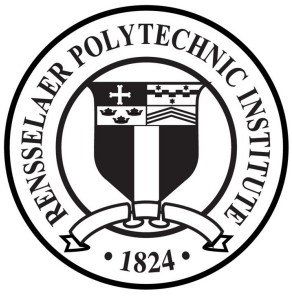
INTRODUCING NON-MARKOVIAN AND EMPIRICAL EFFECTS INTO SOCIAL INTERACTION MODELS

Casey Doyle

Submitted in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Approved By:
Gyorgy Korniss, Chair, Dissertation Adviser
Bołeslaw Szymański, Dissertation Adviser
Vincent Meunier, Member
Humberto Terrones, Member



Department of Physics, Applied Physics, and Astronomy
Rensselaer Polytechnic Institute
Troy, New York

[August 2018]
Submitted July 2018

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

This work was supported in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, (the ARL Network Science CTA) and by the Office of Naval Research (ONR) Grant Nos. N00014-09-1-0607, N00014-15-1-2640. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the Army Research Laboratory or the US Government.

My time at RPI has been influenced by many individuals and organizations, all of whom have helped me get to where I am. There are, however, a few individuals that I would like to thank whose contributions stand out above all others.

First, I would like to thank my committee members for their time and efforts to shape my work into its final form. Dr. Vincent Meunier, who has been an adviser of mine since I was an undergrad. His thoughtfulness and enthusiasm has helped me feel at home at RPI from my very first day. Dr. Bolesław Szymański, who I have known since I began my first research project. Not only has his center (the Social Cognitive Networking Academic Research Center) given me opportunities to explore new fields and ideas I otherwise would have never found, but his friendly nature and intellectual creativity have helped me solve even the most stubborn challenges I have faced. Dr. Humberto Terrones, who has offered a valuable outside perspective to my studies. And finally Dr. Gyorgy Korniss, who was my very first professor at RPI and has helped guide me through all the hurdles of graduate studies. Without their help and perspectives, this work would never have been completed.

Second, I would like to thank my collaborators; Derrik Asher and the Army Research Laboratories as well as Zala Herga and the Jožef Stefan Institute. Their work in data collection and experimental design has provided much of the empirical data used here, and their contributions in analyzing that data have helped broaden the scope of my work.

Finally, I would like to thank my family back home. They have been extraordinarily patient and helpful despite me living across the country from them, and are always there to listen and offer advice when I need it. The support of my parents, Lynne and David Doyle, as well as my sister Shelby Doyle, is the only reason I was able to make it here in the first place and their contributions to my life have only grown since then.

ABSTRACT

Stochastic models of opinion spread are a popular method for simulating and predicting the social behaviors of large populations. Though classical models in this field have proven to be accurate towards their intended purpose, often they fall short when applied to more specific scenarios. Many of the assumptions made in these base models have proven to be quite different from the natural behavioral patterns of real people, making further updates and extensions of the original models imperative to understand these shortcomings. This work presents two such model extensions, building off of the basic examples of the naming game and voter models to create more in depth systems and describe complex phenomenon.

The first of the two models presents a system in which opinions maintain a set inertia value that dictates the degree to which a node holding that opinion will resist switching opinions. The second replaces the speaker selection mechanic to allow for non-exponential waiting time distributions that vary the activity patterns of the nodes. In both of these scenarios it is shown that the symmetry of the system is broken, creating well defined tipping points where the advantaged opinion is able to build a consensus quickly and consistently. Further, despite both extensions breaking the Markov property maintained in the more basic models, analytic approximations that accurately describe the behavior of the systems are provided.

Finally, in addition to the new model extensions, a brief overview of relevant empirical investigations is provided to inform on future work in this area. First, data mining techniques are employed to find frequent response patterns in a large survey data set on opinion formation with regards to media consumption. These results serve to identify both groups of individuals that behave similarly and the general trends that shape their responses. Then, a large scale cell phone data set is analyzed for its capability to provide an empirical social network. Two separate network building schemes are compared and used to provide effective networks that may serve as the setting for future simulations.

CHAPTER 1
INTRODUCTION

The spread of ideas through large populations is a subject that is deeply intertwined with how societies change and evolve. This process controls far more than just the spread of opinions; it defines the growth of groups through innovation, language, religion, and politics. Governments rise and fall based on public opinion, and societal functions shift based on what people know. It is difficult to overstate the importance of such a field to understanding how people behave, yet challenges in designing studies make it somewhat poorly understood. Behavior is not deterministic and is thus hard to generalize, yet a more comprehensive description presents its own issues due to the large scales on which opinion spread occurs. Fortunately, the modern world offers new avenues to bridge the gap between large scale modeling and individual behavior patterns. As complex computations become more viable, models of human behavior are able to become more intricate and can better approximate behavior on large scales. Meanwhile, on the individual scale it is constantly becoming easier to study human behavior. As cell phones and online communication increase in popularity and bring with them detailed documentation of human interactions, new avenues open up for a more detailed and comprehensive look into human communication than has ever been possible before. By connecting these two fields, high performing models of behavior can be achieved, both describing the major contributions that lead to past events and creating a better understanding for the future.

Before wading into the details of how these models are structured and behave, an understanding of the major social phenomenon that they look to capture is necessary. For the large scale stochastic models discussed in this paper, the main focus is on the propensity of social systems to exhibit tipping points leading to rapid state changes in public opinion [1]. These tipping points can be seen in many different areas of human behavior, such as innovation where the number of individuals adopting new technologies tends to follow an S-curve over time. Societies tend to exhibit a rapid gain in technology adoption once a certain threshold is met that only slows once dominance has been established and the society reaches the final relaxation stage where small stubborn communities affect progress [2]. Further modern advancements show tipping points throughout online media as emotional and

contextual bursts [3]. In this case there are distinct mechanisms that lead to these sentiment bursts within communities such as excessive negative activity in indirect communication and high arousal states with external influence for direct communications. The central theme of cascading emotions and viewpoints when certain conditions are met remains the same, though.

