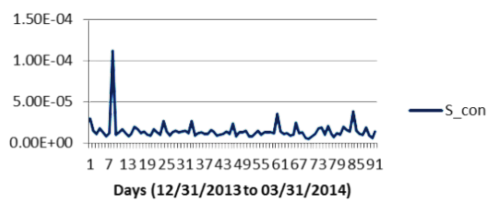
Figure 6: Scale Free Metric--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



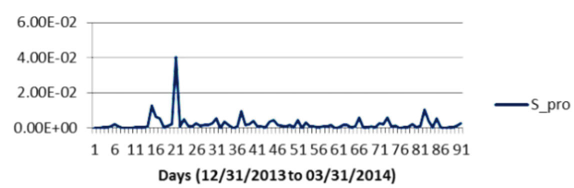
(a)



(b)



(c)



(d)

Figure 6 shows the Scale Free Metric for the undirected, directed, consumption and propagation networks (S_ud, S_d, S_con, S_pro). The Scale Free Metrics for the undirected (S_ud) and the directed network (S_d) are similar, but the Scale Free Metrics for the consumption (S_con) and propagation (S_pro) networks are very different. The propagation network is more Scale Free than the consumption network (S_pro > S_con). The values of the Scale Free Metric ranges between 0 and 1. When the values are closer to 1, it means that the networks are more Scale Free. None of the networks are Scale Free in nature. This means that these networks have hubs in them. However, there is not just one hub that is the center of the community.

A.1.4.2 The Assortativity

Figure 7: Music Assortativity--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

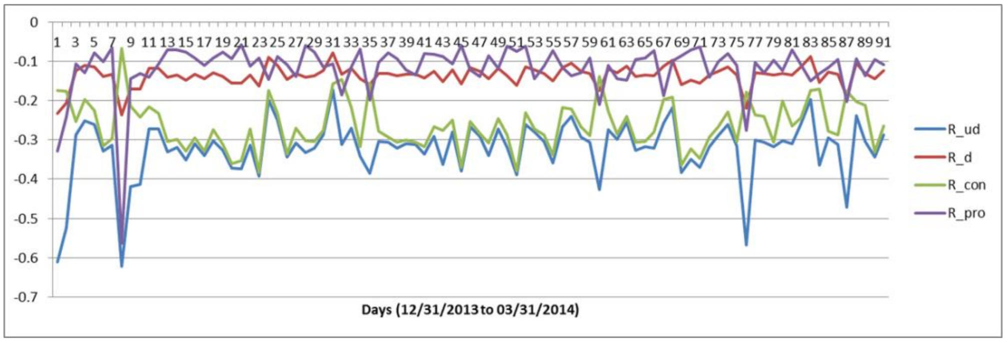


Figure 7 shows the Assortativity for the undirected, directed, consumption and propagation networks of music conversations (R_ud, R_d, R_con, R_Pro). The values of the Assortativity ranges between -1 and 1. When the values are closer to -1, it means that networks are Disassortative. The undirected network is more Disassortative than the directed network (R_d > R_ud). Among the directed networks, the consumption network is more Disassortative than the propagation network (R_pro > R_con). Disassortative means that the nodes in the network connect to nodes that are very similar to themselves. This is true more so in the undirected network and in the consumption network than it is in the directed network and the propagation network. This implies that disassortativeness of consumption contributes more to the disassortativeness of the directed network than the disassortativeness of the propagation does.

A.1.4.3 The Small World Metric

Figure 8: Small World Metric --(a) Undirected Network, (b) Directed Network.

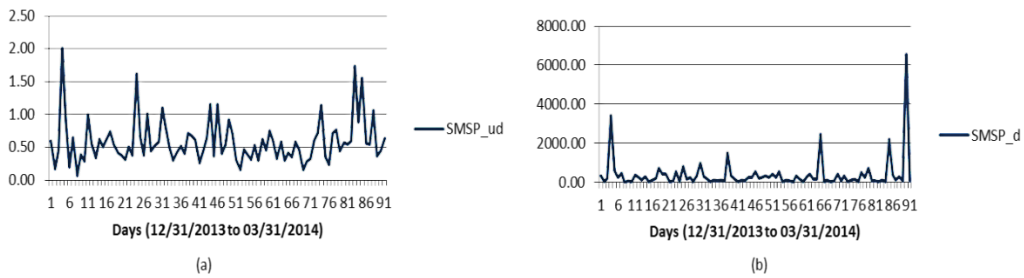


Figure 8 shows the Small World Metric for the undirected (SMSP_ud) and directed networks (SMSP_d). The Small World Metrics for the consumption and propagation networks are the same as the ones for the directed network. The directed networks show stronger Small World behavior than the undirected networks (SMSP_d > SMSP_ud). This means that in directed networks there are more nodes that act as hubs that facilitate communication between other nodes of the network.

A.1.4.4 Paths and Shortest Paths Power law Distribution per Node

Figure 9: Power Law Distribution of Paths and Shortest Paths in (a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

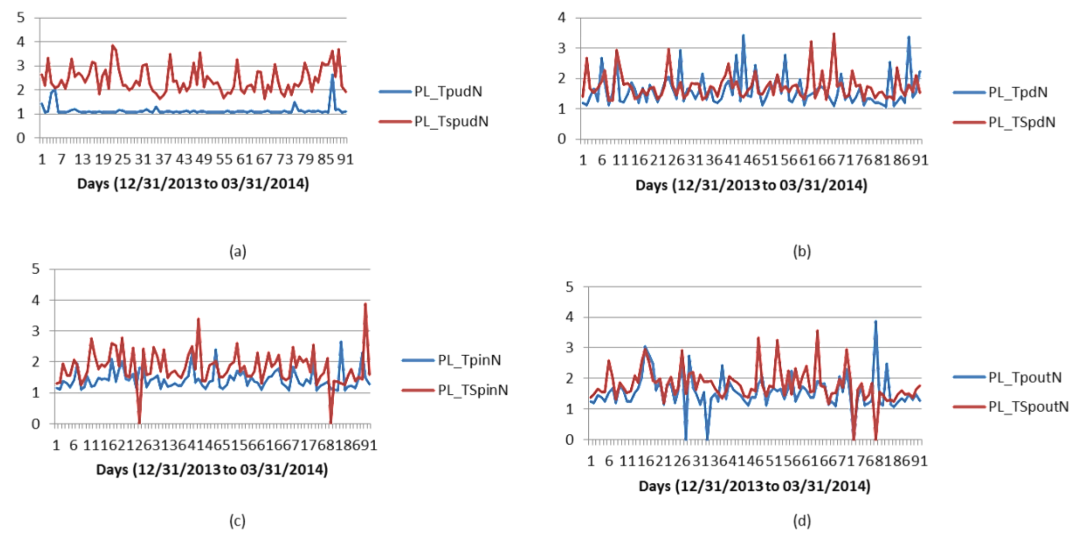


Figure 9 shows that, in the undirected network, paths are more uniformly distributed among nodes (PL_TpudN) than shortest paths are distributed among nodes (PL_TSpudN). This means that fewer nodes are responsible for more of the shortest paths in the undirected network. A similar, albeit less pronounced, trend for the consumption network is seen in figure 9 (c). In the directed and propagation networks, there are no such patterns.

A.1.5 Network Flow Variables (MV2)

Figure 10: Network Flow Variables-- (a) Total Paths and Total Shortest Paths, (b) Average Paths and Average Shortest Paths, (c) Undirected and Directed Network Graph Diameter.

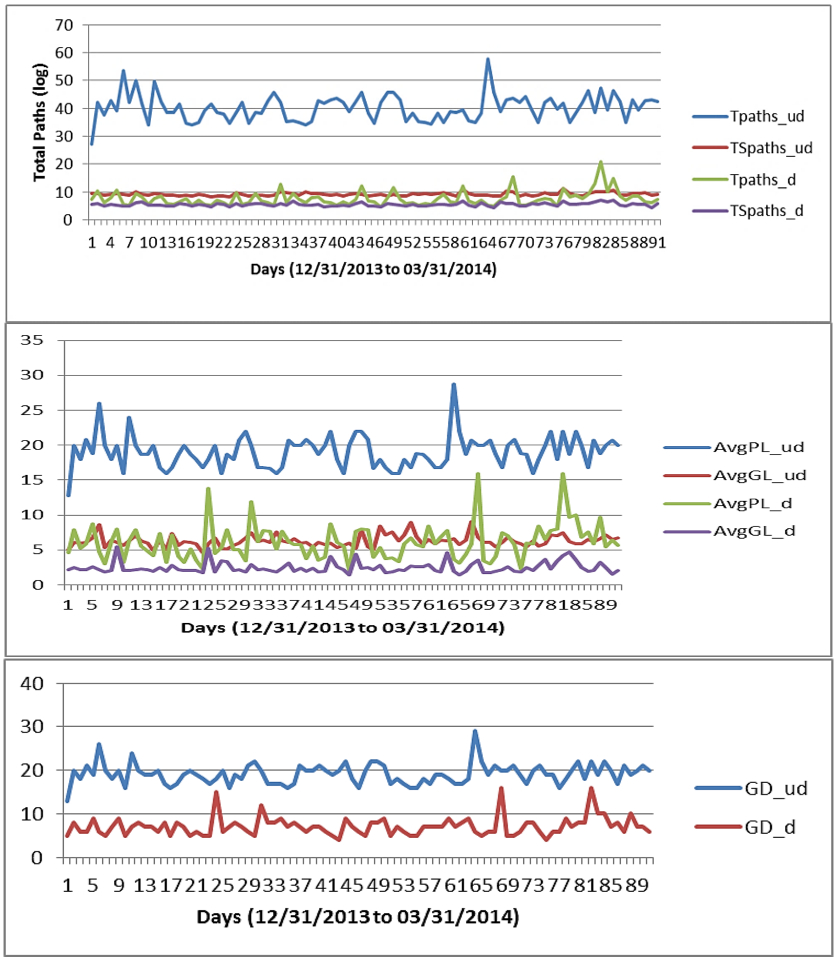


Figure 10 (a), shows that Total Number of Paths in the undirected network (Tpaths_ud) is orders of magnitude higher than the Total Number of Shortest Paths (TSpaths_ud). The Total Number of Paths (Tpaths_d) and the Total Number of Shortest Paths (TSpaths_d) map more closely in the directed network. In figure 10 (b), a similar trend is observed in the Average Path Lengths and the Average Geodesic Lengths of the undirected and directed networks (AvgPL_ud, AvgPL_d, AvgGL_ud, AvgGL_d). In figure

10 (c), the Graph Diameter of the undirected network (GD_ud) is larger than the Graph Diameter of the directed network (GD_d). It is also noteworthy that, in figure 10 (b) and in figure 10 (c), the Graph Diameter and the Average Path Length of the undirected and directed networks (GD_ud, AvgPL_ud, GD_d, AvgPL_d) track pretty closely.

A.1.6 Dependent Variables

A.1.6.1 Eigenvector Centralization

Figure 11: Eigenvector Centralization in the Undirected, Directed, Consumption and Propagation Networks

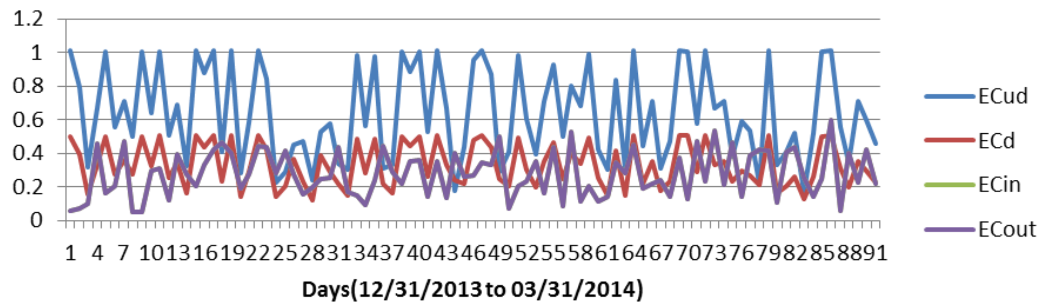


Figure 11 shows that nodes with influence are lot more central in the undirected network than in the directed, consumption and propagation networks (ECud > (ECd, ECin, ECout)). The consumption and propagation networks exhibit same level of centralization.

A.1.6.2 Power law Distribution of Eigenvector Centrality per Node

Figure 12: Power Law Distribution of Eigenvector Centrality in Undirected, Directed, Consumption and Propagation Network

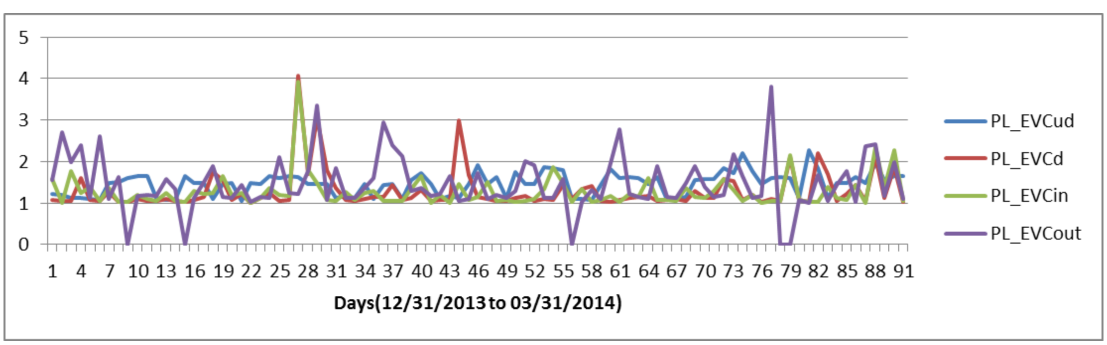
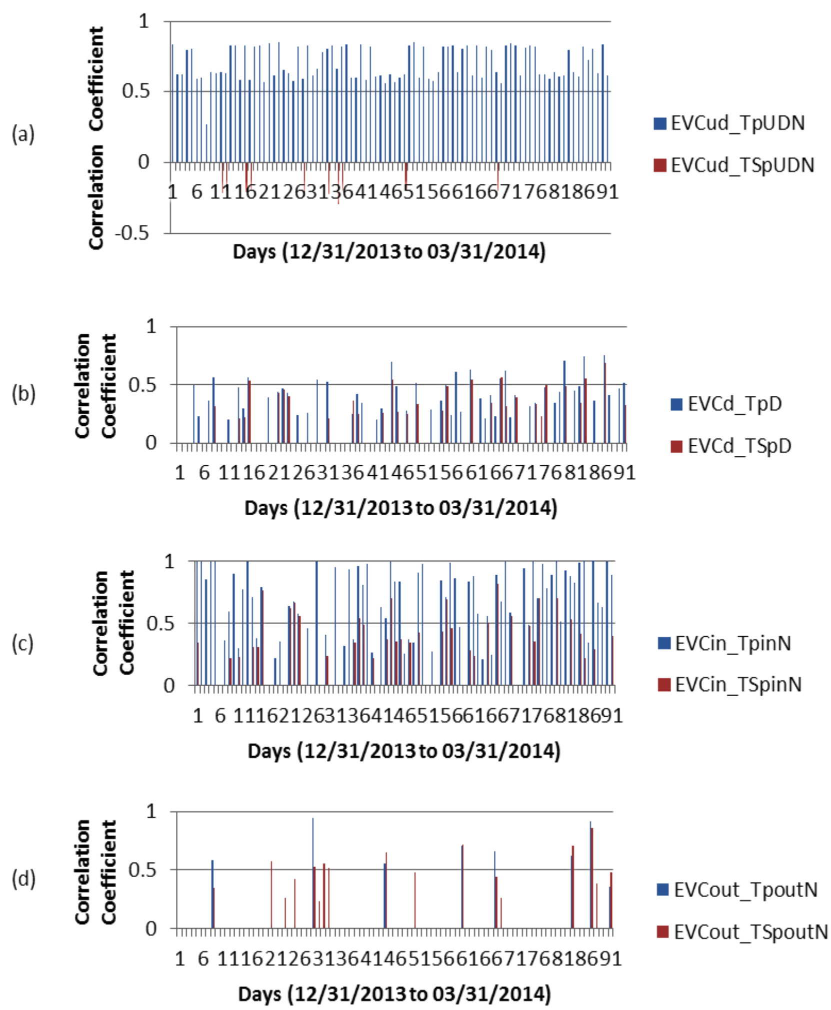


Figure 12 shows that in undirected, directed, consumption and propagation network the distribution of Eigenvector Centrality amongst nodes have similar Power Law patterns.

A.1.6.3 Correlation Coefficients of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node

Figure 13: Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



In figure 13, only those correlation coefficients with a significance value lower than 0.05 are shown. In figure 13 (a), there is a significant correlation between the Eigenvector Centrality of a node and the number of paths from a node in undirected network (EVCud_TpUDN). There is no significant correlation between Eigenvector Centrality of a node and shortest paths from a node in undirected network (EVCud_TSpUDN). Similarly, in figure 13 (c), there is a significant correlation between the in-Eigenvector Centrality of a node and the number of paths ending on a node in the consumption network (EVCin_TpinN). The correlation between the in-Eigenvector Centrality of a node and the number of shortest paths is less significant (EVCin_TSpinN). In figure 13(b) and figure 13 (d) the directed-Eigenvector Centrality and the out-Eigenvector Centrality have no significant correlation with either the number of paths or the number of shortest paths.

A.1.7 Statistical Analysis

A.1.7.1 The Undirected Network

A.1.7.1.1 Correlation Analysis

In Table 1, the statistically significant Correlation Coefficients for the undirected network are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 1: Correlation Coefficients of Undirected Network

		Nodes	Edges_ud	Den_ud	CC_ud	GD_ud	Tpaths_ud
Nodes	Pearson Correlation	1					
	Sig. (2-tailed)						
	N	91					
Edges_ud	Pearson Correlation	.989**	1				
	Sig. (2-tailed)	.000					
	N	91	91				
Den_ud	Pearson Correlation	-.888**	-.870**	1			
	Sig. (2-tailed)	.000	.000				
	N	91	91	91			
Tpaths_ud	Pearson Correlation	.255*	.245*	-.246*	-.003	.949**	1
	Sig. (2-tailed)	.015	.019	.019	.976	.000	
	N	91	91	91	91	91	91
TSpaths_u	Pearson Correlation	.727**	.767**	-.739**	.014	-.005	.112
	Sig. (2-tailed)	.000	.000	.000	.897	.964	.290
	N	91	91	91	91	91	91
AvgPL_ud	Pearson Correlation	.022	.021	-.094	.066	.999**	.955**
	Sig. (2-tailed)	.837	.843	.378	.537	.000	.000
	N	91	91	91	91	91	91
S_ud	Pearson Correlation	-.612**	-.569**	.445**	.419**	-.055	-.269**
	Sig. (2-tailed)	.000	.000	.000	.000	.603	.010
	N	91	91	91	91	91	91
R_ud	Pearson Correlation	-.221*	-.134	.049	.542**	.169	.048
	Sig. (2-tailed)	.035	.205	.644	.000	.109	.653
	N	91	91	91	91	91	91
SMSP_ud	Pearson Correlation	.147	.195	-.287**	.912**	.080	.041
	Sig. (2-tailed)	.165	.064	.006	.000	.449	.698
	N	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

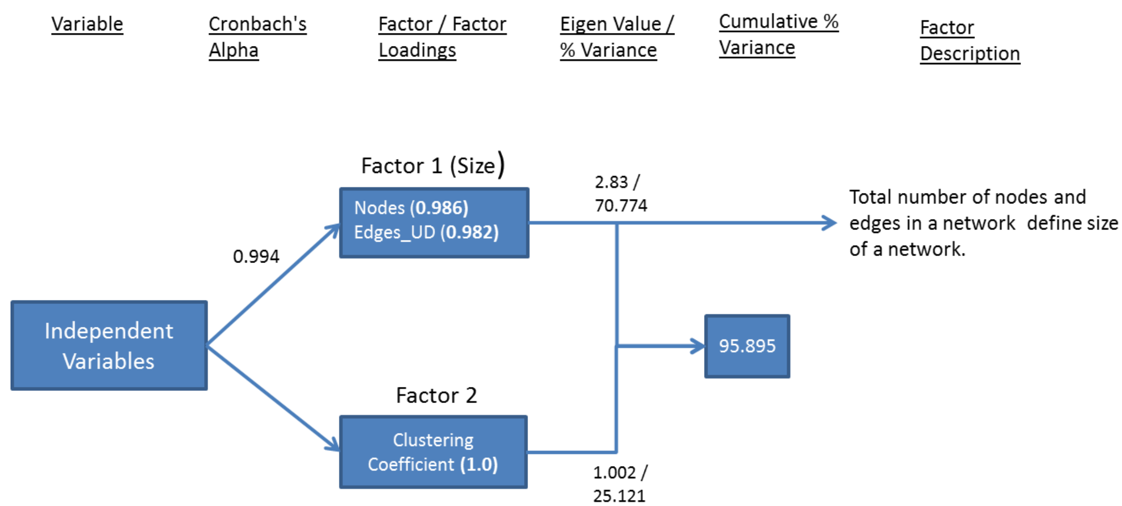
In Table 1, the number of nodes (Nodes) and the number of ties (Edges_ud) have a strong positive correlation. As the number of nodes (Nodes) increases, the number of ties (Edges_ud) also increases. The Density (Den_ud) of this network has a strong negative correlation with both the number of nodes (Nodes) and the number of ties (Edges_ud). As the number of nodes (Nodes) or number of ties increases (Edges_ud), Density (Den_ud) decreases. This is because Density is a measure of the total number of ties that exists in the network vs. the number of all possible ties. As the number of nodes increases (Nodes), the total number of possible ties also increases, pushing down Density (Den_ud). The Total Number of Paths (Tpaths_ud) in the network, the Average Path Length (AvgPL_ud) and Graph Diameter (GD_ud) correlate strongly with each other in this network. The Total Number of Shortest Paths (TSpaths_ud) correlates strongly with the number of nodes (Nodes) and the number of ties (Edges_ud), but it correlates negatively with Density (Den_ud). The possible number of shortest paths increases as the number of nodes (Nodes) and the number of ties (Edges_ud) increases. Since Density (Den_ud) shares a negative relationship with the possible increase in the number of nodes (Nodes) and ties (Edges_ud) (explained above), it also shares a negative relationship with the Total Number of Shortest Paths (Tpaths_ud). The Scale Free (S_ud) metric seems to share a negative relationship with the number of nodes (Nodes) and the number of ties (Edges_ud). Assortativity (R_ud) and the Small World (SMSP_ud) metric share a positive relationship with the Clustering Coefficients (CC_ud).

A.1.7.1.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.1.7.1.2.1 Independent Variables

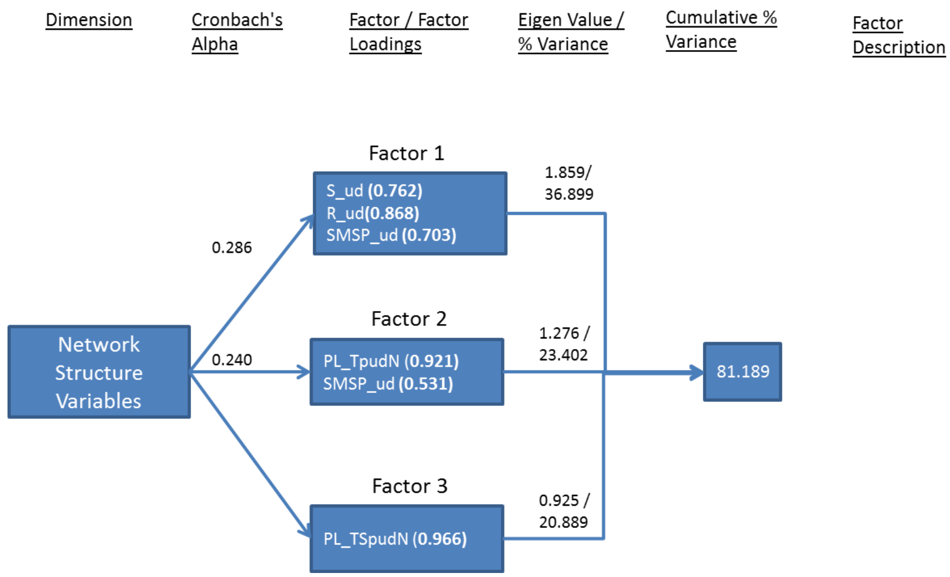
Figure 14: Factor Analysis Independent Variables Music Undirected Network



The factor analysis generated two factors that explain 95% (greater than 80%) of the cumulative variance. Both factors have eigenvalues above one. Nodes and ties (Edges_ud) have significant factor loadings in factor 1. Density (Den_ud) had a negative loading in factor 1, hence it was removed. Only the Clustering Coefficient (CC_ud) has a significant loading in factor 2.Cronbach’s alpha for factor 1 has a value of 0.994. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as “Size”.

A.1.7.1.2.2 Network Structure (MV1)

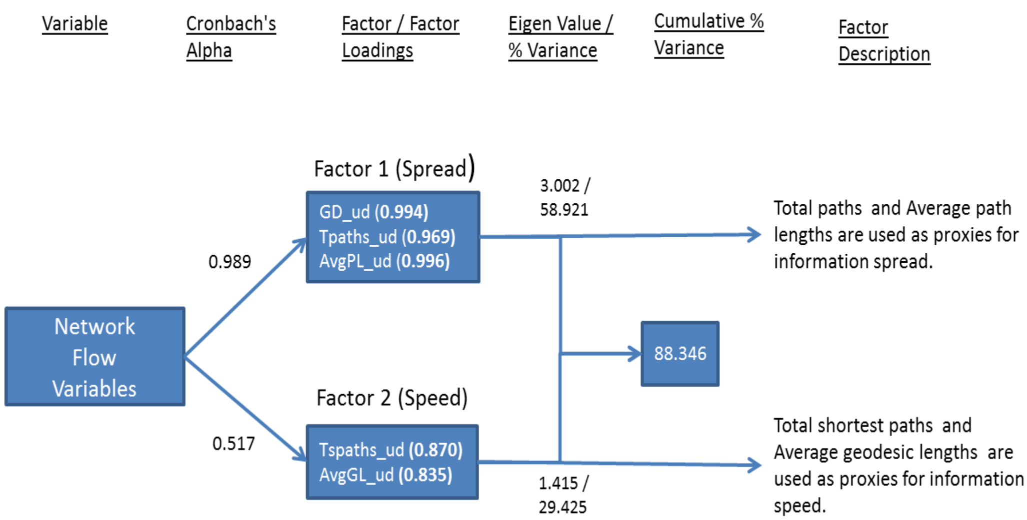
Figure 15: Factor Analysis of Network Structure Variables



The factor analysis generated three factors that explain 81.189% (greater than 80%) of the cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor 3 has eigenvalues below 1. Scale Free Metric (S_ud), Assortativity (R_ud) and Small World Metric (SMSP_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.286. Scale Free Metric (S_ud), Assortativity (R_ud) and Small World Metric (SMSP_ud) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. Power Law Distribution of Total Paths per Node (PL_TpudN) and Small World Metric (SMSP_ud) have significant factor loadings in factor 2. Cronbach's alpha for facto1 has a value of 0.246. PL_TpudN and SMSP_ud are measuring different constructs within factor 2. Hence, they should not be considered as a factor. All other variables load independently.

A.1.7.1.2.3 Network Flow (MV2)

Figure 16: Factor Analysis of Network Flow Variables



Graph Diameter (GD_ud), Total Paths (Tpaths_ud) and Average Path Length (AvgPL_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.989. Graph Diameter (GD_ud), Total Paths (Tpaths_ud) and Average Path Length (AvgPL_ud) are measuring the same construct within factor 1. Factor 1 is named as "Spread", as the AvgPL_ud and Tpaths_ud are being used as proxies for information spread.

Total Shortest Paths (TSpaths_ud) and Average Geodesic Length (AvgGL_ud) have significant factor loadings on factor 2. Cronbach's alpha for factor 2 has a value of 0.517, which indicates poor internal consistency. Therefore, Total Shortest Paths

(TSpaths_ud) and Average Geodesic Length (AvgGL_ud) maybe measuring different constructs in factor 2. Hence, they should not be considered as a factor. But if they had better internal consistency, I would name factor 2 as "Speed", since Total Shortest Paths (TSpaths_ud) and Average Geodesic Length (AvgGL_ud) are being used as proxies for information speed.

A.1.7.1.2.4 Dependent Variables

The value of Kaiser-Meyer-Olkin measure of sampling adequacy was 0.429 (less than 0.5), and the significance Bartlett's test of sphericity is 0.346. This data does not satisfy the measure of appropriateness for factor analysis. Therefore, all the variables are considered independently.

A.1.7.1.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Music.pdf".

A.1.7.1.3.1 Impact of Network Structure on Network Flow

Table 2: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud
Music	NA	(0.093/0.005)[3,4]	(0.254/0.000)[3,5,2]	NA	(0.287/0.000)[1,2,4,5]

Table 2 shows that the network structure variables have a significant impact on the network flow variables. Network structure variables explain 9.3%, 25.4% and 28.7% variation in Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Tpaths_ud is not taken into consideration, as the p-value of 0.005 is greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.1.3.2 Impact of Network Flow on Network Structure

Table 3: Impact of Network Flow on Network Structure

Predictors: **(6) GD_ud, (7) Tpaths_ud, (8) TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpudN	PL_TSpudN	S_ud	R_ud	SMSP_ud
Music	(0.040/0.032)[7]	(0.0105/0.001)[10]	NA	(0.167/0.000)[10,8]	NA

Table 3 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 4%, 3.2%, and 16.7% variation in the PL_TpudN, PL_TSpudN and R_ud, respectively. The impact of network flow variables on PL_TpudN and PL_TSpudN is not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.1.3.3 Impact of Network Structure on Network Phenomenon

Table 4: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Music	(0.079/0.010)[2,5]	NA	NA	NA

Table 4 shows that the network structure variable impacts only Eigenvector Centralization (ECud), explaining only 7.9% variation. The impact of network structure variables on (ECud) is not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.1.3.4 Impact of Network Flow on Network Phenomenon

Table 5: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_ud, (7)Tpaths_ud, (8)TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Music	(0.062/0.010)[8]	(0.045/0.024)[10]	(0.045/0.025)[7]	NA

Table 5 shows that the network flow variable impacts Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 6.2%, 2.4% and 2.5% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.1.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 6: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud, (6),GD_ud (7) Tpaths_ud (8), TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud, (11) Nodes, (12) Edges_ud, (13) Den_ud, (14) CC_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Music	(0.157/0.000) [12,4,2]	(0.045/0.024)[10]	(0.046/0.024)[7]	NA

Table 6 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 15.7%, 4.5% and 4.6% variation respectively. The collective impact of independent variables and the moderating variables on Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.2 The Directed Network

A.1.7.2.1 Correlation Analysis

Significant Correlations Coefficients for directed network are shown below in table 7. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 7: Correlation coefficients of directed network

Correlations										
		Nodes	Edges_d	Reciprocity	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	EVCd_TpD
Nodes	Pearson Corr	1								
	Sig. (2-tailed)									
	N	91								
Edges_d	Pearson Corr	.988**	1							
	Sig. (2-tailed)	.000								
	N	91	91							
Den_d	Pearson Corr	-.888**	-.870**	-.075						
	Sig. (2-tailed)	.000	.000	.478						
	N	91	91	91						
Tpaths_d	Pearson Corr	.461**	.508**	.605**	-.112	.764**	1			
	Sig. (2-tailed)	.000	.000	.000	.292	.000				
	N	91	91	91	91	91	91			
TSpaths_d	Pearson Corr	.466**	.488**	.415**	-.294**	.515**	.752**	1		
	Sig. (2-tailed)	.000	.000	.000	.005	.000	.000			
	N	91	91	91	91	91	91	91		
AvgPL_d	Pearson Corr	.378**	.438**	.352**	-.018	.960**	.847**	.591**	1	
	Sig. (2-tailed)	.000	.000	.001	.865	.000	.000	.000		
	N	91	91	91	91	91	91	91	91	
AvgGL_d	Pearson Corr	.159	.223*	.305**	-.088	.683**	.560**	.626**	.686**	
	Sig. (2-tailed)	.132	.034	.003	.406	.000	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	
S_d	Pearson Corr	-.612**	-.569**	-.022	.044	.020	-.205	-.211*	-.013	
	Sig. (2-tailed)	.000	.000	.839	.675	.854	.051	.045	.906	
	N	91	91	91	91	91	91	91	91	
SMSP_d	Pearson Corr	-.098	-.107	-.018	.945**	-.113	-.142	-.339**	-.073	
	Sig. (2-tailed)	.354	.313	.862	.000	.286	.181	.001	.489	
	N	91	91	91	91	91	91	91	91	
EVCd_TSpD	Pearson Corr	.049	.054	.042	-.159	-.070	.111	.237*	-.061	.717**
	Sig. (2-tailed)	.643	.612	.695	.131	.511	.293	.024	.568	.000
	N	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

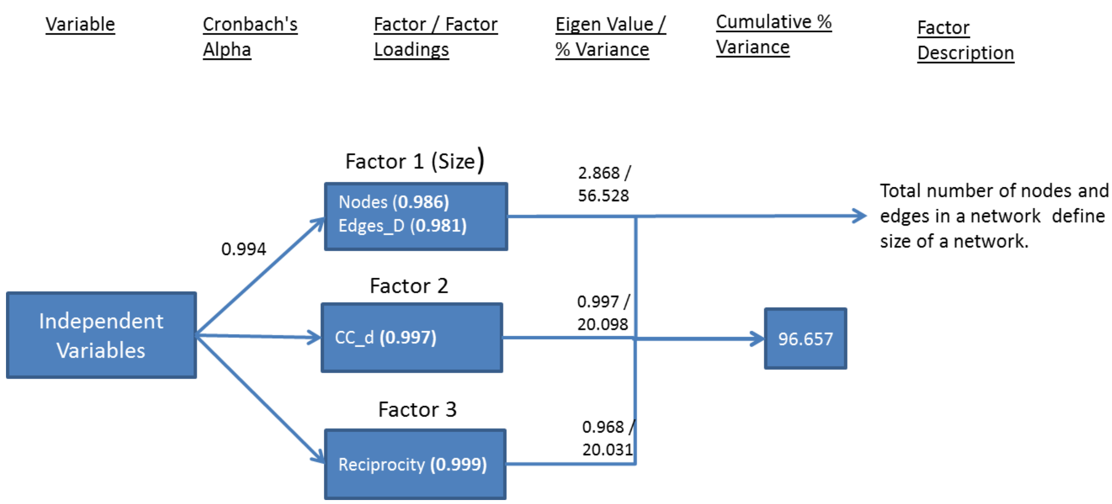
Table 7 shows that nodes (Nodes) and ties (Edges_d) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Shortest Paths (TSpaths_d) correlates strongly with Graph Diameter (GD_d) and Total Paths (Tpaths_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Scale Free Metric (S_d) seems to share a negative relationship with number of nodes (Nodes) and number of ties (Edges_d). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) and Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) correlate strongly with each other.

A.1.7.2.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.1.7.2.2.1 Independent Variables

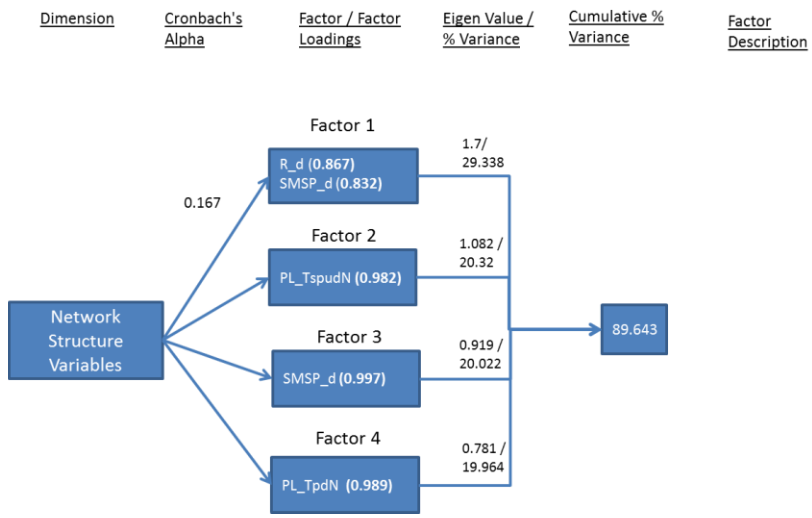
Figure 17: Factor Analysis of Independent Variables



Factor analysis generated three factors that explain 96.65% (greater than 80%) of cumulative variance. Factor 1 has eigenvalue over 1. Factor2 and factor3 have eigenvalues little less than 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had negative loading of -0.941 in factor 1, hence it was removed. Only Clustering Coefficient (CC_d) and Reciprocity have significant loadings in factor 2 and factor 3.Cronbach's alpha for factor 1 has a value of 0.994. This means Nodes and ties (Edges_d) are measuring same construct within factor 1. Hence, I name factor 1 as "Size".

A.1.7.2.2.2 Network Structure (MV1)

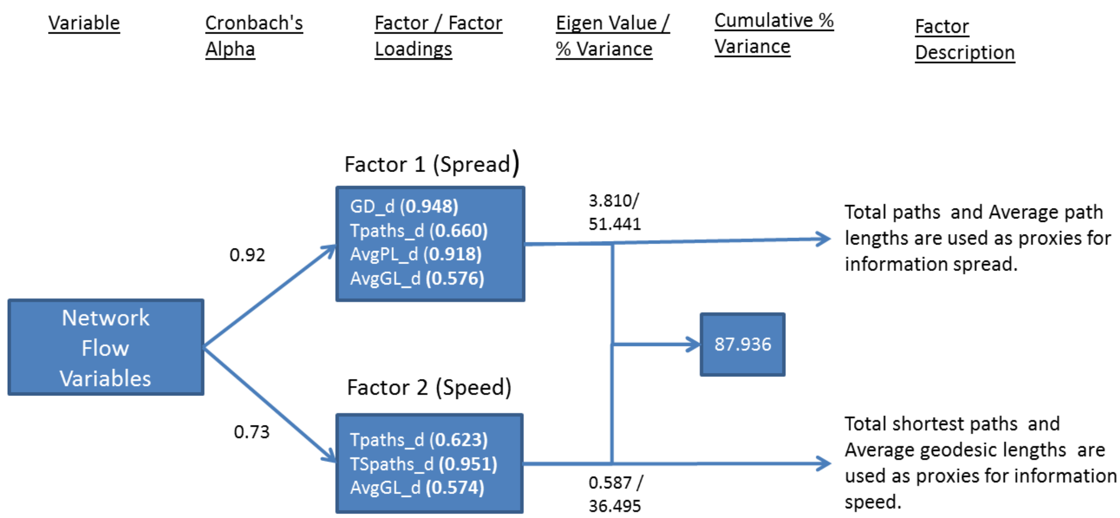
Figure 18: Factor Analysis of Network Structure Variables



Factor analysis generated four factors that explain 89.64% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor 2 and factor 3 have eigenvalues below 1. Assortativity (R_d) and Small World Metric (SMSP_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.167. Assortativity (R_d) and Small World Metric (SMSP_d) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.1.7.2.2.3 Network Flow (MV2)

Figure 19: Factor Analysis of Network Flow Variables



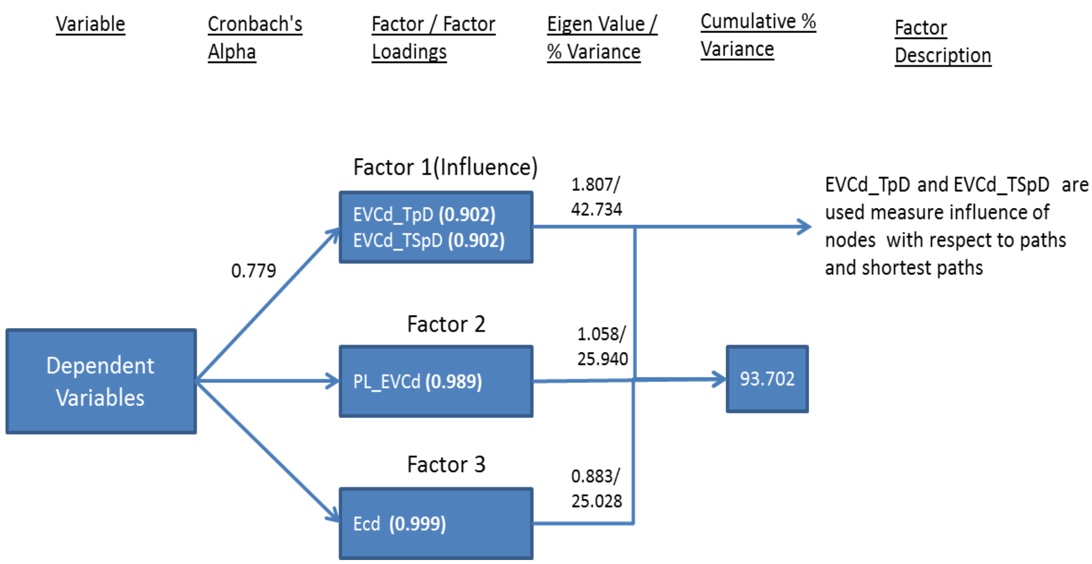
Factor analysis generated two factors that explain 87.936% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.92. Factor 1 is named as "Spread" as Total Paths (Tpaths_d), Average Path Length (AvgPL_d) are being used as proxies for information spread.

Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings on factor2. Cronbach's alpha for factor2 has a value of 0.73. Factor 2 is named as "Speed" since Total Shortest Paths

(TSpaths_d) and Average Geodesic Length (AvgGL_d) are being used as proxies for information speed.

A.1.7.2.2.4 Dependent Variables

Figure 20: Factor Analysis of Dependent Variables



The value of Kaiser-Meyer-Olkin measure of sampling adequacy was 0.474(less than 0.5) but the significance Bartlett's test of sphericity is 0. Factor analysis generated three factors that explain 93.702% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCd_TpD) and Shortest Paths (EVCd_TSpD) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.779. I name factor 1 as "Influence".

A.1.7.2.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Music.pdf".

A.1.7.2.3.1 Impact of Network Structure on Network Flow

Table 8: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Music	(0.127/0.000)[4]	(0.052/0.017)[1]	(0.227/0.000) [1,5]	(0.156/0.000) [1,4]	(0.079/0.010) [4,5]

Table 8 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 12.7%, 5.2%, 22.7%, 15.6% and 7.9% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Total Shortest Paths (TSpaths_d), and Average Geodesic Length (AvgGL_ud) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.2.3.2 Impact of Network Flow on Network Structure

Table 9: Impact of Network Flow on Network Structure

Predictors: **(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpdN	PL_TSpdN	S_d	R_d	SMSP_d
Music	(0.103/0.003) [8,10]	(0.163/0.000) [8,10]	(0.142/0.000) [8]	(0.226/0.000) [6,8]	(0.090/0.002) [8]

Table 9 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 10.3%, 16.3%, 14.2%, 22.6% and 9.0% variation in the PL_TpdN, PL_TSpdN, S_d, R_ud, and SMSP_d, respectively. The impact of network flow variables on PL_TpdN and SMSP_d are not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.2.3.3 Impact of Network Structure on Network Phenomenon

Table 10: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Music	(0.040/0.032)[4]	(0.140/0.001)[1,2,5]	NA	(0.061/0.010)[2]

Table 10 shows that the network structure variable impacts Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN), explaining 4%, 14% and 6.1% variation respectively. The impact of network flow variables Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.2.3.4 Impact of Network Flow on Network Phenomenon

Table 11: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Music	(0.090/0.002)[9]	(0.051/0.018)[10]	NA	(0.098/0.004)[8,9]

Table 11 shows that the network structure variable impacts Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpdN), explaining 9%, 5.1% and 9.8% variation respectively. The impact of network flow variables Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpdN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.2.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 12: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d, (6)GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCud_TpdN	EVCud_TSpdN
Music	(0.090/0.002) [9]	(0.303/0.000) [1,7,14,13]	(0.060/0.011) [15]	(0.061/0.010)[2]

Table 12 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 9%, 30.3%, 6%and 6.1% variation respectively. The collective impact of independent variables and the moderating variables on (EC_d), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.3 The Consumption Network

A.1.7.3.1 Correlation Analysis

Table 13: Correlation coefficients of directed network

Significant correlations coefficients for consumption network are shown below in table 13. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d
Edges_d	Pearson Co	.988**	1							
	Sig. (2-tailed	.000								
	N	91	91							
Den_d	Pearson Co	-.888**	-.870**	-.075	1					
	Sig. (2-tailed	.000	.000	.478						
	N	91	91	91	91					
Tpaths_d	Pearson Co	.461**	.508**	.605**	-.489**	-.112	.764**	1		
	Sig. (2-tailed	.000	.000	.000	.000	.292	.000			
	N	91	91	91	91	91	91	91		
TSpaths_d	Pearson Co	.466**	.488**	.415**	-.474**	-.294**	.515**	.752**	1	
	Sig. (2-tailed	.000	.000	.000	.000	.005	.000	.000		
	N	91	91	91	91	91	91	91	91	
AvgPL_d	Pearson Co	.378**	.438**	.352**	-.434**	-.018	.960**	.847**	.591**	1
	Sig. (2-tailed	.000	.000	.001	.000	.865	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	91
AvgGL_d	Pearson Co	.159	.223*	.305**	-.180	-.088	.683**	.560**	.626**	.686**
	Sig. (2-tailed	.132	.034	.003	.087	.406	.000	.000	.000	.000
	N	91	91	91	91	91	91	91	91	91
S_con	Pearson Co	.734**	.699**	.075	-.467**	-.128	.132	.286**	.401**	.169
	Sig. (2-tailed	.000	.000	.478	.000	.227	.212	.006	.000	.110
	N	91	91	91	91	91	91	91	91	91
R_con	Pearson Co	.804**	.828**	.124	-.809**	-.145	.413**	.517**	.518**	.482**
	Sig. (2-tailed	.000	.000	.242	.000	.171	.000	.000	.000	.000
	N	91	91	91	91	91	91	91	91	91
SMSP_d	Pearson Co	-.098	-.107	-.018	.065	.945**	-.113	-.142	-.339**	-.073
	Sig. (2-tailed	.354	.313	.862	.544	.000	.286	.181	.001	.489
	N	91	91	91	91	91	91	91	91	91
PL_TpinN	Pearson Co	-.273**	-.241*	-.387**	.340**	.051	-.354**	-.527**	-.459**	-.491**
	Sig. (2-tailed	.009	.021	.000	.001	.632	.001	.000	.000	.000
	N	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

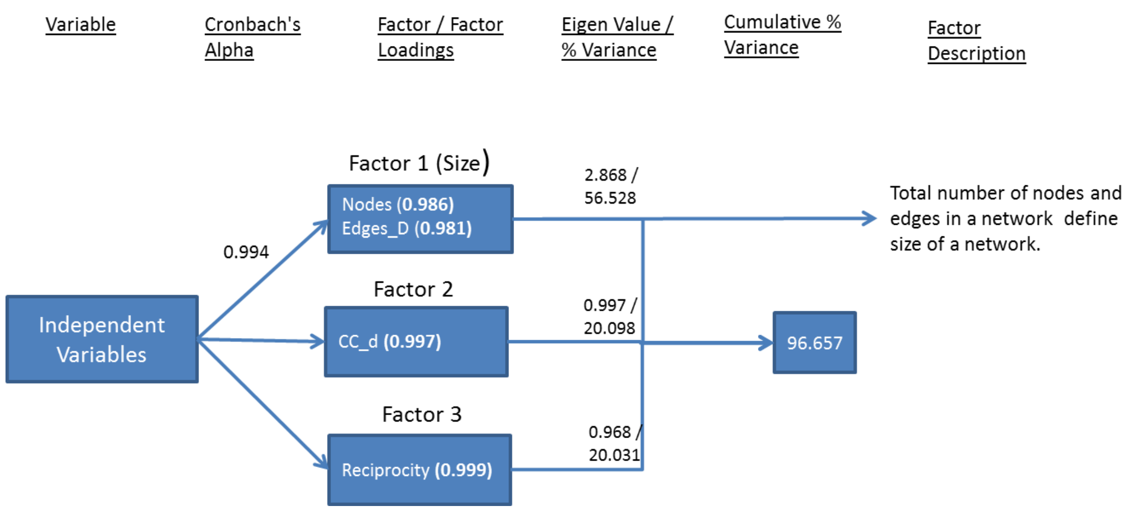
Tables 13 show that nodes (Nodes) and ties (Edges_ud) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Shortest Paths (TSpaths_d) correlates strongly with Graph Diameter (GD_d) and Total Paths (Tpaths_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Scale Free Metric (S_con) and Assortativity (R_con) seems to share a positive relationship with number nodes (Nodes) and ties (Edges_ud). Assortativity (R_con) correlates with Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) but has a negative correlation with Density (Den_d). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Power Law Distribution of Paths per Node (PL_TpinN) correlates negatively with Total Paths (Tpaths_d).

A.1.7.3.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.1.7.3.2.1 Independent Variables

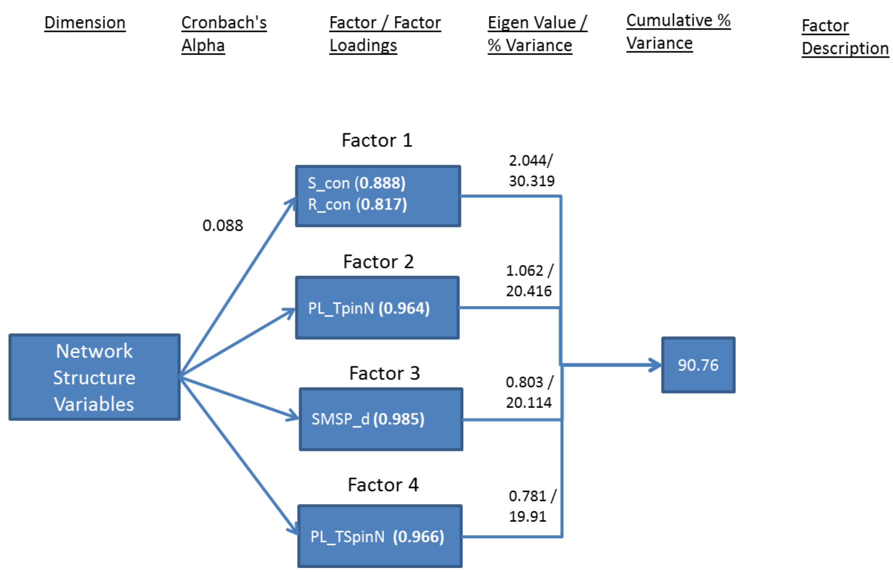
Figure 21: Factor Analysis of Independent Variables



Factor analysis generated three factors that explain 96.65% (greater than 80%) of cumulative variance. Factor 1 has eigenvalue over 1. Factor2 and factor3 have eigenvalues little less than 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had negative loading of -0.941 in factor 1, hence it was removed. Only Clustering Coefficient (CC_d) and Reciprocity have significant loadings in factor 2 and factor 3.Cronbach's alpha for factor 1 has a value of 0.994. This means nodes and ties (Edges_d) are measuring same construct within factor 1. Hence, I name factor 1 as “Size”.

A.1.7.3.2.2 Network Structure (MV1)

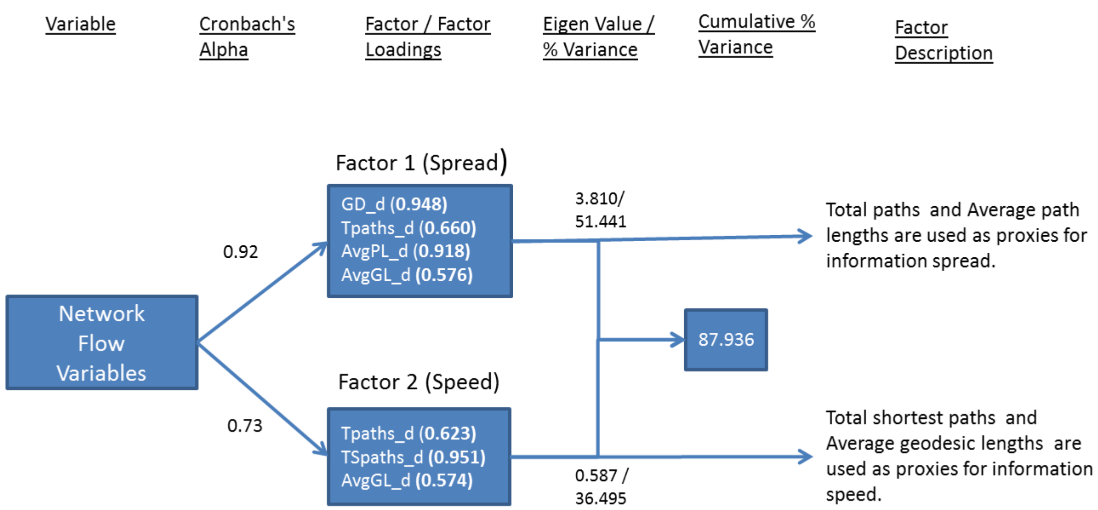
Figure 22: Factor Analysis of Network Structure Variables



Factor analysis generated four factors that explain 90.76% (greater than 80%) of cumulative variance. Factor 1 and factor 2 have eigenvalues above 1. Factor 3 and factor 4 have eigenvalues below 1. Assortativity (R_con) and Small World Metric (SMSP_d) have significant factor loadings in factor 1. Cronbach’s alpha for factor1 has a value of 0.088. Assortativity (R_con) and Small World Metric (SMSP_d) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.1.7.3.2.3 Network Flow (MV2)

Figure 23: Factor Analysis of Network Flow Variables

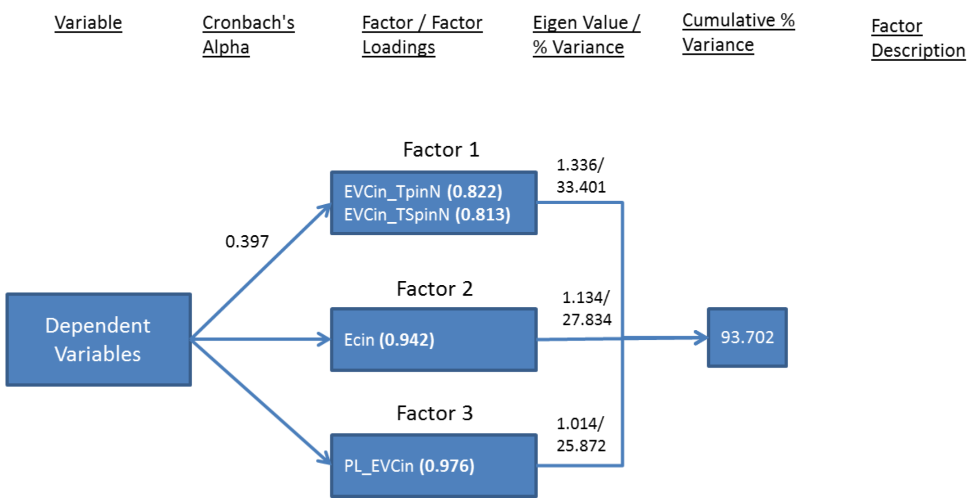


Factor analysis generated two factors that explain 87.936% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.92. Factor 1 is named as "Spread" as Total Paths (Tpaths_d), Average Path Length (AvgPL_d) are being used as proxies for information spread.

Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings on factor2. Cronbach's alpha for factor2 has a value of 0.73. Factor 2 is named as "Speed" since Total Shortest Paths (Tpaths_d) and Average Geodesic Length (AvgGL_d) are being used as proxies for information speed.

A.1.7.3.2.4 Dependent Variables

Figure 24: Factor Analysis of Dependent Variables



The value of Kaiser-Meyer-Olkin measure of sampling adequacy was 0.467(less than 0.5) but the significance Bartlett's test of sphericity is 0.002. Factor analysis generated three factors that explain 93.702% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.397. Eigenvector Centrality with respect to both, Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN), seem to measuring different constructs within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.1.7.3.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled “RegressionAnalysis_Music.pdf”.

A.1.7.3.3.1 Impact of Network Structure on Network Flow

Table 14: Impact of Network Structure on Network Flow

Network Structure - Network Flow
Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Music	(0.204/0.000) [1,4]	(0.411/0.000) [1,2,4]	(0.595/0.000) [1,2,3,5]	(0.407/0.000) [1,4]	(0.076/0.005) [4]

Table 14 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 20.4%, 41.1%, 59.5%, 40.7% and 7.6% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on AvgGL_ud is not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.3.3.2 Impact of Network Flow on Network Structure

Table 15: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpinN	PL_TSpinN	S_con	R_con	SMSP_d
Music	(0.270/0.000) [7]	(0.265/0.000) [8]	(0.360/0.000) [8,10]	(0.300/0.000) [7,9]	(0.087/0.003) [8]

Table 15 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 27%, 26.5%, 36%, 30% and 8.7% variation in the PL_TpinN, PL_TSpinN, S_con, R_con, and SMSP_d, respectively. The impact of network flow variables on SMSP_d is not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.3.3.3 Impact of Network Structure on Network Phenomenon

Table 16: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Music	(0.085/0.003)[4]	NA	(0.234/0.000)[1,2, 5]	NA

Table 16 shows that the network structure variable impacts Eigenvector Centralization (EC_in), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN), explaining 8.5% and 23.4% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_in), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 17: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Music	(0.034/0.044)[8]	NA	(0.274/0.000)[6,7]	(0.101/0.003)[6,8]

Table 17 shows that the network structure variable impacts Eigenvector Centralization (EC_in), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 3.4%, 27.4% and 10.1% variation respectively. The impact of network flow variables on EC_in and EVCin_TSpinN are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.3.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 18: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d, (6)GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Music	(0.199/0.000) [4,15,1]	(0.064/0.009) [14]	(0.274/0.000)[6,7]	(0.144/0.001) [4,6,7]

Table 18 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_in), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 19.9%, 6.4%, 27.4%and 14.4% variation respectively. The collective impact of independent variables and the moderating variables on Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.4 The Propagation Network

A.1.7.4.1 Correlation Analysis

Significant correlations coefficients for propagation network are shown below in table 19. All correlations between all variables are shown in supplemental file titled "Correlations.pdf".

