June 2019

Measuring Influence Across Social Media Platforms: Empirical Analysis Using Symbolic Transfer Entropy

Abhishek Bhattacharjee
University of South Florida, abhishekb1@mail.usf.edu

Measuring Influence Across Social Media Platforms: Empirical Analysis Using Symbolic Transfer Entropy

by

Abhishek Bhattacharjee

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Adriana Iamnitchi, Ph.D.
John Skvoretz, Ph.D.
Giovanni Luca Ciampaglia, Ph.D.

Date of Approval:
April 29, 2019

Keywords: Social Networks, Cross-platform influence

DEDICATION

Dedicated to my parents, brother, girlfriend and those five humans.

ACKNOWLEDGMENTS

I would like to pay my thanks to my advisor Dr Adriana Iamnitchi for her constant support and feedback throughout this project. The work wouldn't have been possible without Leidos who provided us the data, and DARPA for funding this research work. I would also like to acknowledge the contributions made by Dr. Nazim Choudhury whose feedback, discussions and ideas have enhanced the project. I also want to thank Sameera Horawalavithana for letting me use his cascade generation code. Finally, I would like to thank all the members of DSG lab including my advisor Dr. Adriana Iamnitchi for creating a conducive work environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Social media platforms are interconnected environments that influence each other. Information from one social media platform spreads to another. This thesis proposes a platform-independent framework to analyze information transfer across social media platforms. This thesis uses Symbolic Transfer Entropy and Statistical Significance Test to measure influence and optimize the time window of influence between different platforms. To validate the framework, the thesis analyses the temporal activity dynamics and the information transfer across three different platforms, Reddit, Twitter and GitHub.

Two data driven studies are described in this thesis. The first study finds the optimum time windows of influence between the three platforms during two different cyber attack events on cryptocurrency exchanges. It finds that specific types of activities are more influential than others, and optimum time interval changes with pre, during, and post event days. The second study applies information revealed in the first study and specifically the optimal time window to link cross-platform information cascades from Twitter and Reddit. The case-study is a heuristic that, we show, can reduce the search space for connecting information cascades across different platforms.

CHAPTER 1: INTRODUCTION

Today's social media ecosystem is formed of many interacting platforms: videos on YouTube are commonly shared on Facebook and Twitter; snapshots of tweets are frequently posted on Facebook; 73% of the American public uses at least two social media platforms, with the median using three [1]. Information thus, of personal or public value, travels between social media platforms. Even technical information traverses the boundaries of any one platform. For example, software vulnerabilities are often publicized on Twitter [2, 3], soon discussed on Reddit or other forums, and potentially lead to software development activities on GitHub.

Quantifying the transfer of information from one platform to another can have predictive power: the activities on one platform can forecast activities on a different platform. This predictive power can guide intervention techniques such as recruitment of volunteer software developers to fix an urgent vulnerability, or limit misinformation campaigns [4].

Measuring information transfer has been done using transfer entropy [5] in contexts as diverse as social media [6], epidemiology [7], neuroscience [8] and economics [9]. Measuring transfer entropy depends on a chosen time scale. Simply stated, the events in each platform need to be represented by a time series that records the number of particular events per time window. The choice of the time window is arbitrary at best [10], even for measuring the transfer entropy between parts of the same system.

The choice of the time window becomes a more clear challenge when comparing platforms with very different activity rates. For example, some platforms such as Twitter promote fast information spread. Other platforms facilitate contributions that take longer to create, such as software bug fixes, as in GitHub, or book reviews, as on Amazon or Goodreads. Consequently, the rate of activities generated on two platforms can be at opposite ends of the spectrum. Therefore, choosing an arbitrary time window, although widely practiced, to measure information transfer between such platforms becomes questionable.

This work seeks to propose a generic framework which can be applied to determine influence between two social media platforms at the macro-level, and the fitting time window of information transfer between two platforms. The framework proposed in this work not only addresses the selection of an fitting time window and the direction of influence between the platforms, but also handles the scale difference between the platforms. This work is based on the information-theoretic concept of transfer entropy, thus it is agnostic of the type of platforms under analysis.

To show the applicability of our methodology, the thesis presents two case-studies in the subsequent chapters. The first is an empirical study that involves three highly different platforms: GitHub, Reddit, and Twitter. The challenges of measuring the transfer entropy between these platforms is mainly due to different activity rates. Users in GitHub make programming-related contributions via actions on software repositories, such as IssueComment, Push/Pull, Create/Delete [11]. Users on Reddit engage in subject-focused conversations via comments to original posts or other users' comments. At the other extreme, in Twitter users broadcast short messages with a push of a button. This experiment analyses the information dynamics between the platforms during two different cyber-attacks on crypto-currency exchanges. The second is using the framework to connect information cascades between Twitter and Reddit. These information cascades are discussions containing mentions of various Common Vulnerabilities and Exposures (CVE) references.

This thesis makes the following contributions:

1. It proposes a generic methodology for determining the fitting time interval for measuring information transfer between two social media platforms;
2. It uses a statistical significance test to build trust on determining influence between the activities on different platforms;
3. It studies how a platform influences another during an external event, and how different activities within a platform influence activities on another platform;
4. To the best of our knowledge, this is the first study to apply symbolic transfer entropy to measure cross-platform influence;
5. It demonstrates that different time windows are best for different epochs in the duration of an event;

6. It demonstrates that, at least during the intervals and the events considered, primary activities on a platform drive cross-platform influence;

7. It presents a heuristic to connect information cascades across social media platforms.

The rest of the thesis is structured as follows. Chapter 2 presents the state of the art in comparing the different social media platforms. Chapter 3 provides the background information on Transfer Entropy and Symbolic Transfer Entropy. In addition, it introduces the use of a statistical significance test to reject the null hypothesis that the observed influence are not coincidental, and explains the framework thus conceived. Chapter 4 explains in detail the datasets used for the two case-studies. In chapters 5 and 6, the studies are presented conveying the applicability of the framework. Finally, the thesis concludes in Chapter 7.

CHAPTER 2: LITERATURE REVIEW

Studies that focus on multiple social media platforms can be classified in two categories: (i) comparative studies on user behavior on different platforms, and (ii) multi-platform analyses to understand the interrelation patterns across diverse platforms.

Comparative studies of multiple social platforms aim to identify the different emerging user behaviors on each platform. Silvestri et al.[12] analyzed user attributes, platform-specific services, and different matching strategies to propose a methodology to link user accounts in Twitter, GitHub and Stack Overflow. The authors also perform a comparative study on user interaction networks in these three platforms, including the correlations between interactions across different networks.

By analyzing different types of content (e.g., movie, games, news) from different domains in popular social media platforms such as Reddit, Twitter, Facebook, YouTube and Google Plus, Haralabopoulos et al.[13] framed the multi-platform social media analysis as a multi-layer social networks analysis. The authors demonstrate that information flow, content diffusion propagation and virality differ in amount, rate and impact across layers of social networks.

Waterloo et al.[14] examined the emotion expression in Facebook, Twitter, Instagram and WhatsApp to compare the social norms of behavior across different platforms. Shane-Simpson et al. [15] focused on the reasons for choosing a particular social media by college students and the consequences of these choices. The authors used different types of social capital (i.e., bonding and bridging [16]) and performed both content and statistical analyses to answer questions such as *who* is drawn to popular social media sites, *why* they prefer each site, and the *what are the social consequences* of these site preferences.

Studies of interactions between social platforms aim to identify how the use of one platform might depend upon that of other platforms. For example, Vasilescu et al. [17] analyzed the interplay of involvement and productivity of developers in GitHub and StackOverflow. The authors showed that active GitHub com-

mitters ask fewer questions and provide more answers than others and the activity in the StackOverflow correlates the coding change activity in GitHub despite any interruptions incurred. Similarly, Badashian et al. [18] demonstrated the correlation between contributions, management/editorial activities and popularity both within and across GitHub and Stack Overflow.

To correlate the social media usage with the increase of psychological well-being, Pittman and Reich [19] analyzed Instagram, Snapchat, Twitter and Yik Yak to identify how image-based social media (Instagram) can offer enhanced intimacy over text-based social media (Twitter). Quan-Haase and Young [20] conducted a comparative study to examine the level of gratifications achieved by university students from Facebook in comparison to instant messaging. They show that Facebook is predominantly used for having fun and knowing about social activities, whereas instant messaging is used for relationship maintenance and development. Papacharissi [21] provides a comparative discourse analysis using Facebook, LinkedIn and ASmallWorld to examine the symbolic representations of daily communicative routines that users of these social networks experience. Karapanos et al. [22] deployed both quantitative and qualitative analyses to characterize WhatsApp as a social media that unlocks new opportunities for intimate communications and Facebook as a social media primarily used for non-social purposes, despite the fact that both platforms support powerful life-logging tools. Finally, Boczkowski et al. [23] analyzed WhatsApp, Facebook, Twitter, Instagram and Snapchat with the help of four different frameworks focusing on two critical dimensions of social media practice: audience type and temporal dynamics. The authors report that users use one platform in ways related to their usage of the others and users' perceptions and understanding of each platform often include recursive references to the other social media options.