Tipping points can also been seen in many historical events where social and political movements boil over and push societies towards rapid change. Large scale societal changes such as the American civil rights movement and women's suffrage are both important examples of committed minority activist groups formulating the right strategies to push society past its stagnation and change widely held beliefs in a relatively short time [4], [5]. More recently, the prevalence of social media has made it easier to track these sorts of movements, and in fact has played a large role in organizing and pushing many of them to fruition creating a whole new type of grass roots movement termed "cyberprotests" [6]. Specifically, campaigns such as the Occupy movement of 2011 give stark examples of how even ideas that have existed for long periods of time but are not in favor can explode onto the scene and create large scale effect so long as they trigger the right catalysts along the way [7]. Additional examples such as the Arab Spring and protests against Guatemalan president Álvaro Colom show the power of these new cyber movements to not only bring their ideas rapidly to the attention of the masses, but cause rapid government upheaval and leadership changes [8], [9].

Upon establishing the pattern of tipping points in social systems throughout history, the next logical step is to define the catalysts that cause these situations. This is an ambitious undertaking, and really gets to the heart of what this document is about. One method of studying this problem is observing the phenomenon that commonly surround social unrest, such as natural and economic disasters [10]. In these cases, the political landscape is shaken, and the movements of various political groups with regards to the rights and liberties of their people are magnified and have the potential to either solidify or end regimes. A similar driving effect can be seen in the context of opinion dynamics in online communities, where large scale events can force an idea into the collective consciousness quickly and create a widespread discussion [11]. In these sorts of communities, rare resonances can be identified that lead to explosions of opinions and discussion over a particular topic far beyond what would be expected normally. While these can on occasion occur with no identifiable event

leading into them, they are far more commonly caused by a singular important event that starts the conversation, then reach the resonance point and take over discussion from there. Despite being clear catalysts for tipping points, though, these various high-effect events or disasters are by no means the only ways to push societies towards rapid change. In fact, the propensity for societies to experience such events without constant upheaval indicates that there is more contributing to these states than just the initial push; societies must be in a position already where they are susceptible to changes. To this end, various "early warning signs" can be identified to predict the likelihood of a community to tip over [9]. These signs include aspects of the community structure and opinion state, where properties such as high heterogeneity, certain connectivity patterns, and lowered individual thresholds for change can prime a society.

Many of these identified empirical traits mirror common theoretical model parameters, allowing for the creation of toy systems that can be used to study the fine details that put a society most at risk. Such models often begin as simple stochastic models of social systems, studied in-depth by mathematicians to understand the random processes at work in human-based systems [12], [13]. This naturally leads to a connection with epidemiology, which is also commonly studied as a stochastic spreading system and shares many common and convenient elements with opinion spread [14]. In fact, the same cascade behavior that is observed in social tipping points is also seen in the large scale outbreaks that are characteristic of epidemics, indicating that these social effects can be recreated using an epidemiological framework [15]. As a result, ideas are frequently treated as a sort of "social contagion" since this framework provides a good approximation of real behavior as well as a natural intuition into how ideas can spread [16]. The applications of this approach are particularly relevant to the world of modern marketing and campaigning, where various methods for maximizing spread for minimal cost have been developed, exploiting the ability of viral marketing to rapidly spread concepts to large numbers of people [17].

Using these spreading focused models, the study of real world tipping points and mechanics can be brought into the theoretical realm, both offering predictions around the specific parameters required for system change and allowing for the tuning of models to fit the systems of greatest interest. For instance, a basic model might include a very simple rule set that describes how an idea jumps from one person to another that works over a large group in a fairly generic scenario, but the question then becomes whether the conclusions

from this simple model still hold true when the system is not generic. How does the model behave if you are trying to spread an opinion in a system with a different underlying network structure, or where people tend to be far more skeptical of each other? What about a system with a small group of people that behave differently from the masses to sway the larger group towards one side or the other? How large would this group have to be, and how should they behave to optimize their actions? These are the questions that this work addresses on a basic, theoretical level. Understanding how various facets of behavior can alter the eventual outcome is of the utmost importance to predicting and influencing the evolution of society, and the theoretical underpinnings characteristic of stochastic models represent an elegant solution to this problem.

It should be clear, though, that these sorts of experiments do not normally extend to attempting to create a comprehensive model of human dynamics capable of explaining every facet of behavior; such a model is unfeasible in its complexity [18]. Instead, there is a careful balancing act at work here as new features are added to well understood frameworks to extend the knowledge base in a given scenario without complicating the model so much that it is beyond mathematical description. In this work, this balance is carefully maintained. The pre-existing models are adjusted in ways that break their basic Markovian nature, but this is done in such a way that viable approximations still exist to describe the systems analytically. This back and forth of complexity versus analytic capabilities allows these models to capture elements of human behavior that have largely been neglected due to the difficulties they present in the underlying mathematics and presents the opportunity to connect this abstract theory to empirically derived data.

CHAPTER 2
BACKGROUND

As alluded to in the previous chapter, computational modeling of social dynamics is a prolific field; the importance of finding proper descriptors for the many different factors that make up human communication and decision making lead to innumerable distinct models [19], [20]. Many of these models seek to amplify certain processes or features of human behavior, placing them under a microscope and seeing how they affect the dynamics on a larger scale. Due to this, the various computational models have become somewhat fragmented as each seeks to strike their own balance of accuracy and complexity. There are, however, a few basic models that are very well understood and are considered classical to the field. These are the models that are most often modified to fit specific use cases. The high degree of understanding of these base cases allows for a very fine tuned approach to understanding how large the effects of different changes are. In general, these basic models strive to either improve upon one another or simply take different approaches to the subject and are thus so distinct that they can coexist naturally. This section covers three different basic model cases: the threshold model [21], the voter model [22], and the naming game [23]. All three of these models focus on consensus dynamics, studying deeply how tipping points and cascades occur within systems with various structures and behaviors.

The first model, the threshold model, is distinct from the others in that it focuses on network structure and the herding mentality of people [21]. This model seeks to describe personal choices that are heavily affected by peer pressure and the actions of others, a common factor in decision making where perceived social hierarchies or generalized standards are prominent [24]. While this type of model is not the focus of this paper, a basic understanding is helpful for interpreting the results of the dynamics based models that are presented. These threshold models highlight some important features of behavior, showing that the cascading behavior of social interactions can have many different unique drivers and can be seen even in otherwise unconnected models of human behavior.

