Applications of Computational Methods in Political Science

by

Andrew Conway

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Politics
New York University
May, 2013

Advisor: Michael Laver

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Dedications

I would like to dedicate this to my wife Kristen, who without her unflinching committtment to this goal I would have never completed. And to our newborn daughter, Tatum Laursen Conway. The completion of writing this work represents the closing of a long and wonderful chapter in our lives, but also the beginning of a new chapter that we will never finish writing.

I love you both, and I look forward to embarking on this journey together.

Acknowledgments

I would like to thank all of the faculty and graduate students in the Department of Politics at NYU that have contributed to my research, education, and thinking over the course of time at the university. In addition, all of my friends and colleagues at the university, and beyond its city blocks – too numerous to enumerate – but nonetheless, contributed in so many ways to getting this research completed.

I also owe a great deal of gratitude to Kenneth Benoit, for both his considerable input on the crowd-sourcing project; but also, his financial support it getting the experiments deployed.

I would especially like to thank my dissertation committee for their tireless dedication to my research, and all of their effort in helping me make the most out of my research. This thanks goes to committee member Nathaniel Beck and Shankar Satyanath; and readers, Cyrus Samii, and Jennifer Larson.

Finally, I would like to highlight the contributions of my advisor, Michael Laver. Working with Mik over the past several years has been my fondest memory of graduate school. Without him none of this work would have been possible. His mentorship has been invaluable, and it is with my deepest sincerity and gratitude that I want to thank him for all of his effort in helping me achieve this accomplishment.

Mik took a chance on me, and I can only hope that the results have been as fruitful for him as they have been for me.

Abstract

The discipline of political science is undergoing a great methodological transformation. As is often the case, this transformation is being spurred by changes in technology. The application of computational methods to the study of political science has a long tradition, but only in the last few years has there been so much opportunity for disciplinary innovation at the confluence of core political science research problems and computational methods. The following thesis reviews three primary technologies that are rapidly changing the way political science research is being conducted, and through explicit experimentation attempts to highlight their value to the discipline.

The first chapter attempts to understand how the *ex ante* structure of social networks can influence how agents play collective action games. Traditionally, one might interrogate this line of research through fieldwork, or lab experiments; however, my approach is to build a computer simulation to test many more networks. These simulations show that these structures have a strong and meaningful effect on how agents play these games.

The second chapter seeks to understand how social networks change over time. To do this, I specify a new method for modeling this change based on graph motifs, and introduce software for generating networks this way. First, the model is tested against a set of classic generative

network models. Then, a time-series dataset of a large social network is collected to test how well the method can model change in a real-world network. The method performs well at generating both theoretical and real-world networks.

In the final chapter I address the classic political science problem of generating quantitative values from qualitative text data. The innovative technology introduced here is crowd-sourcing, in which a large pool of non-experts collectively contribute small amounts of work to a large coding project. In contrast to traditional methods of hand-coding, or automated machine coding, I show how crowd-sourcing is a viable – and in some ways – superior method for encoding text data.

It is critical that the discipline begin to engage with new computational methods now in order to further develop methodologically.

Contents

List of Figures

List of Tables

List of Appendices

Applications of computational Methods in Political Science

The popular narrative within much of political science is that the application of computational methods is a relatively new phenomenon. The fact is, however, computational methods have been apart of the discipline as long as computers have been readily available to researchers. Thirty years ago John Chamberlin used linear programming to approximate the solutions to social choice theories of voting manipulation (Chamberlin, 1982). Later in the same decade, theories of computational complexity were used to to test what mechanism could be implemented to protect the integrity of voting models (Bartholdi III et al., 1989).

Since then, there have been countless examples of computational methods being used to pursue questions of collective action and decision making. Simulations have been designed to test the empirical implications of Condorcet's voting paradox (Tangian, 2000). Likewise, computational models have been designed to test how classic political science theories, such as the spread of the "democratic peace" (Cederman, 2001), or the size and magnitude of wars (Cederman, 2003) can evolve over time and space.

Generative methods, such as agent-based models (ABM), have become very popular as a method for testing the emergence of many different kinds of observed behavior. ABM have been used to model various types of voting behaviors; including, the emergence of political party competition (Laver and Benoit, 2003; Laver and Sergenti, 2011); to test voting systems (Barkan et al., 2006); and, to analyze voter calculus (Kim et al., 2010). Generative models have also been developed to test the assumptions of experts forecasting the outcomes of international disputes (Sylvan et al., 2004), and when individuals participate in rebellion (Bhavnani and Ross, 2003).

Despite all of this research, however, there remains a conspicuous lack of computational methodology being applied and developed for political science; especially, when compared to the application and development of other techniques. Moreover, these methods are almost never taught as part of core graduate training, and where available they are relegated to special-edition seminars for advanced graduate students or faculty. Given their demonstrated value; why then, does there continue to be a lack of momentum for widespread adoption and training?

Early in the history of their application a shared belief emerged from the discipline that asserted computational methods were used as a substitute for well-specified formal models (Taber and Timpone, 1996). This narrative is typically assembled in the following way: computational methods are incorporated when the reseacher(s) ignore well-known theoretical solutions, or fail to adequately deliberate on the construction of a formal model to represent the phenomenon of interest. As such, the computational model is at best a temporary scaffolding easily torn down by an as-yet specified formal model – or at worst – the direct result of carelessness by the research in his or her review of the formal literature.

It is in large part because of these beliefs that computational methodology remains a sparsely populated island within political science. This belief, however, is the unfortunate result of a lack of understanding as to the purposeful motivation for the use of these tools by their critics; while at the same times, the non-systematic implementation of these methods by some entrepreneurial researchers. The development and application of computational methods are independent research lines within the discipline, and should be considered as such. Rather than replacements or substitutes, computational tools will continue to provide alternative means for addressing current problems of interest. In addition, these methods may well be better suited to address many future research areas of interest.

The application of computational methods to political science is the focus of this thesis. In the following chapters several different types of computational methods are used to address core research questions, and methodological challenges, to political science. The focus of this introduction is to begin by defining a broad set of tools that belong to computational methodology. The narrow focus on generative models deeply harms the discipline, and is unwarranted given the breadth of possible applications. In the following sections I discuss various methods that are a part of computational methods in political science.

Within each of these sections, the method being discussed is linked to the work of one of the these chapters. By way of specific political science problems, each chapter highlights how these methods can be implemented and applied. The thesis conclusion posits how this cumulative work reflects on the future application of computational methods in political science.

Generative Political Science

As was alluded to, for many political scientists "computational methods" is synonymous with "computational models", which itself is simply a more general term for "agent-based modeling". Given their reputation as an inferior replacement for an otherwise well-specified formal model, this association pulls down with it all of computational political science. First, agent-based models are an extremely powerful tool for social scientists, and their construction and execution has been studied widely in the social sciences (Epstein and Axtell, 1996; Epstein, 2006; Miller, 2007). They represent, however, only a fraction of the entire computational methodology toolkit.

Second, as with any tool, ABM are best used when the comparative advantage features of the tool match the designs of the experiment or research area. ABM, or generative models, are best used under two general cases: as a bridge to theory when observed phenomenon escape immediate modeling; or, when the reality of collecting observation proves very difficult, or impossible.

Theoretical models are often – and rightly – preferred because their results are static. As scientist we prefer a result that will not change with the addition of new variables, or reimplementation of a model. There are many phenomena, however, that are difficult or intractable, to solve analytically. In these cases a generative model can be developed to model the phenomenon, then observe behavior, and allow researchers to move closer to a well-specified model. Rather than the "all, or nothing" narrative often cited by those claiming computational models are careless replacements, ABM are extraordinarily useful instruments for moving research forward, and closer to a formal model if that is the final goal.

For example, in (Bianco et al., 2004), a generative model is specified to explore the possible sets of "enactable" outcomes in social choice situations. Upon reviewing the results of the model, the authors then specify a formal model based on learning gleaned from the generative model, and support it with a proof. Later, (Miller, 2007) extend this work by exploring the results of the generative model further to specify additional formal models of social choice situations. Here the generative model provides a bridge to theory, and produces additional theoretical observations in the process. This pattern is very valuable, and imminently repeatable.

In the second situation, where collecting observational data is very difficult or impossible, generative models provide stylized insight where none could be provided before. This type of modeling is standard operating procedure for many disciplines to which social scientists often look with great reverence; such as, physics, biology, chemistry, where the phenomenon of interest simply cannot be observed or instrumented. Instead, generative models and simulations are used to generate approximations of the phenomenon. The social sciences should be equally as welcoming of these methods, as many phenomena of interest are considerably more difficult to observe than those of our colleagues in the hard sciences.

In Chapter 1, "Networks, Collective Action, and State Formation," the focus is on how *ex ante* social structure imposed on agents affects their behavior when playing a stylized provision point public goods game. In this case, it is not possible to impose social structure on individuals naturally, and then observe behavior. As such the methodological choice for such work is between laboratory experiment or a generative model. In this case I chose a generative model because it allowed me to test a far greater number of agents, with many more network configurations, than would have been possible in a laboratory.

The work in Chapter 1 also deals with the study of complex networks. In the next section I discuss how the study of networks provides ample opportunity for the incorporation of computational methods.

Complex Networks

Throughout this thesis there are references to the increasing amount of political science research focusing on the study of complex networks. The length of Bibliography is a testament to the amount of research being done in this sub-field. Much of this work is either based on the statistical analysis of observed networks, or the advancement of statistical methods for analyzing static networks.

In both cases the network structure is often taken as given and static. The structure of real-world networks; particularly, those of interest to political science, have complex dynamic properties. It is extremely difficult to model the structural dynamics of complex networks because of the dependency structures present in these systems. Rather than relying on statistical models, computational models can be used to specify generative models of network dynamics.

In Chapter 1, "Modeling Network Structure Using Graph Motifs," a novel approach to modeling complex network dynamics is proposed. This approach relies on computational methods for counting the number of isomorphic components that exists in a given network for a set of graph motifs. Using the graph motif modeling approach described in the chapter, a series of simulations are conducted as a test of the model's ability to approximate the growth dynamics of real-world networks.

While simulation is central to the work of both Chapters 1 and 2 – and much of computational political science – there are many computational methods beyond simulation. In this final section I describe the use of crowd-sourcing as an alternative to the traditional laboratory.

Crowd-sourcing

A large portion of empirical political science relies on the use of data that has been generated via human interpretation. This often takes that form of some coding task, wherein a research assistant reads some text, or views some picture/video, and then categorizes or codes this item. This process has proved extremely valuable, as a vast amount of empirical work as been done using data generated this way.

Along with this value, however, come many costs. Both from a resources and time perspective; but also the reliability of the data generated. One of the primary benefits of incorporating computational methods into any research agenda is their low-cost and scalability, relatively to human labor. A significant negative of purely computational methods for data coding are the validity of codings. Thus, the challenge for innovative computational method of data coding is finding ways of using the tools that improve the scalability of human coders, without losing the benefits of reliability and interpretation that come from human intelligence.

In Chapter 3, “Methods for Collecting Large-scale Non-expert Text Coding,” I describe a series of experiments conducted to assess the viability of using crowd-sourcing as an alternative laboratory for conducting large-scale coding tasks such as this. In this case there is no “model,” but rather an exposition on methodology for using crowd-sourcing itself as a computational method. This serves as a single example of how computational methods, not simply models, can be applied to the study of political science.

Chapter 1

Networks, Collective Action, and State Formation

The notion of state building is a challenging concept to represent in concrete terms. It is most often thought of either relating to the levying of taxes (Besley, 2009; Besley and Persson, 2008) – particularly with respect to building a military – or through the ability of a state to successfully enforce contracts (Besley and Persson, 2010). Many authors have focused on these issues, and this scholarship has done well to show the importance of investment in various infrastructures to sustain markets. While valuable in their contribution to our general understanding of the role of state capacity in stable governance there are preliminary stages of the process that must also be considered.

Many models examining state capacity assume the state itself as a primitive element. That is, the implicit assumption is that a state upon which to build capacity already exists. Clearly this assumption is reasonable, however, it is important to consider that at some early point the collective decision to form a state had to be made. Part of this decision involves capacity, as there are varying levels to which individuals can contribute to the public good of the state; but more importantly, it involves the group dynamics among individuals and their influence over the collective act. As such, when considering how individuals collectively decide to establish a formal institution – such as a state – the role of informal institutions, e.g., social networks, is paramount. Abstractly, the germination of a state may be thought of as the transition of these informal institutions into formal ones; therefore, the mechanisms of this transition are of interest.

More generally, is the question of why do some efforts to make this transition succeed, while others fail? There are many examples of citizens emerging from relative destitution to establish and build state capacity. Likewise, there are examples of failure, where it seems states are incapable of collectively establishing and building a state. From the establishment of ancient Greek (Hornblower, 1992) and Roman republics (Lintott, 2003), to the reemergence of Japan and Germany after World War II and the post-colonial success of India's democracy (Khilnani, 1992), there is vast historical evidence for the ability of citizens collectively transition informal institutions to formal states, and build capacity. For many other countries, however, this transition has yet to succeed.

While there are far fewer examples of countries that have not successfully made this transition, the challenge of "failed states" remains paramount in both academic and policy research. Unfortunately, the mechanisms of this variance remain a puzzle. Perhaps the difference is third-party intervention or the presence of a "great power" in facilitating the transition? The conflicts in the former Yugoslavia and Somalia, however, provide contrary evidence. In both cases the United Nations initiated large-scale interventions, and while the success of these interventions can be disputed, what cannot is that nations of the former Yugoslavia have successfully transitioned, while Somalia has clearly now.

Alternatively, it has been proposed that variation in colonial history, and differences in institutional investment by colonial powers, explain the comparative levels of development (Acemoglu et al., 2001) among former colony states. While there is strong evidence to support this claim, the comparison begins with the assumption that colonial powers initiated the collective action to build a state; albeit with high variance in quality. This, however, does not explain the transition to formal institutions in the context of the history of countries like Afghanistan, which has a rich and mixed history of colonialism but currently languishes with a weak state bolstered by a foreign military.

Likewise, differences in ethnic diversity (Alesina, 2005), intrastate violence (Condra et al., 2010), or region show equally mixed outcomes. What has lacked exploration up to this point is how the structure of informal institutions, i.e., actors' social networks, contribute to these transitions. The initial conditions of informal institutions may play a critical role in the eventual success or failure of a burgeoning state. Social networks are a useful representation of informal institutions, in the context of state formation the initial conditions for social networks include the structural features and types of individuals within these networks. There is a dearth of micro-level data on social networks – particularly in places that lack formal institutions – making it very difficult to test the significance of networks in these circumstances.

The notion that individuals do not make political decision in a vacuum is well known, but only recently has social stricture been considered as a meaningful factor in the outcome of collective action (Siegel, 2009; Fowler, 2005). The interactions among individuals facilitated by networks can help actors form beliefs about the process of transition to formalized institutions, which forms the central focus of this research. In the following paper I consider the collective decision to form a state as a threshold public goods game. First, as as simple static model to establish a basic framework for the interactions, and then as network variant of the game. To begin, the role of informal networks in Afghanistan is discussed as an example of their influence in the process of state germination. Next, I introduce the basic provision point model, which is followed by a brief discussion of equilibria. Then, the network variant is introduced and a computational model is developed to test it. Computational modeling is used here as a means to incorporate additional assumptions about the behavior of the agents. Specifically, how signals transmitted over social networks affects the game's outcome. These network assumptions create higher-order complexity that is best modeled computationally. In the final sections the results of these simulations are presented, with a discussion and conclusions.

1.1 The role of networks in state building

In a 2010 white paper published by the Afghanistan Research and Evaluation Unit (AREU)[1], entitled "Afghanistan Livelihood Trajectories: Evidence from Badakhshan" the author describes the importance of informal institutions in Afghanistan (Pain, 2010):

> First, from a conceptual viewpoint, the key components of a country's institutional landscape can be characterized as the state, the market, the village and the household. However, Afghanistan has a complex and spatially variable institutional structure outside of the state and the market; this includes structures that may

[1] The Afghanistan Research and Evaluation Unit (AREU) is an independent research institute based in Kabul. http://www.areu.org.af/

> unite or divide villages such as ethnicity, tribe, *qawm* and *mantiqua*. A *qawm* is a form of solidarity that may be based on kinship, residence or occupation that can cross tribal and even ethnic boundaries; a *mantiqua* is a variable unit of social and territorial space that may unite people across villages. All these entities, to varying degrees and in different ways, establish rules and norms that regulate people's actions and moderate the workings and influence of other institutions. Much of the state-building effort in Afghanistan has focused on the architecture and performance of the state and its constituent parts and to a lesser degree on markets. But the nature and role of the multilayered institutions outside the state and market, and particularly those of the village and household, have been of less interest.

The above quotation indicates that the importance of informal institutions in Afghanistan may be equal or superior to that of formal institutions. While the focus here is on their role in Afghanistan, informal institutions are equally powerful in other parts of the developing world (Thies, 2009; Levitsky and Helmke, 2006). Given their prominence, the study of state germination should begin by considering the structure of informal institutions, and how these structures affect outcomes. As stated above, social networks are a natural representation of these institutions, and thus are useful focal point for an analysis. Social networks can have multiple institutional functions, such as simple neighbor relationships; wherein local information is exchanged, or as market relationships; wherein goods and services are exchanged via network ties.

In Afghanistan, social networks form the foundation of political and social life. In a different report compiled by AREU on the power of social networks in determining the political outcomes the author notes (Coburn, 2010):

> ...personal relationships are still the most important aspect of local politics. Individual ties with local elders and *maliks* are often based upon marriage, land use and business. As long as these ties continue to have economic and political import in the daily lives of voters, there is little incentive for these voters to approach other networks with which they do not have such intimate relationships.

By focusing on networks it is possible to investigate what factors in their structure are most influential in the transition from informal to formal institutions. Below, various network types are used to explore how social structure affects the level of contribution to a provision point public goods game. This game is presented as a model of state formation where players must decide how much of their own personal wealth to contribute to the state as function of their type and the information they receive from their network neighbors.

The decision to model the network variant of the game computationally is based on the desire to explicitly model actor networks and measure variation in outcomes based on these networks. While analytical results are always preferred to simulated estimates, the basic provision point model described below is very simple and lacks the assumptions required to model these relationships. The relationships described in the quotations above involve high-order complexity at varying levels of social interaction. The assumption that actor networks contribute directly to the outcome of the game, and that these can vary; therefore, requires a more flexible modeling framework.

In an attempt to more fully model this inherent complexity a computational model is developed. An effort is made to minimize the addition of unnecessary parameters and assumptions in the computational model in order to keep it as close to the analytical foundation that it is built upon.

1.2 Basic model of state building

In the following section I construct a basic threshold public goods; or more precisely, provision-point public goods game. The following game follows closely the form introduced by Palfrey and Rosenthal (Palfrey and Rosenthal, 1984). Unlike the Palfrey and Rosenthal game, but following the experimental work of Cadsby and Maynes (Cadsby and Maynes, 1999), this game incorporates a continuous – but bound – contribution choice for each player.

1. For some state with N citizens, each has an equal endowment of personal wealth denoted y.

2. Citizens can elect to contribute some proportion of their wealth to a public good, denoted c_i where $c \in [0, 1]$ where an individual's contribution is yc_i.

3. The public good will only be provided if common-knowledge w contributions are made, e.g., the threshold for public good provision is $w \geq m$ where $w = \sum_{i=1}^{N} c_i$. Also, the quality of the public good is increasing in the amount of wealth contributed by its citizens, but contributions that fails to meet the threshold are not "refunded".[2]

There are a few additional assumptions; first, the population wide shared cost of providing this public good is always less than the benefit gained from the good, or $\frac{w}{N} < 1 \leq \frac{m}{w}$. Also, assume that $y < m$, or that the contribution of a single citizen will not be sufficient to meet the threshold. From this game, the following payoff matrix for agent i is specified as a function of both c_i and m in Table 1.1.

[2] For a discussion of threshold public goods game using a refund mechanism, see (Marks and Croson, 1998; Rondeau et al., 1999)

	$c_i > 0$	$c_i = 0$
$w \geq m$	$\frac{m}{w} - yc_i$	$\frac{m}{w}$
$w < m$	$-yc_i$	0

Table 1.1: Payoff matrix for U_i given some individual level of contribution (c_i) and collective resources put toward public good (m)

In this case, the public good is represented by $\frac{m}{w}$, so as the amount of contributors increases beyond the threshold so does the shared benefit, and hence the provision-point specification. A citizen's decision, therefore, is whether or not to contribute, and if so how much. Given the duality of this choice, a citizen's utility is dependent first on whether to contribute to state building, which itself depends on whether the sum of contribution by all citizens was adequate to meet the provision point. Using the continuous contribution specification there are an infinite set of possible equilibria, however, below several are discussed.

1.2.1 Equilibria of basic game

As in Cadsby and Maynes, there are two straightforward symmetric pure-strategy Nash equilibria. The first is the complete defection equilibrium where no citizens contribute, or $c_i = 0 \forall i \in N$. Clearly, if a citizen believes that all others will attempt to free-ride and shirk contributing then that citizen will be made strictly worse off by contributing any $c_i > 0$, and thus complete defection is an equilibrium. Likewise, the second symmetric equilibria is the fully cooperative situation, wherein each citizen contributes exactly $c_i = \frac{w}{N}$. As before, if a citizen believes that all other players are contributing $\frac{w}{N}$, consider the decision of that citizen. The level of contributions is $\left(\frac{w}{N}\right)(N-1)$; therefore, that citizen's contribution is pivotal in providing the public good. By assumption $\frac{w}{N} < 1 \leq \frac{m}{w}$, and as such the pivotal citizen will be made strictly better off by contributing exactly what is necessary to meet the threshold.

Moving beyond the symmetric equilibria to the asymmetric where some set of citizens are not contributing while others are, with possible heterogeneity among contribution levels in the latter set. With the possibility of continuous choice there are an infinite number of asymmetric equilibria that meet the threshold constraint. Likewise, there are an infinite number of mixed-strategy equilibria wherein players randomize contributions. Each case, however, is a so-called knife-edge equilibrium, where some pivotal agent will only contribute given some supporting beliefs about the others players contribution levels and $c_i^* \leq \frac{m}{w}$. Put more simply, an agent will only rationally contribute under asymmetric conditions when the expected benefit from contributing is greater than the expected cost, which in this case is denoted with the previous inequality.

The motivation for this simple model is to understand the dynamics under which we might expect collective contributions to building a national defense. The above discussion has reiterated previous findings that such collective action is expected under a wide range of conditions. As discussed in the introduction, however, contemporary attempts at state formation have not enjoyed such success. Furthermore, the experimental results of provision-point models of public goods support the claim that over-provision is more likely than under (Epple and Romano, 1996). As such, it may be necessary to delve further into the model in order to understand the factors contributing to these sub-optimal outcomes.

1.2.2 Moving toward a more complex model of public goods provisions

The linchpin in this model is a citizen's belief about the contributions of other players in the game. As stated, there are an infinite number of equilibrium distributions of contribution among players, and therefore a rich set of outcomes for which the public good is provided. A key assumption of the model that allows for these equilibria to arise is that all citizens have uniform endowments. In practice, however, this is not true, with empirically observed distributions of wealth following something akin to a Pareto (Newman, 2005). In addition, while it may be true that citizens are aware of the necessary level of contribution needed to meet the threshold for a public good, it is not clear that they are aware of their own level of wealth relative to others, or the disposition of their neighbors to contribute. One way to model the means by which these beliefs are generated is to use the information transmitted over social networks to inform citizen prior to deciding if and how much to contribute.

Suppose that rather than existing in a vacuum, citizens exist on some plane wherein they can create connections to other citizens. These connections then allow for the transmission of information about both individual wealth and disposition to contribute to a public good. An extension of the above model to include these dynamics would allow information from the networks to alter a citizen's belief about the state of the world with respect to its network neighbors. Depending on the level of connectedness of a citizen, this information can be highly constructive in leading to efficient equilibrium, or highly detrimental.

In studying how these interactions affect outcomes, however, there is a necessary change in the analytical paradigm. The elegance of the closed form solutions proposed by Palfrey and Rosenthal; and Cadsby and Maynes provides deep insight with minimal complication by making broad simplification about agent interactions. As discussed above, however, these assumption are too limiting when extending the game's framework to include the transmission of information through social networks. One approach to extending the model would be to incorporate networks in terms of a cooperative game, and solve for equilibrium as such. This approach has been used extensively to model economic networks, and has illustrated how equilibrium network structures can arise in games with heterogenous agent utility functions (Jackson and van den Nouweland, 2005; Jackson, 2005 2008). The focus of this research, however, approaches the question in reverse; where the cooperative games framework focuses on how player types influence stable network structures, here the focus is on how both network structures and player types influence provision outcomes.

In fact, the equilibrium behavior observed in cooperative network games does not apply here, as this is a non-cooperative setting where the item of interest is the games' outcomes without appealing to equilibrium behavior. An alternative is to use generative models to estimate the effects of various network structures, and the interactions among agent within them on public goods provisions. As such, the networked variant described below does not rely on equilibria; instead, a set of agent and network types common in the network literature are used in a generative model to simulate a large repository of data for game outcomes given variation in these network types. From this repository the effect of different combinations of agent and network parameters can be measured.

1.3 Network variant of threshold public goods game

One of the primary contributions of social networks on a citizen's calculus to contribute to a public good is how the information exchanged alters their belief about the state of the world with respect to their neighbors.[3] In the case of the base threshold public goods model this information may change a citizen's belief about the necessary amount of contribution needed

[3] Here, the term 'neighbors' is used in the graph theoretic context, e.g., actors with whom an actor shares a direct tie.

to meet the threshold. As described above, an advantage of computational methods in this context is the ability to explicitly mode agent networks; therefore, it is necessary to define a citizen's network and the the nature of the information that is exchanged.

An agent i's network neighborhood is defined as the set of citizens with geodesic distance one from i, i.e., a direct tie. All ties are symmetric, meaning that out-degree implies reciprocal in-degree and vice-a-versa.[4] Formally, a citizen i has n neighbors where $n \subseteq N$, and $n = \{j_1, j_2, ..., j_n\}$. Note, agents only receive information from their immediate neighbors. While in practice people may update their beliefs via information relayed from others, the strength of localized information on both individual and collective decision making in social networks has been shown to be tremendously strong, and thus an assumption well supported in both theoretical and empirical studies of network influence (Granovetter, 1973; Christakis and Fowler, 2007).[5] The complete social network of all citizens is thus defined as the composition of all individual network neighborhoods. With these basic structural assumptions in place it is now possible for actors to ascertain two critical pieces of information via these relationships: neighbor wealth and disposition to contribute.

A fundamental aspect of social interaction is the building of an individuals's beliefs about their relative standing within a group (Kerckhoff, 1995). As shown in the discussion of informal networks in Afghanistan, this information can be critical when making political decisions, such as voting. Within the context of personal or family wealth these networks are often used as a means by which social stratification is assessed (Lin, 1999). Heterogeneity of wealth

[4] This model will be constructed as an undirected graph for simplicity; however, it is possible to implement this model as a directed graph form. Though, it is unclear that this additional assumption would add insight, as it is difficult to interpret the substantive meaning of a one-way social interactions.

[5] Extending the model to include global network information is non-trivial, and requires further assumptions about the degradation and absorption of information as it passes at different depths through a network. Though an abstraction, this localized model reduces the number of assumptions and allows for reasonable tractability.

among citizens in any given population is natural, therefore, this process of belief updating is reasonable. Then, it may be that as human social networks grow so too does a citizen's belief's about their relative standing. The veracity of a citizen's beliefs may in turn be a function of their structural position within a social network, i.e., "better connected" citizens have more informative beliefs about their economic status. In practice the revelation of this information is noisy; as it is impossible to every know with certainty ones social standing. For simplicity, however, in the networked model proposed here I assume a network tie provides a citizen with perfect information about their neighbors' endowments. A citizen, therefore, receives $y_j \forall j \in n$.

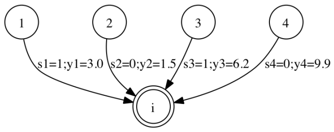
Further, consider how the game changes if along with economic information, prior to making their decision each agent receives some signal about the disposition of their neighbors to contribute. Again, borrowing the notation of Palfrey and Rosenthal, suppose that every neighbor of agent i provides a signal as to their dispositions, denoted $s_j \in \{0, 1\}$ where $s_i = 1$ indicates the intention to contribute. As before, this dynamic is meant to model the process of social communication, and its affect on individual decision making. While these signals are clearly cheap talk, what is worth exploring is how various distributions of these signals and their accompanying endowment affect decision making. To explore these dynamics it will be necessary to model the interpretation of any combination of signal and information.

Given that each citizen is fully informed about the endowments of their neighbors, define the total wealth within some citizen i's network as $Y_i = \left(\sum_{j=1}^{n} y_j\right) + y_i$. To use this information to inform beliefs, assume the strategy of a networked citizen is based on the beliefs about the contributions of their neighbors, or c_j; and thus, define $c_j = \left(\frac{y_j}{Y_i}\right) y_j$. The implicit assumption is that citizen i believes a neighbor will contribute an amount proportional to their own wealth vis-a-vis the neighborhood. It is now possible to rewrite networked beliefs, in terms of both s_j

and c_j, as noted below in Equation 1.

$$m_{net} = \sum_{j=1}^{n} s_j c_j \tag{1.1}$$

In this case, a citizen "takes their neighbor at his word", and plays a strategy that assigns a contribution of zero to those neighbors who signal they intend not to contribute and the network neighborhood proportion to those that do. Figure 1.1 below is provided as a numerical example how how networks affect beliefs.



(a) Agent i's network

If y_i=5.0, then $Y_i = 5.0 + 3.0 + 1.5 + 6.2 + 9.9 = 25.6$, which we use to calculate m_{net}...

$$\begin{aligned} m_{net} &= \left[(1)(3.0)\frac{3.0}{25.6}\right] + ... + \left[(0)(9.9)\frac{9.9}{25.6}\right] \\ &= 0.35 + 0.0 + 1.50 + 0 \\ &= 1.85 \end{aligned}$$

(b) Calculation of m_{net} for agent i

Figure 1.1: Numerical example of how network information affects beliefs

In the abstract this simple variant may provide little additional insight to how social networks affect public goods provision. In its application, however, it is possible to observe very different outcomes given different initial condition of the model, such as the structures of social networks and the distributions of wealth and disposition among citizens. Given the sensitivity of outcomes to initial conditions, however, it is extremely important to note that result from the computational model are not necessarily equilibrium based. Rather, by simulating a very large number of these initial conditions and comparing the results across a wide set of parameterizations it is possible to estimate the effect of different social structures and agent types on public goods provisions. In order to generate the data necessary to measure these effects, in the next section I present a generative computational model, where agents' networks are primitives of the game, which allows for a very thorough testing of possible initial conditions and subsequent outcomes.

1.4 Testing the Network Variant

To test the network variant of this model I have designed a simple computational experiment, which contains two basic objects: the agent, and the environment. Agent objects contain all of the information described in the previous sections; specifically, some random endowment of wealth drawn from Pareto distribution with the shape parameter $\alpha = 3.0$, and a random disposition to contribute drawn from $\sim U\{0, 1\}$.[6] Here I am also interested in the affect of networks on beliefs and public goods provision, agents also form "ties" with other agents, a topic I will discuss in more detail next. Finally, agents also have the ability to decide how much contribution to make to the public good.

As is often the case in multi-agent systems approaches, it is useful to have heterogeneity of types among agents (Laver and Benoit, 2003; Axelrod, 2006; Epstein, 2007). In doing so, the model is able to capture the higher order complexity of differing decision making criteria for actors, and also prevents the model from path dependent outcomes based on a single parameterization of actor type (Bonabeau, 2002). The agent object in this model can take on one of five types; with either a disposition to give to the public good, or not. The first is

[6] This particular parameterization of the Pareto is used as rough approximation of wealth distributions.

the *Altruistic* type, which always gives $\frac{w}{2}$ of its wealth to the public good, regardless of the information it is receiving from the network. This type is meant to model those individuals who are committed to contributing to a public good no matter the group dynamics. Next, is the *Community* type, which is equally giving but does so using information from their network. In this case, however, rather than always giving the same fraction of wealth, a Community agent will give whatever fraction of wealth is needed to supplement what they expect their neighbors to give and reach the threshold.[7]

The next two types set their contribution levels explicitly as a function of information from their network neighbors. The *Max-match* and *Min-match* set their contribution level to that of their most and least generous neighbors respectively. These agents are meant to model the social mimicry often observed in group dynamics – particularly with respect to charity (J.L., 2003). These types are the most consequential in the experiments; and become the focus of the analysis in the next section, as their contribution level is most closely related to their network strucure. Finally, the Miserly type always give some small random fraction of their wealth. Specially, the contribution is draw uniformly from $c_{Miserly} \in [0.0, 0.05]$. To summarize, Table 1.2 below describes each agent types' decision criterion for giving.