Recent studies also considered the perception of information transfer within single platform or between multiple platforms. Ver Steeg and Galstyan [24] suggested a model-free information-theoretic approach capitalizing on the notion of Transfer Entropy (TE) to characterize and quantify the causal information flow for any pair of users in Twitter. Kim et al. [6] used TE to observe the influence of one platform onto the other for information diffusion. The authors divided different platforms into three categories, namely News, Social Networking Sites(SNS), and Blogs, and used TE to determine the direction of influence across heterogeneous systems at macro level (population). However, similar to others, the authors consider

the propagation of an information piece (e.g., news item) and the number of adopters of that information item, which is a fragment of the population related to a particular news item. Further, like most of the approaches analyzing information transfer across multiple platforms, the authors ignored the variation of the time component in generating time series information.

Borge-Holthoefer et al. [25] use Symbolic Transfer Entropy (STE) to define and measure the temporal and structural signatures of collective social events as they emerge and gain attention in disparate geographical locations. The authors consider a detailed empirical analyses on five case studies and micro-blogging time series constructed from Twitter. Their objectives were to determine the intrinsic time scale of the information flow, to extract directed networks of influence among geo-localized sub-units in social systems, and to characterize the sampling intervals required to understand the evolution of social events. However, all these objectives were pursued within the same platform, Twitter.

This thesis is predominantly built on these last three studies. In addition to focusing on the influence between multiple platforms, it accounts for different types of activities within each platform. Moreover, it focuses on a building a framework that can be used to ascertain influence or information transfer between platforms while taking the scale of the platforms into account as well as evaluating different time intervals to determine the fitting (most relevant) time window for information transfer between these platforms. It also evaluates different time scales to understand the amount and rate of information transfer before, during, and after an event across different types of social media platforms. The work also emphasized on statistical significance test to preclude the possibility of influence (or STE) being coincidental. This study uses the proposed framework to present empirical results from information-driven collective phenomena (i.e. cyber attacks on a cryptocurrency) and demonstrates how the choice of characteristic time scale impacts the measure of information transfer between platforms (Chapter 5), and linking cascades between Reddit and Twitter (Chapter 6). The proposed framework is discussed in the next chapter.

CHAPTER 3: PROPOSED FRAMEWORK

Due to their inherent properties and nature of interaction within different social media platforms, information transfers faster in one platform than the others. Similarly, both the quality and quantity of information vary across heterogeneous platforms. Although it is widely believed that one platform influences the others, the temporal variation in measuring how much one platform excites the others has drawn little attention. In the following sections, we present an information-theoretic approach which is widely used in measuring information transfers across temporal systems.

3.1 Background

Cross-correlation or mutual information are the commonly used techniques to estimate inter-dependencies and causal relationships among multiple variables in time series analysis. However, these metrics are symmetric and do not allow for a causal interpretation. Schreiber proposed a non-parametric approach to test the causality based on information theory by introducing the concept of transfer entropy (TE) [5]. It is also a widely used approach to quantify the flow of information between time series [26]. In contrast to the Granger causality, TE is framed in terms of resolution of uncertainty [27]. As mentioned in Chapter 2, TE has been widely used in quantifying information transfer both in single and multi-platform analyses.

3.1.1 Transfer Entropy

Given two stochastic processes, X and Y, where the occurrences of an event in the process is a function of time, the process can be denoted as:

$$X = \{X_t : 0 < t_1 < t_2 < ...\}$$

$$Y = \{Y_t : 0 < t_1 < t_2 < ...\}$$

Now, Transfer Entropy ($TE_{X\to Y}$) can be defined as the reduction in the uncertainty of Y_t given the history of both the processes X and Y. Consider a time series produced by one of these processes, say X, then the sample probabilities of each behaviour $x \in X$ can be computed. The entropy of the process X can be calculated as:

$$H(X) = -\sum_{x\in X} p(x)\log p(x) \tag{3.1}$$

The entropy, $H(X)$, for the time series generated by process X (from Eq. 3.1) can be interpreted as the information gain when a new value of X is observed. In our case studies, processes $XandY$, represent the event counts at specified time intervals for each platform.

Next is the conditional entropy, $H(Y|X)$, which refers to the information gain when a new value in Y is observed given that the value in X is already known. It can be calculated as:

$$H(Y|X) = -\sum_{\substack{x\in X\\ y\in Y}} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{3.2}$$

Eq. 3.2 quantifies the dependency of Y on X. If $H(Y|X)$ is 0, then it can be said that there is no information gain when a new value of Y is observed given that X is already known. This suggests that Y is completely dependent on X. On the other hand, if we have a scenario where $H(Y) = H(Y|X)$, then the information gain when a new event in Y is seen without knowing X is the same as the information gain when X is already known, clearly suggesting that Y is independent of X.

From [5], if we assume that the system can be represented as a stationary Markov process of order k for both processes Y and X, where k is the past number of events affecting the present, then the transition probabilities of these events can be used to infer a dependency between the processes. Firstly, assume that the transition probabilities in processes X and Y are independent of each other, i.e.:

$$p(Y_t|Y_{t-1}) = p(Y_t|Y_{t-1},X_{t-1}) \tag{3.3}$$

As proposed in [5] the degree to which the assumption in Eq. 3.3 is violated can be measured by Eq. 3.1 (also known as Kullback entropy):

$$TE_{X\to Y} = \sum p(Y_t, Y_{t-1}, X_{t-1}) \log \frac{p(Y_t|Y_{t-1}, X_{t-1})}{p(Y_t|Y_{t-1})} \tag{3.4}$$

Transfer entropy, as defined in Eq. 3.4 gives the degree to which the assumption in Eq. 3.3 can be broken. Intuitively, transfer entropy is a measure of reduction in uncertainty of an event given the history of another event.

3.1.2 Symbolic Transfer Entropy

Considering the importance and applications of TE, various techniques have been proposed to estimate TE from observed data. However, most techniques experience the encumbrance of its parameter tuning, high sensitivity to noise contribution, and the inefficiency of the binning methods attributing bias in estimating entropies [28]. For example, when analyzing the influence between social platforms of different size, the difference in the magnitude of activities on the platforms can be a confounding factor.

Staniek and Lehnertz [29] introduced Symbolic Transfer Entropy (STE) that replaces numerical values in the time series with symbols. STE defines symbols by reordering the amplitude values of time series. For a time series $X = x(i), x(i+l), ..., x(i+(n-1)l)$ with m amplitude values and l as the time delay/lag, STE ensures that every $X_i = x(i), x(i+l), ..., x(i+(m-1)l)$ is uniquely mapped onto one of the $m!$ possible permutations of symbols. Thus, a timeseries of length n will become a timeseries of length $n-m+1$. Each of the small arrays of size m will be assigned a symbol.

For clarity, consider the following example. Assuming a timeseries of length, $n = 9$:

$$X = \{20, 15, 45, 32, 17, 29, 43, 10, 12\}$$

Figure 3.1 visually demonstrates the process of symbolization adopted from [25]. Considering $m = 3$, there are $m! = 6$ different symbols. Thus the time series X can be mapped into seven buckets of three numbers (the X_i) representing a symbol.

Thus, given the symbol sequences, the STE is defined as:

$$T^S_{Y,X} = \sum p(X_{i+\delta}, X_i, Y_i)\ log\ \frac{p(X_{i+\delta}|X_i, Y_i)}{p(X_{i+\delta}|X_i)}$$

where the sum operates over all symbols and δ denotes a time scale or sampling interval. The direction of information transfer is measured by the directionality index $T^S = (T^S_{X,Y} - T^S_{Y,X})$ where positive value of T^S exhibits unidirectional coupling with X being the influencer and the negative value signifies Y as driving X. For symmetric bi-directional coupling, $T^S = 0$.

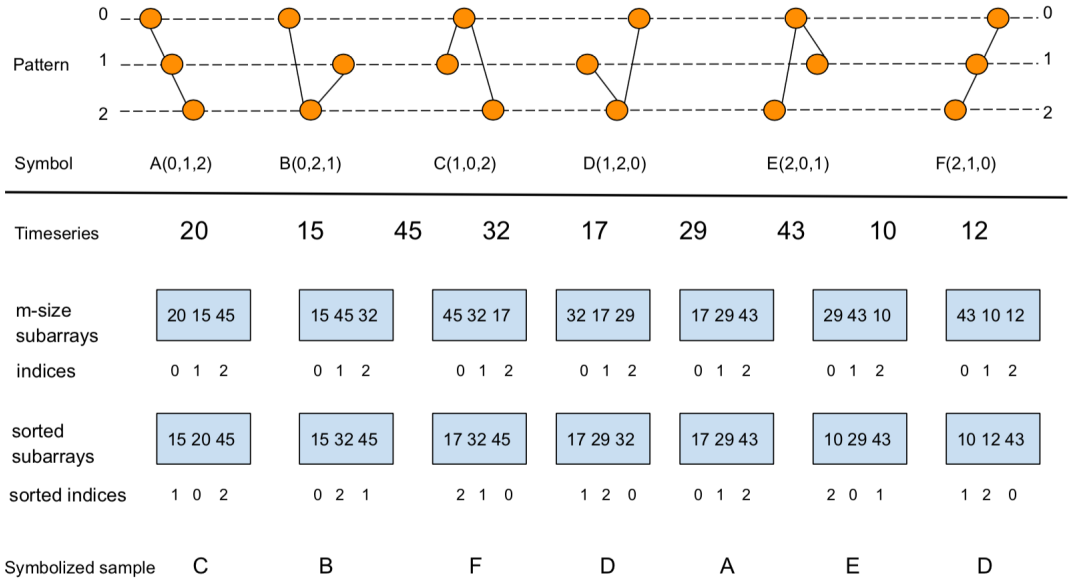


Figure 3.1: Visual illustration of the symbolization process used in STE. The time series x has an embedding dimension or amplitude $m = 3$ and a sliding window size $\omega = 1$.