The later two models, the voter model and the naming game, are both more dynamics based models that look at ideas as invasive processes while the individuals within the system are treated similarly to particles [22], [23]. These models are stochastic, treating interactions

as random occurrences following specific rules that move the particles into and out of different states. In fact, these models can be entirely separated from any true network structure and run on complete graphs to get a view of how the system behaves in a mean-field consideration, focusing entirely on the interaction rules and their effects on the time scaling and behavior of the system. The majority of this work focuses on modified versions of these two models and the interaction rules that they put into place, challenging some of the key assumptions of the basic models and modifying them to fit better with the current understanding of empirical human behavior.

2.1 Threshold Model

The threshold model focuses on a herd mentality decision process, updating individual node states based on the opinions of their neighbors [21]. In this model, if the percentage of a selected node's neighbors holding a given opinion is higher than the preset threshold, then the node is given that opinion. This update process lends itself naturally to structural questions of information cascades that seek to determine which network structures and node locations make the network most susceptible. These questions have been rigorously studied, showing that optimal initiators are better determined by network location than more intuitive features such as the degree of the node [25]. Further, the total volume of influencers has been shown to be far more important than location, and thus the question of node location is focused more on minimizing the number of initiators rather than optimizing single node starting locations [26]. Despite the heavier focus on structure, though, there are still efforts that dive into different behavior aspects of spreading processes. For instance, the behavior of the nodes can be altered to include multiple initiators [27], heterogeneous thresholds, and other modifications [28], [29].

The general focus of the threshold model and its derivatives are important to understanding how some innovations and spreading mechanics are affected by societal pressures and group-think, but they are not suitable for many other aspects. When studying many marketing, political, or opinion competition scenarios, it often makes more sense to look at individual interactions as seen in the voter model and naming game. These models are *agent based*, and are able to capture the same tipping point behavior as the threshold model for more specific scenarios of human interaction.

2.2 Voter Model

The first of the two agent based models discussed here is the voter model, where spreading is treated as a very simple stochastic process that has its roots in statistical physics and contact processes [22]. In the voter model, the system is defined by a set of nodes that are each given a particular state and allowed to interact as a means of changing their states. Commonly, the voter model deals only with binary states, but it can be generalized to have any number of possible states. Interactions occur by choosing a node at random and updating its state to match the state of one of its neighbors (which are defined by the network structure underlying the system). As such, it can be simplified and analyzed on a complete graph to remove the structural aspects from the analysis entirely, or combined with various network models to investigate their effects on the simulation outcomes. This transition rule defined behavior allows the model to be described succinctly via its Markov transition matrix (defining the probabilities that the system will switch from one state to another), and eventually solved to find exact solutions for many interesting parameters of the system [30]. From this analysis, the expectation value for the consensus time when the total possible number of states m is equal to the system size N is shown to be $< T_c >= (1/N)(N-1)^2$. The scaling in this result holds for not only the base case of a complete graph, but also for Erdős-Rényi (ER) random graphs and sparse networks. Additionally, it confirms other less exhaustive studies into the nature of the consensus time [31]. Similarly, this solution can be leveraged to calculate many other quantities, such as the variance in the consensus time (which scales as N^2) and even the expected number of states in the system at a given time (which decays as an inverse of time).

Such a thorough understanding of the base case leads to an interesting building block for more in depth investigations, altering portions of the model to understand the differences that arise. One of the most popular and fruitful ways to do this is to study the voter model applied to different, more complex network structures. This can bring the model either back to its statistical physics roots in the form of studying the behavior on regular lattices, or bring it further into the realm of social dynamics by understanding the impact of social networks and community structures on the system. To date, both of these scenarios have been studied thoroughly. In the simple cases of lattices up to two dimensions, the voter model is able to create order solely through the propagation of opinions via noise across the interfaces [32]. On small-world, community-driven networks, however, the voter model

is unable to reach consensus in the infinite system size limit, instead remaining trapped in an active steady state whose escape time diverges with the system size [33]. Interestingly, when applied to finite systems, the small-world networks allow for a *faster* consensus than is seen for regular lattices. These counter-intuitive results indicate strong effects on the model from network topology, leading to even further study on various structures. For instance, on uncorrelated networks of arbitrary degree, sparse networks ($\langle k \rangle \leq 2$) show exponentially fast consensus times, but when the average degree is greater than two the system gains an active steady state that again diverges and becomes stationary in the infinity system size limit [34]. To connect all of these different structural effects, the mathematics behind the voter model have been studied for arbitrary complex networks of finite size to show that the ordering mechanism of the voter model can be exactly described by a single diffusion equation [35]. In fact, the spatial structures involved can simply be represented via a scaling to a new "effective" system size within the mean-field equations to fully describe the system.

2.3 Naming Game

The naming game is fundamentally very similar to the voter model, except with some added complexity and very different roots. This model attempts to model consensus dynamics as via language spreading, in the base considering different dialects instead of opinions. Similar to prior models, however, it models the spread of languages as invading processes, solving the system for various linguistic behaviors to take advantage of the convenience and simplicity that this framework provides [36]. The naming game is one of the most common models to come out of the combination of linguistics and computational simulations, notable in its success at recreating self organization of systems via simple pairwise interactions [23], [37]. The result is a very simple model for linguistic order, starting with the idea of a large group of N autonomous agents that are introduced to a new object that none are familiar with. In the intuition behind the original model, each agent comes up with a name or word that they associate with that object, then interact with each other to reach a consensus. As the individuals within the system interact with each other, they slowly agree on what the object should be called and the system is returned to order. The interaction rules that the game follows are similarly simple; each node in the system has a list of words which they know are associated with the object. Classically, at the start of the game, each node has only one word in their list and it is unique from all the words each other node has.

Then, for each interaction a random node is chose to be the "speaker", and that node shares a random word from its list with one of its neighbors (the "listener"). If the listener does not already have that word, the word is simply added to the listeners list. If, however, the listener does already have the word, then the speaker and listener are in agreement and both remove all other words from their lists. These dynamics make for stages of growth and elimination, as at early times it is extremely unlikely that the listener will already have any word shared with it and the average number of words each agent holds grows quickly. Eventually, the word lists will be large enough that it becomes more likely than not that the speaker's chosen word will already be in the listener's vocabulary, and the system will go through a slow elimination process until consensus on a single word is achieved [38]. Depending on the initial conditions, however, these transitions become sharp and the system undergoes rapid phase changes towards total consensus [39].