Agent Type	Decision Criterion
Altruistic	$c_i = 0.5$
Community	$c_i = m'$ such that $m' + m_{net} = w$
Min-match	$c_i = min(m_{net})$ for all neighbors of agent i
Max-match	$c_i = max(m_{net})$ for all neighbors of agent i
Miserly	$c_i \in [0.0, 0.05]$

Table 1.2: Agent types used in model, with contribution level decision criterion

The environment object is simply a container of agents, and is the abstract plane upon which the game is played. From a practical standpoint, this is the primary computational object of

[7] In the case where this amount does not reach the threshold a Community agent will give all of their wealth.

interest, as in each instance of the model an environment is filled with agents, wherein they set their contribution levels. Environment objects also generate all of relevant data for each play of the game. While a more technical description of the computational mechanics of the model is provided in Appendix A, it is worth describing at a high-level briefly.[8] The model is designed to create a simulated "state" populated by isolated individuals. In order to model how social networks affect an individual's decision to contribute to a public good, at its instantiation the model generates network ties among all the individuals in the model.

After this process of network formation is complete, the agents use their type- and network-dependent decision criteria to set their contribution levels. These contributions are summed, and if that sum is greater than or equal to the provision point the good is provided, otherwise it is not. By default, the threshold point for all of the experiments described below is $m = 0.25(Y)$, or $\frac{1}{4}$ of the total wealth of a population. The model, however, supports any parameterization of the provision point. In the following sections the implementation of the network experiments are described, followed by a discussion of the results.[9]

1.4.1 The Computational Model: Five Experiments

Technically, the model has been designed to accommodate any network formed by a countable vertex set[10], but for the purposes of this paper I have limited the experiments to five different types of network structures – each well known in the networks literature. In the first experiment the binomial random graph model was used generate ties (Erdos and Renyi, 1961). This is the

[8] The software designed for this research is freely available for inspection and download here: `http://github.com/drewconway/StateBuilding`. This repository includes the full model `Python` classes, as well as the `R` scripts used to analyze the results. All classes are **fully unit tested**, to ensure consistency and accuracy of output.

[9] For each of the experiments described below 75,000 agent observations are generated.

[10] A configuration model is used to generate network ties from any valid degree sequence for the number of actors being modeled. This method is a version of the configuration model described by Newman (Newman, 2003)

most studied of all random graph models, and thus is a useful starting point and benchmark for the remaining experiments. In this case, the purely random binomial model is used; the probability of a dyad forming between any two actors is exactly 0.5. With this configuration binomial random graphs have degree distributions with a “bell-shape” and exhibit high tie density. It should be noted that these types of networks are very rarely observed in real social networks, but again, given their well-known properties are useful benchmarks.

As a logical balance to the binomial network experiments, the next series generates networks where degree is drawn from a uniform distribution. That is, every actor in the network has an equal probability of having some some degree k such that $0 < k < N$ ties in the population. Similarly to binomial networks, networks with degree sequences drawn from a uniform distribution are very rarely observed in real social networks; however, networks with this property are occasionally observed in biological networks (Arita, 2005). In the context of this research, such networks may also be thought of as those existing in areas where local conditions, such as rugged terrain, prevent groups from forming ties based on preference, but rather on some external path dependence. This may be a good model for the types of country-wide networks observed in places like Afghanistan, where such constraints are present.

In the next experiment the Pareto distribution was used to generate networks with “small-world” properties (Watts and Strogatz, 1998).[11] This is the first of the experiments where the model has structural features observed in real social networks. The properties exhibit by these networks are short diameters and high degree of clustering – often around central actors. Likewise, the following experiments use a power-law distribution to generate networks with the so-called “scale-free” property, which are also very often observed in large complex networks

[11] In this experiments the Pareto exponent was 1.0, which provides the desired structural features.

(e.g., citation networks (Newman, 2001), sexual contact (Ergun, 2002) and the World Wide Web (Albert et al., 1999)).

In the final experiment a specific preferential attachment mechanism is used to generate the networks. Here, agents with high wealth parameters relative to the total wealth of the population have a higher probability of forming a tie. These networks will exhibit similar structural features to the power-law and Pareto networks, but rather than hubs forming randomly they do so as a function of a specific agent parameter – wealth. This experiment is particularly important as it models one natural mechanism by which a complex network of informal institutions might form. As a well studied mechanism for edge formation, preferential attachment has been observed in many networks of informal institutions, but is perhaps exemplified in the seminal work in this area by Padgett on the informal networks of the Medici family in Florence, Italy during the 15th century (Padgett and Ansell, 1993). This Table 1.3 below summarizes the network types used in the experiments.

Network Type	Structural Properties	Observed in Social Networks
Binomial	"Bell-shaped" degree dist.	No, a.k.a. Erdos-Renyi random graph
Uniform	Actor $\sim$ equal	No, occasionally in biological networks
Pareto	High clustering, short diameter	Yes, a.k.a "small-world" graph
Power-law	Degree dist. left-skewed	Yes, a.k.a. "scale-free" graph
Pref. attach	Similar to power-law	Yes, nodal attribute drives degree dist.

Table 1.3: Description of network types used in experiments

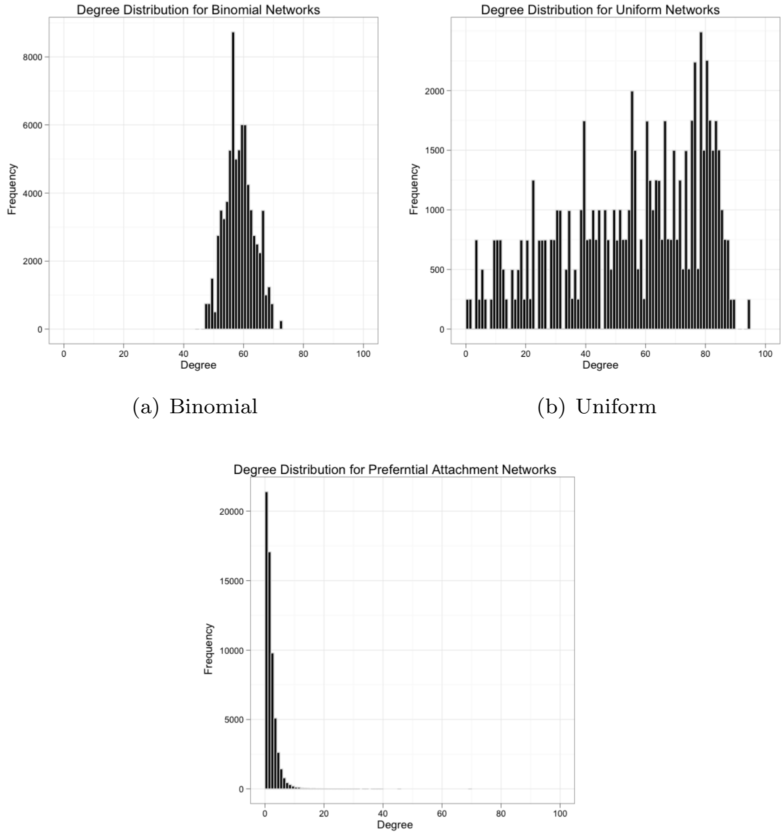
It is also important to note that these various network configuration were not chosen simply for their well known properties. Rather, each represents possible structural features present in local or tribal networks (save the binomial network) where information about public goods provisions may be transmitted via social networks. In the case of Afghanistan discussed in the introduction, the tribal networks could take any or of these forms, depending on the context of relationships. As mentioned, given the harsh terrain and long distances networks with uniform degree distribution might form; whereby, the actors in each tribe or location are densely connected to each other, but are not connected well outside of their immediate physical space. Likewise, a more common hub-spoke structure could be present, such that certain individuals or tribes are central to the network, acting as go-betweens to for other actors, while the majority of actors have limited connectivity. The experiments presented here are meant to provide evidence as to the effects of these structures on the outcomes of public goods provisions.

The first useful descriptive statistic to inspect from the experimental data is to check that the degree distributions of the models fit the expectations described in Table 1.3. In Figure 1.2 three of the overall degree distributions from the five experiments are provided.[12] As can be seen from this figure, the degree distributions from all of these networks matches the expectation. The binomial networks have an overall distribution with a rough bell-shape, with a modal degree of 60. The degree distribution generated by degree sequences drawn from a uniform distribution is a bit less consistent, though generally following expectation.[13] Finally, the preferentially attachment networks have the highly skewed degree distribution expected from this tie formation mechanism. The vast majority of agents have very low degree, with modal degree of just one, while in the tail agents have up to 70 ties.

Next, to understand how networks affect public goods provision it may be useful to examine how contribution levels varied across the experiments, and disaggregate this data by agent types

[12] The degree distributions for Pareto and power-law experiments are excluded because they follow almost identically to that of the preferential attachment networks. For considerations of space, the graphs for these networks have been omitted from all figures, however, high-resolution versions of all graphs generated from these data are available here: `http://github.com/drewconway/StateBuilding/tree/master/images/`

[13] The reason for this discrepancy is the algorithm used to generate the uniform graphs does so by creating tie sequences from a uniform distribution. As the number of agents here is fixed at 150, this adds random error into the uniformity, which then shows up in the degree distributions.



(a) Binomial

(b) Uniform

(c) Pref. Attach

Figure 1.2: Degree distribution for all network types

and whether the public good was provided (henceforth referred to as the provision point). In the next series of plots, data on the frequency of contribution levels for the three focal network experiments are provided. In Figure 3 below, the frequency of giving (binned at intervals of 0.1) is plotted for each of the five agent types (vertical), and split by the provision point (horizontal).

The data on Altruistic types is forgone, as they will either give nothing – when their disposition is such – or they will give exactly half. Likewise, Miserly types either give nothing or very little, while Community types give nothing or everything. The latter observation is interesting, as it indicates that no Community agent received information that its neighbors would contribute enough to the public good that would require them to give anything less than all of their wealth. More interesting, however, are the variations in contributions from the Min- and Max-match types across the different network configurations.

Starting with the binomial networks, we see that the density of the ties in these networks causes predictable behavior from these types. In the Min-match cases, the data shows that giving for these types is driven by neighbors with a disposition not to give, or Miserly neighbors giving only very little. For the Max-match types the outcomes are similar, however, the results are driven by neighbors unwilling to give, or Community type neighbors giving all of their wealth. Moving onto the Uniform networks similar giving is observed for Altruistic, Community and Miserly, types; but a very different pattern for the Matching types. Here, there is more variance among these types; however, each showing distinct trends. For the Min-match types the distribution still peak or 0 and 1, but between these values there is clear favoring toward lower contribution levels, which gradually flatten out around 0.5. For the Max-match the peaks are the same, however, the frequency of contribution between 0 and 1 are much more uniform. These patterns seem to suggest that when degree is uniform across actors the levels of giving are a result of “local” norms, which in this case are the minimum and maximum contribution amount with each of these network clusters.

Finally, the preferential attachment networks exhibit a completely different contribution pattern for the Matching types. Agents of this type, where wealthy agents have more connections, contributions are much higher for both types. In addition, note that this pattern is the same whether or not the provision point is met. From the perspective of inducing public goods provisions, this is very illuminating. That is, all things being equal, preferential attachment networks of this kind induce the highest level of giving from agents most likely to vary their contribution amount. These descriptive statistics, however, do not give clear evidence to the likelihood of meeting a provision point in any of these network types, or how agent types affect these outcomes.

Following from this analysis, therefore, I specify a basic probit regression model to measure the effect of these variables on the probability of meeting the provision point. In the data each observation includes a dummy variable indicating whether it was drawn from an instance of the model that met the provision point. This allows for the use of probit regression to measure the effect of all model variables across all network types. In Table 1.4 are the results of these models.

The most striking observation from this table is that all things being equal, i.e., holding all independent variables constant at zero, preferential attachment networks have the highest probability of successfully reaching the provision point, while power-law networks have the lowest. On the surface this may appear to be a contradiction, as these network types have roughly the same structural configuration (as explained in Table 1.2); however, where they diverge is the mechanism by which that structure is generated. In the preferential attachment networks wealthy agents have the most ties, and thus the highest amount of potential influence on the outcome of the simulation. If these wealthy and highly connected agents contribute, the likelihood of reaching the provision point greatly increase as a result of network effects. On the other hand, in the power-law networks these highly central actors become so at random, and thus their influence over the network is driven only by their type.

This observation can have significant impact when considered in the context of state building

Parameters	**Network Types**				
	Binomial	Uniform	Pareto	Power-law	Pref. Attach
(Intercept)	0.65^{***}	0.92^{***}	0.29^{***}	-0.72^{***}	1.40^{***}
	(0.07)	(0.04)	(0.03)	(0.03)	(0.04)
Contribution	0.12^{***}	0.12^{***}	0.17^{***}	0.15^{***}	0.17^{***}
	(0.00)	(0.02)	(0.02)	(0.02)	(0.02)
Wealth					
Neighbors	-0.00^{**}				
	(0.00)				
Threshold	-0.03^{***}	-0.05^{***}	-0.01^{***}	0.04^{***}	-0.06^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Disposition†					-0.04^{**}
					(0.01)
Miserly†					
Community†	-0.03^{*}		-0.05^{**}		
	(0.01)		(0.01)		
Min-match†		-0.03^{*}			
		(0.01)			
Max-match†			-0.04^{**}		
			(0.01)		
AIC	103442.10	103017.48	103422.46	103370.67	98140.69
BIC	103811.11	103386.49	103791.47	103739.65	98509.70
$\log L$	-51681.05	-51468.74	-51671.23	-51645.33	-49030.34

$N = 75000$ in all models

† indicates dummy variable, and agent type variables *labeled in italics*

Agent type *Altruistic* not defined because of singularities, and thus are excluded

Standard errors in parentheses

*** indicates significance at $p < 0.001$, ** at $p < 0.01$, and * at $p < 0.05$

Table 1.4: Probit regressions for provision point achievement in three network types

and public goods provision. If affluent members of a society can leverage their influence via social networks to commit their neighbors to contribute to a public good then it is much more likely that the provision point with be reached – regardless of that highly central actor's type. When that actor becomes central by a different mechanism, modeled here as randomness, the positive outcome is much less likely. Another strong indication of the strength of network structure on the outcome of these simulations are the minimal impacts agent type and number of neighbors have on outcomes across all network types. In the former case, only in uniform networks does being a Min-match type make you slightly less likely to reach the provision point, while in the Pareto networks being a Max-match have a similar effect. On the latter, the number of neighbors has no significant effect in any of the models, providing evidence that the mechanism by which networks form is much more important than number of ties alone.

After examining the distribution of contribution frequencies for all agents in the various network types, and considering what network configuration had the highest probability of successfully reaching the provision point; next, I consider how contributions vary as a function of wealth and what factors of the model influence this contribution. In Figure 1.4 the data are disaggregated by agent type and provision point – as before – however, in these plots the x-axis represents agent wealth, and the y-axis their contribution. Also, only those agents with a disposition to give are plotted in Figure 1.4, which provides a slightly more detailed view of the variation in giving across the network and agent types than provided by the plots in Figure 1.3.

Beginning with the binomial network, note the very tight clustering of giving for all agent types, regardless of wealth or provision point. Again, note that for both the Min- and Max-match types their is very little variation in giving.[14] It is also clear how much of the Max-match giving behavior were driven by the Community members in those densely connected binomial networks, as the distributions of giving are nearly identical. Moving on to the uniform networks, we see a similar pattern, but with slightly less clustering – particularly with the matching agent types. Here, while he general pattern is followed, there are many instances where there agents are giving far more or less than would be expected in a binomial network setup. Again, this

[14] Recall, agents decision criterion is not a function of their wealth for any types, so the clustering is expected.

variation comes from the uniformity of ties and the wealth and types present in the local clusters formed by the uniform network types.

Finally, the variation in giving for the matching agents in the preferential attachment is extremely clear in this figure. Regardless of wealth or provision point, matching agents are densely scattered across all giving levels. That said, there do appear to be more outliers with high wealth giving relatively little of Min-match type when the provision point is not met, compared to when it is. This is notable, as it may indicate the strength of miserly wealthy agents in preferential attachment networks of this kid; wherein, their unwillingness to give has a network-wide affect causing the threshold to not be met. As before, however, it is useful to take this analysis one step further and measure the impact of specific aspect of the model on the level of contribution. From these plots it is clear that wealth had little effect, therefore, in the final analysis a generalized linear model (GLM) is used to measure how model parameters affected contribution levels across all network types.

To measure the change in a proportional dependent variable, as is the case for the level of contribution is this model, a generalized linear model of the binomial family with a logit link is used. Given the specifications of the dependent variable in this model, such a GLM provides a quality parametric fit. As such, Table 1.5 provides the results of this GLM run for all network types. The first items of note from the table are the coefficients on the intercepts for all of the models. Note, as before, all things being equal contribution levels will be higher in preferential attachment networks than any others. The statistical significance of these coefficients, however, are is extremely low, and likewise the standard errors on these estimates are enormous. As such, it is difficult to draw any conclusion from coefficients. Fortunately, the estimates on the independent variables of the model do provide a greater degree of insight.

Parameters	**Network Types**				
	Binomial	Uniform	Pareto	Power-law	Pref. Attach
(Intercept)	−48.54	−44.51	−41.18	−41.21	−31.01
	(1865.51)	(691.30)	(280.50)	(279.53)	(120.71)
Wealth					-0.23^{***}
					(0.02)
Neighbors		-0.00^{**}	0.01^{***}		0.11^{***}
		(0.00)	(0.00)		(0.01)
Threshold					0.02^{**}
					(0.00)
Provision†			0.10^{***}	0.08^{**}	0.09^{***}
			(0.03)	(0.03)	(0.03)
Disposition†					
Miserly†	-3.66^{***}	-3.67^{***}	-3.67^{***}	-3.66^{***}	-3.67^{***}
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Community†	1.16^{**}				10.97^{***}
	(3.82)				(2.74)
Min-match†	-4.41^{***}	-3.53^{***}	-0.60^{***}	-0.73^{***}	
	(0.11)	(0.07)	(0.03)	(0.03)	
Max-match†		4.89^{***}	0.70^{***}	0.88^{***}	0.86^{***}
		(0.14)	(0.03)	(0.03)	(0.03)
AIC	11057.68	12293.94	29640.28	28778.61	30102.51
BIC	11426.69	12662.95	30009.29	29147.59	30471.52
log L	−5488.84	−6106.97	−14780.14	−14349.30	−15011.26

$N = 75000$ in all models

† indicates dummy variable, and agent type variables *labeled in italics*

GLM with binomial family and logit link used for proportion dependent variable

Agent type *Altruistic* not defined because of singularities, and thus are excluded

Standard errors in parentheses

*** indicates significance at $p < 0.001$, ** at $p < 0.01$, and * at $p < 0.05$

Table 1.5: Generalized least squares regression for contribution level in all network types

First, because the Miserly type has only a small range of possible contribution, the table shows that no matter what the network configuration these types are always going have negative affect on contribution. In fact, that effect is equal for all network types. Next, Community types are likely to give more in binomial and preferential attachment networks, with a strong effect present in the latter. The effect of matching agents is also within expectation, as Min-match have negatively signed coefficients and Max-match with positive. In the case of uniform networks, Max-match agents have a very strong effect. Finally, these result support previous evidence that the preferential attachment mechanism is provides that best outcome for contributions, and thus reaching the provision point. The positive effects are strongest in this network type, and in this model even the number of neighbors on an agent has a positive affect. This is likely due to the scarcity of ties outside of those to the main hub, which when they are present work to reinforce positive network effects.

1.5 Conclusion

The following design and experimentation of a computational model for a network variant of a provision point public goods game has revealed many interesting aspects of how networks can influence outcomes. Most notably, the evidence suggests that preferential attachment mechanisms for creating network structure have a positive outcome on the provision of public goods, regardless of the decision criterion being used by players to contribute to the public good. There are important consequences for this result from the perspective of state building through public goods.

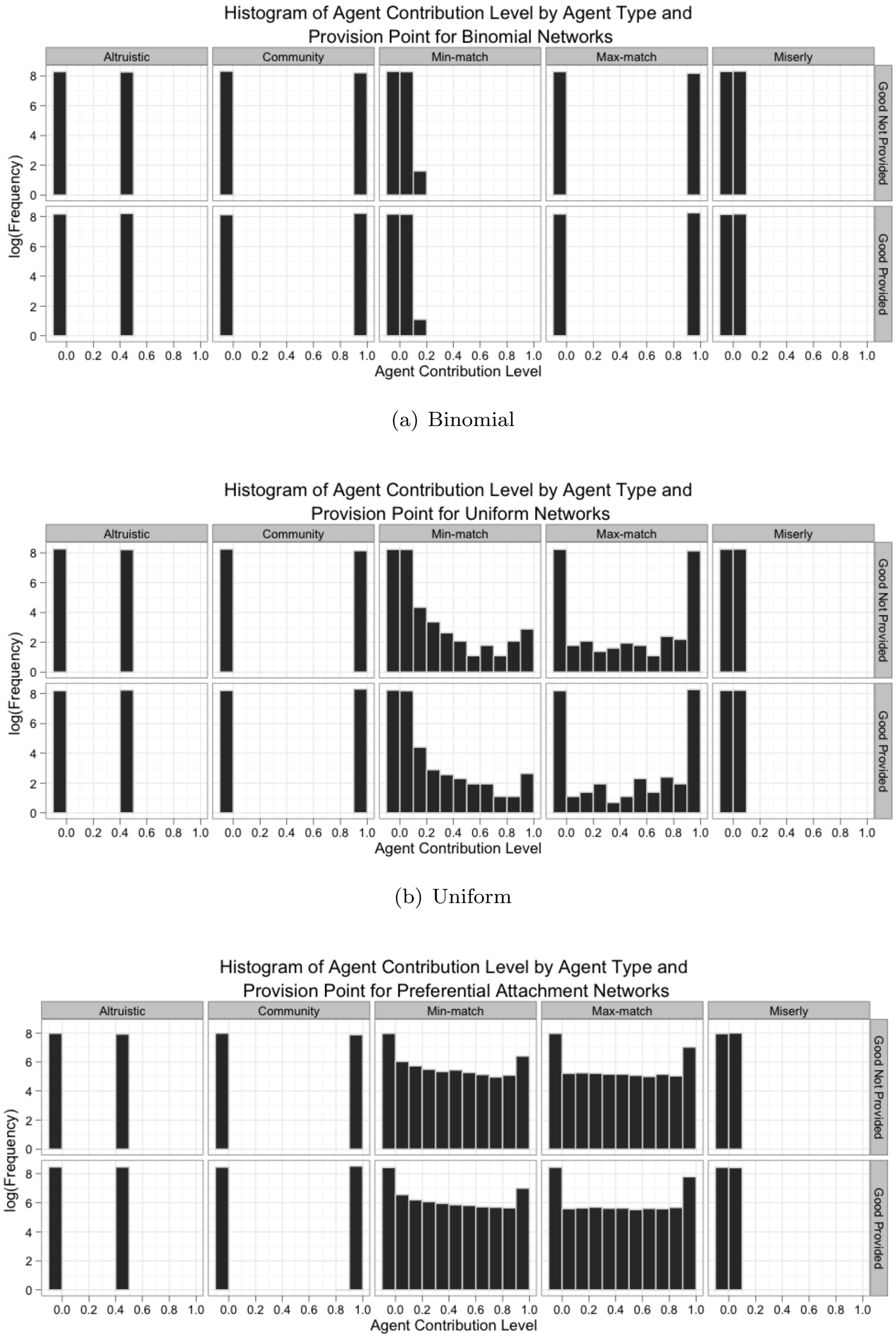
As is noted in the networks literature, a preferential attachment mechanism is often observed in naturally forming social networks, where the cost or barrier to creating a tie is negligible. In places where state formation and public goods provision have been difficult, such as Afghanistan, the cost of creating these ties may be high. As a result, the types of networks needed to help support the provision of public goods, such as those generated via a preferential attachment mechanism, do not have a natural path to formation.

Path dependence based on terrain, conflict, and historical boundaries can all act as hinderances to the natural formation of these networks. As such, the networks are much more likely to form as one of the less supportive structures discussed above; such as the uniform networks – where tight clustering occurs locally – or the power-law – where hubs are defined be some exogenous factor.

Perhaps more interestingly, however, is how the design of these preferential attachment networks – relatively to the typical generation – affect outcomes. In this case, an agent's preference is based on wealth, rather than the number of ties. The strong positive results here for preferential attachment networks may also provide evidence that the centrality of wealthy individuals within an informal institutions also plays a central roll to the likelihood that formal institutions will follow. Furthermore, because agent-types populate each network at random, these experiments do not provide insight as to how the structural position of agent types, as a function of wealth, contribute to the outcomes.

Additional experiments are needed to disambiguate how wealth, agent-type, and structural position contribute to agent decision making; along with the types of networks these agents exist in.

As a first step step toward state building, it may be necessary to break these barriers, or lower the cost of creating ties more naturally. By doing so, networks may form in a way that support the flow of positive information to actors, which can then result in a higher probability of the public good being provided. One most be cautious, as negative outcomes from networks can occur; however, the evidence from the research presented here does not indicate that such negative effects are likely.


Histogram of Agent Contribution Level by Agent Type and Provision Point for Binomial Networks

(a) Binomial

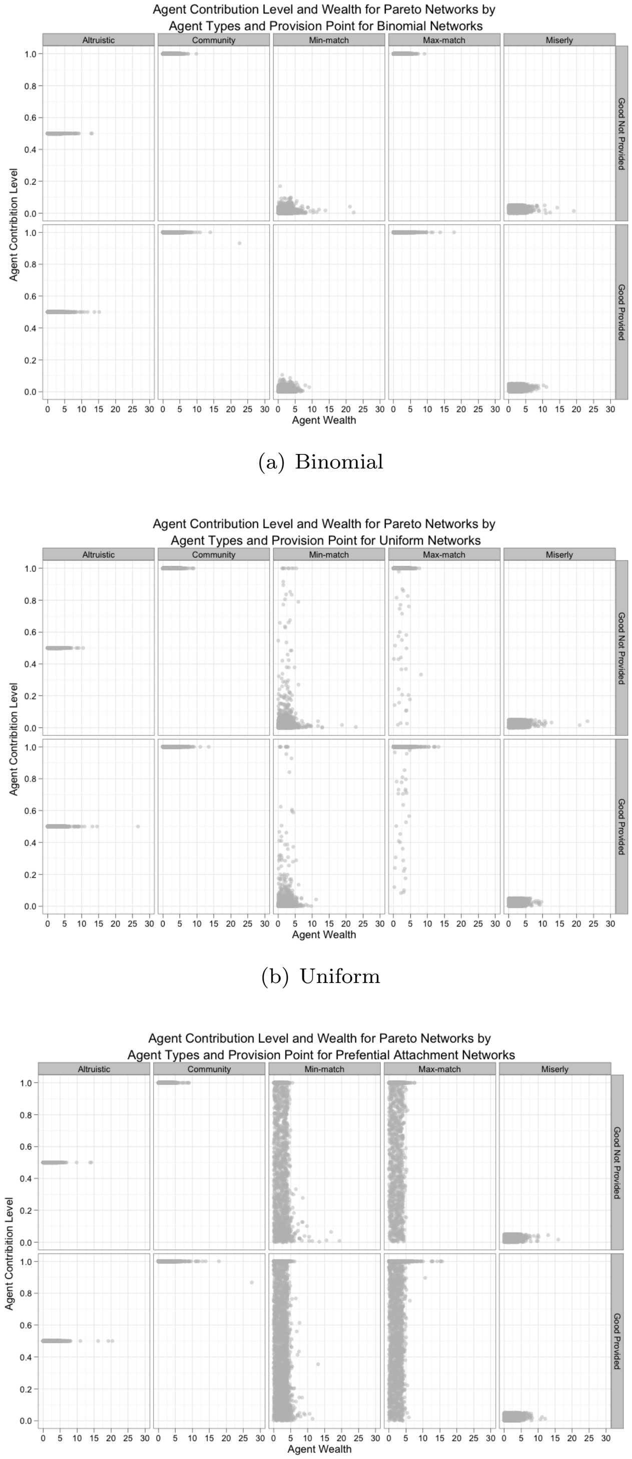
Histogram of Agent Contribution Level by Agent Type and Provision Point for Uniform Networks

(b) Uniform

Histogram of Agent Contribution Level by Agent Type and Provision Point for Preferential Attachment Networks

(c) Pref. Attach

Figure 1.3: Agent Contribution level by type and provision point in three network types



(a) Binomial

(b) Uniform

(c) Pref. Attach

Figure 1.4: Agent contribution level by wealth with provision point and type for all network types

Chapter 2

Modeling Network Structure Using Graph Motifs

2.1 Modeling Networks

The study of networks is perhaps one of the most interdisciplinary fields of study in contemporary scholarship. Many types of data can be represented as a network, therefore, the application of network analysis and modeling have been applied across a wide breadth of disciplines. In the social sciences, the units of analysis are most often human interactions, which by their very nature are difficult to model. More specifically, network structures are extremely important to the study of political science. In many sub-fields there are examples of data that can be represented as networks; including, trade, diplomatic and conflict relationships. The social structure of various networked organization is also of interest to many researchers—particularly terrorist or criminal networks. In most cases, however, there are subtle structural dynamics present in these data, making them difficult to model using traditional methods.

The research and development of random graph models that consistently characterize structural phenomenon observed empirically in social and complex networks dates back to the work of Paul Erdõs and Alfrèd Rènyi in their seminal work on random graph models (Erdos and Rènyi, 1959). More recently, a family of models developed from the work of several scholars in the network science discipline have emerged as the preferred class for the study of complex networks. This class of models; known as the exponential random graph models (ERGM), or p^*, have become the preeminent method for the development and study of complex social structures. These models are the product of several critical observations accumulated through previous research in networks.

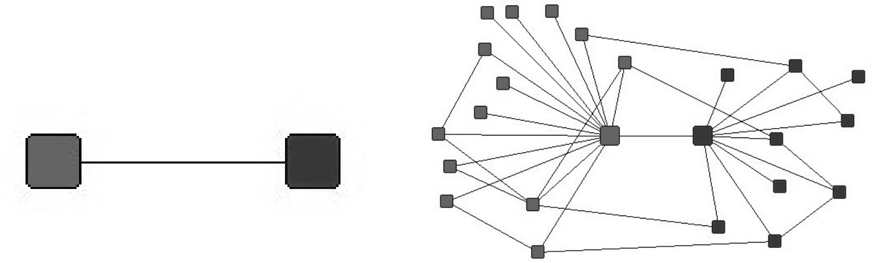
Prior to this work, the so-called "small-world" network model was introduced by Watts and Strogatz (Watts and Strogatz, 1998), and was predicated on two important observations in social networks: short average path length between nodes, and a high level of localized clustering among nodes. These structural phenomena were often observed in relatively small networks, but as technology improved so did our ability to study large complex networks. Following the Watts-Strogatz model was the work of Barabási and Albert (Albert and Barabasi, 2002), which noted that structure within complex networks exhibited "preferential attachment," meaning a limited number of nodes drew in disproportionally more edges than the vast majority of others, creating the now well-studied finding of "heavy-tailed" degree distributions in complex networks (Barabasi and Albert, 1999; Albert and Barabasi, 2002; Clauset et al., 2009). The ERGM class of models retains the structural consistency of these previously developed models; however, ERGM assume a fixed number of nodes, and structure among these nodes are modeled as random variables in a stochastic process (Robins et al., 2007).

In addition to these models, over the past several years there has been an explosion of highly tailored network models developed to address specific structural features of networks. Many of these models are more closely related to graph theoretic concepts, attempting to bridge the gap between the classical concepts of Erõds and Renyi (Bollobás, 2001; Newman, 2003). Likewise, an alternative class of contemporary models take an agent-level approach, evolving structural growth as a decision process occurring endogenously through the nodes themselves (Leskovec et al., 2005; Steglich et al., 2010). While the massive and growing literature on random graph models has provided enormous insight into the general structural dynamics of networks, these models are limited in both their underlying assumption about the means by which structure is generated, and the types of networks they can model.

To be sure, given the volume of literature in this field the number of proposed network models are nearly innumerable. For some, using a model specifically designed to approximate the dynamics of interest may be adequate. The rigidity of these models, however, makes them much less useful for modeling less well-understood network dynamics. Likewise, the ERGM class of models can theoretically model any countable graph, which itself constitutes a monumental and unifying result from this work. In practice, however, the Markov-Chain Monte Carlo (MCMC) methods used to estimate ERGM models provide a much sparser landscape of possible graphs. The problem of model degeneracy is well known in the ERGM literature, and attempts have been made to address these problems (Handcock, 2003). Unfortunately, the practical implications are quite limiting, with many models of interests degenerating into complete or empty graphs.

The primary shortcoming of these models is their treatment of the atomistic component of a network—the node. In all of the models mentioned above, and in fact in the vast majority of random graph models of social networks, actors are modeled as entering the system in a vacuum—free of any pre-existing structure. In real networks, however, it is clear that this is not the case. Except in the simplest of cases, whenever an actor enters a network system that actor is bringing some degree of exogenous structure, which will have an immediate impact on the growth trajectory of that network. This is particularly true of human social networks, which exist in a rich, complex, and often hidden fabric of social ties.