3.2 Framework for Influence Measurement

This work uses STE to propose a generic framework to infer influence between different platforms, and the intrinsic time interval for the transfer of information or influence between these platforms. The framework is agnostic of the types of platforms, and also independent of any micro-level entities(such as URLs, images etc) to determine information flow between two different platforms. After obtaining the

STE value between two different platforms, this work applies a statistical significance test to reject the null hypothesis that the obtained STE value is not a mere coincidence.

3.2.1 Statistical Significance Test

The method described in 3.1.2 calculates STE to determine the amount or direction of information transfer across platforms, but this study is also interested in statistical significance of the calculated results to preclude the hypothesis that the influence measure could be a mere coincidence. To accomplish this, for each temporal scale or sampling interval considered, a randomized trial is also considered where random symbols are generated for each influencing time series. For example, in the case of $STE_{X\to Y}$, which quantifies how much a dynamic process X is driving or influencing another process Y, this study generates random symbols for process X and calculates the sample STE for $STE_{\hat{X}\to Y}$. This trial is repeated 100 times and a two tailed t-test is performed to measure the statistical significance of the observed STE $STE_{X\to Y}$ against the mean of the randomized sample STEs $STE_{\hat{X}\to Y}$. If μ denotes STE $STE_{X\to Y}$ and m_0 denotes mean of 100 trials STE $STE_{\hat{X}\to Y}$, then mathematical representations of the null and alternative hypotheses are:

$H_0 : \mu = m_0$

$H_1 : \mu > m_0$ (two-tailed)

Table 3.1: Statistical significance test example

Window (min)	STE	p-value	t-statistic	Pass
15	1.09	$3.88e^{-12}$	-5.71	Yes
20	1.14	$3.66e^{-8}$	-5.89	Yes
25	1.24	0.94	0.28	No
30	1.40	$4.60e^{-29}$	16.03	Yes

Table 3.1 shows an example of applying the statistical significance test for symbolic transfer entropy from GitHub contributions to Reddit comments (see Chapter 5) for different time window sizes. In this analysis the time windows that lead to a p-value larger than 0.05 (i.e. <95% confidence interval) are ignored, it can be seen that the time windows that pass this test have a negligible p-value, suggesting a very high confidence of the statistical significance of the STEs. A window 30 minutes is the fitting window of influence for GitHub contributions to Reddit comments.

3.2.2 Putting it Together

This work uses the concepts explained above to propose a framework that can be used to determine influence across different social media platforms, and the optimum time window for the transfer of information between two platforms. Assume two timeseries X and Y. To determine the influence of X on Y and the fitting time window of influence the following steps are carried out:

- Sample both X and Y into multiple time windows, such that the length of the timeseries is at least $5m!$ [30], where m is the amplitude for symbolization.
- For each of the time window do the following
  - Calculate $STE_{X_w \to Y_w}$, that is the influence of X_w on Y_w, where X_w and Y_w are the timeseries sampled by time window of w units.
  - Randomize the influencing timeseries (X_w) 100 times and calculate STE values for all the randomized timeseries
  - Run the upper-tailed one-sample-t-test for the actual STE and the randomized STE values
  - If the actual STE value is statistically significant with a confidence of 95% (i.e. p-value <= 0.05), then the STE value is selected and the time window is considered.
- From all the statistically significant STE values find the time window resulting in the highest STE value. This time window is considered to be the fitting time window of influence from X to Y, and the STE is used to quantify the influence.

The next chapter explains the datasets used for the two case studies that use the framework described above to find cross-platform influence.

CHAPTER 4: DATASETS

This chapter explains the datasets used for the two case studies explained in chapters 5 and 6. As mentioned earlier the first study is about two cyber attack events on cryptocurrency exchanges. In this study, we look at three platforms, GitHub, Reddit and Twitter. The dataset and the context is explained in Section 4.1. The second study is about cross-linking of information cascades on Reddit and Twitter. These cascades contain discussions on various Common Vulnerability and Exposure (CVE). The dataset for this study is presented in Section 4.2.

4.1 Cyber Attack Events

As cryptocurrencies have become more mainstream, there has been a rise in attacks on cryptocurrency exchanges. Exchanges are the services through which cryptocurrencies are traded. There are vulnerabilities between these cryptocurrency exchanges and individual user wallets [31]. These exchanges are vulnerable to thefts and phishing attacks, as we will see in this section. For our analysis, we have selected two such major attacks on two popular cryptocurrency exchanges.

4.1.1 Bitfinex

The first event happened on August 3, 2016, when 'Bitfinex', a digital currency exchange located in Hong Kong, was hit by an 'insider' attack resulting in the theft of 119,756 BTC worth of about $61M. Previously, Bitfinex had adopted a new BTC settlement system and security architecture based on the multi-signature technology supported by 'BitGo', a California-based blockchain security company. Each account in Bitfinex was protected by three keys. One key belonged to the user, one to Bitfinex, and one to Bitgo. Two out of the three keys were considered sufficient to move funds. Bitfinex used a special API key to instruct BitGo to provide its signature programmatically. According to a senior finance and economic contributor

at 'Forbes' [32], to compromise a Bitfinex account, the attacker had the keys belonging to the Bitfinex exchange, and the Bitfinex-to-BitGo API key that allowed the hacker to instruct BitGo to sign the transaction. This also enabled the attacker to turn a compromise at one Bitfinex location into many withdrawals.

4.1.2 Bithumb

In the second event, 'Bithumb', the largest bitcoin exchange in South Korea and the fourth largest in the world, was subjected to a phishing attack on June 19, 2017. Personal data of some 30,000 customers were stolen due to the personal computer of an employee being compromised. The attack was reported to the authorities on June 29 and the news emerged on Reddit and Twitter on the 5th of July 2017. Bithumb issued compensation worth of 100,000 won ($86.83) to each user whose data had fallen to hackers.

4.1.3 Cyber Attack Dataset

Table 4.1: Activties, repositories and users for different event epochs

	GitHub							
	Contribution	Popularity	Repos	Users	Contribution	Popularity	Repos	Users
	Attack on Bitfinex in 2016				Attack on Bithumb in 2017			
Pre	17	74	28	64	24	236	65	197
During	19	75	30	70	14	190	49	170
Post	29	77	32	66	38	225	52	207
	Reddit							
	Posts	Comments	Users		Posts	Comments	Users	
	Attack on Bitfinex in 2016				Attack on Bithumb in 2017			
Pre	23	396	267		0	88	64	
During	41	690	424		5	114	84	
Post	16	366	251		5	126	91	
	Twitter							
	Attack on Bitfinex in 2016				Attack on Bithumb in 2017			
	Tweets	Retweets	Users		Tweets	Retweets	Users	
Pre	1,022	491	929		346	301	462	
During	4,575	2,088	4,669		1,824	1,530	2,246	
Post	1,341	457	1,021		1,128	814	1,321	

The study in Chapter 5 is based on data from three platforms, GitHub, Twitter and Reddit. GitHub and Reddit data are publicly available on their respective websites, while Twitter data was collected using the Twitter API.

To focus on the data representative for the study, two sets of keywords were used. One contains keywords related to cyber security issues, such as *attack, cybersecurity, ddos, encryption, hack, malware, phishing, ransomware, security, vulnerability*. For the Twitter dataset, the tweets that include any one of these keywords are selected, then a second set of keywords that focuses on Bitcoin-related attacks: *bitcoin* and *btc*, is applied. For Reddit, posts and comments in subreddits related to crypto-currencies: *Bitcoin, ethereum, Monero, Lisk, pivx, DopeCoin, paycon, orocoin, Donationcoin* are first selected and then the filtration based on the two lists of keywords is carried out in the same order as above. For GitHub, the repositories of interest are obtained by filtering the repository descriptions using the crypto currency keyword list (i.e., *bitcoin and btc*). The events related to this subset of repositories were further filtered based on the event description texts using the first list of keywords.

Chapter 5 analyzes three 1-day periods around each cyber attack event: the day before the event, the event day, and the day after the event. All the three days are also analyzed together. Thus, for Bitfinex, which took place on August 3, 2016, the analysis is focused on the three days from August 2 to August 4, 2016. For Bithumb, it is from July 4 to July 6, 2017.

Table 4.1 presents the size of data. The total counts of all types of activities in the platforms are divided in three different epochs for each event. 'Repos' represent the number of unique GitHub repositories involved. 'Users' represent the number of unique users involved in all the platforms.

Further, the events are divided into different categories in GitHub, *contribution* events and *popularity* events which are explained in the Table 4.2 below, Twitter events are of two types: *tweets* and *retweets*, and Reddit events are: *posts* and *comments*.

4.2 Information Cascades of Common Vulnerabilities and Exposures

Common Vulnerabilities and Exposures (CVE) are used to refer to the security vulnerabilities in various libraries and software. CVE mentions can be found in social media platforms like Reddit and Twitter discussing software security vulnerabilities [33, 34]. We explain the extraction of Twitter and Reddit messages in the subsequent subsections; the cascade recreation processes are explained before that.