For the purposes of opinion dynamics, the large numbers of possible words within the standard naming game are often considered superfluous. It is often far more convenient to look at the direct competition between two opinions and find the behaviors that can make one dominant over the other while also drastically reducing the consensus times and general system complexity [40]. On complete graphs, the time reduction due to this simplification has been shown asymptotically, via simulations, and through mean-field approaches to bring the system to only $\mathcal{O}(\ln(N))$ for the binary model [31], [38], [41]. In comparison, the standard naming game is $\mathcal{O}(N^{1/2})$ to reach order [39], [42].

Of course, just like the voter model previously discussed, the naming game dynamics can show different behavior patterns when run on different network topologies, making for many popular avenues of study. For example, in low dimensional networks ($d \leq 4$), the consensus time has been shown to scale as $\mathcal{O}(N^{2/d})$ [42]. In more detailed network structures such as small-world networks, the system breaks down into two different time regimes with a crossover time scaling similarly to the case of a complete graph with some additional dependence on the nature of the small-world rewiring scheme [43]. Before this transition point, the system behaves similarly to the naming game on one dimensional networks due to the tight clustering, but in the long time regime it reverts to mean-field behavior; essentially it starts very slowly but with low memory requirements, and then speeds up and converges towards consensus quickly. In fact, this topology has been shown to be the most efficient setting for the naming game, both in terms of average memory use (list size) for the nodes

and consensus times [44]. In some cases, simply adding small-world effects to unrelated networks such as random geometric graphs can lead to low memory usage and fast scaling (reaching consensus in $\mathcal{O}(N^{0.4})$, considerably faster than similar low dimensional networks without small-world effects) [45]. On other complex networks with small-world properties such as ER and Barabási-Albert (BA) graphs this effect remains, scaling in consensus time as $\mathcal{O}(N^{1.4\pm0.1})$ which is again faster than low dimensional lattices while maintaining lower memory requirements of the nodes [46]. This is not to say that community structure in inherently beneficial to spreading, however, as studies on the naming game on social networks with strong community structure show a hindrance to overall consensus formation as small pockets of opinions form within the communities and can survive indefinitely on large networks [41].

2.4 Models with Committed Agents

The simplicity of the models discussed thus far is, in many ways, both their greatest strength and weakness. Many of the complete and exact solutions to various interesting quantities are made vastly easier to find due to the careful design of the models. These designs simplify the underlying mathematics, but also create limitations as far as the applicability to many real world situations. For instance, many of the above models would describe the natural spreading processes of an idea with little outside influence pushing it in one direction or another, but many of the most interesting aspects of real world tipping points is how they can arise from dedicated groups supporting the idea. To this end, many different versions of committed agent models have been proposed, often building off of the base cases of the other more generalized models. The common thread among committed agent models is including a small subset of the network that refuses to change their opinion regardless of opposing interactions (or other update processes as described by the base model). For instance, committed agents in non-pairwise interaction models such as the Galam model (where small groups update based on local majorities), change the dynamics as expected with the number of propagators required for consensus decreasing depending on the portion of those individuals that are committed [47]. However, the model is not entirely straight forward, as there exists a critical proportion of committed agents, $p_c = 0.17$, that guarantees consensus for the committed opinion regardless of initial conditions.

Similar results can be seen in pairwise interaction models such as the naming game,

where critical populations can be seen transitioning the system to states where certain outcomes are not possible [48]. In this case, the critical population p_c of committed agents designates the population at which the opinion favored by the committed agent dominates and it is not possible for any other opinion to have more supporters. This critical point also shows resilience to network structure, as the population remains fixed regardless of dimension. Further, in the case where there are multiple committed populations supporting opposing opinions, a critical point p_c still exists, but it instead represents the population at which a "stalemate" is the only possible outcome and the value of p_c decays logarithmically as the number of opinions grow.

These phase transitions in the system dynamics can be simplified and studied in a more focused matter by applying them to the binary version of the naming game as well [49]. In this case, the system is generally set up such that the small fraction of nodes that are committed are the *only* nodes initially propagating their opinion, and the goal is to obtain the minimum population required to build a consensus for their opinion. Of course, in the case of infinitely committed nodes, this consensus will always eventually occur since a consensus on the committed state is the only absorbing state in the system. Any other active steady state will always still contain the committed nodes advocating for their opinion, allowing for a large random fluctuation to eventually push the system towards that consensus. In this case, a "success" for the committed agents is defined by the system time taken to reach this absorbing point. At the critical population of $p_c \approx 0.10$, the consensus time undergoes a sharp transition; beneath this critical value, the system is extraordinarily slow and takes $\mathcal{O}(\exp(N))$ time to reach consensus, while beyond the critical point the consensus time undergoes a discontinuous change and scales as $\mathcal{O}(\ln(N))$. Further, via analysis of the mean-field equations, this phase change can be viewed more acutely. Below the critical population there are three fixed points in the system, one active steady state that corresponds to the committed nodes in their opinion and all others opposing, one saddle point corresponding the requisite number of nodes swayed by the committed nodes to push the system over the edge, and finally the absorbing state of all nodes following the committed opinion. When the committed population is taken above the p_c, however, the first two fixed points are entirely removed and the system is always biased to move quickly towards consensus regardless of other system conditions. On sparse networks, the effect is amplified and the critical point is lowered. Finally, competing committed groups also maintain the phase transitions, creating

an interplay between the committed population sizes that determines whether the system remains in an active steady state with coexisting opinions or moves quickly towards a stable state of single opinion dominance [50].

These models (in particular the cases of committed agents within the naming game) are critical to much of the analysis in this work, and are referenced multiple time in the coming chapters as different tweaks to the underlying naming game rule set are studied for their effects on the value of p_c. Understanding exactly how large activist populations must be and what behaviors affect them the most is extremely important to describing the tipping point behavior in real world networks, as many of the biggest and most influential changes to society are caused by such committed individuals attempting to spread their ideals and innovations to others.