Consider the network dynamics when two people meet each other for the first time. Upon meeting, these individuals have changed the structure of their social networks by creating an edge between them, but with that structure they have also brought with them their pre-existing social structure. All of the people they already know; friends, family, co-workers, competitors, etc. This meeting has not simply created a dyad existing in isolation, but rather it has connected two large components, and increased the probability that the single bridge created by this dyad will in turn become a cluster of shared relationships. Figure 2.1 visually depicts the difference between these concepts.



(a) Dyadic model

(b) Motif model

Figure 2.1: Competing models of social interaction

Clearly, however, there is considerable nuance and ambiguity with respect to how to model the relationships in the right panel of Figure 2.1; whereas, the dyadic relationship on the left is simply a binary event. The dependencies related to these ties can be a function of network-level metrics, e.g., diameters, centralization, or density; node- and edge-level attributes, e.g., centrality, type, or direction; or any combination therein. The plethora of potential modeling parameters have lead to a literature full of rigorous, yet limited models for network growth. Current random graph models of social networks are useful, but are limited by oversimplified assumptions that ignore the inherent complexities of social structure. This research attempts to close the gap between the theoretical assumptions of current models and the self-evident reality of natural network interactions by providing a more flexible framework capable of modeling a much larger set of networks.

The paper proceeds as follows: first, the graph motif model is introduced. Next, as this method relies on a combination of various machine learning and computation techniques, a detailed description of the algorithmic implementation is provided. This is followed by a brief introduction to the `gmm Python` package for specifying graph motif models, with a simple example. The final section describes a new network data set on co-authorship within the Political Science discipline, collected specifically for this analysis. This data is then used in a more thorough example of graph motif modeling with a discussion of the results. The conclusion focuses on the advantages of this framework over current methods, its shortcomings, and necessary future research.

2.2 Graph Motif Model

To overcome the limitations of current random graph models of social networks the concept of a graph motif model (GMM) is introduced. This new framework is predicated on two key assumptions that distinguish it from other network modeling techniques. First, new actors entering a network do not do so in a vacuum, i.e., actors bring exogenous structure to a network when entering it and thus models of social networks should build new structure in an analogous way. To model networks this way, however, it is necessary to form a priori beliefs about these exogenous structures and the process by which they will enter the network. As such, the second assumption is new network structure will resemble currently observable structure in type and frequency.

Given this second assumption, current structure can then be used to form these necessary beliefs. This, however, forces a strict requirement for GMM that is not shared by other random graph models—specifically—the need for some base structure from which to form structural beliefs about the network being modeled. It could be argued, however, that all random graph models require base structure in that they all require some fixed number of nodes to model. A set of nodes without structure still constitutes a base graph, despite its degenerate form. This is particularly true of Barabasi-Albert model of preferential attachment, which always begins with the same base structure.[1]

The observation that networks perpetuate self-similarity as they grow has been noted several times in the empirical literature. In fact, complex networks exhibit significant fractal scaling (Song et al., 2005; Kim et al., 2006; Kim and Jeong, 2006). This literature also shows that fractal scaling persists in networks at both the micro- and macro-scales (Kim et al., 2007);

[1] In practice this is often modeled as a single dyad or a three-node line graph.

therefore, it is natural to use this observation to form the critical bridge between the first and second assumptions. To be clear, there are several assumptions that could be used to form beliefs about network structure, such as node-level metrics, stochastic processes, etc.; however, self-similarity is preferable in that it is minimal and not dependent on a graph's type. For example, if the model used a node-level metric as a foundational assumption the model would also require that this metric described all networks, which may in fact be a contradiction for some graphs, i.e., directed vs. undirected graphs, weighted vs. unweighted graphs, weakly connected vs. strongly connected graphs, and so on. Furthermore, the burgeoning literature on the effects of social networks on political outcomes and collective actions suggests a strong relationship (McClurg, 2003; Scholz and Wang, 2006; Siegel, 2009). As such, it may be the case that individuals are using social networks to overcome problems inherent in collective action, and informational or efficiencies gains may be made through the repeating of certain network structures in various contexts.

To begin, the GMM framework described here applies to undirected and directed graphs with an arbitrary set of node or edge attributes. While this allows for an extremely rich set of possible models, it precludes some graph forms; specifically, multigraphs and hypergraphs. This restriction is done to limit the complexity of the model in this initial form. It may be possible to incorporate multigraph models into the GMM framework; however, the abstract nature of hypergraphs makes their applicability to social network models unclear. For example, consider a hypergraph wherein a single edge is incident on many nodes. This is not a construction that models social interactions naturally, and thus incorporating them into the model has limited value.[2] With these restrictions, the model proceeds as follows: given some graph G of arbitrary

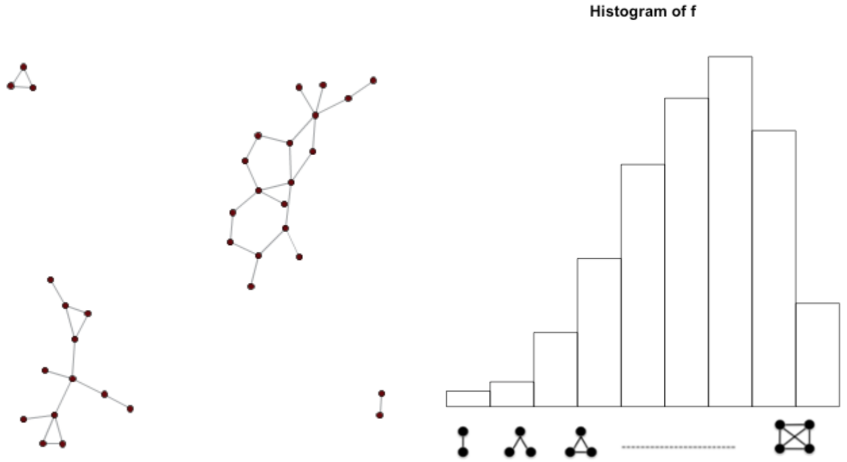
[2] A "multigraph" is defined as a graph where any two nodes may have multiple edges between them. Conversely, a "hypergraph" is defined as graph where a single edge may be incident on any number of nodes.

size and some integer $\tau > 1$ count all of the subgraph isomorphisms in G of graphs $i \in I$, where I is the set of all *single-component graphs formed by τ nodes.*[3] These single component graphs are the motifs on which the entire GMM framework rests. For example, suppose $\tau = 3$, then $I = [\{V = 2, E = 1\}, \{V = 3, E = 2\}, \{V = 3, E = 3\}]$ where V is the number of nodes and E the number of edges of graphs $i \in I$. In this example, therefore, $\{V = 3, E = 1\} \notin I$, as this graph contains two components: a dyad and an isolate. Also, note that these motifs have a natural ordering given their number of nodes and edges, although as V and E increase this ordering is not strict as motif can have equal numbers and nodes and edges and not be isomorphic. This will become become critical to how new network structure enters the model.

Next, let $f(i_n, G)$ be a function that describes the number of subgraph isomorphisms of i_n contained in G, and S be an ordered n-tuple where $S = \{i_1, i_2, ..., i_n\}$, such that i_n is increasing in number of nodes and edges. For two graphs to be isomorphic there must be a one-to-one correspondence among the nodes and edges of two graphs. A subgraph isomorphism between two graphs G and H, therefore, is defined as such a correspondence for graph G in an induced subgraph of H. This construction is very useful, as it allows for the quantification of motif frequency in any given base structure, i.e., the composition of a graph given some set of possible "constituent parts." While the subgraph isomorphism problem is known to be NP-complete, certain cases can be solved in polynomial time and several algorithmic approximations have been proposed (Ullmann, 1976).

Figure 2.2 is an illustrative example of how S can be ordered by complexity, and the counts of subgraph isomorphism. The left panel shows some base structure consisting of four connected components. The right panel depicts how the function $f(i_n)$ generates counts over all motifs

[3] The restriction that τ be strictly greater than one accounts for the fact that a $\tau \geq 1$ would allow for a singleton element, which would violate the first assumption of GMM for exogenous network structure.



(a) Example base structure

(b) Subgraph isomorphism histogram

Figure 2.2: Natural ordering of S with subgraph isomorphism counts for some base graph

in S. Note, however, that the frequencies represented in the example are purely illustrative, and do not reflect actual subgraph isomorphism counts of motifs in the left panel. The tuple S thus constitutes the observable variables of interest, ordered by structural complexity; i.e., the types of structure the second assumption predicts will likely be observed as the network grows given some value of τ. This tuple can then be used to generate beliefs about the type of structure entering the network as it grows in size and complexity. In order to generate these beliefs some function over S must be defined. The function, however, can take several forms.

2.2.1 Generating Beliefs About Structure

As discussed in the introduction, a critical aspect of modeling network growth using motifs is specifying what types of motifs enter the graph. Using subgraph isomorphism it is possible to calculate the frequency of these motifs, but it is still necessary to explicitly specify how these discrete counts are used to generate the probabilities. A straightforward way to do so is to simply define a discrete probability mass function over these counts.

Given S, therefore, define a probability mass function (PMF) such that $\sum_i^S Pr(X = i) = 1$, where i is element of the tuple S with discrete probability, and the sum of probabilities for all elements in S is equal to one. As the number of elements in S is dependent on τ, it is not necessary that the PMF relate exclusively to the elements of S. For the purposes of discussion in this paper the PMF defined are both exclusive to this set as well as meaningful beyond it. For example, recall the simple set of motifs described for $\tau = 3$. In this case, a GMM would not require a PMF that satisfied the above requirement for the complete graph mode of four nodes because this motif would be excluded from S by definition.

As stated, this allows for a large set of possible PMF to be used to specify the probability a given motif will enter the network. This function may rely explicitly on the subgraph isomorphism counts, wherein zero probability mass is defined for any motif with no subgraph isomorphism in the base structure. Alternatively, it is also possible to specify a PMF that models the probability of motifs as a discrete probability distribution over all elements of S. Below I will describe two examples of PMF; the first an explicit function over the elements in S , and the second a function that provides positive probability mass for all elements of S regardless of whether any were observed as subgraph isomorphisms in the base structure.

$$F(i) = \frac{S_i}{\displaystyle\sum_{n=1}^{S} S_n} \tag{2.1}$$

Figure 2.3: Explicit PMF for motif probability

The probability that some subgraph $i_n \in S$ will by the next structural component of graph G may be given by F, the discrete probability distribution over S defined in Equation 2.1. This function states that the probability i_n will be the next structural component of G is given by the proportion of subgraph isomorphisms found for $i_n \in G$ given the total number of subgraph isomorphisms counted $\forall i \in S$. F thus defines a discrete probability distribution over S, and provides the necessary prior beliefs to generate new structure in G. Again, this function will assign zero probability mass to any motifs that are not observed as subgraph isomorphisms in the base structure. This can be problematic, as it presumes that certain motifs will never enter a graph—clearly limiting the possible networks it can model. In other cases, however, this limitation may be necessary and is the case when modeling the co-authorship described below.

Given this limitation, it is also useful to have a PMF that assigns positive probability to all motifs. Here, we may utilize the natural ordering of elements in S by their structural complexity to fit a canonical discrete probability distribution. Specifically, in this case I define an alternative PMF for the elements of S in terms of the Poisson distribution in Equation 2.2.

$$F(i;\lambda) = \frac{\lambda^i e^{-\lambda}}{i!} \tag{2.2}$$

Figure 2.4: Poisson PMF for motif probability

In this construction the "mean" of the distribution, represented by the shape parameter λ, is the mean of all motif counts in S. The natural ordering of motifs by complexity fit the motivation of the Poisson distribution to model event counts, as we may consider the occurrence of increasingly complex motifs within a given base structure as a diminishing rare event. Likewise, the most likely motifs to enter graph may have probabilities centered around the motif with mean complexity in the base structure. If these assumptions do not reflect the data generating process present in the base structure, however, such a specification is misplaced and an alternative specification of the PMF should be used.

It is important to note that these, or any PMF defined over S has a direct effect on the nature of the GMM specified. The function defined in Equation 2.1 requires the least amount of assumptions about the probability of motifs in the model, relying explicitly on subgraph isomorphism counts. This, however, is limiting and alternative methods may be desirable. As such, it may be useful to define a PMF from the canon of discrete probability mass function given that the assumption of those distribution are relevant to the GMM, as I have done using the Poisson distribution in Equation 2.2. Next, once probabilities are defined over S methods for adding these structures to the graphs must be defined.

2.2.2 Generating New Structure

With these beliefs generated, the next step in the model is to draw some motif from S using the probability distribution and add it to the network structure by some growth rule. This rule is denoted $R(\cdot)$ and is defined as a mapping $R : i_n \rightarrow G$, which is restricted only by the graph theoretic constructs assumed by G. That is, the decision rule must be applicable to the fundamental constructs of G and subgraph elements of S, but is otherwise open to the particularities of a model's design. For clarity, the basic steps of the GMM framework are listed in Figure 2.5, and in following section a simple growth rule is defined.

After each iteration of growth the process for forming structural beliefs is repeated, and the probability distribution is recalculated. The continual updating of beliefs as the network grows allows for a certain degree of path dependance in the model, as probability mass may converge to the most likely motifs as the network grows. This may, or may not, be viewed as an advantage of the model, but future version will allow for both static and dynamic probabilities over S.

Now, this process continues until the model has satisfied some termination rule denoted as $T(\cdot)$, which is restricted as $R(\cdot)$. The methods by which structure is added to the network and the model is terminated are intentionally left open, as the GMM framework is meant to support any number of possible growth models. Beyond the limited restrictions described above, it is completely up to the discretion of the modeler. In the following section the algorithmic implementation of this method is described in detail, and one simple implementation of a GMM is specified. Before proceeding, however, a review of the core elements of the method for modeling networks using graph motifs:

The framework for modeling network structure using graph motifs described above attempts to overcome the limitations of current methods by supporting a much larger set of possible network types, allowing for flexible specifications, and modeling network growth more naturally by requiring exogenous structure to enter the network. To achieve this, two key assumption are made: the presence of some base structure upon which to form beliefs about the type of structures present in the graph being modeled; and that the constituent parts of this base structure—modeled as graph motifs—act as a meaningful proxy for the data generating process present in the network being modeled. These assumptions are in stark contrast to those of many traditional network models.

The GMM framework brings with it a different set of limitations, many of which will be discussed in the conclusion. As much of the model hinges on subgraph isomorphisms counts, this model requires a sophisticated computational implementation. In the following section an implementation is described using the `Python` programming language to develop the `gmm` package for graph motif modeling.

1. Begin with some base graph G of arbitrary complexity
2. Given some integer $\tau > 1$, the set I contains all single-component subgraphs formed by τ nodes
3. Define S as an ordered n-tuple containing all $i \in I$
4. Define the function $f(i_n)$ to count the number of subgraph isomorphisms of $i_n \in G$, and a PMF over all elements in S
5. Draw structure from this probability distribution and add that structure to the network by some growth rule $R(\cdot)$
6. Repeat steps 4-5 until the some termination rule $T(\cdot)$ is satisfied

Figure 2.5: The basic steps of the GMM framework

2.3 Algorithmic implementation: the `gmm` `Python` Package

Before any implementation of the GMM can proceed, it will be necessary to have a means for representing complex networks computationally in the `Python` language.[4] Fortunately, the `NetworkX` package is a highly-developed Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks (Hagberg et al., 2008).[5] `NetworkX` is capable of representing graphs of arbitrary complexity, including both node and edge attribute data, making it ideally suited to be the computational foundation for an algorithmic implementation of the GMM framework.

The `gmm` package consists of two object classes. The first is the "gmm" class itself, which is the essential element of any model. Given the specification of the GMM framework outlined above, this class requires three elements: a `NetworkX` graph object as the model's base structure; and growth and termination rules, as `Python` functions. With these elements in place, the "gmm" object's primary function is to verify that all model parameters are valid to

[4] For more information on the Python language see `http://www.python.org/`

[5] The `NetworkX` package exploits existing code from high-quality legacy software in `C`, `C++`, `Fortran`, etc., is open-source, and fully unit-tested. For more information on NetworkX see `http://networkx.lanl.gov/`

a GMM model, store these parameters appropriately, and provide functionality for storing and retrieving information about a given GMM simulation.

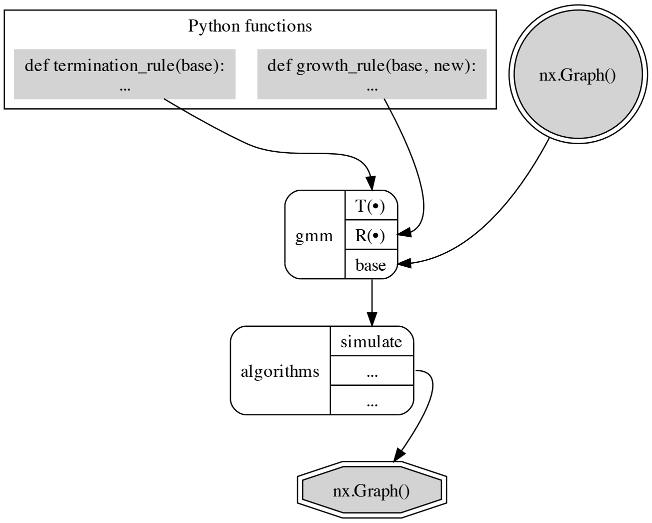


Figure 2.6: Implementation of gmm with dependencies

The second class is "algorithms," which provide all of the functionality for properly running a simulation and generating network structure from a given "gmm" object. Within this class are all of the functions needed to create a set of graph motifs, generate beliefs about how that structure enters the model, and the actual generation of new network structure. With respect to generating beliefs, the two PMF discussed in the previous section are included in the current version of this class. The Poisson function requires SciPy—a third-party scientific computing package—to generate the correct probabilities. NetworkX also requires this package as a dependency, therefore, its use in the gmm package does not compound software requirements.[6]

[6] For more information on SciPy see http://www.scipy.org/

Finally, as stated previously, the subgraph isomorphism problem is known to be NP-complete and therefore a sophisticated approximation is needed given the computational complexity inherent in counting the number of subgraph isomorphisms for arbitrarily large G and τ. Specifically, the VF2 algorithm—the most commonly used algorithm to evaluate subgraph isomorphism—is used to perform the necessary calculation for matching subgraph isomorphism (Junttila and Kaski, 2007; Cordella et al., 2001). With these two simple classes, it is possible to specify a rich set of GMM. Figure 2.6 illustrates the basic computational framework described here. Inheriting the representation of graphs from `NetworkX`, the "gmm" class requires three parameters: a graph object and two properly specified `Python` functions. This object is then passed to the "algorithms" class and a simulation is run, which results in a new graph object.

Following the example of high quality scientific `Python` packages, such as `NetworkX` and `SciPy`, the `gmm` package is open-source and fully unit-tested. All of the code is free to inspect and download at this website: `https://github.com/drewconway/gmm`, which includes all unit-tests to verify all function's execution. In the next section a very basic GMM model is specified and simulated using this software, leading to a more complex specification used to model the growth of collaboration within the SSRN *Conflict Studies eJournal*.

2.3.1 A Simple GMM with Random Growth

Given the requirements of the GMM, the first steps are to determine the base structure, and the growth and termination rules to be used in the model. In this example I will use the well-known Petersen graph as the base structure, and two very simple rules.[7] The termination rule will be a "node ceiling," whereby the model will terminate growth once the network contains

[7] The Petersen graph, often denoted as K_{10}, is a ten node graph with uniform degree of three. It is a well-studied graph for its many known properties, such as being a non-planar.

at least 250 nodes. For growth a random attachment rule will be used such that when a graph motif enters the graph a random node from the motif will be connected to a random node from the current base structure. These algorithms are implemented in pseudo-code below, and are explicitly defined as `Python` function in the documentation for the `gmm` class in Appendix A.

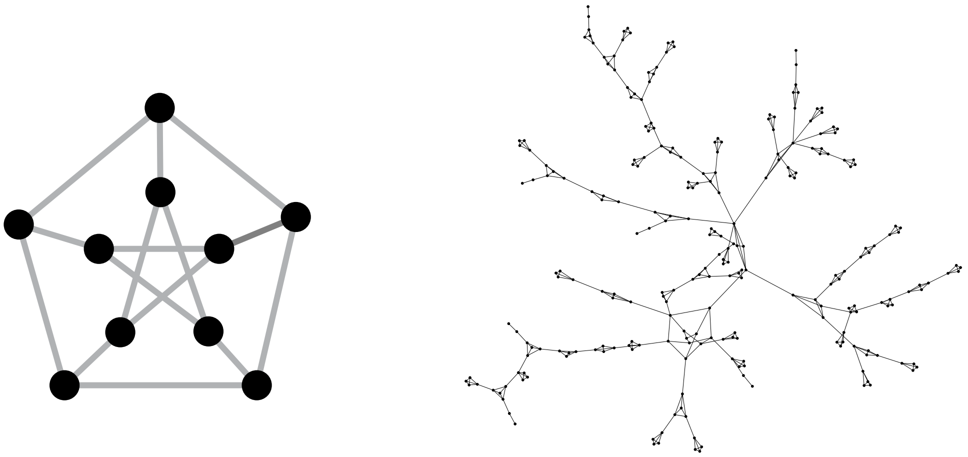
Algorithm 1 Pseudo-code "node ceiling" termination rule

```
Require: $G$
  if $G >= 250$ then
    return true
  else
    return false
  end if
```

Algorithm 2 Pseudo-code random growth rule

```
Require: $G, H$
  $G = G[H]$ {Compose $H$ with $G$}
  $r_1 = RAN(G)$; $r_2 = RAN(H)$ {Select random nodes from each graph}
  $G = EDGE(G, r_1, r_2)$ {Create edge}
  return $G$
```

This example is—of course—not a model of any particular network growth mechanism, but a useful example in that it shows the power of the GMM framework with such simple rules. The Petersen graph is made of all closed motifs, i.e., there are no pendants or pendant chains present in the graph. As such, the random growth rule will be connecting these structures by single edges, which will create simple chains of whatever motifs are drawn from the probability distributions. Figure 2.7 illustrates this, with the Petersen graph shown at the left, and the resulting simulation on the right. The elongated structure in the simulation result show the chains of motifs growing in different directions, as the random selection of nodes caused growth to occur in along several paths.



(a) Petersen graphs as base structure

(b) Simulated GMM results

Figure 2.7: Result of simple GMM with random growth rule

Using this same framework it is possible to model much more complex networks. In the following section I show how the GMM method can be used to model scholarly collaboration within the SSRN *Conflict Studies eJournal*. This begins with a brief description of how the data were collected, the data itself, and then a detailed description of the model and results.

2.4 Modeling collaboration in the SSRN *Conflict Studies eJournal*

The Social Science Research Network is a digital library that indexes non-peer reviewed, self-submitted, research within the social sciences. From their online mission statement, "Social Science Research Network (SSRN) is devoted to the rapid worldwide dissemination of social science research and is composed of a number of specialized research networks in each of the social sciences...Each of SSRN's networks encourages the early distribution of research results by publishing submitted abstracts and by soliciting abstracts of top quality research papers around the world."[8] This is similar to other online digital libraries in the hard sciences, such as arXiv.org, which indexes new research in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics, among other disciplines.

The website has been operating for over 16 years, and contains over 290,000 documents from over 148,000 authors, and has received over 40.9 million downloads at the time of writing. It is a rich repository of working papers and new research in a plethora of fields. The library is divided by discipline, for which Political Science has its own section. Within each discipline, there are several eJournals tailored to specific areas of study—either substantive or methodological. For this study, I focused on the co-authorship network of the *Conflict Studies eJournal*. From its own description the eJournal, "distributes working and accepted paper abstracts on the theoretical or empirical study of conflict. This includes both the causes, processes, and termination of conflict as well as approaches used to prevent and stop conflicts."

The decision to focus on this single eJournal was made for several reasons. First, the entire network is far too large in scale to be used. It was impractical for many practical reasons, such as the amount of available storage and processing resources available. Given these limitations it was necessary to focus on a smaller subset of the data. Perhaps more importantly, however, this was a network I had some contextual knowledge of. My own sub-disciplinary studies focus on conflict, and therefor it was logical to choose an eJournal where I had domain some knowledge about the authors and their papers. As with any modeling endeavor, it is very

[8] For more information on the Social Science Research Network see `http://www.ssrn.com/update/general/ssrn_faq.html`

useful to have some base understanding of the system being studied in order to interpret the results. This is especially useful when specifying the GMM; as having contextual knowledge about the data generating process can help design the growth and termination rules, but also when looking deeper in the data to understand its structural characteristics and dynamics.

Before any analysis can begin, however, it is necessary to generate a network from the document pages on the SSRN website. Figure 2.8 shows an example of a typical document page within the Political Science section of the SSRN. As you can see, there is a tremendous amount of contextual information beyond the simple author-article relations that can be added to the network. First, the co-authorship network is a natural bipartite network, a specific class of graphs. In a bipartite graph there are two types of nodes, which cannot form edges to nodes of the same type. In the case of a co-authorship network the node-types are authors and articles, and authors can only make connections to articles; making this a directed bipartite graph.

Beyond type, each article page contains additional information specific to each node. For authors, there are internal SSRN statistics, such as a download-based ranking; as well as institutional information. For articles there is the title, abstract, keywords and similar internal SSRN statistics. Each of these article pages also contains the date when the paper was posted to SSRN, making this a temporal network where edges have chronological information. This edge data will be particularly important as the GMM is constructed. To generate the network data I developed a series of `R` and `Python` scripts to scrape all of this data from SSRN and construct a data-rich network. For this study only the *Conflict Studies eJournal* was scraped, but the scripts were intentionally designed to be capable of building networks from any SSRN eJournal. To download and view the code see: `https://github.com/drewconway/gmm/tree/master/`

examples/ssrn. Figure 2.9 is a stylized illustration of how the network was constructed using the page from Figure 2.8 as an example.

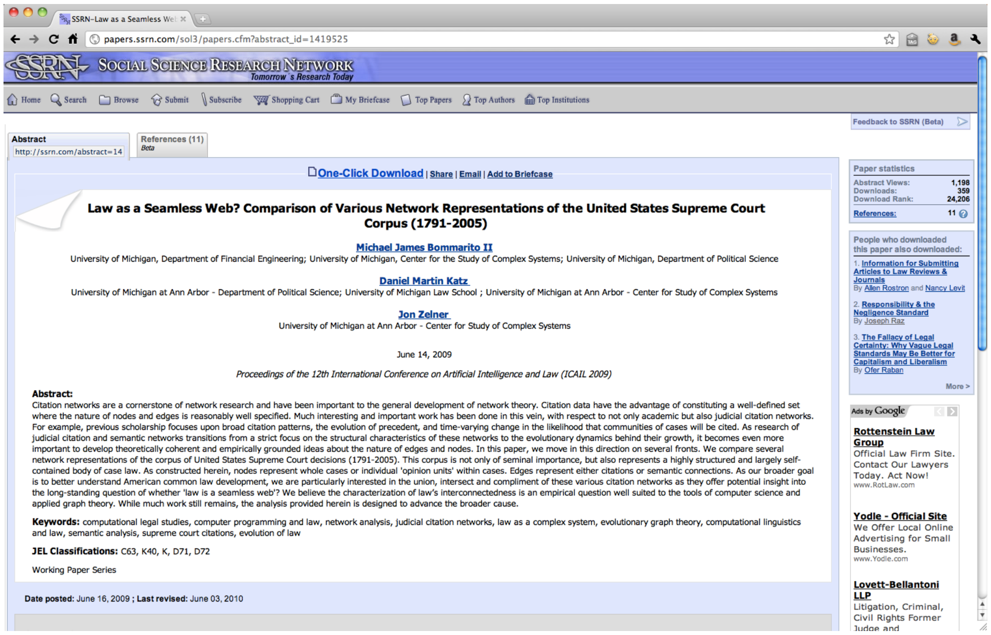
Nodes	Edges	Mean Degree	Density	Weakly Connected Components
5,515	4,457	1.616	$1.5e^{-4}$	1,493

Table 2.1: Descriptive statistics of entire SSRN network

By extracting this data for all 2,416 articles in the *Conflict Studies eJournal* at the of writing, the resulting network is a large and complex representation of academic collaboration within this sub-discipline. Table 2.1 provides some basic descriptive statistics for the network. The most striking revelation from this data is the sparsity of the network. The vast majority of connectivity occurs in small weakly connected components made up of dyads or star graphs of various sizes. This indicates that most authors in this journal submit a single paper and do not continue to submit new, collaborative, works. That said, the largest weakly connected component is sizable, with 1,190 nodes. This scale indicates a core group of scholars that are actively collaborating within the *Conflict Studies eJournal.* This main component of the network is explored in more detail in Appendix B, but the general finding is that this core is made up primarily of legal scholars—rather than political scientists. As such, this data may not be the most useful for understanding collaboration dynamics within Political Science as a discipline, but is still very valuable as a means to build and test a GMM.

2.4.1 Modeling the Growth of the *Conflict Studies eJournal*

As stated, one of the primary motivations for using graph motifs to model networks is to allow for exogenous structure to enter the model, rather than only allowing for endogenous edge formation. In terms of modeling co-authorship this is precisely the dynamic by which


SSRN-Law as a Seamless We
papers.ssrn.com/sol3/papers.cfm?abstract_id=1419525
SOCIAL SCIENCE RESEARCH NETWORK
Tomorrow's Research Today
Home
Search
Browse
Submit
Subscribe
Shopping Cart
My Briefcase
Top Papers
Top Authors
Top Institutions
Feedback to SSRN (Beta)
Abstract
http://ssrn.com/abstract=14
References (11)
Beta
One-Click Download | Share | Email | Add to Briefcase

Law as a Seamless Web? Comparison of Various Network Representations of the United States Supreme Court Corpus (1791-2005)

Michael James Bommarito II
University of Michigan, Department of Financial Engineering; University of Michigan, Center for the Study of Complex Systems; University of Michigan, Department of Political Science

Daniel Martin Katz
University of Michigan at Ann Arbor - Department of Political Science; University of Michigan Law School ; University of Michigan at Ann Arbor - Center for Study of Complex Systems

Jon Zelner
University of Michigan at Ann Arbor - Center for Study of Complex Systems

June 14, 2009

Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)

Abstract:
Citation networks are a cornerstone of network research and have been important to the general development of network theory. Citation data have the advantage of constituting a well-defined set where the nature of nodes and edges is reasonably well specified. Much interesting and important work has been done in this vein, with respect to not only academic but also judicial citation networks. For example, previous scholarship focuses upon broad citation patterns, the evolution of precedent, and time-varying change in the likelihood that communities of cases will be cited. As research of judicial citation and semantic networks transitions from a strict focus on the structural characteristics of these networks to the evolutionary dynamics behind their growth, it becomes even more important to develop theoretically coherent and empirically grounded ideas about the nature of edges and nodes. In this paper, we move in this direction on several fronts. We compare several network representations of the corpus of United States Supreme Court decisions (1791-2005). This corpus is not only of seminal importance, but also represents a highly structured and largely self-contained body of case law. As constructed herein, nodes represent whole cases or individual 'opinion units' within cases. Edges represent either citations or semantic connections. As our broader goal is to better understand American common law development, we are particularly interested in the union, intersect and compliment of these various citation networks as they offer potential insight into the long-standing question of whether 'law is a seamless web'? We believe the characterization of law's interconnectedness is an empirical question well suited to the tools of computer science and applied graph theory. While much work still remains, the analysis provided herein is designed to advance the broader cause.

Keywords: computational legal studies, computer programming and law, network analysis, judicial citation networks, law as a complex system, evolutionary graph theory, computational linguistics and law, semantic analysis, supreme court citations, evolution of law

JEL Classifications: C63, K40, K, D71, D72

Working Paper Series

Date posted: June 16, 2009 ; **Last revised:** June 03, 2010

Paper statistics

Abstract Views:	1,198
Downloads:	359
Download Rank:	24,206
References:	11

People who downloaded this paper also downloaded:

1. Information for Submitting Articles to Law Reviews & Journals
By Allen Rostron and Nancy Levit

2. Responsibility & the Negligence Standard
By Joseph Raz

3. The Fallacy of Legal Certainty: Why Vague Legal Standards May Be Better for Capitalism and Liberalism
By Ofer Raban

More >

Figure 2.8: Example article page on SSRN

structure is generated, as new articles enter the a journal or library the minimal structure allowed is dyad, i.e., a single author article. Another advantage of using this technique to model networks is it provides a method for modeling the growth of a network over time based on some current structure. That is, given some state of a network we may specify growth and termination rules that attempt to model how that network will evolve overtime using the current state as the base structure of a GMM.

In the case of the *Conflict Studies eJournal* co-authorship network, because each edge contains date information it is possible to construct a time-series of the network as new articles are posted to the eJournal. One question we might ask is: can we specify a GMM to model to growth in the *Conflict Studies eJournal* from 2008 to 2009? Figures 2.10 and 2.11 below illustrate how the network evolves from year to year, and there are several notable changes in the network's structure over this time period.