Table 4.2: Categories of GitHub events

Popularity Events	
WatchEvent	Raised when an user starts watching a repo
ForkEvent	Creating an independent copy of a repository
Contribution Events	
CreateEvent	When a repository is created
PushEvent	When code commits are pushed to a repository
CommitCommentEvent	When a comment is made, edited or deleted on a code commit
PullRequestEvent	When a code review and merge request is raised on a repo
PullRequestReviewCommentEvent	When a comment is made, edited or deleted on a pull request
IssuesEvent	When an issue is opened or closed for a repository
IssueCommentEvent	When a comment on an issue is made, edited or deleted
ReleaseEvent	When a release for a repository is published

Cascade can be defined as a tree of messages where the root will be the message that sparks off an discussion. For Reddit, a cascade will start from a post and then grow into a tree of comments, where comments could be replied with further comments and so on. For Twitter, cascade can be seen as the re-tweets of a tweet. A tweet can be retweeted by someone, and can be seen by the followers who can then go on to retweet it further thus propagating a tweet in a recursive manner. This can be represented as tree of users propagating the same messages. The root of this tree will be a tweet and then the children will be retweets and their children will be the retweets of the retweets.

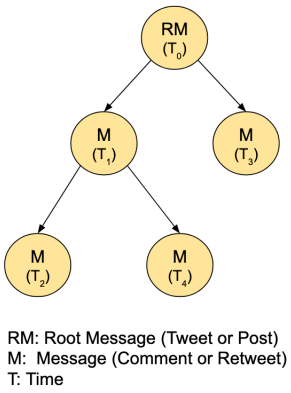


Figure 4.1: Example of a cascade

Figure 4.1 shows an example of a cascade tree. The root of a cascade is a tweet in case of Twitter and a post for Reddit, the message nodes are retweets for Twitter and comments for Reddit. One cascade could

belong to one or more CVE depending on the appearance of CVE keywords in the post or comments, and tweets or retweets which are part of the cascade.

4.2.1 Reddit Cascades Recreation

For the study in Chapter 6, Reddit cascades are created in a top down approach. We start by reading a post which will be the root of a cascade, then the comments whose parent id matches the post are added as children to the post (root). Then for each comment the comments whose parent id match the comment are added as children to that comment. This is a regular breadth first search approach, where each message is treated as a parent and its children are selected and added as nodes.

4.2.2 Twitter Cascades Recreation

Twitter cascades, on the other hand, are not easy to recreate because of the nature of the Twitter API. From the Twitter API only the original tweet, that is, the top level tweet, can be found from the retweets. Thus, we use an approximation method to attach retweets to older retweets by using a follower network. It is assumed that a tweet is visible to the followers of the tweeting person, and the followers retweet to make the tweet visible to their own followers. Thus, the time of retweeting and the follower network are used to create the Twitter cascades. We use PNNL's source code [1] to create these cascades which is based on the work of Vosoughi et. al. [35]

4.2.3 CVE Cascades Dataset

The data for both the platforms, Twitter and Reddit, have been collected for a period of 17 months, from March 2016 to August 2017. The dataset contains tweets and retweets containing any CVE keyword. CVEs have a predefined structure (like: CVE-2002-1337), the tweets and retweets containing such terms are filtered for the analysis. Then the cascades are recreated from the data. For Reddit, the posts containing CVE terms are filtered and then all comments for those posts are also extracted. Similarly, if a comment contains one or more CVE terms, then the post to which the comment was made and all other comments

[1] https://github.com/pnnl/socialsim/tree/master/december-measurements/cascade_reconstruction

belonging to that post are also filtered for the analysis. The basic characteristics of the data are shown in Table 4.3.

Table 4.3: Basic data characteristics for Reddit and Twitter CVE data

Twitter			
Users	Tweets	Retweets	CVEs
8,882	51,777	28,543	11,401

Reddit				
Users	Posts	Comments	CVEs	Subreddits
65,080	3,815	163,731	3,010	490

Table 4.4 shows the number of unique CVEs and total trees (or cascades) in both the platforms. The common CVEs between the two platforms are extracted and trees belonging to these CVEs from both the platforms are the ones to be linked. Table 4.5 shows the common CVEs and the corresponding trees on both platforms. The trees pertaining to these common CVE references are the ones which are sought to be linked.

Table 4.4: CVE counts and tree counts on both platforms

	Unique CVEs	Trees
Twitter	5,211	10,259
Reddit	1,300	2,333

Table 4.5: Common CVEs and tree counts

Common CVEs	421
Reddit Trees	1,128
Twitter Trees	2,252

4.3 Tools for Handling Data

We used various tools to store and curate the data for different needs. Twitter messages, and Reddit messages were converted from huge JSON strings to smaller JSON strings using Go, a general purpose compiled language. GitHub events were stored in ElasticSearch, a document indexing database known for handling big corpora of data, as JSON documents. GitHub events were queried from ElasticSearch for the required dates using Python, and stored as files containing JSON strings. Python has been used heavily to read these trimmed down JSON files into Pandas dataframes. The Pandas library in Python has been used

to change the data into different timeseries of counts. The STE values are calculated for different timeseries using an open source library from GitHub [2].

The next two chapters are the case-studies explaining the use of the generic framework for finding cross-platform influence. Sections 4.1 and 4.2 explain the datasets used in the Chapters 5 and 6, respectively.

[2]https://github.com/mariogutierrezroig/smite

CHAPTER 5: CYBER ATTACKS ON BITCOIN EXCHANGES

This chapter presents an in-depth analysis on how information transfer can be determined using the framework proposed in this thesis. This study looks at influence dynamics between different platforms during an external event, and how different kinds of platform can be tackled using the proposed methodology. We explain the problem in the next section. The context of the selected real-world events is explained in the subsequent section. We find that the time window of influence changes for different epochs of an event. Moreover, we observe that certain types of activities are more influential than others for cross-platform impact. We discuss our observations in the last section of the chapter.

5.1 Problem

While transfer entropy has been used to measure the influence between events, we identified problems with the approaches. First, the selection of the time window for sampling the activities from different platforms with different activity rates is not obvious. Second, it is unclear how this time window should vary over the duration of an event.

This chapter presents an empirical study that involves three highly different platforms: GitHub, Reddit, and Twitter. The challenges of measuring the transfer entropy between these platforms is mainly due to different activity rates. Users in GitHub make programming-related contributions via actions on software repositories, such as IssueComment, Push/Pull, Create/Delete [11]. Users on Reddit engage in subject-focused conversations via comments to original posts or to comments of other users. At the other extreme, in Twitter users broadcast short messages with a push of a button.

During an external event these three social media platforms act differently: there is more traffic and, information is consumed and disseminated at a higher rate. This should affect the time window of information transfer between different social media platforms. In this chapter we look at how the intrinsic time windows

of influence change during three different epochs of an event. These epochs are the day before an event, the day of the event and the day after the event. We also look at the influence dynamics for all the 3 days together.

To ground this empirical analysis, we analyze two cyber attacks on two different cryptocurrency exchanges. The first is the hacking of an exchange called Bitfinex, and the second is a phishing attack on another exchange called Bithumb. Different user activities in these platforms (i.e., GitHub, Twitter and Reddit) are analyzed considering different temporal intervals. The dataset used for the study is described in Section 4.1.

5.2 Measuring Cross-platform Influence

As explained in Section 4.1, to measure the information transfer between platforms using STE this study considers two types of activities in each platform, namely: *contribution and popularity* events in GitHub, *tweet and retweet* events in Twitter, and *post and comment* events in Reddit. Further, we are interested in measuring the pre-event, during-event and post-event impact on information transfer across platforms. Before describing our findings in the following section, we emphasize that different sampling intervals or time scales were considered to generate time series information in this study. The choice of the candidate window sizes depends on the study's context. However, a precautionary measure was taken to select the minimum time scale such that the corresponding time series is not sparse (containing mostly 0). Simultaneously, to consider the maximum value of the time scale we need to define a threshold for the length of the time series to make meaningful interpretations of the STE values. With $m = 3$, the minimum length of a time series to calculate STE should be $N > 30$ (see Section 3.2)

Considering 1-day duration of pre-event, during-event and post-event arrangements in this study, the defined maximum time scale is 30 minutes, which will result in $N = 48$. The minimum time scale is chosen as 1 minute considering the temporal activity counts in each platform. Thus, considering 1-day event duration (pre, during, post) the choice of time scales in this study are, 1, 5, 10, 20, 25, 30 minutes. In addition, a time scale of 60 minutes was also considered in case of whole three days of event duration. Using the aforementioned time scales, time series for all activities across platforms were constructed. STE

was calculated by considering every possible combinations of these activities and the corresponding STE values calculated using the random timeseries were also collected. Then the p-value was obtained using the statistical significance test, as explained in Chapter 3. The detailed values of the STEs that passed the p-value test are listed in the Tables 5.1 and 5.2.