Table 19: Correlation coefficients of directed network

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	EVCout_TpoutN
Edges_d	Pearson C	.988**	1								
	Sig. (2-taile	.000									
	N	91	91								
Den_d	Pearson C	-.888**	-.870**	-.075	1						
	Sig. (2-taile	.000	.000	.478							
	N	91	91	91	91						
Tpaths_d	Pearson C	.461**	.508**	.605**	-.489**	-.112	.764**	1			
	Sig. (2-taile	.000	.000	.000	.000	.292	.000				
	N	91	91	91	91	91	91	91			
TSpaths_d	Pearson C	.466**	.488**	.415**	-.474**	-.294**	.515**	.752**	1		
	Sig. (2-taile	.000	.000	.000	.000	.005	.000	.000			
	N	91	91	91	91	91	91	91	91		
AvgPL_d	Pearson C	.378**	.438**	.352**	-.434**	-.018	.960**	.847**	.591**	1	
	Sig. (2-taile	.000	.000	.001	.000	.865	.000	.000	.000		
	N	91	91	91	91	91	91	91	91	91	
AvgGL_d	Pearson C	.159	.223*	.305**	-.180	-.088	.683**	.560**	.626**	.686**	
	Sig. (2-taile	.132	.034	.003	.087	.406	.000	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	91	
R_pro	Pearson C	-.734**	-.700**	.050	.522**	.098	-.039	-.200	-.273**	-.088	
	Sig. (2-taile	.000	.000	.640	.000	.356	.715	.057	.009	.406	
	N	91	91	91	91	91	91	91	91	91	
SMSP_d	Pearson C	-.098	-.107	-.018	.065	.945**	-.113	-.142	-.339**	-.073	
	Sig. (2-taile	.354	.313	.862	.544	.000	.286	.181	.001	.489	
	N	91	91	91	91	91	91	91	91	91	
EVCout_TSpoutN	Pearson C	.351**	.360**	.028	-.319**	-.079	.159	.196	.189	.181	.749**
	Sig. (2-taile	.001	.000	.795	.002	.456	.131	.063	.073	.086	.000
	N	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 19 shows that nodes and ties have a strong positive correlation. As the number of nodes (Nodes) increase, the number of ties (Edges_d) also increases. Density

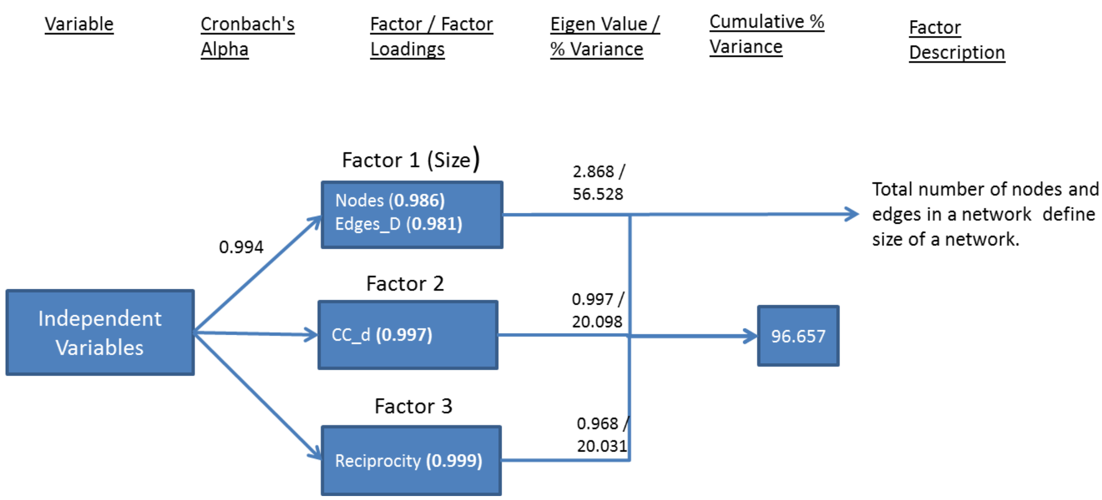
(Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Shortest Paths (TSpaths_d) correlates strongly with Graph Diameter (GD_d) and Total Paths (Tpaths_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Assortativity (R_pro) correlates negatively with Nodes and ties (Edges_d) but has a positive correlation with Density Den_d). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN) and Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) correlate strongly with each other.

A.1.7.4.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.1.7.4.2.1 Independent Variables

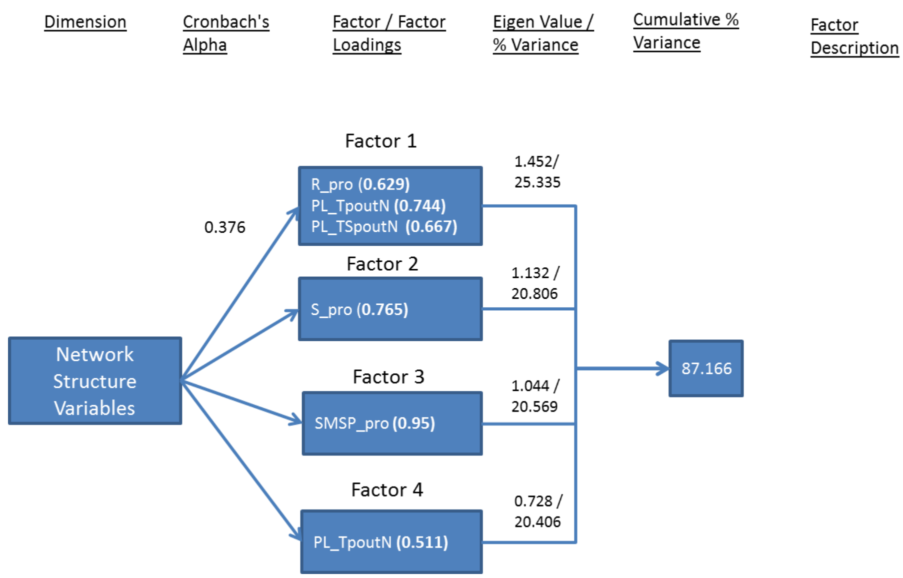
Figure 25: Factor Analysis of Independent Variables



Factor analysis generated three factors that explain 96.65% (greater than 80%) of cumulative variance. Factor 1 has eigenvalue over 1. Factor2 and factor3 have eigenvalues little less than 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had negative loading in factor 1, hence it was removed. Only Clustering Coefficient (CC_d) and Reciprocity have significant loadings in factor 2 and factor 3.Cronbach's alpha for factor 1 has a value of 0.994. This means nodes and ties (Edges_d) are measuring same construct within factor 1. Hence, I name factor 1 as "Size".

A.1.7.4.2.2 Network Structure (MV1)

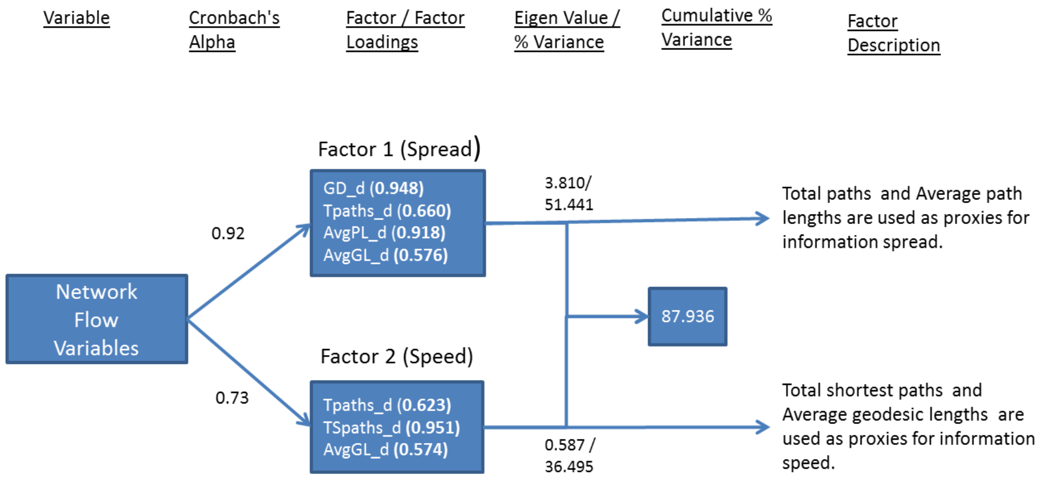
Figure 26: Factor Analysis of Network Structure Variables



Factor analysis generated four factors that explain 87.16% (greater than 80%) of cumulative variance. Factor1, factor2 and factor3 have eigenvalues above 1. Factor 4 has Eigenvalues below 1. Assortativity (R_pro), Power Law Distribution of Paths per Node (PL_TpoutN) and Power Law Distribution of Shortest Paths per Node (PL_TSpoutN) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.376. Assortativity (R_pro), Power Law Distribution of Paths per Node (PL_TpoutN) and Power Law Distribution of Shortest Paths per Node (PL_TSpoutN) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.1.7.4.2.3 Network Flow (MV2)

Figure 27: Factor Analysis of Network Flow Variables

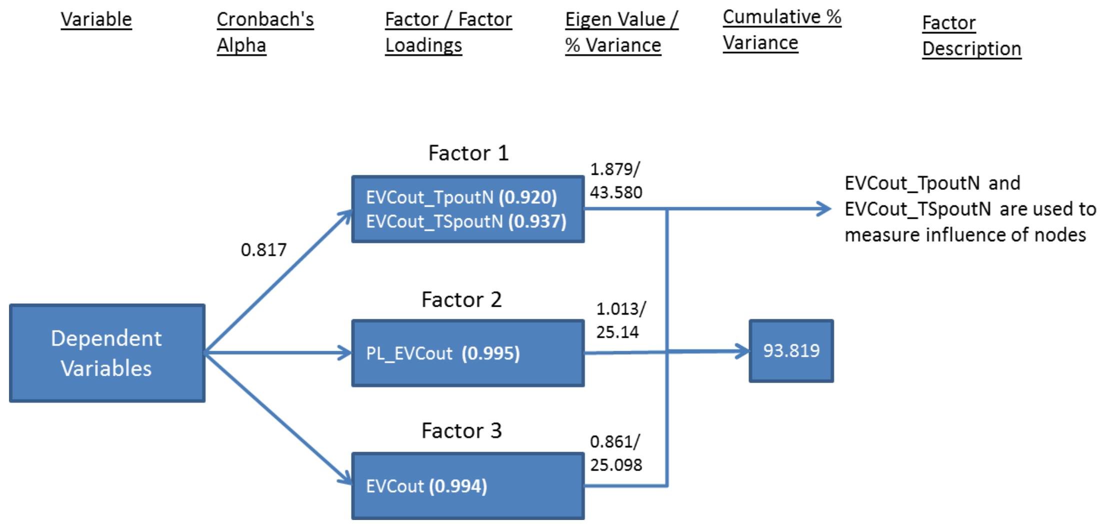


Factor analysis generated two factors that explain 87.936% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.92. Factor 1 is named as "Spread" as average path length and total paths are being used as proxies for information spread.

Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings on factor2. Cronbach's alpha for factor2 has a value of 0.73. Factor 2 is named as "Speed" since total shortest paths and average geodesic length are being used as proxies for information speed.

A.1.7.4.2.4 Dependent Variables

Figure 28: Factor Analysis of Dependent Variables



Factor analysis generated three factors that explain 93.81% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.779. I name factor 1 as "Influence" as both, Eigenvector centralities with respect to paths and shortest paths, are being used measure of influence.

A.1.7.4.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled “RegressionAnalysis_Music.pdf”.

A.1.7.4.3.1 Impact of Network Structure on Network Flow

Table 20: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Music	(0.234/0.000) [4]	(0.034/0.045) [4]	(0.035/0.043) [3]	(0.240/0.000)[4]	(0.416/0.000) [3,4]

Table 20 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 23.4%, 3.4%, 3.5%, 24% and 41.6% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.4.3.2 Impact of Network Flow on Network Structure

Table 21: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpoutN	PL_TSpoutN	S_pro	R_pro	SMSP_d
Music	(0.050/0.019)[6]	(0.058/0.013)[6]	(0.392/0.000)[7,10]	(0.386/0.000)[6,7,10]	NA

Table 21 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 5%, 5.8%, 39.2%, and 38.6% variation in the PL_TpoutN, PL_TSpoutN, S_pro, and R_pro, respectively. The impact of network flow variables on PL_TpoutN, PL_TSpoutN are not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.4.3.3 Impact of Network Structure on Network Phenomenon

Table 22: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Music	NA	(0.034/0.44)[4]	(0.076/0.005)[4]	(0.080/0.004)[4]

Table 22 shows that the network structure variable impacts Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 3.5%, 7.6% and 8% variation respectively. The impact of network flow variables on Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.4.3.4 Impact of Network Flow on Network Phenomenon

Table 23: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Music	(0.064/0.009)[9]	(0.142/0.000)[6,7]	(0.058/0.013)[7]	(0.119/0.001)[7]

Table 23 shows that the network structure variable impacts Eigenvector Centralization (ECout), Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 6.4%, 14.2%, 5.8% and 11.9% variation respectively. The impact of network flow variables on EC_out, EVCout_TpoutN and EVCout_TSpoutN are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.1.7.4.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 24: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d, (6),GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Music	(0.328/0.000) [15]	(0.487/0.000) [8,14,15]	(0.316/0.000) [4,9,11]	(0.358/0.000) [4,9,11]

Table 24 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_out), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 32.8%, 48.7%, 31.6%and 35.8% variation respectively.

A.2 Case 2--Entertainment

A.2.1 Case Overview

Data for keyword "YouTube + Entertainment" was collected over a period of 91 days (31/12/2013 to 31/03/2014). As shown in table 9, overall 44,984 tweets were collected, out of which 10,762 were broadcast tweets and 34,222 were engaged tweets respectively. Out of 34,222 engaged tweets only 16,356 tweets formed the largest community. Similarly, 45,236 daily unique people tweeted overall, out of which 16,670 daily unique people were engaged in broadcast activity whereas 28,566 daily unique people were engaged in conversations. Out of 28,566 daily unique people only 15,822 daily unique people formed the largest community. Data for the largest community was analyzed at a daily interval. The overall trends for the entertainment data are shown below in figure 1 and figure 2.

Figure 1: Overall Tweets

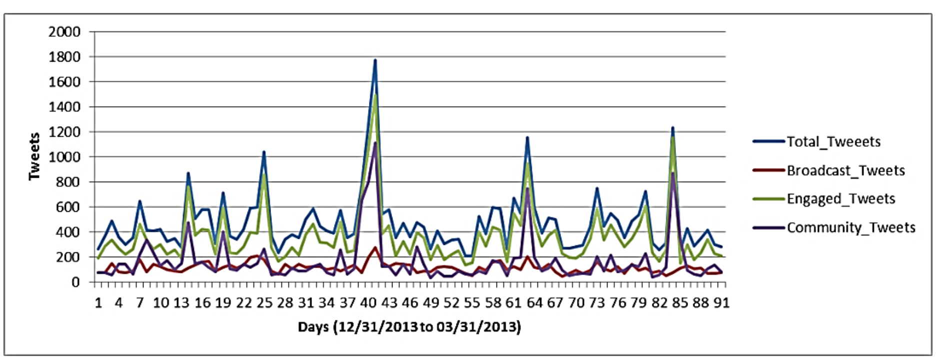


Figure 2: Overall People

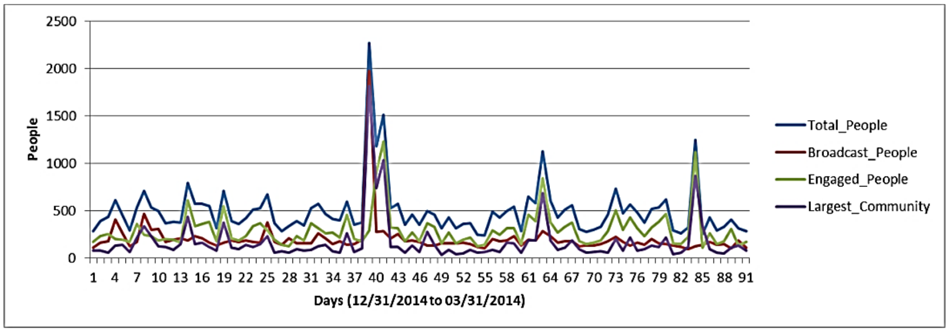


Figure 1 and figure 2 shows that both the total tweets and total people involved are very dynamic and their magnitude changes on a daily basis. The maximum of the total number of daily tweets and the maximum of the total number of daily unique people observed on a single day (the daily uniques) are 1,771 and 2,263 respectively. Similarly, the minimum of the total number daily tweets and the minimum of the number daily unique are 207 and 243, respectively. The size of the largest community on a particular day and the largest number of community tweets on that day also seem to follow the trend of total people and total tweets. The largest number of daily community tweets and the largest number of daily unique people are 1,113 and 1,812, respectively. Similarly, the smallest number of daily community tweets and the smallest number of daily unique people are 35 and 35, respectively. As the total number of daily unique people tweeting increases, so does the size of the community. Most of the engaged people are engaged in the collective conversation forming the largest community.

A.2.2 Random or Not Random

As explained in section 4.4.1, in order to eliminate α- error and β- error, I compare the Clustering Coefficients of both undirected and directed networks with their corresponding random (Erdös-Rényi, E-R) networks. If the Clustering Coefficients of the undirected and directed (CC_ud, CC_d) entertainment networks are equal to those of the E-R random network (CCudran, CCdran), then the directed and undirected networks are considered to be random, if they are not equal, then they are not random.

Figure 3: Comparison of Clustering Coefficients of Undirected Network with E-R Networks

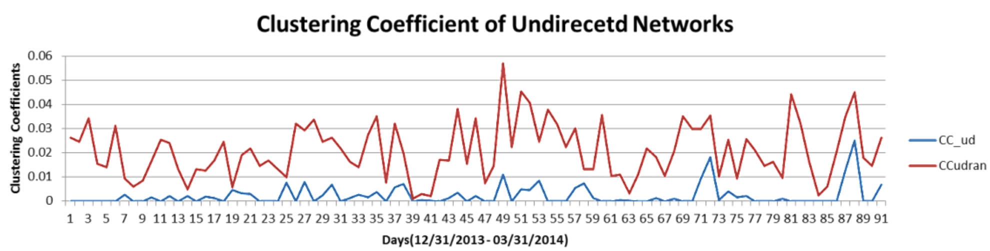
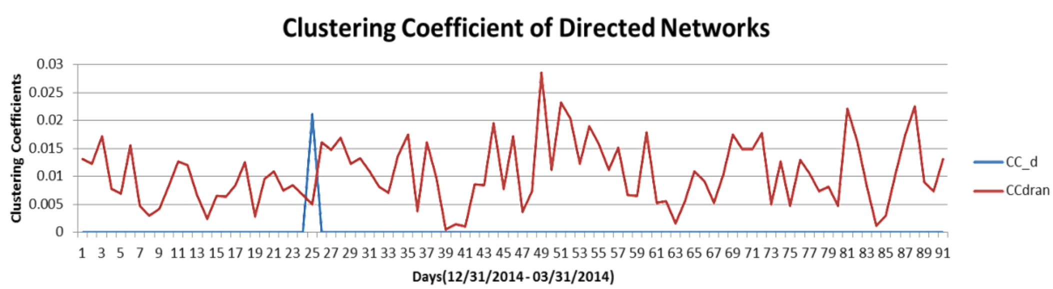


Figure 4: Comparison of Clustering Coefficients of Directed Network with E-R Networks



As seen in figure 3 and figure 4 Clustering Coefficients of both, directed and undirected networks (CC_ud, CC_d) follow very different pattern from their corresponding E-R networks (CCudran, CCdran). Therefore, both these networks are considered to be non-random networks, and the variables computed are a true reflection of network's features.

A.2.3. Independent Variables

The values of the independent variables for both the undirected and the directed entertainment network are shown in figure 5 below.

Figure 5: Independent Variables--(a) Nodes and Edges (Undirected and Directed networks), (b) Reciprocity (Directed Networks), (c) Density (Undirected and Directed Networks), (d) Clustering Coefficient Undirected Network, (e) Clustering Coefficient Directed Network.

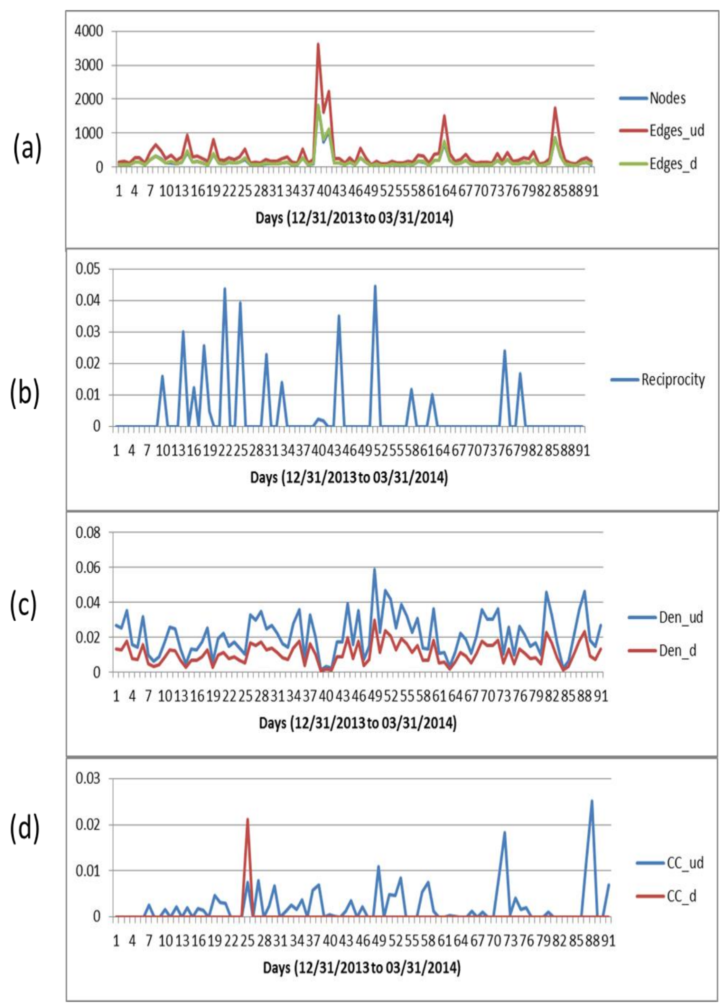


Figure 5 (a) shows that the number of directed ties (Edges_d) in the network and the total number of nodes (Nodes) overlap with each other. The numbers of undirected ties (Edges_ud) is greater than the number of directed ties (Edges_d), because in an undirected network every directed tie is considered to be symmetric. Therefore it is counted twice, except for the ones that are symmetric in a directed network. Reciprocity in figure 5 (b) indicates the presence of symmetric ties in a directed network (in an undirected network 100% are symmetric). The value of 0.01 is equal to 1% of all the ties. Figure 5(c) shows the difference between the densities of the undirected (Den_ud) and the directed (Den_d) networks. The undirected network is denser than the directed network (Den_ud > Den_d). Figure 5 (d) and figure 5 (e) show that the directed networks have higher Clustering Coefficients than the undirected networks (CC_d > CC_ud).

A.2.4 Network Structure Variables (MV1)

A.2.4.1 The Scale Free Metric

Figure 6: Scale Free Metric--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

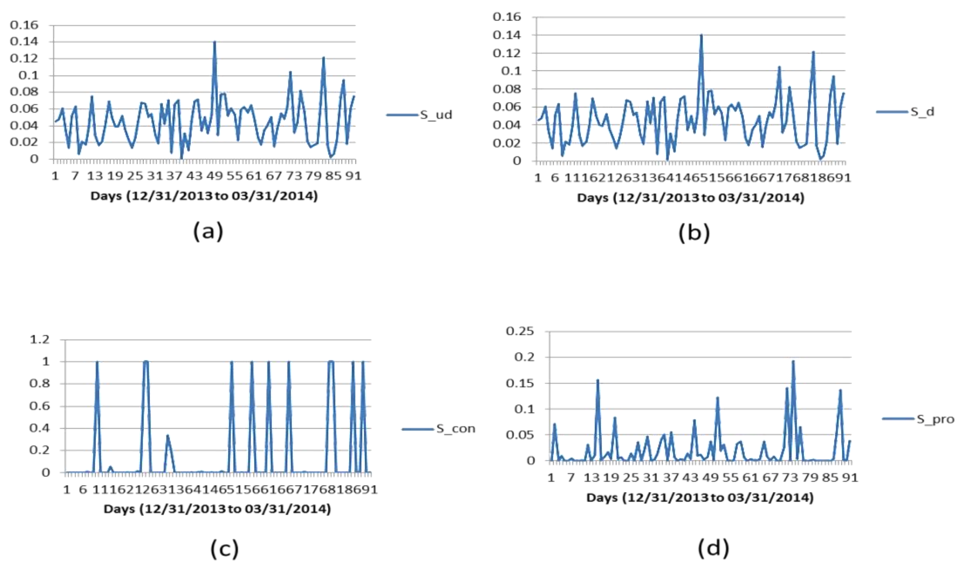
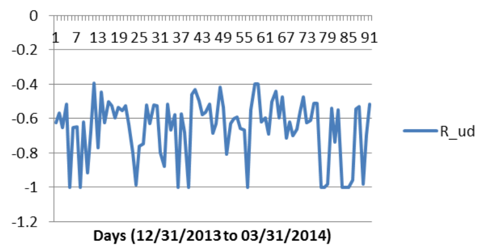


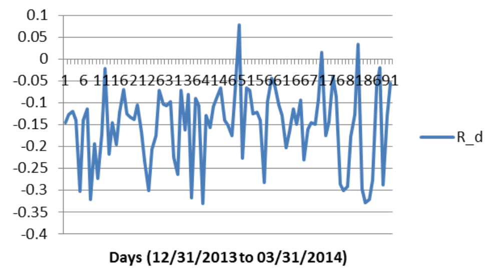
Figure 6 shows the Scale Free Metric for the undirected, directed, consumption and propagation networks (S_ud, S_d, S_con, S_pro). The Scale Free Metrics for the undirected (S_ud) and the directed network (S_d) are similar, but the Scale Free Metrics for the consumption (S_con) and propagation (S_pro) networks are very different. The propagation network is more Scale Free than the consumption network (S_pro > S_con). The values of the Scale Free Metric ranges between 0 and 1. When the values are closer to 1, it means that the networks are more Scale Free. None of the networks are Scale Free in nature. This means that these networks have hubs in them. However, there is not just one hub that is the center of the community.

A.2.4.2 The Assortativity

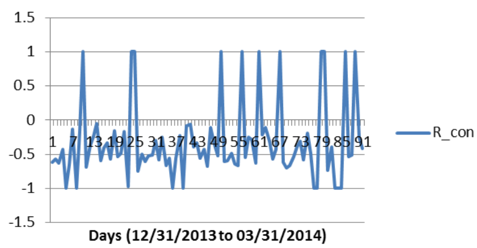
Figure 7: Assortativity--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



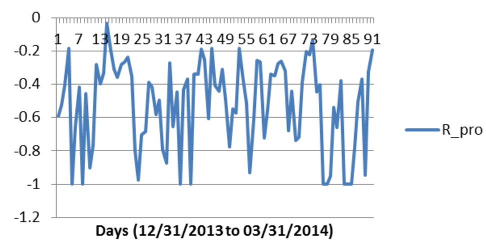
(a)



(b)



(c)



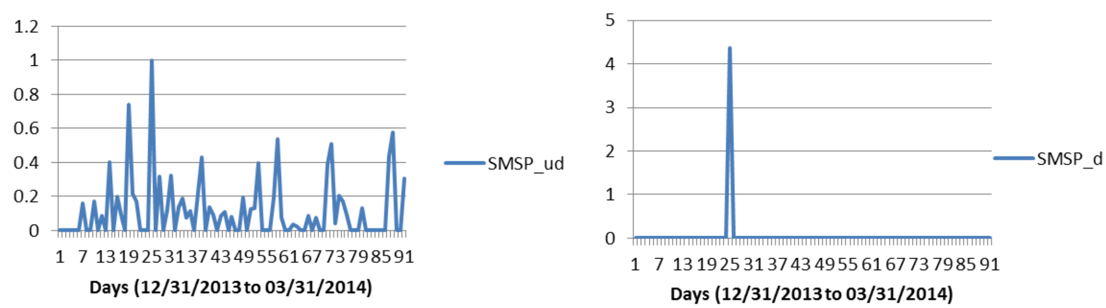
(d)

Figure 7 shows the assortativity metric for the undirected, directed, consumption and propagation networks (R_ud, R_d, R_con, R_Pro). The value of the assortativity metric ranges between -1 and 1. When the values are closer to -1, it means that networks are disassortative. The undirected network is more Disassortative than the directed network (R_d > R_ud). Among the directed networks, the consumption network is more Disassortative than the propagation network (R_pro > R_con). Disassortative means that the nodes in the network connect to nodes that are very similar to themselves. This is true more so in the undirected network and in the consumption network than it is in the directed network and the propagation network.

This implies that disassortativeness of consumption contributes more to the disassortativeness of the directed network than the disassortativeness of the propagation does.

A.2.4.3 The Small World Metric

Figure 8: Small World Metric -- (a) Undirected Network, (b) Directed Network.

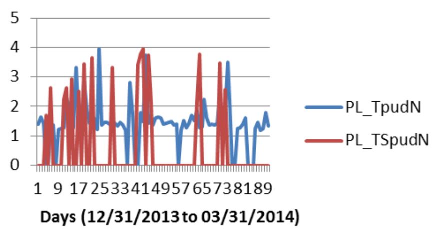


(a) (b)

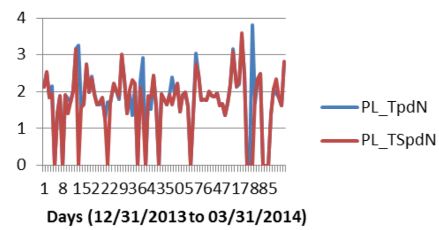
Figure 8 shows the Small World Metric for the undirected (SMSP_ud) and directed networks (SMSP_d). The Small World Metrics for the consumption and propagation networks are the same as the ones for the directed network. The directed networks don't show any small world behavior. Contrary to the directed networks, undirected networks show some small world behavior but not significantly enough. This means that in undirected networks there are more nodes that act as hubs that facilitate communication between other nodes of the network.

A.2.4.4 Paths and Shortest Paths Power law Distribution per Node

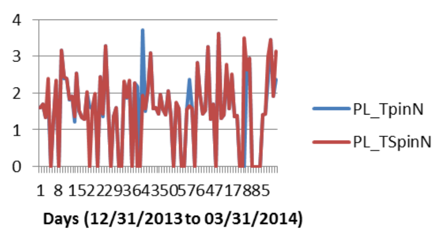
Figure 9: Power Law Distribution of Paths and Shortest Paths in (a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



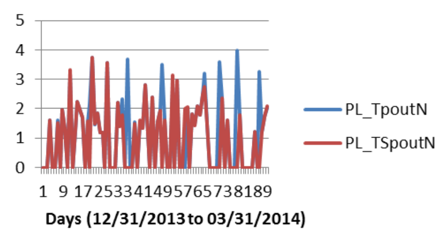
(a)



(b)



(c)



(d)

Figure 9 (a) shows that, in the undirected network, paths are more uniformly distributed among nodes than shortest paths are distributed among nodes. This means that fewer nodes are responsible for more of the shortest paths in the undirected network. There are fewer instances of shortest path following power law distribution in undirected (figure 9 (a)) and consumption (figure 9 (c)) networks. In the directed (figure 9 (b)) and propagation (figure 9 (d)) networks, there are no such patterns.

A.2.5 Network Flow Variables (MV2)

Figure 10: Network Flow Variables-- (a) Total Paths and Total Shortest Paths, (b) Average Paths and Average Shortest Paths, (c) Undirected and Directed Network Graph Diameter.

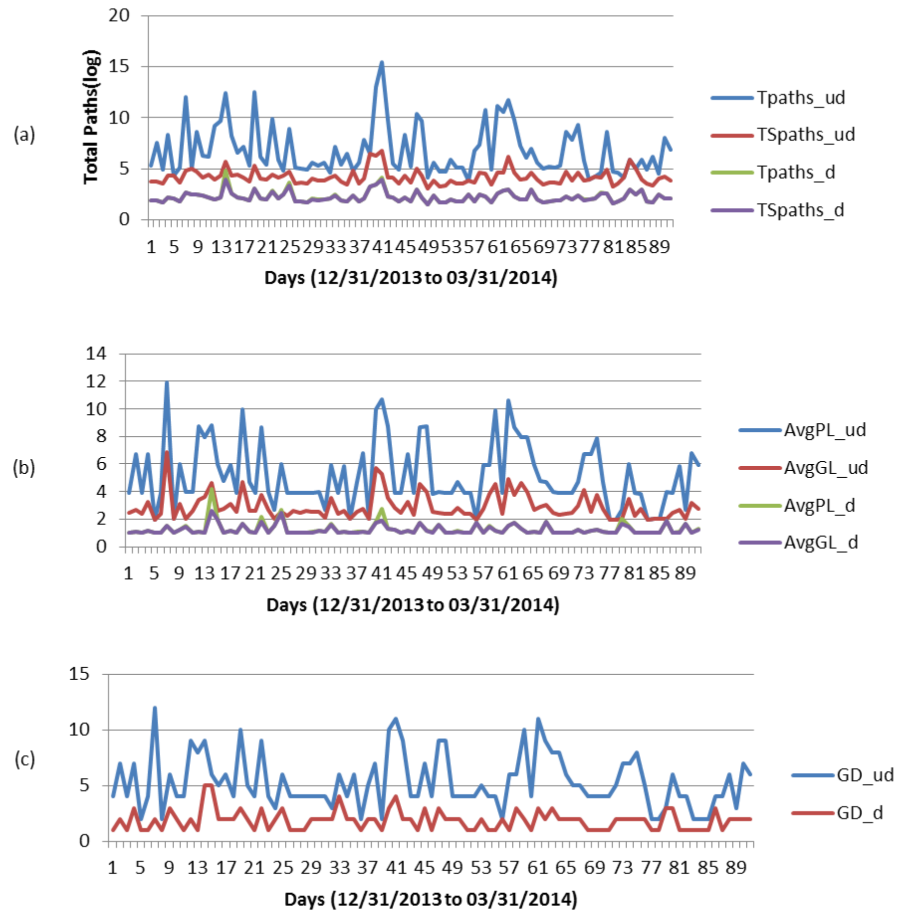


Figure 10 (a), shows that Total Number of Paths in the undirected network (Tpaths_ud) is orders of magnitude higher than the Total Number of Shortest Paths (TSpaths_ud). The Total Number of Paths (Tpaths_d) and the Total Number of Shortest Paths (TSpaths_d) map more closely in the directed network. In figure 10 (b), a similar trend is observed in the Average Path Lengths and the Average Geodesic Lengths of the undirected and directed networks (AvgPL_ud, AvgPL_d, AvgGL_ud, AvgGL_d). In figure

10 (c), the Graph Diameter of the undirected network (GD_ud) is larger than the Graph Diameter of the directed network (GD_d). It is also noteworthy that, in figure 10 (b) and in figure 10 (c), the Graph Diameter and the Average Path Length of the undirected and directed networks (GD_ud, AvgPL_ud, GD_d, AvgPL_d) track pretty closely.

A.2.6 Dependent Variables

A.2.6.1 Eigenvector Centralization

Figure 11: Eigenvector Centralization in the Undirected, Directed, Consumption and Propagation Networks

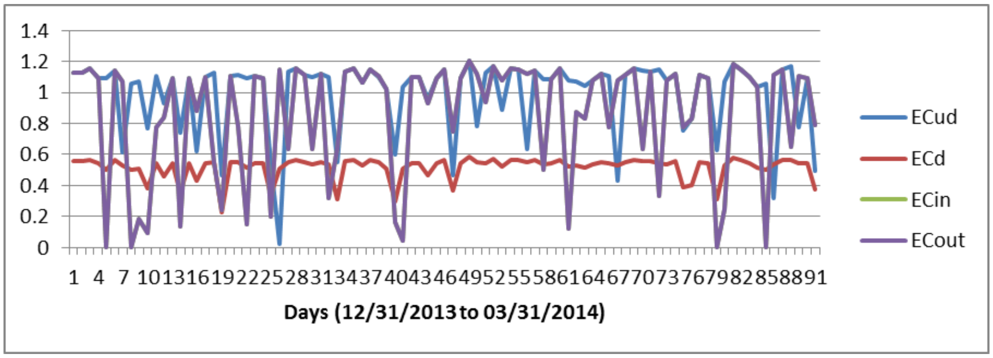


Figure 11 shows that nodes with influence are lot more central in the undirected network than in the directed, consumption and propagation networks (ECud > (ECd, ECin, ECout)). The consumption and propagation networks exhibit same level of centralization.

A.2.6.2 Power law Distribution of Eigenvector Centrality per Node

Figure 12: Power Law Distribution of Eigenvector Centrality in Undirected, Directed, Consumption and Propagation Network

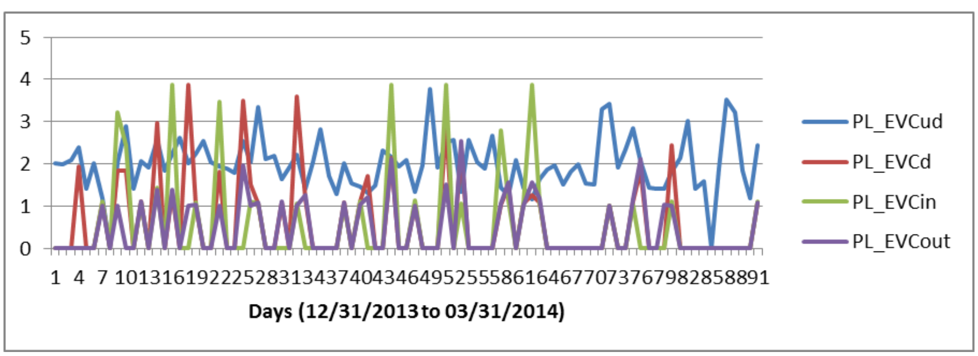
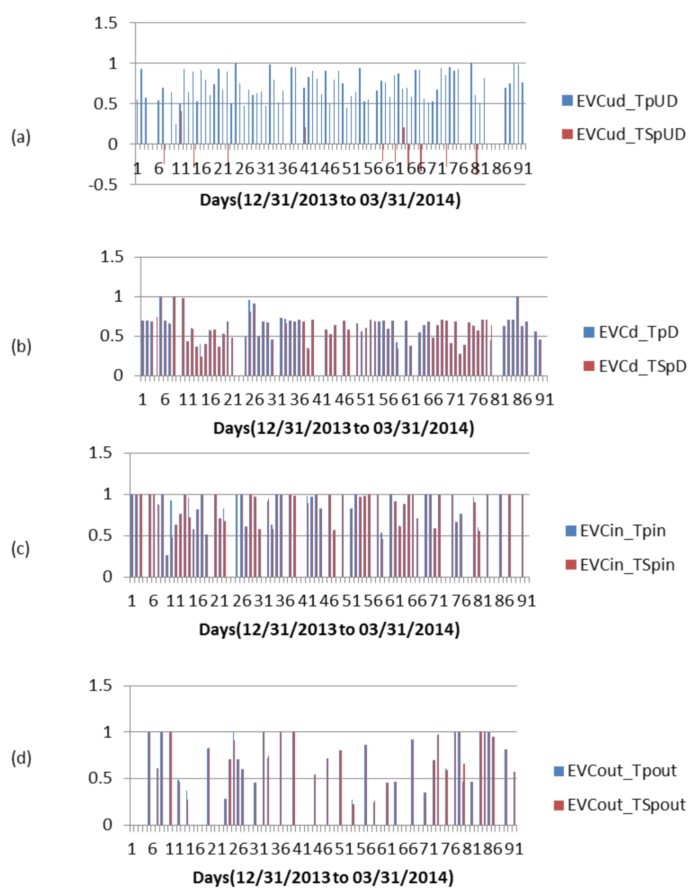


Figure 12 shows that in undirected, directed, consumption and propagation network the distribution of Eigenvector Centrality amongst nodes have similar Power Law patterns.

A.2.6.3 Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node

Figure 13: Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



In figure 13, only those correlation coefficients with a significance value lower than 0.05 are shown. In figure 13 (a), there is a significant correlation between the eigenvector centrality of a node and the number of paths from a node in undirected network (EVCud_TpUDN). There is no significant correlation between eigenvector centrality of a node and shortest paths from a node in undirected network (EVCud_TSpUDN). In figure 13 (b), there is a significant correlation between the directed-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the directed network (EVCd_TpDN, EVCud_TSpUDN). In figure 13 (c), there is a significant correlation between the in-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the consumption network (EVin_TpinN, EVCin_TSpinN). The correlation between the out-eigenvector centrality of a node and the number of shortest paths is less significant figure 13 (d) (EVCout_TpoutN, EVCout_TSpoutN).

A.2.7 Statistical Analysis

A.2.7.1 The Undirected Network

A.2.7.1.1 Correlation Analysis

In Table 1, the statistically significant Correlation Coefficients for the undirected network are marked in yellow. All correlations between all variables are shown in supplemental file titled "Correlations.pdf".

Table 1: Correlation Coefficients of Undirected Network

Correlations

		Nodes	Edges_ud	Den_ud	CC_ud	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud	PL_TpudN	PL_TSpudN	S_ud	R_ud
Edges_ud	Pearson C	.999**	1											
	Sig. (2-tail	.000												
	N	91	91											
Den_ud	Pearson C	-.573**	-.576**	1										
	Sig. (2-tail	.000	.000											
	N	91	91	91										
Tpaths_ud	Pearson C	.410**	.435**	-.587**	-.106	.928**	1							
	Sig. (2-tail	.000	.000	.000	.315	.000								
	N	91	91	91	91	91	91							
TSpaths_ud	Pearson C	.840**	.853**	-.866**	-.291**	.380**	.676**	1						
	Sig. (2-tail	.000	.000	.000	.005	.000	.000							
	N	91	91	91	91	91	91	91						
AvgPL_ud	Pearson C	.139	.161	-.344**	.017	.999**	.933**	.389**	1					
	Sig. (2-tail	.187	.127	.001	.870	.000	.000	.000						
	N	91	91	91	91	91	91	91	91					
AvgGL_ud	Pearson C	.265*	.288**	-.441**	-.028	.907**	.898**	.535**	.912**	1				
	Sig. (2-tail	.011	.006	.000	.792	.000	.000	.000	.000					
	N	91	91	91	91	91	91	91	91	91				
S_ud	Pearson C	-.444**	-.443**	.727**	.555**	.119	-.170	-.610**	.115	.049	.390**	.004	1	
	Sig. (2-tail	.000	.000	.000	.000	.262	.106	.000	.279	.646	.000	.971		
	N	91	91	91	91	91	91	91	91	91	91	91	91	
R_ud	Pearson C	-.145	-.123	.180	.323**	.631**	.469**	-.067	.629**	.555**	.530**	.380**	.579**	1
	Sig. (2-tail	.170	.245	.087	.002	.000	.000	.527	.000	.000	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91
SMSP_ud	Pearson C	-.034	-.019	.059	.733**	.178	.160	.007	.181	.132	.466**	.119	.275**	.301**
	Sig. (2-tail	.746	.856	.581	.000	.091	.129	.947	.085	.212	.000	.261	.008	.004
	N	91	91	91	91	91	91	91	91	91	91	91	91	91
PL_EVCudN	Pearson C	-.280**	-.277**	.546**	.547**	-.035	-.176	-.440**	-.038	-.137	.353**	.021	.596**	.361**
	Sig. (2-tail	.007	.008	.000	.000	.741	.096	.000	.718	.195	.001	.844	.000	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91
EVCud_TpudN	Pearson C	-.217*	-.203	.076	.257*	.538**	.362**	-.095	.515**	.408**	.705**	.200	.411**	.629**
	Sig. (2-tail	.039	.054	.474	.014	.000	.000	.369	.000	.000	.000	.058	.000	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

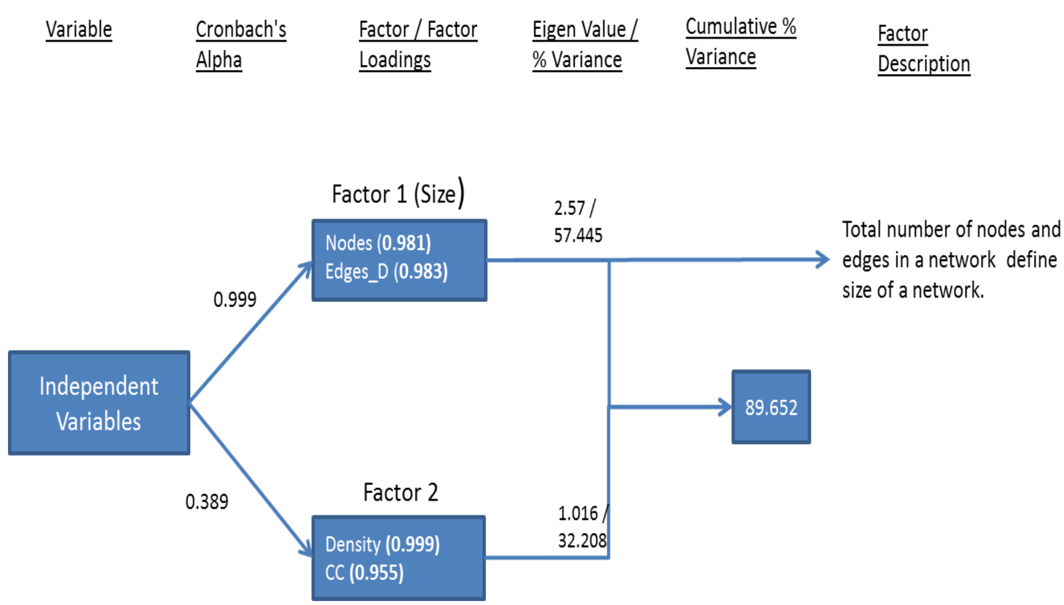
Table 1 shows that the number of nodes (Nodes) and the number of ties (Edges_ud) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. The Density (Den_ud) of this network has a strong negative correlation with both the number of nodes (Nodes) and the number of ties (Edges_ud). Total Paths (Tpaths_ud) has a negative correlation with Density (Den_ud) but has a strong positive correlation with Graph Diameter (GD_ud). Total Shortest Paths (TSpaths_ud) share a positive correlation with the number of nodes (Nodes), number of ties (Edges_ud) and Total Paths (Tpaths_ud) but share a negative correlation with Density (Den_ud). Average Path Length (AvgPL_ud) and Average Geodesic Length (AvgGL_ud) share a strong positive correlation with Graph Diameter (GD_ud) and Total Paths (Tpaths_ud). Average Geodesic Length (AvgGL_ud) shares a positive correlation with Total Shortest Paths (TSpaths_ud) and a strong correlation with Average Path Length (AvgPL_ud). Scale Free Metric (S_ud) shares a positive correlation with Density (Den_ud) and Clustering Coefficient (CC_ud) and a negative correlation with Total Shortest Paths (TSpaths_ud). Assortativity (R_ud) shares positive correlation with Graph Diameter (GD_ud), Average Path Length (AvgPL_ud), Average Geodesic Length (AvgGL_ud), Paths Power Law Distribution per Node (PL_TpudN) and Scale Free Metric (S_ud). Small World Metric (SMSP_ud) shares a positive correlation with Clustering Coefficient (CC_ud). Power law Distribution of Eigenvector Centrality per Node (PL_EVCudN) shares a positive correlation with Density (Den_ud), Clustering Coefficient (CC_ud) and Scale Free Metric (S_ud). Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) correlates strongly with Graph Diameter (GD_ud), Average Path Length (AvgPL_ud), Paths Power Law Distribution per Node (PL_TpudN) and Assortativity (R_ud).

A.2.7.1.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.2.7.1.2.1 Independent Variables

Figure 14: Factor Analysis Independent Variables Music Undirected Network

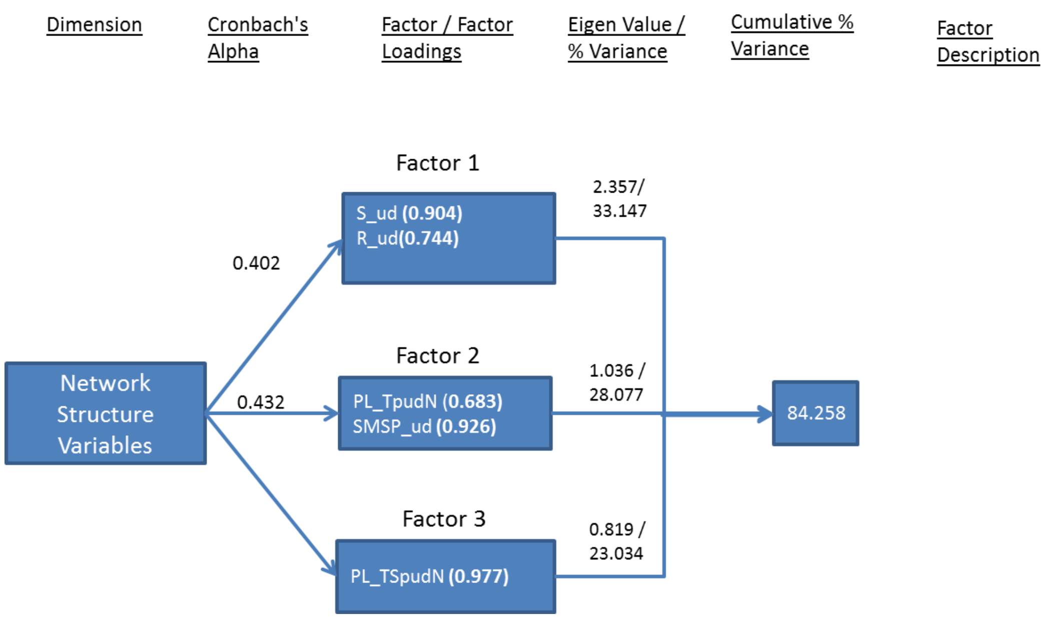


The factor analysis generated two factors that explain 89.65% (greater than 80%) of the cumulative variance. Both factors have eigenvalues above one. Nodes and ties (Edges_ud) have significant factor loadings in factor 1. Density (Den_ud) and Clustering

Coefficient (CC_ud) have significant loading in factor 2.Cronbach's alpha for factor 1has a value of 0.999 and factor2 has a value of 0.389. This means nodes and ties are measuring same construct within factor 1 whereas density and clustering coefficient are not measuring the same construct. Hence, I name factor 1 as "Size".

A.2.7.1.2.2 Network Structure (MV1)

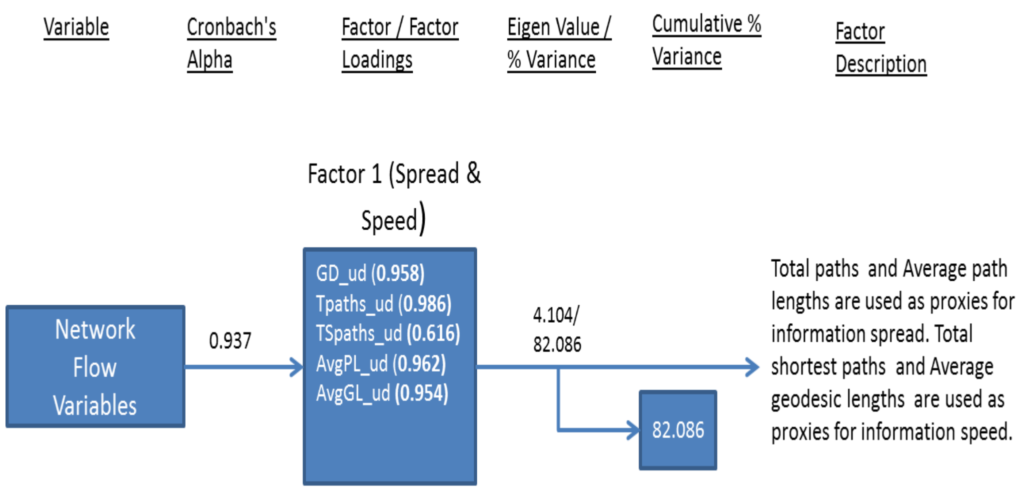
Figure 15: Factor Analysis of Network Structure Variables



The factor analysis generated three factors that explain 84.25% (greater than 80%) of the cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor 3 has eigenvalue below 1. Scale Free Metric (S_ud) and Assortativity (R_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.402. Scale Free Metric (S_ud) and Assortativity (R_ud) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. Power Law Distribution of Total Paths per Node (PL_TpudN) and Small World Metric (SMSP_ud) have significant factor loadings in factor 2. Cronbach's alpha for factor1 has a value of 0.432. Power Law Distribution of Total Paths per Node (PL_TpudN) and Small World Metric (SMSP_ud) are measuring different constructs within factor 2. Hence, they should not be considered as a factor. All other variables load independently.

A.2.7.1.2.3 Network Flow (MV2)

Figure 16: Factor Analysis of Network Flow Variables



The factor analysis generated one factor that explains 82.08% (greater than 80%) of the cumulative variance. Factor1 has eigenvalues above 1. All variables have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.937. Hence, they should be considered as a factor.

A.2.7.1.2.4 Dependent Variables

The value of Kaiser-Meyer-Olkin measure of sampling adequacy was 0.480 (less than 0.5), and the significance Bartlett's test of sphericity is 0.108. This data does not satisfy the measure of appropriateness for factor analysis. Therefore, all the variables are considered independently.

A.2.7.1.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Entertainment.pdf".

A.2.2.1.1.3.1 Impact of Network Structure on Network Flow

Table 2: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud
Entertainment	(0.477/0.000) [3,4]	(0.523/0.000) [1,3,4]	(0.501/0.000) [2,3,4]	(0.477/0.000) [3,4]	(0.406/0.000) [3,4]

Table 2 shows that the network structure variables have a significant impact on the network flow variables. Network structure variables explain 47.7%, 52.3%, 50.1%, 47.7% and 40.6% variation in Graph Diameters (GD_ud), Total Paths (Tpaths_ud), Total

Shortest Paths (TSpaths_ud) Average Path Length (AvgPL_ud), and Average Geodesic Length (AvgGL_ud), respectively.