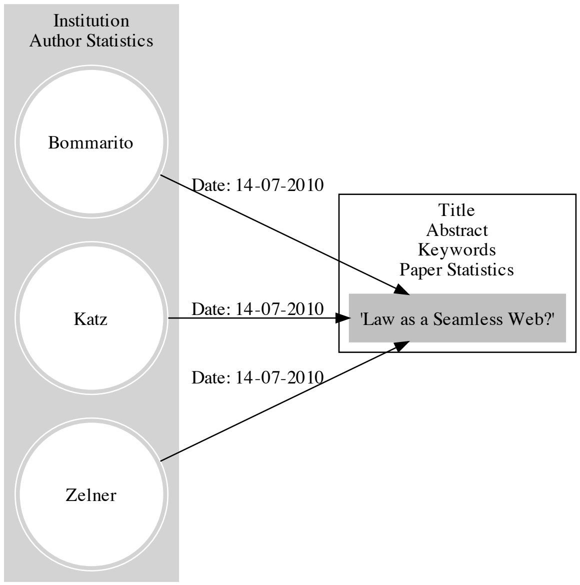


Figure 2.9: Stylized illustration of SSRN network data collected

First, the network becomes considerably larger. In 2008 it contains only 403 nodes with 375 edges, while in 2009 that increases to 1,446 nodes with 1,596 edges. More importantly, however, we see the emergence of a large weakly connected component at the network's center with 208 nodes, making up approximately 14% of the total structure. In addition to the large component, Figure 2.11 reveals smaller but meaningful components where collaboration is clearly occurring. Finally, a critical characteristic of this network is its bipartite structure, as any deviation from this would violate a principle feature of a co-authorship network. When specifying the growth and termination rules for this network these are key features of the network growth dynamic that the GMM should capture.

To begin, the simplest way to model the increase in scale of the network from one year to the next is to use a "node ceiling" termination rule, as was done before. In this case, the termination rule will halt the simulation once the base structure has grown to have at least 1,446 nodes. One concern with this rule is that while it will accurately model the evolution in terms of number of nodes, it may not capture the edge growth. The use of motifs, however, will overcome this as new edges will form as they have before, and here this means structure characteristics present in the 2008 network will carry over into the simulated 2009 network. That is, by allowing for exogenous structure to enter the network as it is growing the increase in edges will be commensurate with structure of the base graph, and thus the expectation is the resulting simulation will have a similar edge counts.

Next, a growth rule must be constructed that models the features described above. As stated, the largest connected component of the 2009 SSRN network contains about 14% of the total nodes in the network. To capture this dynamic on every iteration of the model the motif entering the network will have a 14% probability of fusing two weakly connected components together, while also maintaining the bipartite structure of the network. This is meant to model the building of the network through collaboration, wherein a new article can enter the network and create connectivity between two previous disparate components. Likewise, in order to maintain the large number of disconnected components still present in the network in 2009, if the 14% probability test fails the motif will simply enter the graph and make no connections to current structure.

To be clear, this growth rule does not explicitly specify the generation of a large central component. Furthermore, it is not calculating any metric or attribute by which a preferential attachment dynamic would induce hub structure. Instead, using a simple network-level statistic calculated from the base structure, the rule is pulling disparate components of the network together at random. This parsimony is design gives the model more power, and there is less concern that the results of the model are caused by over-specification of the GMM.

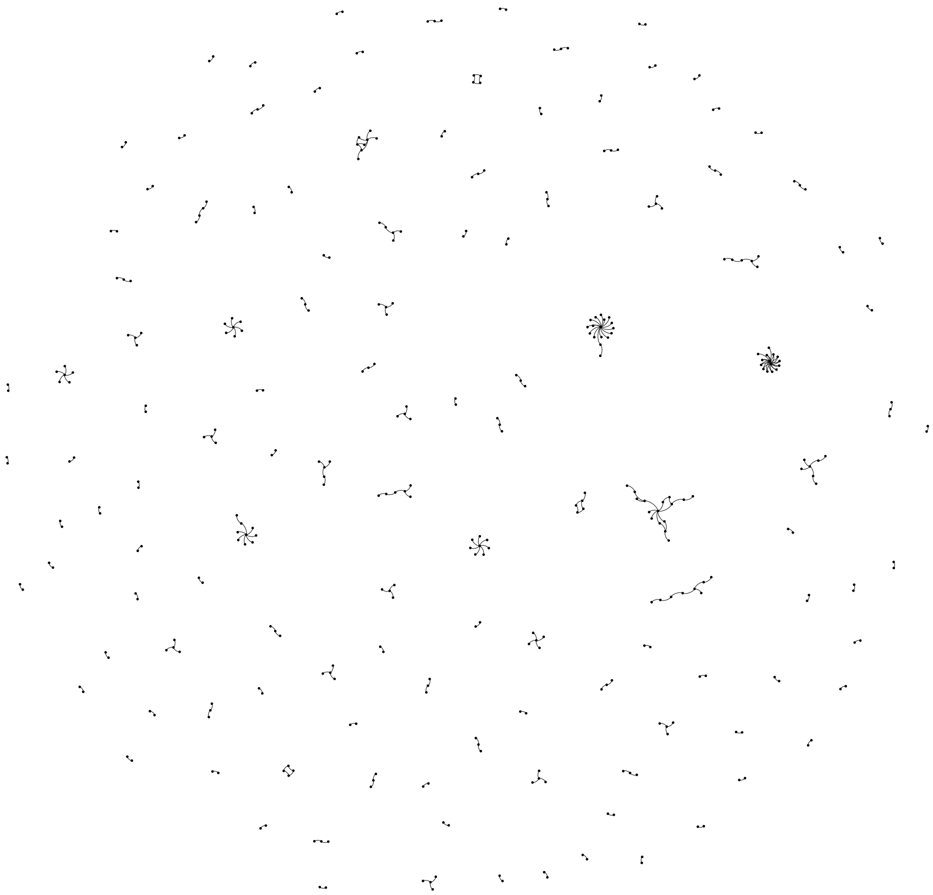


Figure 2.10: SSRN *Conflict Studies eJournal* co-authorship network, circa 2008

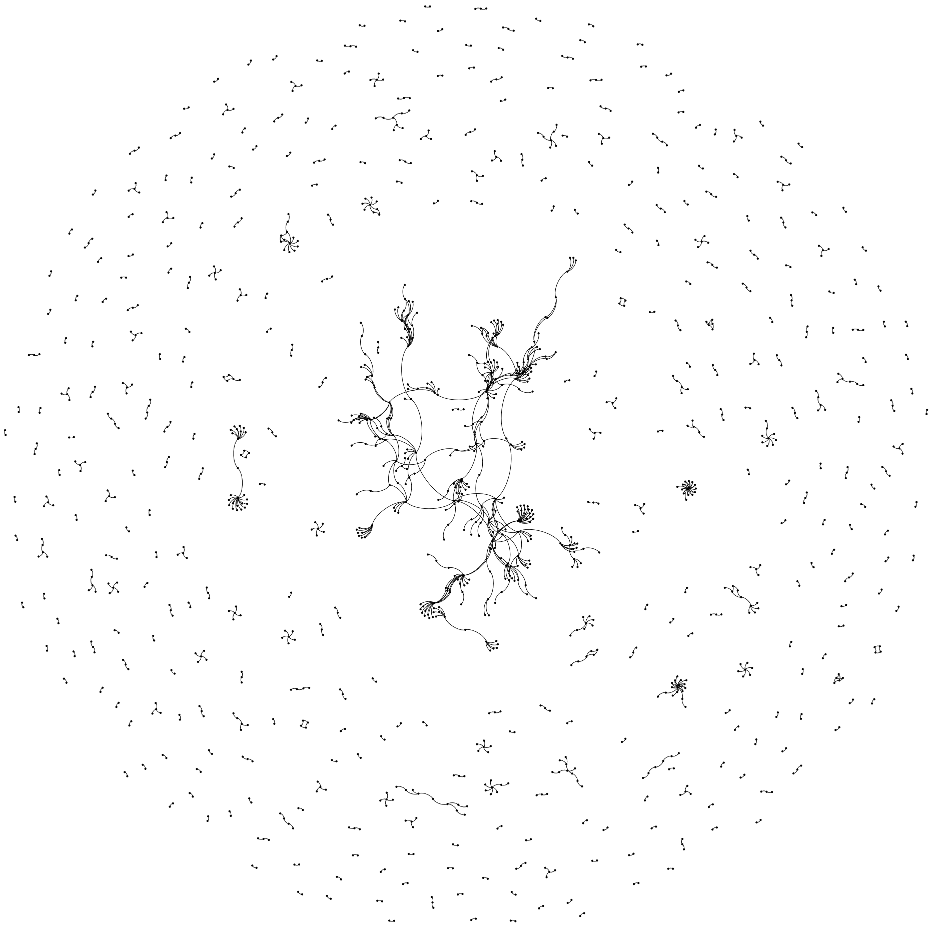


Figure 2.11: SSRN *Conflict Studies eJournal* co-authorship network, circa 2009

Finally, because the model must explicitly maintain the bipartite structure required of a co-authorship network the explicit probability mass function from Equation 1 is used to generate probabilities over the set of motifs. In this case, there are some motifs in the set that can never occur in a co-authorship network, such as a triangle or square, and therefore these motifs must have zero probability mass placed on them in order to maintain the bipartite structure of the final network. The result of this algorithm should be a bipartite network with an emergent central component, but also one with many disconnected articles. Both the termination and growth rules were implemented in Python as described here, and this code is available online (`https://github.com/drewconway/GMM/blob/master/examples/ssrn/ssrn_evolution.py`). Next, I discuss the results of this simulation and provide a brief statistical comparison of the actual and simulated 2009 co-authorship networks.

2.4.2 Results of Simulations

The GMM was specified with $\tau = 5$, a relatively large value that allowed for a fairly comprehensive set of motifs to be used in the model. The results of the simulation are illustrated in Figure 2.12. Immediately, it is clear the model was able to capture the emergence of the large weakly connected component and the persistence of many small disconnected components. What is less clear, however, is how similar the structural feature of the simulated network—particularly the large component—are to the actual 2009 co-authorship network. Unfortunately, there are limited methods for directly measuring the similarity of graphs. By way of matrix correlation or quadratic approximation procedure it is possible to compare structural similarity between networks with exactly the same numbers of nodes, where the nodes themselves have a similar identity or role in the network. In this case, however, the simulation does not guarantee that the simulation will have exactly the same number of nodes; as is in the case here. Furthermore, these simulations are abstractions from the base structure, so nodes are not given an identity in the simulated graph that corresponds to a node in the real network.

Descriptive statistics		
	Actual Network	Simulated Network
Nodes	1,446	1,448
Edges	1,158	1,160
Weakly Connected Components	338	305
Mean In-degree	2.207	1.602
Goodness-of-fit test of degree distributions		
χ^2	115.323	
Degree of freedom	88	
p-value	0.0267	

Table 2.2: Comparative statistics between the actual and simulated 2009 networks

As such, the first basis for comparison are simple descriptive statistics. The upper-panel of Table 2.2 shows this comparison between the two networks. For these basic structural characteristics the simulated network appears to have done quite well at capturing the growth dynamics in the SSRN network. The number of nodes and edges are nearly identical, and while the former is to be expected the latter provides evidence to the strength of modeling networks with motifs. There are slightly more connected components in the actual network, but not by a notable margin. The divergence in mean-degree, however, is somewhat troubling. Clearly the simple growth rule used in this example does not fully capture the cohesiveness nature of the SSRN network, as the simulation is considerably sparser than the actual network. One possible way to capture this would be to require that the growth rule minimize the diameter of the emergent large component as it was being built. Short diameters are very common in real social networks, and by enforcing such structure the component would grow with more local clustering, e.g., a "small-world" effect.

An alternative metric for comparing the networks is to test the goodness-of-fit between their degree distributions. The actual network has a maximum degree of 30, but the simulated network only has a maximum of 15. In order to perform a χ^2 goodness-of-fit test the degree distribution for the simulated network is padded with 15 extra zero entries in order to match the length of the actual degree distribution. Performing the calculation results in the values on the lower-panel of Table 2.2. Here we can see that the distributions are a poor fit; but given the p-value, at relatively high confidence levels we cannot reject the null hypothesis that the simulated degree distribution was created by the same process as the actual network's.

These results are very encouraging. Using a very simple algorithm to model the key growth features in the SSRN *Conflict Studies eJournal* from 2008 to 2009 the GMM was able to simulate a network from the 2008 data that has a very similar structure to that of the actual 2009 co-authorship. Clearly, however, there are many aspects of the actual evolutionary process that are not being captured by this specific model. The advantage to this framework is that a more refined GMM could be easily created, but that itself is not without possible danger.

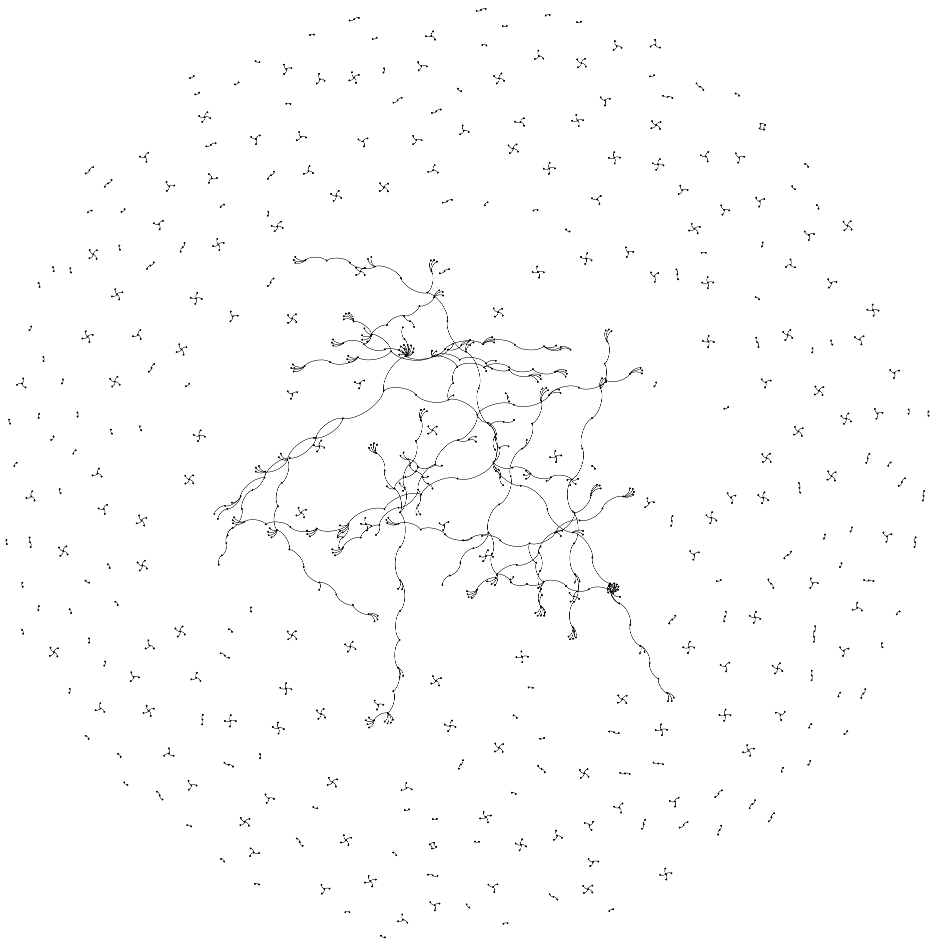


Figure 2.12: Simulated GMM of 2009 co-authorship network with $\tau = 5$

2.5 Conclusions

In this paper I have introduced an alternative technique for modeling networks called the "graph motif model." This method differs greatly from current models in its core assumptions, and how those assumptions are implemented. The first assumption is that GMM requires exogenous network growth. That is, when new actors enter the network they do so with some degree of preexisting structure; therefore, this structure should be present in the model. Second, future structure in a network will resemble current structure. This assumption relies on the observation that networks exhibit considerable fractal scaling as they increase in complexity. Using these assumptions, the GMM is constructed as computational framework for simulating network growth using some base structure, and subgraph isomorphism counts of a set of graph motifs to measure the frequency of various network structures within a network.

The basic GMM framework has been implemented as the `gmm` package in the `Python` programming language. Relying on high-quality scientific computing packages already available in `Python`, this package allows for the specification of a near boundless set of GMM to model any number of networks. To test this modeling framework a new data set is introduced: the co-authorship network of the SSRN *Conflict Studies eJournal*. Dividing this network into a time-series, a GMM is designed to model the growth of collaboration in this network from 2008 to 2009. The results of this simulation were very promising, as the simple GMM proposed was able to capture many key features of the network's evolution over this time period.

This work has many potential contributions to Political Science, and social science more generally. As stated, much of the data studied in the social sciences is relational. More specifically, this data often represents relationship among people. While there have been great advances in techniques for modeling these relationships, much of the current work relies on a set of models founded on very limiting assumptions. The dynamics of human interaction are both complex and subtle, and by attempting to fit this into a overly simple models results in a massive reduction in the types of networks that can be modeled. By using a more flexible framework, such as the GMM proposed here, social science researchers may be able to specify models that capture these elusive dynamics, and explore deeper how the ramifications of these dynamics affect social outcomes.

The technique proposed here, however, has many of its own limitations. Perhaps most pressing is it is poorly suited to model non-human networks. There are many networks for which exogenous growth is a contradiction; such as physical networks like transportation or telecommunication. Also, many biological networks are also poorly modeled with exogenous network growth, such as protein-interaction or neural networks.

Furthermore, what is meant by "human networks" are those in which actors interact with little to no cost associated with that interaction. We might think of these as the standard types of social networks, either existing as face-to-face relationships, or those existing in online social networks – such as Facebook or Twitter. There are many human interactions for which tie formation may be difficult or costly. Consider – for example – terrorist or criminal networks that are always working to hide their affiliations. It is much less likely that these types of networks would form through graph motifs. That said, the primary motivation for this work is to contribute to the modeling of a subset of networks relevant in the social sciences; therefore, while this is certainly a limitation of the GMM framework it does not discount its value.

In its current form the growth and termination rules are exogenous primitives, which remain static for a model. Conceptually, this is useful because it simplifies the construction of a model and provides a basis for interpretation of the results. In some case, however, it may be useful to endogenize these rules given the initial state of the base structure. Consider the case where the base structure is unknown to the modeler at the outset. Here, we may want rules that emerge as the result of this structure given the context of our modeling task, in which case endogenous rules generation will be necessary. Furthermore, the notion of "learning rules," is a potential extension; whereby, rules dynamically alter given the evolution of the network through the iterations of the model. These adaptations, however, make model interpretation more difficult—as is the case in techniques such as neural networks—and consideration therein must be made before proceeding.

Beyond these theoretical limitations, there are also some technical. A linchpin of the model is the need to count subgraph isomorphism in order to form beliefs about future network structure. As stated, this problem is known to be NP-complete, which means the method scales very poorly as either τ or the complexity of the base structure increase. As such, in practice both of these model parameters must be relatively small in order for the method to compute results in a reasonable amount of time. Improving the speed of the VF-2 algorithm is a Computer Science problem, and one that I am not qualified to address. With current technology, however, there are methods for improving runtime as the networks scale. First, rather than recomputing the probability distribution at every iteration this could remain static, meaning that subgraph isomorphism would only need to calculated once. Additionally, these counts could easily be done in parallel in a high-performance computing environment. Future version of the `gmm` package will allow for such distributed computing.

Additional improvements need to be made to future versions of the software. Better accounting for simulation statistics need to be made, including runtime, growth metrics, probability mass convergence and iterative changes to the base network. For example, given the path dependence of the model, it would be very useful to have some knowledge of the distribution of motifs used in a given simulation in order to compare and interpret results from multiple runs of the same model. This will allow for more precise comparison among various GMM specifications.

Finally, utilizing the ERGM literature, considerations for the quality of “model fit” within the context of the GMM must be made. A large advantage of ERGM models is the ability to compare model fitness. Likewise, a conspicuous omission from this research is a direct comparison between the GMM specified for the co-authorship network above and a comparable ERGM. In order to more fully understand how these two modeling techniques differ direct comparisons must be made across a large class of networks. This type of research, therefore, will constitute a large portion of the future effort in the work.

In this paper a co-authorship network within a Political Science sub-discipline was used to illustrate how a graph motif model could be used to model the growth of a network over time. Clearly, however, the potential for this method goes well beyond simple models of scholarly collaboration. In my own work, studying the dynamics of covert and illicit social networks is confounded by a lack of data (Sandler and Enders, 2004). With the above method it is possible to theorize network models with a limited amount of information and study how these theoretical models change. In populations where more direct sampling can be done; such as sexual contact (Hamilton et al., 2008) or drug-user networks (Weeks et al., 2002), and the task may be to uncover hidden populations this method may allow for the testing of a broader set of theoretical growth dynamics (Gile and Handcock, 2010). Finally, within the context of economic models of networks a cooperative game theoretic approach has often been used (Jackson, 2008). With GMM, however, it may be possible to incorporate non-cooperative decision criterion, and thus opening the possibility of studying an entirely different class of network games.

Using graph motifs to model networks represents an alternative method with a broad potential for application. By utilizing computational methods to specify these models the technique has inherent flexibility, which is useful when attempting to model the naturally complex nature of human social interactions.

Chapter 3

Methods for Collecting Large-scale Non-expert Text Coding

Much of the data used in empirical political science is not directly measured. This is because there are rarely any straightforward – or standard – ways to directly measure many of the phenomena of interest in the discipline. As such, these phenomena must be quantified in indirect ways. One of the most common methods for this is through the encoding of information gleaned from text.

Examples abound from all sub-fields in political science: casualty statistics from a battle description, categorization of a news stories, or the policy content of a politician's statements. In all of these cases, it is nearly always the case that the task of encoding the text is given to one, or a limited number of expert coders; or more recently, machine automated methods of encoding are used. Given that so much of the data within political science are generated via an intermediated encoding mechanism, it follows that there may be ample opportunity for innovation with respect to the methods used to encode text.

In this paper I explore the viability of using a large number of non-expert human coders as an alternative method for coding text. To do this Amazon's Mechanical Turk platform is used to crowd-source the coding from an on-demand labor pool of non-experts. Crowd-sourcing platforms have been used to generate data across many academic disciplines, and with these experiments I hope to highlight its applicability and value to political science. The paper proceeds as follows: first; by way of example, I provide a review of current methods for coding text with the discipline. This includes a discussion of several shortcomings of these methods observed in the literature. Next, an introduction to crowd-sourcing is given, with specific emphasis on Mechanical Turk. Examples of previous work done on that platform are provided. The remainder of the paper describes the design and implementation of a series of experiments conducted to assess the viability of using MT for coding political text, and how variations in the mechanism used on that platform affects results. This includes a discussion of the experimental design, technical implementation, results, and conclusions.

3.1 Coding political text

There are various types of qualitative data that may be subject to quantitative encoding by political scientists. For example, one may want to rate a photo on an emotional scale, or encode a social network by observing individuals' interactions. These endeavors require careful and unique coding methods, but are not the focus of this research. Here, I am interested in coding text. The vast majority of data, both historic and current, are recorded as text. In addition, the confluence of technologies on the Internet; such as e-mail, social media, with modern distributed storage and data processing tools has increased the scale of text data available to social scientists for analysis exponentially.

Access to text data is not a problem, but rather the increased scale of its availability poses a challenge to researchers with limited resources to analyze it. Researchers must match their coding method to the dataset of interest, while at the same time optimizing this decision based on costs, time, and methodological preference. At present, the methods employed by political scientists to perform text encoding can roughly be divided into two modes: human encoding, or automated machine encoding.

3.1.1 Human Coding

Human encoding is, for all intents and purposes, the classic mode of political text coding within the discipline. The most common data generation process in this mode involves employing one or more "experts", typically in the form of research assistants – graduate or undergraduates students. This pool of labor is then provided with some codebook, or rubric, for how the text are to be encoded. They are then provided with the text documents and begin the task. In most cases this process happens once, and the resulting data are included in the sample.

As this is the classic method of text encoding, there are countless examples from the literatures. Well known examples include the Policy Agendas Project [1], the Comparative Par-

[1] http://www.policyagendas.org/

liamentary Democracy Data Archive [2], and the Militarized Interstate Dispute data [3]. In all of these projects a detailed codebook and schema are used to attempt to ensure consistency among coders.

Perhaps the most long-standing, and widely cited, data set based on this method in political science is the Comparative Manifestos Project (CMP). At the time of writing, this dataset includes 3,611 party manifestos, from 905 parties, across 55 countries. The manifestos are encoded using a detailed handbook, and coders must be trained and tested before their data are accepted into the project. Briefly, the coding method for CMP proceeds as follows: humans "unitize" each manifesto by breaking the text into quasi-sentences. Human coders then assign one of seven policy domains and one of fifty-six policy categories to each unit. Percentage totals of categories are used to estimate the policy scores for each manifesto. This process is is done for each document once, and the summation of these codings becomes the CMP.

The contribution of the CMP cannot be understated, as it is one of the most valuable resources in existence to the comparative politics sub-discipline. In fact, the experiments that follow use raw manifesto data pulled directly from the CMP. It is, however, not without problems. Given its prominence in the discipline, there has been a significant amount of research done to examine the validity and reliability of the baseline data itself, and the resulting encodings. This research has uncovered non-systematic errors in encoding (Klingemann et al., 2006; Benoit et al., 2007 2009), misclassification of text (Mikhaylov et al., 2012), and the inclusion of non-manifesto documents into the corpus of data (Gemenis, 2012). All of these findings raise serious concerns given that so much research is empirically based on the CMP.

It is also important to note that because the CMP has such a high-profile in the discipline

[2] `http://www.erdda.se/index.php/projects/cpd`
[3] `http://www.correlatesofwar.org/`

it attracts the most scrutiny. These issues of bias, error, and misclassification are part of many – if not all – research based on human encoded text documents. Human are (quite) fallible. In fact, large-scale hand-coding projects of this kind are now much less common due to the salience of coding issues raised by those examining the CMP and other large-scale human coding projects.

In recent years, more technically sophisticated methods of text coding have been developed that attempt to address some of the issues arising from hand coding. These algorithmic methods are designed to overcome the issue of reliability, non-systematic bias, and data volume that dog human coding methods.

3.1.2 Automated Machine Coding

As the previously cited research has shown, humans are unreliable coders of text documents that require a moderate degree of interpretation. Part of the reason for this is that human coded data are typically produced by experts, and by definition experts are biased. Having long historical context, or strongly held opinions – features considered desirable for experts – may bias them to interpret an otherwise innocuous sentence as having some policy relevance that a different expert, or non-expert, would either ignore or interpret differently.

Furthermore, as the volume of text being coded by experts increases these errors become compounded. Increasing volume presents additional problems for coding capacity. Even in a world where human coders did not inject bias and non-systematic errors, it is simply impossible to expect human coders to handle the task when the amount of text data needing coding is measured in terabytes.

In an attempt to address issues of reliability, and increase the capacity of text possible to encode, researchers have turned to automated, algorithmic, machine coding methods. The contrast is quite clear, both from a resource constraint and reliability perspective. Rather than having to limit the amount of text coded based on budget constraints for hiring research assistants or time limitations, machine coding is constrained only be the vastly cheaper resource of computing power. Likewise, because algorithmic coding methods vary in implementation between deterministic and probabilistic, the coding does not suffer any non-systematic bias or errors.

Examples of automated machine coding include using text mining techniques to extract policy position (Laver and Benoit, 2003), differentiation between political party's policy positions (Kidd, 2008), categorical assignment of text documents (Hopkins and King, 2010), and time-series of international conflict (King and Lowe, 2003) and party position (Slapin and Proksch, 2008). The results of this research highlight the value of these methods when coding reliability is paramount. Unlike human coders, properly implemented machine coding methods will always code identical pieces of text the same way. Also, since the volume of coding a machine is capable of handling is constrained only by bandwidth and processing power, these methods scale much more reasonably with the pace text is being generated.

Among the best known, and most widely cited, examples of high-volume automated machine text coding is the Penn State Events Data Project (PSED), formerly the Kansas Events Data System (Gerner et al., 1994). PSED uses a dictionary-based method to identify and extract entities from a stream of new articles. These dictionaries contain a very large number of proper nouns and phrases that allow the system to match these dictionary items in text and code them as event data. Because this method relies so heavily on the dictionary itself, additional research has been done attempting to sharpen and formalize methods for defining the dictionary's content (Schrodt and Gerner, 2012, Chap. 2).

The reliability gained by using algorithmic methods, however, is often counterbalanced by losses in validity. While most of the machine coding methods used have been shown to code text at least as "good" as human coders, the fact is to reach these comparable levels a great deal of effort must be exerted to fine-tune the algorithm. In the case of PSED, the project explicitly states that researchers are encouraged to generate their own dictionaries when using the software for their own research[4].

Taken to an extreme, this fine-tuning can lead to over-fitting, which in turn can cause errors of false-positive or -negative coding. In a less extreme case this practice of fine-tuning can limit the generalization of these methods, making them less appealing to researchers whom either lack the requisite volume of text to code, or the acumen in algorithm design.

Most recently researchers have sought to combine algorithmic, or statistical, methods with human coders as a step-wise approach to generating data (Simon and Xenos, 2004; Grimmer, 2011; Ahlquist and Breunig, 2012). The benefits of combining methods is clear: the algorithmic approach provides a reliable and unbiased method for minimizing the set of possible codes or categories a given piece of text could belong to. Then, human coders can be used to verify or recode all or some of the text that a machine has processed. This approach provides the benefits of fully-automated coding, but also optimizes for human time when humans need to be in the loop.

A popular coding method that operates at this intersection is the generation of topics from text using Latent Dirichlet Allocation (LDA)(Blei et al., 2003). Recently, several researchers have been experimenting with LDA for political text coding (Monroe et al., 2008; Quinn et al.,

[4] `http://eventdata.psu.edu/data.html`

2010; Grimmer, 2010). Briefly, the method, also known as topic modeling, probabilistically generates a set of discrete topics from a text corpus. This is useful for coding political text because these topics can be used as categories, or the words in each topic can be used in a dictionary. Unfortunately, due to the probabilistic nature of the method, topics can vary significantly from each run of the model. This has lead to several new lines of research that attempt to adapt the basic LDA model to perform more reliably (Ramage et al., 2009; Titov and McDonald, 2008; Hospedales et al., 2011).

All of these methods: purely human coded, fully-automated machine coded, or a statistical or hybrid method, present advantages and disadvantages to researchers needing qualitative text data coded into quantitative units. In the following sections I describe an alternative method for coding political text data by crowd-sourcing non-experts using Amazon's Mechanical Turk (MT) platform (`http://www.mturk.com`). Before proceeding to a discussion of the design and implementation of the experiments, I first introduce the technology of crowd-sourcing.

3.2 Crowd-sourcing

The notion of the "wisdom of crowds" has deeply penetrated the popular consciousness. With the success of Wikipedia, and the popularity of mainstream books by Surowiecki, Gladwell, and Shirky on the subject, the idea that high-quality output can be generated by the effort of many people is well known. Here the term "crowd-sourcing" refer to the idea of collective intelligence, or that through a natural populist mechanism the best, or most accurate, results or answer emerge from a large pool of labor.

While this type of crowd-sourcing is well known, this context is too general to describe how it is used to code data here. Crowd-sourcing platforms, like MT, are not designed to collectively produce some tome, or generate an eveolving document that is constantly iterated over. Instead, MT is designed for individuals to request work on finite – and often very small – tasks that require human intelligence to complete. The final feature, that human intelligence in required, is very important.

Computers are quite good at performing small (or large) finite tasks, such as arithmetic, recalling a record, or manipulating data. Despite our best efforts, computers still perform much worse than humans on tasks that require interpretation, context, or human vision. One classic example is determining whether a picture is of a cat or a dog. Most toddlers would have no problem telling you if a picture contained a dog or a car, or both, or many. Computers, however, are not very good at this.

This is exactly this type of task MT was designed to crowd-source: small tasks that humans excel over computers at, and can be easily distributed at large scale. In fact, the primary unit in MT is called a HIT, or "human intelligence task". Crowd-sourcing in this context is thus the collection of completed human intelligence tasks, which themselves are specific and finite. For this paper I work with this narrower definition of crowd-sourcing.

Though there are clearly limitations to the types of tasks that can be done on MT; for example, dissertation writing – the bounds are quite large. Tasks posted to MT range from purely commercial endeavors, to academic research, and many things in-between. Common short-form tasks include classifying web sites, performing web searches, or writing brief reviews or summaries; and common long-form tasks include, audio and video transcription, writing essays (500 words or less), submitting original graphic designs.

Many of these examples have clear commercial applications. For example, one may want to classify websites in order to build a filter to remove adult content or offensive content. Likewise, one could generate a massive amount of feedback on a product very quickly by asking MT workers to write short reviews. There has also been an increasing amount of academic work conducted on MT already. Given that the platform as a tool is more familiar to disciplines technical disciplines, such as computer science and information systems, a large amount of research done on MT comes from these fields. Very often researchers who study fields such as machine learning, computer vision, or natural language processing will use MT to build datasets that can be used to inform their models of human intelligence (Sorokin and Forsyth, 2008; Snow et al., 2008; Callison-Burch, 2009; Kittur et al., 2008).