For the sake of brevity, E_{Bx} and E_{Bb} refer to the Bitfinex and Bithumb events, respectively. The symbol E^T_{Bx} will denote the whole duration of three days where $E^T_{Bx} = E^{t-1}_{Bx} \cup E^t_{Bx} \cup E^{t+1}_{Bx}$. Similarly, the symbol E^T_{Bb} will denote the whole duration of three days for the Bithumb event. E^{t-1} will denote the day before the event, E^t will denote the event day, and E^{t+1} will denote the post event day for both events.

5.3 Empirical Results

For result presentation, we group the platform events into two categories: *primary* and *secondary*. For GitHub, the primary events are contribution events, denoted by G_p. The secondary events are popularity events (i.e., Watch and Fork events), denoted by G_s. Similarly, in Twitter, tweets are primary events (T_p), and retweets are secondary (T_s). Finally, in Reddit posts constitute the primary events (R_p) and comments are secondary (R_s). Intuitively, the primary events are the ones which engender a discussion or diffusion, whereas the secondary events are in response to the primary events.

Considering GitHub, Reddit, and Twitter, there are 6 possible combinations of influence between these platforms: GitHub actvities influencing Reddit activities ($G \rightarrow R$), GitHub activities influencing Twitter activities ($G \rightarrow T$), Reddit activities influencing GitHub activities ($R \rightarrow G$), Reddit activties influencing Twitter activities ($R \rightarrow T$), Twitter activities influencing GitHub activities ($T \rightarrow G$), and Twitter activities influencing Reddit activities ($T \rightarrow R$).

Figure 5.1 presents the STE values for two different platform combinations, Twitter activities influencing GitHub activities ($T \rightarrow G$), and Twitter activities influencing Reddit activities ($T \rightarrow R$) before the day of the Bithumb event (E^{t-1}_{Bb}). The plot shows that tweets (T_p) influence Reddit comments (R_s) with an STE of 1.29 and the fitting time window is 20 minutes. This is the highest STE for the combination of Twitter activities influencing Reddit activities. Similarly, the highest STE for ($T_p \rightarrow G_s$) is 1.09 for a fitting time window

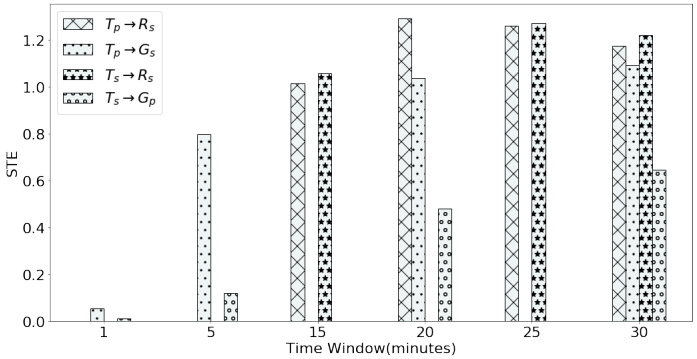
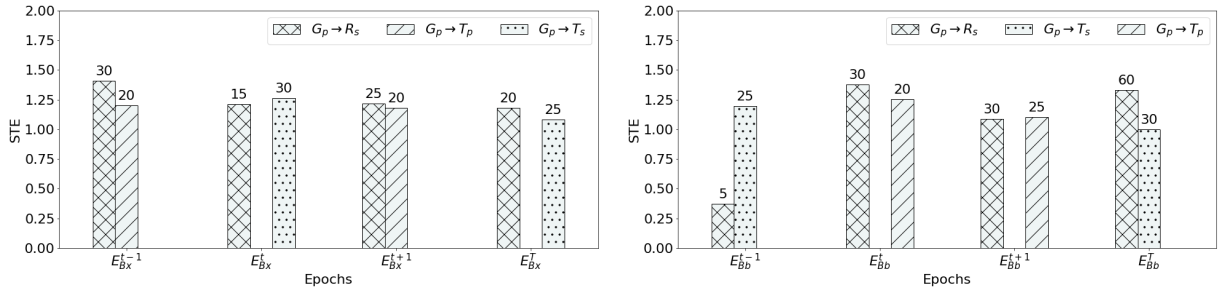


Figure 5.1: Example for finding platform influence by STE values. This is Twitter's influence on Reddit and GitHub as represented by STE values measured over different time windows on the day before the Bithumb event (E_{Bb}^{t-1}).

of 30 minutes. The statistically insignificant STE values are not reported, thus there are missing bars in Figures 5.1, 5.2, 5.3, and 5.4.



(a) E_{Bx} (b) E_{Bb}

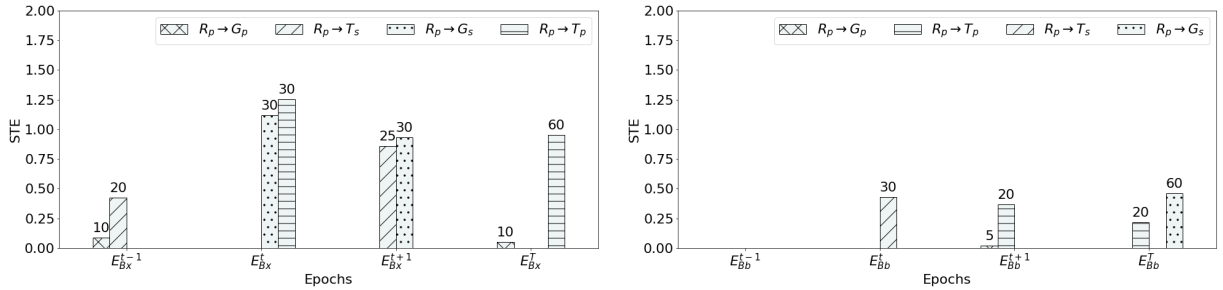
Figure 5.2: GitHub's influence on Twitter and Reddit. Influence is measured by the largest symbolic transfer entropy value (STE) obtained for different time windows. Top figure presents the Bitfinex event and the bottom figure presents the Bithumb event. The different epochs of each event are presented on the x-axis. The labels on bars are the time windows in minutes that led to the highest STE value.

Figures 5.2, 5.3, and 5.4 present the highest STE values for platform combinations. The numbers on the bars represent the time windows that led to the highest STE value. From these figures, we draw the following observations.

First, GitHub contributory events (G_p) influence Reddit commenting activities (R_s) for both Bitfinex and Bithumb throughout all the epochs (Figure 5.2). However, the influence of GitHub to Twitter is manifested on both primary (tweets) and secondary (retweets) activity: for example, while STE for $G_p \rightarrow R_p$ is not sta-

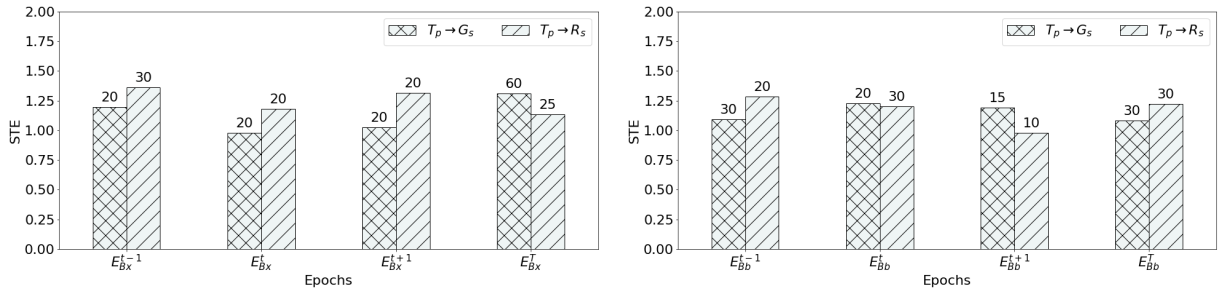
tistically significant, the transfer entropy for $G_p \rightarrow T_p$ and $G_p \rightarrow T_s$ are close in value (about 1.2). Moreover, the influence of GitHub manifests on different activity types at different epochs of an events: e.g., for the Bitfinex event, GitHub influences tweeting activity before the event but it influences retweeting during the event day.

Second, Reddit does not have much influence over the other two platforms. Figure 5.3 shows a lower STE value compared to the STE values in Figures 5.2 and 5.4.



(a) E_{Bx} (b) E_{Bb}

Figure 5.3: Reddit's influence on Twitter and GitHub. Influence is measured by the largest symbolic transfer entropy value (STE) obtained for different time windows. Top figure presents the Bitfinex event and the bottom figure presents the Bithumb event. The different epochs of each event are presented on the x-axis. The labels on bars are the time windows in minutes that led to the highest STE value.



(a) E_{Bx} (b) E_{Bb}

Figure 5.4: Twitter's influence on GitHub and Reddit. Influence is measured by the largest symbolic transfer entropy value (STE) obtained for different time windows. Top figure presents the Bitfinex event and the bottom figure presents the Bithumb event. The different epochs of each event are presented on the x-axis. The labels on bars are the time windows in minutes that led to the highest STE value.

Third, Twitter only influences the secondary activities in the other platforms. Figure 5.4 shows that tweeting activity in Twitter influences GitHub Watch and Fork events, but not the contributory events. That

is, tweets do not often result into software contributions. Similarly, tweeting activity seems to influence commenting activity on Reddit rather than posts.