2.5 Waning Commitment

Models of opinion commitment can be considered as an extreme case of stubbornness in stochastic systems, an idea that has been explored in several other ways as well [51], [52], [53]. Despite its computational advantages, however, the extreme case of infinite commitment also proves to be a limiting factor to the realism of the models. While in some scenarios such as civil rights movements infinite commitment makes sense, in many other social, political, and even marketing applications it is more natural to study the scenarios where individuals have finite commitment and may lose their dedication after repeated failed interactions with others. Thus, the most obvious and natural extension of commitment and stubbornness into a less extreme context is simply to introduce a finite, waning commitment level for nodes [54], [55]. To this end, the *waning commitment* model examines the impact of different commitment levels, altering it so that it can be lost (or gained) based on interactions with others. Specifically, after a committed node hears w consecutive opposing opinions, they lose their commitment and become a normal node that changes states according to the standard binary naming game rules. The addition of finite commitment forces a new state-space on the system to keep track of how close each committed node is to losing its commitment. Whereas in the original binary agreement model there are only three possible states (A, B, and AB), the addition of commitment strength creates multiple substates, A_0, ..., A_w, for the consecutive interaction counts of committed nodes. A visualization of the consequent Markov chain for this model can be seen in Fig. 2.1.

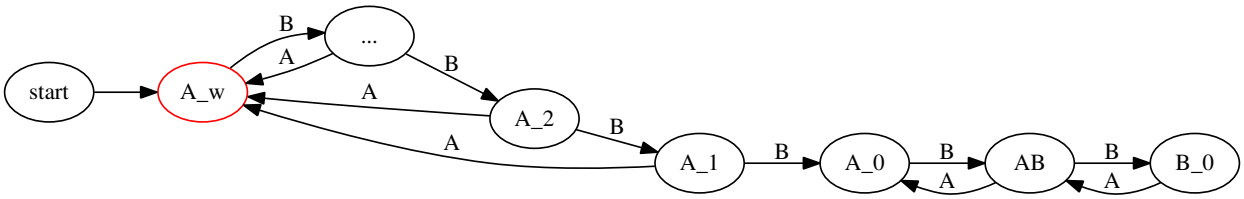


Fig. 2.1: Markov chain visualization for the *waning commitment* model for nodes committed to opinion A [55]. The current commitment level, i, of a node in state A_i corresponds to the number of consecutive B opinions that node would need to hear before losing commitment. Upon losing commitment, the node behaves according to the standard binary naming game rules while any confirming interactions remove all progress towards losing commitment.

By carefully defining each possible state transition in the system, Niu et. al. [55] solve the mean-field equations for a robust analytic solution to fixed points within the system. The main finding is that there still exists a population of committed agents that creates a phase transition within the system similar to that found in the standard committed agent model, but in this case the critical population relies heavily upon the commitment strength of the minority population. In prior work, it was suggested that this relationship is closely approximated by a power law due to simulations on low w systems, but the more recent results show that the relationship is far closer to an exponential decay of the critical population with increasing w when the analysis is extended to larger values of w (Fig. 2.2). Further, it is shown that the system quickly collapses to a critical population very near that of the standard committed agent model for relatively low values of w.

This model is of particular importance in motivating [54] and building off of [55] the work presented in Chap. 3, where the idea of incrementally pushing nodes towards new states is explored in another setting [54], [55]. The persistent existence of tipping points as well as their sensitivity to various commitment levels is also extremely important to overall motivation of this work, as it provides a clear map for devising strategies of optimal opinion spread, especially when combined with empirical data such as that presented in Chap. 5.