This work by researchers in other disciplines has led to many general findings on MT that are valuable across all disciplines. First, due to the open nature of the MT labor pool, it is critically important that researchers design tasks such that they can not be easily abused or SPAM'ed[5] by workers (Ipeirotis et al., 2010). Second, the labor pool participating in MT tasks is much more representative of the general population than the typical convenience sample used in academic experiments (Ross et al., 2010). Finally, workers are highly responsive to variation in task length and compensations, i.e., short/high-value tasks are much more desirable than long/low-value tasks (Buhrmester et al., 2011).

Recently, social scientists have turned to MT as an alternative means for running experiments. These experiments have ranged from specific empirical implementations of theoretical work (Rand, 2012; Sprouse, 2011), to general assessments of the viability of MT for running social and behavioral experiments (Mason and Suri, 2012; Paolacci et al., 2010). Very recently work was done to show that many classic examples of experimental work from political science

[5] SPAM often refers to unsolicited email, but can also refer to any unwanted or unsolicited response or submission.

could be replicated using MT (Berinsky et al., 2012).

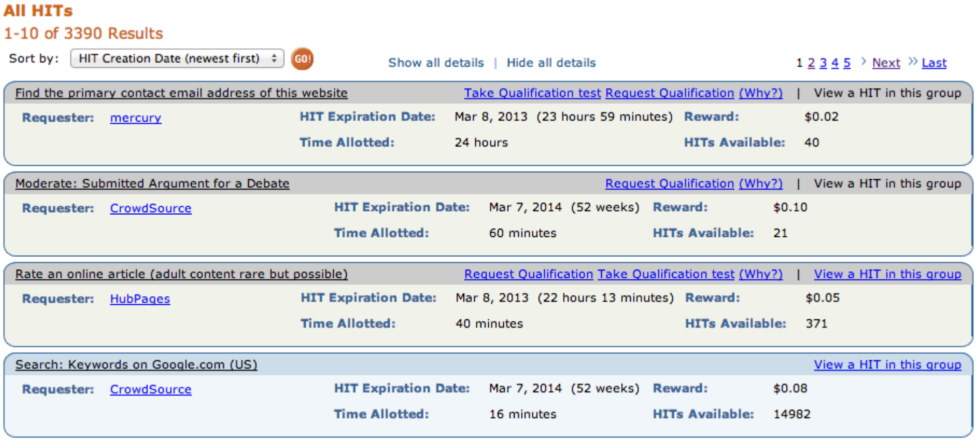
3.2.1 Mechanical Turk

The MT platform itself is setup as a simple web-based interface, and is part of Amazon's larger set of web-services. One can interact with the service via one of two modes: as a requester, or a worker. Workers browser a list of HITs available to participate in. These HITs can be ordered by several dimensions, including reward amount, creation date, or time alloted to complete.

Some HITs may be unavailable to workers because they do not meet the qualification requirements. Requesters specify these requirements when posting HITs, and they can range from age minimums (used often for jobs that may contain adult material), location restrictions, or the completion of a qualification test. The concept of a qualification test requirement is central to the experiments described in later sections, and I will return to this later in greater detail.

While browsing the list of available HITS a worker can preview a HIT. Once inside the preview, a worker can then accept HIT to work on it, or not and return back to the list of available HITs. Once work has been submitted it must then be approved by the requester before the worker is compensated. Figure 3.1 is an example of what the HIT interface looks like for workers. As you can see, information on the requester, the compensation level, time allotted, number of HITs available, etc., can all be seen from this interface. Figure 3.1 also show that the top three jobs in this case were unavailable to the author because I did not meet the qualification requirements.

For the purposes of these experiments I interact only as a requester; however, it proved quite practical and informative to operate as a worker at the outset of this research in order


All HITs
1-10 of 3390 Results
Sort by: HIT Creation Date (newest first) GO!
Show all details | Hide all details
1 2 3 4 5 › Next » Last

Find the primary contact email address of this website — Take Qualification test Request Qualification (Why?) | View a HIT in this group
Requester: mercury — HIT Expiration Date: Mar 8, 2013 (23 hours 59 minutes) — Reward: $0.02
Time Allotted: 24 hours — HITs Available: 40

Moderate: Submitted Argument for a Debate — Request Qualification (Why?) | View a HIT in this group
Requester: CrowdSource — HIT Expiration Date: Mar 7, 2014 (52 weeks) — Reward: $0.10
Time Allotted: 60 minutes — HITs Available: 21

Rate an online article (adult content rare but possible) — Request Qualification Take Qualification test (Why?) | View a HIT in this group
Requester: HubPages — HIT Expiration Date: Mar 8, 2013 (22 hours 13 minutes) — Reward: $0.05
Time Allotted: 40 minutes — HITs Available: 371

Search: Keywords on Google.com (US) — View a HIT in this group
Requester: CrowdSource — HIT Expiration Date: Mar 7, 2014 (52 weeks) — Reward: $0.08
Time Allotted: 16 minutes — HITs Available: 14982

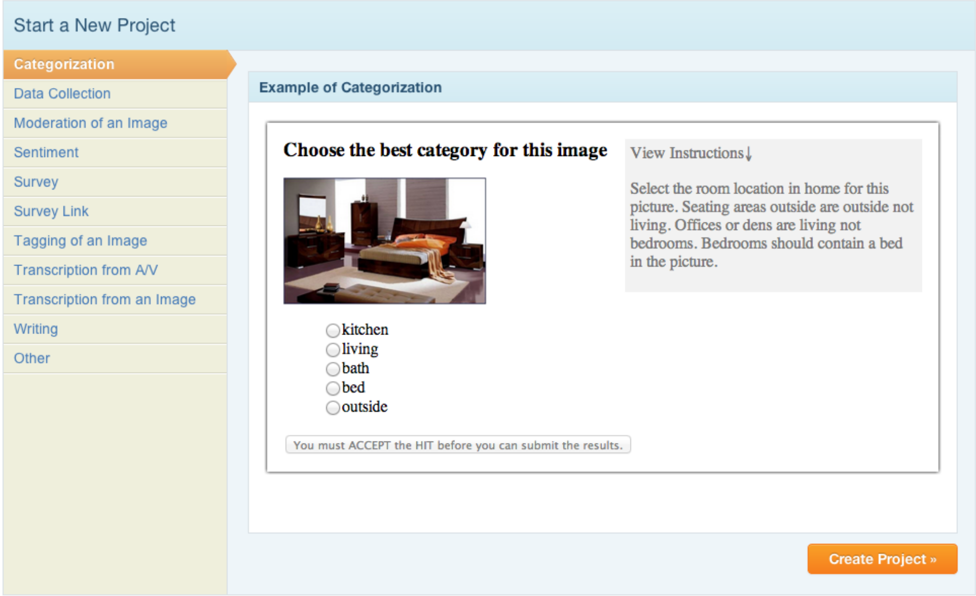
Figure 3.1: Example of HIT Interface for MT Workers

to become more familiar with the system, and standard rapacities within it. Requesters have two primary modes of interacting with the platform: through MT's web-based HIT creation tools, or directly through the application programming interface (API) [6].

The web-based tools provide a useful set of templates and standard formats for creating HITs. Anyone who has every worked with a web-based tool for creating content will have familiarity with this interface. As you can see in Figure 3.2, HIT templates are based on the type of data one may want to collect. Whether it is categorization, image moderation, or a survey, MT provides requesters with a set of templates based on common MT use cases.

Once a task has been created using the web interface, and the HIT has been posted, requesters can manage and track their HITs via additional web-based interfaces. Here a requesters can extend a HIT to get more responses, or disable a HIT if something goes wrong or it is not longer needed. This is also the interface through which requesters can evaluate, and approve, responses. Once a response has been approved, requesters can then download

[6] It should also be noted that Amazon provides both a "production" and "sandbox" version of these interfaces. The sandbox version is used for building and testing HITs before exposing them to real MT workers, and having to compensate them for approved work.



Choose the best category for this image

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

○ kitchen
○ living
○ bath
○ bed
○ outside

You must ACCEPT the HIT before you can submit the results.

Figure 3.2: Example of HIT Creation Interface for MT Requesters

the results to collect the data.

The web-based requester interface is useful for those new to MT, or when a project is relatively small and will not require any iteration. To develop and deploy the experiments described in the following sections I did not use the web-based interface. The experiment described here are of a much larger-scale, which the web-based creation tools are not well suited to support. Instead, custom software was written to interface directly with the MT API.

To create and manage HITs via the API, a requester must be familiar with MT's XML specification for communicating with the API, and have a preferred means of interacting with the API[7]. There are many ways for requesters to interact with the API, and I provide a detailed accounting of the method used for these experiments in Section 3.4.

[7] Documentation on MT API can be found online here `http://awsdocs.s3.amazonaws.com/MechTurk/latest/amt-API.pdf`

In the following section I describe a series of experiments to test the viability of MT for collecting political text coding, and how variations in the mechanism affect the quality of output. Informed by the results of much of the previous research cited here, these experiments have been designed to maximize the quality of responses and incentive workers to complete a large amount of coding in a relatively short amount of time. This begins with a description of the text data set used in these experiments, and how its baseline expert coding was collected. This is followed by a discussion of the technical work-flow used to implement the experiments, including hardware and software components. Finally, a detailed discussion of the experimental design is provided.

3.3 Political Text Coding Experiments

This work exists as part of a larger research project on how different coder types, and coding treatments affect the quality of political text coding (Benoit et al., 2012). The purpose of these experiments is to test whether the collective coding output of political texts from a pool of non-experts compares to the coding of experts. With expert coding our *a priori* assumption is that by virtue of being "experts" the quality of their codings will meet some minimum level. As has been mentioned several times before there are problems with expert codings, however, as a discipline we still believe that a single expert coder is capable of producing valuable output.

With non-expert, however, coders our initial belief is exactly the opposite. Our expectation is that the output of an individual non-expert will be inferior to that of an expert. By the same logic that we believe experts produce quality codings, we therefore believe that non-experts produce lower quality. By definition, experts understand what they are reading and are trained to recognize words or phrase that match some coding schema. Non-experts do not.

Within the context of crowd-sourcing, however, we are not limited to a single non-expert. In fact, our access to the labor pool is limited only by our economic resources, and at the scale of only a few pennies a task this limit is quite high. The the notion of "collective coding," therefore, is pivotal to the hypothesis. That is, if many non-experts code the same piece of text, will the consensus – or modal – coding be as good as a single expert? Moreover, if these individual non-expert codings are then aggregated to the document level, is the modal coding of the whole document also on par with expert coding?

From a resource perspective, the benefits of using crowd-sourced non-experts to perform tasks such as coding are quite clear. Again, given the expected compensation levels in the MT market, a massive amount of work can be completed for very little resources. Also, the time to completion is only a fraction of the usual expected time frame for large-scale coding tasks. In this case, over 48,000 individual text units were coded over the span of a few weeks, for less than $1,500 USD.

Theses benefits, however, have little practical value if the collective output of non-experts does not compare to that of experts. As such, the primary focus of these experiments is to examine how this trade-off between experts and non-experts can be mitigated by careful design of the collection mechanism used for non-experts.

As has been mentioned previously, MT provides requesters the ability to filter workers through the incorporation of qualification requirements. Requesters are highly motivated to design coding task such that only the best coders are permitted to submit work. This is akin to research faculty hand-picking graduate students to work as research assistants on a traditional coding project. For these experiments, variations in the qualification test are used to measure how altering the collection mechanism can affect the quality non-experts' collective coding.

To design these experiments several key components must be addressed. First, a corpus of text data must be specified, and that text must be sampled in a way that can be coded by both experts and non-experts. Next, a general purpose coding schema must be developed such that it can be used to collect codings from both experts and non-experts. Finally, a well-designed mechanism must be developed to collect codings from non-experts.

3.3.1 Manifesto Text Data

The text data used in these experiments is drawn from party manifestos. The texts are sentences from the British Conservative, Labour and Liberal Democrat manifestos issued in 1987 and 1997 (Benoit et al., 2012). These data are chosen for two reasons. First, these text are readily available as part of the CMP; and second, as function of this inclusion these text have already been widely studied (Laver and Hunt, 1992; Laver, 1998). One notable observation is the significant shift to the center by the Labor party over the time period. The presence of this shift provides an additional check on the viability of crowd-codings.

To begin, these manifestos must be distilled into single units of text such that they can be coded by both experts and non-experts alike. Experts are capable, and quite comfortable, with coding large pieces of text. As has been discussed, this is the primary mechanism by which the CMP is coded. To explore how coding works on MT, however, the unit must fit into the "small and finite" framework MT is designed to support and MT workers are accustomed to working in.

As mentioned, the CMP unitizes manifesto text into "quasi-sentences," which are determined by the coders themselves and are meant to represent a unit of text that naturally expresses some policy position. These sentences can consist of whole sentences, parts of sentences, or arbitrary concatenations of sentences and phrases. While these units are concise enough to work within the MT framework, there are problems with this method for our purposes. Besides the awkward breaks, this method of *unitization* is endogenous to the coding process, which for our purposes may insert bias into the coding before coding has even begun.

The units of text drawn from the manifestos for these experiments are natural sentences units [8]. Along with avoiding the endogeneity presented by quasi-sentences, using natural sentences allows for machine automated text dissection of the manifestos into single units. A simple script was developed to generate the single text units for these experiments. Using this method the total set of sentences available for coding is 5,444; again, drawn from six manifestos, across three parties, a decade apart.

Before coding can begin, a sequence and schema must be specified that can be used for both expert and non-expert coders. It is crucial that these work for both types of coders in order to ensure that the resulting codings are comparable.

3.3.2 Coding Sequence and Schema

The sequence by which coders read text and perform coding has significant impact on how text are coded. When a coder is reading an entire manifesto from start to finish something read at the outset can influence codings later in the document. Also, if coders are asked to code only single text units taken completely out of context important information may be lost. For this reason, the sequence of sentence units delivered to coders in these experiments attempts

[8] A "sentence unit" here does not necessarily mean a grammatically concise sentence. Because party manifestos often include bulleted lists, or long enumeration of policy positions, a sentence unit here can be divided by various punctuation.

to compromise these issues.

Sentence units are coded out of sequence from how they appear in the manifestos. Coders are asked to read a chunk of text draw randomly from the corpus of sentence units. This chunk contains the sentence unit being coded, and two preceding and proceeding sentence units appearing in sequence from the manifestos. In practice, the sentence unit being coded is highlighted in red, and the context sentence units are in black.

> ALTERNATIVES TO PRISON. Every effort should be made to ensure that fine defaulters, elderly shoplifters and drunks are not sent to prison. Police cautions and intermediate treatment should be more widely used. Where punishment is appropriate, it should normally be community service rather than prison.

Figure 3.3: Example coding unit.

Providing the context is critical to informing the coder as to the possible policy statements being made by a given sentence unit, but does not appreciably increase the effort required to complete the task. The latter point is of utmost importance on MT, where workers guard their time actively. Figure 3.3 is a replication of how a coding unit would appear to a coder[9]. The sentence being coded is highlighted in red, and the instructions explicitly tell coders to code only the red sentence, using the surrounding sentence units only for context. A copy of the coding instructions given to non-experts is provided in Appendix E.

With the delivery mechanism for sentence units specified, the next step is to develop a coding schema. For these experiments a simple coding schema is used. Unlike the CMP, which uses a 56-category policy handbook, this design uses only three categories. This significant simplification is used for two reasons. First, the much larger specification used by the CMP

[9] The formatting is slightly different here compared to how it appears inside MT due to printing requirements. See Figure 21 in Appendix D for a screen shots the actual HIT interface.

has been shown to be a major source of discrepancy when multiple coders have been asked to code the same manifestos (Mikhaylov et al., 2012). It follows that non-experts would suffer at least as much as experts in this case; therefore, a simpler specification is preferable. Likewise, for these experiments my primary interest is the design of the coding method – and how that affects results – not the coding itself. It is important that the method be applicable to any type of coding task, not only policy positions.

The second reason extends from the requirement that both the experts and non-experts use the same coding schema. It is quite impractical to attempt to provide coding instructions to non-experts on MT for a 10-category schema, let alone 56. Again, MT is well-suited for simple and finite tasks that require human intelligence. In order to develop a schema that would work on MT, and also allow for a reasonable level of instruction for non-experts, three high-level policy categories were specified: social, economic, or neither.

Coders are also asked to code each sentence unit on a policy scale. The policy scale is area dependent. The economic policy ranges on a five-point scale from very liberal to very conservative. Likewise, the social policy ranges on separate five-point scale from very left to very right. Making this distinction allows the coding instruction to be more explicit about what is an economic policy, and how to find one, vice a social policy, or something that contained no statements regarding either of those policy areas.

Figure 3.4 is a visual representation of this coding schema. The first level of the hierarchy is the policy area, which have corresponding numeric values denoted below. The next level down the hierarchy are the policy scales, which also have a corresponding numeric values. These numeric values are what are used to quantify the codings, and represent the primary data generated. Given this schema, each sentence coding produces a numeric pair that corresponds

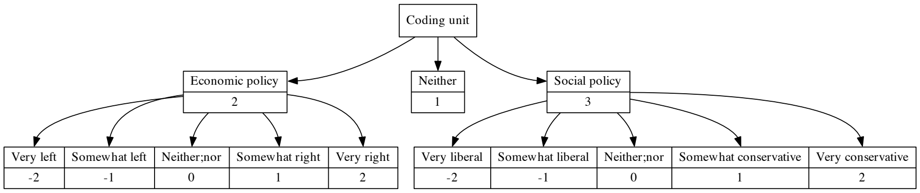


Figure 3.4: Visual Representation of Simplified Coding Schema

to the policy area and scale respectively. For example, a sentence that is coded as denoting a "somewhat conservative economic policy" will have a corresponding code pair of $(2, 1)$ in the dataset: "Economic" → "Somewhat conservative".

With the coding schema specified, the first step is to use it to collect expert codings. In the following section I describe how these expert codings were collected and aggregated.

3.3.3 Baseline Expert Coding

In order to accurately test the viability of crowd-sourcing for political text coding there must be a baseline measure of the "correct" coding for each sentence unit. Once this coding has been determined, it is possible to measure how well the non-expert coders performed. For these experiments this baseline coding was achieved by having multiple experts code the entire corpus. In this case, those experts are three faculty members and three advanced graduate students in political science. All of this coding was done as part of the larger research agenda on political text coding (Benoit et al., 2012), and I am relying on the data generated by that work as a baseline for my experiments.

As has been mentioned many times, using only a single coder has been shown create errors in codings. Also, multiple coders often do not agree, or reach consensus on the coding for a given unit of text. These issues create problems for our experiments, as we need a single coding for each unit to compare with the crowd-codings. To resolve this the notion of a "gold" coding is introduced[10]. A sentence unit is said to be "gold" when a simple majority of expert coders have consensus on the policy area.

This is a very inclusive definition, and is used to maximize the number of sentence units that are available for coding by non-experts. Of the 5,444 total sentence units 4,403 qualified as "gold" sentences by this definition. The expert coders reached consensus on 1,650 economic policy sentences, 507 social policy sentences, and 2,246 that conveyed neither policy area. In Table 1 counts of the sentence units' expert coding are listed by political party and the manifesto year. As you can see from the table, there is a relatively even distribution of sentences among parties, with a slight bias toward sentences from the 1997 manifestos. Manifestos from this year were all longer, so this bias is expected.

	Economic	Social	Neither	Total
Conservative	794	224	752	1,770
Labour	409	108	709	1,226
Liberal Democrats	447	175	785	1,407
1987	885	206	746	1,837
1997	765	301	1,500	2,566

Table 3.1: Counts of "gold" sentence units by party and year

You will note that the policy scale is not used as part of the definition. This decision was made for two reasons. First, to include this in the definition a determination as to the appropriate level of aggregation for each five-point scale would have to be made. That is, would it be most appropriate to leave it as a five-point scale, or define consensus as a majority left/neither/right or liberal/neither/conservative? While this is certainly possible, it is not

[10] The "gold" terminology is borrowed from the crowd-sourcing literature, and is not necessarily meant to denote quality, only conformity.

clear how doing so increases our understanding of the ability of non-expert coders to recognize policy statements in text. Second, as a practical matter, further refining the definition of "gold" significantly reduces the number of sentences available for coding by non-experts. Of the 507 social policy "gold" sentences, only 81 would qualify as gold by requiring the three-point aggregation of policy scale as part of the definition. Interestingly, this provides further evidence as the the inability of experts to reach consensus.

For these reasons, the initial focus for the crowd-sourcing experiments is on non-experts' ability to correctly identify the policy area of a sentence drawn from these subset of "gold" sentences. In the following section a detailed description of the experimental design on the crowd-sourcing experiments in provided. This includes a discussion of their technical implementation.

3.3.4 Experimental Design and Implementation

With a large-scale online experiment of this kind, the design has two distinct aspects. First, the experimental manipulations must be specified and designed. Second, because these experiments require broad integration of various technical platforms, great care must also be taken with the technical design and implementation of the experiments. I begin with the design of the experiments itself, and its manipulations.

First, it is important to consider how the labor pool on a platform like MT differs from participants in a typical university data collection task. In most cases, the data collection effort would be driven by the researcher themselves. That is to say, the researcher conducting the data collection has total control over how the subjects, and how they interact with the collection effort. On MT those dynamics are reversed.

Crowd-sourcing platforms are driven by the labor pool, since without them there would be no platform at all. This reversal presents interesting challenges for researchers attempting to use it as a means of data collection. First, MT workers have a very clear economic incentive to optimize their decision for accepting HITs based on their perceived difficulty; or time consumption, and how that related to their advertised compensation level. The more high-value HITs an MT worker can complete in the shortest amount of time the more money that worker will make. This has significant impact on how the tasks should be designed, how many sentences they are asked to read and code, and how much they are compensated.

Given these economic incentives, workers may also be motivated to "game" a HIT if such a strategy would result in the successful completion of a HIT in a shorter amount of time. Recall, workers are only compensated once a HIT has been approved by a requester. Workers, therefore, receive a constant feedback loop from the requesters as to what they want, or at least workers can interpret it this way. This impulse-response dynamic has implications for the data used here because of the significantly larger number of "gold" sentences coded as not having a policy area. These "Neiter-nor" sentences outnumber economic and social sentences nearly 2-to-1, therefore, it would be entirely possible that workers would rightly infer this from working through multiple HITs and begin to mis-classify based on this bias.

Also, because MT is a distributed web-based platform, the requester has no control over, how, when, or where workers are interacting with the task. All of the text in our corpus is English-language, however, most of the workers on MT do not come from native English speaking counties (Ross et al., 2010). Since workers comes from many different parts of the world, time zone difference can affect who is working on a task when. With respect to delivery of the HITs themselves, because operating systems, browsers, and Internet connections can vary significantly from worker-to-worker it is important that the actual formatting of each HIT be designed in the most standard possible way. This is not an issue when using MT's templates, but is in this case because the experiments used highly customized interfaces.

The experimental design laid out here attempts to address all of these issues. First, to find the optimal balance of sentence units-to-compensation several small trial HITs were deployed. These HITs, which contained various version of the full experimental design, were very valuable in helping guide the final HIT structure. Because they were used as initial tests, and did not include the full experimental design, data generated through them is not included in the results of this study[11]. Based on information gathered through this process the final HIT design includes six sentence units per HIT, with a total compensation of $0.18 per HIT. This breaks down to a $0.03 reward per-sentence, amazingly meager level of compensation when considered relative to traditional human coding projects. This number of sentences and level of compensation struck a balance between the workers' participation calculus observed in the trial experiments, and my desire to get the coding completed in a reasonable amount of time. Given the dynamics described above, and observations in the literature, increasing the compensation would likely have only decreased the time to completion rather than increased the overall quality.

To prevent the workers from coding sentences based on learning the bias toward "Neither" sentences present in the whole corpus, a uniform distribution of sentence types are used for these experiments. The "gold" sentences coded as "Social" by the experts is the smallest subset, containing only 506 sentences. To make the distribution uniform, a random sample

[11] There may be some concern for priming, or contamination of the experiment since workers participating in these initial HITs could also participate in the real experiments. Upon investigation, there were no workers that participated in both. This likely the result of simple good fortune, therefore, in the future it would be preferable to explicitly exclude these workers from participating in both phases of work. This is entirely possible within the MT framework.

of 506 sentences are drawn from the "Economic" and "Neither" subsets. The resulting set of 1,518 sentences becomes the corpus for the crowd-sourcing experiments. For each HIT, a random sample of six sentences are drawn from this corpus, and therefore in expectation each worker will be asked to code a uniform number of "gold" sentence units[12]

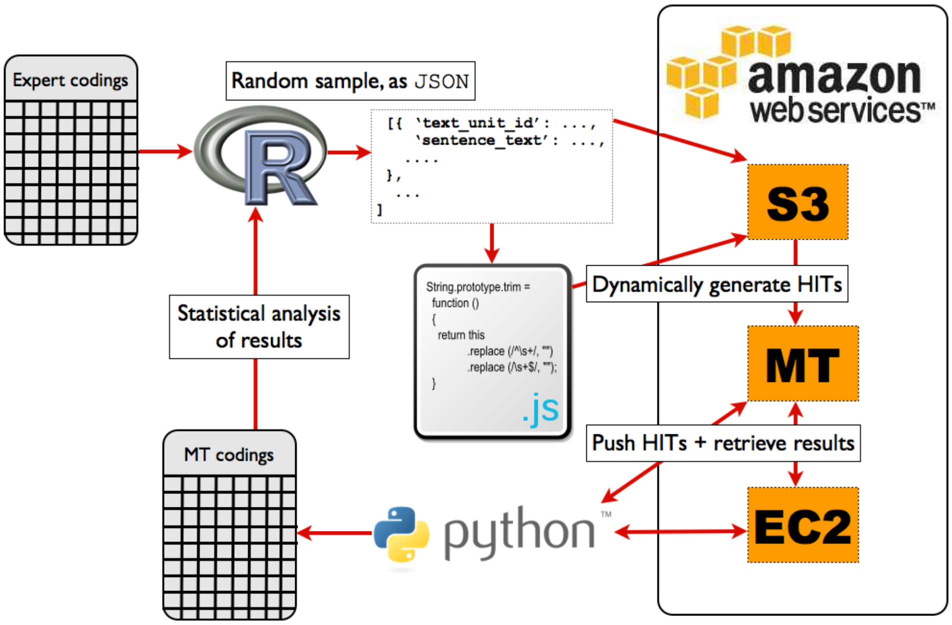
Also, to prevent a very small group of coders, or even a single of coder, from doing all of the coding, the number of HITs an individual MT worker can submit is limited to 50. The maximum number of sentence units an MT worker can code, therefore, is 300. Workers, of course, are free to code far fewer, and as is shown in the next section this is often the case. For each of the experiments, 30 HITs are issued resulting in a total of 1,500 submissions, or 9,000 sentence units coded per experiment.

With the basic experimental design set, the next step is implementation. To implement these experiments several tools from Amazon's Web Services (AWS) toolkit are used. AWS hosts a very large number of cloud-storage and cloud-computing services, along with secondary service like MT. First, the corpus of sentences are hosted on Amazon's Simple Storage Service (S3), which provides simple and scalable storage for the data. In order to make the interface for the HITs as customizable as possible I opted-out of using MT's HIT templates, choosing instead to host the HIT remotely. In this case, custom Javascript was written and hosted on S3 to dynamically render a new HIT ever time a worker requested one.

When a worker requests a HIT, the Javascript hosted on S3 is executed and a new set of sentences are dynamically pulled into the HIT. Workers then do the coding and submit their work, which is stored on the MT platform. Additional Python software was written to

[12] This does not, however, guarantee that each sentence unit is coded the same number of times. The reason for this is the result of a technical limitation of the system. In the future it would be preferable to design a system wherein this uniform distribution is guaranteed.

interact directly with the MT API[13]. This software was used to generate HITs, qualification tests (discussed below), and approve and download coding results.


Expert codings
Random sample, as JSON
amazon web services™
S3
Dynamically generate HITs
MT
Statistical analysis of results
Push HITs + retrieve results
MT codings
python™
EC2
.js

```
[{ 'text_unit_id': ...,
   'sentence_text': ...,
  ....
 },
  ...
]
```

```
String.prototype.trim =
 function ()
 {
  return this
        .replace (/^\s+/, "")
        .replace (/\s+$/, "");
 }
```

Figure 3.5: Diagram of Technology Work-flow for MT Experiments

Because I am interested in testing how well non-experts code political text based only on the design of these experiments, all submitted work is approved. As such, the software written to approve and download the work was hosted on AWS's cloud computing platform called Elastic Cloud Compute (EC2). Scripts would run many times a day to approve any new work submitted, and download the results. Providing the workers with this immediate feedback also improved participation, as MT workers are more likely to participate in jobs they can see are being actively monitored by requesters.

[13] I relied heavily on the open-source boto package for interacting with the MT API (https://github.com/boto/boto).

To promote uniformity of formatting and style, an industry standard style-sheet library was used for the design of all the HIT web pages[14]. This reduces issues of divergence in how workers interact with the interface across platforms, browsers, or Internet service providers. All of the software developed and used to support this research is available freely to inspect and download at `https://github.com/drewconway/mturk_coder_quality`.

Figure 3.5 illustrates the basic technical work-flow of generating, posting, and approving work for these experiments. This constitutes the basic setup of the experiment. What is missing, however, are design manipulations. In this setup, any worker is able to accept a HIT and code data. As has been noted, it is well known that quality control is an important feature to a well-designed MT task. In the following section I described the use of qualification tests to filter workers, and how these tests were manipulated as part of the experiment.

3.3.5 Qualification Tests

As mentioned, MT provides many different types of qualification requirements that requesters can use filter workers. Requesters can also design their own qualifications tests, which workers must take and pass before they are allowed to accept HIT requiring the test. As part of these experiments I wanted to investigate how variations in qualification requirements affected the ability of non-expert coders to match the coding of experts. That is, does a stricter qualification test increase the quality of coding output?

In this case, to create a qualification test that will filter for coding ability the test exactly replicates the coding task. Here the qualification test contains six sentences, uniformly drawn from the three policy area types. The qualification test contains exactly the same coding

[14] `http://twitter.github.com/bootstrap/`

instructions, and asks workers to code them. The difference being workers are instructed that the correct codings are known, and passing the test requires coding them correctly [15]. Figure 22 in Appendix D provides a screen shot of one question from a deployed qualification test.

I deployed two different qualification test, each of which represents its own experiment. Each contains six questions, but the first – "Low-Threshold" test required workers to correctly code 4-out-of-6 sentences correctly. The second test – "High-Threshold" – required workers to correctly code 5-out-of-6 sentences correctly. An individual worker could pass one, both, or neither test, and be eligible to code HITs associated with each test accordingly. The corpus of sentences drawn from for all experiment is the same, however, the sentences in each qualification test are exclusively different.

With the qualification test manipulations specified, there are a total of three experiments deployed to MT: no qualification, low-threshold, and high-threshold. In the following section I describe the results of these experiments.

3.4 Results

We begin by investigating MT coders ability to agree with the expert coders. A pair of coded sentences are said to "agree" when the consensus codings from MT for a given sentence unit matches the "gold" coding from the experts. A consensus requires a simple majority of coders with a minimum of three codings per sentence unit. From the data, the minimum number of codings for any sentence unit is 12, while the median is 31. As such, the minimum number of coders is always reached, however, consensus is not. For all experiments, non-expert coders

[15] To pass the test, workers must only match the "gold" coding, which applied to policy area. They are still asked to code for policy scale, but this is not used to adjudicate their responses.

reach a consensus 89% of the time. In order to contrast these results with the experts the remaining 167 sentence units were dropped.

	Results			Kappa Statistic		
Experiment	Sentences	# MT Coders	% Agree	k^*	Std. Error	z
No Qual.	1,315	89	0.65	0.47	0.13	22.6
Low-Threshold	1,393	56	0.70	0.54	0.12	26.7
High-Threshold	1,250	23	0.62	0.41	0.13	18.3

* A k value between 0.4-0.6 is considered “moderate” agreement

Table 3.2: Inter-rater Reliability Statistics Between Experts and MT Consensus Codings for all Experiments

Table 2 reports the inter-rater reliability statistics when comparing expert codings with the consensus MT codings, by experiment. The table also reports some descriptive statistics about each experiment; specifically, the total number of sentences with consensus codings, the number of coders that participating in each experiment, and the aggregate percent agreement between the expert and MT codings. The inter-rater reliability statistic reported here is the Cohen's kappa statistics (Cohen, 1960). This measure is particularly useful in this case because Cohen's kappa adjusts for the probability that the two raters – in this case the experts and the non-experts – agreed by chance. This is a concern if it is the case the coders; particularly in the no qualification experiment, were randomly coding without reading in order to quickly accrue compensation.

Here we get our first look at the level of participation in each experiment, and the overall performance of the crowd-coders. Starting with participation, it appears that the presence of a qualification test does reduce the number of workers participating in a task. Likewise, by making the test more stringent – in this case going from a Low- to High-Threshold test – participation is further reduced. More interesting, however, is how the presence of these tests seem to affect agreement.