A.2.7.1.3.2 Impact of Network Flow on Network Structure

Table 3: Impact of Network Flow on Network Structure

Predictors: **(6) GD_ud, (7) Tpaths_ud, (8) TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpudN	PL_TSpudN	S_ud	R_ud	SMSP_ud
Entertainment	(0.212/0.000) [6,7]	(0.207/0.000) [7]	(0.559/0.000) [8,10]	(0.496/0.000)[6,8]	NA

Table 3 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 21.2%, 20.7%, 55.9% and 49.6% variation in the PL_TpudN, PL_TSpudN, S_ud and R_ud, respectively. The impact of network flow variables on PL_TSpudN is not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.1.3.3 Impact of Network Structure on Network Phenomenon

Table 4: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Entertainment	(0.091/0.005) [4,5]	(0.402/0.000) [4,5]	(0.0578/0.000) [1,4]	NA

Table 4 shows that the network structure variable Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 9.1%, 40.2% and 5.78% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.1.3.4 Impact of Network Flow on Network Phenomenon

Table 5: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_ud, (7)Tpaths_ud, (8)TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Entertainment	(0.056/0.013) [8]	(0.184/0.000) [8]	(0.282/0.000) [6]	(0.076/0.005)[10]

Table 5 shows that the network flow variable impacts Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 5.6%, 18.4%, 28.2% and 7.6% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud), Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.1.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 6: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud, (6),GD_ud (7) Tpaths_ud (8), TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud, (11) Nodes, (12) Edges_ud, (13) Den_ud, (14) CC_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Entertainment	(0.120/0.001) [14,5]	(0.041/0.000) [3,14]	(0.597/0.000) [1,4,6]	(0.076/0.005) [10]

Table 6 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 12%, 4.1%, 59.7% and 7.6% variation respectively. The collective impact of independent variables and the moderating variables on Eigenvector Centralization (EC_ud) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.2 The Directed Network

A.2.7.2.1 Correlation Analysis

Significant Correlations Coefficients for directed network are shown below in table 7. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 7: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpdN	PL_TSpdN	S_d	EVCd_TpdN
Edges_d	Pearson	.999**	1												
	Sig. (2-ta	.000													
	N	91	91												
Den_d	Pearson	-.572**	-.576**	.027	1										
	Sig. (2-ta	.000	.000	.797											
	N	91	91	91	91										
Tpaths_d	Pearson	.644**	.662**	.249*	-.756**	.261*	.663**	1							
	Sig. (2-ta	.000	.000	.017	.000	.012	.000								
	N	91	91	91	91	91	91	91							
TSpaths_d	Pearson	.680**	.697**	.190	-.803**	.242*	.641**	.988**	1						
	Sig. (2-ta	.000	.000	.071	.000	.021	.000	.000							
	N	91	91	91	91	91	91	91	91						
AvgPL_d	Pearson	.261*	.283**	.410**	-.415**	.313**	.755**	.850**	.785**	1					
	Sig. (2-ta	.013	.007	.000	.000	.002	.000	.000	.000						
	N	91	91	91	91	91	91	91	91	91					
AvgGL_d	Pearson	.216*	.236*	.359**	-.455**	.382**	.765**	.820**	.796**	.944**	1				
	Sig. (2-ta	.040	.025	.000	.000	.000	.000	.000	.000	.000					
	N	91	91	91	91	91	91	91	91	91	91				
PL_TSpdN	Pearson	-.294**	-.284**	-.142	.261*	-.215*	.103	-.311**	-.296**	-.201	-.161	.748**	1		
	Sig. (2-ta	.005	.006	.180	.012	.041	.333	.003	.004	.057	.126	.000			
	N	91	91	91	91	91	91	91	91	91	91	91	91		
S_d	Pearson	-.444**	-.443**	.002	.726**	-.080	-.186	-.564**	-.599**	-.307**	-.334**	.474**	.541**	1	
	Sig. (2-ta	.000	.000	.982	.000	.450	.077	.000	.000	.003	.001	.000	.000		
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
R_d	Pearson	-.340**	-.326**	.070	.533**	-.067	.063	-.364**	-.405**	-.141	-.194	.603**	.664**	.906**	
	Sig. (2-ta	.001	.002	.508	.000	.528	.554	.000	.000	.183	.065	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
SMSP_d	Pearson	.024	.036	.372**	-.099	1.000**	.131	.261*	.242*	.313**	.382**	-.010	-.215*	-.080	
	Sig. (2-ta	.822	.731	.000	.353	0.000	.217	.012	.021	.002	.000	.923	.041	.450	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
ECd	Pearson	-.260*	-.275**	-.392**	.436**	-.263*	-.475**	-.555**	-.548**	-.506**	-.525**	-.172	.126	.187	
	Sig. (2-ta	.013	.008	.000	.000	.012	.000	.000	.000	.000	.000	.103	.232	.075	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
PL_EVCdN	Pearson	.079	.095	.637**	-.166	.315**	.398**	.335**	.297**	.394**	.359**	.178	.019	-.041	
	Sig. (2-ta	.458	.372	.000	.116	.002	.000	.001	.004	.000	.000	.091	.860	.697	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
EVCd_TSpdN	Pearson	-.078	-.092	-.065	.128	.111	-.381**	-.227*	-.220*	-.263*	-.244*	-.351**	-.365**	-.051	.952**
	Sig. (2-ta	.462	.386	.538	.228	.295	.000	.031	.036	.012	.020	.001	.000	.630	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 7 shows that nodes (Nodes) and ties (Edges_d) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of ties (Edges_d), Reciprocity and Total Paths (Tpaths_d) in the network. Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpdN) correlates positively with Paths Power Law Distribution per Node (PL_TpdN). Scale Free Metric (S_d) seems to share a positive relationship with Density (Den_d) and Shortest Paths Power Law Distribution per Node (PL_TSpdN). Scale Free Metric (S_d) seems to share a negative relationship Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d). Assortativity (R_d) shares a positive relationship with (Den_d), Shortest Paths Power Law Distribution per Node (PL_TSpdN), Paths Power Law Distribution per Node (PL_TpdN) and Scale Free Metric (S_d). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centralization (ECd) correlates negatively with Total Paths (Tpaths_d), Total Shortest

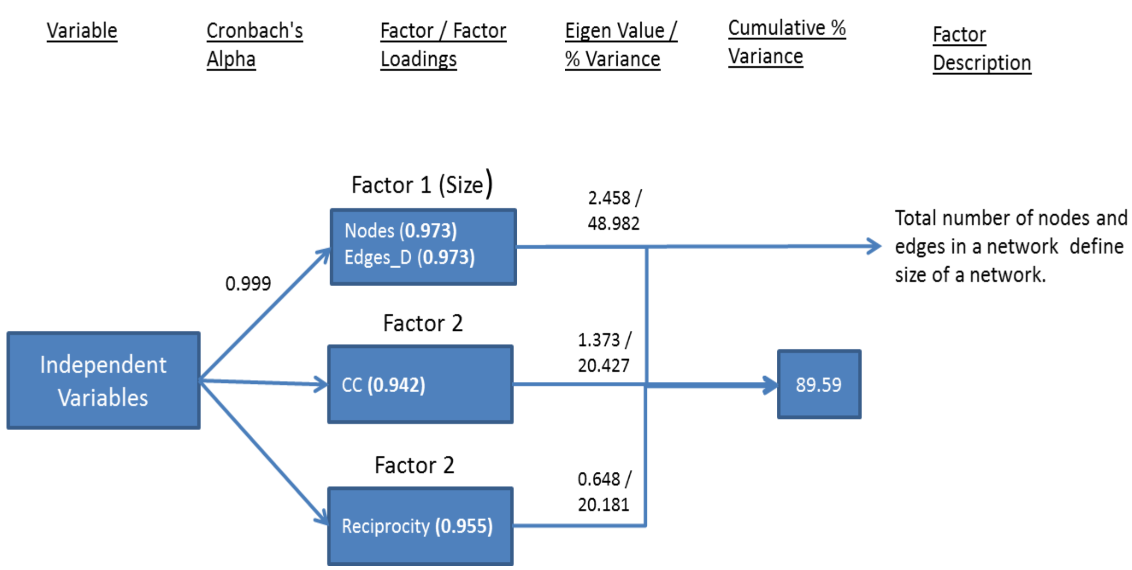
Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) and Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) correlate strongly with each other.

A.2.7.2.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.2.7.2.2.1 Independent Variables

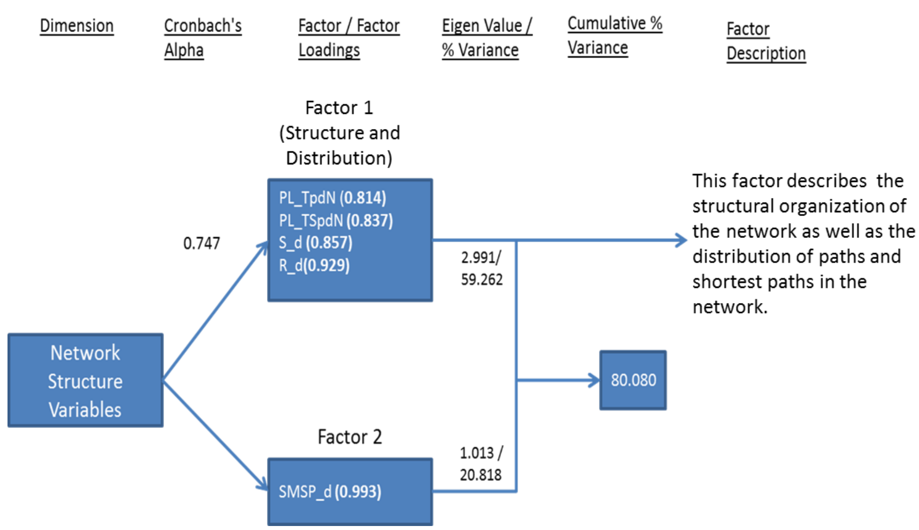
Figure 17: Factor Analysis of Independent Variables



Factor analysis generated three factors that explain 89.59% (greater than 80%) of cumulative variance. Factor 1 and factor2 have eigenvalues over 1. Factor3 has eigenvalue less than 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had negative loading in factor 1, hence it was removed. Only Clustering Coefficient (CC_d) and Reciprocity have significant loadings in factor 2 and factor 3.Cronbach's alpha for factor 1 has a value of 0.999. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as "Size".

A.2.7.2.2.2 Network Structure (MV1)

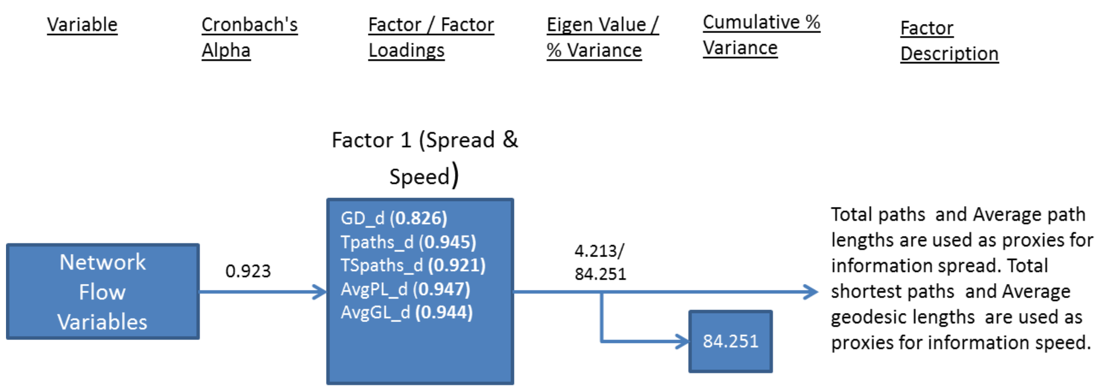
Figure 18: Factor Analysis of Network Structure Variables



Factor analysis generated two factors that explain 80.08% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Paths Power Law Distribution per Node (PL_TpdN), Shortest Paths Power Law Distribution per Node (PL_TSpdN), Assortativity (R_d) and Scale Free Metric (S_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.747. Paths Power Law Distribution per Node (PL_TpdN), Shortest Paths Power Law Distribution per Node (PL_TSpdN), Assortativity (R_d) and Scale Free Metric (S_d) are measuring same construct within factor 1. Hence, they should be considered as a factor. All other variables load independently.

A.2.7.2.2.3 Network Flow (MV2)

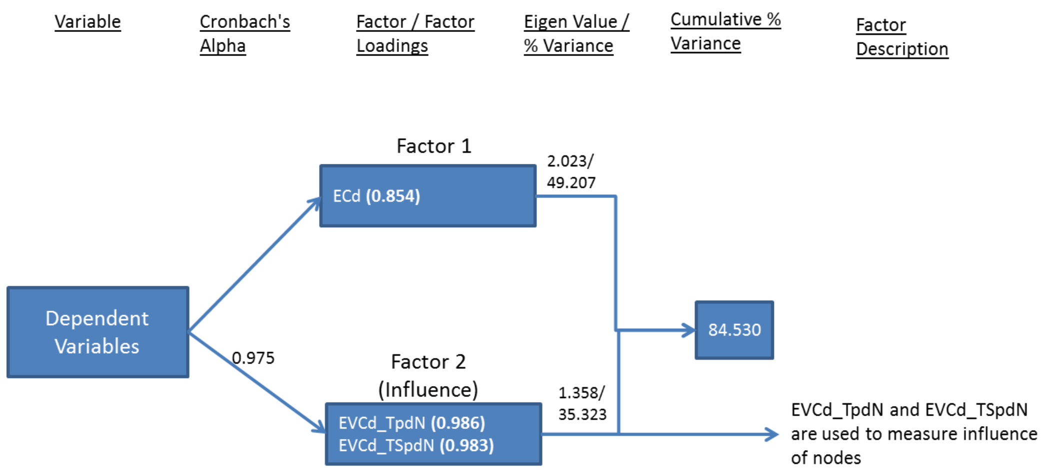
Figure 19: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 84.251% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.923. Factor 1 is named as "Spread and Speed".

A.2.7.2.2.4 Dependent Variables

Figure 20: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 84.53% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCd_TpdN) and Shortest Paths (EVCd_TSpdN) have significant factor loading on factor 2. Factor 2 has a Cronbach's alpha of 0.975. I name the factor 2 as "Influence" as both, Eigenvector Centralities with respect to Paths (EVCd_TpdN) and Shortest Paths (EVCd_TSpdN), are being used measure of influence.

A.2.7.2.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Entertainment.pdf"..

A.2.7.2.3.1 Impact of Network Structure on Network Flow

Table 8: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Entertainment	(0.252/0.000)[1,3]	(0.306/0.000)[3]	(0.351/0.000)[3]	(0.234/0.000)[3,5]	(0.333/0.000)[1,3,5]

Table 8 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 25.2%, 30.6%, 35.1%, 23.4% and 33.3% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively.

A.2.7.2.3.2 Impact of Network Flow on Network Structure

Table 9: Impact of Network Flow on Network Structure

Predictors: **(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpdN	PL_TSpdN	S_d	R_d	SMSP_d
Entertainment	(0.253/0.000)[6,8]	(0.251/0.000)[6,7]	(0.351/0.000)[7]	(0.326/0.000)[6,8]	NA

Table 9 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 25.3%, 25.1%, 35.1%, and 32.6.0% variation in the PL_TpdN, PL_TSpdN, S_d and R_ud, respectively.

A.2.7.2.3.3 Impact of Network Structure on Network Phenomenon

Table 10: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Entertainment	(0.059/0.012)[5]	(0.089/0.002)[5]	0.157/0.000)[2, 3]	(0.123/0.000)[2]

Table 10 shows that the network structure variable impacts Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN), explaining 5.9%, 8.9%, 15.7% and 12.3% variation respectively. The impact of network flow variables Eigenvector Centralization (EC_d) and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.2.3.4 Impact of Network Flow on Network Phenomenon

Table 11: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Entertainment	(0.300/0.000)[7]	(0.149/0.000)[6]	(0.135/0.000)[6]	(0.135/0.000)[6]

Table 11 shows that the network structure variable impacts Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpdN), explaining 30%, 14.9%, 13.5% and 13.5% variation respectively.

A.2.7.2.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 12: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d, (6)GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCud_TpdN	EVCud_TSpdN
Entertainment	(0.362/0.000) [7,15]	(0.456/0.000) [6,15]	(0.239/0.000) [2,6]	(0.235/0.000) [2,6]

Table 12 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 36.2%, 45.6%, 23.9% and 23.5% variation respectively.

A.2.7.3 The Consumption Network

A.2.7.3.1 Correlation Analysis

Significant correlations coefficients for consumption network are shown below in table 13. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 13: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpinN	S_con	EVCin_TpinN
Edges_d	Pearson	.999**	1											
	Sig. (2-ta	.000												
	N	91	91											
Den_d	Pearson	-.572**	-.576**	.027	1									
	Sig. (2-ta	.000	.000	.797										
	N	91	91	91	91									
Tpaths_d	Pearson	.644**	.662**	.249*	-.756**	.261*	.663**	1						
	Sig. (2-ta	.000	.000	.017	.000	.012	.000							
	N	91	91	91	91	91	91	91						
TSpaths_d	Pearson	.680**	.697**	.190	-.803**	.242*	.641**	.988**	1					
	Sig. (2-ta	.000	.000	.071	.000	.021	.000	.000						
	N	91	91	91	91	91	91	91	91					
AvgPL_d	Pearson	.261*	.283**	.410**	-.415**	.313**	.755**	.850**	.785**	1				
	Sig. (2-ta	.013	.007	.000	.000	.002	.000	.000	.000					
	N	91	91	91	91	91	91	91	91	91				
AvgGL_d	Pearson	.216*	.236*	.359**	-.455**	.382**	.765**	.820**	.796**	.944**	1			
	Sig. (2-ta	.040	.025	.000	.000	.000	.000	.000	.000	.000				
	N	91	91	91	91	91	91	91	91	91	91			
PL_TSpinN	Pearson	-.170	-.159	.068	.079	-.011	.382**	.047	.047	.215*	.276**	.866**		
	Sig. (2-ta	.108	.133	.523	.455	.920	.000	.658	.661	.041	.008	.000		
	N	91	91	91	91	91	91	91	91	91	91	91		
S_con	Pearson	-.054	-.057	.067	-.132	.186	.174	.261*	.281**	.372**	.502**	.010	1	
	Sig. (2-ta	.608	.589	.530	.212	.078	.099	.012	.007	.000	.000	.926		
	N	91	91	91	91	91	91	91	91	91	91	91	91	
R_con	Pearson	-.054	-.054	.085	-.153	.151	.253*	.294**	.311**	.419**	.543**	.083	.985**	
	Sig. (2-ta	.609	.612	.425	.148	.154	.015	.005	.003	.000	.000	.433	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	
SMSP_d	Pearson	.024	.036	.372**	-.099	1.000**	.131	.261*	.242*	.313**	.382**	-.020	.186	
	Sig. (2-ta	.822	.731	.000	.353	0.000	.217	.012	.021	.002	.000	.850	.078	
	N	91	91	91	91	91	91	91	91	91	91	91	91	
ECin	Pearson	-.261*	-.276**	-.377**	.421**	-.187	-.410**	-.525**	-.511**	-.467**	-.427**	-.036	-.100	
	Sig. (2-ta	.013	.008	.000	.000	.076	.000	.000	.000	.000	.000	.731	.344	
	N	91	91	91	91	91	91	91	91	91	91	91	91	
PL_EVCinN	Pearson	.011	.017	.526**	-.094	-.054	.190	.170	.167	.185	.200	.237*	-.093	
	Sig. (2-ta	.919	.876	.000	.375	.611	.072	.108	.113	.079	.058	.024	.382	
	N	91	91	91	91	91	91	91	91	91	91	91	91	
EVCin_TSpinN	Pearson	-.219*	-.215*	-.002	.292**	-.154	-.174	-.297**	-.335**	-.192	-.298**	-.012	-.322**	.961**
	Sig. (2-ta	.037	.041	.981	.005	.144	.098	.004	.001	.069	.004	.909	.002	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

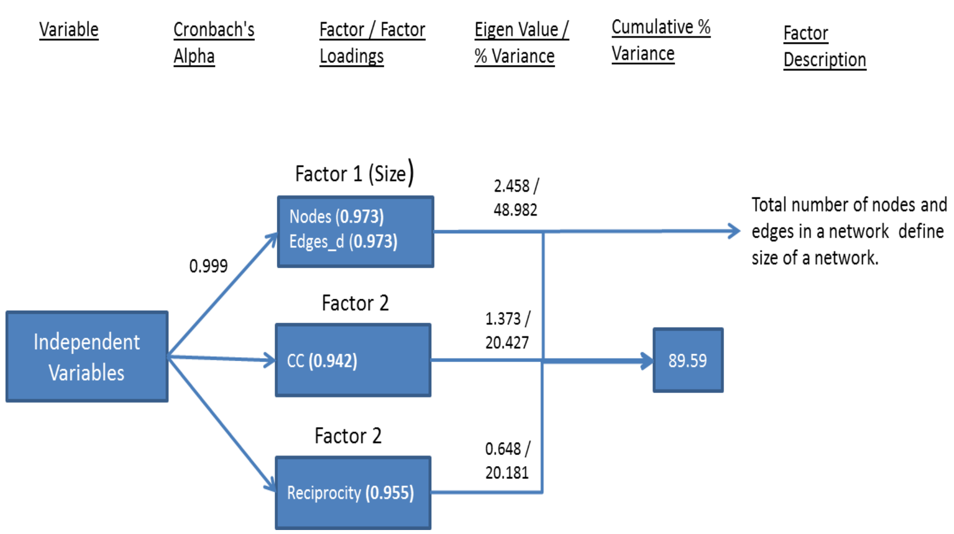
Tables 13 show that nodes (Nodes) and ties (Edges_ud) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes), number of ties (Edges_d), Reciprocity and Total Paths (Tpaths_d) in the network. Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpinN) correlates positively with Paths Power Law Distribution per Node (PL_TpinN). Scale Free Metric (S_con) seems to share a positive relationship with Average Geodesic Length (AvgGL_d). Assortativity (R_con) shares a positive relationship with Average Geodesic Length (AvgGL_d) and Scale Free Metric (S_con). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centralization (ECin) correlates negatively with Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN) and Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) correlate strongly with each other.

A.2.7.3.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.2.7.3.2.1 Independent Variables

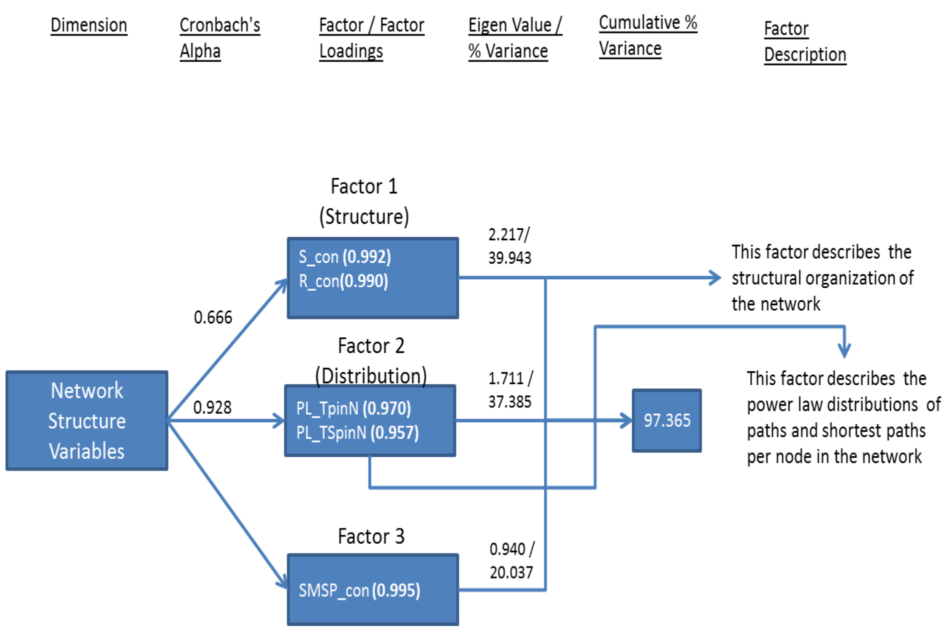
Figure 21: Factor Analysis of Independent Variables



Factor analysis generated three factors that explain 89.59% (greater than 80%) of cumulative variance. Factor 1 and factor2 have eigenvalue over 1. Factor3 has eigenvalue little less than 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had negative loading in factor 1, hence it was removed. Only Clustering Coefficient (CC_d) and Reciprocity have significant loadings in factor 2 and factor 3.Cronbach's alpha for factor 1 has a value of 0.999. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as "Size".

A.2.7.3.2.2 Network Structure (MV1)

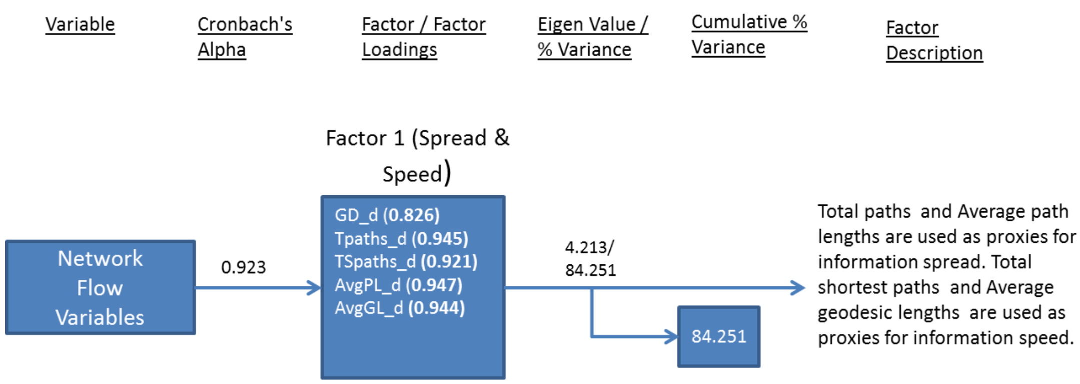
Figure 22: Factor Analysis of Network Structure Variables



Factor analysis generated three factors that explain 97.365% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 has eigenvalue little less than 1. Assortativity (R_con) and Scale Free Metric (SMSP_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.666. Assortativity (R_con) and Scale Free Metric (SMSP_d) are measuring same construct within factor 1. Hence, they should be considered as a factor. Paths Power Law Distribution per Node (PL_TpinN) and Shortest Paths Power Law Distribution per Node (PL_TpinN) have significant factor loadings in factor 2. Cronbach's alpha for factor1 has a value of 0.928. . Paths Power Law Distribution per Node (PL_TpinN) and Shortest Paths Power Law Distribution per Node (PL_TpinN) are measuring same construct within factor 2. Hence, they should be considered as a factor. All other variables load independently. Factor 1 is named as “Structure”. Factor 2 is named as “Distribution”.

A.2.7.3.2.3 Network Flow (MV2)

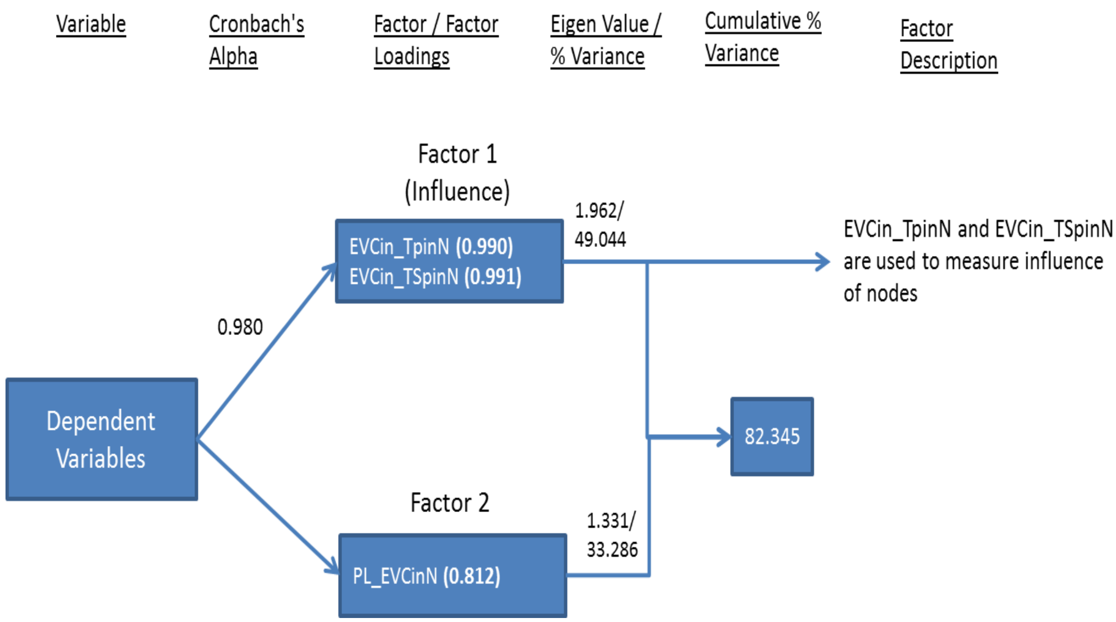
Figure 23: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 84.251% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach’s alpha for factor1 has a value of 0.923. Factor 1 is named as “Spread and Speed”.

A.2.7.3.2.4 Dependent Variables

Figure 24: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 82.345% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.98. I name factor1 as "Influence" as both, Eigenvector centralities with respect to paths and shortest paths, are being used measure of influence.

A.2.7.3.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Entertainment.pdf".

A.2.7.3.3.1 Impact of Network Structure on Network Flow

Table 14: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Entertainment	(0/137/0.000) [2]	(0.115/0.002) [4,5]	(0.116/0.002) [4,5]	(0.223/0.000) [4,5]	(0.373/0.000) [4,5]

Table 14 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 13.7%, 11.5%, 11.6%, 22.3% and 37.3% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.3.3.2 Impact of Network Flow on Network Structure

Table 15: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpinN	PL_TSpinN	S_con	R_con	SMSP_d
Entertainment	(0.181/0.000) [6,7]	(0.205/0.000) [6,7]	(0.345/0.000) [6,10]	(0.287/0.000) [10]	NA

Table 15 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 18.1%, 20.5%, 34.5%, and 28.7% variation in the PL_TpinN, PL_TSpinN, S_con and R_con, respectively.

A.2.7.3.3.3 Impact of Network Structure on Network Phenomenon

Table 16: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Entertainment	NA	(0.046/0.024)[1]	(0.055/0.014)[3]	(0.094/0.002)[3]

Table 16 shows that the network structure variable Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and variable Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN), explaining 4.6%, 5.5% and 9.4% variation respectively. The impact of network flow variables on Power Law Distribution of

Eigenvector Centrality per Node (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and variable Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 17: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Entertainment	(0.268/0.000) [7]	NA	(0.070/0.007)[8]	(0.128/0.000)[8]

Table 17 shows that the network structure variable impacts Eigenvector Centralization (EC_in), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and variable Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN), explaining 26.8%, 7% and 12.8% variation respectively. The impact of network flow variables on EVCin_TpinN are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.3.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 18: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d, (6)GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Entertainment	(0.325/0.000) [7,15]	(0.381/0.000) [1,14,15]	(0.092/0.005) [3,11]	(0.149/0.000) [3,8]

Table 18 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_in), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 32.5%, 38.1%, 9.2%and 14.9% variation respectively. The collective impact of independent variables and the moderating variables on Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.4 The Propagation Network

A.2.7.4.1 Correlation Analysis

Significant correlations coefficients for propagation network are shown below in table 19. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 19: Correlation Coefficients of Directed Network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	PL_TpoutN	EVCout_TpoutN
Edges_d	Pearson	.999**	1									
	Sig. (2-ta	.000										
	N	91	91									
Den_d	Pearson	-.572**	-.576**	.027	1							
	Sig. (2-ta	.000	.000	.797								
	N	91	91	91	91							
Tpaths_d	Pearson	.644**	.662**	.249*	-.756**	.261*	.663**	1				
	Sig. (2-ta	.000	.000	.017	.000	.012	.000					
	N	91	91	91	91	91	91	91				
TSpaths_d	Pearson	.680**	.697**	.190	-.803**	.242*	.641**	.988**	1			
	Sig. (2-ta	.000	.000	.071	.000	.021	.000	.000				
	N	91	91	91	91	91	91	91	91			
AvgPL_d	Pearson	.261*	.283**	.410**	-.415**	.313**	.755**	.850**	.785**	1		
	Sig. (2-ta	.013	.007	.000	.000	.002	.000	.000	.000			
	N	91	91	91	91	91	91	91	91	91		
AvgGL_d	Pearson	.216*	.236*	.359**	-.455**	.382**	.765**	.820**	.796**	.944**		
	Sig. (2-ta	.040	.025	.000	.000	.000	.000	.000	.000	.000		
	N	91	91	91	91	91	91	91	91	91		
PL_TSpoutN	Pearson	.028	.035	-.077	-.389**	.021	.424**	.248*	.268*	.211*	.703**	
	Sig. (2-ta	.793	.740	.471	.000	.843	.000	.018	.010	.045	.000	
	N	91	91	91	91	91	91	91	91	91	91	
SMSP_d	Pearson	.024	.036	.372**	-.099	1.000**	.131	.261*	.242*	.313**	-.001	
	Sig. (2-ta	.822	.731	.000	.353	0.000	.217	.012	.021	.002	.996	
	N	91	91	91	91	91	91	91	91	91	91	
ECout	Pearson	-.261*	-.276**	-.377**	.421**	-.187	-.410**	-.525**	-.511**	-.467**	-.165	
	Sig. (2-ta	.013	.008	.000	.000	.076	.000	.000	.000	.000	.118	
	N	91	91	91	91	91	91	91	91	91	91	
PL_EVCoutN	Pearson	.094	.111	.572**	-.168	.246*	.381**	.319**	.299**	.358**	.193	
	Sig. (2-ta	.375	.294	.000	.112	.019	.000	.002	.004	.000	.067	
	N	91	91	91	91	91	91	91	91	91	91	
EVCout_TSpoutN	Pearson	.260*	.247*	.014	-.370**	.158	-.037	.312**	.347**	.151	-.076	.998**
	Sig. (2-ta	.013	.018	.894	.000	.134	.728	.003	.001	.152	.474	.000
	N	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

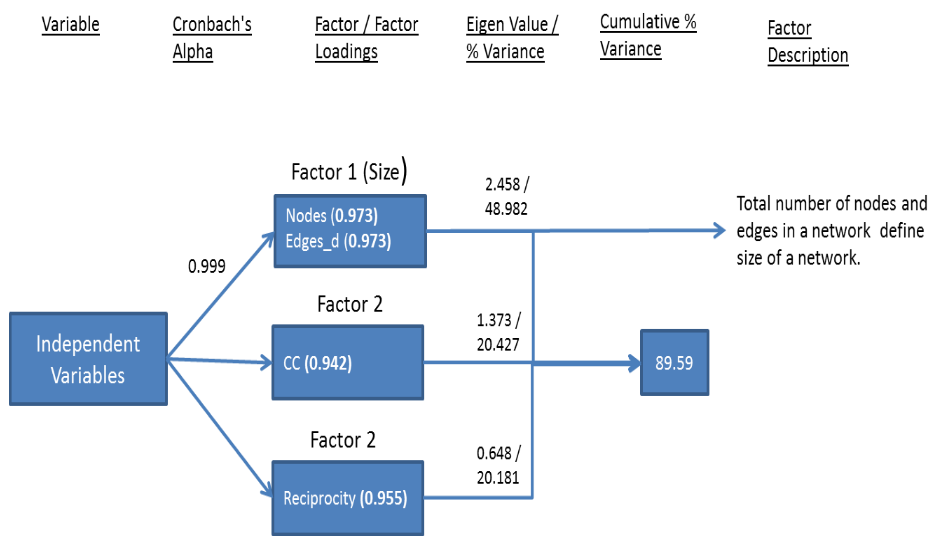
Table 19 shows that nodes and ties have a strong positive correlation. As the number of nodes (Nodes) increase, the number of ties (Edges_d) also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes), number of ties (Edges_d), Reciprocity and Total Paths (Tpaths_d) in the network. Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpoutN) correlates positively with Paths Power Law Distribution per Node (PL_TpoutN). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centralization (ECout) correlates negatively Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN) and Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) correlate strongly with each other.

A.2.7.4.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.2.7.4.2.1 Independent Variables

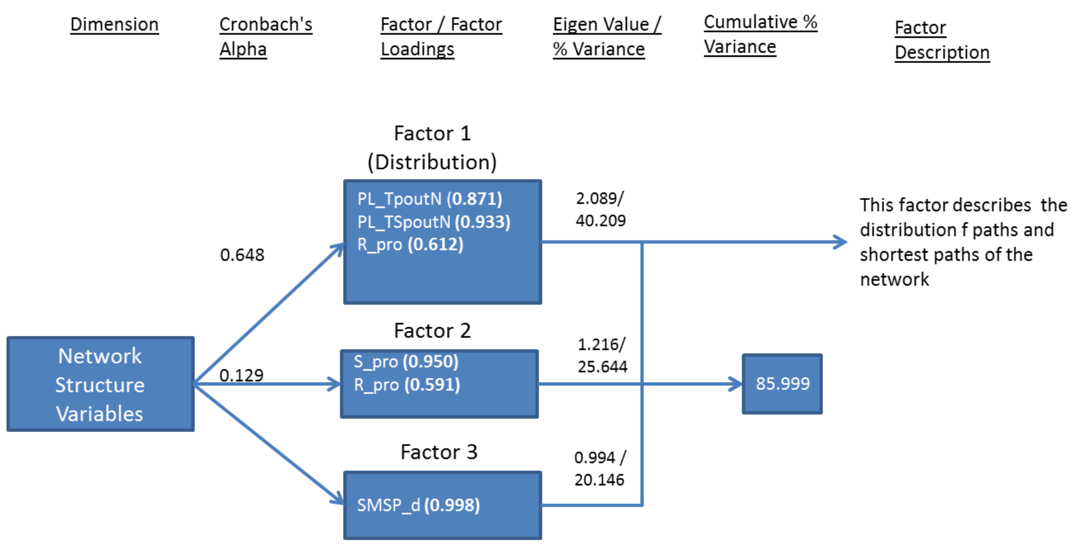
Figure 25: Factor Analysis of Independent Variables



Factor analysis generated three factors that explain 89.59% (greater than 80%) of cumulative variance. Factor 1 and factor 2 have eigenvalues over 1. Factor3 has an eigenvalue that is little less than 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had negative loading in factor 1, hence it was removed. Only Clustering Coefficient (CC_d) and Reciprocity have significant loadings in factor 2 and factor 3.Cronbach’s alpha for factor 1 has a value of 0.999. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as "Size".

A.2.7.4.2.2 Network Structure (MV1)

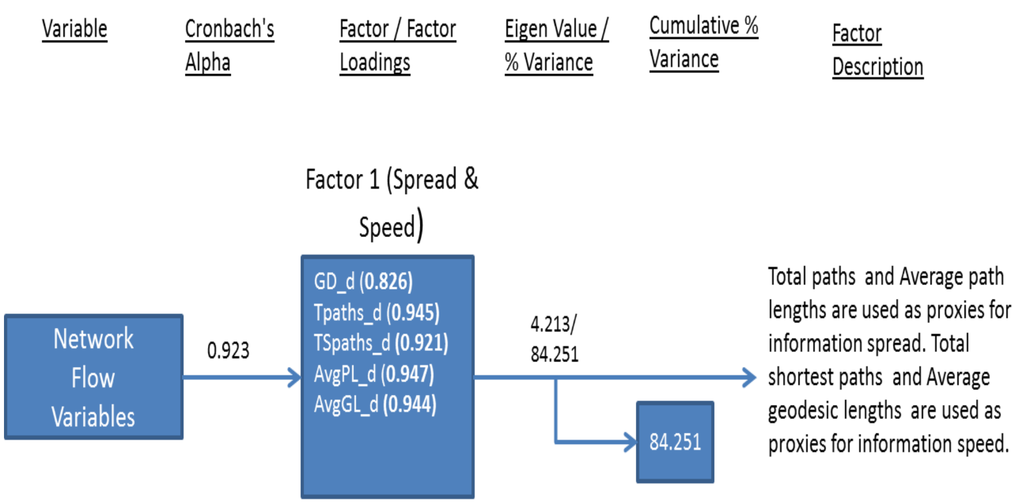
Figure 26: Factor Analysis of Network Structure Variables



Factor analysis generated three factors that explain 85.99% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 has an eigenvalue that is little less than 1. Assortativity (R_pro), Power Law Distribution of Paths per Node (PL_TpoutN) and Power Law Distribution of Shortest Paths per Node (PL_TSpoutN) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.648. Assortativity (R_pro) and Scale Free Metric (S_pro) have significant factor loadings in factor2. Cronbach's alpha for factor1 has a value of 0.129. Assortativity (R_pro), Power Law Distribution of Paths per Node (PL_TpoutN) and Power Law Distribution of Shortest Paths per Node (PL_TSpoutN) are measuring same construct within factor 1. Hence, they should be considered as a factor. All other variables load independently. Factor1 is named as “Structure”. Assortativity (R_pro) and Scale Free Metric (S_pro) are not measuring same construct within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.2.7.4.2.3 Network Flow (MV2)

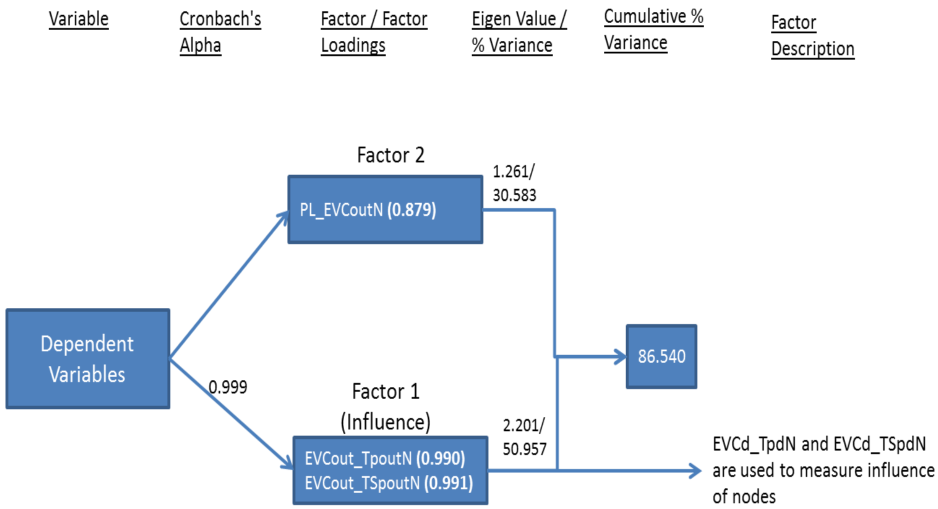
Figure 27: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 84.251% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach’s alpha for factor1 has a value of 0.923. Factor 1 is named as “Spread and Speed”.

A.2.2.1.4.2.4 Dependent Variables

Figure 28: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 86.54% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.999. I name factor1 as "Influence" as both, Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN), are being used measure of influence.

A.2.7.4.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Entertainment.pdf"..

A.2.7.4.3.1 Impact of Network Structure on Network Flow

Table 20: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Entertainment	(0.227/0.000) [1,4]	(0.107/0.003) [2,5]	(0.143/0.001) [2,3,5]	(0.120/0.001) [2,5]	(0.197/0.000) [2,5]

Table 20 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 22.7%, 10.7%, 14.3%, 12% and 19.7% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Total Paths (Tpaths_d) Total Shortest Paths (TSpaths_d) and Average Path Length (AvgPL_d) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.2.7.4.3.2 Impact of Network Flow on Network Structure

Table 21: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpoutN	PL_TSpoutN	S_pro	R_pro	SMSP_d
Entertainment	(0.214/0.000) [6,7]	(0.171/0.000) [6]	(0.204/0.000) [6,8]	(0.315/0.000) [6,8]	NA

Table 21 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 21.4%, 17.1%, 20.4%, and 31.5% variation in the PL_TpoutN, PL_TSpoutN, S_pro, and R_pro, respectively.

A.2.7.4.3.3 Impact of Network Structure on Network Phenomenon

Table 22: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Entertainment	NA	(0.128/0.001) [4,5]	(0.422/0.000) [2,4]	(0.432/0.000) [2,4]

Table 22 shows that the network structure variable impacts Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector

Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 12.8%, 42.2% and 43.2% variation respectively.

A.2.7.4.3.4 Impact of Network Flow on Network Phenomenon

Table 23: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Entertainment	(0.268/0.000) [7]	(0.136/0.000) [6]	(0.226/0.000) [6,8]	(0.217/0.000) [6,8]

Table 23 shows that the network structure variable impacts Eigenvector Centralization (ECout), Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 26.8%, 13.6%, 22.6% and 21.7% variation respectively.

A.2.7.4.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 24: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

<u>Predictors</u>: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d, (6),GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Entertainment	(0.325/0.000) [7,15]	(0.413/0.000) [2,15]	(0.466/0.000) [4,13]	(0.495/0.000) [1,4,13]

Table 24 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_out), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 32.5%, 41.3%, 46.6%and 49.5% variation respectively.

A.3 Case 3--Comedy

A.3.1 Case Overview

Data for keyword "YouTube + comedy" was collected over a period of 91 days (31/12/2013 to 31/03/2014). As shown in table 9, overall 94,111 tweets were collected, out of which 33,350 were broadcast tweets and 60,761 were engaged tweets respectively. Out of 60,761 engaged tweets only 25,624 tweets formed the largest community. Similarly, 83,175 daily unique people tweeted overall, out of which 37,456 daily unique people were engaged in broadcast activity whereas 45,719 daily unique people were engaged in conversations. Out of 45,719 daily unique people only 24,555 daily unique people formed the largest community. Data for the largest community was analyzed at a daily interval. The overall trends for the comedy data are shown below in figure 1 and figure 2.

Figure 1: Overall Tweets

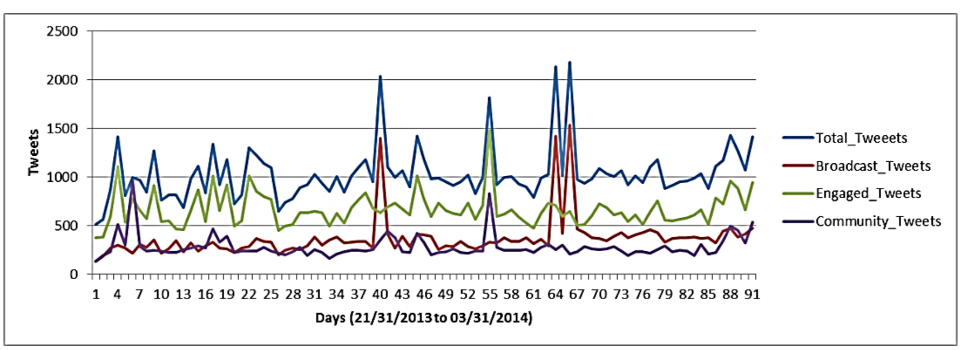


Figure 2: Overall People

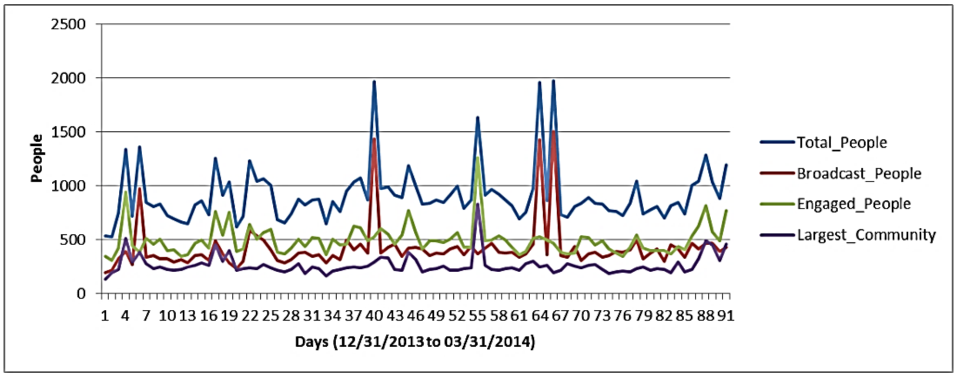


Figure 1 and figure 2 shows that both the total tweets and total people involved are very dynamic and their magnitude changes on a daily basis. The maximum of the total number of daily tweets and the maximum of the total number of daily unique people observed on a single day (the daily uniques) are 2,178 and 1,968, respectively. Similarly, the minimum of the total number daily tweets and the minimum of the number daily unique are 508 and 526, respectively. The size of the largest community on a particular day and the largest number of community tweets on that day also seem to follow the trend of total people and total tweets. The largest number of daily community tweets and the largest number of daily unique people are 832 and 833, respectively. Similarly, the smallest number of daily community tweets and the smallest number of daily unique people are 131 and 130, respectively. As the total number of daily unique people tweeting increases, so does the size of the community.

Most of the engaged people are engaged in the collective conversation forming the largest community.

A.3.2 Random or Not Random

As explained in section 4.4.1, in order to eliminate α- error and β- error, I compare the clustering coefficients of both undirected (CC_ud) and directed networks (CC_d) with their corresponding random (Erdös-Rényi, E-R) networks (CCudran, CCdran). If the clustering coefficients of the undirected and directed networks are equal to those of the E-R random network, then the directed and undirected networks are considered to be random. If they are not equal, then they are not random.

Figure 3: Comparison of Clustering Coefficients of Undirected Network with E-R Networks

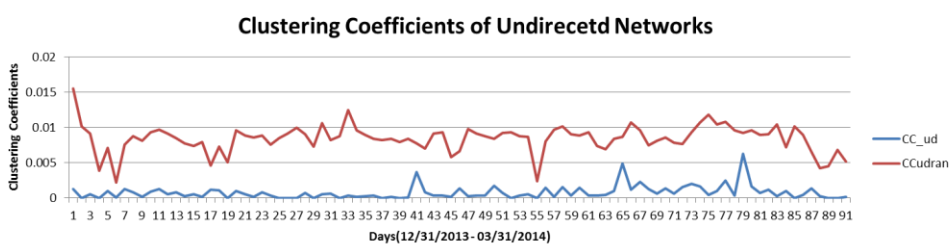
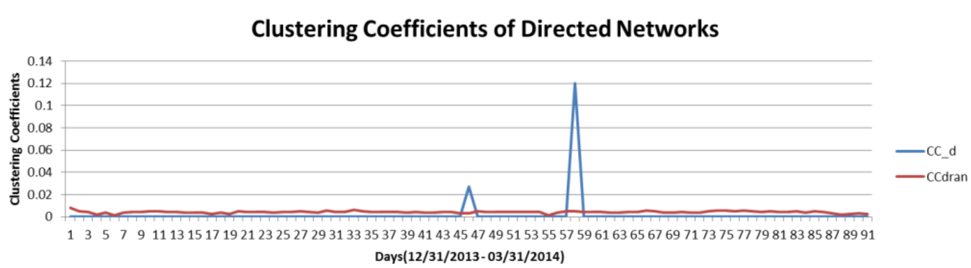


Figure 4: Comparison of Clustering Coefficients of Directed Network with E-R Networks



As seen in figure 3 and figure 4 clustering coefficients of the undirected networks follows a very different pattern from their corresponding E-R networks. Therefore, the undirected network is considered to be non-random networks and the variables computed are a true reflection of network's features. For the direct network the clustering coefficients (CC_d) is zero for the most part. Therefore, the directed networks are random.

A.3.3. Independent Variables

The values of the independent variables for both the undirected and the directed network are shown in figure 5 below.

Figure 5: Independent Variables--(a) Nodes and Edges (Undirected and Directed networks), (b) Reciprocity (Directed Networks), (c) Density (Undirected and Directed Networks), (d) Clustering Coefficient Undirected Network, (e) Clustering Coefficient Directed Network.

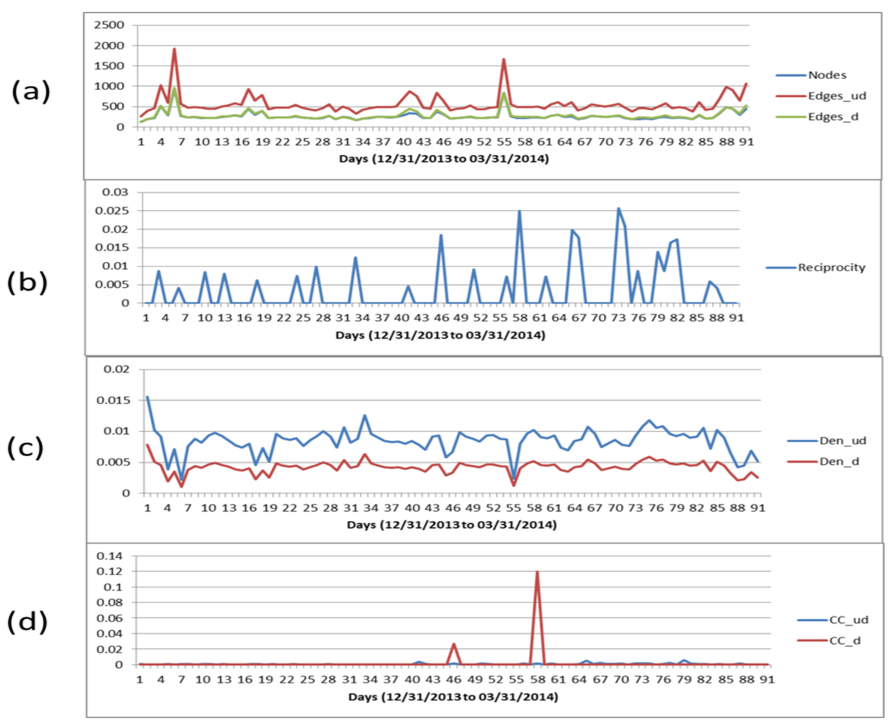


Figure 5(a) shows that the number of directed ties (Edges_d) in the network and the total number of nodes (Nodes) overlap with each other. The numbers of undirected ties (Edges_ud) is greater than the number of directed ties (Edges_d), because in an undirected network every directed tie is considered to be symmetric. Therefore it is counted twice, except for the ones that are symmetric in a directed network. Reciprocity in Figure 5(b) indicates the presence of symmetric ties in a directed network (in an undirected network 100% are symmetric). The value of 0.01 is equal to 1% of all the ties. Figure 5(c) shows the difference between the densities of the undirected (Den_ud) and the directed networks (Den_d). The undirected network is denser than the directed network (Den_ud>Den_d). Figure 5(d) shows that the directed networks have higher Clustering Coefficients than the undirected networks (CC_d >CC_ud).

A.3.4 Network Structure Variables (MV1)

A.3.4.1 The Scale Free Metric

Figure 6: Scale Free Metric--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

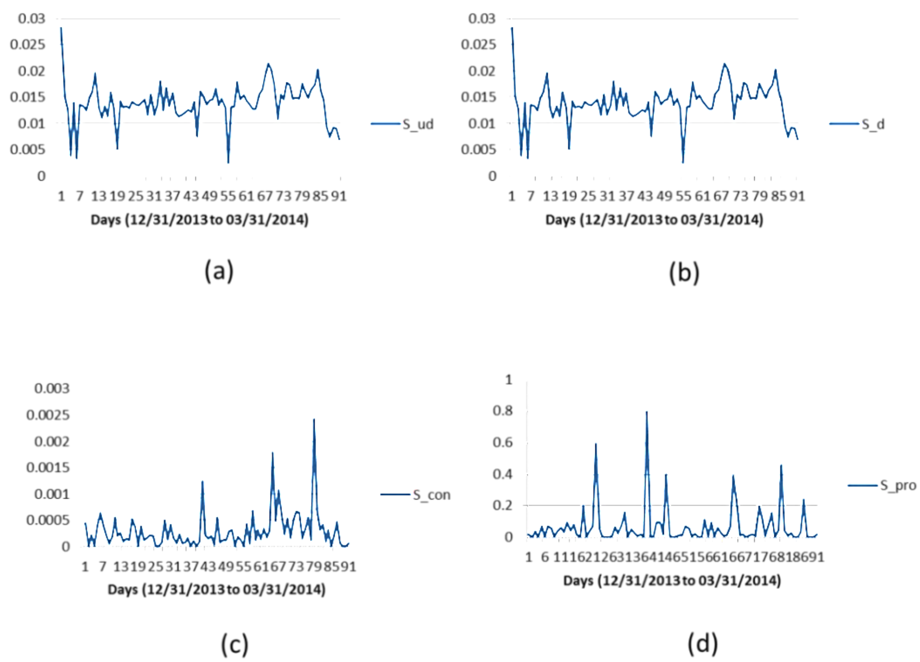
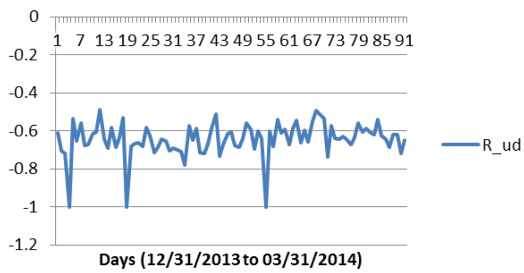


Figure 6 shows the Scale Free Metric for the undirected, directed, consumption and propagation networks (S_ud, S_d, S_con, S_pro). The Scale Free Metrics for the undirected (S_ud) and the directed network (S_d) are similar, but the Scale Free Metrics for the consumption (S_con) and propagation (S_pro) networks are very different. The propagation network is more scale free than the consumption network (S_pro >S_con). The values of the scale free metric ranges between 0 and 1. When the values are closer to 1, it means that the networks are more scale free. Neither the directed (S_d) nor the undirected network (S_ud) is scale free. This means that these networks may have hubs in them.

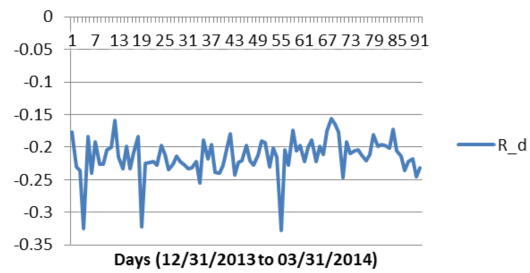
However, there is not just one hub that is the center of the community. As shown in figure 6 (c) the propagation network is more scale free than the consumption network shown in figure 6 (d).

A.3.4.2 The Assortativity

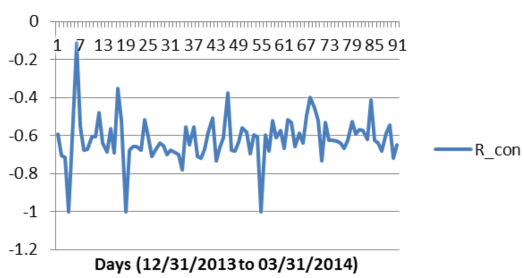
Figure 7: Assortativity--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



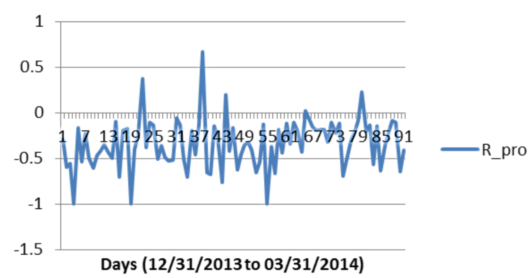
(a)



(b)



(c)



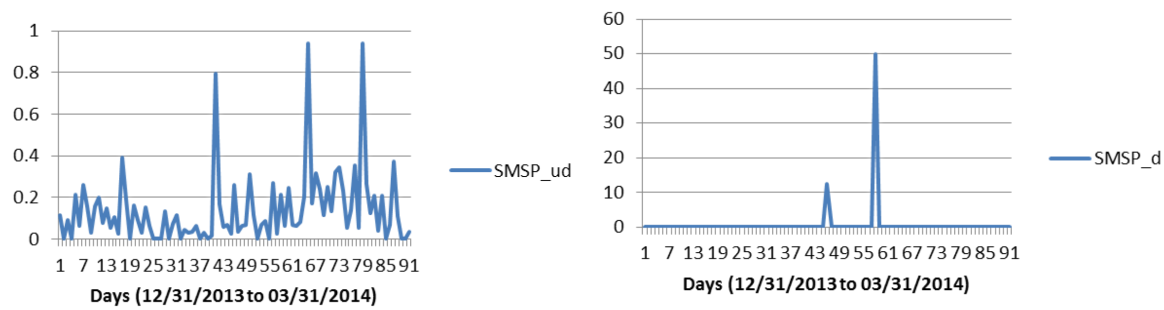
(d)

Figure 7 shows the assortativity metric for the undirected, directed, consumption and propagation networks (R_ud, R_d, R_con, R_Pro). The value of the assortativity metric ranges between -1 and 1. When the values are closer to -1, it means that networks are disassortative. The undirected network is more Disassortative than

the directed network (R_d > R_ud). Among the directed networks, the consumption network is more Disassortative than the propagation network (R_pro > R_con). Disassortative means that the nodes in the network connect to nodes that are very similar to themselves. This is true more so in the undirected network and in the consumption network than it is in the directed network and the propagation network. This implies that disassortativeness of consumption contributes more to the disassortativeness of the directed network than the disassortativeness of the propagation does.

A.3.4.3 The Small World Metric

Figure 8: Small World Metric -- (a) Undirected Network, (b) Directed Network.



(a) (b)

Figure 8 shows the Small World Metric for the undirected (SMSP_ud) and directed networks (SMSP_d). The Small World Metrics for the consumption and propagation networks are the same as the ones for the directed network. The directed networks don't show any small world behavior. Contrary to the directed networks, undirected networks show some small world behavior but not significantly enough. This

means that in undirected networks there are more nodes that act as hubs that facilitate communication between other nodes of the network.