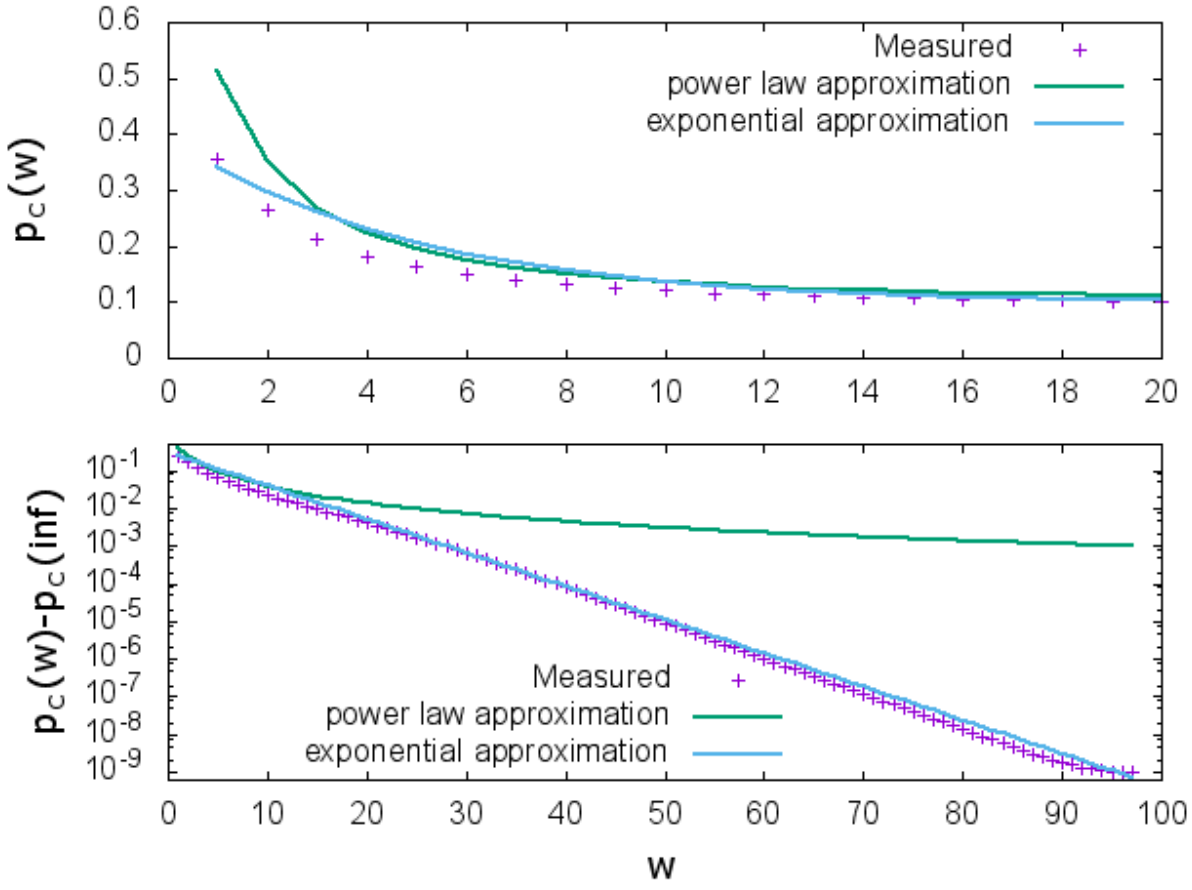


Fig. 2.2: **The scaling of the critical population of committed nodes required to force consensus on the system in the *waning commitment model* [55]. Both proposed fits (the original power law and more recent exponential) are shown. The top figure shows how the value of the critical population changes with the commitment level w while the bottom figure shows the change induced by the waning commitment of the nodes and is extended to extreme values of w and put on a log-normal scale to emphasize the better fit of the new results.**

CHAPTER 3
OPINION INERTIA

3.1 Motivation and Related Work

The works discussed in the previous chapter describe the many ways in which computational models can be adapted to study various aspects of human behavior, but due to the highly complex nature of the subject there will always be more modifications that deserve further attention. For example, finite commitment models such as those seen in [55], [54] simulate how individuals themselves are dynamic in their pliability to change based on their activity. The base idea of these models can be extended, though, as in many ways individual behavior itself is subject to some inertia that opposes any change in the beliefs or opinions adopted by the individual [56]. A well known example of such inertia is the phenomenon of confirmation bias in social psychology, where individuals tend to favor beliefs that conform to their currently held position. Overcoming such individual inertia to change is therefore a primary consideration in campaigns for public opinion change.

Most of the systems that have investigated this phenomenon rely on very balanced initial positions as discussed in the previous chapter, heavily relying on the symmetrical states within the system to study the changes that can be made by simple modifications. In many real world scenarios, however, the competing opinions, products, or innovations are not symmetric and one side is easier to understand or more palatable, and thus easier to convince people to adopt. This suggests that investigations into the inertial nature of people's opinion should not be linked only to the stubbornness inherent to the individual, but also encompass a specific *opinion inertia* that is tied to the opinion itself.

Motivated by this phenomenon, this chapter studies a theoretical model of opinion change similar to that of the *waning commitment* model where each node's state transitions depend upon the current state of the individual as well as the recent history of the opinions they have encountered in interactions with their neighborhood. Specifically, the system has two opinions vying for adoption on a social network, and each individual requires a

Portions of this chapter previously appeared as: C. Doyle, S. Sreenivasan, B. Szymanski, and G. Korniss, "Social consensus and tipping points with opinion inertia," *Physica A: Statistical Mechanics and its Applications*, vol. 443, pp. 316–323, Feb. 2016.

pre-defined threshold number of interactions with the alternative opinion before switching. Thus, each opinion is sticky to its respective extent. Furthermore, in an attempt to capture the effect of confirmation bias, each individuals memory of a stream of encounters with the alternative opinion is erased by a single interaction in which they encounter their currently held opinion.

There are other works devoted studying similar memory-based model of switching between states that this work draws upon. Dodds and Watts studied a model of disease contagion where a susceptible person became infected only when his interactions with infected neighbors within a certain prior time window had led to a pre-defined infection-dosage threshold being exceeded [57]. More pertinently to the current study, DallAsta and Castellano studied a variant of the Naming Game with two pure opinions, where an individual switches to the intermediate state only when the number of times he has encountered the opposing opinion exceeds some pre-defined threshold [58]. The model presented here thus is a special case of Ref. [58] where the memory window is exactly equal to the threshold, and, critically, where no intermediate state is present. In contrast to the work done in Ref. [58], the focus here is to look at the fraction of initiators required to bring about a tipping point. The idea of opinion 'stickiness' has also been studied in the context of the Naming Game in various forms [40], [51]. In these studies, the stickiness parameter quantifies the probabilities with which a node in a mixed-opinion state rejects a pure state that it encounters in an interaction with its neighbors. The introduction of the stickiness parameter for nodes in the mixed-opinion state, gives rise to a phase transition between a regime where the consensus states are stable (when stickiness is low) to one where the consensus states are unstable and the system gravitates to a stable state with a non-zero density of mixed-opinion nodes.

3.2 Description of Model

This section defines in detail the microscopic rules of the *opinion inertia* model. First, each individual within the network initially adopts one of two opinions, designated A and B. The fundamental mechanism in the model for the change in individual states is the interaction of pairs of individuals, which represent speaker-listener pairs. In each such interaction, the speaker conveys their opinion to the listener, and in response to this conveyed opinion, the listener may or may not change their state depending on what other opinions they have heard in their prior interactions. Whether the listener's opinion changes depends on the

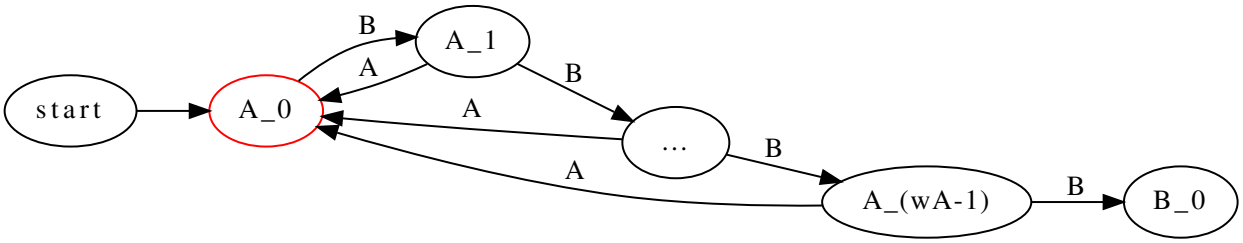


Fig. 3.1: Markov chain visualization of the *opinion inertia* model for a node currently holding opinion A. A_i represents a node holding opinion A with i counts towards switching to opinion B. After w_A consecutive interactions where the node hears opinion B, the switch will occur putting the node in state B with zero counts towards switching. The node will undergo the same process, requiring w_B opposing interactions consecutively before reverting back to state A with no counts towards switching. This model has no intermediate AB state.