Overall, the coders performed moderately well in all of the experiments. First, it is clear that non-experts are able to identify the policy contents of text and code accordingly. Given the distribution of "gold" sentence types used in all the experiments, if the non-experts were randomly coding the sentences we would expect an overall percent agreement of approximately 0.34. We see, however, the coders are doing much better than that, with percent agreement over 0.60 in all three experiments.

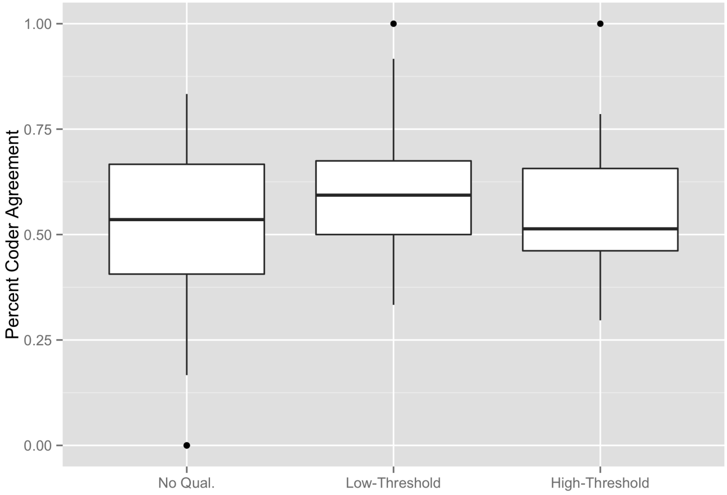


Figure 3.6: Coder Percent Agreement Distribution, by Experiment

In these experiments non-experts, however, fall well short of matching the expert codings. Examining the kappa scores provides a bit more insight here. For all experiments the non-experts match the expert coders "moderately" well. Given the amount of data, all of these estimates are statistically significant, so p-values are not reported. What is more interesting are the standard errors. Given the magnitude of the kappa estimates these standard errors are relatively high. This would indicate that there may be some systematic bias in the non-expert coding.

Perhaps most surprising is that there appears to be no significant affect in quality by increasing the qualification test threshold. Workers in the low-threshold test perform the best, but workers in the high-threshold test perform worse than those where no test is present. In Figure 3.6 the distribution of percent agreement for all non-expert coders is illustrated using box plots. Here the unit of analysis is the percent agreement of each coder in every experiment. We see that while the mean percent agreement in the high-threshold is actually slightly lower than that of the no qualification experiment, the distribution is skewed more heavily toward higher percent agreement.

This is reassuring, as it provides some evidence that the qualification test had a positive affect, relative to the lack of a test, on filtering for higher-quality coders. It is still the case, however, that coders in the low-threshold experiment performed better overall. Given this result, the evidence collected from these experiments suggests that the simple presence of a qualification, low-threshold or otherwise, provides as much of a filter as an increasingly difficult test.

Though he results are very encouraging as to the performance of non-experts, there appears to be some bias among the coders that is negatively affecting performance. In the following section I further examine the results to attempt to uncover how the source of this bias.

3.4.1 Non-expert Bias

A simple way to check for bias in the non-expert codings to tabulate the counts of consensus policy area codings across all experiments. We know that there is a uniform distribution of these sentence types in each experiment, so if the non-experts were coding with some bias that should manifest in these counts.

	No Qual.	Low-Threshold	High-Threshold
Economic	454	524	489
Social	734	774	710
Neither	127	95	51

Table 3.3: MT Consensus Policy Area Coding, by Experiment

Table 3 shows the counts of consensus policy area codings for all experiments. Here we can observe significant biases being exhibited by MT coders. It appears that MT coders are *heavily* biased against coding a sentence as having no policy area. Likewise, the non-expert coders seem to over-code sentences as having to do with social policy.

The bias away from the "no policy" category is very strong; and interestingly, gets stronger as the qualification test gets harder. This latter point is important, because the qualification tests contains two sentences of each type. In the low-threshold test it is possible for a coder to miscode two sentences and still pass, while in the high-threshold this is reduced to one. This means to pass the low-threshold test a coder could qualify by missing both sentences not containing a policy statement, but to pass the high-threshold test a coder would have had to correctly code at least one no policy sentence. The evidence here, however, suggests that this recognition does not carry over into the actual coding task.

To investigate this bias further, in Table 4 contains the percent agreement between MT consensus sentence units by expert codings and experiment. Here the bias is extremely apparent. This data shows that non-experts are – in fact – quite good at identifying policy statements in text; with percent agreement over 0.75 for economic sentences and over 0.90 for social sentences. It is recognizing when a sentence makes no policy statement that they fail.

This highlights a very interesting feature that may be present in the MT labor pool: coders

Experiment	Expert Coding	MT % Agreement
No Qual.	Economic	0.77
	Social	0.92
	Neither	0.22
Low-Threshold	Economic	0.87
	Social	0.98
	Neither	0.20
High-Threshold	Economic	0.77
	Social	0.91
	Neither	0.09

Table 3.4: Agreement Between Experts and MT Coders for Each Sentence Type, by Experiment

want to code. By that I mean, after reading instructions about policy statements, and seeing examples, MT coders are hyper-aware of policy statements. They look for them, perhaps, even when they are not there. Also, as a function of the labor environment of MT, successful workers are "successful" because they are good at identifying patterns or components in a task and labeling as such.

It very well may be the case the MT coders find it difficult to understand why a requesters would even include a "null" category. The work they are accustomed to doing involves quick and repeated categorization, and the subtlety of recognizing when a unit has no category may be beyond the scope of what MT is well-suited to support. To follow this thread further, in the next section I report the results of an additional set of experiments that attempts to separate out this phenomenon.

3.4.2 Separating Social and Economic Sentences

To further explore MT coder's bias against coding text as not making a policy statement two additional experiments were run. The corpus of sentences used in the first three experiments was divided such that one contained only "gold" coded economic and neither sentence units, and the other with only social and neither sentence units. With updated instruction, and the incorporation of the low-threshold qualification test, coders in these experiments have a binary choice: code the sentences of having a specific kind of policy, or not[16].

	Results			Kappa Statistic		
Experiment	Sentences	# MT Coders	% Agree	k^*	Std. Error	z
Econ-only	942	15	0.62	0.23	0.10	4.28
Soc-only	955	32	0.60	0.17	0.09	0.95

* A k value between 0.4-0.6 is considered "moderate" agreement

Table 3.5: Inter-rater Reliability Statistics for "Economic-" and "Social-only" Experiments

Experiment	Expert Coding	MT % Agreement
Economic-Only	Economic	0.92
	Not Econ	0.28
Social-only	Social	0.97
	Not Social	0.19

Table 3.6: Agreement Between Experts and MT Coders for Each Sentence Type for "Economic-" and "Social-only" Experiments

Interestingly, the results reported in Tables 5 and 6 reinforce the observation made in the previous section the MT coders find it very difficult to code for null. In both experiments we see the non-experts performing extremely poorly at identifying sentences that contain no policy statement. Using the same comparative statistics as before, we can see that coders overall performance decreases dramatically in these experiments. Unpacking the results highlights why this is the case.

As before, MT coders are unable to recognize sentences that experts code as having no policy statement. In this case, however, because the experiment presents only a binary choice the MT workers heavily over-code whichever policy area the experiment specifies. While it is

[16] Given the results from the previous section, the decision was made to use the low-threshold test as qualification for these additional test.

the case that MT coders do slightly better at recognizing non-policy sentences, the gains are insignificant.

The results of these experiments, combined with the observations from the previous experiments, show that non-expert coders on MT have difficulty identifying sentences that make no policy statement. Put in a different – perhaps more accurate way – when MT coders are asked to categorize something they seem to have difficulty *not* applying some category to it. The subtlety of a null coding may be beyond the capacity of what non-expert coders are capable of. More likely, however, is that because MT coders are deeply primed by the platform to always code, additional care must best taken ensure that the null coding is perceived as equally valuable and likely.

If it is the case, however, that most MT coders are predisposed against the notion of giving something a null code, then well-designed coding tasks on MT must identify coders that do not have this predisposition quickly. This way those coders can be incentived to continue to perform coding, while those that under-perform are kept out. In the following section I investigate how individual coders perform over the course of many HITs to examine what – if any – affect previous work has on worker performance.

3.4.3 Coder Performance Progression

Recall that in each HIT a worker is asked to code six sentence units, and that an individuals worker is limited to submitting 50 HITs total. It is possible, therefore, to track the cumulative performance of individual coders in each experiment. For each HIT, I can calculate the percent agreement an individual coder achieves with the experts, and as they accept more HITs, recalculate based on additional codings. With this information we can observe how much individual coder performance fluctuates. It is also possible to observe the aggregate performance of coders with a given experiment, and trends therein.

In Figure 3.7, the performance of MT coders in the “no qualification” experiment is plotted. Each red line represents the performance of an individual coder. The vertical dashed gray lines represent the HIT counts, so line segments between these dashed lines represent changes in performance from HIT-to-HIT; and, the longer the red line the more HITs a coder submitted. When a coder only submitted one HIT their work is represented as a single red dot at extreme left of the plot. The darker the dot, the more single-submission were collected at a given performance level. For example, we see a few workers submitted all incorrect codings on their first HIT, but then quit. Also, the darker red a line, the more coders followed that exact performance path. This behavior is particularly prominent for workers starting in the 25th to 75th percentile of performance, and then fades as workers leave the task. The blue line with 95% confidence-intervals is the linear fit to all of the data in the experiment.

There are many interesting features of coder participating and performance highlighted by this analysis. First, most coders submit multiple HITs. In fact, in the no qualification test experiment featured in Figure 3.7 only 32% of coders submitted only a single HIT. This is by far the highest percentage of single-submission workers, with the percentages ranging between 0% – 15% for the other experiments. In addition many coders participate in the maximum number of HIT available, providing an ample historical perspective on their performance.

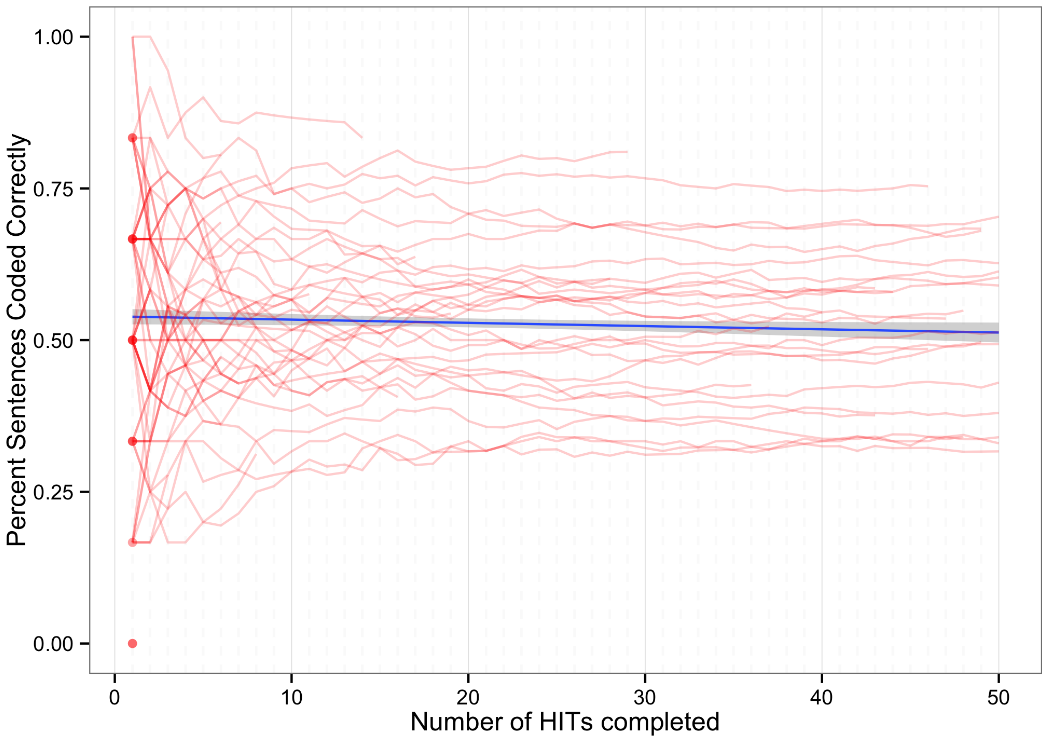
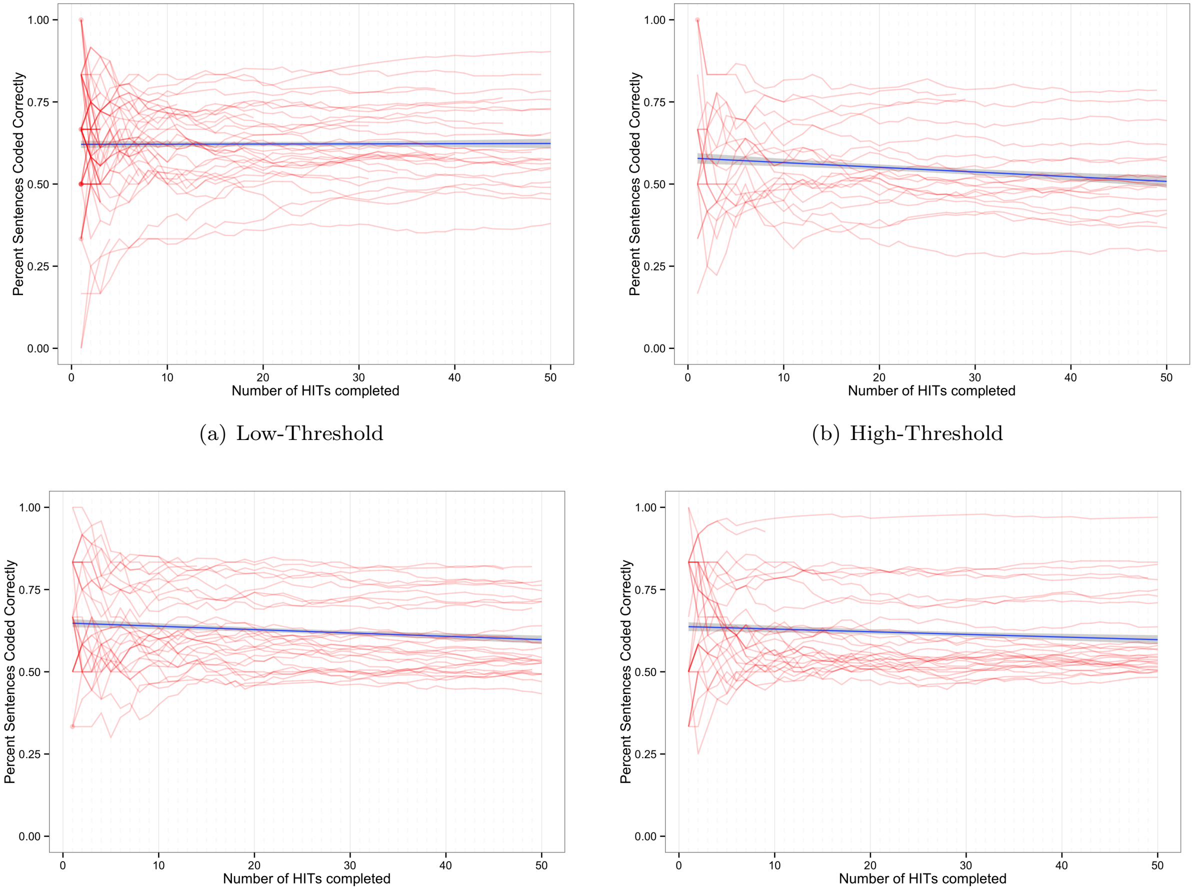


Figure 3.7: Coder Performance for “No Qualification” Experiment



(a) Low-Threshold

(b) High-Threshold

(c) Economic-only

(d) Social-only

Figure 3.8: Coder Performance For All Qualified Experiments

This result is important because it provides evidence that future designs should include a mechanism for assessing coder performance both before they are allowed participations; as with the qualification test, as well as while they are participating in the task. If most coders are submitting multiple HITs, than a well-specified model of adjudication can include a dynamic assessment of coder performance. Also, evidence from these experiments suggests that the number of MT workers submitting coding for given sentence unit does affect the likelihood that the consensus coding will match expert coding[17]. As such, there should be little to no affect from further filtering workers after they are allowed to participate.

Figure 3.8 provides the coder performance graphs for the remaining four experiments, and all highlight a second key observation of the analysis. Coder performance quickly reaches a steady level of quality, and remains at that level for the length of participation. More importantly, this stability is not path dependent, i.e, it is not explicitly a function of their performance at the outset. Many coders across the experiments perform very well – or very poorly – to start, but as they do more work their performance quickly adjusts up or down, and then reaches a steady level. For a more detailed view of coder performance stability see Figures 17 and 18 in Appendix C.

These plots also show several clear groupings of coder quality, where stable coders group in high- mid- and low-quality tiers. This is most apparent in the economic- and social-only experiments. This groups likely separate those coders that are able to identify sentences with no policy statements, from those that are not. In fact, many coders in the high-quality tier very closely resemble expert coders, maintaining percent agreements statistics well above 0.90. It may be that the most well-designed crowd-sourced coding tasks will utilize a combination

[17] See Table 13 in Appendix C

of these observations. If one were able to design a mechanism that leveraged this information, the mechanism would be able to identify and reward high-quality coders systematically. This would almost certainly result in higher-quality output

In the next, and final, exposition of the results I examine the MT coders' ability to identify policy scale as a second-order coding.

3.4.4 Measuring Party Position Shifts

There are three general ideological scaling constant that exist in the manifestos used in these experiments. The Conservative party has makes systematically more conservative/right statements; the Liberal Democrats make systematically more liberal/left statements; and the Labor party makes a dramatic shift to the center from the left between 1987 and 1997. To examine if the MT coders are able to identify these ideological positions, I analyze the distribution of mean policy scale codings for sentence units with consensus economic and social codings for all experiments.

Figures 3.9 and 3.10 illustrate the results of policy scale coding for consensus economic and social sentence units, respectively. It is necessary to limit these analyses to sentences that achieve consensus coding by the non-experts given the hierarchical structure of the coding schema. Given the nuanced nature of ideological statements in party manifestos, and non-expert coder performance in the previous section, our expectations for performance in this regard may be tepid. The results, however, are promising.

The plots in Figures 3.9 and 3.10 are positioned as a grid. Each column corresponds to sentence units from a given party manifesto, and each row are likewise sentence units coded in

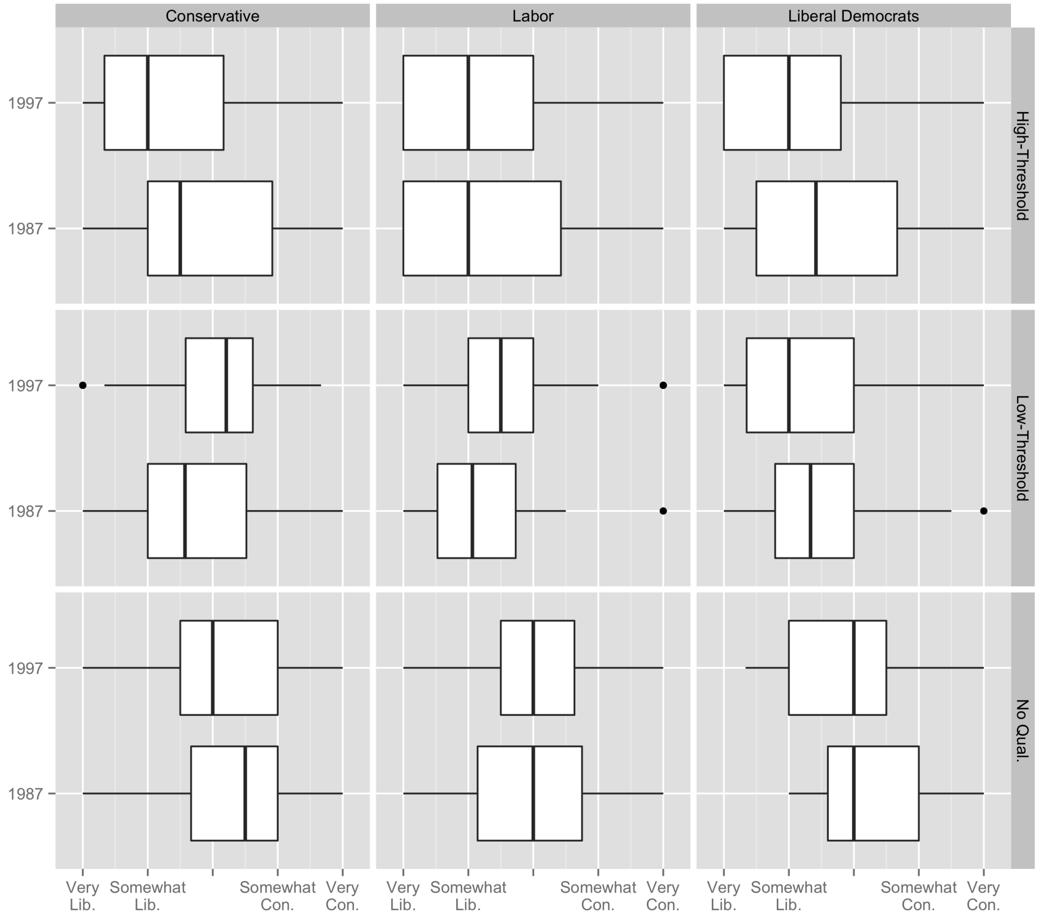


Figure 3.9: Distribution of Average Scales for MT Consensus "Economic" Sentences, by Year and Party

the given experiment[18]. The x-axis corresponds to the five-point scale. The box plots on the y-axis are the distribution of mean policy scale values for consensus coded sentence units.

At first glance, it may appear that coders interpret texts as not having a policy scale, or the zero value on the five-point scale. On the contrary, this is simply the result of the high variance

[18] The economic- and social-only experiments were excluded from this analysis because the results add no additional insight.

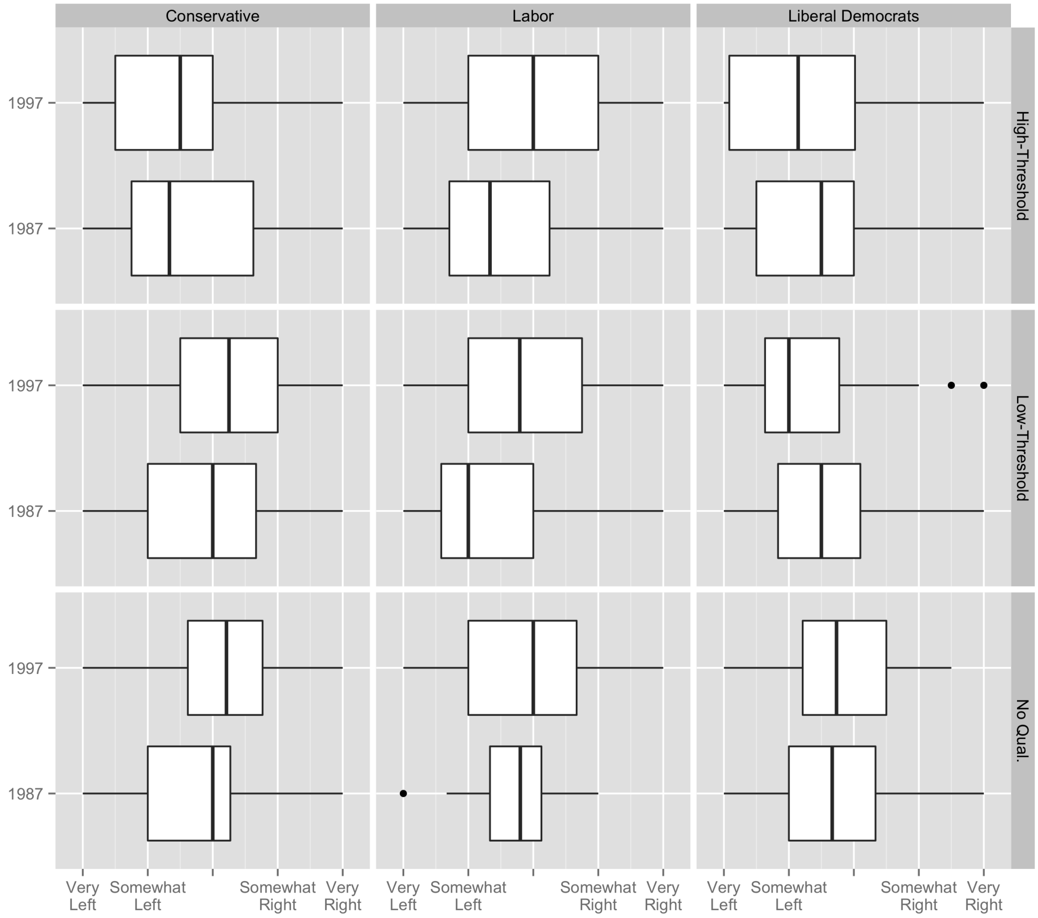


Figure 3.10: Distribution of Average Scales for MT Consensus “Social” Sentences, by Year and Party

in these mean estimates. In fact, the “neither; nor” policy scale coding is the least likely coding across all of the experiments. This follows the pattern observed in the previous analyses: MT coders as biased against a null coding, which carries down the hierarchical scheme to policy scale.

Despite this persistent shortcoming of the MT coders, there are several positive results in these data. First, in both the low- and high-threshold tests, the coders are systematically coding the Liberal Democrat texts as liberal and left for both economic and social policy statements, respectively. Likewise, coders are also systematically coding text from the Conservative as being more conservative/right, except for a significant bias to the left for coders in the high-threshold experiment. It is unclear what caused this bias.

Finally, coders in experiments including a qualification test appear to be picking up on ideological changes in the Labor party between 1987 and 1997. This can be observed in both mean shifts to the center, and the overall increased variance of the mean scales in these results. The magnitude of this shift in the MT codings, however, is considerably smaller than what has been observed by previous expert codings and the codings of experts in this research[19].

3.5 Conclusions

I have presented the results of a series of experiments used to assess the viability of using the Mechanical Turk platform to crowd-source political text coding. Using a simple hierarchical coding schema, multiple expert codings are gathered on party manifestos from three political parties for two different years. The results of these expert codings are used to assess the quality of codings gathered from non-experts on the MT platform using the same coding schema.

The results of these experiments provide considerable evidence to that crowd-sourcing is an effective alternative method to generating quantitative categorization from text. When using a well-designed mechanism for collecting non-expert codings *en masse*, the results compare quite well to the results of multiple experts. There are, however, some important features of the MT environment that must be considered when using it for this type of work.

[19] See Figures 19 and 20 in Appendix C for visualizations of expert coding of policy scales.

First, and most importantly, workers in the MT platform have tremendous difficulty identifying a null category. In these experiments this manifested as the extremely poor performance on identifying sentences as not having a policy area, or not expressing a policy scale. It may be that because of the nature of work on MT, workers are predisposed to finding a categorization even when one does not exist. Though not the majority, many coders were – however – able to accurately do this type of subtle coding.

For this reason, it is crucial that well-designed MT tasks have a mechanism for identifying these high-quality coders quickly, and rewarding them appropriately to incentives them to continue to do work. Additionally, low-quality coders should be identified equally quickly and filtered out of the task. The observation that the vast majority of workers do multiple HITs, and that their performance quickly reaches a – and measurable – state, suggests that a dynamic mechanism for tracking performance should be included to perform this adjudication.

Crowd-sourcing is a cost-effective, highly scalable, and extremely flexible alternative means for generating quantitative coding data. The results of this research indicate its future as a methodological tool in the discipline are quite bright.

Final thoughts

The work presented in this thesis has attempted to highlight the power of computational methods for political science. The focus here has been on expanding the reader's perspective on what constitutes a "computational method" in the discipline. In addition to generative methods, each chapter has presented a core problem from political science, and approached it using computational methods.

In Chapter 1, "Networks, Collective Action, and State Formation," the question is posed: how do social network structure affect how groups make choices about contributing to a public good? Given the nature of public goods, it seems natural that these networks would play a role in their establishment. No previous research, however, had examined this in a systematic way.

One of the primary reasons for the lack of research in this area is it is very difficult to collect and test data on social structure that have naturally formed in meaningfully different ways. As such, in this chapter a generative model is designed to test how network variations affect how agents play a network variant of a provision point public goods game. In this case the computational method is an agent-based model, which is used because of the difficulty presented by collecting real data, along with the flexibility offered to test many different network types in a single model.

The results provides strong evidence that agents playing the game on a network formed using a "preferential attachment" mechanism are much more likely to coordinate on the provision of public goods. This result holds regardless of the decision criterion being used by players to contribute to the public good. In the second chapter the focus shifts from generating classic networks structures, to studying real-world complex networks and attempting to model their dynamics.

In this chapter, "Modeling Network Structure Using Graph Motifs," I introduce a novel approach to modeling the dynamics of networks by studying the distribution of a set of graph motifs constituting a given network. The vast majority of network analysis done within political science – and social science more generally – focuses on statistical measurement and modeling of static networks. It is difficult to study network dynamics analytically because of the intractability of the inherent dependency structures present in these structures.

In order to begin exploring how these dynamics might be studied, the graph motif model is designed computationally, and several models are specified and tested to assess its ability to model network dynamics. A key criterion for dynamic network models is their ability to accurately model the growth of networks over time. To test this, a time-series of a co-authorship from the *Conflict Studies eJournal* was analyzed to illustrate how a graph motif model could be used to model the growth of a network over time.

The results are quite promising and highlight the value of using graph motifs to model networks. By utilizing computational methods to specify these models the technique has inherent flexibility, which is useful when attempting to model the naturally complex nature of human social interactions.

In the final chapter the methodological application moves away from simulation to experimentation. Here, the computational method is crowd-sourcing, and it is used as an alternative means of generating categorical data. In “Methods for Collecting Large-scale Non-expert Text Coding,” the innovation is the application of moving the task of coding text away from single-expert composition, to dispersed, large-scale, collection of non-expert coding. To do this, several experiments were conducted on Amazon’s Mechanical Turk platform to assess the viability of crowd-sourcing for this work.

The results of these experiments provide considerable evidence that crowd-sourcing is an effective alternative method for generating quantitative categorization from text. The key insight from this work is that such efforts must be meticulously designed to account for the particularities of the labor pool available on MT. Most notably, non-expert coding exhibit a significant bias against the identification of a null, or “does not fit,” category. As such, researchers attempting to use crowd-sourcing for text coding must be accurately aware of this bias when designing experiments.

The use of computational methods in political science presents – perhaps – the largest methodological innovation of the discipline to date. Their application is limited only by the imagination of a researcher, and their willingness to learn new tools. This thesis has attempted to highlight the potential of computational methods within the discipline. The results from each chapter show that these methods have a very fruitful present, and an extremely promising future, in the discipline.

Appendicies

Appendix A: Pseudocode for Network Variant of Public Goods Game

All code for the above model can be downloaded here: `http://github.com/drewconway/StateBuilding`. Below is a brief "pseudocode" implementation of the model, provided as an explanatory supplement to the actual code for non-programmers.

Algorithm 3 Agent Class

```
Require: $id$ and $type$
Ensure: $id \in \mathbb{Z}^+$, $type \in \{0,1,2,3,4\}$ and $disposition \in \{0,1\}$
  $network = \emptyset$
  $c = 0.0$
  $mnet = 0.0$
  {...}
  {Also, supporting class operations for getting and setting parameters}
  {...}
```

Algorithm 4 Environment Class

```
Require: $population$
Ensure: $population \in \mathbb{Z}^+$
  $state_{wealth} = \sum_{i=1}^{population} i_{wealth}$
  $w = \frac{1}{4}(state_{wealth})$
  for all $agents \in population$ do
    create networks
  end for
  for all $agents \in population$ do
    set $mnet$
  end for
  for all $agents \in population$ do
    set contribution level based on type, disposition and $mnet$
  end for
  $state_{contributions} = \sum_{i=1}^{population}(i_{contribution})(i_{wealth})$
  if $state_contributions \geq w$ then
    $provision = TRUE$
  else
    $provision = FALSE$
  end if
  {...}
  {Also, supporting class operations for getting and setting parameters, and outputting data}

  {...}
```

Appendix B: Analysis of SSRN *Conflict Studies eJournal*

The network data used in the above paper is a novel contribution to both the network science and political science literatures. As such, it is useful to provide a separate—more detailed—analysis of this data in addendum to the description already provided. As an initial step in the analysis it will be useful to isolate the part of this data with the most meaningful structure. At the time of writing, the entire network contains 5,516 nodes and 4,457 edges, giving it a density of $1.5e^{-4}$. Furthermore, within this sparsely connected network there are 1,493 weakly connected components, many of which are simple dyads or star graphs.