A.3.4.4 Paths and Shortest Paths Power law Distribution per Node

Figure 9: Power Law Distribution of Paths and Shortest Paths in (a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

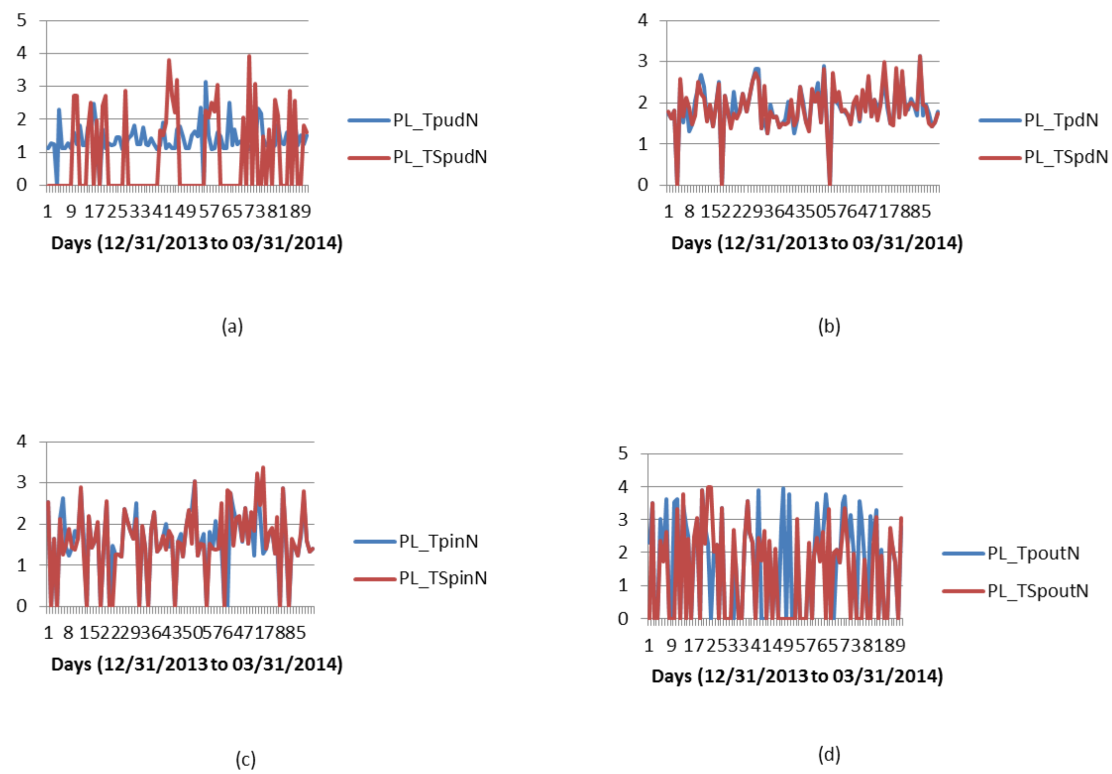


Figure 9 (a) shows that, in the undirected network, paths are more uniformly distributed among nodes than shortest paths are distributed among nodes. This means that fewer nodes are responsible for more of the shortest paths in the undirected network. There are fewer instances of shortest path following power law distribution in undirected (figure 9 (a)) and consumption (figure 9 (c)) networks. In the directed (figure 9 (b)) and propagation (figure 9 (d)) networks, there are no such patterns.

A.3.5 Network Flow Variables (MV2)

Figure 10: Network Flow Variables-- (a) Total Paths and Total Shortest Paths, (b) Average Paths and Average Shortest Paths, (c) Undirected and Directed Network Graph Diameter.

(a)

(b)

(c)

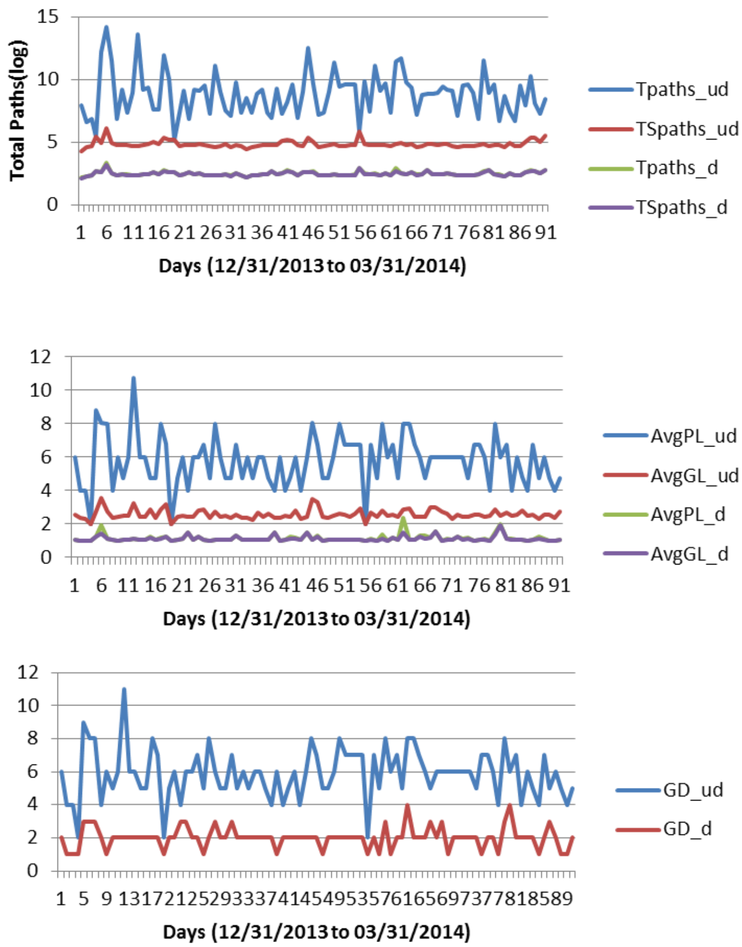


Figure 10 (a), shows that total number of paths in the undirected network (Tpaths_ud) is orders of magnitude higher than the total number of shortest paths (TSpaths_ud). The total number of paths (Tpaths_d) and the total number of shortest paths (TSpaths_d) map more closely in the directed network. In figure 10 (b), a similar trend is observed in the Average Path Lengths (AvgPL_ud, AvgPL_d) and the Average

Geodesic Lengths (AvgGL_ud, AvgGL_d) of the undirected and directed networks. In figure 10 (c), the Graph Diameter (GD_ud) of the undirected network is larger than the graph diameter of the directed network (GD_d). It is also noteworthy that, in figure 10 (b) and in figure 10 (c), the Graph Diameter (GD_ud, GD_d) and the Average Path Length (AvgPL_ud, AvgPL_d) of the undirected and directed networks track pretty closely.

A.3.6 Dependent Variables

A.3.6.1 Eigenvector Centralization

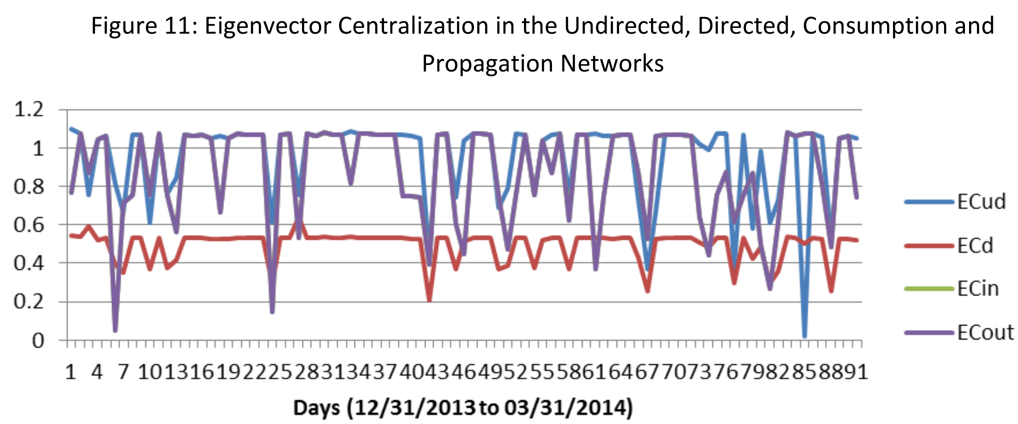


Figure 11 shows that nodes with influence are lot more central in the undirected and propagation networks than in the directed network (ECud > ECd). The consumption and propagation networks exhibit same level of centralization.

A.3.6.2 Power law Distribution of Eigenvector Centrality per Node

Figure 12: Power Law Distribution of Eigenvector Centrality in Undirected, Directed, Consumption and Propagation Network

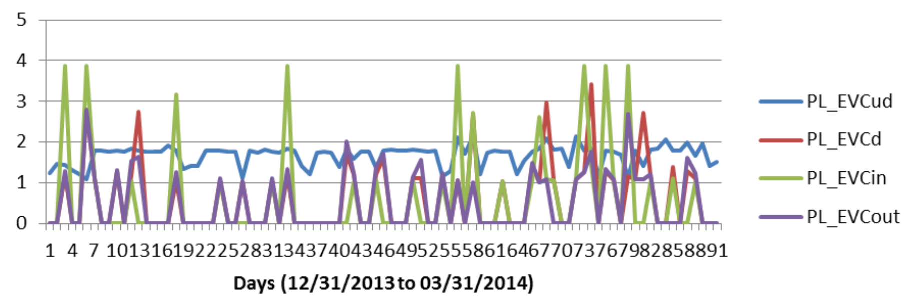
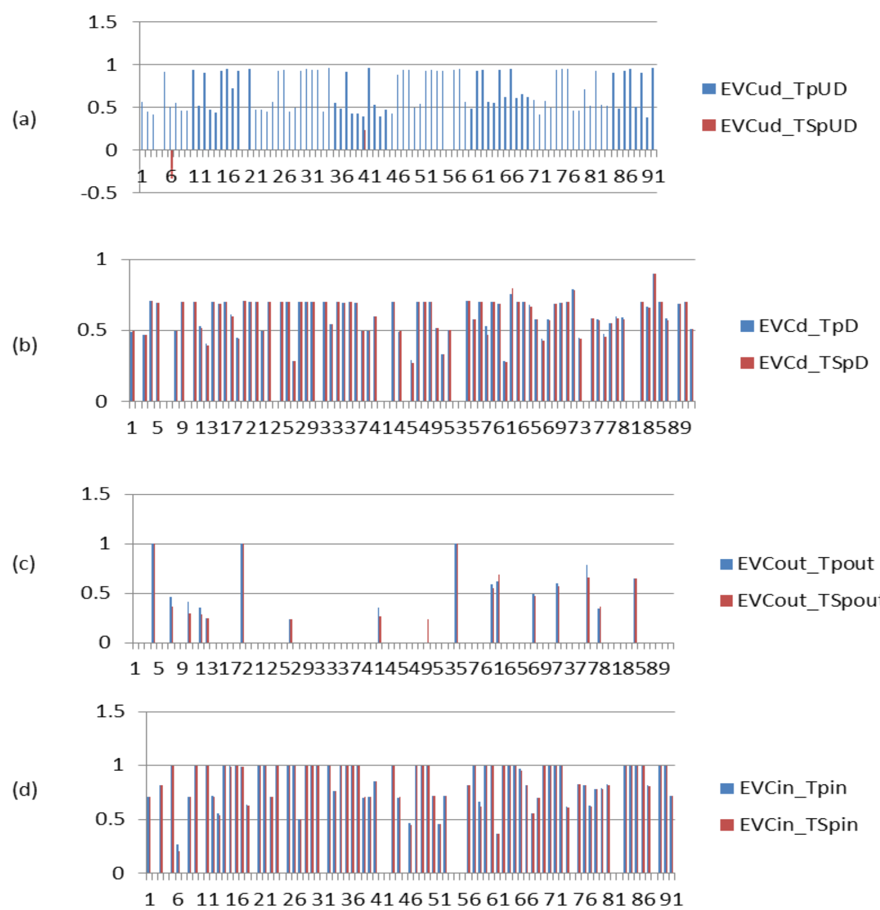


Figure 12 shows that in the undirected network eigenvector centrality values are consistently distributed in a power law distribution pattern (PL_EVCud), over a period of time. In the directed, the consumption and the propagation network the distribution of eigenvector centrality follows a power law distribution (PL_EVCd, PL_EVCin, PL_EVCout) pattern only sometimes.

A.3.6.3 Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node

Figure 13: Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



In figure 13, only those correlation coefficients with a significance value lower than 0.05 are shown. In figure 13 (a), there is a significant correlation between the eigenvector centrality of a node and the number of paths from a node in undirected network (EVCud_TpUDN). There is no significant correlation between eigenvector centrality of a node and shortest paths from a node in undirected network (EVCud_TSpUDN). In figure 13 (b), there is a significant correlation between the directed-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the directed network (EVCd_TpDN, EVCud_TSpUDN). In figure 13 (c), there is a significant correlation between the in-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the consumption network (EVin_TpinN, EVCin_TSpinN). The correlation between the out-eigenvector centrality of a node and the number of shortest paths is less significant figure 13 (d) (EVCout_TpoutN, EVCout_TSpoutN).

A.3.7 Statistical Analysis

A.3.7.1 The Undirected Network

A.3.7.1.1 Correlation Analysis

In Table 1, the statistically significant Correlation Coefficients for the undirected network are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 1: Correlation Coefficients of Undirected Network

Correlations

		Nodes	Edges_ud	Den_ud	CC_ud	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud	PL_TpudN	S_ud
Edges_ud	Pearson	.991**	1									
	Sig. (2-ta	.000										
	N	91	91									
Den_ud	Pearson	-.850**	-.840**	1								
	Sig. (2-ta	.000	.000									
	N	91	91	91								
Tpaths_ud	Pearson	.159	.173	-.155	.227*	.943**	1					
	Sig. (2-ta	.133	.100	.143	.031	.000						
	N	91	91	91	91	91	91					
TSpaths_ud	Pearson	.946**	.964**	-.923**	-.085	-.078	.194	1				
	Sig. (2-ta	.000	.000	.000	.420	.465	.065					
	N	91	91	91	91	91	91	91				
AvgPL_ud	Pearson	-.097	-.085	.082	.249*	.996**	.958**	-.066	1			
	Sig. (2-ta	.363	.422	.438	.017	.000	.000	.536				
	N	91	91	91	91	91	91	91	91			
AvgGL_ud	Pearson	.201	.225*	-.208*	.130	.696**	.730**	.286**	.697**	1		
	Sig. (2-ta	.056	.032	.048	.221	.000	.000	.006	.000			
	N	91	91	91	91	91	91	91	91	91		
S_ud	Pearson	-.688**	-.681**	.753**	.335**	.316**	.083	-.730**	.312**	.183	.296**	1
	Sig. (2-ta	.000	.000	.000	.001	.002	.436	.000	.003	.082	.004	
	N	91	91	91	91	91	91	91	91	91	91	91
R_ud	Pearson	-.318**	-.272**	.231*	.355**	.645**	.549**	-.211*	.643**	.667**	.435**	.620**
	Sig. (2-ta	.002	.009	.027	.001	.000	.000	.045	.000	.000	.000	.000
	N	91	91	91	91	91	91	91	91	91	91	91
SMSP_ud	Pearson	-.035	.034	.037	.971**	.234*	.249*	.020	.244*	.131	.178	.249*
	Sig. (2-ta	.740	.749	.725	.000	.026	.017	.849	.020	.217	.091	.017
	N	91	91	91	91	91	91	91	91	91	91	91
EVCud_TpudN	Pearson	-.277**	-.250*	.214*	.104	.327**	.113	-.216*	.250*	.236*	.777**	.291**
	Sig. (2-ta	.008	.017	.041	.327	.002	.286	.040	.017	.024	.000	.005
	N	91	91	91	91	91	91	91	91	91	91	91
EVCud_TSpudN	Pearson	-.506**	-.467**	.281**	.019	-.199	-.326**	-.362**	-.203	-.331**	.033	.211*
	Sig. (2-ta	.000	.000	.007	.855	.059	.002	.000	.054	.001	.756	.044
	N	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

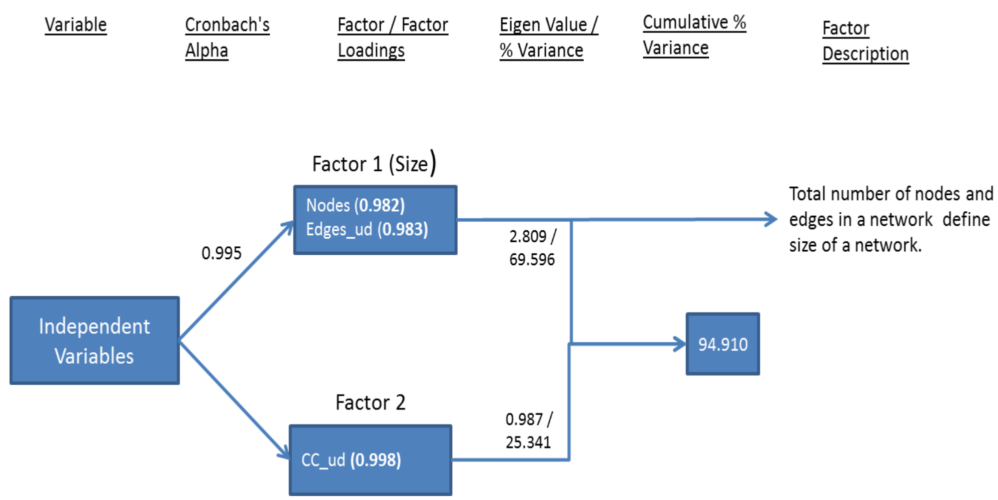
In Table 1, the number of nodes (Nodes) and the number of ties (Edges_ud) have a strong positive correlation. As the number of nodes (Nodes) increases, the number of ties (Edges_ud) also increases. The Density (Den_ud) of this network has a strong negative correlation with both the number of nodes (Nodes) and the number of ties (Edges_ud). Total Paths (Tpaths_ud) have a strong positive correlation with Graph Diameter (GD_ud). The Total Number of Shortest Paths (TSpaths_ud) correlates strongly with the number of nodes (Nodes) and the number of ties (Edges_ud), but it correlates negatively with Density (Den_ud). Average Path Length (AvgPL_ud) and Average Geodesic Length (AvgGL_ud) share a strong positive correlation with Graph Diameter (GD_ud) and Total Paths (Tpaths_ud). Average Geodesic Length (AvgGL_ud) shares a strong correlation with Average Path Length (AvgPL_ud). Scale Free Metric (S_ud) shares a positive correlation with Density (Den_ud) and negative correlations with number of number of nodes (Nodes), the number of ties (Edges_ud) and Total Number of Shortest Paths (TSpaths_ud). Assortativity (R_ud) shares positive correlations with Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Average Path Length (AvgPL_ud), Average Geodesic Length (AvgGL_ud) and Scale Free Metric (S_ud). Small World (SMSP_ud) metric share a positive relationship with the Clustering Coefficients (CC_ud). Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and correlate strongly Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN). Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN) correlates negatively with number of nodes.

A.3.7.1.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.3.7.1.2.1 Independent Variables

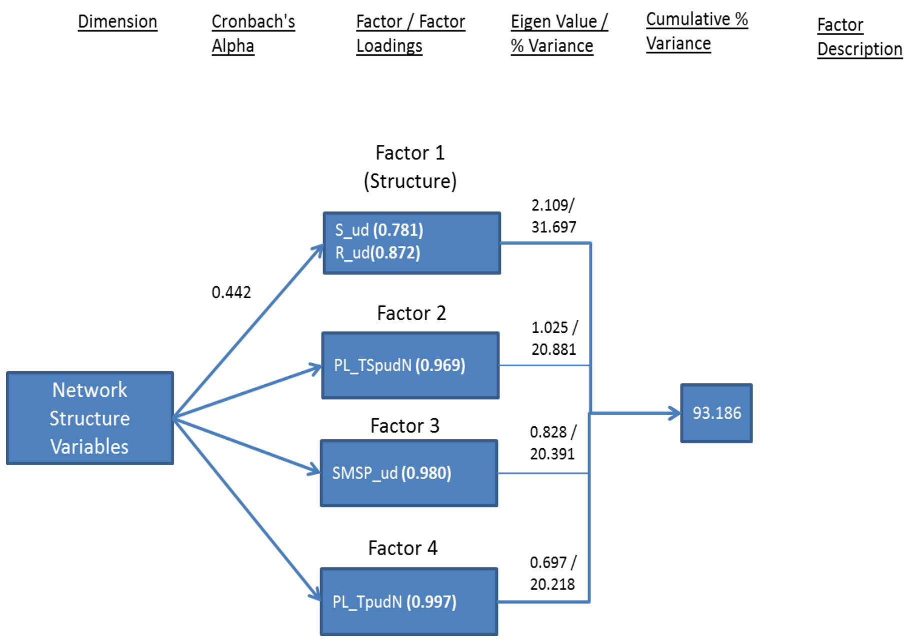
Figure 14: Factor Analysis Independent Variables Music Undirected Network



The factor analysis generated two factors that explain 94.91% (greater than 80%) of the cumulative variance. Both factors have eigenvalues above one. Nodes and ties (Edges_ud) have significant factor loadings in factor 1. Density (Den_ud) had a negative loading in factor 1, hence it was removed. Only the Clustering Coefficient (CC_ud) has a significant loading in factor 2.Cronbach's alpha for factor 1 has a value of 0.995. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as "Size".

A.3.7.1.2.2 Network Structure (MV1)

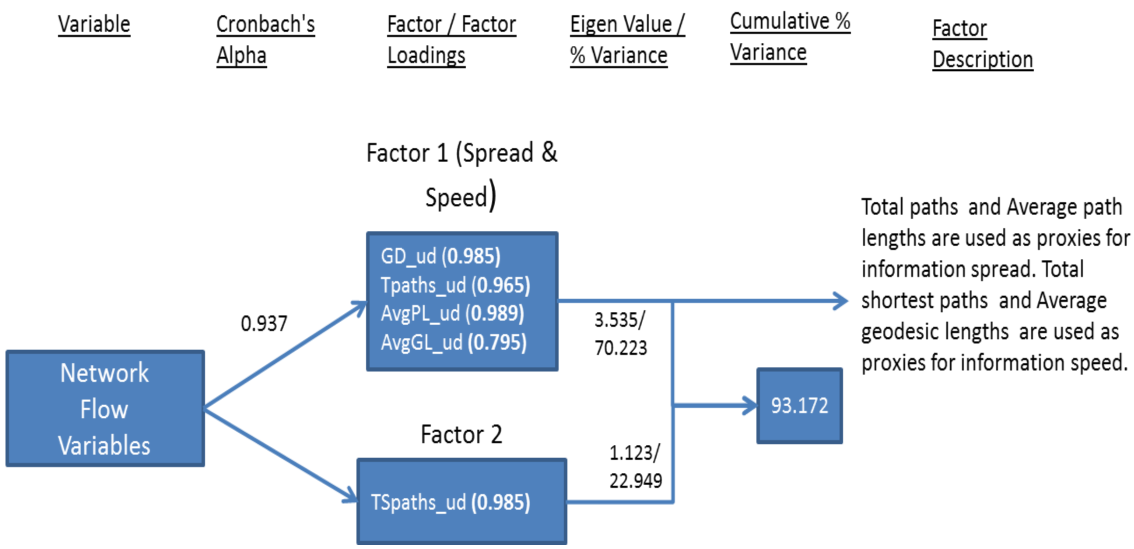
Figure 15: Factor Analysis of Network Structure Variables



The factor analysis generated four factors that explain 93.186% (greater than 80%) of the cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 and factor4 have eigenvalues below 1. Scale Free Metric (S_ud) and Assortativity (R_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.442. Scale Free Metric (S_ud) and Assortativity (R_ud) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.3.7.1.2.3 Network Flow (MV2)

Figure 16: Factor Analysis of Network Flow Variables



The factor analysis generated two factors that explain 93.172% (greater than 80%) of the cumulative variance. Factor1 has eigenvalue above 1. Factor2 has eigenvalue below 1. Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Average Geodesic Length (AvgGL_ud) and Average Path Length (AvgPL_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.937. Hence, they should be considered as a factor.

A.3.7.1.2.4 Dependent Variables

The value of Kaiser-Meyer-Olkin measure of sampling adequacy was 0.475 (less than 0.5), and the significance Bartlett's test of sphericity is 0.133. This data does not satisfy the measure of appropriateness for factor analysis. Therefore, all the variables are considered independently.

A.3.7.1.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Comedy.pdf".

A.3.7.1.3.1 Impact of Network Structure on Network Flow

Table 2: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud
Comedy	(0.410/0.000) [4]	(0.396/0.000) [3,4]	(0.634/0.000) [3,4,5]	(0.407/0.000) [4]	(0.521/0.000) [3,4]

Table 2 shows that the network structure variables have a significant impact on the network flow variables. Network structure variables explain 41%, 39.6%, 63.4%, 40.7% and 52.1% variation in Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud), Average Path Length (AvgPL_ud) and Average Geodesic Length (AvgGL_ud), respectively.

A.3.7.1.3.2 Impact of Network Flow on Network Structure

Table 3: Impact of Network Flow on Network Structure

Predictors: **(6) GD_ud, (7) Tpaths_ud, (8) TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpudN	PL_TSpudN	S_ud	R_ud	SMSP_ud
Comedy	(0.138/0.000) [6]	NA	(0.694/0.000) [8,10]	(0.612/0.000) [8,10]	(0.052/0.017) [7]

Table 3 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 4%, 3.2%, and 16.7% variation in the PL_TpudN, S_ud R_ud, and SMSP_ud respectively. The impact of network flow variables on SMSP_ud is not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.1.3.3 Impact of Network Structure on Network Phenomenon

Table 4: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Comedy	(0.105/0.003) [1,4]	(0.082/0.004) [4]	(0.640/0.000) [1]	NA

Table 4 shows that the network structure variable Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 10.5%, 8.2% and 6.4% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud) and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.1.3.4 Impact of Network Flow on Network Phenomenon

Table 5: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_ud, (7)Tpaths_ud, (8)TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Comedy	(0.076/0.005) [10]	(0.033/0.048) [8]	(0.097/0.000) [6]	NA

Table 5 shows that the network flow variable impacts Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 7.6%, 3.3% and 9.7% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud), and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.1.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 6: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud, (6),GD_ud (7) Tpaths_ud (8), TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud, (11) Nodes, (12) Edges_ud, (13) Den_ud, (14) CC_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Comedy	(0.157/0.000) [12,4,2]	(0.045/0.024) [10]	(0.046/0.024) [7]	NA

Table 6 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 15.7%, 4.5% and 4.6% variation respectively. The collective impact of independent variables and the moderating variables on Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.2 The Directed Network

A.3.7.2.1 Correlation Analysis

Significant Correlations Coefficients for directed network are shown below in table 7. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 7: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	PL_TpdN	PL_TSpdN	S_d	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Edges_d	Pearson	.990**	1														
	Sig. (2-ta	.000															
	N	91	91														
Den_d	Pearson	-.849**	-.838**	.209*	1												
	Sig. (2-ta	.000	.000	.047													
	N	91	91	91	91												
Tpaths_d	Pearson	.803**	.824**	.028	-.763**	.030	.352**	1									
	Sig. (2-ta	.000	.000	.789	.000	.779	.001										
	N	91	91	91	91	91	91	91									
TSpaths_d	Pearson	.827**	.846**	-.030	-.807**	-.008	.278**	.987**	1								
	Sig. (2-ta	.000	.000	.780	.000	.937	.008	.000									
	N	91	91	91	91	91	91	91	91								
AvgPL_d	Pearson	.201	.206	.249*	-.139	.124	.694**	.674**	.584**	1							
	Sig. (2-ta	.057	.051	.017	.190	.243	.000	.000	.000								
	N	91	91	91	91	91	91	91	91	91							
AvgGL_d	Pearson	.111	.114	.161	-.098	.019	.626**	.600**	.573**	.891**							
	Sig. (2-ta	.296	.280	.128	.355	.860	.000	.000	.000	.000							
	N	91	91	91	91	91	91	91	91	91							
PL_TSpdN	Pearson	-.397**	-.372**	.242*	.376**	.075	.301**	-.241*	-.273**	.054	.938**	1					
	Sig. (2-ta	.000	.000	.021	.000	.479	.004	.021	.009	.609	.000						
	N	91	91	91	91	91	91	91	91	91	91	91					
S_d	Pearson	-.688**	-.680**	.225*	.754**	.127	.150	-.510**	-.547**	.024	.449**	.473**	1				
	Sig. (2-ta	.000	.000	.032	.000	.229	.155	.000	.000	.819	.000	.000					
	N	91	91	91	91	91	91	91	91	91	91	91	91				
R_d	Pearson	-.426**	-.385**	.228*	.376**	.158	.334**	-.145	-.182	.195	.644**	.629**	.769**				
	Sig. (2-ta	.000	.000	.030	.000	.134	.001	.171	.084	.064	.000	.000	.000				
	N	91	91	91	91	91	91	91	91	91	91	91	91				
SMSP_d	Pearson	-.036	-.020	.419**	.073	1.000**	.163	.032	-.006	.125	.027	.075	.128				
	Sig. (2-ta	.737	.850	.000	.490	.000	.123	.764	.956	.240	.797	.481	.226				
	N	91	91	91	91	91	91	91	91	91	91	91	91				
PL_EVCdN	Pearson	.111	.136	.660**	.000	.289**	.324**	.272**	.225*	.360**	.305**	.302**	.140	-.504**	1		
	Sig. (2-ta	.297	.198	.000	.998	.005	.002	.009	.032	.000	.003	.004	.184	.000			
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91		
EVCd_TpdN	Pearson	-.135	-.165	-.335**	.066	-.015	-.139	-.166	-.167	-.088	-.159	-.106	.064	.577**	-.530**	1	
	Sig. (2-ta	.201	.118	.001	.533	.886	.189	.117	.114	.407	.133	.317	.544	.000	.000		
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	
EVCd_TSpdN	Pearson	-.133	-.164	-.347**	.062	-.040	-.141	-.163	-.165	-.084	-.165	-.113	.058	.584**	-.540**	.999**	1
	Sig. (2-ta	.208	.121	.001	.557	.705	.183	.123	.118	.428	.118	.285	.586	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 7 shows that nodes (Nodes) and ties (Edges_d) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes) and ties (Edges_d). Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) correlate positively with each other. Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpdN) correlates positively with Paths Power Law Distribution per Node (PL_TpdN). Scale Free Metric (S_d) seems to share a positive relationship with Density (Den_d). Scale Free Metric (S_d) shares a negative relationship with the number nodes (Nodes), ties (Edges_d), Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d). Assortativity (R_d) shares a positive relationship with Shortest Paths Power Law Distribution per Node (PL_TSpdN), Paths Power Law Distribution per Node (PL_TpdN) and Scale Free Metric (S_d). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) shares a positive correlation with Reciprocity and a negative correlation with the Eigenvector Centralization (ECd). Eigenvector Centrality with respect to Total Shortest Paths per

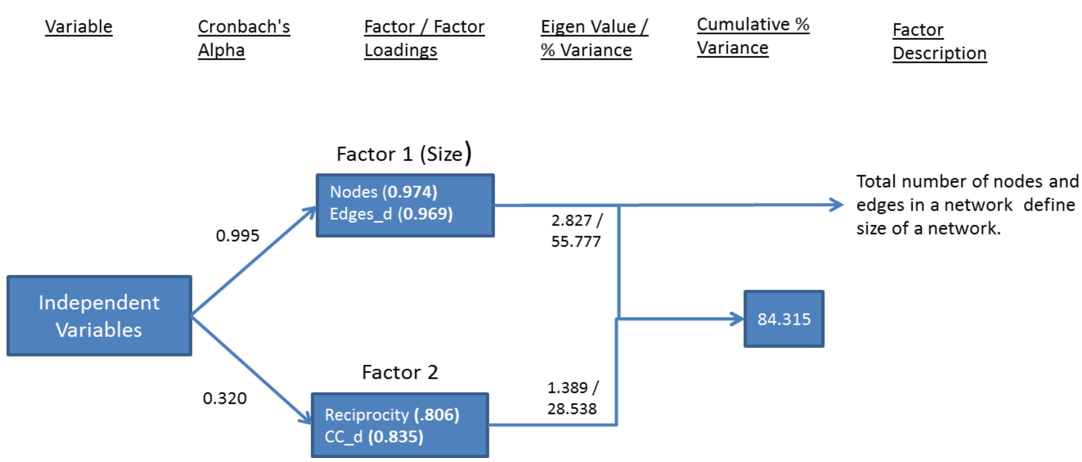
Node (EVCd_TSpdN) and Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) share a positive correlation with Eigenvector Centralization (ECd) and a negative correlation with the Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN). Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) and Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) correlate strongly with each other.

A.3.7.2.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.3.7.2.2.1 Independent Variables

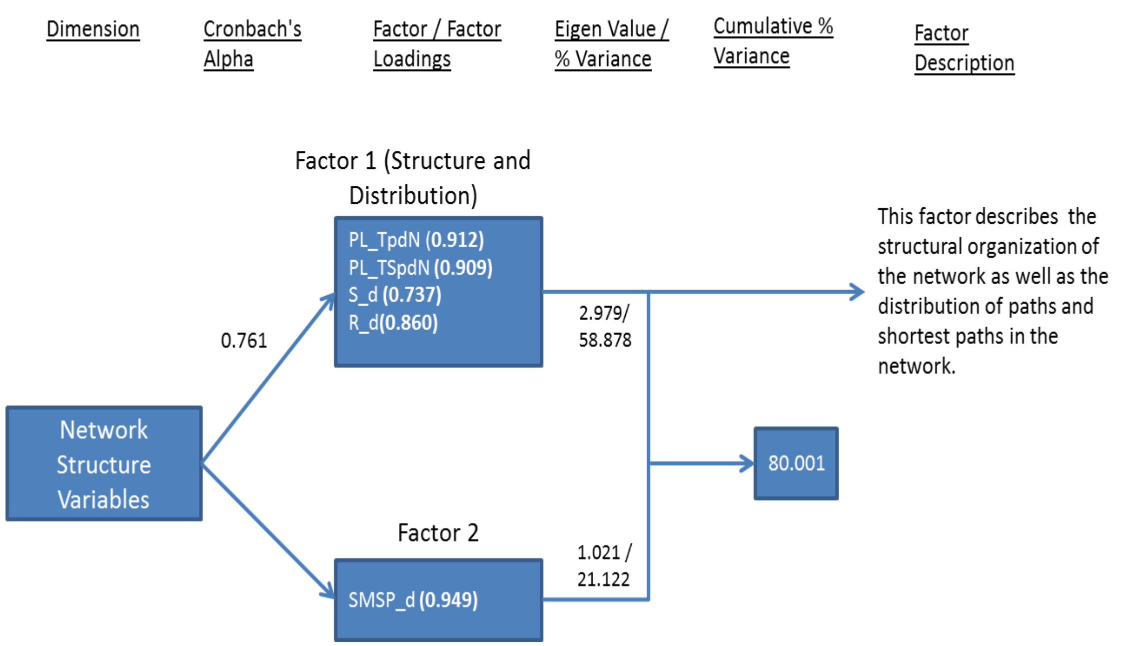
Figure 17: Factor Analysis of Independent Variables



The factor analysis generated two factors that explain 84.315% (greater than 80%) of the cumulative variance. Both factors have eigenvalues above one. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had a negative loading in factor 1, hence it was removed. Clustering Coefficient (CC_d) and Reciprocity have a significant loading in factor 2.Cronbach's alpha for factor1 has a value of 0.994. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as "Size". Cronbach's alpha for factor2 has a value of 0.32. This means reciprocity and clustering coefficient s are not measuring same construct within factor2.

A.3.7.2.2.2 Network Structure (MV1)

Figure 18: Factor Analysis of Network Structure Variables

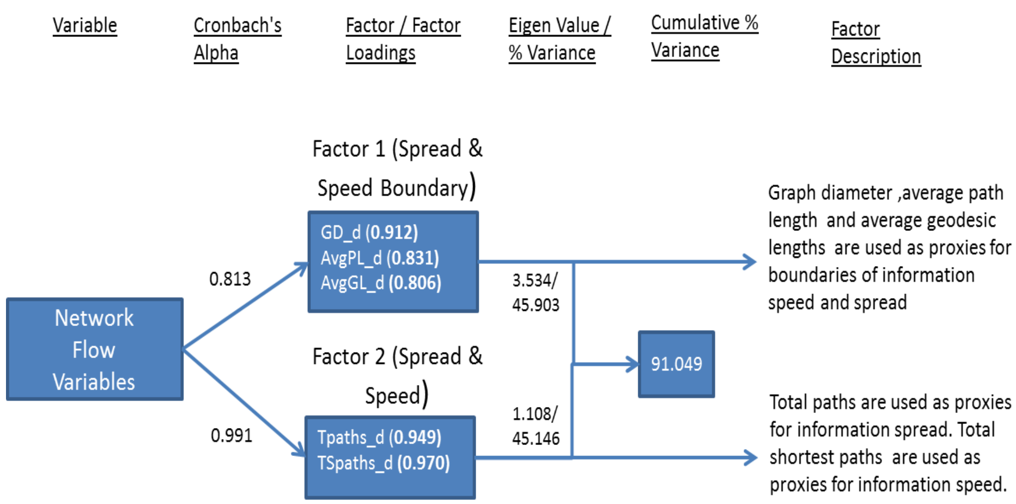


Factor analysis generated two factors that explain 80.001% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Paths Power Law Distribution per Node (PL_TpdN), Shortest Paths Power Law Distribution per Node

(PL_TpdN), Assortativity (R_d) and Scale Free Metric (S_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.761. Paths Power Law Distribution per Node (PL_TpdN), Shortest Paths Power Law Distribution per Node (PL_TpdN), Assortativity (R_d) and Scale Free Metric (S_d) are measuring same construct within factor 1. Hence, they should be considered as a factor. All other variables load independently.

A.3.7.2.2.3 Network Flow (MV2)

Figure 19: Factor Analysis of Network Flow Variables

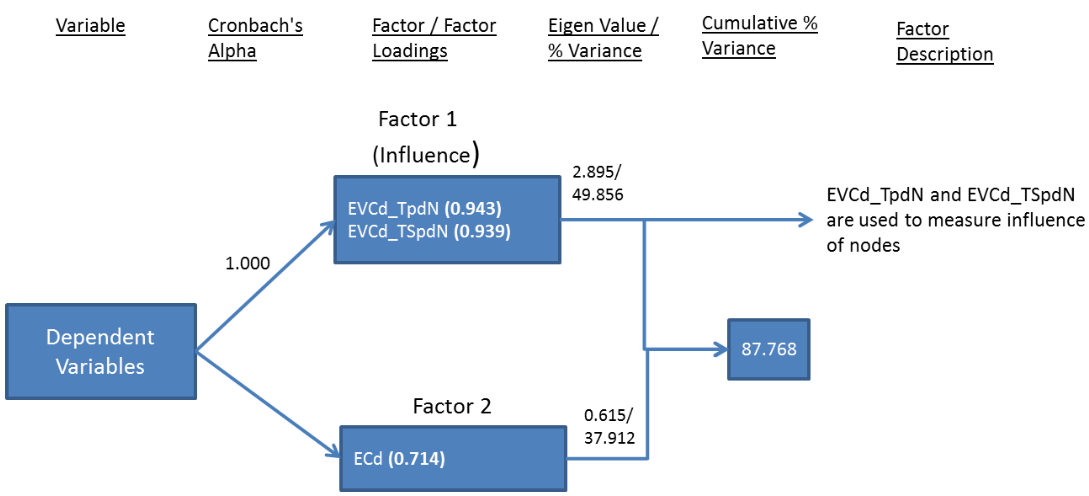


Factor analysis generated two factors that explain 80.001% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Graph Diameter (GD_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.813. Graph Diameter (GD_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) are measuring the same construct within factor 1. Factor 1 is named as "Spread and Speed Boundary".

Total Paths (Tpaths_d) and Total Shortest Paths (Tpaths_d) have significant factor loadings on factor2. Cronbach's alpha for factor2 has a value of 0.991. Total Paths (Tpaths_d) and Total Shortest Paths (Tpaths_d) are measuring the same construct within factor2. Factor2 is named as "Spread and Speed".

A.3.7.2.2.4 Dependent Variables

Figure 20: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 87.768% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCd_TpD) and Shortest Paths (EVCd_TSpD) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 1. I name factor1 as "Influence" as both, Eigenvector Centralities with respect to Paths (EVCd_TpD) and Shortest Paths (EVCd_TSpD), are being used measure of influence. All other variables load independently.

A.3.7.2.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Comedy.pdf".

A.3.7.2.3.1 Impact of Network Structure on Network Flow

Table 8: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Comedy	(0.102/0.001) [4]	(0.398/0.000) [3,4]	(0.426/0.000) [3,4]	NA	(0.037/0.039) [4]

Table 8 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 10.2%, 39.8%, 42.6%, 3.7% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d),) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Graph Diameter (GD_d) and Average Geodesic Length (AvgGL_ud) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.2.3.2 Impact of Network Flow on Network Structure

Table 9: Impact of Network Flow on Network Structure

Predictors: **(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpdN	PL_TSpdN	S_d	R_d	SMSP_d
Comedy	(0.208/0.000) [6,7]	(0.211/0.000) [6,8]	(0.529/0.000) [8,10]	(0.204/0.000) [6,8,10]	NA

Table 9 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 20.8%, 21.1%, 52.9%, and 20.4% variation in the PL_TpdN, PL_TSpdN, S_d, and R_ud, respectively.

A.3.7.2.3.3 Impact of Network Structure on Network Phenomenon

Table 10: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Comedy	(0.155/0.000) [3,4]	(0.086/0.003) [1]	NA	NA

Table 10 shows that the network structure variable impacts Eigenvector Centralization (EC_d) and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN). Network structure variables explain 15.5% variation in Eigenvector Centralization (EC_d).

A.3.7.2.3.4 Impact of Network Flow on Network Phenomenon

Table 11: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Comedy	NA	(0.095/0.002) [9,10]	NA	NA

Table 11 shows that the network structure variable Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), explaining 9.5% variation respectively. The impact of network Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.2.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 12: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d, (6)GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCud_TpdN	EVCud_TSpdN
Comedy	(0.246/0.000) [3,4,15]	(0.546/0.000) [1,7,15]	(0.132/0.001) [12,15]	(0.140/0.000) [12,15]

Table 12 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 24.6%, 54.6%, 13.2% and 14% variation respectively. The collective impact of independent variables and the moderating variables on Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.3 The Consumption Network

A.3.7.3.1 Correlation Analysis

Significant correlations coefficients for consumption network are shown below in table 13. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled "Correlations.pdf".

Table 13: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	PL_TpinN	ECin
Edges_d	Pearson	.990**	1									
	Sig. (2-ta	.000										
	N	91	91									
Den_d	Pearson	-.849**	-.838**	.209*	1							
	Sig. (2-ta	.000	.000	.047								
	N	91	91	91	91							
Tpaths_d	Pearson	.803**	.824**	.028	-.763**	.030	.352**	1				
	Sig. (2-ta	.000	.000	.789	.000	.779	.001					
	N	91	91	91	91	91	91	91				
TSpaths_d	Pearson	.827**	.846**	-.030	-.807**	-.008	.278**	.987**	1			
	Sig. (2-ta	.000	.000	.780	.000	.937	.008	.000				
	N	91	91	91	91	91	91	91	91			
AvgPL_d	Pearson	.201	.206	.249*	-.139	.124	.694**	.674**	.584**	1		
	Sig. (2-ta	.057	.051	.017	.190	.243	.000	.000	.000			
	N	91	91	91	91	91	91	91	91	91		
AvgGL_d	Pearson	.111	.114	.161	-.098	.019	.626**	.600**	.573**	.891**		
	Sig. (2-ta	.296	.280	.128	.355	.860	.000	.000	.000	.000		
	N	91	91	91	91	91	91	91	91	91		
PL_TSpinN	Pearson	-.161	-.142	.266*	.141	-.026	.169	-.066	-.109	.062	.808**	
	Sig. (2-ta	.127	.178	.011	.184	.805	.109	.534	.305	.559	.000	
	N	91	91	91	91	91	91	91	91	91	91	
SMSP_d	Pearson	-.036	-.020	.419**	.073	1.000**	.163	.032	-.006	.125	.075	
	Sig. (2-ta	.737	.850	.000	.490	.000	.123	.764	.956	.240	.482	
	N	91	91	91	91	91	91	91	91	91	91	
ECin	Pearson	-.199	-.231*	-.549**	.042	-.143	-.223*	-.272**	-.238*	-.273**	-.179	1
	Sig. (2-ta	.058	.028	.000	.691	.175	.033	.009	.023	.009	.089	
	N	91	91	91	91	91	91	91	91	91	91	91
PL_EVCinN	Pearson	.104	.120	.553**	.054	.181	.173	.172	.125	.239*	.230*	-.362**
	Sig. (2-ta	.326	.258	.000	.608	.087	.101	.102	.236	.022	.028	.000
	N	91	91	91	91	91	91	91	91	91	91	91
EVCin_TpinN	Pearson	-.313**	-.327**	-.284**	.234*	-.023	.000	-.256*	-.276**	-.031	.042	.580**
	Sig. (2-ta	.002	.002	.006	.026	.827	.997	.014	.008	.770	.691	.000
	N	91	91	91	91	91	91	91	91	91	91	91
EVCin_TSpinN	Pearson	-.323**	-.336**	-.291**	.238*	-.035	-.008	-.266*	-.284**	-.041	.038	.588**
	Sig. (2-ta	.002	.001	.005	.023	.740	.943	.011	.006	.700	.720	.000
	N	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

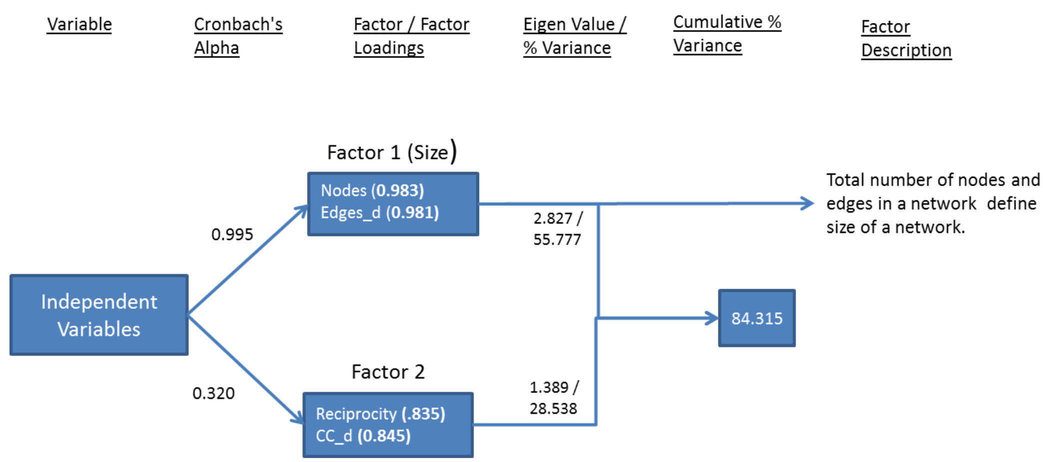
Tables 13 show that nodes (Nodes) and ties (Edges_ud) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d) and Reciprocity. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes) and ties (Edges_ud). Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpinN) correlates positively with Paths Power Law Distribution per Node (PL_TpinN). Small World Metric (S_con) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centralization (ECin) correlates negatively with Reciprocity. Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSPinN) and Eigenvector Centrality with respect to Total Paths per Node (EVCin_TPinN) correlate strongly with Eigenvector Centralization (ECin).

A.3.7.3.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.3.7.3.2.1 Independent Variables

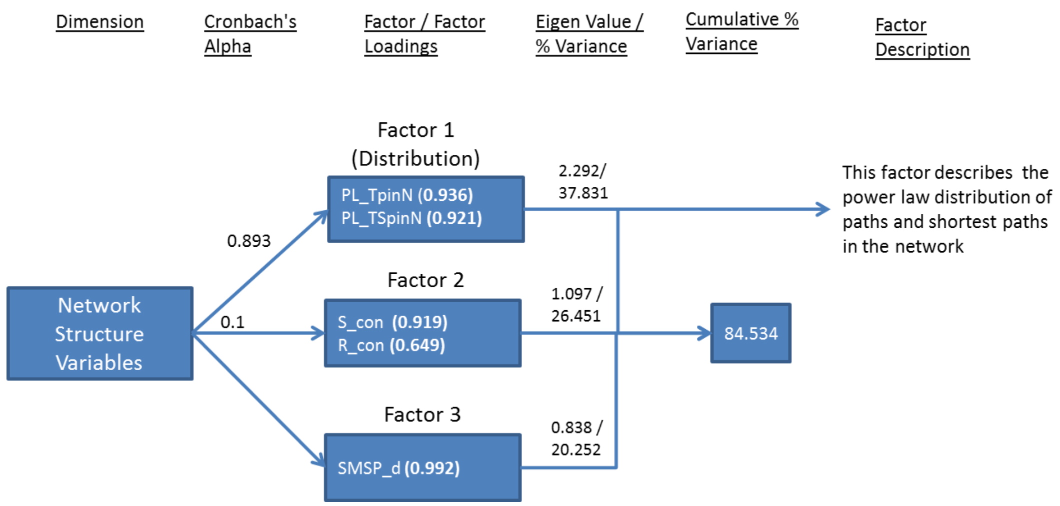
Figure 21: Factor Analysis of Independent Variables



The factor analysis generated two factors that explain 84.315% (greater than 80%) of the cumulative variance. Both factors have eigenvalues above one. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had a negative loading in factor 1, hence it was removed. Clustering Coefficient (CC_d) and Reciprocity have a significant loading in factor 2.Cronbach’s alpha for factor1 has a value of 0.994. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as “Size”. Cronbach’s alpha for factor2 has a value of 0.32. This means Clustering Coefficient (CC_d) and Reciprocity are not measuring same construct within factor2.

A.3.7.3.2.2 Network Structure (MV1)

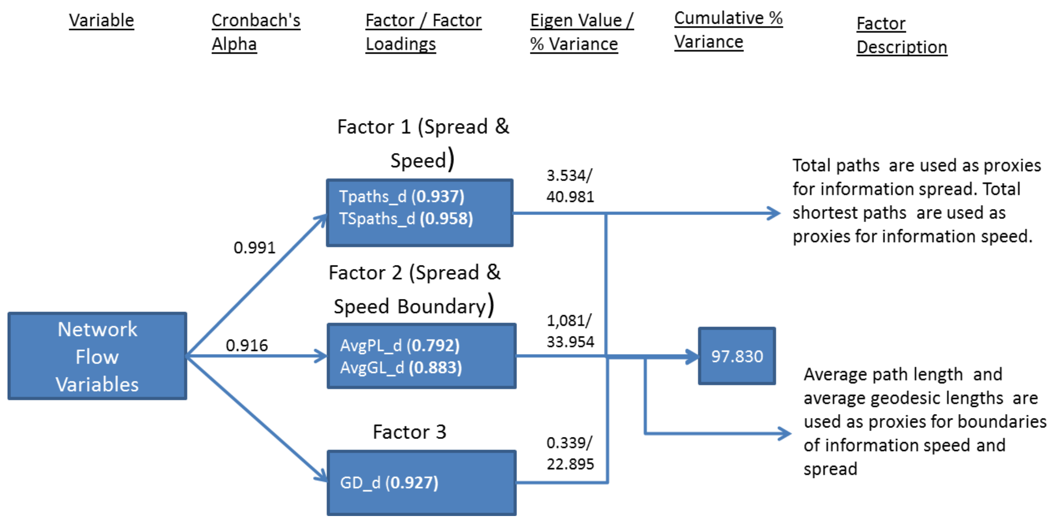
Figure 22: Factor Analysis of Network Structure Variables



Factor analysis generated three factors that explain 84.534% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 has eigenvalue little less than 1. Assortativity (R_con) and Scale Free Metric (S_con) have significant factor loadings in factor 2. Cronbach's alpha for factor2 has a value of 0.1. Assortativity (R_con) and Scale Free Metric (S_con) are not measuring same construct within factor 2. Hence, they should not be considered as a factor. Paths Power Law Distribution per Node (PL_TpinN) and Shortest Paths Power Law Distribution per Node (PL_TpinN) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.893. Paths Power Law Distribution per Node (PL_TpinN) and Shortest Paths Power Law Distribution per Node (PL_TpinN) are measuring same construct within factor 1. Hence, they should be considered as a factor. All other variables load independently. Factor 1 is named as "Distribution".

A.3.7.3.2.3 Network Flow (MV2)

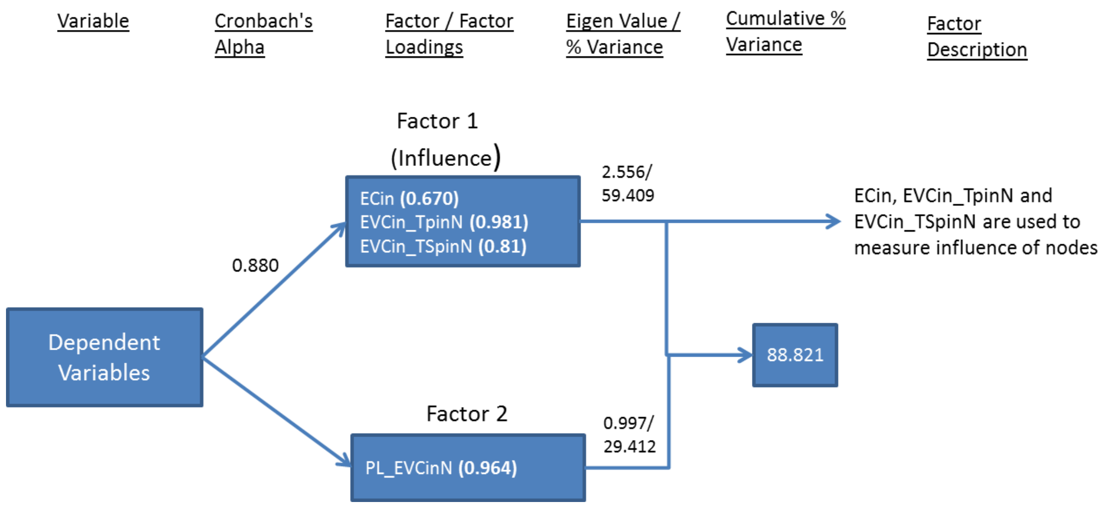
Figure 23: Factor Analysis of Network Flow Variables



Factor analysis generated three factors that explain 97.83% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 has eigenvalue little less than 1. Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 2. Cronbach's alpha for factor2 has a value of 0.916. Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) are measuring the same construct within factor2. Factor2 is named as "Spread and Speed Boundary". Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d) have significant factor loadings on factor1. Cronbach's alpha for factor2 has a value of 0.991. Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d) are measuring the same construct within factor1. Factor1 is named as "Spread and Speed".

A.3.7.3.2.4 Dependent Variables

Figure 24: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 88.821% (greater than 80%) of cumulative variance. Eigenvector Centralization (ECin), Eigenvector Centralities with respect to Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.880. I name factor1 as "Influence" as both, Eigenvector Centralities with respect to Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN), are being used measure of influence.

A.3.7.3.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Comedy.pdf".

A.3.7.3.3.1 Impact of Network Structure on Network Flow

Table 14: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Comedy	(0.210/0.000) [3,4]	(0.095/0.000) [1,3]	NA	0,230/0.000) [3]	(0.281/0.000) [1,3,4]

Table 14 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 21%, 9.5%, 23%, and 28.1% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively.

A.3.7.3.3.2 Impact of Network Flow on Network Structure

Table 15: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpinN	PL_TSpinN	S_con	R_con	SMSP_d
Comedy	NA	NA	(0.157/0.000)[6]	(0.100/0.001)[6]	NA

Table 15 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 15.7% and 10% variation in the S_con and R_con, respectively. The impact of network flow variables on R_con is not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.3.3.3 Impact of Network Structure on Network Phenomenon

Table 16: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Comedy	(0.124/0.000) [4]	(0.145/0.000) [3]	NA	NA

Table 16 shows that the network structure variable impacts Eigenvector Centralization (EC_in), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), explaining 12.4% and 14.5% variation respectively.

A.3.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 17: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Comedy	(0.064/0.009) [9]	(0.047/0.022) [9]	(0.066/0.008) [8]	(0.071/0.006) [8]

Table 17 shows that the network structure variables do not impact network phenomenon variables as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

.

A.3.7.3.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 18: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d, (6)GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Comedy	(0.409/0.000)[4,12,15]	(0.325/0.000)[4,15]	(0.188/0.000)[12,15]	(0.200/0.000)[12,15]

Table 18 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (ECin), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 40.9%, 32.5%, 18.8% and 20% variation respectively.