inertia of their currently held opinion, a pre-defined value for each opinion and designated as w_A and w_B respectively. In terms of the model, the inertia w_A (w_B) of an individual in state A (B) is the number of consecutive times they must hear the opposing opinion B (A) before switching. For the special case $w_A = w_B = 1$, the model becomes the well-known voter model [19], [22]. To keep track of this, each individual keeps a counter dedicated to counting the number of times they encounter the alternative opinion, which resets to zero either when the required number of consecutive interactions of the opposing opinion are heard, or whenever the current opinion is heard. Note that in the former case, the counter also switches the opinion that it is keeping track of. A visualization of the Markov chain describing the state transitions in this model can be seen in Fig. 3.1. In this implementation, exposure to a different/same opinion only impacts the individuals counter when their role is the listener in a pairwise interaction. Naturally, one may consider the scenario where both the speakers and listeners counters are affected by the interactions (i.e., the speaker can also be reinforced in their view). Explorations on this generalization of the model have shown that the effects on the critical points in the system are negligible, as seen in Fig. 3.2.

Each simulation, the individuals (nodes) in the network are initially assigned opinions to fill the prescribed fractions p_A and $p_B = 1 - p_A$ of the total population in states A and B respectively. Then at each microscopic time step, a random node is chosen from the system and designated as the speaker. A random node is selected from among the speakers neighbors and designated as the listener. If the listeners opinion is the same as the speakers,

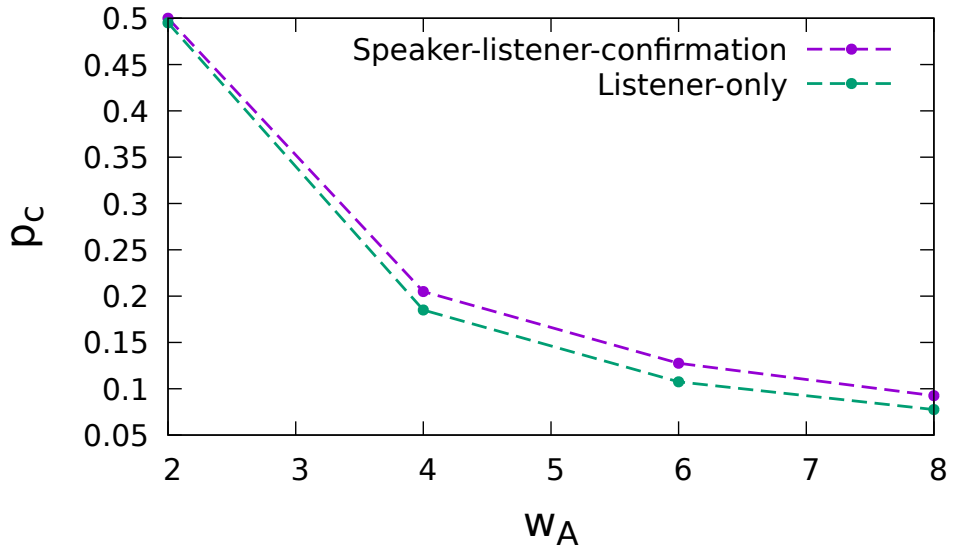


Fig. 3.2: Comparison of the critical populations for consensus on opinion *A* in the *listener-only* and *speaker-listener-confirmation* versions of the *opinion inertia* model. The alterations to the interaction rules have a negligible effect on the model performance.

its progress towards switching is reset to zero. If the listeners opinion is different from the speakers, the listeners count towards switching increases by one. If the listeners count becomes equal to its opinions inertia, it adopts the alternative opinion and begins a fresh count. Every N such microscopic time steps constitute one unit time step, where N is the network size. Thus, the event that a node is selected as a speaker is a Poisson process with rate one.

3.3 Results

3.3.1 Complete Graph

The first step to investigating the outcome of these rules is to study the system on a complete graph through Monte-Carlo simulations. Shown in Fig. 3.3, varying the fraction of nodes initially holding opinion A (p_A) and measuring the resulting fraction of simulation runs (over a total of 500) for which the system reaches consensus on opinion A reveals the same tipping point behavior expected from other committed agent style models. In these simulations, the inertia of opinion B is kept fixed at $w_B = 2$ while different values of w_A are tested to reveal the dependence of the critical populations on the minority opinion's inertia.

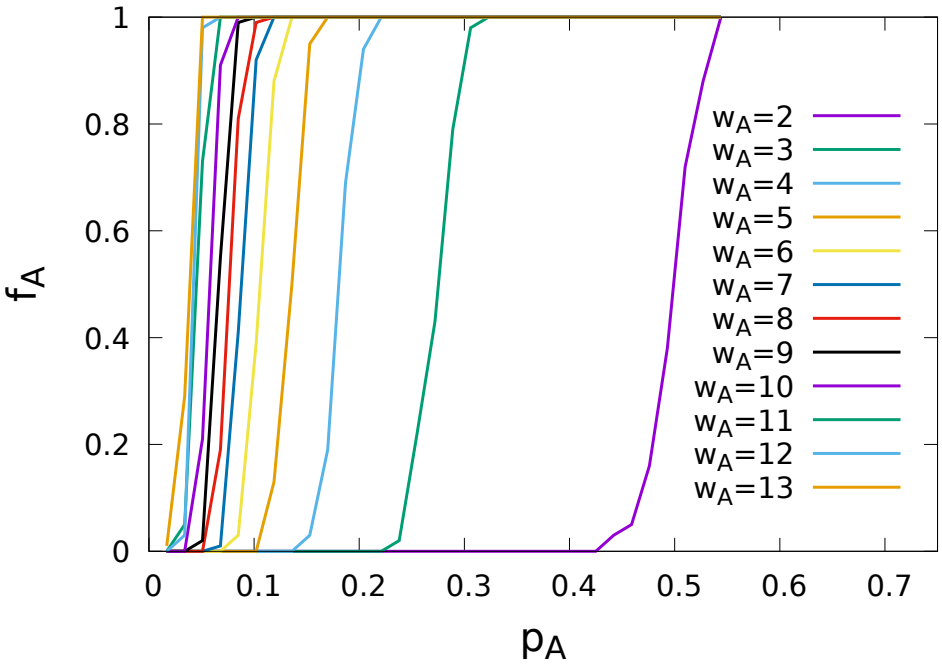
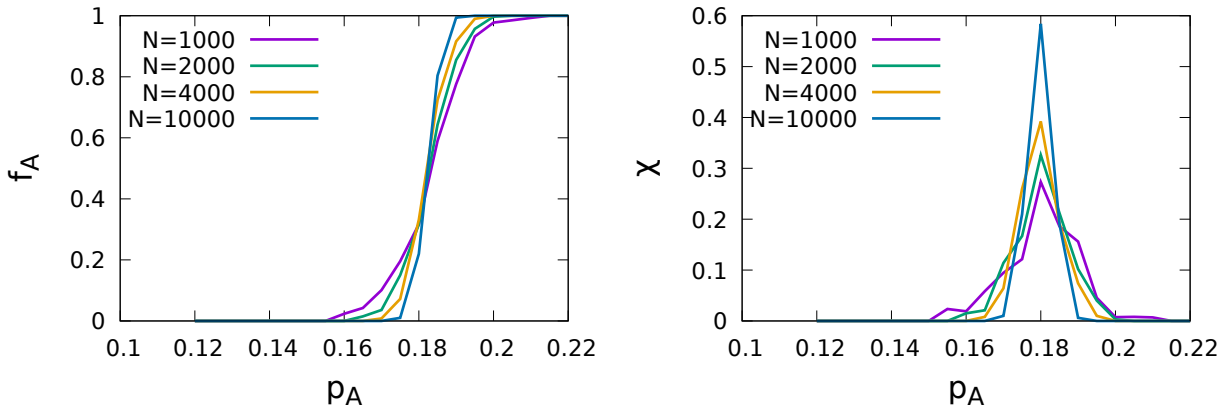