As such, for the remainder of this analysis the focus will be on the largest weakly connected component, also referred to as the main component of the network. Focuses only this subgraph is helpful as it allows for traditional network analysis techniques to be applied to a single fully-connected component. Unfortunately, many graph theoretic metrics cannot be applied to networks for which there are some set of nodes that cannot be reached through graph traversal. For example, betweenness centrality; defined as the proportion of shortest paths that pass through a given node, is undefined for graphs with multiple components because there is infinite distance between nodes in different components.

Also, because the co-authorship network is a natural bipartite graph with authors connected to papers it is possible to produce projections of the graph into affiliations networks. In addition to the main component of the full network, we may also analyze the network formed by author-to-author and article-to-article affiliations. This is done in order to understand how these two modes of the bipartite network interact with one another, which allows for comparison between the two networks. Table 3 below provides basic descriptive statistics of these three networks.

	Main Component	Authors	Articles
Number of Nodes	1,190	745	445
Number of Edges	1,522	4,574	2,728
Mean Degree*	1.279	12.279	12.261
Density	0.001	0.017	0.028

* For the main component this is mean in-degree

Table 7: Descriptive statistics for entire network and projections

As can be seen from the mean degree values for the affiliations networks, despite the relatively sparsity of the full-network there is actually quite a lot of collaboration occurring within the *Conflict Studies eJournal*. Clearly, however, mean degree does not fully capture the collaborative dynamics of this network. A better measure is the degree distribution, as this will indicate the frequencies of various collaborations, i.e., number of times individuals have co-authored articles in the network. Figures 13 and 14 illustrate the degree distributions for the authors and articles affiliations networks respectively.

There are many interesting features of this network that become apparent when examining these distributions. First, there is significant variation in collaboration among authors within the network. The vast majority of authors have co-authored with up to twenty different individuals within the population of the network. Having more than twenty different co-authors, however, is extremely rare. Likewise, the articles network shows that most articles share up to ten authors, but having more than ten is considerable rarer. Interestingly, sharing the same authors thirty-five times is nearly as common as having three or fewer. Upon further examination of the data this appears to be the result of a single entity, the International Council on Human Rights Policy (ICHRP), single-authoring thirty-five articles, which resulted in a completely connected clique within the graph for this single author. Fortunately, this was the only occurrence of such an anomaly in the data.

	Authors	Articles
χ^2	1,672	1,083
Degrees of freedom	1,650	1,064
p-value	0.347	0.336

Table 8: χ^2 goodness-of-fit test of affiliations networks to Poisson

In addition to exploring the degree distributions to understand the collaboration mechanisms presents in the network it is also illuminating to test the fit of these distributions against a theoretical distribution. Specifically, purely random graphs—such as the Erdõs-Rènyi model—are well known to follow a Poisson distribution (Newman et al., 2001). As such, using the mean degree values from Table 3 as the theorized shape parameters of a Poisson it is possible to test the fit of both the authors and articles affiliations networks. Table 4 above shows the results of a χ^2 goodness-of-fit test for the degree distributions to the theorized Poisson. As can be seen, these data are fairly poor fit to these distributions, which provides evidence that the connectivity dynamics are not random.

This is encouraging, if the data were a good fit it would indicate unnatural skewing of the degree, calling into question the validity of the data for study. The poor fit of these data to the Poisson is highlighted more starkly in Figures 15 and 16, which show the differences between the theorized Poisson count and the empirical degree distribution for the authors and articles networks respectively. What is clear from these figures is that both networks have considerably more nodes with degree in the head of the distributions than would be predicted by the theorized Poisson, leading to their poor fits. As to be expected, the distributions have a heavy tail, a characteristic present in many large social networks.[20]

Finally, a popular method in the statistical analysis of networks is to identify so-called "key

[20] See (Clauset et al., 2009) for of evidence of this empirical result.

actors," based on various centrality metrics. There are many metrics by which this centrality can be defined, including degree (number of edges), betweenness (presence in shortest paths), closeness (geodesic distance to all nodes), or Eigenvector Centrality (Freeman, 1979). For the purposes of this analysis I will be using Page Rank, which is closely related to Eigenvector Centrality and was originally introduced by Sergey Brin and Lawrence Page—the founders of Google—as the primary metric by which their World Wide Web search engine ranked the relevance of web pages (Brin and Page, 1998). This metric is well suited for this analysis as it was designed to overcome the variational limitations present in Eigenvector Centrality and is tailored to ranking inter-related documents.

As there are so many centrality measures available, it is often useful to show the relationship between two metrics within a single network. Given the high levels of interdependence among these metrics, the expectation is that in most cases they will have a roughly linear relationship, i.e., actors with high values for one centrality metric will also have high values for another. The variation, however, is also useful because deviation from this linear relationship can indicate a unique structural position for certain nodes. In this case, because Page Rank is the primary metric I will use betweenness centrality as the secondary metric. The former metric focuses on structural centrality, i.e., weights those nodes at a network's core higher, while the latter focuses on structural singularity, i.e., weights those nodes whose position represent cut-points in the network.

For example, actors that are outliers on the PageRank dimension are those that are at the core of the network, but likely only have a few connections into that core; hence they are on a fewer number of shortest paths. Such actors may be those relatively new to the network, but having connections to the core indicates high-value. In a co-authorship network, these actors may be those scholars or subjects emerging within a sub-discipline. Likewise, outliers on the betweenness centrality dimension indicate actors that act as bridges between two disparate region of a network. In a co-authorship network these actors may be those scholars or subjects that blend research from multiple areas of the sub-discipline, or newly emergent research areas.

Figures 17 and 18 are scatter plots of the Page Rank and betweenness centralities for all nodes in the authors and articles networks respectively. A naive linear model was fit to these metrics, and the absolute values of the residuals of these regressions were used to size the points in these plots. This is done to highlight those outliers on either dimension, which may be considered the key-actors of these networks. The point labels themselves correspond to the SSRN ID numbers for each author and paper. In the authors network the linear model has an R^2 value of 0.542 and a Pearson's correlation value of 0.736, indicating a fairly weak linear fit, but relatively high correlation—precisely what we would expect for such an analysis given the many outliers. The relationship is a bit weaker in the articles network but still worthy of this analysis; with an R^2 value of 0.477 and a correlation of 0.691.

In both figures there are several actors that are clear outliers on both dimensions, as well as actors that have high values for both metrics. For convenience, Tables 5 and 6 identify the top ten outliers on each dimension for both networks. To buttress these data, Tables 7 and 8 list the top fifty authors and articles by Page Rank respectively. These tables tell an interesting story about the type of authors using *Conflict Studies eJournal* to distribute research, and the structure of their collaboration.

For the authors network there is a clear dominance by legal scholars Thirty-two out of the fifty highest ranked authors are from law schools. Perhaps more shocking, none of the authors in this table are political scientists. In fact, with the exception of Daron Acemoglu, none of the scholars listed in Table 7 have ever published in a top-tier political science journal.[21]. Looking at the highest ranked papers, however, it is clear that legal issues related to conflict are the most prevalent within this eJournal. Given the time period in which these data were collected this follows logically; with two on-going wars wherein many domestic and international legal issues were raised this was a topic of great relevance.

Finally, the bottom panel of Table 6 is interesting in that it indicates security issues related to climate change acts as a bridge between other areas of study within the network—with three articles specifically focused on the topic. As time proceeds, it will be interesting to see if this area of research becomes most central as the legal issues related to America's foreign wars subside. Unfortunately, these results provide evidence that this particular eJournal may not be the most relevant to the study of collaboration with political science as a discipline given the lack of political scientists at its core.

[21] A search of *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *Quarterly Journal of Political Science* was done for each author in Table 7.

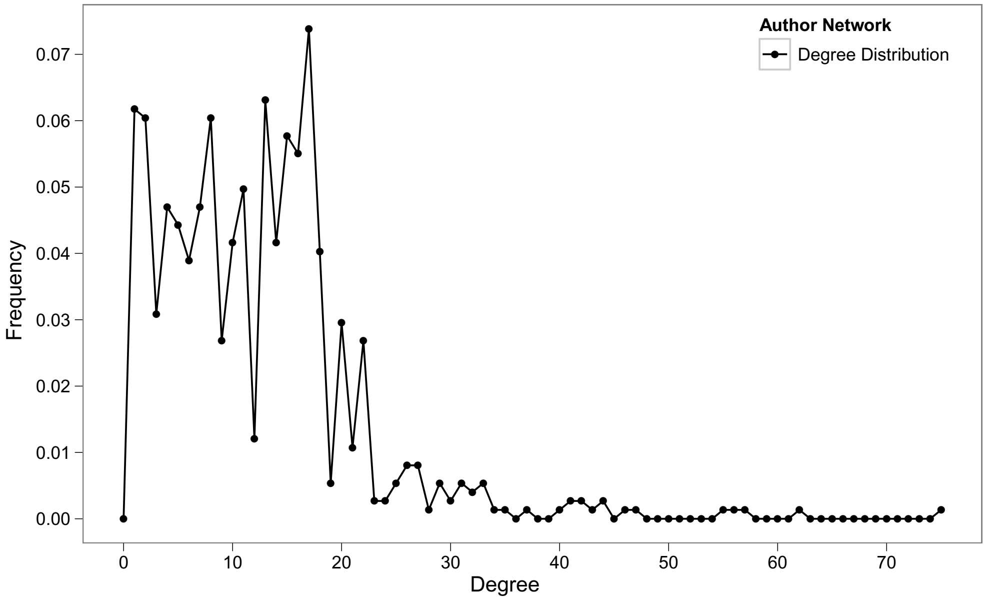


Figure 11: Degree distribution of authors affiliations network

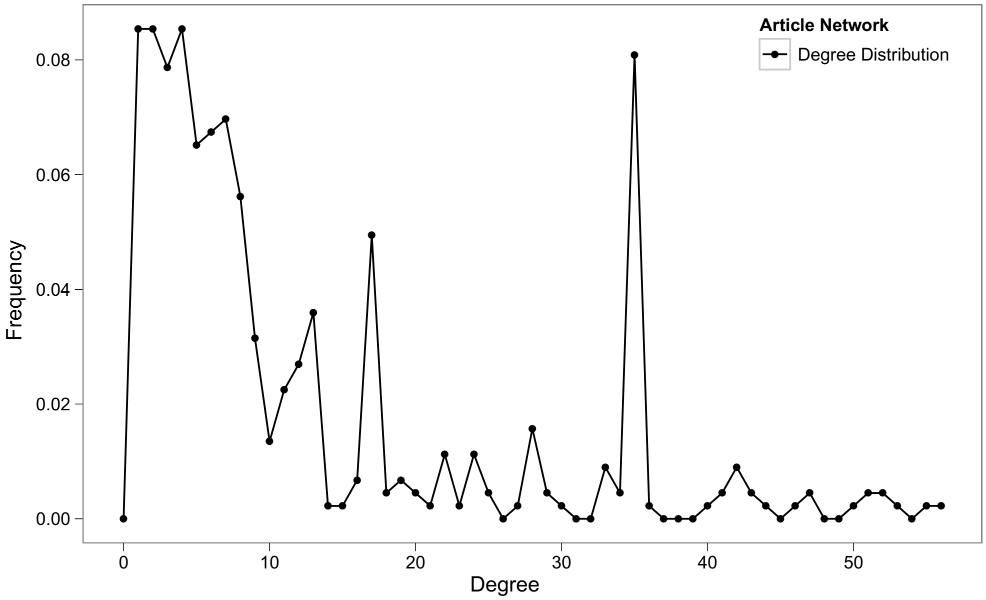


Figure 12: Degree distribution of articles affiliations network

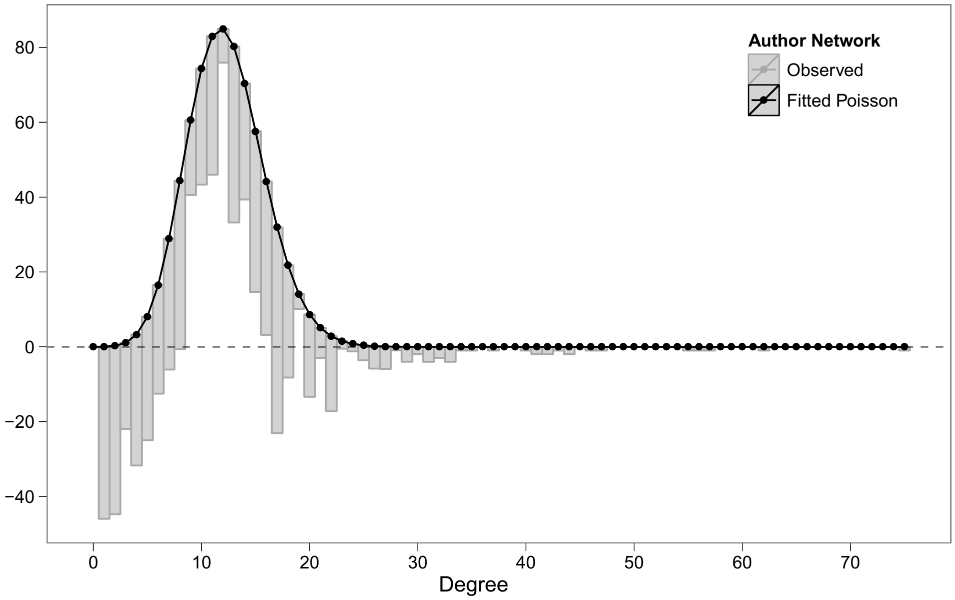


Figure 13: Goodness-of-fit plot for authors network degree to Poisson distribution

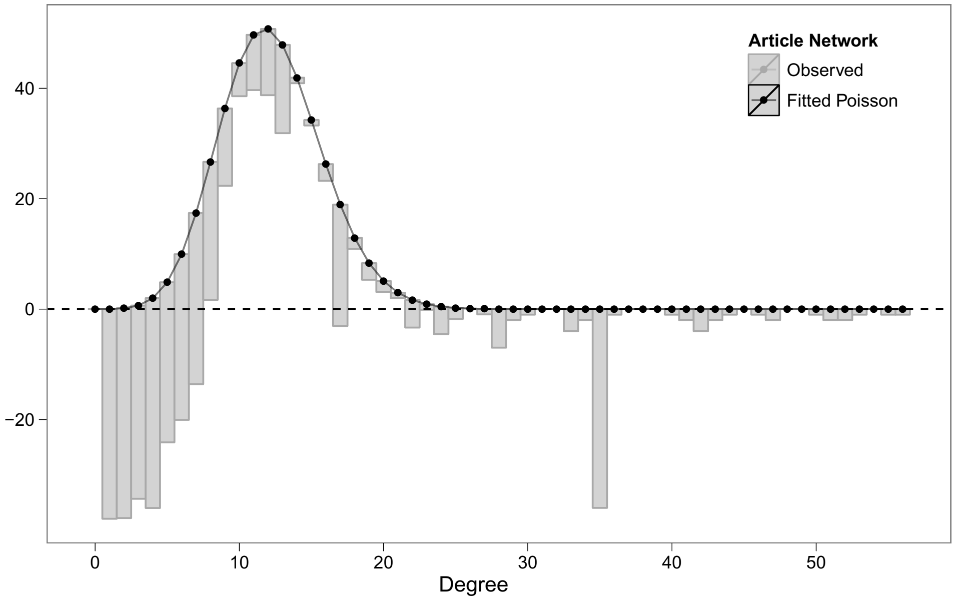


Figure 14: Goodness-of-fit plot for articles network degree to Poisson distribution

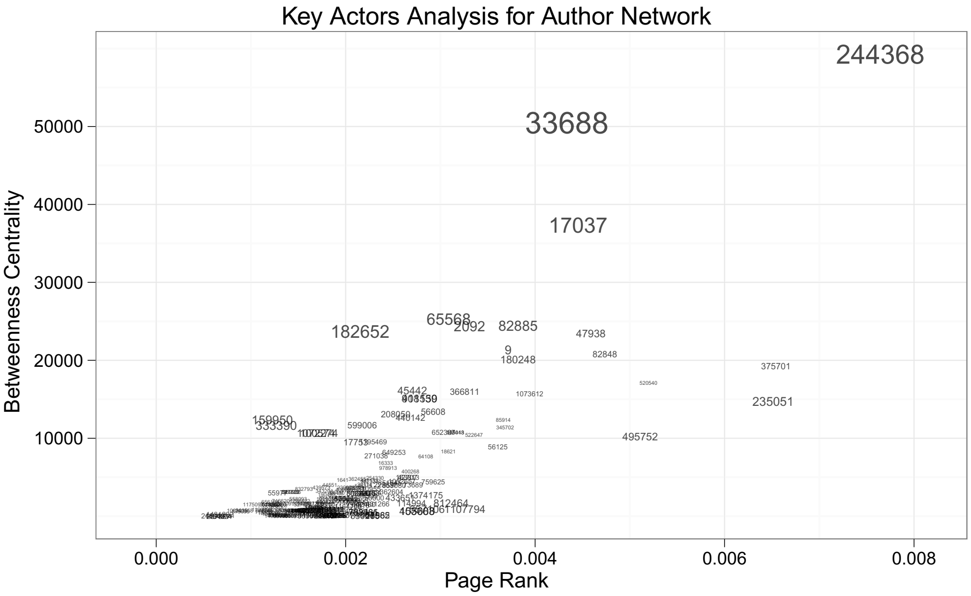


Figure 15: Key actor plot for authors network

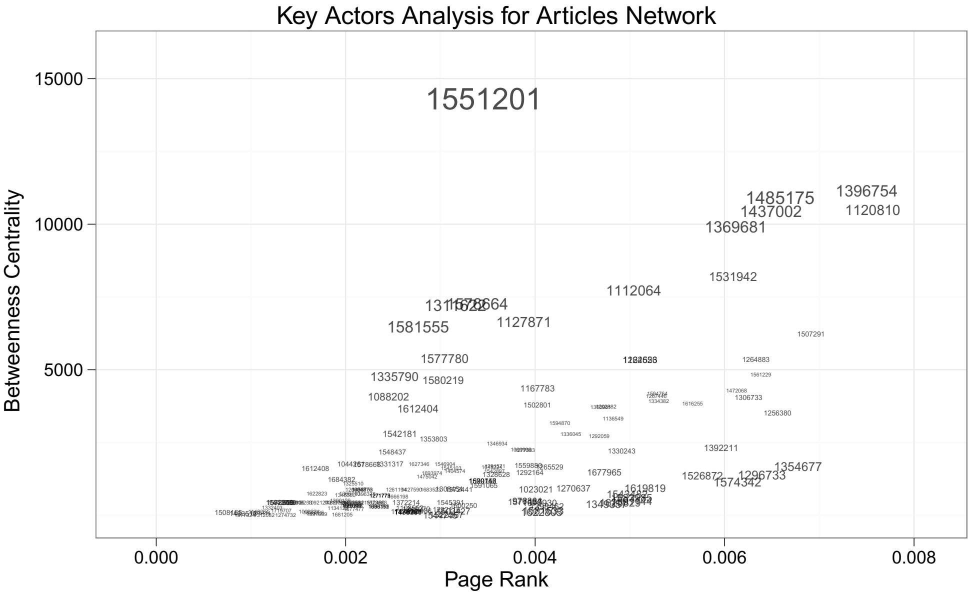


Figure 16: Key actor plot for articles network

	Author Name	Institution	SSRN ID	Residual
Outliers on Page Rank dimension				
1	Kenneth Anderson	Washington College of Law, American University	235051	-9442.03
2	Ganesh Sitaraman	Harvard Law School	1107794	-8886.40
3	Amos N. Guiora	University of Utah - S.J. Quinney College of Law	495752	-7770.47
4	Eric Talbot Jensen	Fordham University - School of Law	812464	-7472.41
5	Geoffrey S. Corn	South Texas College of Law	557106	-7093.83
6	Richard W. Murphy	Texas Tech University - School of Law	105603	-6863.07
7	Afsheen John Radsan	William Mitchell College of Law	453088	-6863.07
8	Curtis J. Milhaupt	Columbia Law School	63865	-5588.80
9	Ronald J. Gilson	Stanford Law School	17982	-5588.80
10	Eric Neumayer	London School of Economics and Political Science	114994	-5585.64
Outliers on betweenness centrality dimension				
1	Eric A. Posner	University of Chicago - Law School	33688	35949.53
2	Simon Chesterman	New York University - School of Law, Singapore Programme	244368	30071.85
3	Lucian A. Bebchuk	Harvard Law School	17037	22280.72
4	Peter J. Boettke	George Mason University - Department of Economic	182652	18827.85
5	Austin Murphy	Oakland University - School of Business Administration	65568	16272.93
6	Luigi Zingales	University of Chicago Booth School of Business	2092	14351.57
7	Jeremy Waldron	New York University (NYU) - School of Law	82885	12156.01
8	Paul Schiff Berman	Sandra Day O'Connor College of Law	159950	11588.07
9	Fabien Gelinas	McGill University	333390	10690.32
10	Michael C. Jensen	Harvard Business School	9	9570.30

Table 9: Key outliers analysis of authors network

	Article Title (truncated at 60 characters)	# of Authors	SSRN ID	Residual
Outliers on Page Rank dimension				
1	Counterinsurgency, the War on Terror, and the Laws of War	1	1354677	-3500.62
2	The International Legality of U.S. Military Cross-Border Op...	1	1296733	-3417.46
3	States of Exception: Regulating Targeted Killing in a Globa...	1	1574342	-3412.21
4	Unlawful Killing with Combat Drones: A Case Study of Pakist...	1	1501144	-2970.06
5	Measure Twice, Shoot Once: Higher Care for CIA Targeted Kil...	2	1625829	-2895.95
6	Legality of Lethality: Paradigm Choice and Targeted Killing...	1	1583985	-2846.91
7	A Trial to End All Terrorism: How America Could Have Won th...	1	1526872	-2827.02
8	Due Process and Targeted Killing of Terrorists	2	1349357	-2809.58
9	The Dark Sides of Convergence: A Pro-Civilian Critique of t...	1	1543482	-2683.09
10	'Drones II' - Kenneth Anderson Testimony Submitted to U.S. ...	1	1619819	-2661.58
Outliers on betweenness centrality dimension				
1	Climate Change and Human Rights: A Rough Guide	1	1551201	12397.30
2	What Causes Terrorism?	2	1148682	8652.95
3	Climate Change, Conflict and Security: International Law Ch...	1	1485175	5914.01
4	The Story of El Masri v. Tenet: Human Rights and Humanitari...	1	1311622	5593.85
5	Cyberwarfare: Law & Policy Proposals for U.S. & Global Gove...	1	1437002	5550.85
6	Implementing the Responsibility to Protect	1	1576664	5406.87
7	The Consequences of Radical Reform: The French Revolution	4	1369681	5387.69
8	The 'Bush Doctrine': Can Preventive War Be Justified?	2	1396754	5254.80
9	Climate Change and Human Rights: Unpacking the Issues	1	1581555	5234.77
10	Two Crises of Confidence: Securing Non-Proliferation and th...	1	1120810	4539.87

Table 10: Key outliers analysis of articles network

	Author Name	Institution	SSRN ID	Page Rank
1	Simon Chesterman	New York University - School of Law, Singapore Programme	244368	0.0076
2	Mary Ellen O'Connell	Robert & Marion Short Chair in Law	375701	0.0065
3	Kenneth Anderson	Washington College of Law, American University	235051	0.0065
4	Kishore Mahbubani	Lee Kuan Yew School of Public Policy	520540	0.0052
5	Amos N. Guiora	University of Utah - S.J. Quinney College of Law	495752	0.0051
6	Joseph Raz	Columbia Law School	82848	0.0047
7	Jordan J. Paus	University of Houston Law Center	47938	0.0046
8	Lucian A. Bebchuk	Harvard Law School	17037	0.0044
9	Eric A. Posner	University of Chicago - Law Schoo	33688	0.0043
10	Stuart Malawer	George Mason University - School of Public Policy	1073612	0.0039
11	Jeremy Waldron	New York University (NYU) - School of Law	82885	0.0038
12	John Yoo	University of California at Berkeley School of Law	180248	0.0038
13	Michael C. Jensen	Harvard Business School	9	0.0037
14	Jeremy J. Sarkin	Hofstra University - School of Law	345702	0.0037
15	Gregory Shaffer	University of Minnesota - Twin Cities - School of Law	85914	0.0037
16	Ali Khan	Washburn University - School of Law	56125	0.0036
17	Robert J. Delahunty	University of St. Thomas School of Law (Minnesota)	522647	0.0034
18	Luigi Zingales	University of Chicago Booth School of Business	2092	0.0033
19	Ganesh Sitaraman	Harvard Law School	1107794	0.0033
20	Daniel Bodansky	Arizona State University Sandra Day O'Connor College of Law	366811	0.0033
21	Jonatan Pinkse	University of Amsterdam - Amsterdam Business School (ABS)	363448	0.0032
22	Ans Kolk	University of Amsterdam - Amsterdam Business School (ABS)	105013	0.0032
23	Eric Talbot Jensen	Fordham University - School of Law	812464	0.0031
24	Austin Murphy	Oakland University - School of Business Administration	65568	0.0031
25	Daron Acemoglu	MIT - Department of Economics	18621	0.0031

Table 11: Top 25 Authors by Page Rank

	Article Title (truncated at 60 characters)	# of Authors	Article SSRN ID	Page Rank
1	What Causes Terrorism?	2	1148682	0.0088
2	Two Crises of Confidence: Securing Non-Proliferation and th...	1	1120810	0.0076
3	The ‘Bush Doctrine’: Can Preventive War Be Justified?	2	1396754	0.0075
4	Teaching an Old Dog New Tricks: Operationalizing the Law of...	2	1507291	0. 0069
5	Counterinsurgency, the War on Terror, and the Laws of War	1	1354677	0.0068
6	Climate Change, Conflict and Security: International Law Ch...	1	1485175	0.0066
7	Transnational Armed Conflict: A ‘Principled’ Approach to th...	2	1256380	0.0066
8	Cyberwarfare: Law & Policy Proposals for U.S. & Global Gove...	1	1437002	0.0065
9	The International Legality of U.S. Military Cross-Border Op...	1	1296733	0.0064
10	Predators Over Pakistan	1	1561229	0.0064
11	Anticipatory Self-Defence and International Law - A Re-Eval...	1	1264883	0.0063
12	Terrorism, Criminal Prosecution, and the Preventive Detenti...	1	1306733	0.0063
13	States of Exception: Regulating Targeted Killing in a Globa...	1	1574342	0.0061
14	The Language of Law and the Practice of Politics: Great Pow...	1	1472068	0.0061
15	The Consequences of Radical Reform: The French Revolution	4	1369681	0.0061
16	A Global Model for Forecasting Political Instability	8	1531942	0.0061
17	Defining Armed Conflict	1	1392211	0.0060
18	A Trial to End All Terrorism: How America Could Have Won th...	1	1526872	0.0058
19	The Structure of Terrorism Threats and the Laws of War	1	1616255	0.0057
20	The Kosovo Crisis: A Dostoievskian Dialogue on Internationa...	2	1334382	0.0053
21	Mapping the Concepts Behind the Contemporary Liberalization...	1	1594764	0.0053
22	International Common Law: The Soft Law of International Tri...	2	1267446	0.0053
23	‘Drones II’ - Kenneth Anderson Testimony Submitted to U.S. ...	1	1619819	0.0052
24	Leashing the Dogs of War: The Rise of Private Military and ...	1	1162526	0.0051
25	Privatizing Peacekeeping: The Regulatory Preconditions for ...	1	1224653	0.0051

Table 12: Top 25 Articles by Page Rank

Appendix C: Additional Statistical Analysis of MT Coding Data

A simple logistic regression model is fit to data from all of the experiments to test of the number of MT coders for each sentences affects the likelihood that the consensus coding will match the experts. The results for all experiments is effectively null, support the claim that the number of coders does not affect the consensus coding.

Experiment Type						
	(Intercept)	# of MT Coders	Sentences	AIC	BIC	log L
No Qual.	0.31	0.05*	1,315	1,697.73	1,739.18	-840.86
	(0.16)	(0.02)				
Low-Threshold	0.99*	-0.02	1,393	1,705.04	1746.95	-844.52
	(0.16)	(0.02)				
High-Threshold	0.36*	0.03	1,250	1,666.09	1,707.14	-825.05
	(0.13)	(0.03)				
Econ-only	0.55**	-0.01	945	1,262.66	1,301.47	-623.33
	(0.20)	(0.02)				
Social-only	0.66**	-0.03	955	1287.75	1326.64	-635.87
	(0.21)	(0.02)				

Standard errors in parentheses

* indicates significance at $p < 0.05$, ** at $p < 0.01$

Table 13: Logistic Regression for Number of MT Coders Per-Sentence on Agreement, by Experiment

To understand how volatile coder performance is over time a 5-HIT moving volatility index is defined as the standard deviation of each coder's cumulative performance. As the standard deviation on cumulative performance decreases, the more stable our estimates of coder performance become. From this analysis it is clear that after 20 HITs, coder performance becomes very stable.

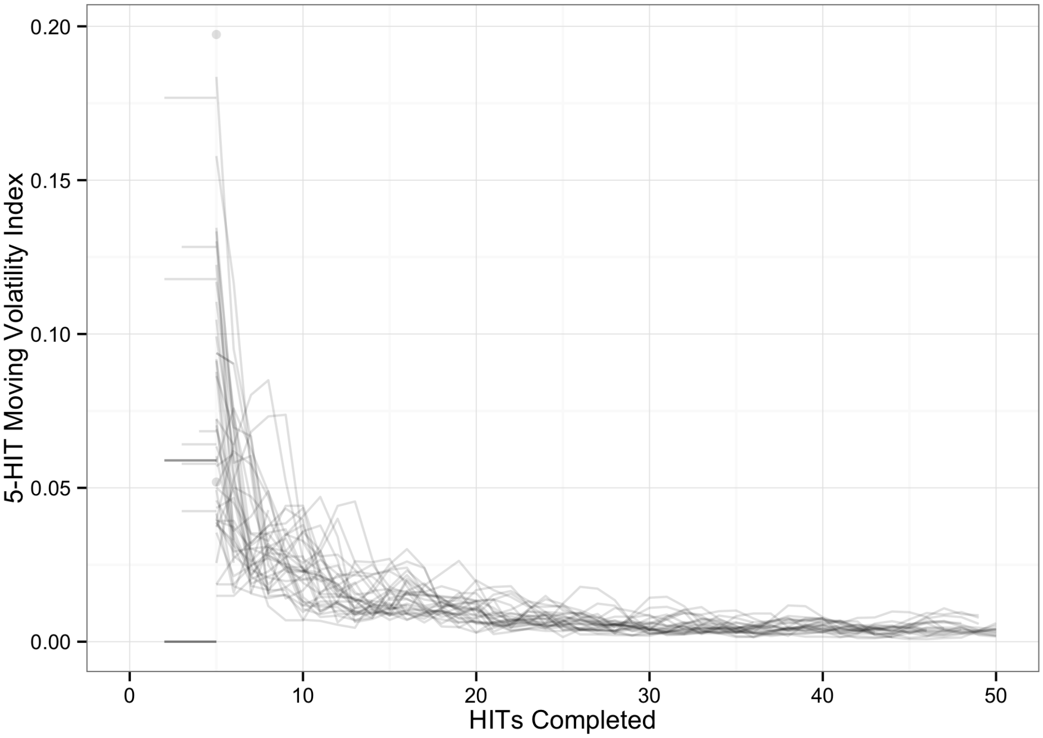
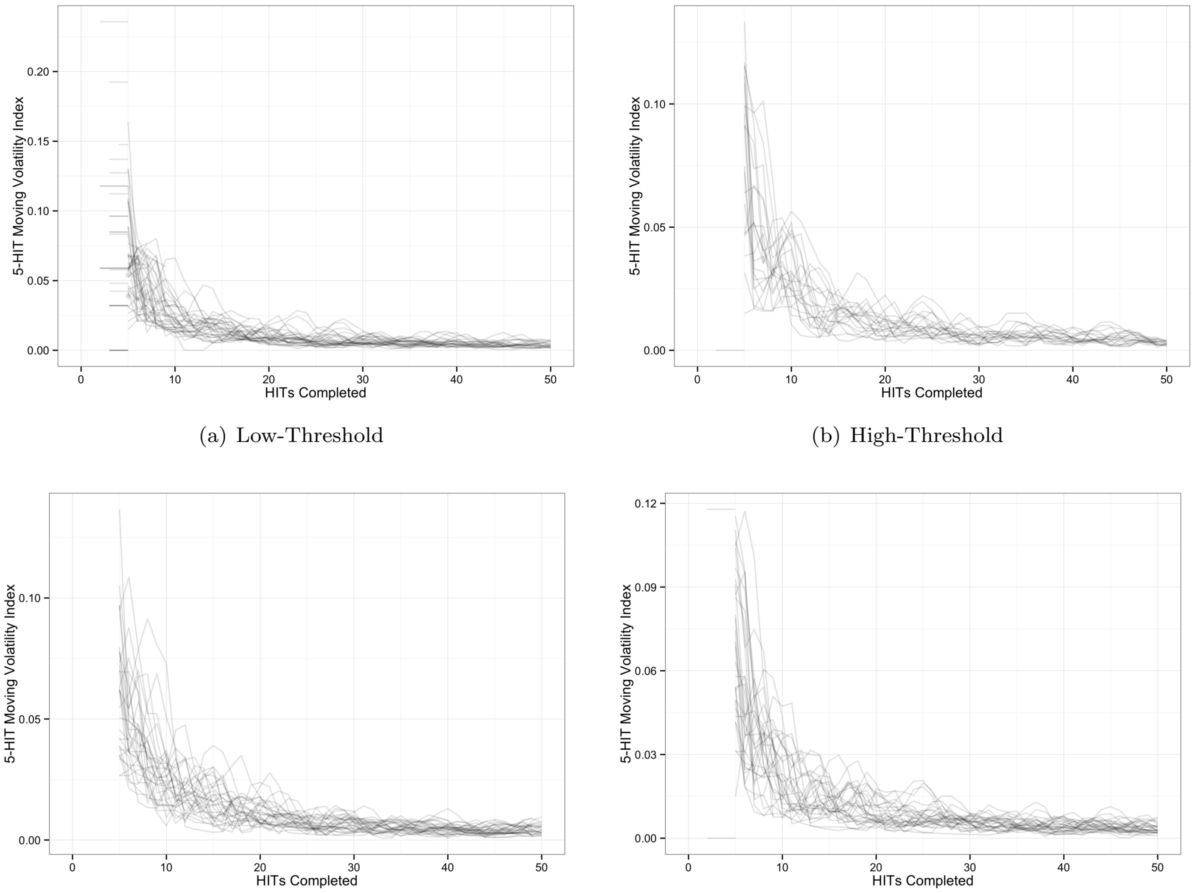


Figure 17: Coder Volatility for “No Qualification” Experiment



(a) Low-Threshold

(b) High-Threshold

(c) Economic-only

(d) Social-only

Figure 18: Coder Volatility For All Qualified Experiments

While coder's reach a relative steady-state in performance, it is not the case that restricting codings to those received after the "burn-in" period improves overall coding quality. As can be seen from the coder performance plots in Figures 3.7 and 3.8, those steady-state codings can still be of a mid-to low quality. Table 14 reports the percent agreement for all of the experiments when using only those codes from coders submitted after their 20th HIT. There are modest gains, but by any meaningful comparison they are unchanged.