A.3.7.4 The Propagation Network

A.3.7.4.1 Correlation Analysis

Significant correlations coefficients for propagation network are shown below in table 19. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 19: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpoutN	S_pro	ECout	EVCout_TpoutN
Edges_d	Pearson	.990**	1												
	Sig. (2-ta	.000													
	N	91	91												
Den_d	Pearson	-.849**	-.838**	.209*	1										
	Sig. (2-ta	.000	.000	.047											
	N	91	91	91	91										
Tpaths_d	Pearson	.803**	.824**	.028	-.763**	.030	.352**	1							
	Sig. (2-ta	.000	.000	.789	.000	.779	.001								
	N	91	91	91	91	91	91	91							
TSpaths_d	Pearson	.827**	.846**	-.030	-.807**	-.008	.278**	.987**	1						
	Sig. (2-ta	.000	.000	.780	.000	.937	.008	.000							
	N	91	91	91	91	91	91	91	91						
AvgPL_d	Pearson	.201	.206	.249*	-.139	.124	.694**	.674**	.584**	1					
	Sig. (2-ta	.057	.051	.017	.190	.243	.000	.000	.000						
	N	91	91	91	91	91	91	91	91	91					
AvgGL_d	Pearson	.111	.114	.161	-.098	.019	.626**	.600**	.573**	.891**	1				
	Sig. (2-ta	.296	.280	.128	.355	.860	.000	.000	.000	.000					
	N	91	91	91	91	91	91	91	91	91	91				
PL_TSpoutN	Pearson	.053	.072	.006	-.169	.081	.261*	.202	.207*	.162	.189	.559**			
	Sig. (2-ta	.616	.500	.958	.109	.445	.013	.054	.049	.124	.072	.000			
	N	91	91	91	91	91	91	91	91	91	91	91			
R_pro	Pearson	-.199	-.182	.219*	.086	.071	.493**	.211*	.196	.499**	.615**	.379**	.636**		
	Sig. (2-ta	.058	.085	.037	.419	.504	.000	.045	.062	.000	.000	.000	.000		
	N	91	91	91	91	91	91	91	91	91	91	91	91		
SMSP_d	Pearson	-.036	-.020	.419**	.073	1.000**	.163	.032	-.006	.125	.021	.028	.008		
	Sig. (2-ta	.737	.850	.000	.490	.000	.123	.764	.956	.240	.844	.791	.940		
	N	91	91	91	91	91	91	91	91	91	91	91	91		
ECout	Pearson	-.199	-.231*	-.549**	.042	-.143	-.223*	-.272**	-.238*	-.273**	-.202	-.154	.017	1	
	Sig. (2-ta	.058	.028	.000	.691	.175	.033	.009	.023	.009	.055	.146	.873		
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
PL_EVCoutN	Pearson	.182	.221*	.623**	-.054	.112	.346**	.313**	.280**	.313**	.275**	.026	.074	-.678**	
	Sig. (2-ta	.084	.035	.000	.614	.291	.001	.003	.007	.003	.008	.808	.488	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
EVCout_TSpoutN	Pearson	.320**	.296**	-.090	-.287**	-.053	-.071	.280**	.274**	.123	.049	-.103	-.124	-.026	.989**
	Sig. (2-ta	.002	.004	.396	.006	.620	.503	.007	.009	.245	.643	.330	.241	.807	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

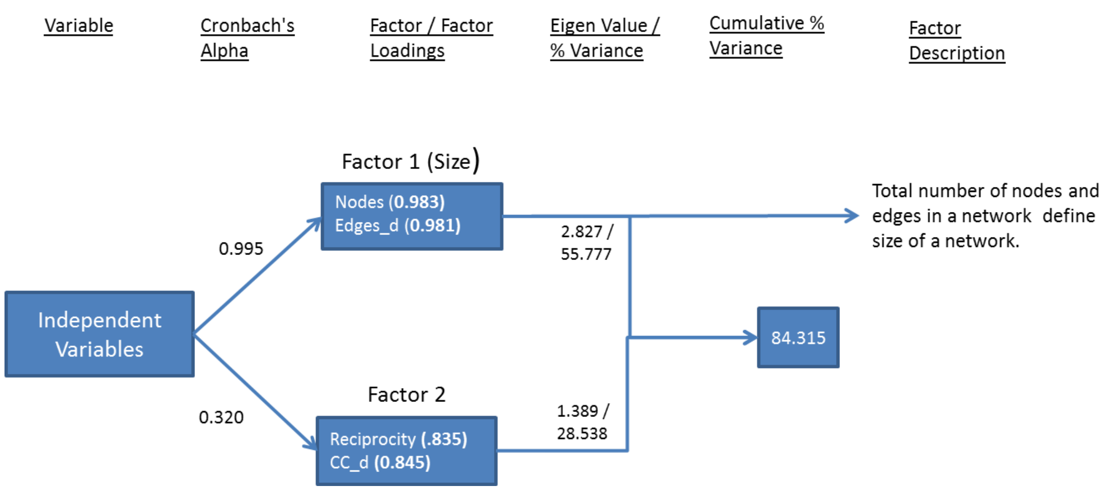
Table 19 shows that nodes (Nodes) and ties (Edges_ud) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d) and Reciprocity. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes) and ties (Edges_ud). Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpoutN) correlates positively with Paths Power Law Distribution per Node (PL_TpoutN). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centralization (ECout) correlates negatively with Reciprocity. Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSPoutN) and Eigenvector Centrality with respect to Total Paths per Node (EVCoutN_TPoutN) correlate strongly with Eigenvector Centralization (ECout).

A.3.7.4.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.3.7.4.2.1 Independent Variables

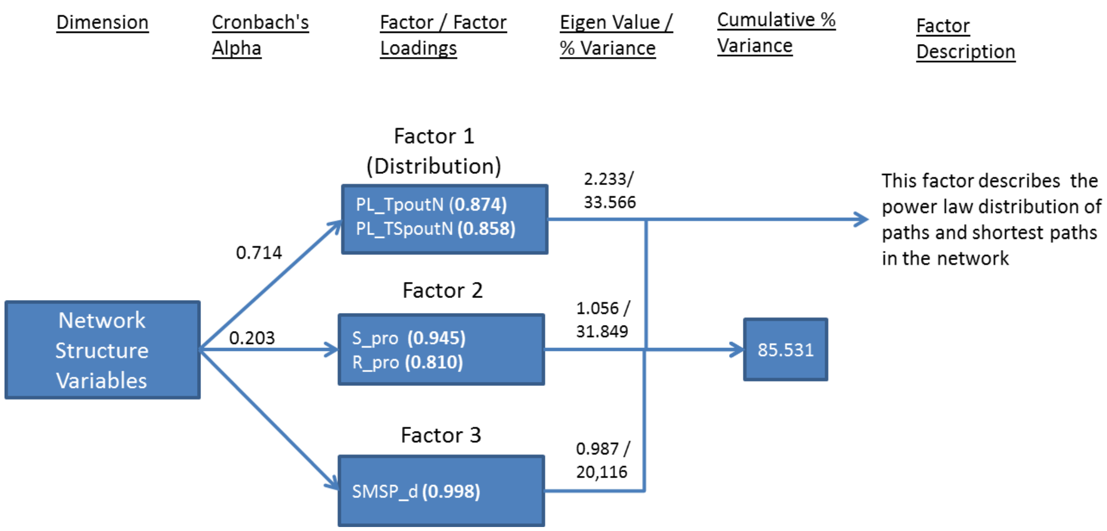
Figure 25: Factor Analysis of Independent Variables



The factor analysis generated two factors that explain 84.315% (greater than 80%) of the cumulative variance. Both factors have eigenvalues above one. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) had a negative loading in factor 1, hence it was removed. Reciprocity and Clustering Coefficient (CC_d) have a significant loading in factor 2.Cronbach’s alpha for factor1 has a value of 0.995. This means nodes and ties are measuring same construct within factor 1. Hence, I name factor 1 as “Size”. Cronbach’s alpha for factor2 has a value of 0.32. This means Reciprocity and Clustering Coefficient (CC_d) are not measuring same construct within factor 2.

A.3.7.4.2.2 Network Structure (MV1)

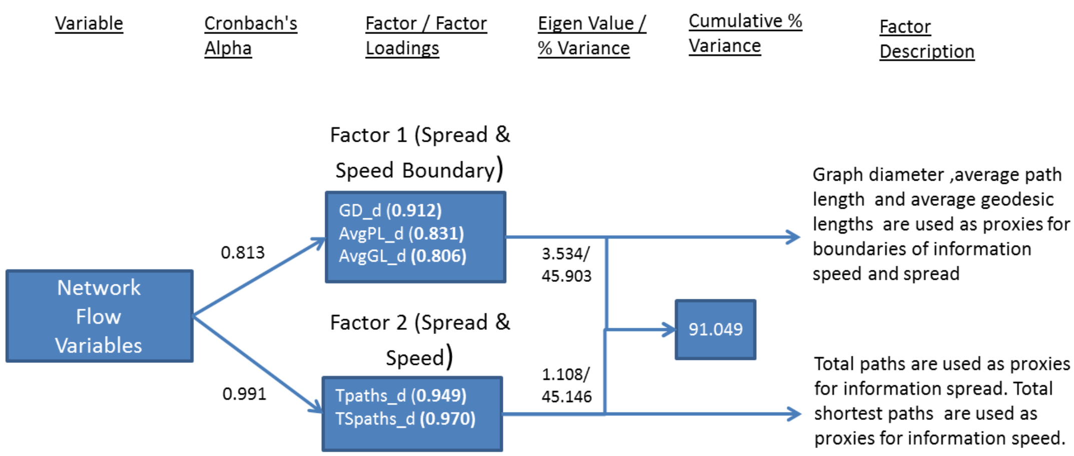
Figure 26: Factor Analysis of Independent Variables



Factor analysis generated three factors that explain 85.531% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 has eigenvalue little less than 1. Assortativity (R_pro) and Scale Free Metric (S_pro) have significant factor loadings in factor 2. Cronbach's alpha for factor2 has a value of 0203. Assortativity (R_pro) and Scale Free Metric (S_pro) are not measuring same construct within factor 2. Hence, they should not be considered as a factor. Power Law Distribution of Paths per Node (PL_TpoutN) and Power Law Distribution of Shortest Paths per Node (PL_TSpoutN) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.714. Power Law Distribution of Paths per Node (PL_TpoutN) and Power Law Distribution of Shortest Paths per Node (PL_TSpoutN) are measuring same construct within factor 1. Hence, they should be considered as a factor. All other variables load independently. Factor 1 is named as "Distribution".

A.3.7.4.2.3 Network Flow (MV2)

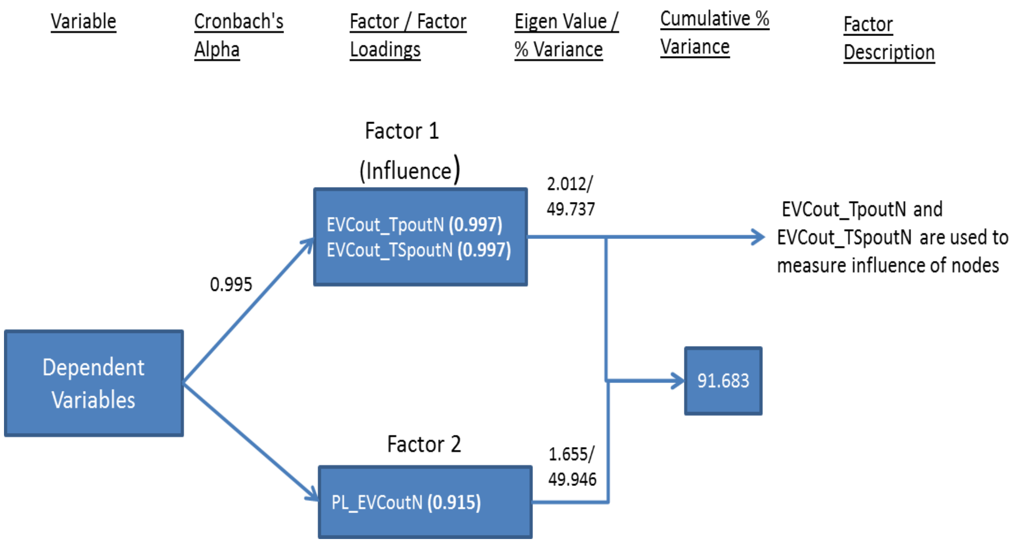
Figure 27: Factor Analysis of Network Flow Variables



Factor analysis generated two factors that explain 91.049% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Graph Diameter (GD_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor2 has a value of 0.813. Graph Diameter (GD_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) are measuring the same construct within factor1. Factor1 is named as "Spread and Speed Boundary". Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) have significant factor loadings on factor2. Cronbach's alpha for factor2 has a value of 0.991. Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) are measuring the same construct within factor2. Factor2 is named as "Spread and Speed".

A.3.7.4.2.4 Dependent Variables

Figure 28: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 91.683% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.995. I name factor1 as "Influence" as both Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN), are being used measure of influence.

A.3.7.4.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled “RegressionAnalysis_Comedy.pdf”.

A.3.7.4.3.1 Impact of Network Structure on Network Flow

Table 20: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Comedy	(0.240/0.000) [4]	(0.034/0.045) [4]	(0.035/0.043) [3]	(0.240/0.000) [4]	(0.416/0.000) [3,4]

Table 19 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 24%, 3.4%, 3.5%, 24% and 41.6% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.4.3.2 Impact of Network Flow on Network Structure

Table 21: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpoutN	PL_TSpoutN	S_pro	R_pro	SMSP_d
Comedy	(0.050/0.019) [6]	(0.058/0.013) [6]	(0.114/0.001) [6]	(0.386/0.000) [6,7,10]	NA

Table 21 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 5%, 5.8%, 11.4%, and 38.6% variation in the PL_TpoutN, PL_TSpoutN, S_pro, and R_pro, respectively. The impact of network flow variables on PL_TpoutN, PL_TSpoutN and S_pro, are not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.4.3.3 Impact of Network Structure on Network Phenomenon

Table 22: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Comedy	NA	(0.034/0.044) [4]	(0.076/0.005) [4]	(0.080/0.004)[4]

Table 22 shows that the network structure variables do not impact network phenomenon variables as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 23: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Comedy	(0.064/0.009) [10]	(0.142/0.000) [6,7]	(0.058/0.013) [7]	(0.119/0.001) [7]

Table 23 shows that the network structure variable impacts Eigenvector Centralization (ECout), Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 6.4%, 14.2%, 5.8% and 11.9% variation respectively. The impact of network flow variables on EC_out, EVCout_TpoutN and EVCout_TSpoutN are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.3.7.4.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 24: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d, (6),GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Comedy	(0.328/0.000) [15]	(0.487/0.000) [8,14,15]	(0.316/0.000) [4,9,11]	(0.358/0.000) [4,9,11]

Table 24 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_out), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 32.8%, 48.7%, 31.6%and 35.8% variation respectively.

A.4 Case 4 --Sports

A.4.1 Case Overview

Data for keyword “YouTube + sports” was collected over a period of 91 days (31/12/2013 to 31/03/2014). As shown in table 9, overall 129,182 tweets were collected, out of which 67,476 were broadcast tweets and 61,706 were engaged tweets respectively. Out of 61,706 engaged tweets only 32,778 tweets formed the largest community. Similarly, 77,617 daily unique people tweeted overall, out of which 25,776 daily unique people were engaged in broadcast activity whereas 51,841 daily unique people were engaged in conversations. Out of 51,841 daily unique people only 29,998 daily unique people formed the largest community. Data for the largest community was analyzed at a daily interval. The overall trends for the sports data are shown below in figure 1 and figure 2.

Figure 1: Overall Tweets

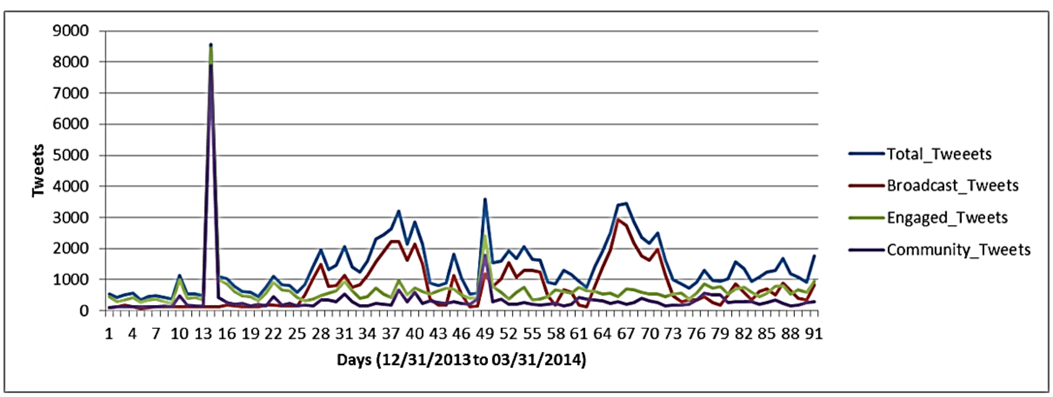


Fig.2: Overall People

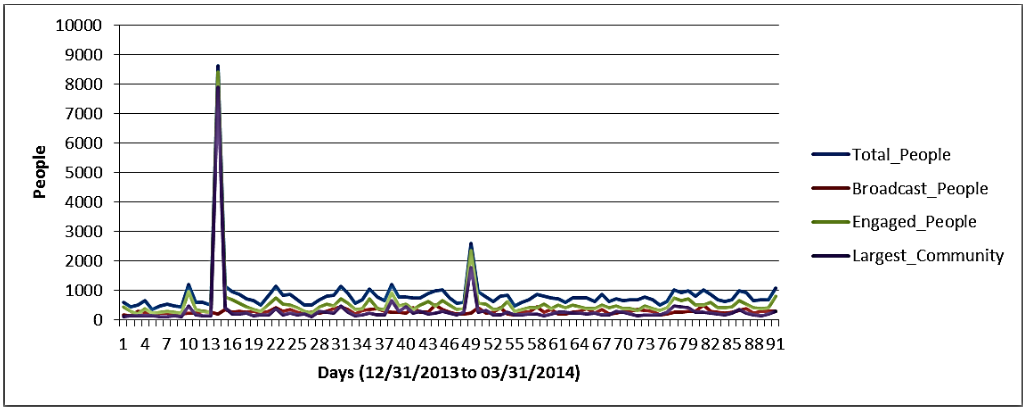


Figure 1 and figure 2 shows that both the total tweets and total people involved are very dynamic and their magnitude changes on a daily basis. The maximum of the total number of daily tweets and the maximum of the total number of daily unique people observed on a single day (the daily uniques) are 8,562 and 8,624, respectively. Similarly, the minimum of the total number daily tweets and the minimum of the number daily unique are 333 and 360, respectively. The size of the largest community on a particular day and the largest number of community tweets on that day also seem to follow the trend of total people and total tweets. The largest number of daily community tweets and the largest number of daily unique people are 7,881 and 7,882, respectively. Similarly, the smallest number of daily community tweets and the smallest number of daily unique people are 108 and 108, respectively. As the total number of daily unique people tweeting increases, so does the size of the community. Most of the engaged people are engaged in the collective conversation forming the largest community.

A.4.2 Random or Not Random

As explained in section 4.4.1, in order to eliminate α- error and β- error, I compare the clustering coefficients of both undirected (CC_ud) and directed networks (CC_d) with their corresponding random (Erdös-Rényi, E-R) networks (CCudran, CCdran). If the clustering coefficients of the undirected and directed networks are equal to those of the E-R random network, then the directed and undirected networks are considered to be random, if they are not equal, then they are not random.

Figure 3: Comparison of Clustering Coefficients of Undirected Network with E-R Networks

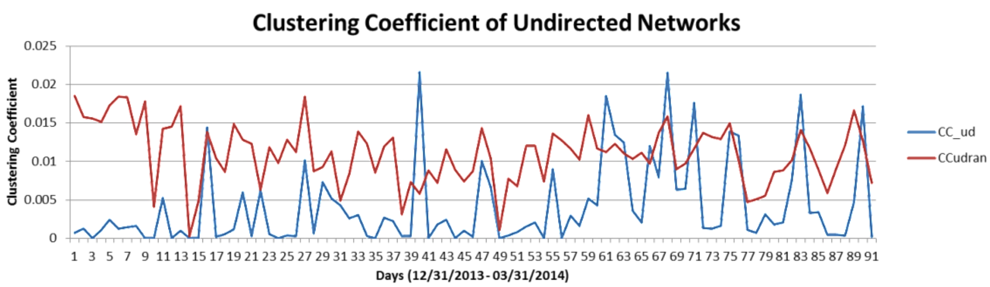
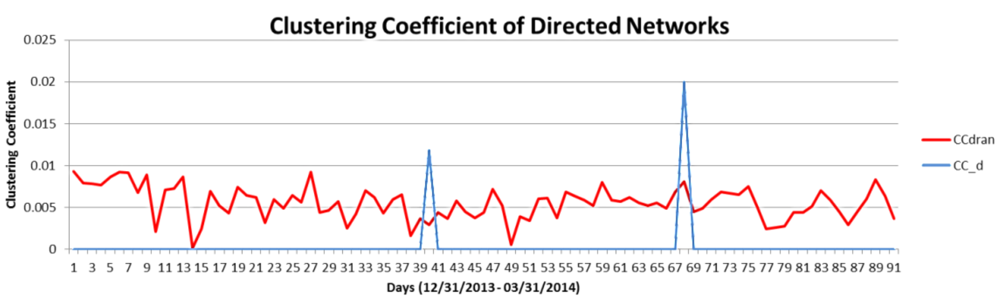


Figure 4: Comparison of Clustering Coefficients of Directed Network with E-R Networks



As seen in figure 3 and figure 4 Clustering Coefficients of the undirected networks (CC_ud) follows a very different pattern from their corresponding E-R networks. Therefore, the undirected network is considered to be non-random networks and the variables computed are a true reflection of network's features. For the direct network the Clustering Coefficients (CC_d) is zero for the most part. Therefore, the directed networks are random.

A.4.3. Independent Variables

The values of the independent variables for both the undirected and the directed network are shown in figure 5 below.

Figure 5: Independent Variables--(a) Nodes and Edges (Undirected and Directed networks), (b) Reciprocity (Directed Networks), (c) Density (Undirected and Directed Networks), (d) Clustering Coefficient Undirected Network, (e) Clustering Coefficient Directed Network.

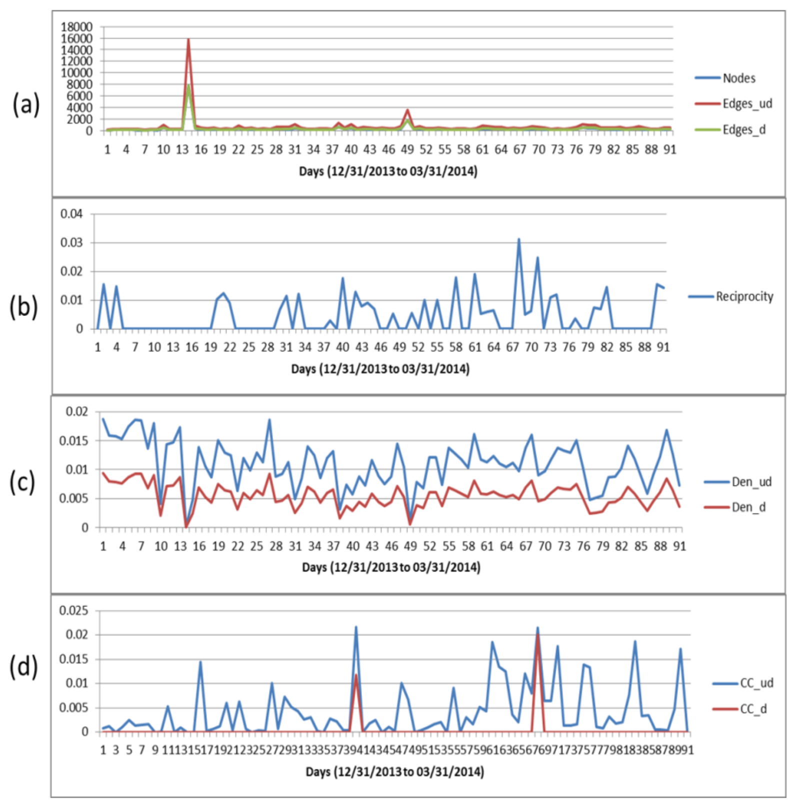
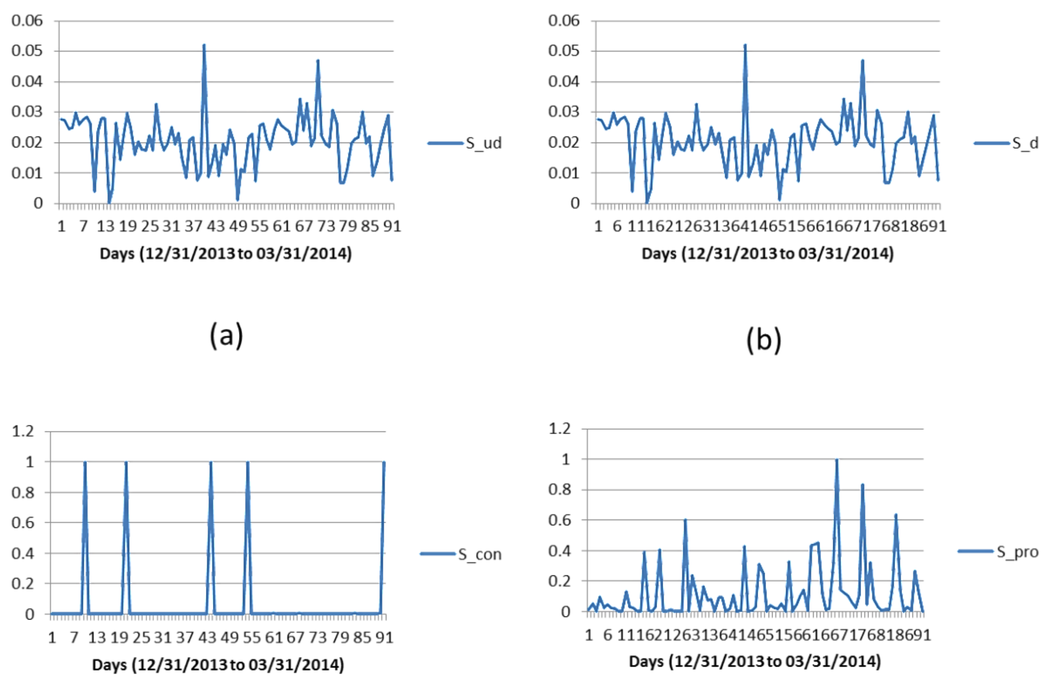


Figure 5 (a) shows that the number of directed ties (Edges_d) in the network and the total number of nodes (Nodes) overlap with each other. The numbers of undirected ties (Edges_ud) is greater than the number of directed ties (Edges_d), because in an undirected network every directed tie is considered to be symmetric. Therefore it is counted twice, except for the ones that are symmetric in a directed network. Reciprocity in Figure 5(b) indicates the presence of symmetric ties in a directed network (in an undirected network 100% are symmetric). The value of 0.01 is equal to 1% of all the ties. Figure 5(c) shows the difference between the densities of the undirected (Den_ud) and the directed networks (Den_d). The undirected network is denser than the directed network (Den_ud>Den_d). Figure 5(d) shows that the directed networks have higher Clustering Coefficients than the undirected networks (CC_d >CC_ud).

A.4.4 Network Structure Variables (MV1)

A.4.4.1 The Scale Free Metric

Figure 6: Scale Free Metric--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



(a) (b) (c) (d)

Figure 6 shows the Scale Free Metric for the undirected, directed, consumption and propagation networks (S_ud, S_d, S_con, S_pro). The Scale Free Metrics for the undirected (S_ud) and the directed network (S_d) are similar, but the Scale Free Metrics for the consumption (S_con) and propagation (S_pro) networks are very different. The propagation (S_pro) network is more scale free than the consumption network (S_con). The values of the scale free metric ranges between 0 and 1. When the values are closer to 1, it means that the networks are more scale free. Neither the directed nor the undirected network is scale free. This means that these networks may have hubs in them. However, there is not just one hub that is the center of the community. As shown in figure 6 (c) and figure 6 (d) the consumption network and the propagation network are scale free in some instances.

A.4.4.2 The Assortativity

Figure 7: Assortativity--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

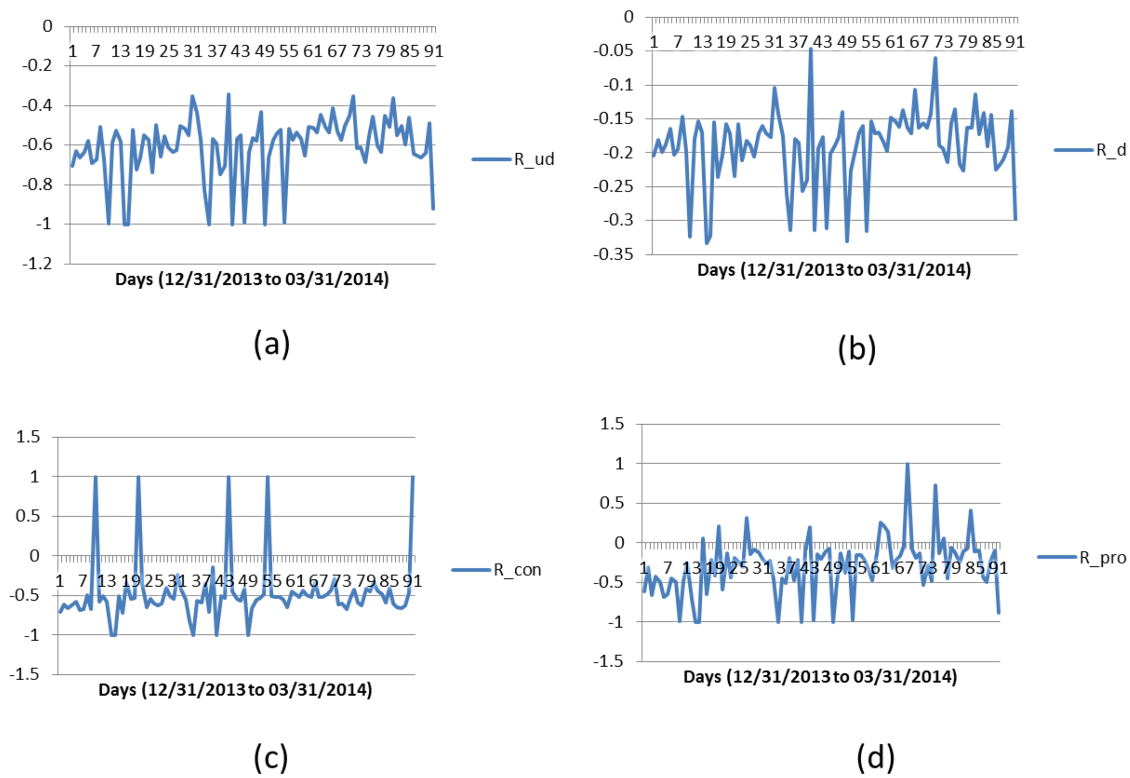
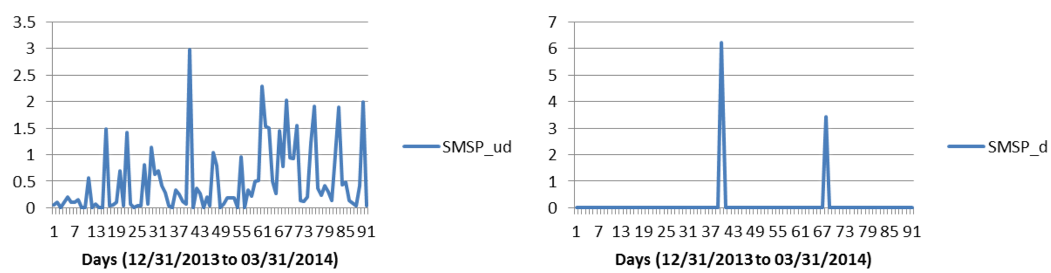


Figure 7 shows the assortativity metric for the undirected, directed, consumption and propagation networks (R_ud, R_d, R_con, R_Pro). The value of the assortativity metric ranges between -1 and 1. When the values are closer to -1, it means that networks are disassortative. The undirected network is more Disassortative than

the directed network (R_d > R_ud). Among the directed networks, the consumption network is more Disassortative than the propagation network (R_pro > R_con). Disassortative means that the nodes in the network connect to nodes that are very similar to themselves. This is true more so in the undirected network and in the consumption network than it is in the directed network and the propagation network. This implies that disassortativeness of consumption contributes more to the disassortativeness of the directed network than the disassortativeness of the propagation does.

A.4.4.3 The Small World Metric

Figure 8: Small World Metric -- (a) Undirected Network, (b) Directed Network.



(a) (b)

Figure 8 shows the Small World Metric for the undirected (SMSP_ud) and directed networks (SMSP_d). The Small World Metrics for the consumption and propagation networks are the same as the ones for the directed network. The directed networks don't show any small world behavior. Contrary to the directed networks, undirected networks show some small world behavior but not significantly enough. This

means that in undirected networks there are more nodes that act as hubs that facilitate communication between other nodes of the network.

A.4.4.4 Paths and Shortest Paths Power law Distribution per Node

Figure 9: Power Law Distribution of Paths and Shortest Paths in (a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

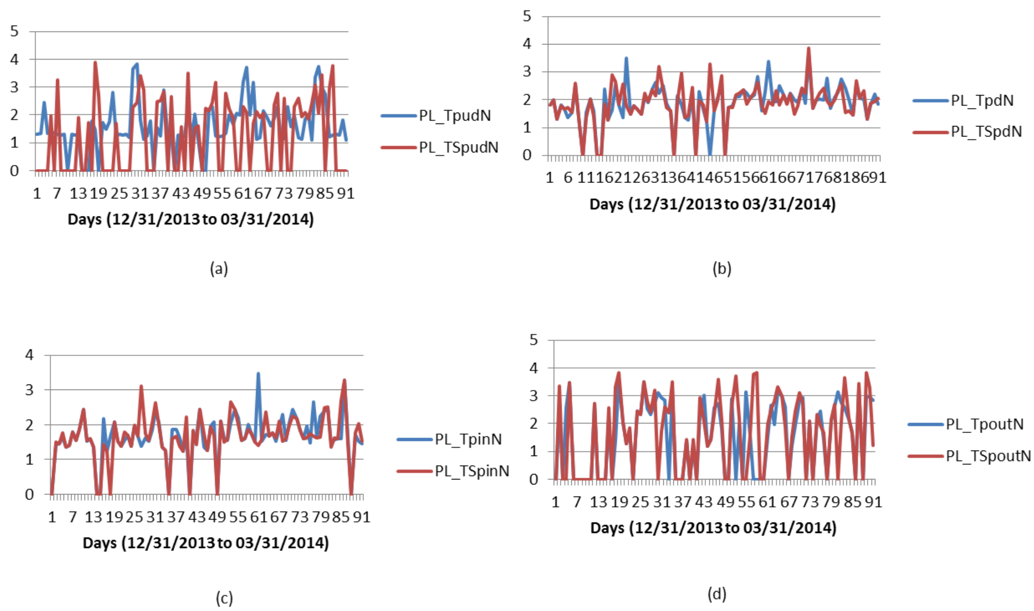


Figure 9 (a) shows that, in the undirected network, paths are more uniformly distributed among nodes than shortest paths are distributed among nodes. This means that fewer nodes are responsible for more of the shortest paths in the undirected network. There are fewer instances of shortest path following power law distribution in undirected (figure 9 (a)) and consumption (figure 9 (c)) networks. In the directed (figure 9 (b)) and propagation (figure 9 (d)) networks, there are no such patterns.

A.4.5 Network Flow Variables (MV2)

Figure 10: Network Flow Variables-- (a) Total Paths and Total Shortest Paths, (b) Average Paths and Average Shortest Paths, (c) Undirected and Directed Network Graph Diameter.

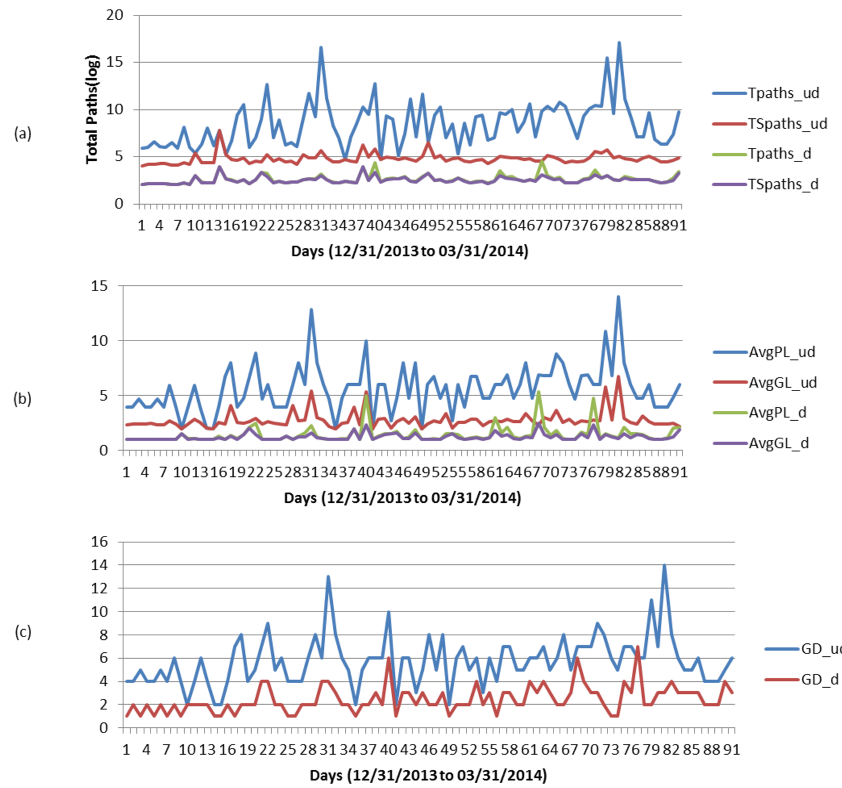


Figure 10 (a), shows that total number of paths in the undirected network (Tpaths_ud) is orders of magnitude higher than the total number of shortest paths (TSpaths_ud). The total number of paths (Tpaths_d) and the total number of shortest paths (TSpaths_d) map more closely in the directed network. In figure 10 (b), a similar trend is observed in the Average Path Lengths (AvgPL_ud, AvgPL_d) and the Average Geodesic Lengths (AvgGL_ud, AvgGL_d) of the undirected and directed networks. In figure 10 (c), the Graph Diameter (GD_ud) of the undirected network is larger than the graph diameter of the directed network (GD_d). It is also noteworthy that, in figure 10

(b) and in figure 10 (c), the Graph Diameter (GD_ud, GD_d) and the Average Path Length (AvgPL_ud, AvgPL_d) of the undirected and directed networks track pretty closely.

A.4.6 Dependent Variables

A.4.6.1 Eigenvector Centralization

Figure 11: Eigenvector Centralization in the Undirected, Directed, Consumption and Propagation Networks

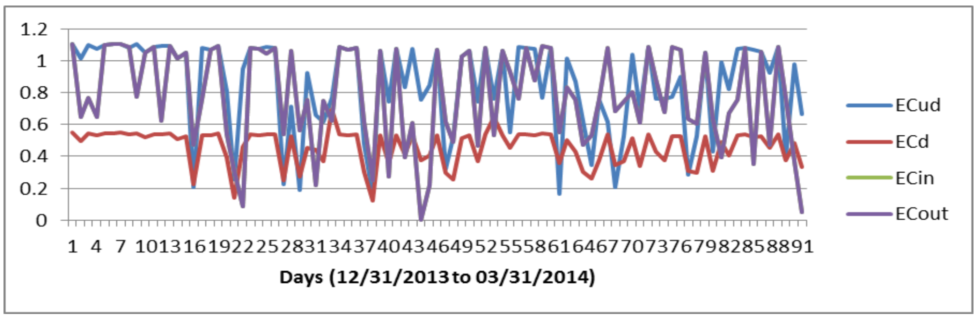


Figure 11 shows that nodes with influence are lot more central in the undirected (ECud) and propagation networks (ECout) than in the directed network (ECd). The consumption (ECin) and propagation (ECout) networks exhibit same level of centralization.

A.4.6.2 Power law Distribution of Eigenvector Centrality per Node

Figure 12: Power Law Distribution of Eigenvector Centrality in Undirected, Directed, Consumption and Propagation Network

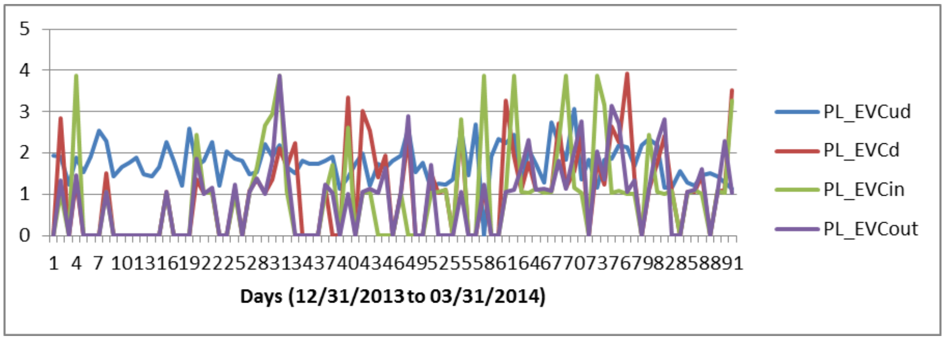
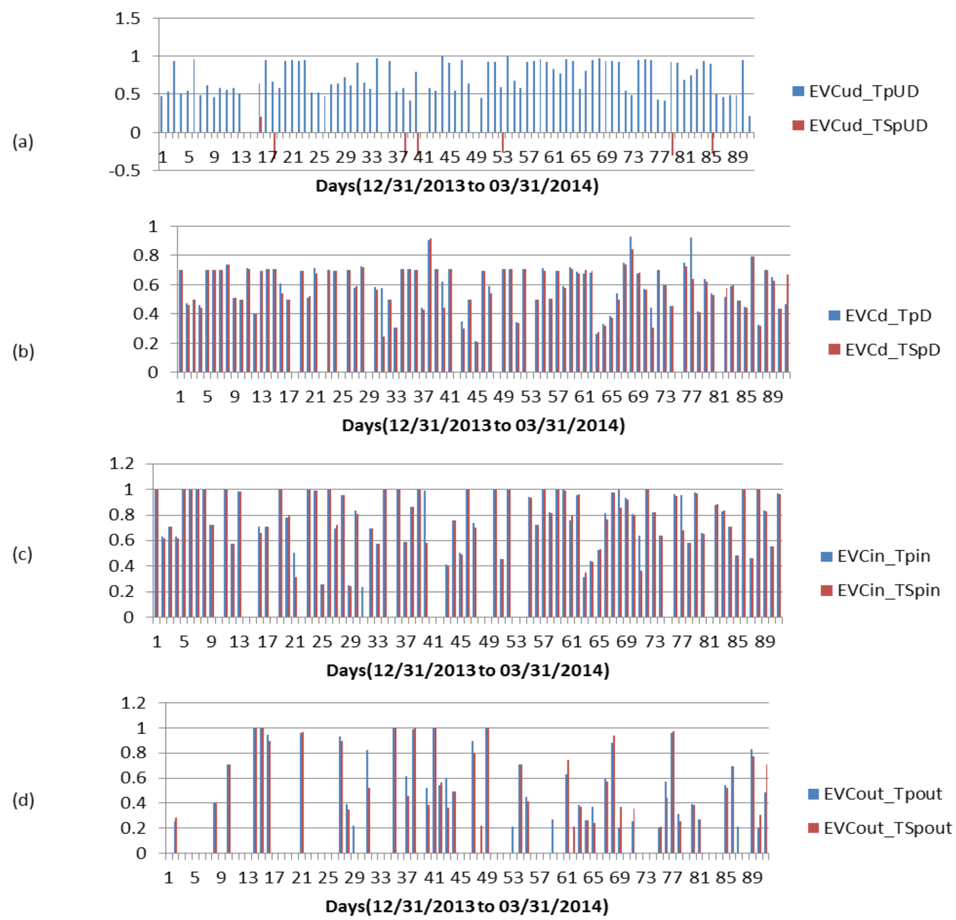


Figure 12 shows that in the undirected network eigenvector centrality values are consistently distributed in a power law distribution pattern (PL_EVCud), over a period of time. In the directed, the consumption and the propagation network the distribution of eigenvector centrality follows a power law distribution (PL_EVCd, PL_EVCin, PL_EVCout) pattern only sometimes.

A.4.6.3 Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node

Figure 13: Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



In figure 13, only those correlation coefficients with a significance value lower than 0.05 are shown. In figure 13 (a), there is a significant correlation between the eigenvector centrality of a node and the number of paths from a node in undirected network (EVCud_TpUDN). There is no significant correlation between eigenvector centrality of a node and shortest paths from a node in undirected network

(EVCud_TSpUDN). In figure 13 (b), there is a significant correlation between the directed-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the directed network (EVCd_TpDN, EVCud_TSpUDN). In figure 13 (c), there is a significant correlation between the in-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the consumption network (EVin_TpinN, EVCin_TSpinN). The correlation between the out-eigenvector centrality of a node and the number of shortest paths is less significant figure 13 (d) (EVCout_TpoutN, EVCout_TSpoutN).

A.4.7 Statistical Analysis

A.4.7.1 The Undirected Network

A.4.7.1.1 Correlation Analysis

In Table 1, the statistically significant correlation coefficients for the undirected network are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 1: Correlation Coefficients of Undirected Network

Correlations

		Nodes	Edges_ud	Den_ud	CC_ud	GD_ud	Tpaths_ud	AvgPL_ud	AvgGL_ud	PL_TpudN	S_ud	R_ud
Edges_ud	Pearson	.999**	1									
	Sig. (2-ta	.000										
	N	91	91									
Tpaths_ud	Pearson	.014	.036	-.410**	.269**	.950**	1					
	Sig. (2-ta	.898	.733	.000	.010	.000						
	N	91	91	91	91	91	91					
TSpaths_ud	Pearson	.728**	.741**	-.854**	.010	.139	.399**					
	Sig. (2-ta	.000	.000	.000	.927	.189	.000					
	N	91	91	91	91	91	91					
AvgPL_ud	Pearson	-.174	-.155	-.186	.299**	.999**	.956**	1				
	Sig. (2-ta	.098	.143	.077	.004	.000	.000					
	N	91	91	91	91	91	91	91				
AvgGL_ud	Pearson	-.072	-.056	-.264*	.170	.810**	.787**	.815**	1			
	Sig. (2-ta	.500	.597	.011	.108	.000	.000	.000				
	N	91	91	91	91	91	91	91	91			
S_ud	Pearson	-.338**	-.326**	.635**	.633**	.296**	.083	.294**	.259*	.270**	1	
	Sig. (2-ta	.001	.002	.000	.000	.004	.433	.005	.013	.010		
	N	91	91	91	91	91	91	91	91	91	91	
R_ud	Pearson	-.317**	-.293**	.232*	.497**	.693**	.576**	.693**	.564**	.427**	.673**	1
	Sig. (2-ta	.002	.005	.027	.000	.000	.000	.000	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	91	91	91
SMSP_ud	Pearson	-.081	-.047	.055	.969**	.361**	.353**	.361**	.216*	.383**	.580**	.528**
	Sig. (2-ta	.448	.657	.605	.000	.000	.001	.000	.040	.000	.000	.000
	N	91	91	91	91	91	91	91	91	91	91	91
EVCud_TpudN	Pearson	-.338**	-.325**	.227*	.398**	.419**	.250*	.380**	.214*	.614**	.410**	.503**
	Sig. (2-ta	.001	.002	.030	.000	.000	.017	.000	.042	.000	.000	.000
	N	91	91	91	91	91	91	91	91	91	91	91
EVCud_TSpudN	Pearson	-.017	-.020	.263*	.009	-.261*	-.273**	-.262*	-.508**	-.152	-.052	-.101
	Sig. (2-ta	.877	.853	.012	.930	.013	.009	.012	.000	.152	.624	.339
	N	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

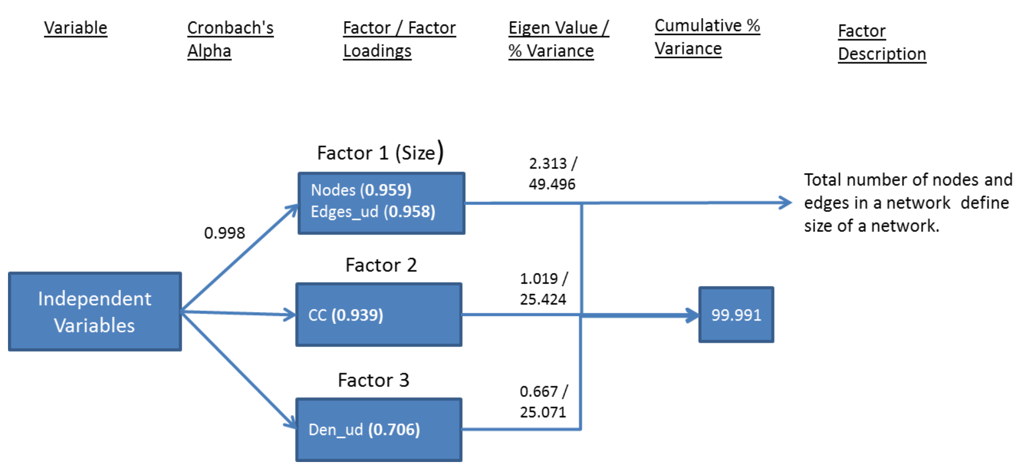
In Table 1, the number of nodes (Nodes) and the number of ties (Edges_ud) have a strong positive correlation. As the number of nodes (Nodes) increases, the number of ties (Edges_ud) also increases. Total Paths (Tpaths_ud) have a strong positive correlation with Graph Diameter (GD_ud). The Total Number of Shortest Paths (TSpaths_ud) correlates strongly with the number of nodes (Nodes) and the number of ties (Edges_ud), but it correlates negatively with Density (Den_ud). Average Path Length (AvgPL_ud) and Average Geodesic Length (AvgGL_ud) share a strong positive correlation with Graph Diameter (GD_ud) and Total Paths (Tpaths_ud). Average geodesic length shares a strong correlation with average path length. Average Geodesic Length (AvgGL_ud) shares a strong correlation with Average Path Length (AvgPL_ud). Scale Free Metric (S_ud) shares a positive correlation with Density (Den_ud) and Clustering Coefficients (CC_ud). Assortativity (R_ud) shares positive correlations with Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Average Path Length (AvgPL_ud), Average Geodesic Length (AvgGL_ud) and Scale Free Metric (S_ud). Small World (SMSP_ud) metric share a positive relationship with the Clustering Coefficients (CC_ud), Scale Free Metric (S_ud) and Assortativity (R_ud). Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and correlate strongly Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Assortativity (R_ud). Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN) correlates negatively with number of nodes.

A.4.7.1.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.4.7.1.2.1 Independent Variables

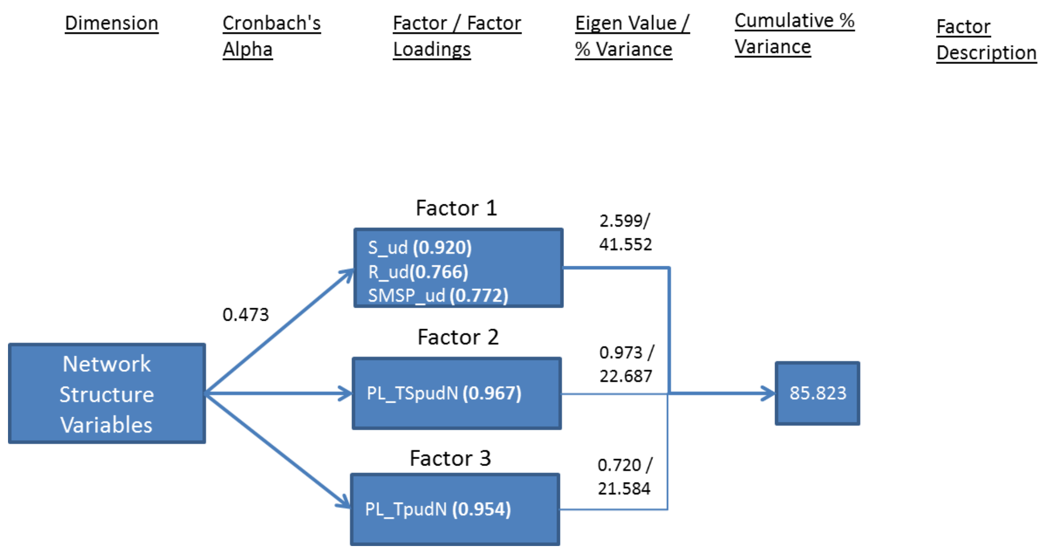
Figure 14: Factor Analysis Independent Variables Music Undirected Network



The factor analysis generated three factors that explain 99.991% (greater than 80%) of the cumulative variance. Factor1 and factor 2 have eigenvalues above 1. Factor3 has eigenvalue below 1. Nodes and ties (Edges_ud) have significant factor loadings in factor 1. Clustering Coefficient (CC_ud) has significant loading in factor 2. Density (Den_ud) has significant loading in factor 3.Cronbach's alpha for factor 1has a value of 0.998. This means nodes and ties are measuring same construct within factor 1 whereas Den_ud and CC_ud load independently on factor2 and factor 3. Hence, I name factor 1 as "Size".

A.4.7.1.2.2 Network Structure (MV1)

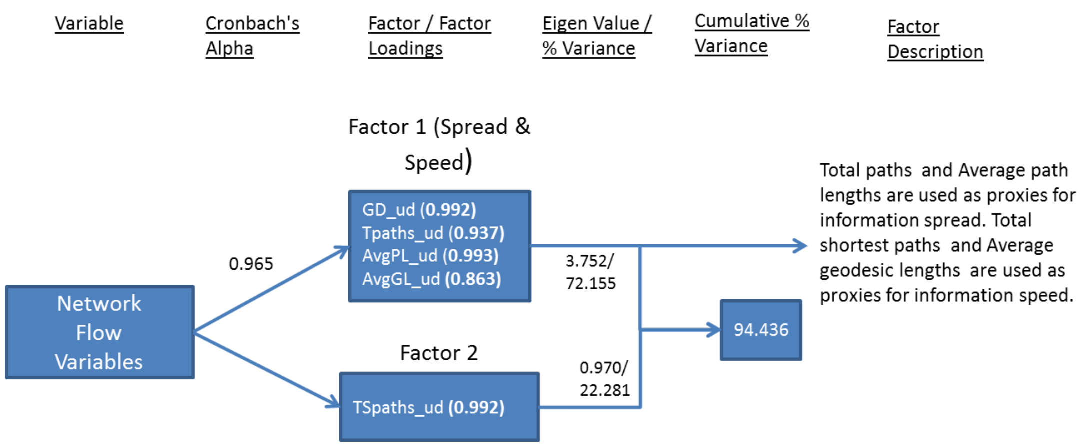
Figure 15: Factor Analysis of Network Structure Variables



The factor analysis generated three factors that explain 85.823% (greater than 80%) of the cumulative variance. Factor1 has eigenvalue above 1. Factor2 and factor3 have eigenvalues below 1. Scale Free Metric (S_ud), Assortativity (R_ud) and Small World Metric (SMSP_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.473. Scale Free Metric (S_ud), Assortativity (R_ud) and Small World Metric (SMSP_ud) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.4.7.1.2.3 Network Flow (MV2)

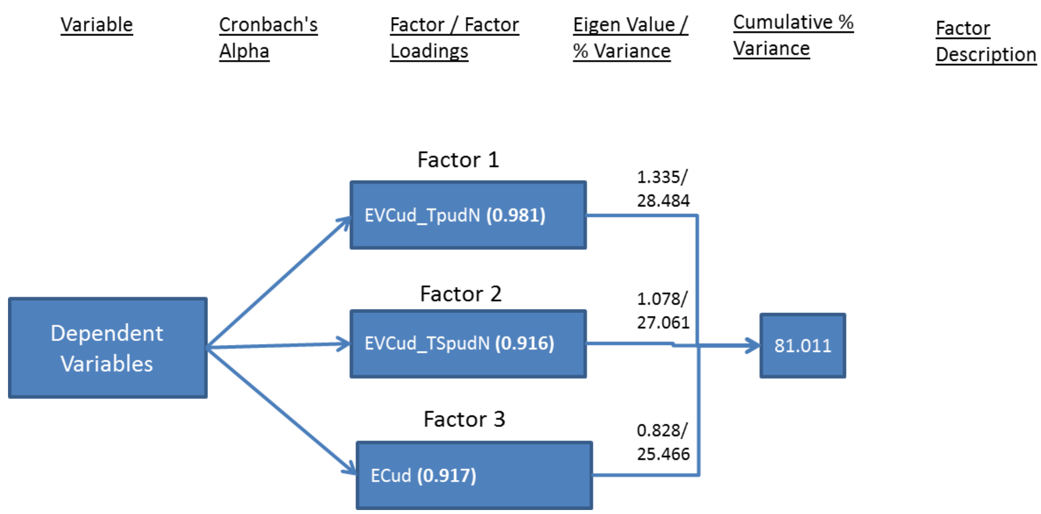
Figure 16: Factor Analysis of Network Flow Variables



The factor analysis generated two factors that explain 94.436% (greater than 80%) of the cumulative variance. Factor1 has eigenvalues above 1. Factor2 has eigenvalue below 1. Graph Diameter (GD_ud), Total Paths (Tpaths_ud) Average Path Length (AvgPL_ud) and Average Geodesic Length (AvgGL_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.937. Hence, they should be considered as a factor.

A.4.7.1.2.4 Dependent Variables

Figure 17: Factor Analysis of Dependent Variables



Factor analysis generated three factors that explain 81.011% (greater than 80%) of cumulative variance. All variables load independently. No significant factors were formed.

A.4.7.1.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Sports.pdf".

A.4.7.1.3.1 Impact of Network Structure on Network Flow

Table 2: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud
Sports	(0.522/0.000) [3,4]	(0.534/0.000) [3,4,5]	(0.462/0.000) [1,2,3,5]	(0.524/0.000)[3,4]	(0.337/0.000) [1,4]

Table 2 shows that the network structure variables have a significant impact on the network flow variables. Network structure variables explain 52.2%, 53.4%, 46.2%, 52.4% and 33.7% variation in Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud), Average Path Length (AvgPL_ud), and Average Geodesic Length (AvgGL_ud), respectively.

A.4.7.1.3.2 Impact of Network Flow on Network Structure

Table 3: Impact of Network Flow on Network Structure

Predictors: **(6) GD_ud, (7) Tpaths_ud, (8) TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpudN	PL_TSpudN	S_ud	R_ud	SMSP_ud
Sports	(0.083/0.000)[6,9]	(0.146/0.000)[10]	(0.572/0.000)[8,10]	(0.474/0.000)[9]	(0.120/0.000)[9]

Table 3 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 8.3%, 14.6%, 57.2%, 47.4% and 12% variation in the PL_TpudN, PL_TSpudN, S_ud, R_ud and SMSP_ud, respectively.