Fourth, different epochs have different fitting time windows of influence. For example, in Figure 5.4(b) the influence of tweets on GitHub popularity activities is best measured for the time window of 30 minutes before the Bithumb event, 20 minutes during the event, and 15 minutes after the event. Moreover, different events have different fitting time windows for the same epoch. Figure 5.4 shows the fitting time window for influence of tweets on GitHub popularity activities over the 3-day evaluation is 60 minutes for Bitfinex, and 30 minutes for Bithumb. The same could be seen for all combinations except one (the day of the event for Twitter to GitHub).

Finally, an important observation of these experiments is that the influencing events on all platforms are the primary events (G_p, R_p, or T_p). This can be observed clearly from the Tables 5.1 and 5.2. This observation suggests that the cross-platform influence is driven by the primary events of a platform, rather than by the secondary events. That is, contributory activities of GitHub will influence Reddit and Twitter more than the GitHub popularity activities do. Similarly, Twitter tweets will have more impact on Reddit and GitHub activities than the retweets have. An interpretation of this result is that while secondary events play an important role in disseminating information within the platform, the primary events are the ones generating influence outside the platform.

Table 5.1: Highest STE values for Bitfinex event passing the statistical significance test. The values are listed for each platform combination for all the three epochs and the 3 days together

E_{Bx}^{t-1}				
Direction	Time Window (min)	STE	p-value	t-statistic
$G_p \rightarrow R_s$	30	1.40	$4.6e^{-29}$	-16.03
$G_p \rightarrow T_p$	20	1.20	$9.22e^{-9}$	-3.77
$R_p \rightarrow G_p$	10	0.08	$4.95e^{-37}$	-17.33
$R_p \rightarrow T_s$	20	0.42	$8.51e^{-12}$	-5.67
$T_p \rightarrow G_s$	20	1.19	$8.51e^{-12}$	-4.55
$T_p \rightarrow R_s$	30	1.36	$4.80e^{-23}$	-15.03
E_{Bx}^{t}				
$G_p \rightarrow R_s$	15	1.21	$2.05e^{-41}$	-17.91
$G_p \rightarrow T_s$	30	1.26	3.68^{-19}	-13.35
$R_p \rightarrow G_s$	30	1.11	$2.08e^{-22}$	-11.23
$R_p \rightarrow T_p$	30	1.25	$5.86e^{-17}$	-10.26
$T_p \rightarrow G_s$	20	0.98	$4.11e^{-7}$	-3.88
$T_p \rightarrow R_s$	20	1.18	$1.18e^{-32}$	-17.02
E_{Bx}^{t+1}				
$G_p \rightarrow R_s$	25	1.21	$7.79e^{-7}$	-4.4
$G_p \rightarrow T_p$	20	1.18	$1.02e^{-16}$	-9.6
$R_p \rightarrow G_s$	30	0.93	0.02	-11.23
$R_p \rightarrow T_s$	25	0.85	$2.94e^{-35}$	-21.03
$T_p \rightarrow G_s$	20	1.02	$1.07e^{-17}$	-10.93
$T_p \rightarrow R_s$	20	1.31	$2.88e^{-27}$	-19.05
E_{Bx}				
$G_p \rightarrow R_s$	20	1.18	$1.98e^{-14}$	-12.04
$G_p \rightarrow T_s$	25	1.08	$6.58e^{-34}$	-16.04
$R_p \rightarrow G_p$	10	0.05	0.006	-1.55
$R_p \rightarrow T_p$	60	0.95	$8.97e^{-16}$	-10.56
$T_p \rightarrow G_s$	60	1.31	$7.63e^{-45}$	-22.77
$T_p \rightarrow R_s$	25	1.13	$4.01e^{-53}$	-30.99

Table 5.2: Highest STE values for Bithumb event passing the statistical significance test. The values are listed for each platform combination for all the three epochs and the 3 days together

E_{Bb}^{t-1}				
Direction	Time Window (min)	STE	p-value	t-statistic
$G_p \rightarrow R_s$	5	0.37	$1.7e^{-25}$	-17.76
$G_p \rightarrow T_s$	25	1.19	$1.04e^{-6}$	-6.25
$T_p \rightarrow G_s$	30	1.09	$3.17e^{-35}$	-20.52
$T_p \rightarrow R_s$	20	1.28	$1.28e^{-30}$	-14.91
E_{Bb}^{t}				
$G_p \rightarrow R_s$	30	1.37	$8.44e^{-39}$	-21.96
$G_p \rightarrow T_s$	20	1.25	7.48^{-23}	-12.03
$R_p \rightarrow T_s$	30	0.43	0.046	-3.99
$T_p \rightarrow G_s$	20	1.22	$5.15e^{-7}$	-5.88
$T_p \rightarrow R_s$	30	1.20	$6.19e^{-16}$	-15.89
E_{Bb}^{t+1}				
$G_p \rightarrow R_s$	30	1.08	$1.95e^{-21}$	-11.15
$G_p \rightarrow T_p$	25	1.10	$2.05e^{-19}$	-10.57
$R_p \rightarrow G_p$	5	0.02	$9.02e^{-9}$	-7.15
$R_p \rightarrow T_p$	20	0.37	$3.29e^{-20}$	-9.50
$T_p \rightarrow G_s$	15	1.19	$8.01e^{-22}$	-14.66
$T_p \rightarrow R_s$	10	0.97	$4.77e^{-5}$	-5.74
E_{Bb}				
$G_p \rightarrow R_s$	60	1.33	$2.20e^{-16}$	-14.66
$G_p \rightarrow T_s$	30	1.00	$8.72e^{-25}$	-13.26
$R_p \rightarrow G_s$	60	0.46	0.0004	-5.03
$R_p \rightarrow T_p$	20	0.21	$1.83e^{-8}$	-7.58
$T_p \rightarrow G_s$	30	1.08	$1.10e^{-27}$	-15.92
$T_p \rightarrow R_s$	30	1.22	$2.14e^{-49}$	-29.19

CHAPTER 6: LINKING CASCADES IN HETEROGENEOUS PLATFORMS

Information cascades are dynamical processes in social media platforms. An information cascade can describe the spreading dynamics of rumour, disease, memes, or marketing campaigns, which initially start from a node or a set of nodes in the network [36]. As we have discussed that the information is not contained within a single social media platform, it can be assumed that cascades in one platform influence cascades in another. In this chapter we look at linking such cascades from two different social media platforms, Reddit and Twitter.

6.1 Problem

Information cascades could be found on different social media platform, where a cascade on one platform can be linked to one on another. This can have applications in predictive algorithms to predict cascades in one platform using another, or diffusion of information across platforms in form of cascades. The challenge in this task is to connect one cascade with another from a different platform to build cross-platform cascades. The framework proposed in this thesis could be used to bind links between cascades from different social media platforms by using the optimum time window of influence. The dataset used for this study is explained in detail in Section 4.2.

6.2 Linking Cascade Trees

As social media platforms are not isolated environments, an information cascade in one platform can engender a cascade in another platform. To connect a pair of trees, one tree from Twitter and one from Reddit, (T_t, T_r), we need to find a pair of nodes (N_t, N_r) which can link the two platforms, where N_t is a message in Twitter (tweet or retweet) and N_r is a message in Reddit (post or comment). This pair can be

determined by picking any pair from all the possible pairs randomly. Considering T_t has n nodes and T_r has m nodes, then there will be $n \cdot m$ pairs.

The proposed framework can be used to choose between these possible pairs. A simple formula can be applied to find the best possible pair in terms of STE. A pair whose time difference is closest to the optimum time window can be used as the candidate pair for connecting the trees. For instance, a tree from Twitter is to be connected to a tree from Reddit, then we use the following equation

$$\delta_i = (\left|(\left|N_t^i[timestamp] - N_r^i[timestamp]\right|) - \omega_{STE_{t\to r}}\right|)/\omega_{STE_{t\to r}} \tag{6.1}$$

In this equation, $\omega_{STE_{t\to r}}$ is the optimum time window obtained by applying the framework.

Thus, a pair with smallest δ is the candidate pair. In the process of connecting trees across platforms, the tree that appears first can be connected to trees which appear later in the other platform. δ will range from 0 to 1, where $\delta = 0$ means that the time difference between the two nodes exactly matches the optimum time window.

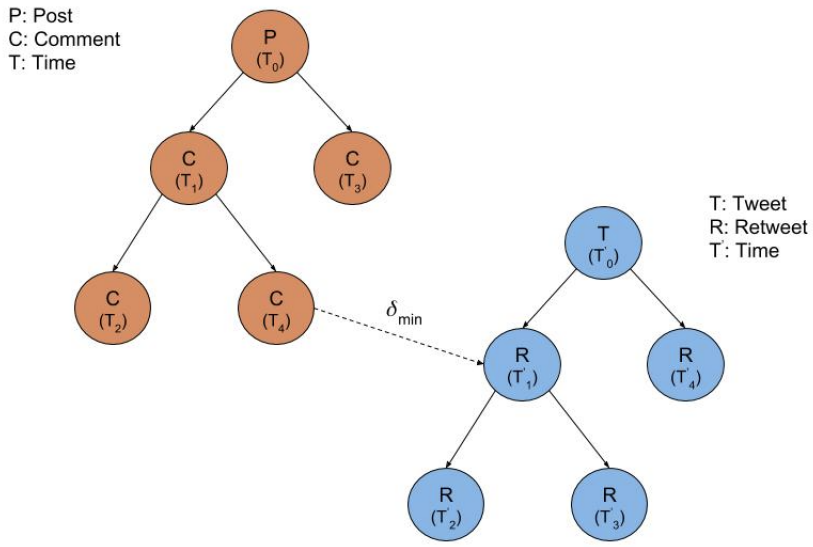


Figure 6.1: Linking trees

Figure 6.1 shows the linking of a Reddit tree to Twitter tree, here the assumption is that $T_0 < T_0^{'}$ i.e. the Reddit cascade starts before the Twitter cascade, δ_{min} is the smallest δ obtained using the Eq 6.1.