Experiment	Expert Coding	MT % Agreement
No Qual.	Economic	0.70
	Social	0.85
	Neither	0.28
Low-Threshold	Economic	0.83
	Social	0.96
	Neither	0.24
High-Threshold	Economic	0.71
	Social	0.86
	Neither	0.12
Economic-only	Economic	0.92
	Neither	0.25
Social-only	Social	0.25
	Neither	0.95

Table 14: Agreement Between Experts and MT Coders for Each Sentence Type for coders submitting more than 20 HITs, by Experiment

In section 3.4.4 the distribution of non-expert coding of policy scales are analyzed. The results are not impressive, however, it could be the case that the random sample of sentences used for the experiments were unrepresentative of policy scale shifts included in each manifesto. If this occurred, then the policy scale coding of experts for these subsets should also show meager shifts in policy among the parities over the time period.

Figures 19 and 20 replicate Figures 3.9 and 3.10 from section 3.4.4 using expert codings. Here, it is clear that the experts are picking up the expected policy shifts, thus reinforcing the observation of non-experts poor performance in coding policy scales.

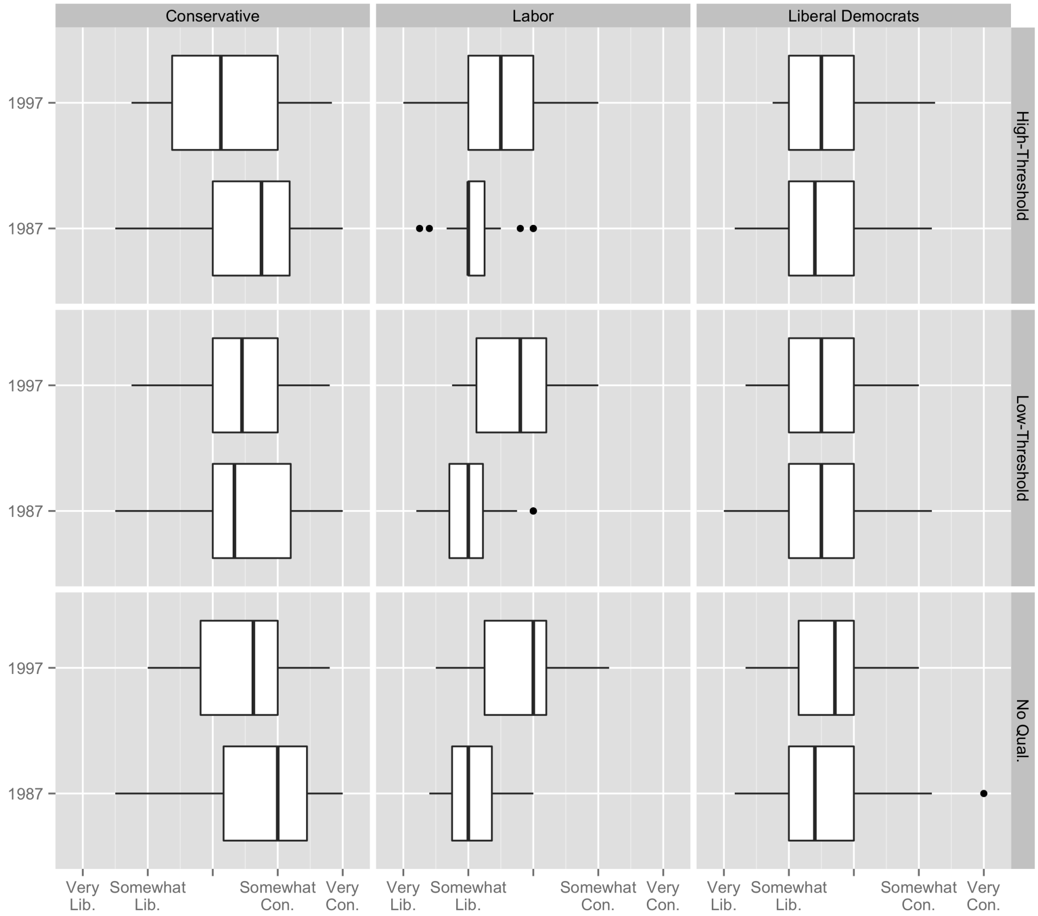


Figure 19: Distribution of Average Scales for Experts “Economic” Sentences Used in Experiments, by Year and Party

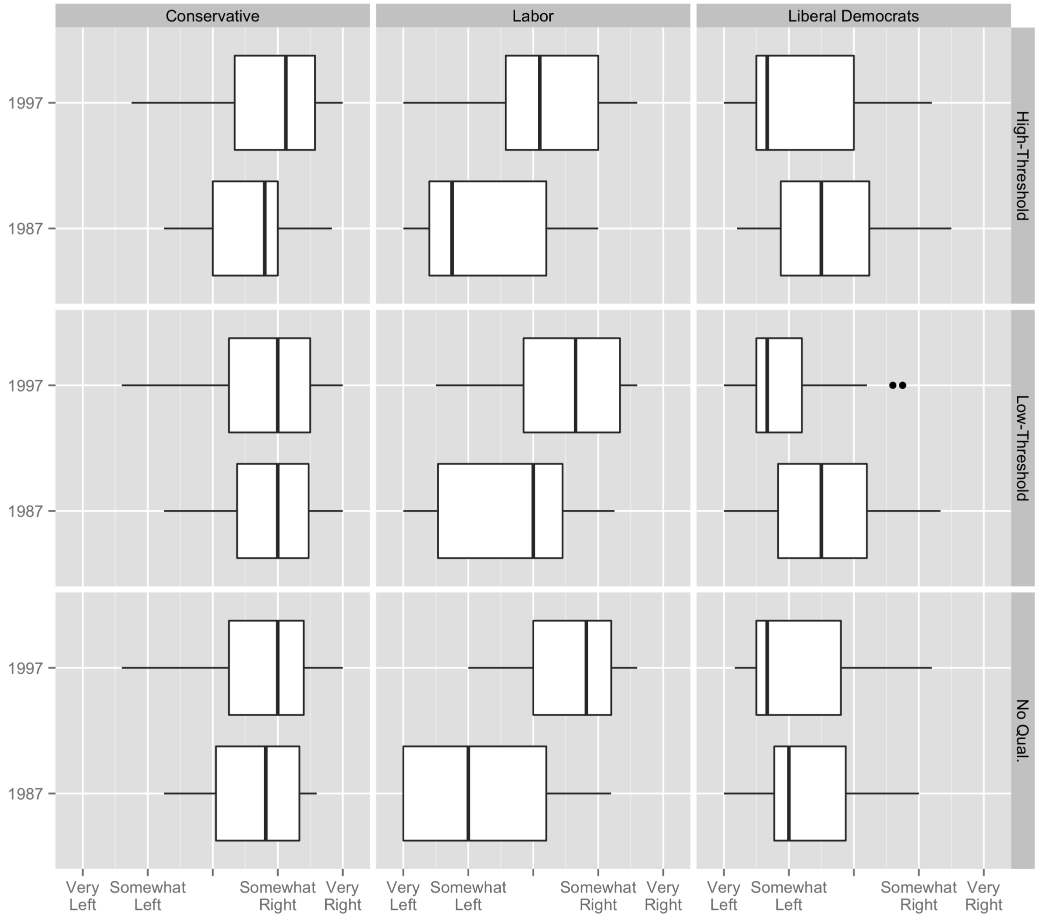
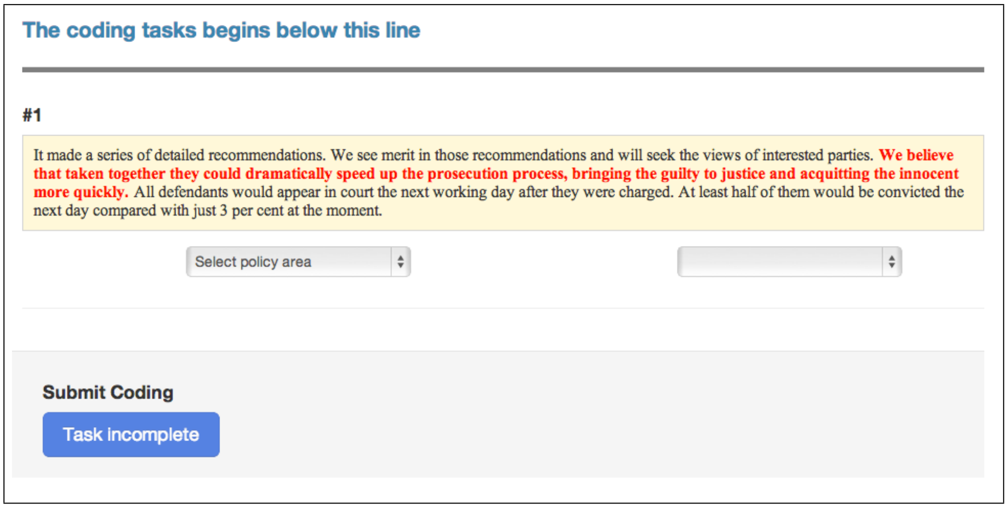


Figure 20: Distribution of Average Scales for Experts "Economic" Sentences Used in Experiments, by Year and Party

Appendix D: Qualification Test and HIT

Below are screen shots of both the HIT and qualification test interface used for these experiments. You will note that the formatting differs between these two interfaces. The reason for this is MT does not allow qualification tests to be hosted externally, meaning that they must conform to MT's templates. The HIT is hosted externally, and is therefore completely customizable.

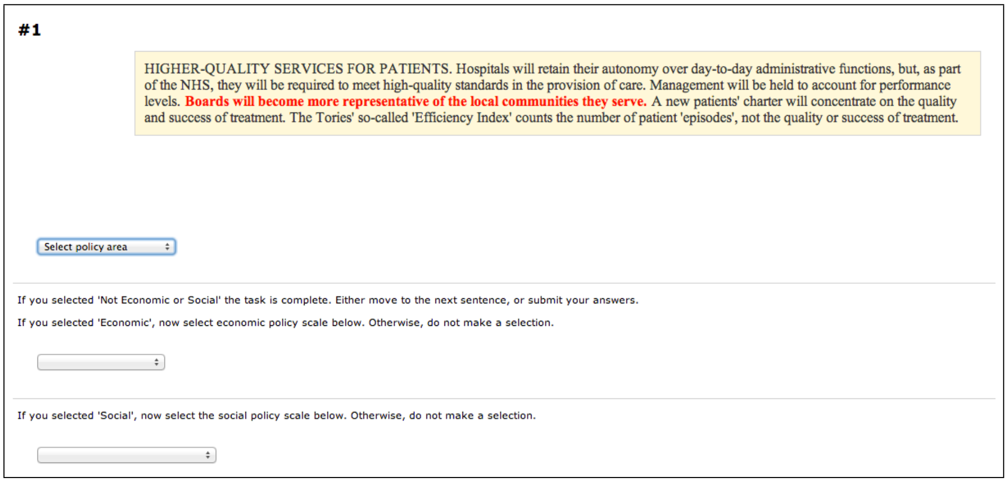


The coding tasks begins below this line

#1

It made a series of detailed recommendations. We see merit in those recommendations and will seek the views of interested parties. **We believe that taken together they could dramatically speed up the prosecution process, bringing the guilty to justice and acquitting the innocent more quickly.** All defendants would appear in court the next working day after they were charged. At least half of them would be convicted the next day compared with just 3 per cent at the moment.

Figure 21: Example HIT Interface



#1

HIGHER-QUALITY SERVICES FOR PATIENTS. Hospitals will retain their autonomy over day-to-day administrative functions, but, as part of the NHS, they will be required to meet high-quality standards in the provision of care. Management will be held to account for performance levels. **Boards will become more representative of the local communities they serve.** A new patients' charter will concentrate on the quality and success of treatment. The Tories' so-called 'Efficiency Index' counts the number of patient 'episodes', not the quality or success of treatment.

If you selected 'Not Economic or Social' the task is complete. Either move to the next sentence, or submit your answers.

If you selected 'Economic', now select economic policy scale below. Otherwise, do not make a selection.

If you selected 'Social', now select the social policy scale below. Otherwise, do not make a selection.

Figure 22: Example Qualification Test Interface

Appendix E: Coding Instructions

The text below are a verbatim copy of the instructions given to MT workers before accepting a HIT in these experiments. The formatting of this text varies from that used in the HIT due to printing requirements. To view the text as it was formatted for MT worker visit: `http://s3.amazonaws.com/aws.drewconway.com/mt/experiments/cmp/html/instructions.html`.

What this task is about

This task involves reading sentences from political texts and judging whether these deal with economic or social policy.

The sentences you will be asked to interpret come from political party manifestos. Some of these sentences will deal with economic policy; some will deal with social policy; other sentences will deal with neither economic nor social policy. We tell you below about what we mean by "economic" and "social" policy.

First, you will read a short section from a party manifesto. For the sentence highlighted in red, enter your best judgment about whether it mainly refers to economic policy, to social policy, or to neither.

If the sentence refers to economic policy, select "economic" in the drop down menu; if it refers to social policy, select "social". If the sentence does not refer to either policy area, select Not Economic or Social – in this case you will move directly to the next sentence.

If you select economic or social, you will be shown a policy scale. Use this to give your best judgment of the sentence in terms of how much it is left or right wing (for economic policy) or liberal or conservative (for social policy). We tell you below about what we mean by left and right etc..

For example, if you see a sentence containing what you think is very right wing economic policy, select the economic policy area from the drop down menu and click very right on the economic policy scale. If you think it has a position that is left-wing, but not very left wing, click left.

If you believe the sentence expresses a centrist position on economic or social policy OR concerns economic or social policy but does not express any clear position, select the appropriate policy category from the drop down menu, as above, and then click the neither. . . nor. . . position on the scale.

Now we need to tell you about what we mean by economic and social policy, and by left and right, liberal and conservative.

What is economic policy? What are left or right economic policies?

Economic policies deal with all aspects of the economy, including:

- Taxation;
- Government spending;
- Services provided by the government or other public bodies;
- Pensions, unemployment and welfare benefits, and other state benefits;
- Property, investment and share ownership, public or private;
- Interest rates and exchange rates;
- Regulation of economic activity, public or private;

- Relations between employers, workers and trade unions.

Left economic policies tend to favor one or more of the following:

- High levels of services provided by the government and state benefits, even if this implies high levels of taxation;
- Public investment. Public ownership or control of sections of business and industry;
- Public regulation of private business and economic activity;
- Support for workers/trade unions relative to employers.

Right economic policies tend to favor one or more of the following:

- Low levels of taxation, even if this implies low levels of levels of services provided by the government and state benefits;
- Private investment. Minimal public ownership or control of business and industry;
- Minimal public regulation of private business and economic activity;
- Support for employers relative to trade unions/workers.

Social policies deal with aspects of social and moral life, relationships between social groups, and matters of national and social identity, including:

- Policing, crime, punishment and rehabilitation of offenders;
- Immigration, relations between social groups, discrimination and multiculturalism;
- The role of the state in regulating the social and moral behavior of individuals.

Liberal social policies tend to favor one or more of the following:

- Policies emphasizing prevention of crime, rehabilitation of convicted criminals;
- The right of individuals to make personal moral choices on matters such as abortion, gay rights, and euthanasia;
- Policies penalizing discrimination against particular social groups and/or favoring a multicultural society.

Conservative social policies tend to favor one or more of the following:

- Policies emphasizing more aggressive policing, increasing police numbers, conviction and punishment of criminals, building more prisons;
- The right of society to regulate personal moral choices on matters such as abortion, gay rights, and euthanasia;
- Policies favoring restriction of immigration, and/or opposing explicit provision of state services for minority cultures.

Text Examples

Below we provide two examples of text from the manifestos and instructions on how they should be coded, and why.

Example 1: Right economic policy:

> With a Conservative Government, all that has been changing. We were determined to make share-ownership available to the whole nation. Just as with cars, television sets, washing machines and foreign holidays, it would no longer be a privilege of the few

The highlighted text should be coded as economic because it references ownership. In addition, the text is right because it is promoting private ownership.

Example 2: Liberal social policy:

> ALTERNATIVES TO PRISON. Every effort should be made to ensure that fine defaulters, elderly shoplifters and drunks are not sent to prison. Police cautions and intermediate treatment should be more widely used. Where punishment is appropriate, it should normally be community service rather than prison

The highlighted text should be coded as having to do with social policy because it references policing. In addition the text is liberal because it promotes alternative punitive measures to prison.

Bibliography

Acemoglu, D., S. Johnson, and J. A. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *The American Economic Review 91*(5), pp. 1369–1401.

Ahlquist, J. S. and C. Breunig (2012). Model-based clustering and typologies in the social sciences. *Political Analysis 20*(1), pp. 92–112.

Albert, R. and A. Barabasi (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74*(1), 47. Copyright (C) 2009 The American Physical Society; Please report any problems to prola@aps.org.

Albert, R., H. Jeong, and A.-L. Barabasi (1999, Sep). Internet: Diameter of the world-wide web. *Nature 401*(6749), 130–131. 10.1038/43601.

Alesina, A. (2005, September). Ethnic diversity and economic performance. *Journal of Economic Literature 43*, 762–800(39).

Arita, M. (2005). Scale-Freeness and Biological Networks. *J Biochem 138*(1), 1–4.

Axelrod, R. (2006). *The Evolution of Cooperation*. Basic Books.

Barabasi, A. and R. Albert (1999, October). Emergence of scaling in random networks. *cond-mat/9910332*. Science 286, 509 (1999).

Barkan, J. D., P. J. Densham, and G. Rushton (2006). Space matters: Designing better electoral systems for emerging democracies. *American Journal of Political Science 50*(4), pp. 926–939.

Bartholdi III, J. J., C. A. Tovey, and M. A. Trick (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare 6*(3), pp. 227–241.

Benoit, K., D. Conway, M. Laver, and S. Mikhaylov (2012). Crowd-sourced data coding for the social sciences: massive non-expert coding of political texts. In *Prepared for the third annual New Directions in Analyzing Text as Data conference, at Harvard University*.

Benoit, K., M. Laver, and S. Mikhaylov (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science 53*(2), pp. 495–513.

Benoit, K., M. L. Laver, and S. Mikhaylov (2007, September). Estimating party policy positions with uncertainty based on manifesto codings. In *Presented at the 2007 Annual Meeting of the American Political Science Association*.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (Summer 2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis 20*(3), 351–368.

Besley, T. (2009, September). The origins of state capacity: Property rights, taxation, and politics. *The American Economic Review 99*.

Besley, T. and T. Persson (2008). Wars and state capacity. *Journal of the European Economic Association 6*(2-3), 522–530.

Besley, T. J. and T. Persson (2010, January). State capacity, conflict and development. *Econometrica 78*, 1–34.

Bhavnani, R. and M. Ross (2003). Announcement, credibility, and turnout in popular rebellions. *The Journal of Conflict Resolution 47*(3), pp. 340–366.

Bianco, W. T., I. Jeliazkov, and I. Sened (2004). The uncovered set and the limits of legislative action. *Political Analysis 12*(3), 256–276.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003, March). Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

Bollobás, B. (2001). *Random Graphs* (Second ed.). Cambridge University Press.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America 99*(Suppl 3), 7280–7287.

Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30*(1-7), 107 – 117. Proceedings of the Seventh International World Wide Web Conference.

Buhrmester, M., T. Kwang, and S. D. Gosling (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science 6*(1), 3–5.

Cadsby, C. B. and E. Maynes (1999). Voluntary provision of threshold public goods with continuous contributions: experimental evidence. *Journal of Public Economics 71*(1), 53 – 73.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, Stroudsburg, PA, USA, pp. 286–295. Association for Computational Linguistics.

Cederman, L.-E. (2001). Modeling the democratic peace as a kantian selection process. *The Journal of Conflict Resolution 45*(4), pp. 470–502.

Cederman, L.-E. (2003). Modeling the size of wars: From billiard balls to sandpiles. *The American Political Science Review 97*(1), pp. 135–150.

Chamberlin, J. R. (1982). A mathematical programming approach to assessing the manipulability of social choice functions. *Political Methodology 8*(4), pp. 25–38.

Christakis, N. A. and J. H. Fowler (2007, July). The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine 357*(4), 370–379.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-law distributions in empirical data. *SIAM Review 51*, 661–703.

Coburn, N. (2010, May). Connecting with kabul: The importance of the wolesi jirga election and local political networks in afghanistan. White paper, Afghanistan Research and Evaluation Unit.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Condra, L. N., J. H. Felter, R. K. Iyengar, and J. N. Shapiro (2010, July). The effect of civilian casualties in afghanistan and iraq. Working Paper 16152, National Bureau of Economic Research.

Cordella, L., F. P, C. Sansone, and M. Vento (2001). An improved algorithm for matching large graphs. In *Proceedings of the 3rd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition*, pp. 149–159.

Epple, D. and R. E. Romano (1996). Public provision of private goods. *The Journal of Political Economy 104*(1), 57–84.

Epstein, J. (2007). *Generative Social Science: Studies in Agent-Based Computational Modeling.* Princeton University Pres.

Epstein, J. M. (2006). *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.

Epstein, J. M. and R. L. Axtell (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. A Bradford Book.

Erdos, P. and A. Rènyi (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen) 6*, 297, 290.

Erdos, P. and A. Renyi (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica 12*, 261–267.

Ergun, G. (2002). Human sexual contact network as a bipartite graph. *Physica A: Statistical Mechanics and its Applications 308*(1-4), 483 – 488.

Fowler, J. H. (2005). *Social Logic of Politics*, Chapter Turnout in a Small World, pp. 269–287. Temple University Press.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *3*, 215–239.

Gemenis, K. (2012). Proxy documents as a source of measurement error in the comparative manifestos project. *Electoral Studies 31*(3), 594 – 604. ¡ce:title¿Special Symposium: Economic Crisis and Elections: The European Periphery¡/ce:title¿.

Gerner, D. J., P. A. Schrodt, R. A. Francisco, and J. L. Weddle (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly 38*(1), pp. 91–119.

Gile, K. J. and M. S. Handcock (2010). Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology 40*, 285–327.

Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology 78*(6), pp. 1360–1380.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis 18*(1), pp. 1–35.

Grimmer, J. (2011). An introduction to bayesian inference via variational approximations. *Political Analysis 19*(1), pp. 32–47.

Hagberg, A. A., D. A. Schult, and P. J. Swart (2008, August). Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, and J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pp. 11–15.

Hamilton, D. T., M. S. Handcock, and M. Morris (2008). Degree distributions in sexual networks: A framework for evaluating evidence. *Sexually Transmitted Diseases 35*, 30–40.

Handcock, M. S. (2003). *Dynamic Social Network Modeling and Analysis*, Chapter Statistical Models for Social Networks: Inference and Degeneracy, pp. 229–240.

Hopkins, D. J. and G. King (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science 54*(1), pp. 229–247.

Hornblower, S. (1992). *Democracy: The Unfinished Journey (508 BC to AD 1993)*, Chapter Creation and Development of Democratic Institutions in Ancient Greece, pp. 1–17. Oxford University Press.

Hospedales, T., J. Li, S. Gong, and T. Xiang (2011, Dec.). Identifying rare and subtle behaviors: A weakly supervised joint topic model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

Ipeirotis, P. G., F. Provost, and J. Wang (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, New York, NY, USA, pp. 64–67. ACM.

Jackson, M. O. (2005). Allocation rules for network games. *Games and Economic Behavior 51*(1), 128 – 154.

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton University Press.

Jackson, M. O. and A. van den Nouweland (2005). Strongly stable networks. *Games and Economic Behavior 51*(2), 420 – 444. Special Issue in Honor of Richard D. McKelvey.

J.L., L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior 27*, 145–162(18).

Junttila, T. and P. Kaski (2007). Engineering an efficient canonical labeling tool for large and sparse graphs. In *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithms and Combinatorics*.

Kerckhoff, A. C. (1995). Institutional arrangements and stratification processes in industrial societies. *Annual Review of Sociology 21*, pp. 323–347.

Khilnani, S. (1992). *Democracy: The Unfinished Journey (508 BC to AD 1993)*, Chapter India's Democratic Career, pp. 189–206. Oxford University Press.

Kidd, Q. (2008, 6). The real (lack of) difference between republicans and democrats: A computer word score analysis of party platforms, 1996–2004. *PS: Political Science & Politics 41*, 519–525.

Kim, D.-H. and H. Jeong (2006). Inhomogeneous substructures hidden in random networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 73*(3), 037102.

Kim, J. S., K. I. Goh, B. Kahng, and D. Kim (2007). A box-covering algorithm for fractal scaling in scale-free networks. *cond-mat/0701504*. Chaos 17, 026116 (2007).

Kim, J. S., K. I. Goh, G. Salvi, E. Oh, B. Kahng, and D. Kim (2006, May). Fractality in complex networks: critical and supercritical skeletons. *Phys. Rev. E 75, 016110 (2007)*..

Kim, S.-y., C. S. Taber, and M. Lodge (2010). A computational model of the citizen as motivated reasoner: Modeling the dynamics of the 2000 presidential election. *Political Behavior 32*(1), pp. 1–28.

King, G. and W. Lowe (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization 57*, 617–642.

Kittur, A., E. H. Chi, and B. Suh (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, New York, NY, USA, pp. 453–456. ACM.

Klingemann, H.-D., A. Volkens, J. Bara, I. Budge, and M. McDonald (2006). *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003*. Oxford University Press.

Laver, M. (1998). Party policy in britain 1997: Results from an expert survey. *Political Studies 46*(2), 336–347.

Laver, M. and K. Benoit (2003). The evolution of party systems between elections. *American Journal of Political Science 47*(2), 215–233.

Laver, M. and B. W. Hunt (1992). *Policy and party competition*. Routledge.

Laver, M. and E. Sergenti (2011). *Party Competition: An Agent-Based Model*. Princeton University Press.

Leskovec, J., J. Kleinberg, and C. Faloutsos (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, New York, NY, USA, pp. 177–187. ACM.

Levitsky, S. and G. Helmke (2006). *Informal Institutions and Democracy: Lessons from Latin America*. Baltimore: Johns Hopkins University Press.

Lin, N. (1999). Social networks and status attainment. *Annual Review of Sociology 25*, pp. 467–487.

Lintott, A. (2003). *The Constitution of the Roman Republic*. Oxford University Press.

Marks, M. and R. Croson (1998). Alternative rebate rules in the provision of a threshold public good: An experimental investigation. *Journal of Public Economics 67*(2), 195 – 220.

Mason, W. and S. Suri (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods 44*, 1–23.

McClurg, S. D. (2003). Social Networks and Political Participation: The Role of Social Interaction in Explaining Political Participation. *Political Research Quarterly 56*(4), 449–464.

Mikhaylov, S., M. Laver, and K. R. Benoit (Winter 2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis 20*(1), 78–91.

Miller, N. R. (2007). In search of the uncovered set. *Political Analysis 15*(1), pp. 21–45.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis 16*(4), pp. 372–403.

Newman, M. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics 46*(5), 323–351.

Newman, M. E. J. (2001, Jun). Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E 64*(1), 016131.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Newman, M. E. J., S. H. Strogatz, and D. J. Watts (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E 64*.

Padgett, J. F. and C. K. Ansell (1993). Robust action and the rise of the medici, 1400-1434. *The American Journal of Sociology 98*(6), 1259–1319.

Pain, A. (2010, February). Afghanistan livelihood trajectories: Evidence from badakhshan. White paper, Afghanistan Research and Evaluation Unit.

Palfrey, T. R. and H. Rosenthal (1984). Participation and the provision of discrete public goods: a strategic analysis. *Journal of Public Economics 24*(2), 171 – 193.

Paolacci, G., J. Candler, and P. G. Ipeirotis (2010, 08). Running experiments on amazon mechanical turk. *Judgement and Decision Making 5*(5).

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*(1), pp. 209–228.

Ramage, D., D. Hall, R. Nallapati, and C. D. Manning (2009). Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, Stroudsburg, PA, USA, pp. 248–256. Association for Computational Linguistics.

Rand, D. G. (2012). The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology 299*(0), 172 – 179. ¡ce:title¿Evolution of Cooperation¡/ce:title¿.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks 29*(2), 173 – 191. Special Section: Advances in Exponential Random Graph (p*) Models.

Rondeau, D., W. D. Schulze, and G. L. Poe (1999). Voluntary revelation of the demand for public goods using a provision point mechanism. *Journal of Public Economics 72*(3), 455 – 470.

Ross, J., L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, New York, NY, USA, pp. 2863–2872. ACM.

Sandler, T. and W. Enders (2004). An economic perspective on transnational terrorism. *European Journal of Political Economy 20*(2), 301 – 316. The Economic Consequences of Terror.

Scholz, J. T. and C.-L. Wang (2006). Cooptation or transformation? local policy networks and federal regulatory enforcement. *American Journal of Political Science 50*(1).

Schrodt, P. A. and D. J. Gerner (2012). *Analyzing International Event Data (2001/2012).*

Siegel, D. A. (2009, January). Social networks and collective action. *American Journal of Political Science 53*, 122–138(17).

Simon, A. F. and M. Xenos (2004). Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis. *Political Analysis 12*(1), pp. 63–75.

Slapin, J. B. and S.-O. Proksch (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science 52*(3), pp. 705–722.

Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, Stroudsburg, PA, USA, pp. 254–263. Association for Computational Linguistics.

Song, C., S. Havlin, and H. A. Makse (2005). Self-similarity of complex networks. *Nature 433*(7024), 392–395.

Sorokin, A. and D. Forsyth (2008). Utility data annotation with amazon mechanical turk. In *Proceedings of the Computer Vision and Pattern Recognition Workshops*, CVPRW '08, pp. 1–8. IEEE.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods 43*, 155–167.

Steglich, C., T. A. Snijders, and M. Pearson (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology 40*, 329–393.

Sylvan, D. A., J. W. Keller, and Y. Z. Haftel (2004). Forecasting israeli-palestinian relations. *Journal of Peace Research 41*(4), pp. 445–463.

Taber, C. S. and R. J. Timpone (1996). Beyond simplicity: Focused realism and computational modeling in international relations. *Mershon International Studies Review 40*(1), pp. 41–79.

Tangian, A. S. (2000). Unlikelihood of condorcet's paradox in a large society. *Social Choice and Welfare 17*(2), pp. 337–365.

Thies, C. G. (2009). National design and state building in sub-saharan africa. *World Politics 61*(4), 623–669.

Titov, I. and R. McDonald (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, New York, NY, USA, pp. 111–120. ACM.

Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *J. ACM 23*(1), 31–42.

Watts, D. J. and S. H. Strogatz (1998, June). Collective dynamics of 'small-world' networks. *Nature 393*(6684), 440–442. PMID: 9623998.

Weeks, M. R., S. Clair, S. P. Borgatti, K. Radda, and J. J. Schensul (2002). Social networks of drug users in high-risk sites: Finding the connections. *AIDS and Behavior 6*, 193–206. 10.1023/A:1015457400897.