A.4.7.1.3.3 Impact of Network Structure on Network Phenomenon

Table 4: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Sports	(0.133/0.000) [5]	(0.032/0.049)[4]	(0.435/0.000) [1,4]	(0.045/0.025)[2]

Table 4 shows that the network structure variable Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN), Eigenvector Centrality with respect to Paths per Node (EVCud_TudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 13.3%, 3.2%, 43.5% and 4.5% variation respectively. The impact of network flow variables on Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.1.3.4 Impact of Network Flow on Network Phenomenon

Table 5: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_ud, (7)Tpaths_ud, (8)TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Sports	(0.054/0.015) [7]	NA	(0.167/0.000) [6]	(0.0539/0.000) [9,10]

Table 5 shows that the network flow variable impacts Eigenvector Centralization (EC_ud), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 5.4%, 16.7% and 53.9% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud) is not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.1.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 6: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud, (6),GD_ud (7) Tpaths_ud (8), TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud, (11) Nodes, (12) Edges_ud, (13) Den_ud, (14) CC_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Sports	(0.181/0.000) [14,3]	(0.032/0.049)[4]	(0.476/0.000) [1,8,6]	(0.631/0.000) [10,4,1,7]

Table 6 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN), Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TpudN) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN), explaining 18.1%, 3.2%, 47.6 and 63.1% variation respectively. The collective impact of independent variables and the moderating variables on Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) is not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.2 The Directed Network

A.4.7.2.1 Correlation Analysis

Significant correlations coefficients for directed network are shown below in table 7. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 7: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpdN	PL_TSpdN	S_d	EVCd_TpdN
Edges_d	Pearson	.999**	1												
	Sig. (2-ta	.000													
	N	91	91												
GD_d	Pearson	-.109	-.082	.540**	-.181	.447**	1								
	Sig. (2-ta	.302	.437	.000	.086	.000									
	N	91	91	91	91	91	91								
Tpaths_d	Pearson	.381**	.404**	.477**	-.561**	.563**	.666**	1							
	Sig. (2-ta	.000	.000	.000	.000	.000	.000								
	N	91	91	91	91	91	91	91							
TSpaths_d	Pearson	.511**	.529**	.303**	-.719**	.266*	.488**	.926**	1						
	Sig. (2-ta	.000	.000	.004	.000	.011	.000	.000							
	N	91	91	91	91	91	91	91	91						
AvgPL_d	Pearson	-.022	.003	.585**	-.179	.707**	.846**	.809**	.565**	1					
	Sig. (2-ta	.833	.976	.000	.089	.000	.000	.000	.000						
	N	91	91	91	91	91	91	91	91	91					
AvgGL_d	Pearson	-.033	-.012	.544**	-.264*	.532**	.790**	.834**	.708**	.900**	1				
	Sig. (2-ta	.758	.909	.000	.011	.000	.000	.000	.000	.000					
	N	91	91	91	91	91	91	91	91	91	91				
PL_TSpdN	Pearson	-.347**	-.337**	.166	.200	-.008	.365**	-.058	-.117	.141	.190	.702**	1		
	Sig. (2-ta	.001	.001	.116	.057	.944	.000	.584	.271	.181	.071	.000			
	N	91	91	91	91	91	91	91	91	91	91	91	91		
S_d	Pearson	-.338**	-.325**	.297**	.638**	.328**	.238*	-.095	-.319**	.244*	.106	.411**	.431**	1	
	Sig. (2-ta	.001	.002	.004	.000	.002	.023	.372	.002	.020	.316	.000	.000		
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
R_d	Pearson	-.337**	-.315**	.282**	.358**	.196	.423**	-.011	-.187	.261*	.133	.655**	.622**	.827**	
	Sig. (2-ta	.001	.002	.007	.000	.063	.000	.919	.076	.013	.209	.000	.000	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
SMSP_d	Pearson	.004	.019	.401**	-.057	.859**	.444**	.546**	.292**	.682**	.509**	.053	.047	.412**	
	Sig. (2-ta	.972	.861	.000	.594	.000	.000	.000	.005	.000	.000	.617	.660	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
PL_EVCdN	Pearson	-.100	-.077	.659**	-.090	.269**	.648**	.444**	.313**	.582**	.544**	.393**	.347**	.240*	
	Sig. (2-ta	.347	.471	.000	.397	.010	.000	.000	.003	.000	.000	.000	.001	.022	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
EVCd_TSpdN	Pearson	.093	.087	-.174	.095	.108	-.150	.017	.003	-.027	-.075	-.214*	-.344**	-.128	.968**
	Sig. (2-ta	.382	.412	.098	.371	.306	.157	.873	.979	.796	.477	.041	.001	.227	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 7 shows that nodes (Nodes) and ties (Edges_d) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Graph Diameter (GD_d) correlates positively with Reciprocity. Total Paths (Tpaths_d) correlates positively with Clustering Coefficient (CC_d) and Graph Diameter (GD_d) but correlates negatively with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes) and ties (Edges_d). Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) correlate positively with each other. Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Reciprocity, Clustering Coefficient (CC_d), and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpdN) correlates positively with Paths Power Law Distribution per Node (PL_TpdN). Scale Free Metric (S_d) seems to share a positive relationship with Density (Den_d). Assortativity (R_d) shares a positive relationship with Shortest Paths Power Law Distribution per Node (PL_TSpdN), Paths Power Law Distribution per Node (PL_TpdN) and Scale Free Metric (S_d). Small World Metric (SMSP_d) is strongly correlated with clustering coefficient and shares a positive relationship with Total Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) shares a positive correlation with Reciprocity, Graph Diameter (GD_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d).

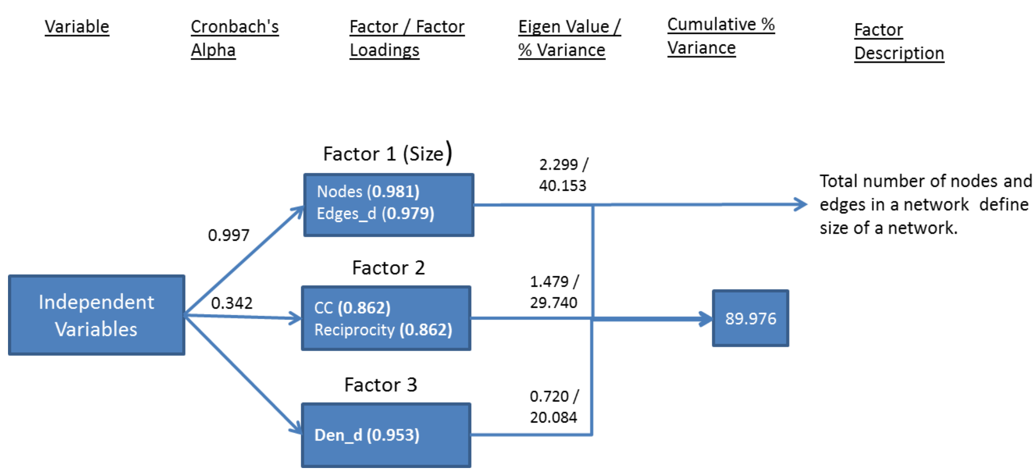
Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) and Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) correlate strongly with each other.

A.4.7.2.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.4.7.2.2.1 Independent Variables

Figure 18: Factor Analysis of Independent Variables

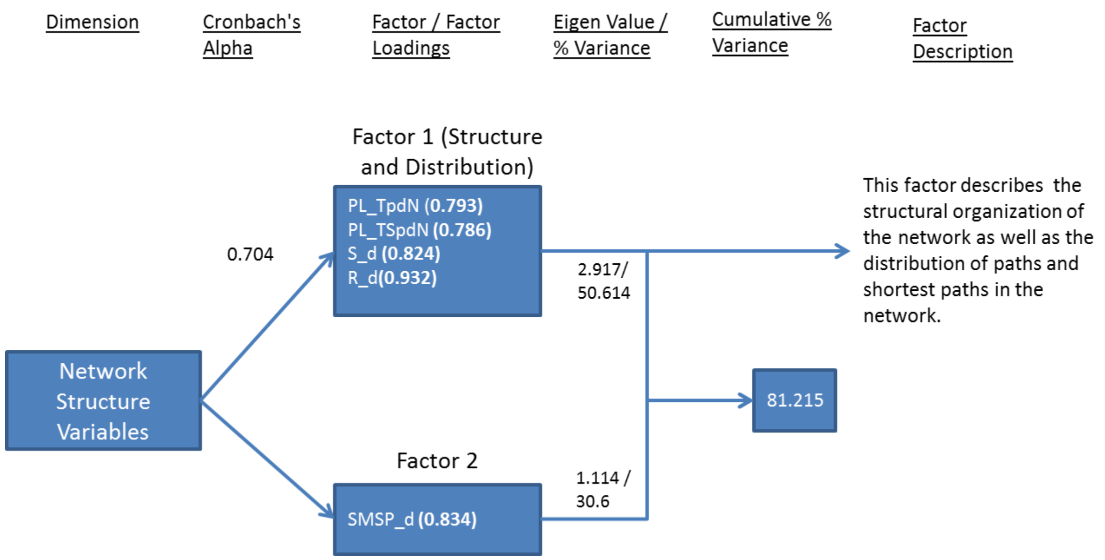


The factor analysis generated three factors that explain 89.976% (greater than 80%) of the cumulative variance. Factor1 and factor2 have eigenvalues above one. Factor3 has eigenvalue below 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) has a significant factor loading in factor 3. Reciprocity and Clustering Coefficient (CC_d) have a significant loading in factor

2.Cronbach's alpha for factor1 has a value of 0.997. This means Nodes and ties (Edges_d) are measuring same construct within factor 1. Hence, I name factor 1 as "Size". Cronbach's alpha for factor2 has a value of 0.342. This means reciprocity and clustering coefficient s are not measuring same construct within factor2.

A.4.7.2.2.2 Network Structure (MV1)

Figure 19: Factor Analysis of Network Structure Variables

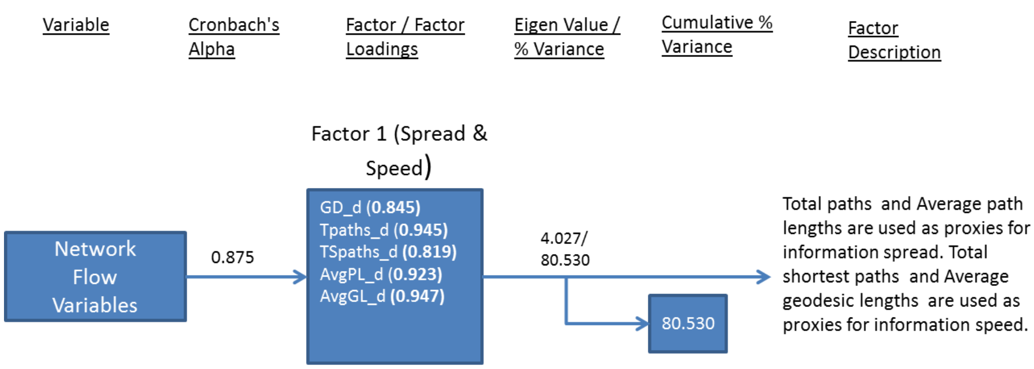


Factor analysis generated two factors that explain 81.215% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Paths Power Law Distribution per Node (PL_TpdN), Shortest Paths Power Law Distribution per Node (PL_TSpdN), Assortativity (R_d) and Scale Free Metric (S_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.704. . Paths Power Law Distribution per Node (PL_TpdN), Shortest Paths Power Law Distribution per Node

(PL_TSpdN), Assortativity (R_d) and Scale Free Metric (S_d) are measuring same construct within factor 1. Hence, they should be considered as a factor. All other variables load independently.

A.4.7.2.2.3 Network Flow (MV2)

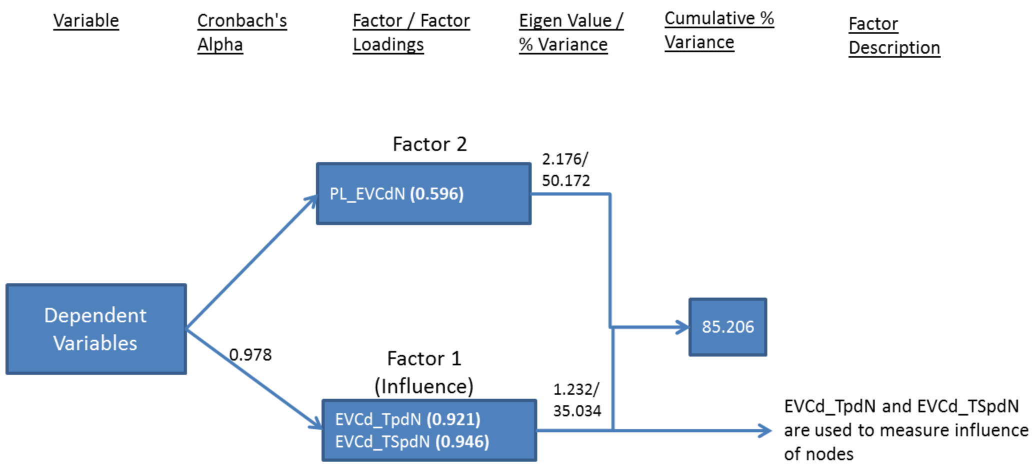
Figure 20: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 80.530% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.875. Factor 1 is named as "Spread and Speed".

A.4.7.2.2.4 Dependent Variables

Figure 21: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 85.206% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCd_TpD) and Shortest Paths (EVCd_TSpD) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.978. I name the factor 1 as "Influence" as both Eigenvector Centralities with respect to Paths (EVCd_TpD) and Shortest Paths (EVCd_TSpD), are being used measure of influence.

A.4.7.2.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Sports.pdf"..

A.4.7.2.3.1 Impact of Network Structure on Network Flow

Table 8: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Sports	(0.312/0.000) [1,5]	(0.336/0.000) [3,5]	(0.303/0.001) [3,5]	(0.426/0.000) [1,5]	(0.256/0.000) [2,3,5]

Table 8 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 31.2%, 33.6%, 30.3%, 42.6% and 25.6% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Total Shortest Paths (TSpaths_d) is not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.2.3.2 Impact of Network Flow on Network Structure

Table 9: Impact of Network Flow on Network Structure

Predictors: **(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpdN	PL_TSpdN	S_d	R_d	SMSP_d
Sports	(0.350/0.000) [6,7]	(0.281/0.000) [6,7]	(0.472/0.000) [8,9]	(0.386/0.000) [6,8]	NA

Table 9 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 35%, 28.1%, 47.2%, and 38.6% variation in the PL_TpdN, PL_TSpdN, S_d, and R_ud, respectively.

A.4.7.2.3.3 Impact of Network Structure on Network Phenomenon

Table 10: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Sports	NA	(0.211/0.000) [1,5]	(0.077/0.005) [2]	(0.108/0.001)[2]

Table 10 shows that the network structure variable impacts Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Shortest per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN), explaining 21.1%, 7.7% and 10.8% variation respectively. The impact of network flow variables on Eigenvector Centrality with respect to Total Shortest per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.2.3.4 Impact of Network Flow on Network Phenomenon

Table 11: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Sports	(0.229/0.000) [10]	(0.413/0.000)[6]	NA	NA

Table 11 shows that the network structure variable impacts Eigenvector Centralization (EC_d), and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), explaining 22.9% and 41.3% variation respectively.

A.4.7.2.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 12: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d, (6)GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCud_TpdN	EVCud_TSpdN
Sports	(0.229/0.000) [10]	(0.077/0.005)[2]	(0.108/0.001) [2]	(0.561/0.000) [6,15]

Table 12 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 22.9%, 7.7%, 10.8%and 56.1% variation respectively. The collective impact of independent variables and the moderating variables on Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) and Eigenvector Centrality with respect to Total Paths per Node

(EVCd_TpdN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.3 The Consumption Network

A.4.7.3.1 Correlation Analysis

Significant correlations coefficients for consumption network are shown below in table 13. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled "Correlations.pdf".

Table 13: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpinN	PL_TSpinN	S_con	EVCin_TpinN
Edges_d	Pearson	.999**	1												
	Sig. (2-ta	.000													
	N	91	91												
GD_d	Pearson	-.109	-.082	.540**	-.181	.447**	1								
	Sig. (2-ta	.302	.437	.000	.086	.000									
	N	91	91	91	91	91	91								
Tpaths_d	Pearson	.381**	.404**	.477**	-.561**	.563**	.666**	1							
	Sig. (2-ta	.000	.000	.000	.000	.000	.000								
	N	91	91	91	91	91	91	91							
TSpaths_d	Pearson	.511**	.529**	.303**	-.719**	.266*	.488**	.926**	1						
	Sig. (2-ta	.000	.000	.004	.000	.011	.000	.000							
	N	91	91	91	91	91	91	91	91						
AvgPL_d	Pearson	-.022	.003	.585**	-.179	.707**	.846**	.809**	.565**	1					
	Sig. (2-ta	.833	.976	.000	.089	.000	.000	.000	.000						
	N	91	91	91	91	91	91	91	91	91					
AvgGL_d	Pearson	-.033	-.012	.544**	-.264*	.532**	.790**	.834**	.708**	.900**	1				
	Sig. (2-ta	.758	.909	.000	.011	.000	.000	.000	.000	.000					
	N	91	91	91	91	91	91	91	91	91	91				
PL_TSpinN	Pearson	-.302**	-.294**	.190	.101	.034	.361**	.001	-.054	.161	.193	.779**	1		
	Sig. (2-ta	.004	.005	.071	.342	.752	.000	.991	.610	.128	.066	.000			
	N	91	91	91	91	91	91	91	91	91	91	91	91		
R_con	Pearson	-.053	-.058	.179	-.189	-.007	.098	.241*	.311**	.150	.361**	.177	.185	.988**	
	Sig. (2-ta	.616	.588	.089	.072	.949	.357	.021	.003	.157	.000	.094	.079	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
SMSP_d	Pearson	.004	.019	.401**	-.057	.859**	.444**	.546**	.292**	.682**	.509**	.057	.076	-.028	
	Sig. (2-ta	.972	.861	.000	.594	.000	.000	.000	.005	.000	.000	.593	.475	.792	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
ECin	Pearson	.059	.046	-.526**	.213*	-.120	-.426**	-.431**	-.419**	-.401**	-.504**	-.240*	-.319**	-.271**	
	Sig. (2-ta	.580	.665	.000	.043	.257	.000	.000	.000	.000	.000	.022	.002	.009	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
PL_EVCinN	Pearson	-.084	-.066	.533**	-.054	.183	.385**	.284**	.219*	.308**	.285**	.246*	.291**	.108	
	Sig. (2-ta	.431	.533	.000	.609	.082	.000	.006	.037	.003	.006	.019	.005	.309	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	
EVCin_TSpinN	Pearson	-.262*	-.262*	-.122	.422**	.044	-.066	-.256*	-.338**	-.081	-.161	.123	.110	-.039	.982**
	Sig. (2-ta	.012	.012	.251	.000	.677	.536	.014	.001	.447	.128	.246	.299	.716	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 13 shows that nodes (Nodes) and ties (Edges_d) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Graph Diameter (GD_d) correlates positively with Reciprocity. Total Paths (Tpaths_d) correlates positively with Clustering Coefficient (CC_d) and Graph Diameter (GD_d) but correlates negatively with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes) and ties (Edges_d). Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) correlate positively with each other. Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Reciprocity, Clustering Coefficient (CC_d), and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpinN) correlates positively with Paths Power Law Distribution per Node (PL_TpinN). Scale Free Metric (S_d) seems to share a positive relationship with Density (Den_d). Assortativity (R_con) shares a positive relationship with Shortest Paths Power Law Distribution per Node (PL_TSpinN), Paths Power Law Distribution per Node (PL_TpinN) and Scale Free Metric (S_con). Small World Metric (SMSP_d) is strongly correlated with clustering coefficient and shares a positive relationship with Total Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN) shares a positive correlation with Reciprocity, Graph Diameter (GD_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d).

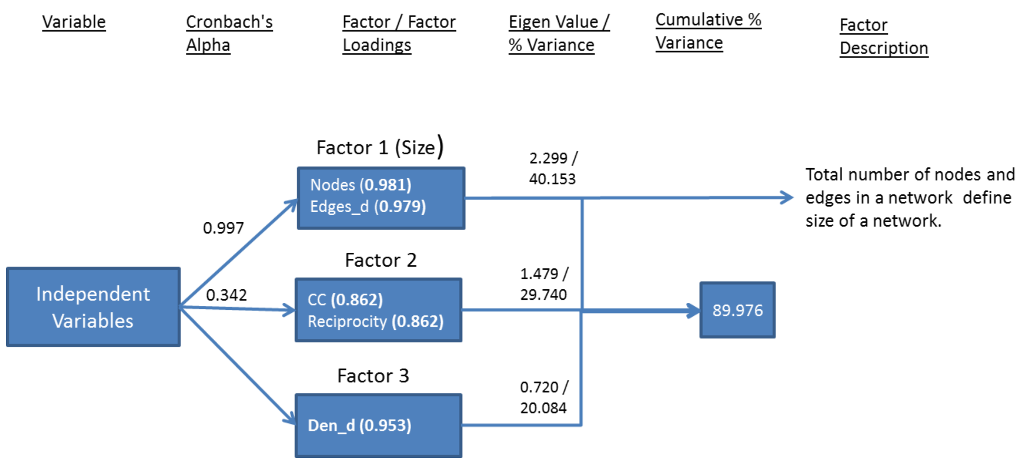
Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN) and Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) correlate strongly with each other.

A.4.7.3.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.4.7.3.2.1 Independent Variables

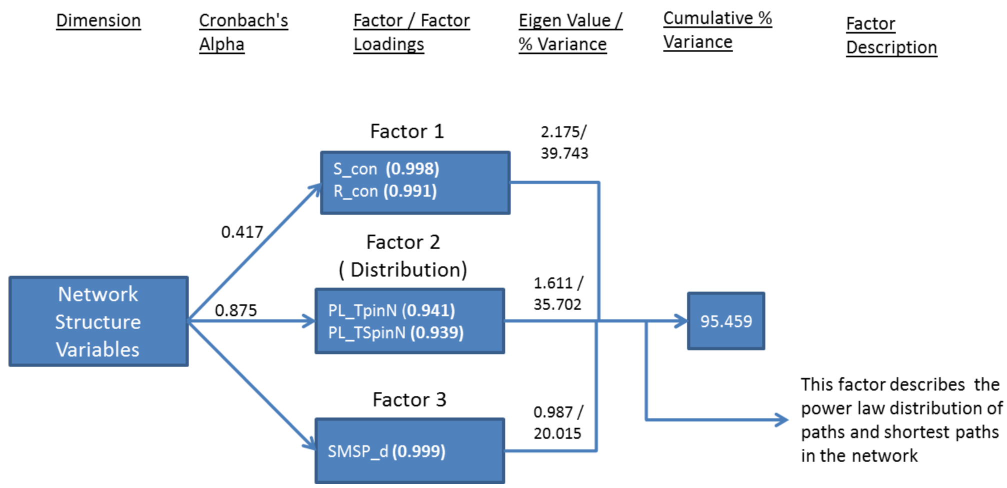
Figure 22: Factor Analysis of Network Structure Variables



The factor analysis generated three factors that explain 89.976% (greater than 80%) of the cumulative variance. Factor1 and factor2 have eigenvalues above one. Factor3 has eigenvalue below 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) has a significant factor loading in factor 3. Reciprocity and Clustering Coefficient (CC_d) have a significant loading in factor 2.Cronbach’s alpha for factor1 has a value of 0.997. This means Nodes and ties (Edges_d) are measuring same construct within factor 1. Hence, I name factor 1 as “Size”. Cronbach’s alpha for factor2 has a value of 0.342. This means reciprocity and clustering coefficient s are not measuring same construct within factor2.

A.4.7.3.2.2 Network Structure (MV1)

Figure 23: Factor Analysis of Network Structure Variables

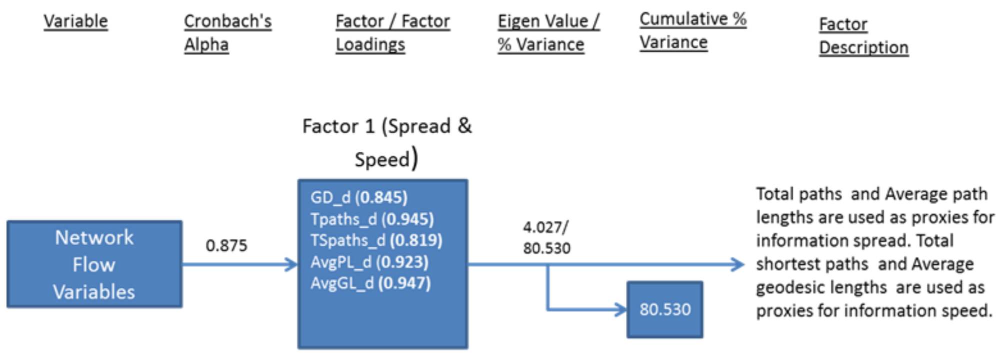


Factor analysis generated three factors that explain 84.534% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 has eigenvalue little less than 1. Assortativity (R_con) and Scale Free Metric (SMSP_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.417. . Assortativity (R_con) and Scale Free Metric (SMSP_d) are not measuring same construct within factor 1. Hence, they should not be considered as a factor.

PL_TpdN and PL_TSpdN have significant factor loadings in factor 2. Cronbach's alpha for factor1 has a value of 0.893. PL_TpdN and PL_TSpdN are measuring same construct within factor 2. Hence, they should be considered as a factor. All other variables load independently. Factor2 is named as "Distribution".

A.4.7.3.2.3 Network Flow (MV2)

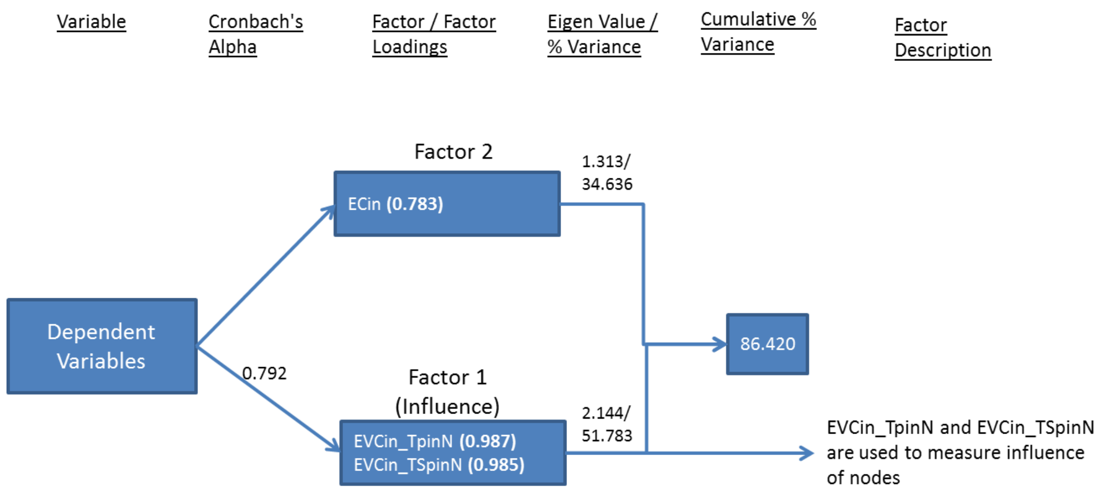
Figure 24: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 80.530% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.875. Factor 1 is named as "Spread and Speed".

A.4.7.3.2.4 Dependent Variables

Figure 25: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 86.420% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.792. I name the factor1 as "Influence" as both, Eigenvector Centralities with respect to Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN), are being used measure of influence.

A.4.7.3.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Sports.pdf".

A.4.7.3.3.1 Impact of Network Structure on Network Flow

Table 14: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Sports	(0.287/0.000)1,5]	(0.267/0.000) [4,5]	(0.162/0.000) [4,5]	(0.450/0.000) [1,5]	(0.327/0.000) [4,5]

Table 14 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 28.7%, 26.7%, 16.2%, 45% and 32.7% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively.

A.4.7.3.3.2 Impact of Network Flow on Network Structure

Table 15: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpinN	PL_TSpinN	S_con	R_con	SMSP_d
Sports	(0.218/0.000) [6,7]	(0.216/0.000) [6,7]	(0.049/0.020) [10]	(0.102/0.000)[8]	NA

Table 15 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 21.8%, 21.6%, 4.9% and 10.2 % variation in the PL_TpinN, PL_TSpinN, S_con, and R_con, respectively. The impact of network flow variables on S_con is not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.3.3.3 Impact of Network Structure on Network Phenomenon

Table 16: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Sports	(0.117/0.002) [2,3]	(0.074/0.005)[2]	NA	NA

Table 16 shows that the network structure variable impacts Eigenvector Centralization (EC_in) and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), explaining 11.7% and 7.4% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_in) and Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 17: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Sports	(0.245/0.000)[10]	(0.139/0.000)[6]	(0.112/0.002)[8,9]	(0.104/0.001)[8]

Table 17 shows that the network structure variable impacts Eigenvector Centralization (EC_in), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 24.5%, 13.9%, 11.2 and 10.4% variation respectively. The impact of network flow variables on Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.4.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 18: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d, (6)GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Sports	(0.441/0.000) [15,8,2,14]	(0.306/0.000) [2,15]	(0.144/0.000) [13]	(0.169/0.000) [13]

Table 18 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_in), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 41.1%, 30.6%, 14.4% and 16.9% variation respectively.

A.4.7.4 The Propagation Network

A.4.7.4.1 Correlation Analysis

Significant correlations coefficients for propagation network are shown below in table 19. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 19: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	CC_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpoutN	PL_TSpoutN	S_pro	ECout	EVCout_TpoutN
Edges_d	Pearson	.999**	1													
	Sig. (2-ta	.000														
	N	91	91													
GD_d	Pearson	-.109	-.082	.540**	-.181	.447**	1									
	Sig. (2-ta	.302	.437	.000	.086	.000										
	N	91	91	91	91	91	91									
Tpaths_d	Pearson	.381**	.404**	.477**	-.561**	.563**	.666**	1								
	Sig. (2-ta	.000	.000	.000	.000	.000	.000									
	N	91	91	91	91	91	91	91								
TSpaths_d	Pearson	.511**	.529**	.303**	-.719**	.266*	.488**	.926**	1							
	Sig. (2-ta	.000	.000	.004	.000	.011	.000	.000								
	N	91	91	91	91	91	91	91	91							
AvgPL_d	Pearson	-.022	.003	.585**	-.179	.707**	.846**	.809**	.565**	1						
	Sig. (2-ta	.833	.976	.000	.089	.000	.000	.000	.000							
	N	91	91	91	91	91	91	91	91	91						
AvgGL_d	Pearson	-.033	-.012	.544**	-.264*	.532**	.790**	.834**	.708**	.900**	1					
	Sig. (2-ta	.758	.909	.000	.011	.000	.000	.000	.000	.000						
	N	91	91	91	91	91	91	91	91	91	91					
PL_TSpoutN	Pearson	-.150	-.138	.164	.083	-.011	.212*	.028	-.005	.091	.100	.686**	1			
	Sig. (2-ta	.156	.194	.121	.434	.915	.044	.795	.963	.389	.345	.000				
	N	91	91	91	91	91	91	91	91	91	91	91	91			
S_pro	Pearson	-.086	-.069	.320**	.290**	.553**	.464**	.344**	.127	.518**	.452**	.168	.175	1		
	Sig. (2-ta	.416	.516	.002	.005	.000	.000	.001	.231	.000	.000	.111	.098			
	N	91	91	91	91	91	91	91	91	91	91	91	91	91		
R_pro	Pearson	-.227*	-.204	.371**	.211*	.456**	.586**	.307**	.102	.504**	.436**	.422**	.434**	.784**		
	Sig. (2-ta	.030	.053	.000	.045	.000	.000	.003	.334	.000	.000	.000	.000	.000		
	N	91	91	91	91	91	91	91	91	91	91	91	91	91		
SMSP_d	Pearson	.004	.019	.401**	-.057	.859**	.444**	.546**	.292**	.682**	.509**	-.032	-.015	.307**		
	Sig. (2-ta	.972	.861	.000	.594	.000	.000	.000	.005	.000	.000	.760	.887	.003		
	N	91	91	91	91	91	91	91	91	91	91	91	91	91		
ECout	Pearson	.059	.046	-.526**	.213*	-.120	-.426**	-.431**	-.419**	-.401**	-.504**	-.384**	-.230*	-.061	1	
	Sig. (2-ta	.580	.665	.000	.043	.257	.000	.000	.000	.000	.000	.000	.028	.567		
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91	
PL_EVCoutN	Pearson	-.093	-.073	.498**	-.103	.099	.479**	.298**	.252*	.319**	.340**	.489**	.257*	.286**	-.513**	
	Sig. (2-ta	.380	.493	.000	.333	.348	.000	.004	.016	.002	.001	.000	.014	.006	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91	
EVCout_TpoutN	Pearson	.293**	.296**	.085	-.345**	.190	.204	.491**	.515**	.318**	.372**	-.074	-.132	.218*	-.224*	1
	Sig. (2-ta	.005	.004	.424	.001	.072	.052	.000	.000	.002	.000	.484	.214	.038	.033	
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91
EVCout_TSpoutN	Pearson	.300**	.305**	.133	-.348**	.192	.216*	.530**	.553**	.345**	.411**	-.084	-.133	.239*	-.218*	.971**
	Sig. (2-ta	.004	.003	.207	.001	.069	.039	.000	.000	.001	.000	.427	.208	.023	.038	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

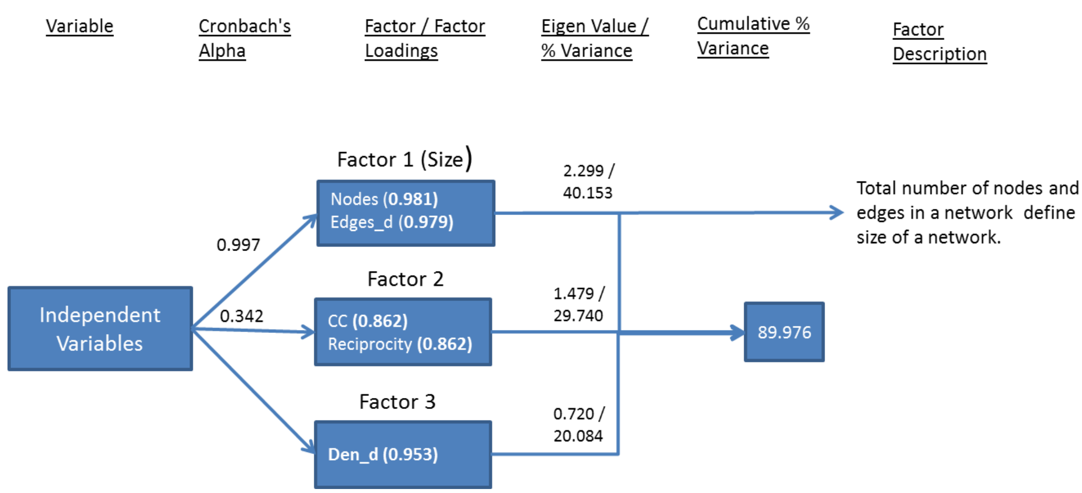
Table 19 shows that nodes (Nodes) and ties (Edges_ud) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Graph Diameter (GD_d) correlates positively with Reciprocity. Total Paths (Tpaths_d) correlates positively with Clustering Coefficient (CC_d) and Graph Diameter (GD_d) but correlates negatively with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of (Nodes) and ties (Edges_ud). Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) correlate positively with each other. Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpoutN) correlates positively with Paths Power Law Distribution per Node (PL_TpoutN). Scale Free Metric (S_pro) correlated with Density (Den_d) and average path length. Assortativity shares a positive relationship with graph diameter, Average Path Length (AvgPL_d) and Scale Free Metric (S_pro). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d), Total Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN) correlated negatively with Eigenvector Centralization (ECout). Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSPoutN) and Eigenvector Centrality with respect to Total Paths per Node (EVCoutN_TPoutN) correlate strongly with each other and correlate positively with Total Shortest Paths (TSpaths_d).

A.4.7.4.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.4.7.4.2.1 Independent Variables

Figure 26: Factor Analysis of Independent Variables

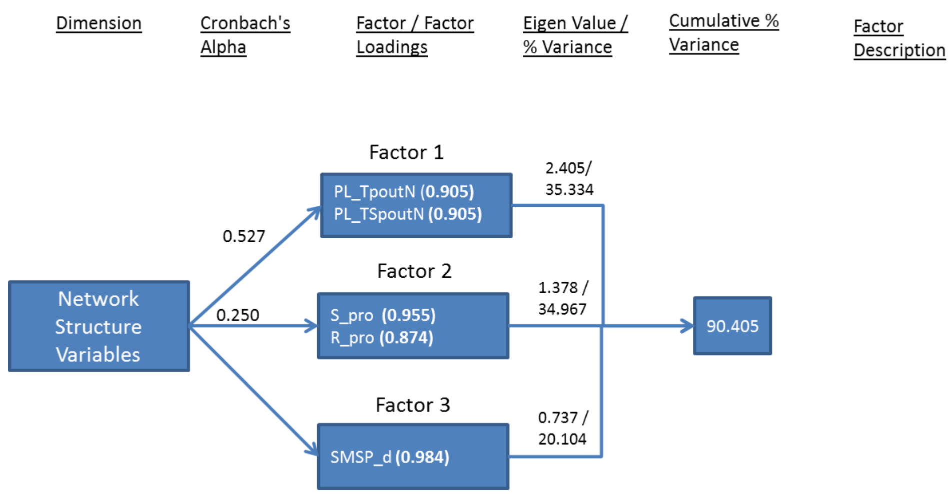


The factor analysis generated three factors that explain 89.976% (greater than 80%) of the cumulative variance. Factor1 and factor2 have eigenvalues above one. Factor3 has eigenvalue below 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Density (Den_d) has a significant factor loading in factor 3. Reciprocity and Clustering Coefficient (CC_d) have a significant loading in factor 2.Cronbach’s alpha for factor1 has a value of 0.997. This means Nodes and ties

(Edges_d) are measuring same construct within factor 1. Hence, I name factor 1 as "Size". Cronbach's alpha for factor2 has a value of 0.342. This means reciprocity and clustering coefficient s are not measuring same construct within factor2.

A.4.7.4.2.2 Network Structure (MV1)

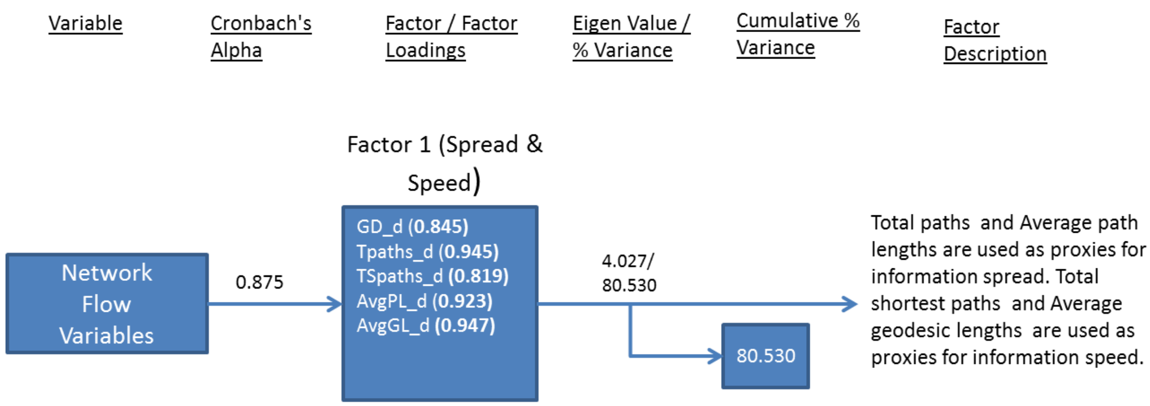
Figure 27: Factor Analysis of Network Structure Variables



Factor analysis generated three factors that explain 90.405% (greater than 80%) of cumulative variance**.** Cronbach's alpha for factor1 has a value of 0.250. Cronbach's alpha for factor2 has a value of 0.527. Therefore, all variables load independently.

A.4.7.4.2.3 Network Flow (MV2)

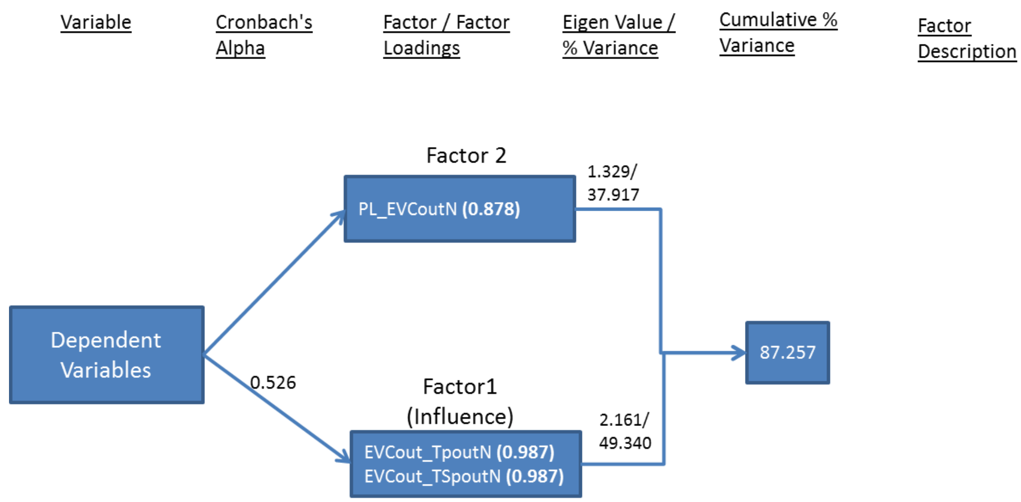
Figure 28: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 80.530% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.875. Factor 1 is named as "Spread and Speed".

A.4.7.4.2.4 Dependent Variables

Figure 29: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 87.257% (greater than 80%) of cumulative variance Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.526. Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN) are not measuring same construct within factor 1.

A.4.7.4.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Sports.pdf".

A.4.7.4.3.1 Impact of Network Structure on Network Flow

Table 20: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Sports	(0.369/0.000) [4]	(0.205/0.000) [5]	(0.075/0.005) [5]	(0.449/0.000) [4,5]	(0.342/0.000) [4,5]

Table 20 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 36.9%, 20.5%, 7.5%, 44.9% and 34.2% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.4.3.2 Impact of Network Flow on Network Structure

Table 21: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpoutN	PL_TSpoutN	S_pro	R_pro	SMSP_d
Sports	(0.057/0.013) [6]	(0.034/0.044) [6]	(0.115/0.001) [6]	(0.354/0.000) [6,8]	NA

Table 21 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 5.7%, 3.411.5%, and 35.4% variation in the PL_TpoutN, PL_TSpoutN, S_pro, and R_pro, respectively. The impact of network flow variables on PL_TpoutN, PL_TSpoutN and S_pro are not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.4.3.3 Impact of Network Structure on Network Phenomenon

Table 22: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Sports	(0.138/0.000) [1]	(0.292/0.000) [1,4]	(0.037/0.038) [3]	(0.046/0.023)[3]

Table 22 shows that the network structure variable impacts Eigen Centralization (ECout) Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 13.8%, 29.2%, 3.7% and 4.6% variation respectively. The impact of network flow variables on Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.4.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 23: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpout N
Sports	(0.245/0.000) [10]	(0.221/0.000)[6]	(0.256/0.000) [8]	(0.298/0.000)[8]

Table 23 shows that the network structure variable impacts Eigenvector Centralization (ECout), Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 24.5%, 22.1%, 25.6% and 29.8% variation respectively.

A.4.7.4.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 24: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d, (6),GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Sports	(0.462/0.000) [1,8,15]	(0.411/0.000) [1,6,15]	(0.256/0.000) [8]	(0.298/0.000)[8]

Table 24 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_out), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 46.2%, 41.1%, 25.6%and 29.8% variation respectively.

A.5 Case 5--Howto

A.5.1 Case Overview

Data for keyword “YouTube + howto” was collected over a period of 91 days (31/12/2013 to 31/03/2014). As shown in table 9, overall 10,856 tweets were collected, out of which 3,213 were broadcast tweets and 7,643 were engaged tweets respectively. Out of 7,643 engaged tweets only 4,299 tweets formed the largest community. Similarly, 10,557 daily unique people tweeted overall, out of which 4,802 daily unique people were engaged in broadcast activity whereas 6,475 daily unique people were engaged in conversations. Out of 6,475 daily unique people only 4,203 daily unique people formed the largest community. Data for the largest community was analyzed at a daily interval. The overall trends for the data are shown below in figure 1 and figure 2.

Figure 1: Overall Tweets

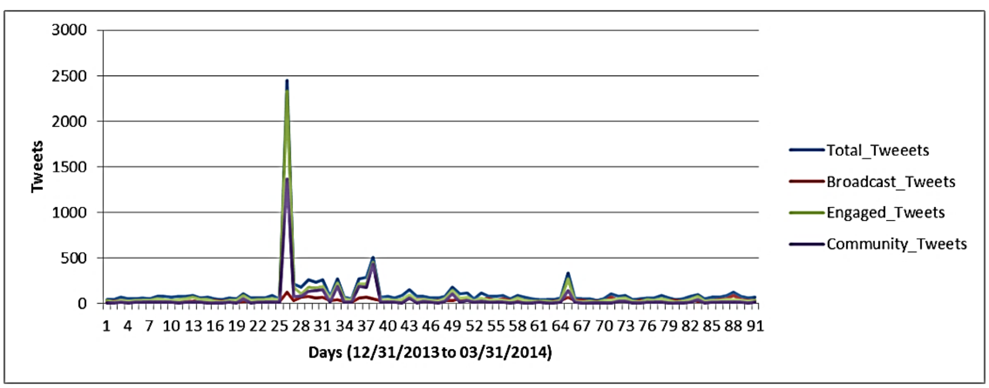


Fig.2: Overall People

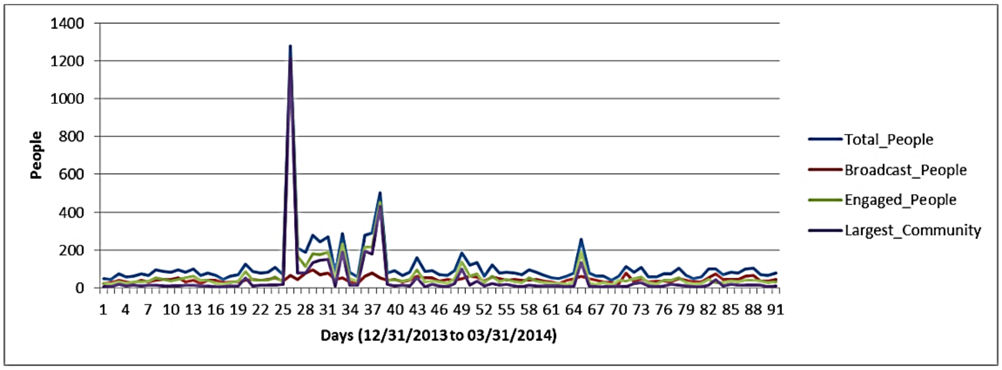


Figure 1 and figure 2 show that both the total tweets and total people involved are very dynamic and their magnitude changes on a daily basis. The maximum of the total number of daily tweets and the maximum of the total number of daily unique people observed on a single day (the daily uniques) are 2,448 and 1,279, respectively. Similarly, the minimum of the total number daily tweets and the minimum of the number daily unique are 37 and 42, respectively. The size of the largest community on a particular day and the largest number of community tweets on that day also seem to follow the trend of total people and total tweets. The largest number of daily community tweets and the largest number of daily unique people are 1,370 and 1,213, respectively. Similarly, the smallest number of daily community tweets and the smallest number of daily unique people are 4 and 5, respectively. As the total number of daily unique people tweeting increases, so does the size of the community. Most of the engaged people are engaged in the collective conversation forming the largest community.

A.5.2 Random or Not Random

As explained in section 4.4.1, in order to eliminate α- error and β- error, I compare the clustering coefficients of both undirected (CC_ud) and directed networks (CC_d) with their corresponding random (Erdös-Rényi, E-R) networks (CCudran, CCdran). If the clustering coefficients of the undirected and directed networks are equal to those of the E-R random network, then the directed and undirected networks are considered to be random, if they are not equal, then they are not random.

Figure 3: Comparison of Clustering Coefficients of Undirected Network with E-R Networks

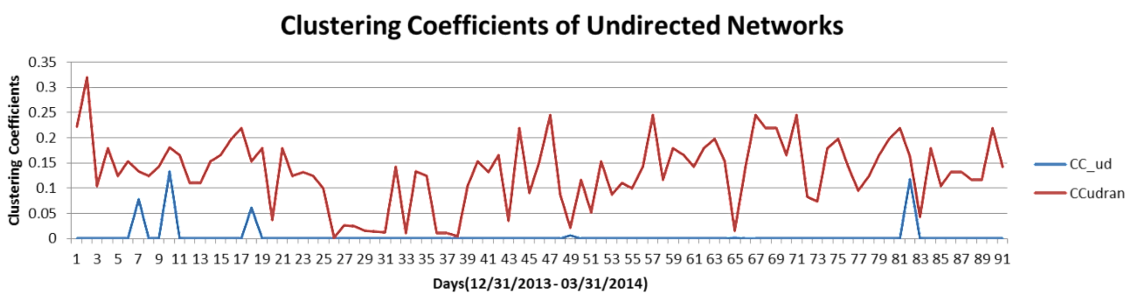
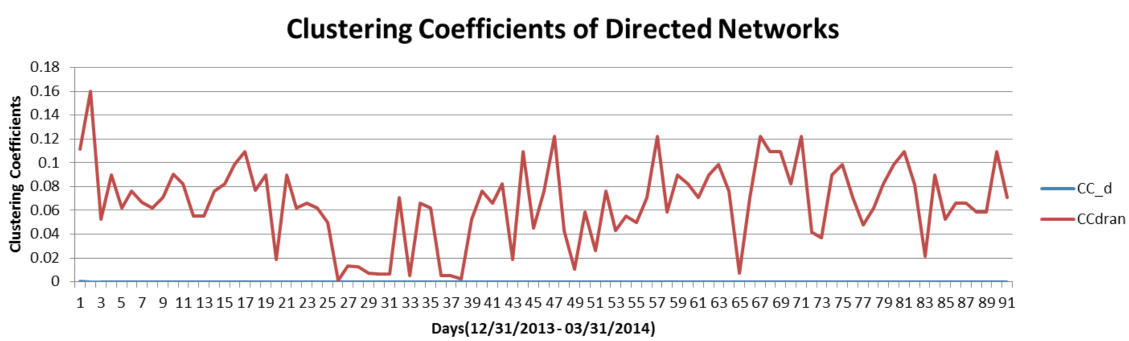


Figure 4: Comparison of Clustering Coefficients of Directed Network with E-R Networks



As seen in figure 3 and figure 4 clustering coefficients of the undirected and directed networks are zero for the most part. Therefore, they are random networks.

A.5.3. Independent Variables

The values of the independent variables for both the undirected and the directed network are shown in figure 5 below.

Figure 5: Independent Variables--(a) Nodes and Edges (Undirected and Directed networks), (b) Reciprocity (Directed Networks), (c) Density (Undirected and Directed Networks), (d) Clustering Coefficient Undirected Network, (e) Clustering Coefficient Directed Network.

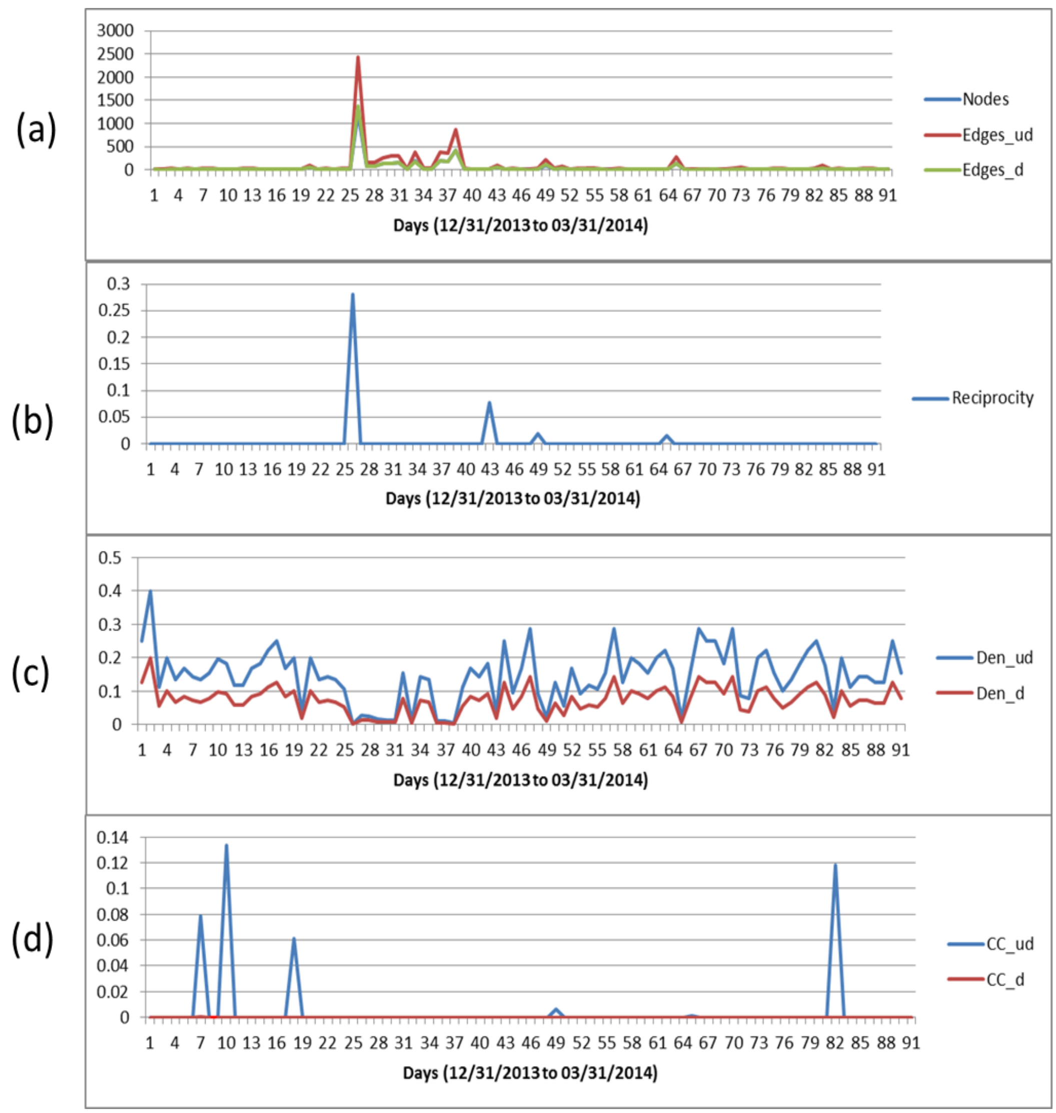
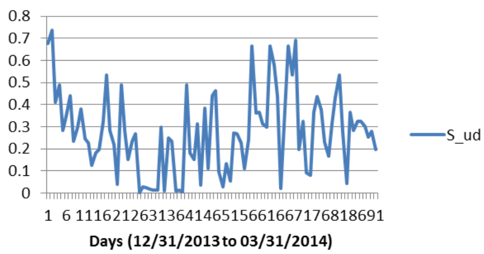


Figure 5 (a) shows that the number of directed ties (Edges_d) in the network and the total number of nodes (Nodes) overlap with each other. The numbers of undirected ties (Edges_ud) is greater than the number of directed ties (Edges_d), because in an undirected network every directed tie is considered to be symmetric. Therefore it is counted twice, except for the ones that are symmetric in a directed network. Reciprocity in Figure 5(b) indicates the presence of symmetric ties in a directed network (in an undirected network 100% are symmetric). The value of 0.01 is equal to 1% of all the ties. Figure 5(c) shows the difference between the densities of the undirected (Den_ud) and the directed networks (Den_d). The undirected network is denser than the directed network (Den_ud>Den_d). Figure 5(d) shows that the directed networks have higher Clustering Coefficients than the undirected networks (CC_d >CC_ud).

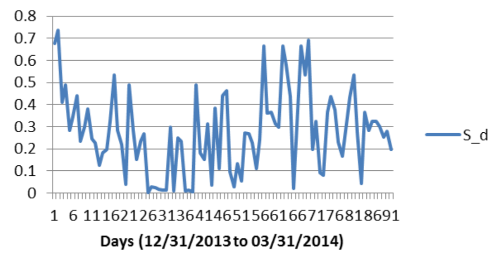
A.5.4 Network Structure Variables (MV1)

A.5.4.1 The Scale Free Metric

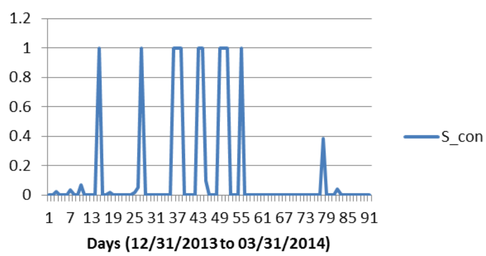
Figure 6: Scale Free Metric--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



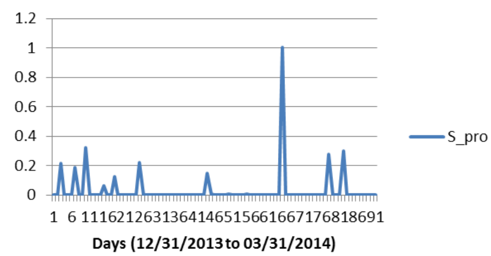
(a)



(b)



(c)



(d)

Figure 6 shows the Scale Free Metric for the undirected, directed, consumption and propagation networks (S_ud, S_d, S_con, S_pro). The Scale Free Metrics for the undirected (S_ud) and the directed network (S_d) are similar, but the Scale Free Metrics for the consumption (S_con) and propagation (S_pro) networks are very different. The propagation (S_pro) network is more scale free than the consumption network (S_con). The values of the scale free metric ranges between 0 and 1. When the values are closer to 1, it means that the networks are more scale free. Neither the directed nor the undirected network is scale free. This means that these networks may have hubs in them. However, there is not just one hub that is the center of the community. As shown in figure 6 (c) and figure 6 (d) the consumption network and the propagation network are scale free in some instances.