The first step in the experiment is to calculate the STE between the two platforms. Following STE values were calculated using the proposed framework: $STE_{R \to T}$, and $STE_{T \to R}$, where R, Eand T represents Reddit activities, and Twitter activities respectively. When connecting a tree from Twitter to Reddit the time window which has the highest value of $STE_{T \to R}$ is considered. Similarly, when connecting a Reddit tree to a Twitter tree we check for the time window which gives the highest value of $STE_{R \to T}$

To calculate the STE between the two platforms events are sampled by following time windows: *30 minutes, 1 hour, 2 hours, 4 hours, 8 hours, 12 hours, 24 hours, 36 hours, 48 hours, 56 hours, 60 hours, 72 hours, 84 hours, and 96 hours*. Again, we ensure that the time intervals are chosen such that the timeseries for the platforms are not mostly zeroes. We stop when we find no influence for both combinations. Then the proposed methodology is applied to find the time windows which pass the statistical significance test and the corresponding STE values.

6.3 Empirical Results

In Figure 6.2, we see that the fitting time window of influence is 84 hours. The STE values for each time window passing the statistical significance test are also presented in Tables 6.2.

To connect the trees between Reddit and Twitter, an 84-hour window is chosen for both scenarios: Twitter tree appearing first and Reddit tree appearing first. As can be noted from Table 4.5 there are 421 common CVEs and 1,128 trees from Reddit, and 2,252 trees from Twitter which can have a corresponding connection (as they are the trees with the common CVEs). Once a tree belonging to a CVE from one platform is connected to a tree (with same CVE) in the other, it cannot be connected to any other trees (with the same CVE) from the other platform. Thus, there could be trees which can get connected multiple times only if they belong to multiple CVE references.

As explained in Section 6.2, the linking node pair is defined by the smallest value from Eq 6.1, per CVE. We find that after calculating δ for each possible node pairs only 27% of the possible pairs have δ in the

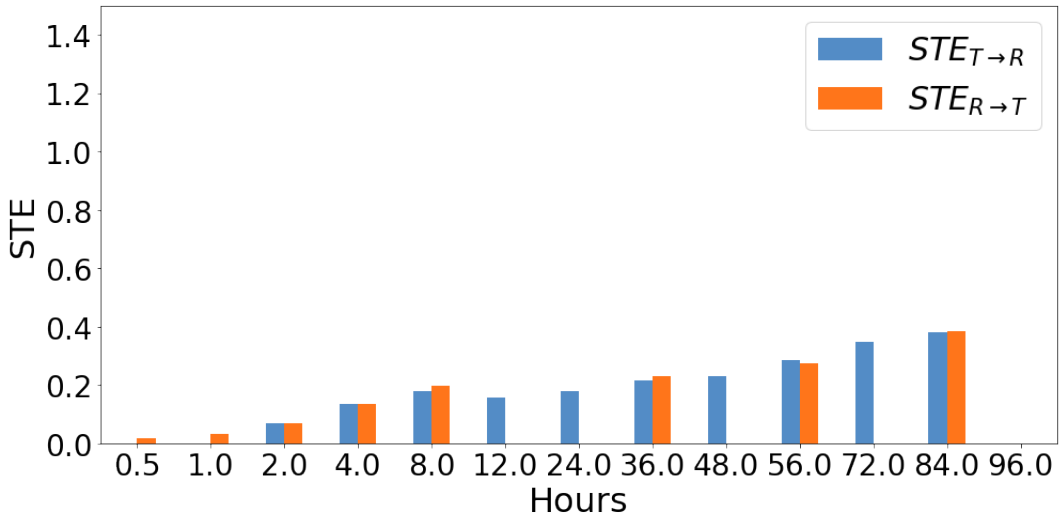


Figure 6.2: Statistically significant STE values between the platforms

range $[0, 1]$. Thus, the use of STE and the fitting time window of influence helps in narrowing down the search space for finding the linking node pairs.

The connecting node pairs are obtained from these 27% of possible δ values. The node pairs with the minimum δ value is considered for linking. The result obtained after connecting these trees is shown in Table 6.1. Using the Eq 6.1 there will be some nodes which will have a time gap different from 36 hours as all pairs falling in the range $[0, 1]$ are considered for pairing. Thus, there could be time gaps ranging from 0 to 72 hours.

Table 6.1: Cross-linked trees statistics

Total Linked Trees	395
Reddit First	155
Twitter First	240

Table 6.1 presents the total cross-platform trees that were obtained after the linking process. Out of the 395 trees, 155 trees have Reddit cascade appearing before the corresponding Twitter cascade, and 240 trees have Twitter cascade appearing before the corresponding Reddit cascade.

Figure 6.3 shows the distribution of time difference between the linking nodes (connecting node pairs) for each cross-platform tree. The first quartile of the distribution is 47.46 hours, the second quartile is 83.56

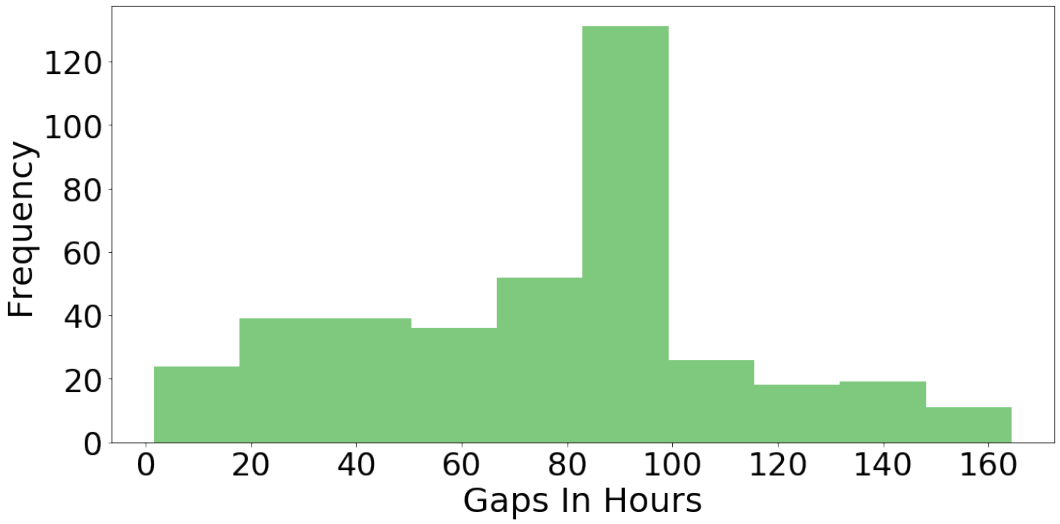


Figure 6.3: Distribution of time difference between linking nodes (node pairs)

hours, and the third quartile is 90.28 hours. It can be deduced that many of the linking node pairs for these 395 linked trees have many time difference close to 84 hours, that is δ, is close to 0.

This experiment provides a technique to connect cross-platform cascades using STE and statistical significance test to obtain the optimum time window of influence. During the empirical analysis we find that the fitting window of influence indeed helps in narrowing down the search space for node pairs which can be linked. In this particular scenario the search space is reduce by 73%.

Table 6.2: STE values for the time windows that pass the statistical significance test. These values represent the influence of Reddit activities on Twitter activities ($R \rightarrow T$), and Twitter activities on Reddit ($T \rightarrow R$)

$R \rightarrow T$			
Time Window(hours)	STE	p-value	t-statistic
0.5	0.017	$7.6e^{-26}$	-14.31
1	0.33	$7.2e^{-6}$	-4.74
2	0.068	$3.62e^{-26}$	-14.47
4	0.134	$6.66e^{-75}$	-53.54
8	0.198	$3.59e^{-75}$	-53.88
36	0.230	0.0002	-3.86
56	0.275	0.003	-3.02
84	0.385	0.015	-2.47
$T \rightarrow R$			
2	0.0686	$3.6e^{-17}$	-10.22
4	0.136	$7.2e^{-67}$	-44.07
8	0.179	$1.7e^{-61}$	-38.61
12	0.156	$3.27e^{-76}$	-32.98
24	0.177	$5.8e^{-42}$	-23.29
36	0.214	$1.06e^{-18}$	-10.92
48	0.232	$1.74e^{-14}$	-9.07
56	0.286	0.0012	-3.34
72	0.346	$1.007e^{-14}$	-9.09
84	0.381	0.002	-3.15

CHAPTER 7: CONCLUSION

This research was motivated by various interrelated questions. First, what is a useful time scale for measuring the information transfer between social media platforms with highly different activity patterns? Second, does a particular time scale work equally well for capturing information transfer before, during and after some event happens? Third, if we distinguish between different groups of activities on each platform, do the answers to the first two questions change? Fourth, how can we develop a method which is generic, independent of the platforms under scrutiny, and can be helpful in answering the above questions? And finally, can we develop a heuristic guided by this generic method to link cross-platform cascades? This thesis addresses these questions first by proposing a generic framework and then empirically by conducting two different studies.