A.5.4.2 The Assortativity

Figure 7: Assortativity--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

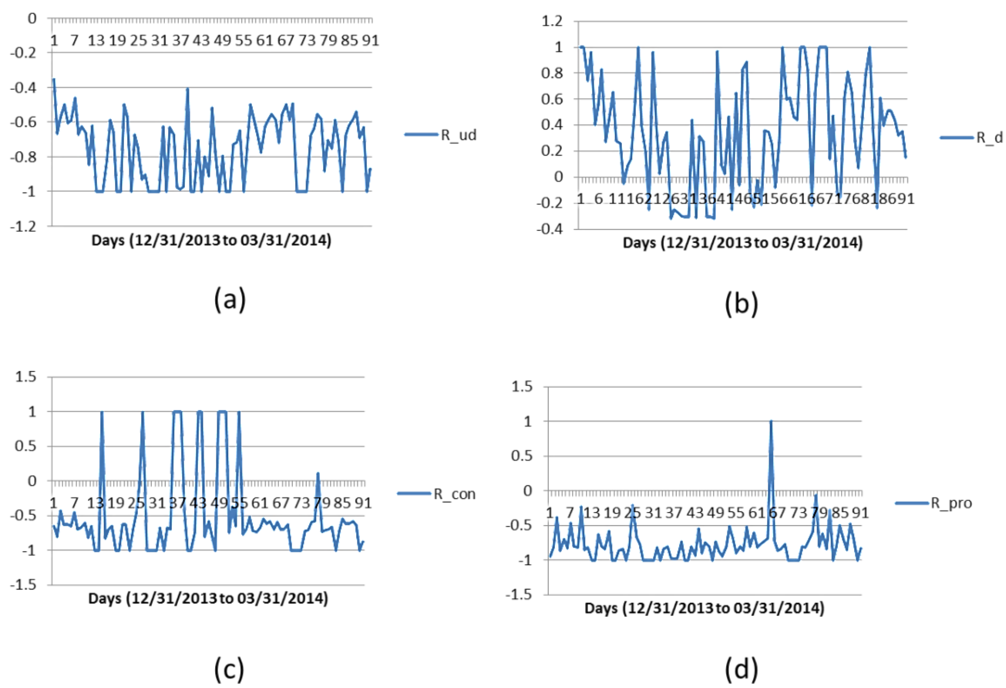
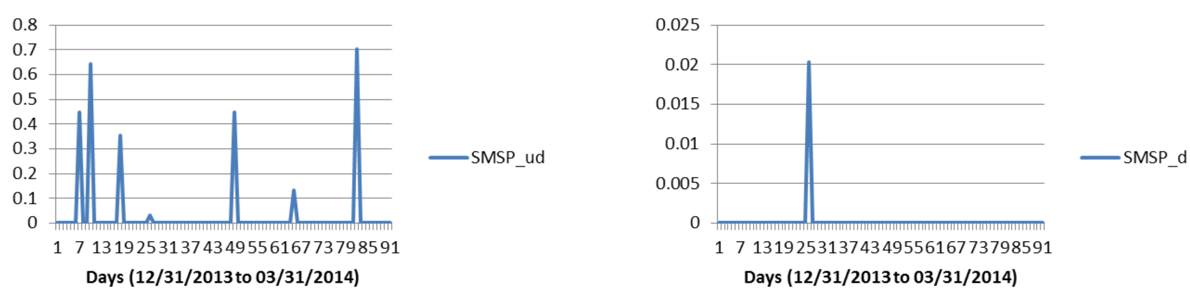


Figure 7 shows the assortativity metric for the undirected, directed, consumption and propagation networks (R_ud, R_d, R_con, R_Pro). The value of the assortativity metric ranges between -1 and 1. When the values are closer to -1, it means that networks are disassortative. The undirected network is more Disassortative than the directed network

(R_d > R_ud). Among the directed networks, the consumption network is more Disassortative than the propagation network (R_pro > R_con). Disassortative means that the nodes in the network connect to nodes that are very similar to themselves. This is true more so in the undirected network and in the consumption network than it is in the directed network and the propagation network. This implies that disassortativeness of consumption contributes more to the disassortativeness of the directed network than the disassortativeness of the propagation does.

A.5.4.3 The Small World Metric

Figure 8: Small World Metric -- (a) Undirected Network, (b) Directed Network.



(a) (b)

Figure 8 shows the Small World Metric for the undirected (SMSP_ud) and directed networks (SMSP_d). The Small World Metrics for the consumption and propagation networks are the same as the ones for the directed network. The directed networks don't show any small world behavior. Contrary to the directed networks, undirected networks show some small world behavior but not significantly enough. This

means that in undirected networks there are more nodes that act as hubs that facilitate communication between other nodes of the network.

A.5.4.4 Paths and Shortest Paths Power law Distribution per Node

Figure 9: Power Law Distribution of Paths and Shortest Paths in (a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

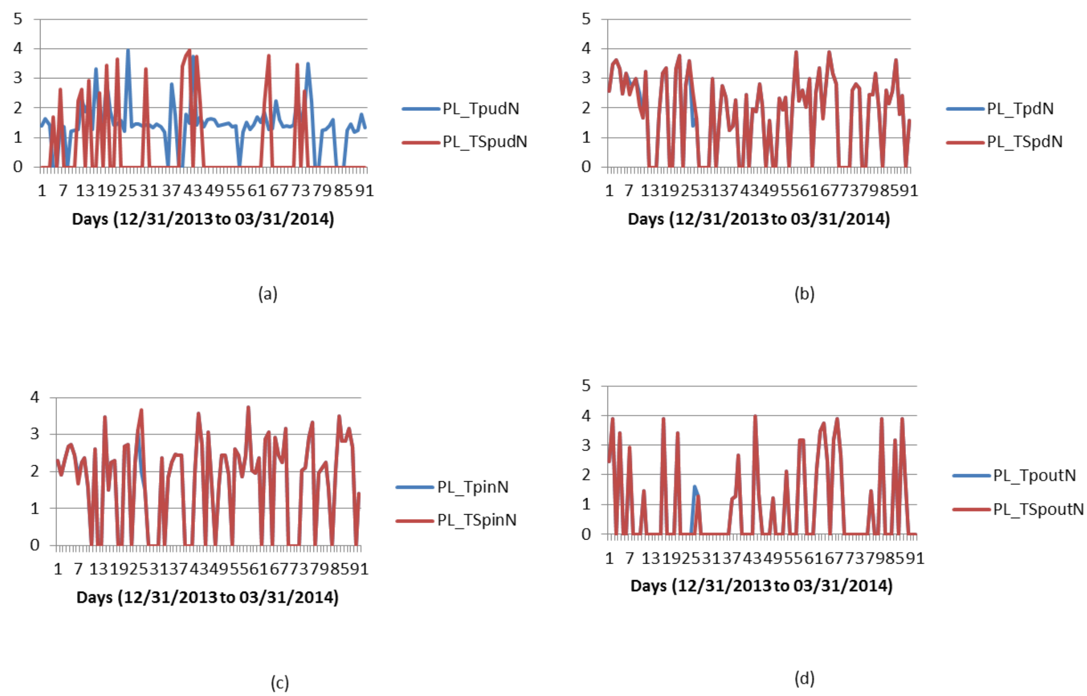


Figure 9 (a) shows that, in the undirected network, paths are more uniformly distributed among nodes than shortest paths are distributed among nodes. This means that fewer nodes are responsible for more of the shortest paths in the undirected network. There are fewer instances of shortest path following power law distribution in undirected (figure 9 (a)) and consumption (figure 9 (c)) networks. In the directed (figure 9 (b)) and propagation (figure 9 (d)) networks, there are no such patterns.

A.5.5 Network Flow Variables (MV2)

Figure 10: Network Flow Variables-- (a) Total Paths and Total Shortest Paths, (b) Average Paths and Average Shortest Paths, (c) Undirected and Directed Network Graph Diameter.

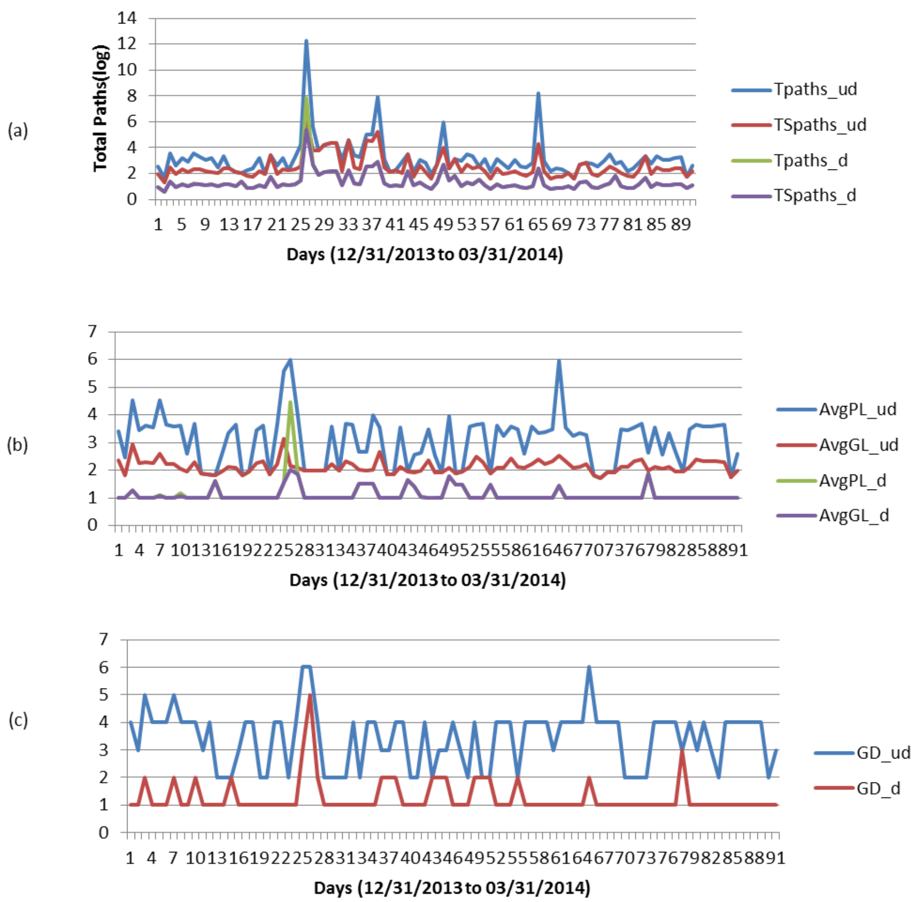


Figure 10 (a), shows that total number of paths in the undirected network (Tpaths_ud) is orders of magnitude higher than the total number of shortest paths (TSpaths_ud). The total number of paths (Tpaths_d) and the total number of shortest paths (TSpaths_d) map more closely in the directed network. In figure 10 (b), a similar trend is observed in the Average Path Lengths (AvgPL_ud, AvgPL_d) and the Average Geodesic Lengths (AvgGL_ud, AvgGL_d) of the undirected and directed networks. In figure 10 (c), the Graph Diameter (GD_ud) of the undirected network is larger than the graph diameter of the directed network (GD_d). It is also noteworthy that, in figure 10 (b) and in figure 10 (c), the Graph Diameter (GD_ud, GD_d) and the Average Path Length (AvgPL_ud, AvgPL_d) of the undirected and directed networks track pretty closely.

A.5.6 Dependent Variables

A.5.6.1 Eigenvector Centralization

Figure 11: Eigenvector Centralization in the Undirected, Directed, Consumption and Propagation Networks

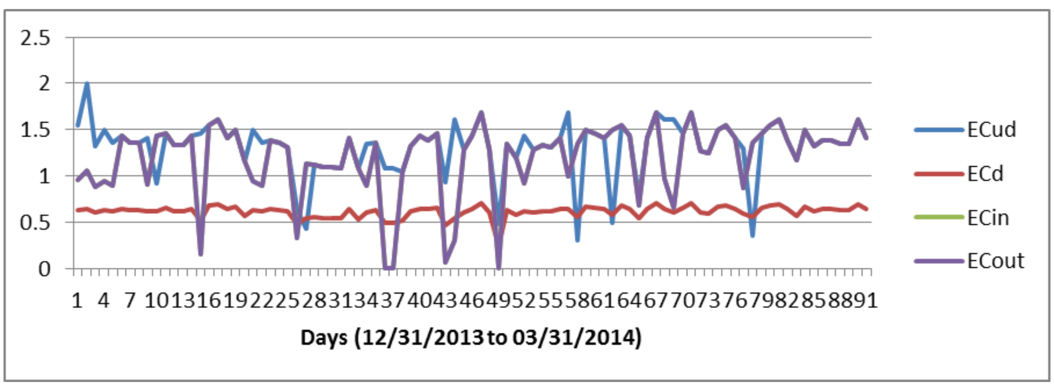


Figure 11 shows that nodes with influence are lot more central in the undirected (ECud) and propagation networks (ECout) than in the directed network (ECd). The consumption (ECin) and propagation (ECout) networks exhibit same level of centralization.

A.5.6.2 Power law Distribution of Eigenvector Centrality per Node

Figure 12: Power Law Distribution of Eigenvector Centrality in Undirected, Directed, Consumption and Propagation Network

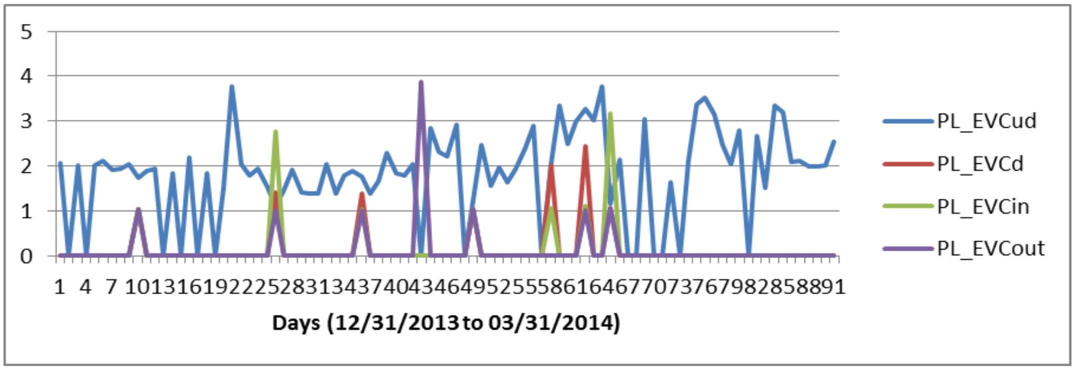
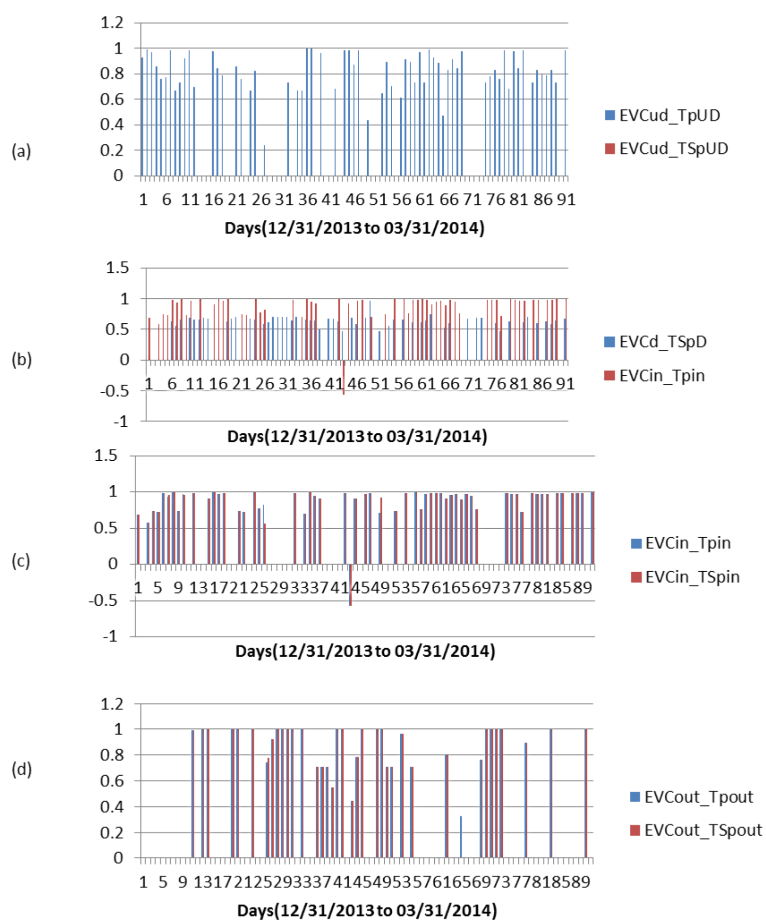


Figure 12 shows that in the undirected network eigenvector centrality values are consistently distributed in a power law distribution pattern (PL_EVCud), over a period of time. In the directed, the consumption and the propagation network the distribution of eigenvector centrality follows a power law distribution (PL_EVCd, PL_EVCin, PL_EVCout) pattern only sometimes.

A.5.6.3 Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node

Figure 13: Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



In figure 13, only those correlation coefficients with a significance value lower than 0.05 are shown. In figure 13 (a), there is a significant correlation between the eigenvector centrality of a node and the number of paths from a node in undirected network (EVCud_TpUDN). There is no significant correlation between eigenvector centrality of a node and shortest paths from a node in undirected network (EVCud_TSpUDN). In figure 13 (b), there is a significant correlation between the directed-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the directed network (EVCd_TpDN, EVCud_TSpUDN). In figure 13 (c), there is a significant correlation between the in-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the consumption network (EVin_TpinN, EVCin_TSpinN). The correlation between the out-eigenvector centrality of a node and the number of shortest paths is less significant figure 13 (d) (EVCout_TpoutN, EVCout_TSpoutN).

A.6 Case 6 --Science

A.6.1 Case Overview

Data for keyword "YouTube + science" was collected over a period of 91 days (31/12/2013 to 31/03/2014). As shown in table 9, overall 49,332 tweets were collected, out of which 13,462 were broadcast tweets and 35,870 were engaged tweets respectively. Out of 35,870 engaged tweets only 22,598 tweets formed the largest community. Similarly, 52,785 daily unique people tweeted overall, out of which 20,157 daily unique people were engaged in broadcast activity whereas 32,628 daily unique people were engaged in conversations. Out of 32,628 daily unique people only 21,277 daily unique people formed the largest community. Data for the largest community was analyzed at a daily interval. The overall trends for the data are shown below in figure 1 and figure 2.

Figure 1: Overall Tweets

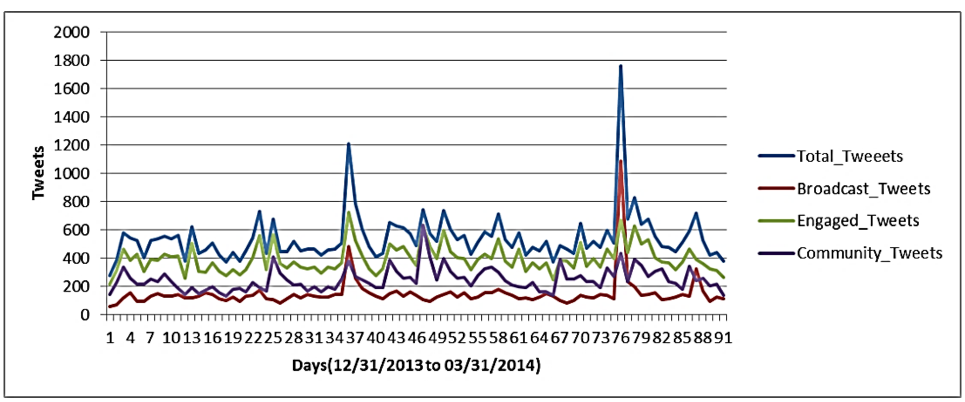


Fig.2: Overall People

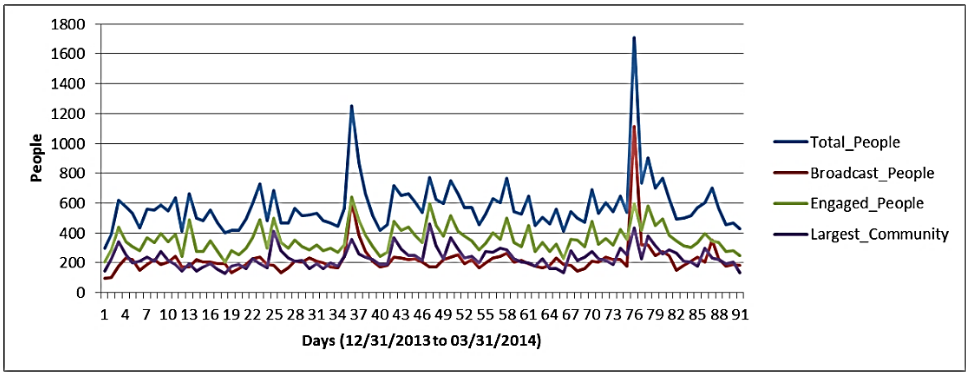


Figure 1 and figure 2 shows that both the total tweets and total people involved are very dynamic and their magnitude changes on a daily basis. The maximum of the total number of daily tweets and the maximum of the total number of daily unique people observed on a single day (the daily uniques) are 1,757 and 1,708, respectively. Similarly, the minimum of the total number daily tweets and the minimum of the number daily unique are 277 and 300, respectively. The size of the largest community on a particular day and the largest number of community tweets on that day also seem to follow the trend of total people and total tweets. The largest number of daily community tweets and the largest number of daily unique people are 634 and 461, respectively. Similarly, the smallest number of daily community tweets and the smallest number of daily unique people are 130 and 130, respectively. As the total number of daily unique people tweeting increases, so does the size of the community.

Most of the engaged people are engaged in the collective conversation forming the largest community.

A.6.2 Random or Not Random

As explained in section 4.4.1, in order to eliminate α- error and β- error, I compare the clustering coefficients of both undirected (CC_ud) and directed networks (CC_d) with their corresponding random (Erdös-Rényi, E-R) networks (CCudran, CCdran). If the clustering coefficients of the undirected and directed networks are equal to those of the E-R random network, then the directed and undirected networks are considered to be random, if they are not equal, then they are not random.

Figure 3: Comparison of Clustering Coefficients of Undirected Network with E-R Networks

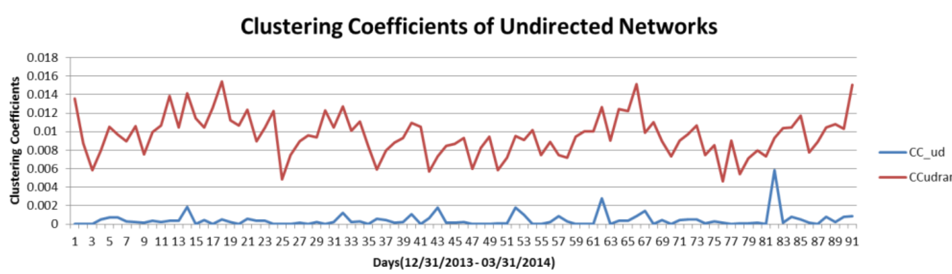
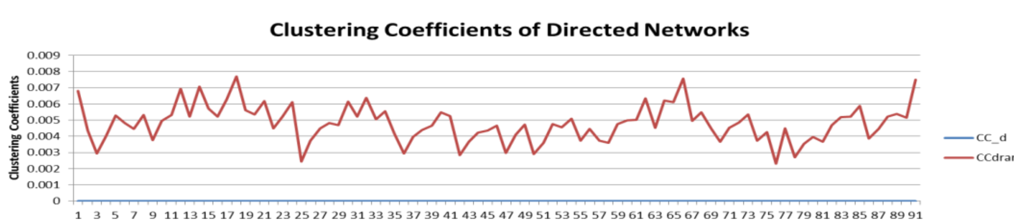


Figure 4: Comparison of Clustering Coefficients of Directed Network with E-R Networks



As seen in figure 3 and figure 4, Clustering Coefficients of the undirected networks (CC_ud) follows a very different pattern from their corresponding E-R networks. Therefore, the undirected network is considered to be non-random networks and the variables computed are a true reflection of network's features. For the direct network the Clustering Coefficients (CC_d) is zero for the most part. Therefore, the directed networks are random.

A.6.3. Independent Variables

The values of the independent variables for both the undirected and the directed network are shown in Figure 5 below.

Figure 5: Independent Variables--(a) Nodes and Edges (Undirected and Directed networks), (b) Reciprocity (Directed Networks), (c) Density (Undirected and Directed Networks), (d) Clustering Coefficient Undirected Network, (e) Clustering Coefficient Directed Network.

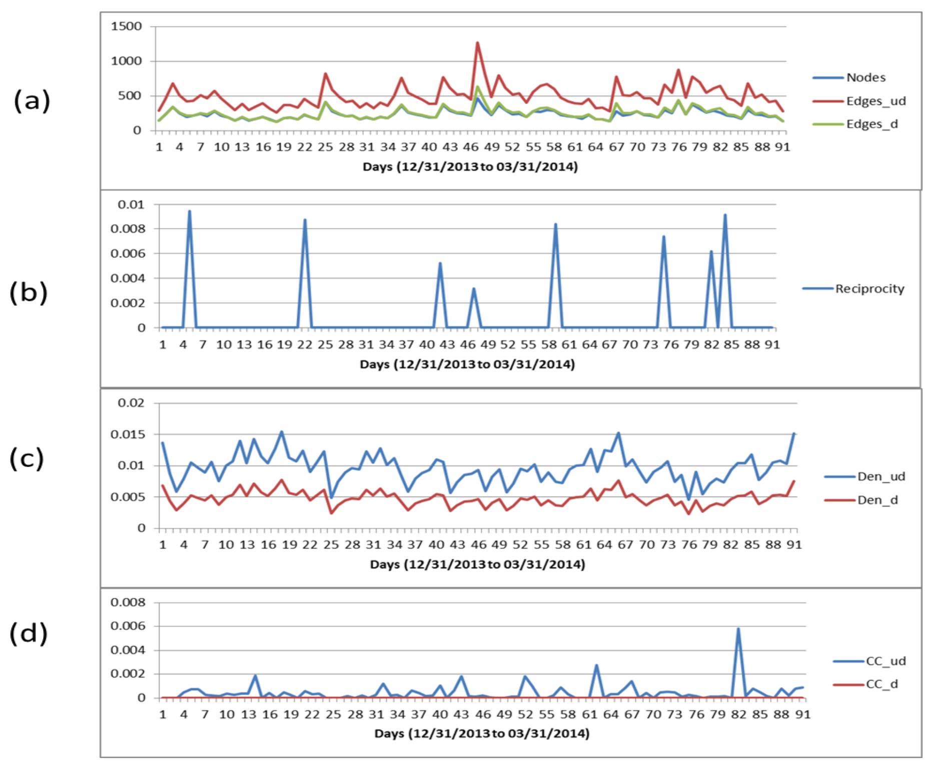


Figure 5 (a) shows that the number of directed ties (Edges_d) in the network and the total number of nodes (Nodes) overlap with each other. The numbers of undirected ties (Edges_ud) is greater than the number of directed ties (Edges_d), because in an undirected network every directed tie is considered to be symmetric. Therefore it is counted twice, except for the ones that are symmetric in a directed network. Reciprocity in Figure 5(b) indicates the presence of symmetric ties in a directed network (in an undirected network 100% are symmetric). The value of 0.01 is equal to 1% of all the ties. Figure 5(c) shows the difference between the densities of the undirected (Den_ud) and the directed networks (Den_d). The undirected network is denser than the directed network (Den_ud>Den_d).

A.6.4 Network Structure Variables (MV1)

A.6.4.1 The Scale Free Metric

Figure 6: Scale Free Metric--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.

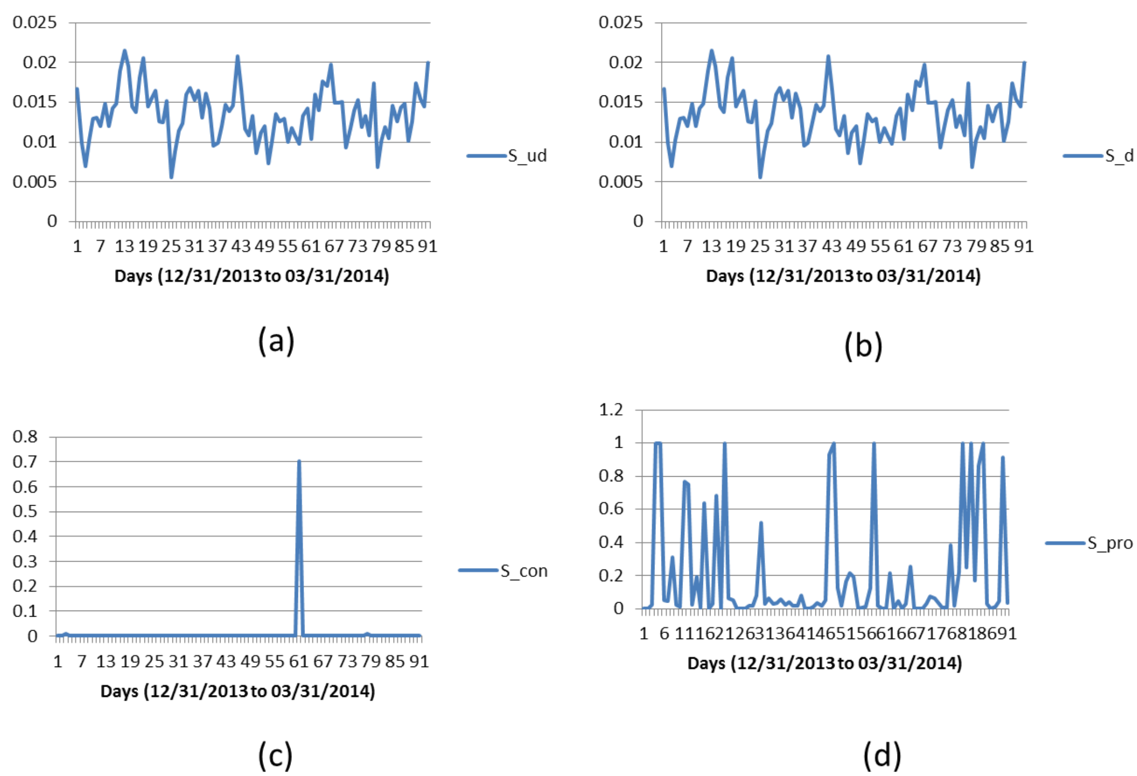
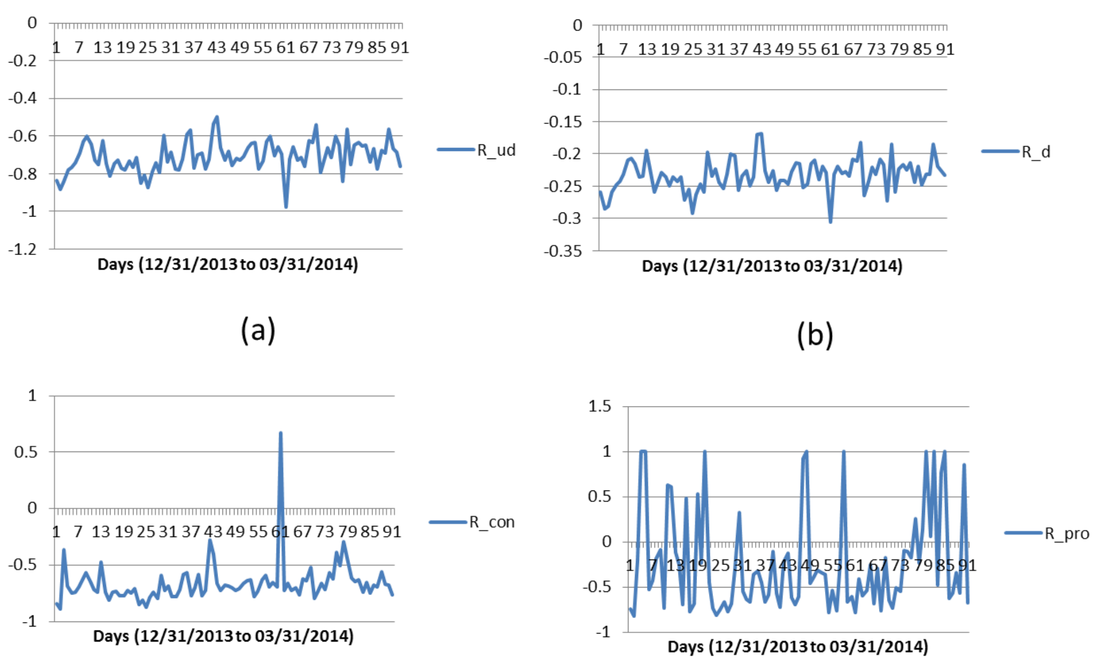


Figure 6 shows the Scale Free Metric for the undirected, directed, consumption and propagation networks (S_ud, S_d, S_con, S_pro). The Scale Free Metrics for the undirected (S_ud) and the directed network (S_d) are similar, but the Scale Free Metrics for the consumption (S_con) and propagation (S_pro) networks are very different. The propagation (S_pro) network is more scale free than the consumption network (S_con). The values of the scale free metric ranges between 0 and 1. When the values are closer to 1, it means that the networks are more scale free. Neither the directed nor the undirected network is scale free. This means that these networks may have hubs in them. However, there is not just one hub that is the center of the community. As shown in figure 6 (c) and figure 6 (d) the consumption network and the propagation network are scale free in some instances.

A.6.4.2 The Assortativity

Figure 7: Assortativity--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



(a) (b) (c) (d)

Figure 7 shows the assortativity metric for the undirected, directed, consumption and propagation networks (R_ud, R_d, R_con, R_Pro). The value of the assortativity metric ranges between -1 and 1. When the values are closer to -1, it means that networks are disassortative. The undirected network is more Disassortative than

the directed network (R_d > R_ud). Among the directed networks, the consumption network is more Disassortative than the propagation network (R_pro > R_con). Disassortative means that the nodes in the network connect to nodes that are very similar to themselves. This is true more so in the undirected network and in the consumption network than it is in the directed network and the propagation network. This implies that disassortativeness of consumption contributes more to the disassortativeness of the directed network than the disassortativeness of the propagation does.

A.6.4.3 The Small World Metric

Figure 8: Small World Metric -- (a) Undirected Network, (b) Directed Network.

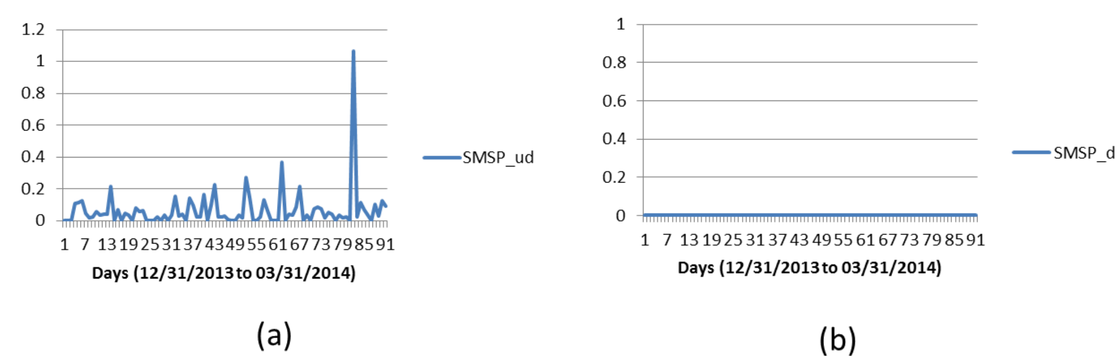
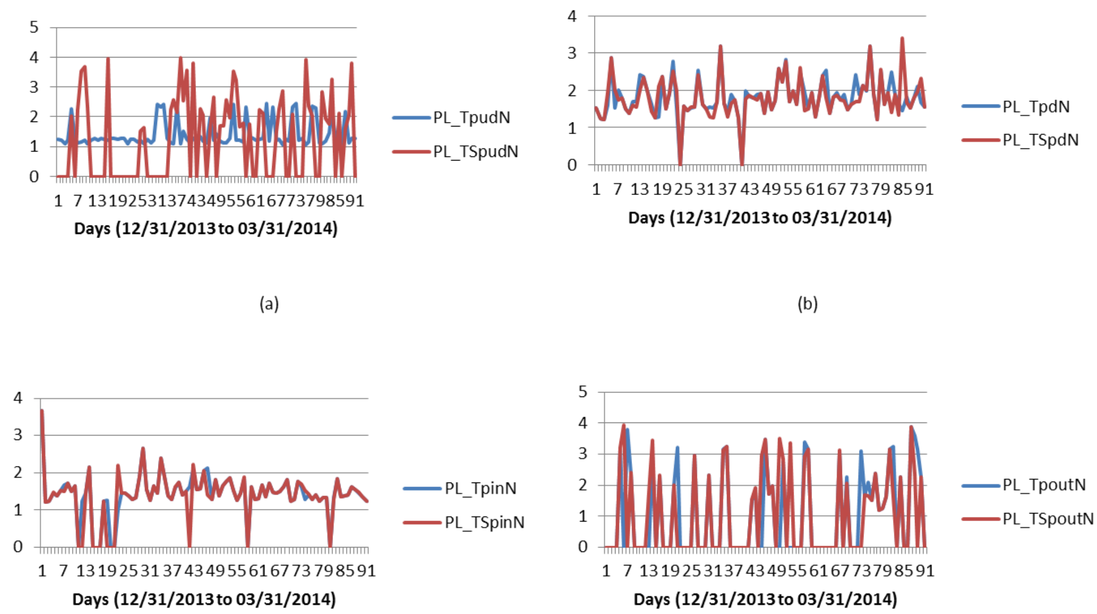


Figure 8 shows the Small World Metric for the undirected (SMSP_ud) and directed networks (SMSP_d). The Small World Metrics for the consumption and propagation networks are the same as the ones for the directed network. The directed networks don't show any small world behavior. Contrary to the directed networks, undirected networks show some small world behavior.

A.6.4.4 Paths and Shortest Paths Power law Distribution per Node

Figure 9: Power Law Distribution of Paths and Shortest Paths in (a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



(a) (b) (c) (d)

Figure 9 (a) shows that, in the undirected network, paths are more uniformly distributed among nodes than shortest paths are distributed among nodes. This means that fewer nodes are responsible for more of the shortest paths in the undirected network. There are fewer instances of shortest path following power law distribution in undirected (figure 9 (a)) and consumption (figure 9 (c)) networks. In the directed (figure 9 (b)) and propagation (figure 9 (d)) networks, there are no such patterns.

A.6.5 Network Flow Variables (MV2)

Figure 10: Network Flow Variables-- (a) Total Paths and Total Shortest Paths, (b) Average Paths and Average Shortest Paths, (c) Undirected and Directed Network Graph Diameter.

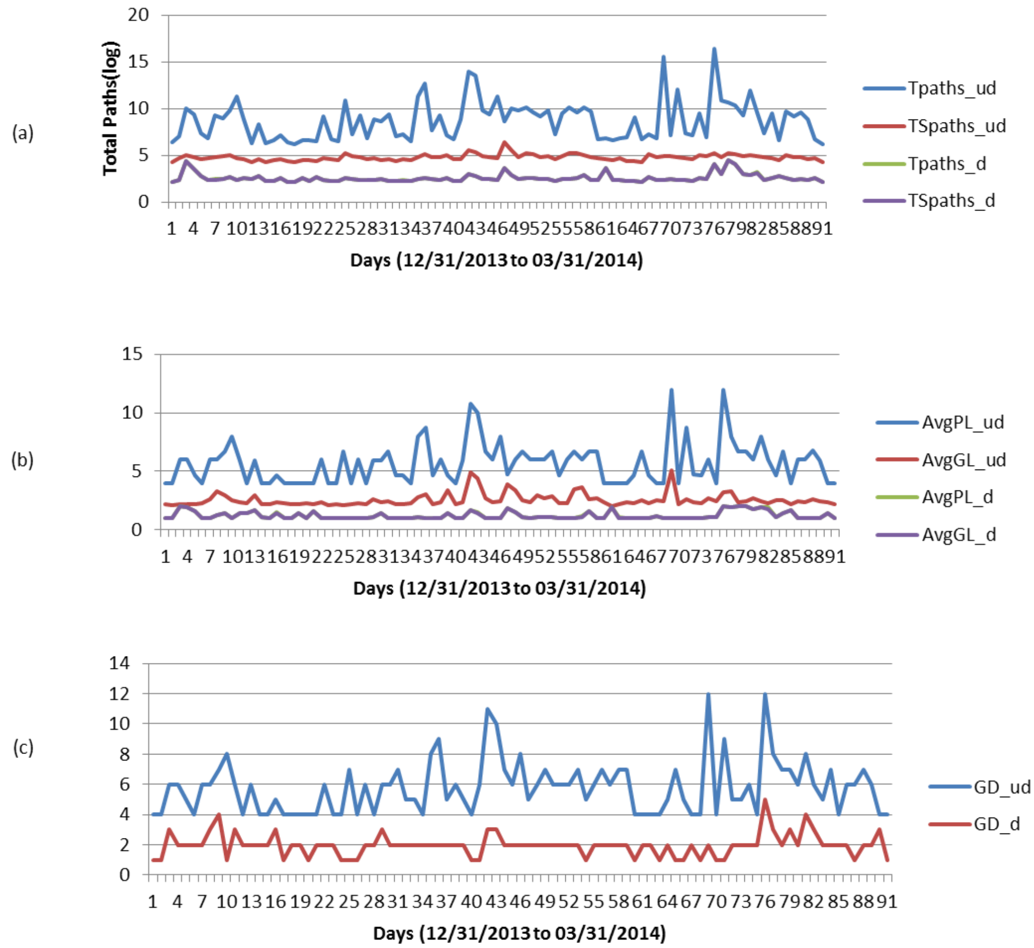


Figure 10 (a), shows that total number of paths in the undirected network (Tpaths_ud) is orders of magnitude higher than the total number of shortest paths (TSpaths_ud). The total number of paths (Tpaths_d) and the total number of shortest paths (TSpaths_d) map more closely in the directed network. In figure 10 (b), a similar trend is observed in the Average Path Lengths (AvgPL_ud, AvgPL_d) and the Average Geodesic Lengths (AvgGL_ud, AvgGL_d) of the undirected and directed networks. In figure 10 (c), the Graph Diameter (GD_ud) of the undirected network is larger than the graph diameter of the directed network (GD_d). It is also noteworthy that, in figure 10 (b) and in figure 10 (c), the Graph Diameter (GD_ud, GD_d) and the Average Path Length (AvgPL_ud, AvgPL_d) of the undirected and directed networks track pretty closely.

A.6.6 Dependent Variables

A.6.6.1 Eigenvector Centralization

Figure 11: Eigenvector Centralization in the Undirected, Directed, Consumption and Propagation Networks

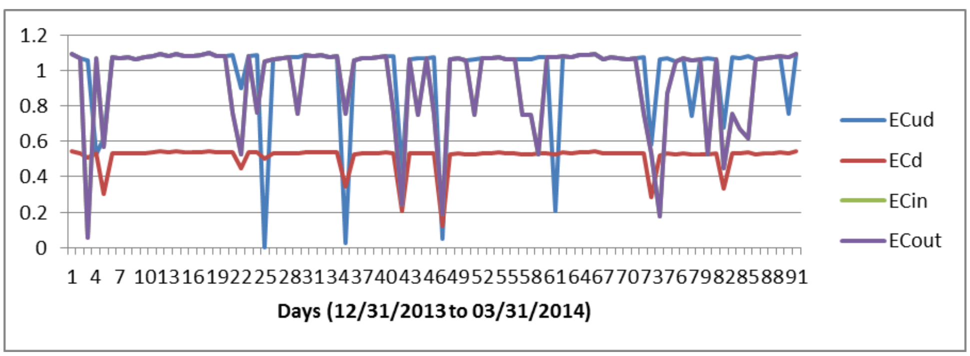


Figure 11 shows that nodes with influence are lot more central in the undirected (ECud) and propagation networks (ECout) than in the directed network (ECd). The consumption (ECin) and propagation (ECout) networks exhibit same level of centralization.

A.6.6.2 Power law Distribution of Eigenvector Centrality per Node

Figure 12: Power Law Distribution of Eigenvector Centrality in Undirected, Directed, Consumption and Propagation Network

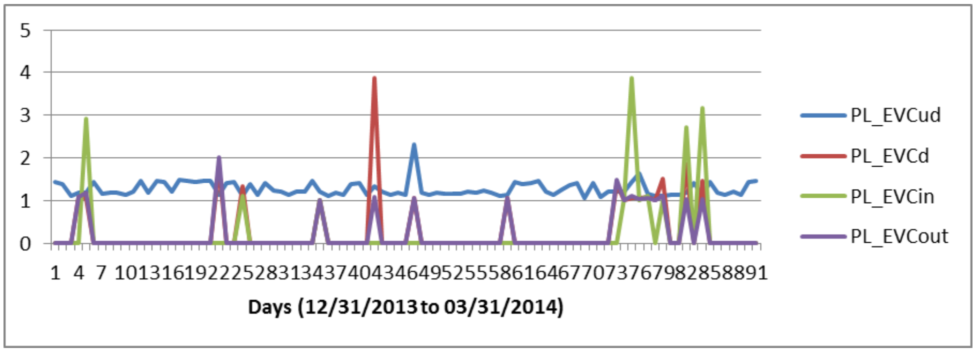
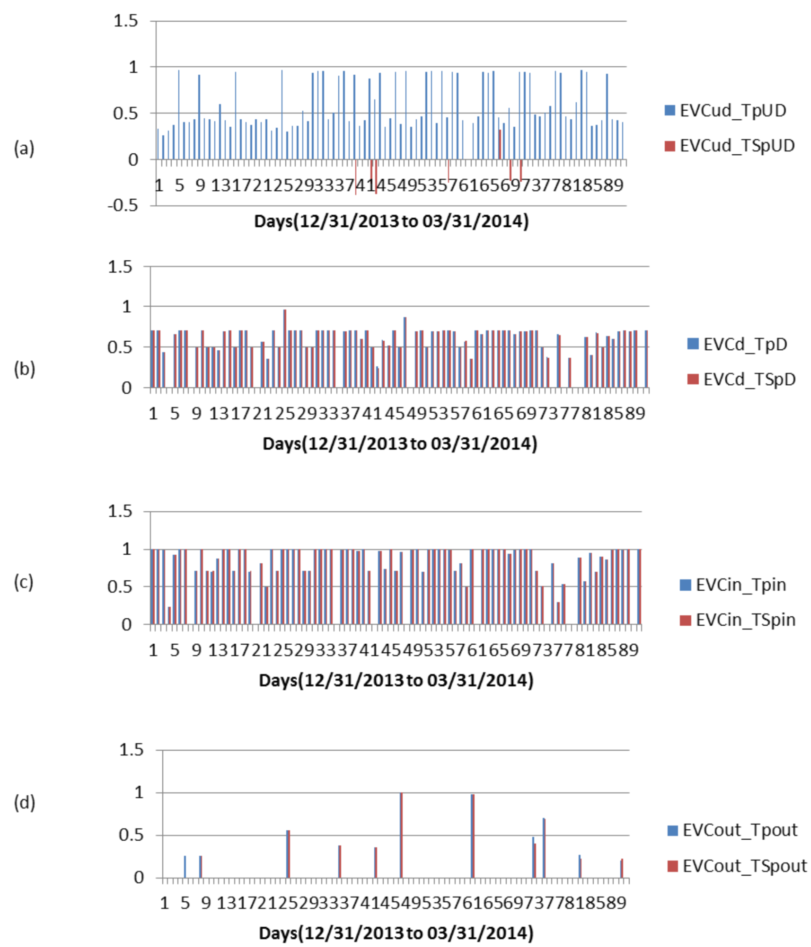


Figure 12 shows that in the undirected network eigenvector centrality values are consistently distributed in a power law distribution pattern (PL_EVCud), over a period of time. In the directed, the consumption and the propagation network the distribution of eigenvector centrality follows a power law distribution (PL_EVCd, PL_EVCin, PL_EVCout) pattern only sometimes.

A.6.6.3 Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node

Figure 13: Correlation Coefficient of Eigenvector Centrality vs. Total Paths per Node and Eigenvector Centrality vs. Total Shortest Paths per Node--(a) Undirected Network, (b) Directed Network, (c) Consumption Network, (d) Propagation Network.



In figure 13, only those correlation coefficients with a significance value lower than 0.05 are shown. In figure 13 (a), there is a significant correlation between the eigenvector centrality of a node and the number of paths from a node in undirected network (EVCud_TpUDN). There is no significant correlation between eigenvector centrality of a node and shortest paths from a node in undirected network (EVCud_TSpUDN). In figure 13 (b), there is a significant correlation between the directed-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the directed network (EVCd_TpDN, EVCud_TSpUDN). In figure 13 (c), there is a significant correlation between the in-eigenvector centrality of a node and the number of paths and shortest paths ending on a node in the consumption network (EVin_TpinN, EVCin_TSpinN). The correlation between the out-eigenvector centrality of a node and the number of shortest paths is less significant figure 13 (d) (EVCout_TpoutN, EVCout_TSpoutN).

A.6.7 Statistical Analysis

A.6.7.1 The Undirected Network

A.6.7.1.1 Correlation Analysis

In Table 1, the statistically significant correlation coefficients for the undirected network are marked in yellow. All correlations between all variables are shown in supplemental file titled "Correlations.pdf".

Table 1: Correlation Coefficients of Undirected Network

Correlations

		Nodes	Edges_ud	Den_ud	CC_ud	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud	PL_TpudN
Edges_ud	Pearson	.965**	1								
	Sig. (2-ta	.000									
	N	91	91								
Den_ud	Pearson	-.920**	-.835**	1							
	Sig. (2-ta	.000	.000								
	N	91	91	91							
GD_ud	Pearson	.505**	.425**	-.529**	-.007	1					
	Sig. (2-ta	.000	.000	.000	.951						
	N	91	91	91	91	91					
Tpaths_ud	Pearson	.642**	.565**	-.658**	-.012	.974**	1				
	Sig. (2-ta	.000	.000	.000	.909	.000					
	N	91	91	91	91	91	91				
TSpaths_ud	Pearson	.917**	.959**	-.828**	-.053	.463**	.583**	1			
	Sig. (2-ta	.000	.000	.000	.619	.000	.000				
	N	91	91	91	91	91	91	91			
AvgPL_ud	Pearson	.505**	.424**	-.531**	-.003	.997**	.980**	.461**	1		
	Sig. (2-ta	.000	.000	.000	.974	.000	.000	.000			
	N	91	91	91	91	91	91	91	91		
AvgGL_ud	Pearson	.434**	.465**	-.385**	.022	.679**	.660**	.622**	.682**	1	
	Sig. (2-ta	.000	.000	.000	.833	.000	.000	.000	.000		
	N	91	91	91	91	91	91	91	91	91	
S_ud	Pearson	-.695**	-.633**	.772**	.252*	-.125	-.284**	-.573**	-.121	.088	.037
	Sig. (2-ta	.000	.000	.000	.016	.237	.006	.000	.252	.407	.724
	N	91	91	91	91	91	91	91	91	91	91
R_ud	Pearson	.173	.222*	-.216*	.189	.508**	.475**	.318**	.512**	.631**	-.087
	Sig. (2-ta	.100	.034	.040	.072	.000	.000	.002	.000	.000	.410
	N	91	91	91	91	91	91	91	91	91	91
SMSP_ud	Pearson	-.011	.059	.084	.982**	.013	.023	-.002	.016	-.002	.002
	Sig. (2-ta	.921	.579	.429	.000	.906	.832	.989	.883	.986	.982
	N	91	91	91	91	91	91	91	91	91	91
EVCud_TpudN	Pearson	.164	.170	-.126	.055	.341**	.215*	.198	.275**	.232*	.666**
	Sig. (2-ta	.120	.107	.235	.606	.001	.040	.059	.008	.027	.000
	N	91	91	91	91	91	91	91	91	91	91
EVCud_TSpudN	Pearson	-.101	-.033	.170	-.053	-.408**	-.356**	-.220*	-.405**	-.597**	.056
	Sig. (2-ta	.339	.758	.106	.620	.000	.001	.036	.000	.000	.597
	N	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

In Table 1, the number of nodes (Nodes) and the number of ties (Edges_ud) have a strong positive correlation. As the number of nodes (Nodes) increases, the number of ties (Edges_ud) also increases. The Density (Den_ud) of this network has a strong negative correlation with both the number of nodes (Nodes) and the number of ties (Edges_ud). Graph Diameter (GD_ud) correlates positively with number of nodes (Nodes) and negatively with Density (Den_ud). Total Paths (Tpaths_ud) and Total Shortest Paths (TSpaths_ud) share a positive correlation with the number of nodes (Nodes), number of ties (Edges_ud) and a negative correlation with Density (Den_ud). Average Path Length (AvgPL_ud) shares a strong positive correlation with number of nodes (Nodes), Graph Diameter (GD_ud) and Total Paths (Tpaths_ud). Average Path Length (AvgPL_ud) shares a negative relationship with Density (Den_ud). Average Geodesic Length (AvgGL_ud) shares a strong correlation with Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud) and Average Path Length (AvgPL_ud). Scale Free Metric (S_ud) shares a positive correlation with Density (Den_ud), and a negative relationship with number of nodes (Nodes), number of ties (Edges_ud) and Total Shortest Paths (TSpaths_ud). Assortativity (R_ud) shares positive correlations with Graph Diameter (GD_ud), Average Path Length (AvgPL_ud) and Average Geodesic Length (AvgGL_ud). Small World Metric (SMSP_ud) shares a positive correlation with Clustering Coefficient (CC_ud). Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and correlate strongly Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN) and Assortativity (R_ud). Eigenvector

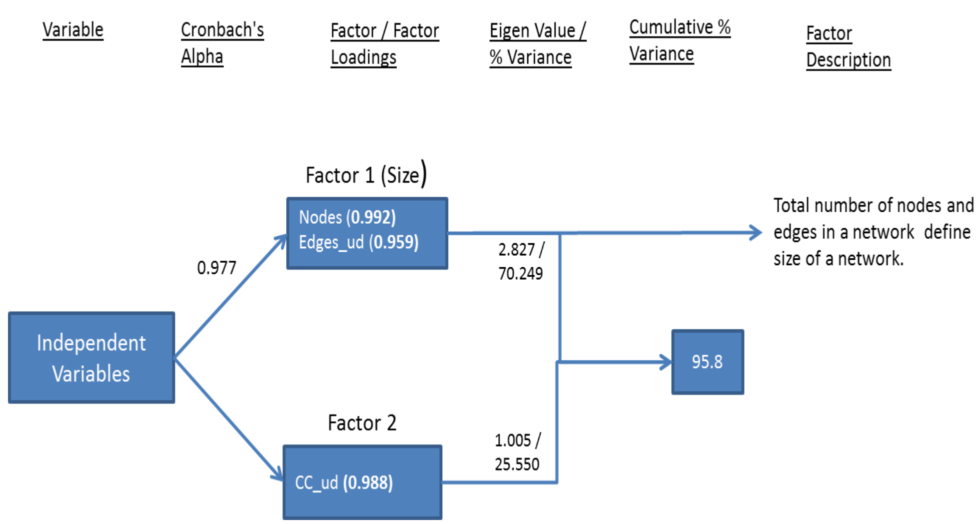
Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN) correlates negatively with Average Geodesic Length (AvgGL_ud).

A.6.7.1.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.6.7.1.2.1 Independent Variables

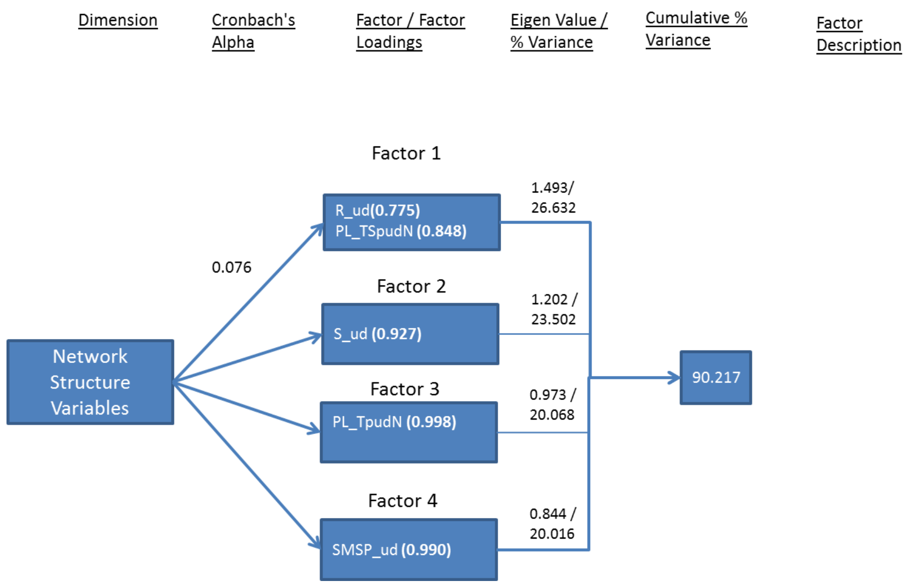
Figure 14: Factor Analysis Independent Variables Music Undirected Network



The factor analysis generated two factors that explain 95.8% (greater than 80%) of the cumulative variance. Factor1 and factor 2 have eigenvalues above 1. Nodes and ties (Edges_ud) have significant factor loadings in factor 1. Clustering Coefficient (CC_ud) has significant loading in factor 2. Cronbach’s alpha for factor 1has a value of 0.977. This means Nodes and ties (Edges_ud) are measuring same construct within factor 1. Hence, I name factor 1 as “Size”.

A.6.7.1.2.2 Network Structure (MV1)

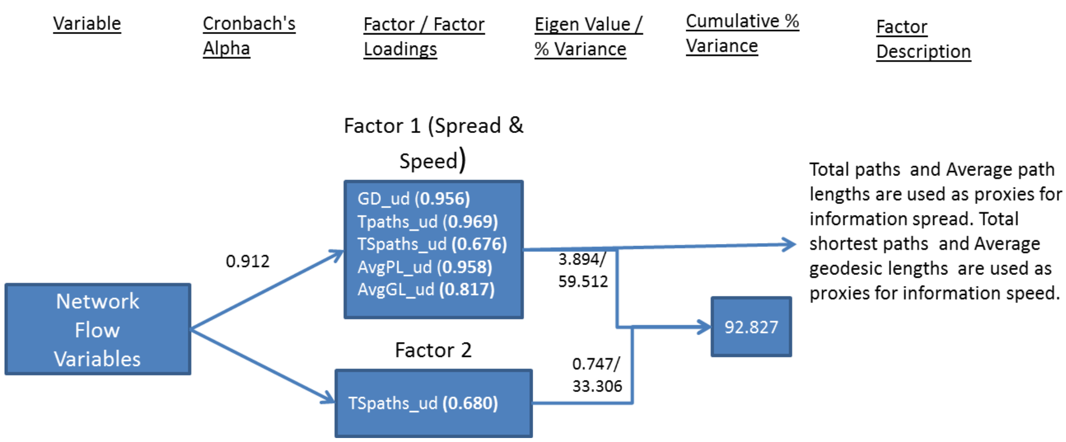
Figure 15: Factor Analysis of Network Structure Variables



The factor analysis generated three factors that explain 90.217% (greater than 80%) of the cumulative variance. Factor1 and factor2 have eigenvalues above 1. Factor3 and factor4 have eigenvalues below 1. Power Law Distribution of Total Paths per Node (PL_TpudN) and Assortativity (R_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.0.76. . Power Law Distribution of Total Paths per Node (PL_TpudN) and Assortativity (R_ud) are measuring different constructs within factor 1. Hence, they should not be considered as a factor. All other variables load independently.

A.6.7.1.2.3 Network Flow (MV2)

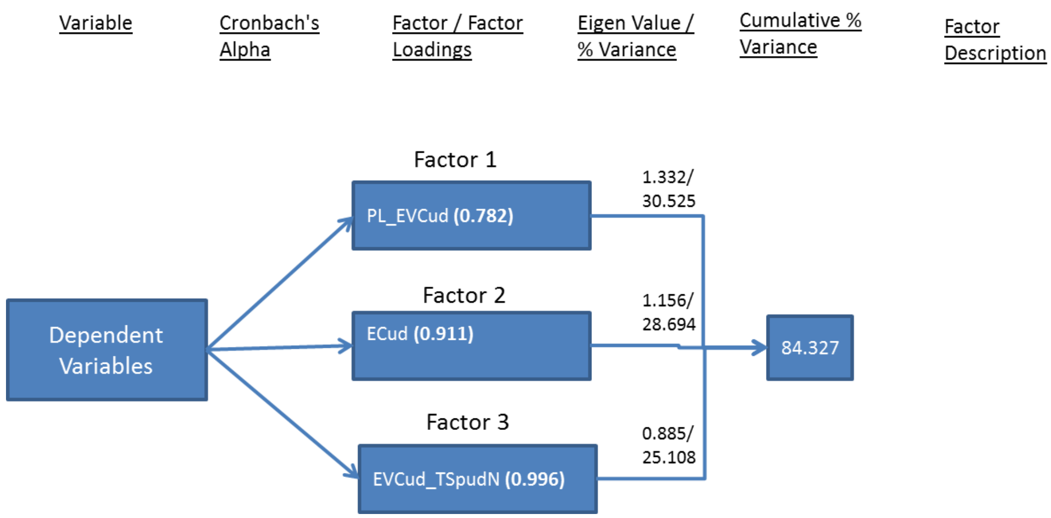
Figure 16: Factor Analysis of Network Flow Variables



The factor analysis generated two factors that explain 92.827% (greater than 80%) of the cumulative variance. Factor1 has eigenvalues above 1. Factor2 has eigenvalue below 1. Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Average Geodesic Length (AvgGL_ud) and Average Path Length (AvgPL_ud) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.912. Hence, they should be considered as a factor.

A.6.7.1.2.4 Dependent Variables

Figure 17: Factor Analysis of Dependent Variables



Factor analysis generated three factors that explain 84.327% (greater than 80%) of cumulative variance. All variables load independently. No significant factors were formed.

A.6.7.1.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Sports.pdf".

A.6.7.1.3.1 Impact of Network Structure on Network Flow

Table 2: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_ud	Tpaths_ud	TSpaths_ud	AvgPL_ud	AvgGL_ud
Science	(0.307/0.000) [3,4]	(0.379/0.000) [3,4]	(0.537/0.000) [3,4]	(0.309/0.000)[3,4]	(0.412/0.000) [2,4]

Table 2 shows that the network structure variables have a significant impact on the network flow variables. Network structure variables explain 30.7%, 37.9%, 53.9%, 30.9% and 41.2% variation in Graph Diameter (GD_ud), Total Paths (Tpaths_ud), Total Shortest Paths (TSpaths_ud), Average Path Length (AvgPL_ud) and Average Geodesic Length (AvgGL_ud), respectively.

A.6.7.1.3.2 Impact of Network Flow on Network Structure

Table 3: Impact of Network Flow on Network Structure

Predictors: **(6) GD_ud, (7) Tpaths_ud, (8) TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpudN	PL_TSpudN	S_ud	R_ud	SMSP_ud
Science	NA	(0.136/0.000)[8]	(0.693/0.000) [7,8,10]	(0.392/0.000)[9]	NA

Table 3 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 13.6%, 69.3%, and 39.2% variation in the PL_TpudN, S_ud and R_ud, respectively.

A.6.7.1.3.3 Impact of Network Structure on Network Phenomenon

Table 4: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Science	(0.042/0.028)[3]	(0.160/0.000)[3,4]	(0.531/0.000)[1,4]	(0.060/0.000)[4]

Table 4 shows that the network flow variable impacts Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 4.3%, 16%, 53.1% and 6% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud) is not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.1.3.4 Impact of Network Flow on Network Phenomenon

Table 5: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_ud, (7)Tpaths_ud, (8)TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud**

	Dependent Variable (Adjusted R Square/ Significance)			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Science	(0.033/0.048)[8]	(0.740/0.000) [6,8,10]	(0.106/0.001)[6]	(0.380/0.000)[8.10]

Table 5 shows that the network flow variable impacts Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 3.3%, 74%, 10.6 and 38% variation respectively. The impact of network flow variables on Eigenvector Centralization (EC_ud) and Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) is not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.1.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 6: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpudN, (2) PL_TSpudN, (3) S_ud, (4) R_ud, (5)SMSP_ud, (6),GD_ud (7) Tpaths_ud (8), TSpaths_ud, (9) AvgPL_ud, (10) AvgGL_ud, (11) Nodes, (12) Edges_ud, (13) Den_ud, (14) CC_ud**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECud	PL_EVCudN	EVCud_TpudN	EVCud_TSpudN
Science	(0.192/0.000) [12,4]	(0.709/0.000) [7,10,1]	(0.595/0.000) [1,6]	(0.458/0.000) [10,4,12]

Table 6 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_ud), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCudN), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpudN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 19.2%, 70.9%, 59.5% and 45.8% variation respectively.

A.6.7.2 The Directed Network

A.6.7.2.1 Correlation Analysis

Significant correlations coefficients for directed network are shown below in table 7. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 7: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	PL_TpdN	ECd	EVCd_TpdN
Edges_d	Pearson C	.965**	1									
	Sig. (2-taile	.000										
	N	91	91									
Den_d	Pearson C	-.920**	-.834**	-.070	1							
	Sig. (2-taile	.000	.000	.510								
	N	91	91	91	91							
TSpaths_d	Pearson C	.647**	.609**	.075	-.576**	.495**	1.000**	1				
	Sig. (2-taile	.000	.000	.482	.000	.000	.000					
	N	91	91	91	91	91	91	91				
AvgPL_d	Pearson C	.363**	.352**	.153	-.304**	.604**	.854**	.852**	1			
	Sig. (2-taile	.000	.001	.147	.003	.000	.000	.000				
	N	91	91	91	91	91	91	91	91			
AvgGL_d	Pearson C	.363**	.348**	.137	-.308**	.594**	.856**	.855**	.998**			
	Sig. (2-taile	.000	.001	.196	.003	.000	.000	.000	.000			
	N	91	91	91	91	91	91	91	91			
PL_TSpdN	Pearson C	.104	.093	.079	-.097	.347**	.109	.109	.169	.821**		
	Sig. (2-taile	.326	.383	.455	.358	.001	.303	.304	.109	.000		
	N	91	91	91	91	91	91	91	91	91		
S_d	Pearson C	-.695**	-.632**	.028	.772**	-.066	-.446**	-.449**	-.124	.055		
	Sig. (2-taile	.000	.000	.790	.000	.536	.000	.000	.243	.604		
	N	91	91	91	91	91	91	91	91	91		
R_d	Pearson C	-.029	.031	.119	.011	.222*	-.185	-.190	-.045	.320**		
	Sig. (2-taile	.786	.773	.262	.914	.035	.079	.071	.669	.002		
	N	91	91	91	91	91	91	91	91	91		
PL_EVCdN	Pearson C	.399**	.371**	.633**	-.343**	.262*	.394**	.389**	.383**	.262*	-.677**	
	Sig. (2-taile	.000	.000	.000	.001	.012	.000	.000	.000	.012	.000	
	N	91	91	91	91	91	91	91	91	91	91	
EVCd_TSpdN	Pearson C	-.205	-.164	-.023	.233*	-.434**	-.445**	-.447**	-.447**	-.219*	.079	1.000**
	Sig. (2-taile	.052	.121	.830	.026	.000	.000	.000	.000	.037	.459	.000
	N	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

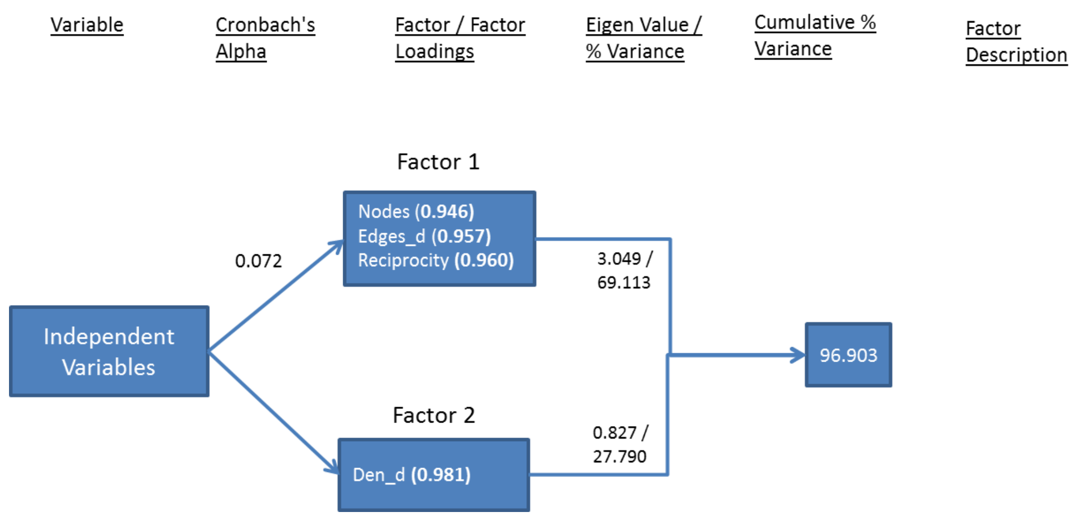
Table 7 shows that nodes (Nodes) and ties (Edges_d) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of ties (Edges_d), Reciprocity and Total Paths (Tpaths_d) in the network. Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpdN) correlates positively with Paths Power Law Distribution per Node (PL_TpdN). Scale Free Metric (S_d) seems to share a positive relationship with Density (Den_d) and Shortest Paths Power Law Distribution per Node (PL_TSpdN). Scale Free Metric (S_d) seems to share a negative relationship Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) and Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) correlate strongly with each other.

A.6.7.2.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled “Factor Analysis.pdf”.

A.6.7.2.2.1 Independent Variables

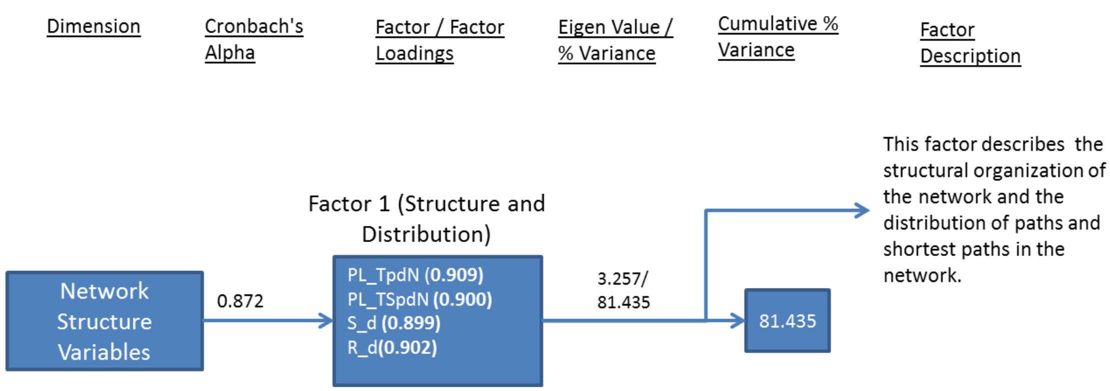
Figure 18: Factor Analysis of Independent Variables



The factor analysis generated two factors that explain 96.903% (greater than 80%) of the cumulative variance. Factor1 has eigenvalues above one. Factor 2 has eigenvalue below 1. Reciprocity, Nodes and ties (Edges_d) have significant factor loadings in factor 1. Cronbach’s alpha for factor1 has a value of 0.072. This means Reciprocity, Nodes and ties (Edges_d) are not measuring same construct within factor 1.

A.6.7.2.2.2 Network Structure (MV1)

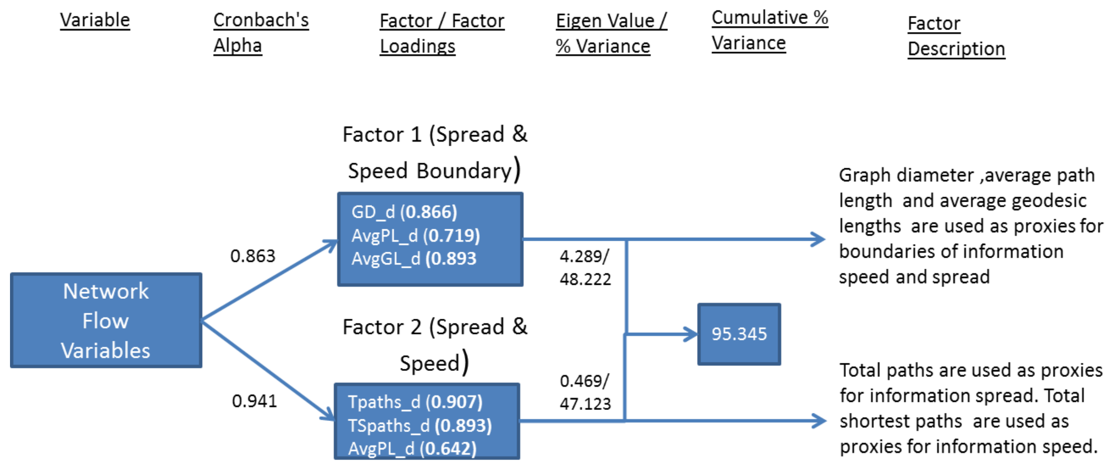
Figure 19: Factor Analysis of Network Structure Variables



Factor analysis generated one factor that explain 81.435% (greater than 80%) of cumulative variance. Factor1 has eigenvalues above 1. Power Law Distribution of Total Paths per Node (PL_TpdN), Power Law Distribution of Shortest Total Paths per Node (PL_TSpdN), Assortativity (R_d) and Scale Free Metric (S_d) have significant factor loadings in actor 1. Cronbach's alpha for factor1 has a value of 0.872. Power Law Distribution of Total Paths per Node (PL_TpdN), Power Law Distribution of Shortest Total Paths per Node (PL_TSpdN), Assortativity (R_d) and Scale Free Metric (S_d) are measuring same construct within factor 1. Hence, they should be considered as a factor. All other variables load independently.

A.6.7.2.2.3 Network Flow (MV2)

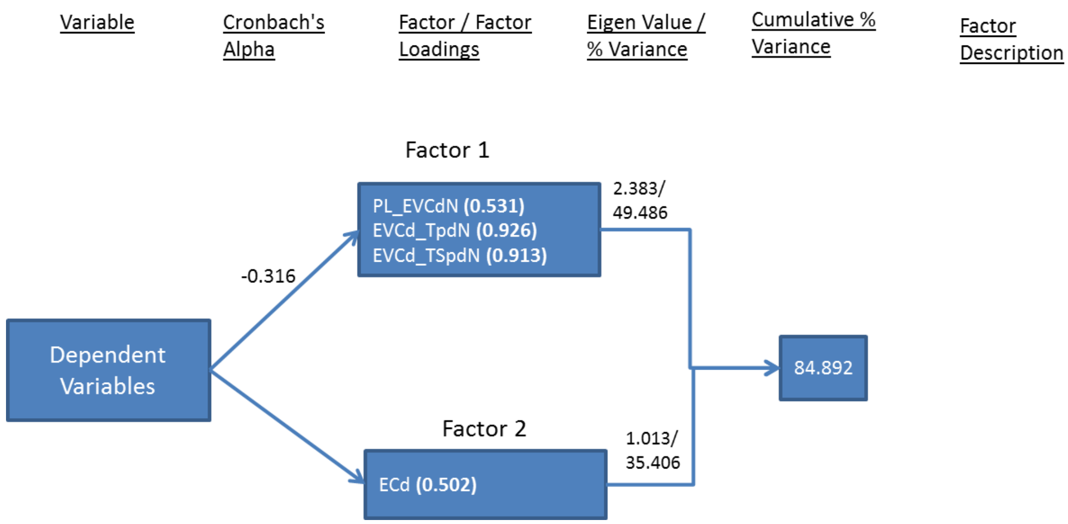
Figure 20: Factor Analysis of Network Flow Variables



Factor analysis generated two factors that explains 95.345% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d) and Average Path Length (AvgPL_d) have significant factor loadings in factor 1. Cronbach’s alpha for factor1 has a value of 0.863. Factor 1 is named as “Spread and Speed Boundary”. Cronbach’s alpha for factor2 has a value of 0.941. Factor 1 is named as “Spread and Speed”.

A.6.7.2.2.4 Dependent Variables

Figure 21: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 84.892% (greater than 80%) of cumulative variance. Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centralities with respect to Paths (EVCd_TpD) and Shortest Paths (EVCd_TSpD) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of -3.16. Eigenvector centralization loads independently.

A.6.7.2.3 Regression Analysis

In this section, only the regressions in which the predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Science.pdf".

A.6.7.2.3.1 Impact of Network Structure on Network Flow

Table 8: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Science	(0.139/0.000)[1]	(0.190/0.000)[3]	(0.193/0.000)[3]	NA	NA

Table 8 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 13.9%, 19% and 19.3% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), and Total Shortest Paths (TSpaths_d).

A.6.7.2.3.2 Impact of Network Flow on Network Structure

Table 9: Impact of Network Flow on Network Structure

Predictors: **(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)				
	PL_TpdN	PL_TSpdN	S_d	R_d	SMSP_d
Science	(0.139/0.000) [6]	(0.111/0.001)[6]	(0.193/0.000) [8]	(0.149/0.000) [6,8]	NA

Table 9 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 13.9%, 11.1%, 19.3%, and 14.9% variation in the PL_TpdN, PL_TSpdN, S_d, and R_ud, respectively. The impact of network flow variables on PL_TpdN is not taken into consideration, as the p-values is greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.2.3.3 Impact of Network Structure on Network Phenomenon

Table 10: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Science	NA	(0.058/0.012)[1]	(0.056/0.014)[2]	(0.056/0.014)[2]

Table 10 shows that the network structure variable impacts Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN), explaining 5.8%, 5.6% and 5.6% variation respectively. The impact of network flow variables Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCd_TSpdN) are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.2.3.4 Impact of Network Flow on Network Phenomenon

Table 11: Impact of Network Flow on Network Phenomenon

Predictors:**(6) GD_d, (7)Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCd_TpdN	EVCd_TSpdN
Science	(0.071/0.006)[9]	(0.146/0.000)[7]	(0.231/0.000) [6,10]	(0.233/0.000) [6,10]

Table 11 shows that the network structure variable impacts Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCud_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpdN), explaining 7.1%, 14.6%, 23.1% and 23.3% variation respectively. The impact of network flow variables Eigenvector Centralization (EC_d) is not taken into consideration, as the p-values is greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.2.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 12: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpdN, (2) PL_TSpdN, (3) S_d, (4) R_d, (5)SMSP_d, (6)GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECd	PL_EVCdN	EVCud_TpdN	EVCud_TSpdN
Science	(0.411/0.000) [3,12,15]	(0.609/0.000) [3,11,15]	(0.231/0.000) [6,10]	(0.233/0.000) [6,10]

Table 12 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_d), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCdN), Eigenvector Centrality with respect to Total Paths per Node (EVCd_TpdN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCud_TSpudN), explaining 41.1, 60.9%, 23.1% and 23.3% variation respectively.

A.6.7.3 The Consumption Network

A.6.7.3.1 Correlation Analysis

Significant correlations coefficients for consumption network are shown below in table 13. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 13: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpinN	S_con	EVCin_TpinN
Edges_d	Pearson C	.965**	1										
	Sig. (2-taile	.000											
	N	91	91										
Den_d	Pearson C	-.920**	-.834**	-.070	1								
	Sig. (2-taile	.000	.000	.510									
	N	91	91	91	91								
Tpaths_d	Pearson C	.646**	.610**	.081	-.575**	.499**	1						
	Sig. (2-taile	.000	.000	.448	.000	.000							
	N	91	91	91	91	91	91						
TSpaths_d	Pearson C	.647**	.609**	.075	-.576**	.495**	1.000**	1					
	Sig. (2-taile	.000	.000	.482	.000	.000	.000						
	N	91	91	91	91	91	91	91					
AvgPL_d	Pearson C	.363**	.352**	.153	-.304**	.604**	.854**	.852**	1				
	Sig. (2-taile	.000	.001	.147	.003	.000	.000	.000					
	N	91	91	91	91	91	91	91	91				
AvgGL_d	Pearson C	.363**	.348**	.137	-.308**	.594**	.856**	.855**	.998**	1			
	Sig. (2-taile	.000	.001	.196	.003	.000	.000	.000	.000				
	N	91	91	91	91	91	91	91	91	91			
PL_TSpinN	Pearson C	-.037	-.011	.013	.027	-.120	-.152	-.153	-.305**	-.310**	.882**		
	Sig. (2-taile	.731	.920	.905	.800	.258	.151	.147	.003	.003	.000		
	N	91	91	91	91	91	91	91	91	91	91		
R_con	Pearson C	.275**	.240*	.036	-.294**	.344**	.582**	.582**	.511**	.513**	.043	.783**	
	Sig. (2-taile	.008	.022	.734	.005	.001	.000	.000	.000	.000	.682	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	
PL_EVCinN	Pearson C	.125	.129	.669**	-.115	.157	.174	.166	.235*	.214*	-.014	-.030	
	Sig. (2-taile	.236	.224	.000	.279	.136	.100	.115	.025	.041	.897	.775	
	N	91	91	91	91	91	91	91	91	91	91	91	
EVCin_TSpinN	Pearson C	-.228*	-.200	-.076	.241*	-.366**	-.446**	-.449**	-.452**	-.462**	.133	-.261*	1.000**
	Sig. (2-taile	.030	.058	.474	.022	.000	.000	.000	.000	.000	.210	.012	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Tables 13 show that nodes (Nodes) and ties (Edges_ud) have a strong positive correlation. As the number of nodes increase, the number of ties also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes), number of ties (Edges_d), Reciprocity and Total Paths (Tpaths_d) in the network. Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpinN) correlates positively with Paths Power Law Distribution per Node (PL_TpinN). Scale Free Metric (S_con) seems to share a positive relationship with Average Geodesic Length (AvgGL_d). Assortativity (R_con) shares a positive relationship with Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d), Average Geodesic Length (AvgGL_d) and Scale Free Metric (S_con). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centralization (ECin) correlates negatively with Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d). Power Law Distribution of

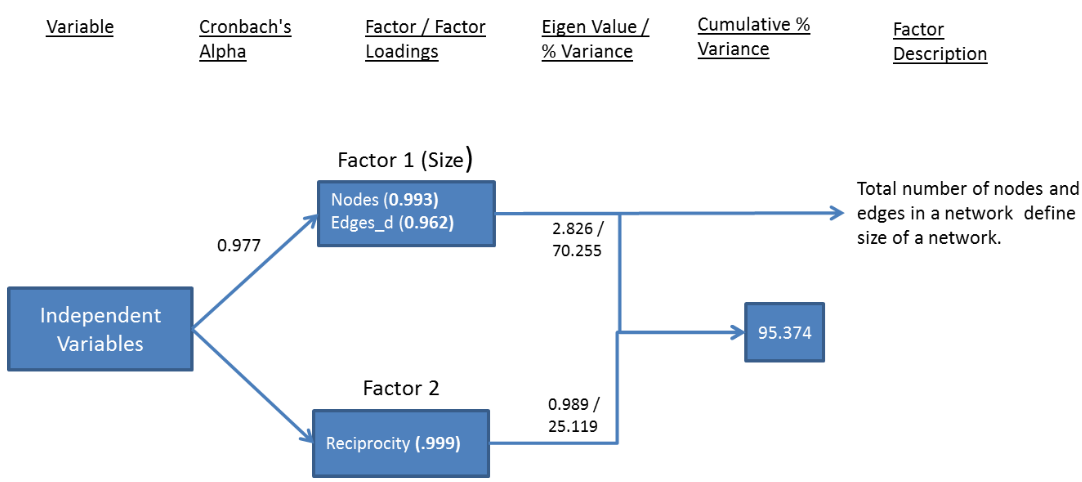
Eigenvector Centrality per Node (PL_EVCinN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN) and Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) correlate strongly with each other.

A.6.7.3.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.6.7.3.2.1 Independent Variables

Figure 22: Factor Analysis of Independent Variables

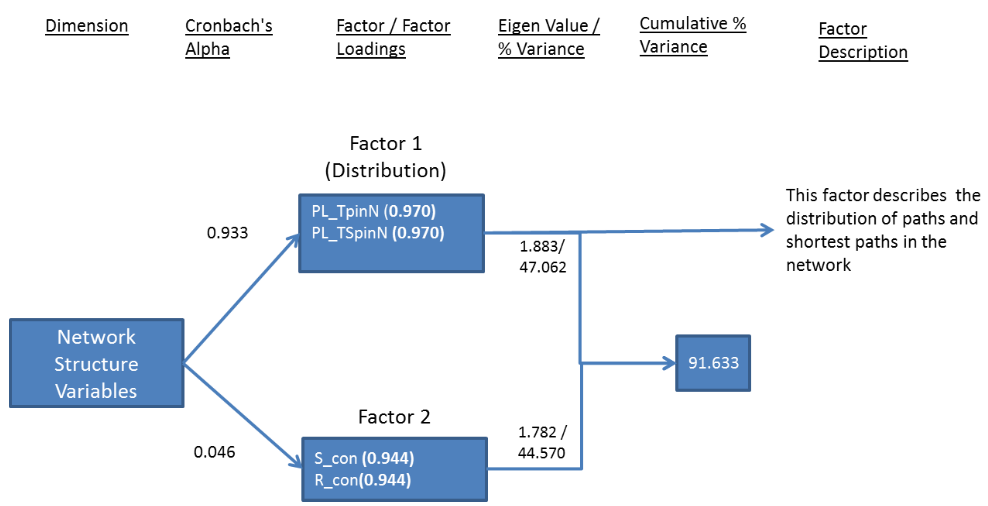


The factor analysis generated two factors that explain 95.374% (greater than 80%) of the cumulative variance. Factor1 has eigenvalues above one. Factor2 has eigenvalue below 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1.

Cronbach's alpha for factor1 has a value of 0.977. This means Nodes and ties (Edges_d) are measuring same construct within factor 1. Factor 1 is named "Size".

A.6.7.3.2.2 Network Structure (MV1)

Figure 23: Factor Analysis of Network Structure Variables

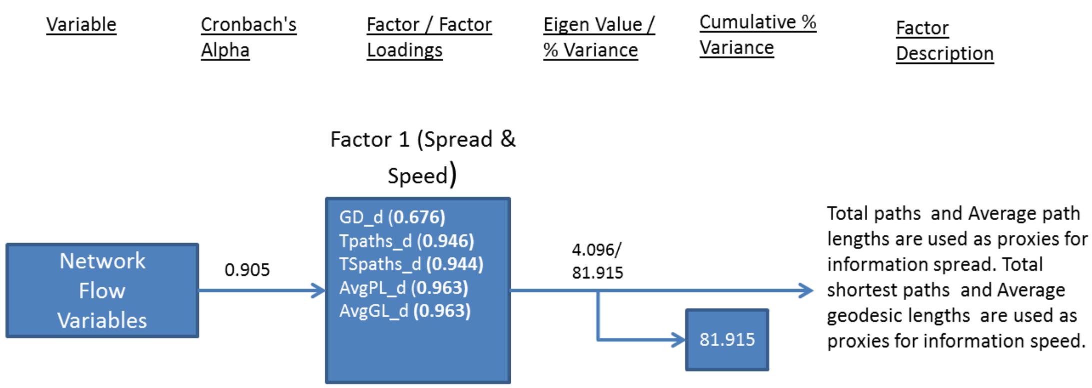


Factor analysis generated two factors that explain 91.633% (greater than 80%) of cumulative variance. Factor1 and factor 2 have eigenvalues above 1. Assortativity and scale free metric have significant factor loadings in factor2. Cronbach's alpha for factor2 has a value of 0.046. Assortativity (R_con) and Scale Free Metric (S_d) are not measuring same construct within factor2. Hence, they should not be considered as a factor. Paths Power Law Distribution per Node (PL_TpinN) and Shortest Paths Power Law Distribution per Node (PL_TSpinN) have significant factor loadings in factor1. Cronbach's alpha for factor1 has a value of 0.933. Paths Power Law Distribution per Node (PL_TpinN) and

Shortest Paths Power Law Distribution per Node (PL_TSpinN) are measuring same construct within factor 2. Hence, they should be considered as a factor. All other variables load independently. Factor2 is named as “Distribution”.

A.6.7.3.2.3 Network Flow (MV2)

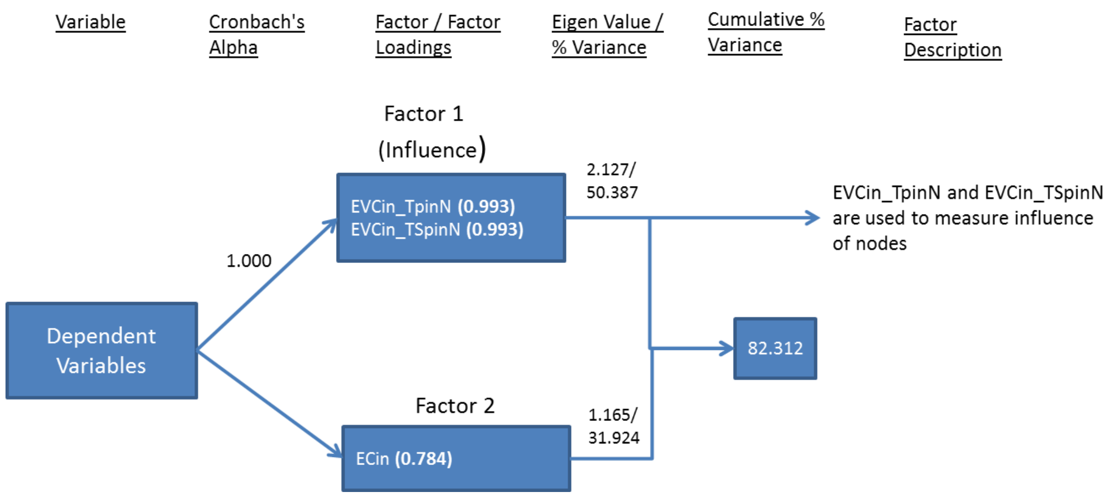
Figure 24: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 81.915% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (Tpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach’s alpha for factor1 has a value of 0.875. Factor 1 is named as “Spread and Speed”.

A.6.7.3.2.4 Dependent Variables

Figure 25: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 82.312% (greater than 80%) of cumulative variance. Eigenvector Centralities with respect to Paths (EVCin_TpinN) and Shortest Paths (EVCin_TSpinN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 1. I name the factor1 as "Influence" as both, Eigenvector centralities with respect to paths and shortest paths, are being used measure of influence.

A.6.7.3.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled “RegressionAnalysis_Science.pdf”.

A.6.7.3.3.1 Impact of Network Structure on Network Flow

Table 14: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Science	(0.285/0.000) [4]	(0.558/0.000) [2,3,4]	(0.543/0.000) [2,3,4]	(0.484/0.000) [2,3,4]	(0.456/0.000) [2,3,4]

Table 14 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 28.5%, 55.8%, 54.3%, 48.4% and 7.6% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on AvgGL_ud is not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.3.3.2 Impact of Network Flow on Network Structure

Table 15: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpinN	PL_TSpinN	S_con	R_con	SMSP_d
Science	(0.037/0.038) [10]	(0.086/0.003) [10]	(0.220/0.000)[9]	(0.475/0.000) [6,7]	NA

Table 15 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 27%, 26.5%, 36%, 30% and 8.7% variation in the PL_TpinN, PL_TSpinN, S_con, and R_con, respectively. The impact of network flow variables on PL_TpinN and PL_TSpinN are not taken into consideration, as their p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.3.3.3 Impact of Network Structure on Network Phenomenon

Table 16: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	ECin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Science	NA	NA	(0.167/0.000)[4]	(0.168/0.000)[4]

Table 16 shows that the network structure variable Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TpinN), explaining 16.7% and 16.8% variation respectively.

A.6.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 17: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Science	(0.097/0.002)[7]	(0.044/0.025)[9]	(0.205/0.000)[10]	(0.205/0.000)[10]

Table 17 shows that the network structure variable impacts Eigenvector Centralization (EC_in), Power Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCinN), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 3.4%, 27.4% and 10.1% variation respectively. The impact of network flow variables on EC_in and PL_EVCinN are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.3.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 18: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: **(1) PL_TpinN, (2) PL_TSpinN, (3) S_con, (4) R_con, (5)SMSP_d, (6)GD_d, (7) Tpaths_d, (8) TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecin	PL_EVCinN	EVCin_TpinN	EVCin_TSpinN
Science	(0.308/0.000) [8,15]	NA	(0.262/0.000) [4,10]	(0.262/0.000) [4,10]

Table 18 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_in), Eigenvector Centrality with respect to Total Paths per Node (EVCin_TpinN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCin_TSpinN), explaining 30.8%, 26.2%, and 26.2% variation respectively.

A.6.7.4 The Propagation Network

A.6.7.4.1 Correlation Analysis

Significant correlations coefficients for propagation network are shown below in table 19. Significant correlations observed are marked in yellow. All correlations between all variables are shown in supplemental file titled “Correlations.pdf”.

Table 19: Correlation coefficients of directed network

Correlations

		Nodes	Edges_d	Reciprocity	Den_d	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d	PL_TpoutN	S_pro	ECout	EVCout_TpoutN
Edges_d	Pearson	.965**	1											
	Sig. (2-ta	.000												
	N	91	91											
Den_d	Pearson	-.920**	-.834**	-.070	1									
	Sig. (2-ta	.000	.000	.510										
	N	91	91	91	91									
Tpaths_d	Pearson	.646**	.610**	.081	-.575**	.499**	1							
	Sig. (2-ta	.000	.000	.448	.000	.000								
	N	91	91	91	91	91	91							
TSpaths_d	Pearson	.647**	.609**	.075	-.576**	.495**	1.000**	1						
	Sig. (2-ta	.000	.000	.482	.000	.000	.000							
	N	91	91	91	91	91	91	91						
AvgPL_d	Pearson	.363**	.352**	.153	-.304**	.604**	.854**	.852**	1					
	Sig. (2-ta	.000	.001	.147	.003	.000	.000	.000						
	N	91	91	91	91	91	91	91	91					
AvgGL_d	Pearson	.363**	.348**	.137	-.308**	.594**	.856**	.855**	.998**	1				
	Sig. (2-ta	.000	.001	.196	.003	.000	.000	.000	.000					
	N	91	91	91	91	91	91	91	91	91				
PL_TSpoutN	Pearson	.237*	.265*	.111	-.226*	.227*	.131	.129	.190	.185	.689**			
	Sig. (2-ta	.024	.011	.295	.031	.031	.214	.223	.071	.080	.000			
	N	91	91	91	91	91	91	91	91	91	91			
S_pro	Pearson	-.017	.044	.251*	.055	.167	.282**	.275**	.552**	.538**	.248*	1		
	Sig. (2-ta	.871	.680	.016	.606	.115	.007	.008	.000	.000	.018			
	N	91	91	91	91	91	91	91	91	91	91	91		
R_pro	Pearson	.049	.097	.268*	-.016	.255*	.351**	.345**	.620**	.606**	.302**	.980**		
	Sig. (2-ta	.642	.358	.010	.881	.015	.001	.001	.000	.000	.004	.000		
	N	91	91	91	91	91	91	91	91	91	91	91		
PL_EVCoutN	Pearson	.300**	.298**	.680**	-.272**	.262*	.406**	.403**	.375**	.367**	.232*	.173	-.522**	
	Sig. (2-ta	.004	.004	.000	.009	.012	.000	.000	.000	.000	.027	.100	.000	
	N	91	91	91	91	91	91	91	91	91	91	91	91	
EVCout_TSpoutN	Pearson	.301**	.349**	.233*	-.203	.051	.276**	.275**	.242*	.239*	.128	.031	-.287**	.989**
	Sig. (2-ta	.004	.001	.027	.053	.630	.008	.008	.021	.023	.226	.769	.006	.000
	N	91	91	91	91	91	91	91	91	91	91	91	91	91

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

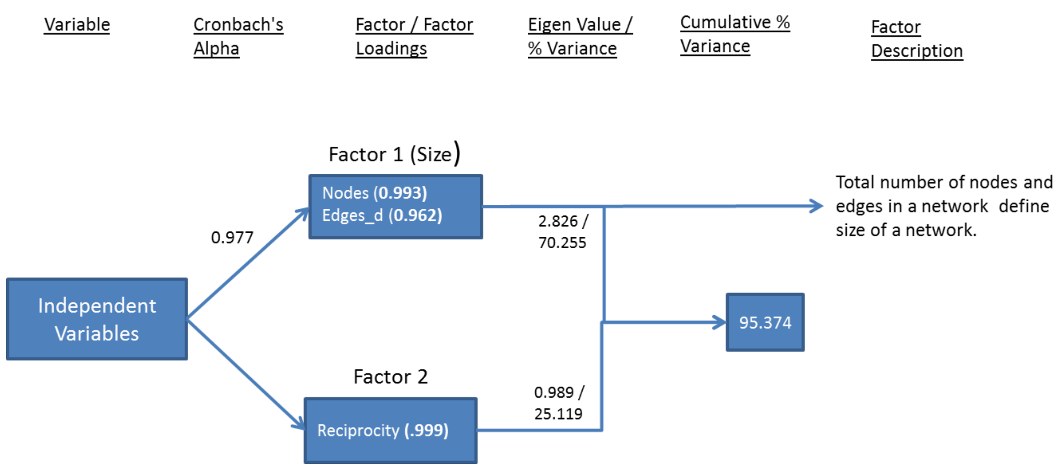
Table 19 shows that nodes and ties have a strong positive correlation. As the number of nodes (Nodes) increase, the number of ties (Edges_d) also increases. Density (Den_d) of this network has a strong negative correlation with both, number of nodes (Nodes) and number of ties (Edges_d). Total Paths (Tpaths_d) in the network correlate with number of ties (Edges_d), Reciprocity and the Graph Diameter (GD_d) of the network. Total Paths (Tpaths_d) in the network share a negative correlation with Density (Den_d). Total Shortest Paths (TSpaths_d) in the network correlate positively with the number of nodes (Nodes), number of ties (Edges_d), Reciprocity and Total Paths (Tpaths_d) in the network. Total Shortest Paths (TSpaths_d) in the network share a negative correlation with Density (Den_d). Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) correlates with Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d) and with each other. Shortest Paths Power Law Distribution per Node (PL_TSpoutN) correlates positively with Paths Power Law Distribution per Node (PL_TpoutN). Small World Metric (SMSP_d) is strongly correlated with Clustering Coefficient (CC_d). Eigenvector Centralization (ECout) correlates negatively Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d). Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN) shares a positive correlation with Reciprocity. Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN) and Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) correlate strongly with each other.

A.6.7.4.2 Factor Analysis

In this section, results of factor analysis are shown. Details of the factor analysis are shown in supplemental file titled "Factor Analysis.pdf".

A.6.7.4.2.1 Independent Variables

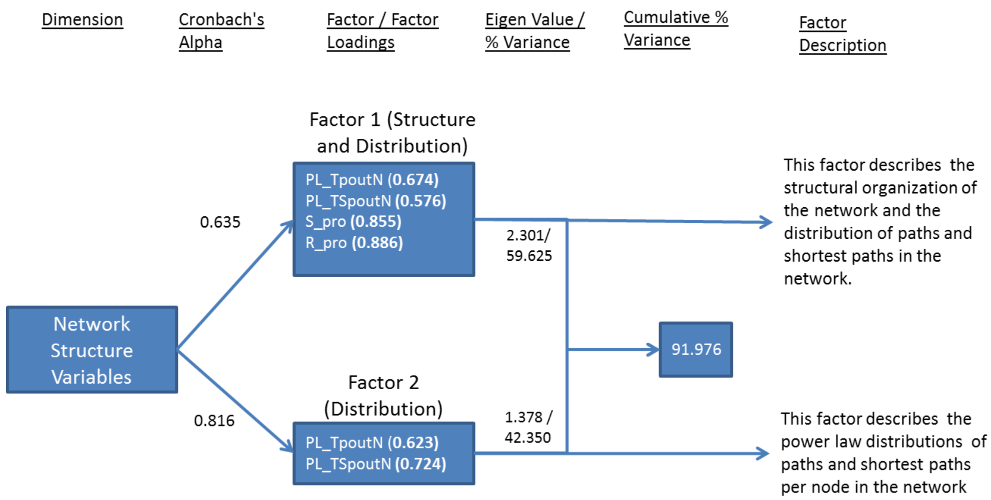
Figure 26: Factor Analysis of Independent Variables



The factor analysis generated two factors that explain 95.374% (greater than 80%) of the cumulative variance. Factor1 has eigenvalues above one. Factor2 has eigenvalue below 1. Nodes and ties (Edges_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.977. This means Nodes and ties (Edges_d) are measuring same construct within factor 1. Factor 1 is named "Size".

A.6.7.4.2.2 Network Structure (MV1)

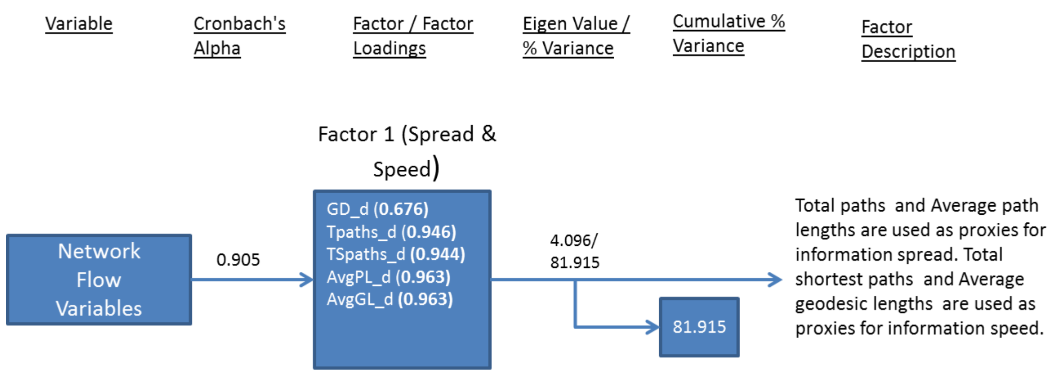
Figure 27: Factor Analysis of Network Structure Variables



Factor analysis generated two factors that explain 91.976% (greater than 80%) of cumulative variance. Factor1 and factor2 have eigenvalues above 1. Assortativity (R_pro), Scale Free Metric (S_pro), Paths Power Law Distribution per Node (PL_TpoutN) and Shortest Paths Power Law Distribution per Node (PL_TSpoutN) have significant factor loadings in factor1. Cronbach's alpha for factor2 has a value of 0.635. Assortativity (R_pro), Scale Free Metric (S_pro), Paths Power Law Distribution per Node (PL_TpoutN) and Shortest Paths Power Law Distribution per Node (PL_TSpoutN) are measuring same construct within factor1. Factor 1 is named "Structure and Distribution". Paths Power Law Distribution per Node (PL_TpoutN) and Shortest Paths Power Law Distribution per Node (PL_TSpoutN) have significant factor loadings in factor2. Cronbach's alpha for factor2 has a value of 0.816. Paths Power Law Distribution per Node (PL_TpoutN) and Shortest Paths Power Law Distribution per Node (PL_TSpoutN) are measuring same construct within factor 2. Factor2 is named as "Distribution".

A.6.7.4.2.3 Network Flow (MV2)

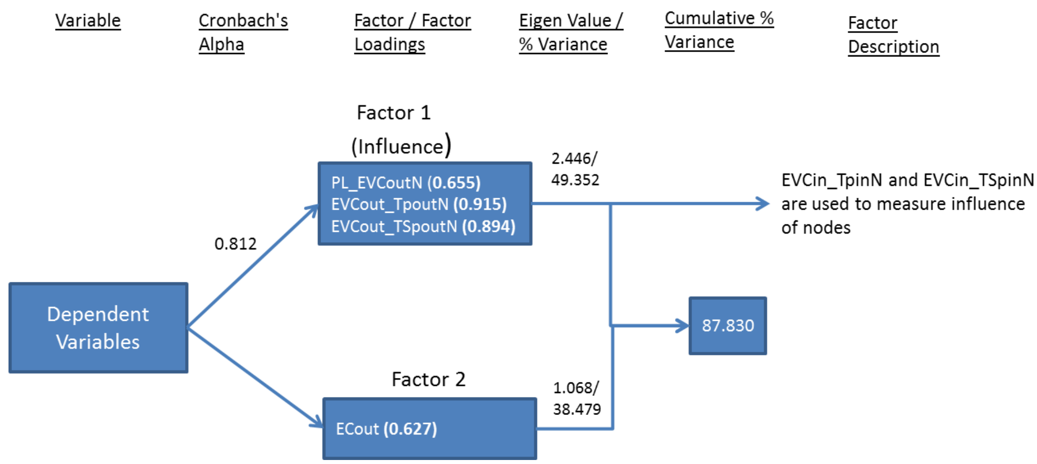
Figure 28: Factor Analysis of Network Flow Variables



Factor analysis generated one factor that explains 81.915% (greater than 80%) of cumulative variance. Graph Diameter (GD_d), Total Paths (TSpaths_d), Total Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_d) have significant factor loadings in factor 1. Cronbach's alpha for factor1 has a value of 0.905. Factor 1 is named as "Spread and Speed".

A.6.7.4.2.4 Dependent Variables

Figure 29: Factor Analysis of Dependent Variables



Factor analysis generated two factors that explain 87.830% (greater than 80%) of cumulative variance. Power Law Distribution of Eigenvector Centrality (PL_EVCoutN), Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN) have significant factor loading on factor 1. Factor 1 has a Cronbach's alpha of 0.812. Power Law Distribution of Eigenvector Centrality (PL_EVCoutN), Eigenvector Centralities with respect to Paths (EVCout_TpoutN) and Shortest Paths (EVCout_TSpoutN) are measuring same construct within factor 1. Factor 1 is named "Influence".

A.6.7.4.3 Regression Analysis

In this section, only the impactful regressions in which predictors had a significant impact on dependent variables are shown. Detailed regressions are shown in supplemental file titled "RegressionAnalysis_Science.pdf".

A.6.7.4.3.1 Impact of Network Structure on Network Flow

Table 20: Impact of Network Structure on Network Flow

Predictors: **(1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	GD_d	Tpaths_d	TSpaths_d	AvgPL_d	AvgGL_d
Science	(0.054/0.015) [4]	(0.113/0.001) [4]	(0.109/0.001) [4]	(0.377/0.000)[4]	(0.361/0.00) [4]

Table 20 shows that network structure variables have a significant impact on network flow variables. Network structure variables explain 5.4%, 11.3%, 10.9%, 37.7% and 36.1% variation in Graph Diameter (GD_d), Total Paths (Tpaths_d), Total Shortest Paths (TSpaths_d), Average Path Length (AvgPL_d) and Average Geodesic Length (AvgGL_ud), respectively. The impact of network structure variables on Graph Diameter (GD_d), Total Paths (Tpaths_d) and Total Shortest Paths (TSpaths_d) are not taken into consideration, as the p-value is greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.4.3.2 Impact of Network Flow on Network Structure

Table 21: Impact of Network Flow on Network Structure

Predictors:**(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]				
	PL_TpoutN	PL_TSpoutN	S_pro	R_pro	SMSP_d
Science	(0.046/0.023)[6]	(0.041/0.031)[6]	(0.297/0.000)[9]	(0.377/0.000)[9]	NA

Table 21 shows that the network flow variables have a significant impact on the network structure variables. Network flow variables explain 4.6%, 4.1%, 29.7%, and 37.7% variation in the PL_TpoutN, PL_TSpoutN, S_pro, and R_pro, respectively. The impact of network flow variables on PL_TpoutN and PL_TSpoutN are not taken into consideration, as the p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.4.3.3 Impact of Network Structure on Network Phenomenon

Table 22: Impact of Network Structure on Network Phenomenon

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Science	(0.126/0.000) [4]	(0.043/0.027) [1]	NA	NA

Table 22 shows that the network structure variable impacts Eigenvector Centralization (ECout) and Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN, explaining 12.6% and 4.3% variation respectively. The impact of network flow variables on Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN) is not taken into consideration, as the p-values is greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.3.3.4 Impact of Network Flow on Network Phenomenon

Table 23: Impact of Network Flow on Network Phenomenon

Predictors: **(6)GD_d, (7)Tpaths_d, (8)TSpaths_d, (9) AvgPL_d, (10) AvgGL_d**

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Science	(0.097/0.002)[7]	(0.155/0.000)[7]	(0.066/0.008)[7]	(0.066/0.008)[7]

Table 23 shows that the network structure variable impacts Eigenvector Centralization (ECout), Powel Law Distribution of Eigenvector Centrality with respect to Nodes (PL_EVCoutN), Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN) and Eigenvector Centrality with respect to Total Shortest Paths per Node (EVCout_TSpoutN), explaining 9.7%, 15.5%, 6.6% and 6.6% variation respectively. The impact of network flow variables on EC_out, EVCout_TpoutN and EVCout_TSpoutN are not taken into consideration, as their respective p-values are greater than the Bonferroni-adjusted p-value of 0.000694.

A.6.7.4.3.5 Collective Impact of Independent Variables, Moderating Variables (Network Structure and Network Flow Variables) on the Network Phenomenon Variables.

Table 24: Collective Impact of Independent Variables, Moderating Variables on the Network Phenomenon Variables

Predictors: (1) PL_TpoutN, (2) PL_TSpoutN, (3) S_pro, (4) R_pro, (5)SMSP_d, (6),GD_d (7) Tpaths_d (8), TSpaths_d, (9) AvgPL_d, (10) AvgGL_d, (11) Nodes, (12) Edges_d, (13) Den_d, (14) CC_d, (15) Reciprocity

	Dependent Variable (Adjusted R Square/ Significance)[Predictors]			
	Ecout	PL_EVCoutN	EVCout_TpoutN	EVCout_TSpoutN
Science	(0.308/0.000) [7,15]	(0.577/0.000) [8,15]	(0.065/0.009) [15]	NA

Table 24 shows the collective impact of independent and moderating variables on the network phenomenon variables. The independent variables and the moderating variables collectively impact Eigenvector Centralization (EC_out), Power Law Distribution of Eigenvector Centrality per Node (PL_EVCoutN), and Eigenvector Centrality with respect to Total Paths per Node (EVCout_TpoutN), explaining 30.8%, 57.7%, and 6.5% variation respectively.

Appendix B: Supplemental Files

Name	Size	Software Requirement	Description
Correlations.pdf	583 kB	Adobe Acrobat Reader	Detailed correlations of all the variables in Comedy, Entertainment, Music, Howto, Science and Sports product categories
Daily_Values_Com_con.pdf	241 kB	Adobe Acrobat Reader	Daily values of all variables of Comedy Network in Consumption Phase for 91 days
Daily_Values_Com_d.pdf	239 kB	Adobe Acrobat Reader	Daily values of all variables of Comedy Network in Directed Phase for 91 days
Daily_Values_Com_pro.pdf	241 kB	Adobe Acrobat Reader	Daily values of all variables of Comedy Network in Propagation Phase for 91 days
Daily_Values_Com_ud.pdf	236 kB	Adobe Acrobat Reader	Daily values of all variables of Comedy Network in Undirected Phase for 91 days
Daily_Values_Ent_con.pdf	241 kB	Adobe Acrobat Reader	Daily values of all variables of Entertainment Network in Consumption Phase for 91 days
Daily_Values_Ent_d.pdf	242 kB	Adobe Acrobat Reader	Daily values of all variables of Entertainment in Directed Phase for 91 days
Daily_Values_Ent_pro.pdf	241 kB	Adobe Acrobat Reader	Daily values of all variables of Entertainment Network in Propagation Phase for 91 days
Daily_Values_Ent_ud.pdf	236 kB	Adobe Acrobat Reader	Daily values of all variables of Entertainment Network in Undirected Phase for 91 days
Daily_Values_Howto_con.pdf	239 kB	Adobe Acrobat Reader	Daily values of all variables of Howto Network in Consumption Phase for 91 days
Daily_Values_Howto_d.pdf	240 kB	Adobe Acrobat Reader	Daily values of all variables of Howto in Directed Phase for 91 days

Daily_Values_Howto_pro.pdf	240 kB	Adobe Acrobat Reader	Daily values of all variables of Howto Network in Propagation Phase for 91 days
Daily_Values_Howto_ud.pdf	234 kB	Adobe Acrobat Reader	Daily values of all variables of Howto Network in Undirected Phase for 91 days
Daily_Values_Mus_con.pdf	245 kB	Adobe Acrobat Reader	Daily values of all variables of Music Network in Consumption Phase for 91 days
Daily_Values_Mus_d.pdf	244 kB	Adobe Acrobat Reader	Daily values of all variables of Music in Directed Phase for 91 days
Daily_Values_Mus_pro.pdf	245 kB	Adobe Acrobat Reader	Daily values of all variables of Music Network in Propagation Phase for 91 days
Daily_Values_Mus_ud.pdf	236 kB	Adobe Acrobat Reader	Daily values of all variables of Music Network in Undirected Phase for 91 days
Daily_Values_Sci_con.pdf	240 kB	Adobe Acrobat Reader	Daily values of all variables of Science Network in Consumption Phase for 91 days
Daily_Values_Sci_d.pdf	241 kB	Adobe Acrobat Reader	Daily values of all variables of Science in Directed Phase for 91 days
Daily_Values_Sci_pro.pdf	240 kB	Adobe Acrobat Reader	Daily values of all variables of Science Network in Propagation Phase for 91 days
Daily_Values_Sci_ud.pdf	234 kB	Adobe Acrobat Reader	Daily values of all variables of Science Network in Undirected Phase for 91 days
Daily_Values_Spo_con.pdf	242 kB	Adobe Acrobat Reader	Daily values of all variables of Sports Network in Consumption Phase for 91 days
Daily_Values_Spo_d.pdf	250 kB	Adobe Acrobat Reader	Daily values of all variables of Sports in Directed Phase for 91 days
Daily_Values_Spo_pro.pdf	242 kB	Adobe Acrobat Reader	Daily values of all variables of Sports Network in Propagation Phase for 91

			days
Daily_Values_Spo_ud.pdf	236 kB	Adobe Acrobat Reader	Daily values of all variables of Sports Network in Undirected Phase for 91 days
Factor Analysis.pdf	7.1 MB	Adobe Acrobat Reader	Detailed Factor Analysis output of all the variables in Comedy, Entertainment, Music, Science and Sports product categories
Meta Data.pdf	248 kB	Adobe Acrobat Reader	Daily values of Meta Data for Comedy, Entertainment, Music, Howto, Science and Sports product categories for 91 days
RegressionAnalysis_Comedy.pdf	11.2 MB	Adobe Acrobat Reader	Detailed Regression Analysis output of all the variables in Comedy product category
RegressionAnalysis_Entertainment.pdf	11.1 MB	Adobe Acrobat Reader	Detailed Regression Analysis output of all the variables in Entertainment product category
RegressionAnalysis_Music.pdf	12.1 MB	Adobe Acrobat Reader	Detailed Regression Analysis output of all the variables in Music product category
RegressionAnalysis_Science.pdf	14.3 MB	Adobe Acrobat Reader	Detailed Regression Analysis output of all the variables in Science product category
RegressionAnalysis_Sports.pdf	14.6 MB	Adobe Acrobat Reader	Detailed Regression Analysis output of all the variables in Sports product category