In the first study, we focus our data-driven analysis on platforms with highly different activity rates. In GitHub, users contribute code or follow the evolution of software repositories, and their (time-involved) activities are often informed by the discovery of bugs or security vulnerabilities. Such bugs and security vulnerabilities are often broadcasted on Twitter and discussed in detail on Reddit. While software contributions may take long to submit, tweeting and retweeting take little skill, time and effort. Reddit conversations, meanwhile, are often involved, leading to long debates that, if described in terms of information cascades, are both wide and deep. For our analysis we isolated two cyber security events related to the Bitcoin cryptocurrency. The first event, Bitfinex, was a vulnerability breach that led to theft of money. The second event, Bithumb, was a privacy breach that led to the exposure of clients personal data. Our results focused on 3-day intervals for each event on each platform. The main message of our work is that the choice of time scale matters both for identifying the information transfer between platforms and for characterizing the transfer influence before, during, and after an event. We also found that the primary activities within the platforms wield cross-platform influence.

In the second study, information cascades from two different platforms are sought to be linked. The problem is solved by finding the fitting time window of influence between the the two platforms, Reddit and Twitter. Then the node pairs with a time difference closest to the fitting time window are linked to form the cross-linked trees. This study is an exercise on how to use the framework, and to show how it can be applied to tackle problems where cross-platform influence is involved. This case-study posits a heuristic to attach cross-platforms cascades, which may not be an fitting way to link these cascades. Moreover, the STE values calculated between Twitter activities and Reddit activities are not specific to CVE keywords. Thus, there could be possibilities of different fitting windows of influence for different CVE references. But such large number of CVE references (421 CVEs) would have led to an impractical amount of STE values to be considered if the STE values were calculated for each CVE separately. Furthermore, there would have been very less activities in the platforms for some CVE references. Here, a question worth asking is whether the heuristic proposed in this study is an optimal way of linking cascades across platforms. This question would warrant future work.

Future work involves developing techniques for real-time data and dynamic generation of STE between platforms, and optimization techniques for various parameters involved in the calculation of STE. Current work optimizes finds the most fitting time window from the choices made, the work could be pursued further to find the optimal time window. It would be interesting to apply the proposed framework in predictive tasks using machine learning techniques.

REFERENCES

[1] P. R. Center, "Social media use in 2018," http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/, [Online; accessed 17-February-2019].

[2] R. Syed, M. Rahafrooz, and J. M. Keisler, "What it takes to get retweeted: An analysis of software vulnerability messages," *Computers in Human Behavior*, vol. 80, pp. 207–215, 2018.

[3] S. Mittal, P. K. Das, V. Mulwad, A. Joshi, and T. Finin, "Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities," in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 2016, pp. 860–867.

[4] G. E. Hine, J. Onaolapo, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn, "Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web," in *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, 2017, pp. 92–101. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670

[5] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[6] M. Kim, D. Newth, and P. Christen, "Macro-level information transfer in social media: Reflections of crowd phenomena," *Neurocomputing*, vol. 172, pp. 84–99, 2016.

[7] A. Sumi, K.-i. Kamo, N. Ohtomo, K. Mise, and N. Kobayashi, "Time series analysis of incidence data of influenza in japan," *Journal of epidemiology*, vol. 21, no. 1, pp. 21–29, 2011.

[8] M. Garofalo, T. Nieus, P. Massobrio, and S. Martinoia, "Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks," *PloS one*, vol. 4, no. 8, p. e6482, 2009.

[9] J. Kim, G. Kim, S. An, Y.-K. Kwon, and S. Yoon, "Entropy-based analysis and bioinformatics-inspired integration of global economic information transfer," *PloS one*, vol. 8, no. 1, p. e51986, 2013.

[10] S. Uddin, N. Choudhury, S. M. Farhad, and M. T. Rahman, "The optimal window size for analysing longitudinal networks," *Scientific reports*, vol. 7, no. 1, p. 13389, 2017.

[11] G. Inc., "Github developer: Event types payloads," https://developer.github.com/v3/activity/events/types/, 2019, [Online; accessed 17-February-2019].

[12] G. Silvestri, J. Yang, A. Bozzon, and A. Tagarelli, "Linking accounts across social networks: the case of stackoverflow, github and twitter." in *KDWeb*, 2015, pp. 41–52.

[13] G. Haralabopoulos, I. Anagnostopoulos, and S. Zeadally, "Lifespan and propagation of information in on-line social networks: A case study based on reddit," *Journal of network and computer applications*, vol. 56, pp. 88–100, 2015.

[14] S. F. Waterloo, S. E. Baumgartner, J. Peter, and P. M. Valkenburg, "Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp," *new media & society*, vol. 20, no. 5, pp. 1813–1831, 2018.

[15] C. Shane-Simpson, A. Manago, N. Gaggi, and K. Gillespie-Lynch, "Why do college students prefer facebook, twitter, or instagram? site affordances, tensions between privacy and self-expression, and implications for social capital," *Computers in Human Behavior*, vol. 86, pp. 276–288, 2018.

[16] R. D. Putnam, "Bowling alone: America's declining social capital," in *Culture and politics*. Springer, 2000, pp. 223–234.

[17] B. Vasilescu, V. Filkov, and A. Serebrenik, "Stackoverflow and github: Associations between software development and crowdsourced knowledge," in *Social computing (SocialCom), 2013 international conference on*. IEEE, 2013, pp. 188–195.

[18] A. S. Badashian, A. Esteki, A. Gholipour, A. Hindle, and E. Stroulia, "Involvement, contribution and influence in github and stack overflow," in *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering*. IBM Corp., 2014, pp. 19–33.

[19] M. Pittman and B. Reich, "Social media and loneliness: Why an instagram picture may be worth more than a thousand twitter words," *Computers in Human Behavior*, vol. 62, pp. 155–167, 2016.

[20] A. Quan-Haase and A. L. Young, "Uses and gratifications of social media: A comparison of facebook and instant messaging," *Bulletin of Science, Technology & Society*, vol. 30, no. 5, pp. 350–361, 2010.

[21] Z. Papacharissi, "The virtual geographies of social networks: a comparative analysis of facebook, linkedin and asmallworld," *New media & society*, vol. 11, no. 1-2, pp. 199–220, 2009.

[22] E. Karapanos, P. Teixeira, and R. Gouveia, "Need fulfillment and experiences on social media: A case on facebook and whatsapp," *Computers in Human Behavior*, vol. 55, pp. 888–897, 2016.

[23] P. J. Boczkowski, M. Matassi, and E. Mitchelstein, "How young users deal with multiple platforms: The role of meaning-making in social media repertoires," *Journal of Computer-Mediated Communication*, vol. 23, no. 5, pp. 245–259, 2018.

[24] G. Ver Steeg and A. Galstyan, "Information transfer in social media," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 509–518.

[25] J. Borge-Holthoefer, N. Perra, B. Gonçalves, S. González-Bailón, A. Arenas, Y. Moreno, and A. Vespignani, "The dynamics of information-driven coordination phenomena: A transfer entropy analysis," *Science advances*, vol. 2, no. 4, p. e1501158, 2016.

[26] C. Wang, H. Yu, R. W. Grout, K.-L. Ma, and J. H. Chen, "Analyzing information transfer in time-varying multivariate data," in *Visualization Symposium (PacificVis), 2011 IEEE Pacific*. IEEE, 2011, pp. 99–106.

[27] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical review letters*, vol. 103, no. 23, p. 238701, 2009.

[28] D. Kugiumtzis, "Partial transfer entropy on rank vectors," *The European Physical Journal Special Topics*, vol. 222, no. 2, pp. 401–420, 2013.

[29] M. Staniek and K. Lehnertz, "Symbolic transfer entropy," *Physical Review Letters*, vol. 100, no. 15, p. 158101, 2008.

[30] M. Riedl, A. Müller, and N. Wessel, "Practical considerations of permutation entropy," *The European Physical Journal Special Topics*, vol. 222, no. 2, pp. 249–262, 2013.

[31] C. Y. Kim and K. Lee, "Risk management to cryptocurrency exchange and investors guidelines to prevent potential threats," in *2018 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 2018, pp. 1–6.

[32] F. Coppola, *Theft And Mayhem In The Bitcoin World*, 6 Aug 2016, https://www.forbes.com/sites/francescoppola/2016/08/06/theft-and-mayhem-in-the-bitcoin-world/#3c3c2d7b644f.

[33] C. Sabottke, O. Suciu, and T. Dumitraș, "Vulnerability disclosure in the age of social media: exploiting twitter for predicting real-world exploits," in *24th {USENIX} Security Symposium ({USENIX} Security 15)*, 2015, pp. 1041–1056.

[34] S. Mittal, A. Joshi, and T. Finin, "Thinking, fast and slow: Combining vector spaces and knowledge graphs," *arXiv preprint arXiv:1708.03310*, 2017.

[35] S. Vosoughi, M. Mohsenvand, and D. Roy, "Rumor gauge: predicting the veracity of rumors on twitter," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 11, no. 4, p. 50, 2017.

[36] M. Jalili and M. Perc, "Information cascades in complex networks," *Journal of Complex Networks*, vol. 5, no. 5, pp. 665–693, 2017.