Fig. 3.3: The fraction of *opinion inertia* simulations that reach consensus on opinion A for varying initial populations of opinion A with multiple w_A inertia values. For these simulations $w_B = 2$. Simulations were run 500 times on a complete graph with $N = 1000$.

The tipping point behavior emerges as every value of w_A shows a typical S-shaped curve in the fraction of runs reaching consensus for opinion A (f_A). For increasing system sizes, these curves become progressively sharper (Fig. 3.4(a)), approaching a discontinuous transition in the infinite system-size limit and indicating the existence of a true and abrupt tipping point at a critical fraction p_c. For a finite system size N, the value of p_c is identified as the population where the (forward) derivative of the fraction of simulations reaching consensus on opinion A, $\chi \equiv df_A/dp_A$, is maximum (Fig. 3.4(b)). These results also indicate that the finite-size effects of the location of the critical point are negligible for this transition as it shows no drift through a full order of magnitude change in the system size.

As demonstrated by the results shown in Fig. 3.3, for values of $w_A \geq 3$, the critical population for opinion A constitutes the minority opinion. Thus, having a inertia even marginally greater than that of the majority opinion allows the minority opinion to tip over the entire population, as long as the minority fraction is greater than p_c. In the case of equal inertias where $w_A = w_B = 2$, the fraction of individuals holding opinion A initially must

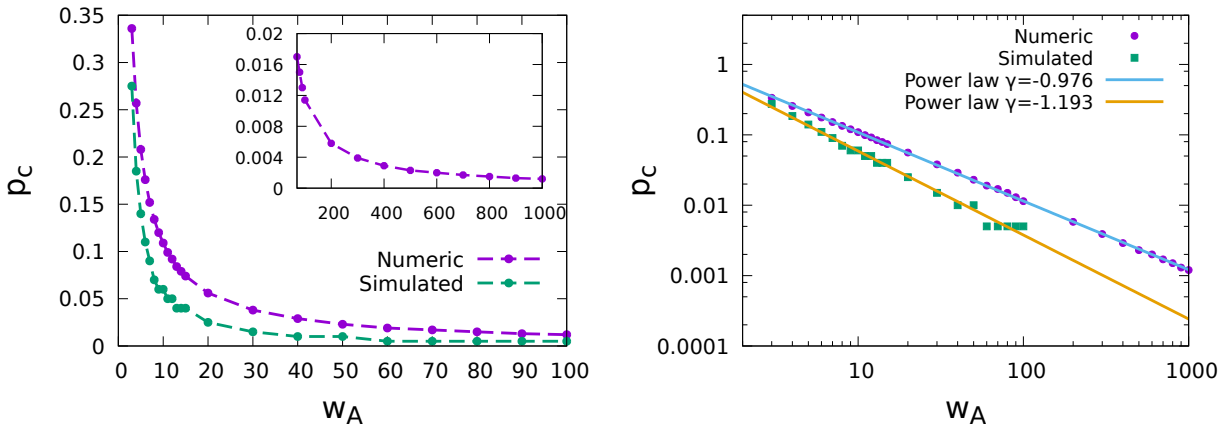


(a) (b)

Fig. 3.4: (a) The fraction of *opinion inertia* simulations that reach consensus on opinion A for varying initial populations of opinion A on multiple system sizes. These simulations are on a complete graph, have $w_A = 4$, $w_B = 2$, and are averaged over 1000 runs. (b) Forward derivative of the fraction of runs reaching consensus on opinion A from (a) vs. the initial population fraction of opinion A.

be in the majority in order to win over the population, as the dynamics become balanced similar to what would be seen in the standard naming game. It is important to note however, that this situation does differ slightly from standard naming game dynamics as there is no intermediate state, meaning nodes on the edge of switching still will *always* share their single opinion when chosen as a speaker.

As the inertia of the minority opinion is increased, the possibility of consensus over the entire network occurs at progressively smaller minority fractions. As shown in Fig. 3.5, p_c appears to converge to zero as $w_A \to \infty$. For the simulations shown here, $N = 1000$, and hence the smallest value that p_c can adopt is 0.001. However, we show using a semi-analytic approach (Sec. 3.3.2) that an upper bound to the critical value p_c converges to 0 as $w_A \to \infty$ (Fig. 3.5(a) inset), which confirms that the critical fraction vanishes for asymptotically large inertia values. Further, both the simulated and semi-analytic approaches can be described by power-law decays, shown in Fig. 3.5(b). The semi-analytic approach converges to 0 slightly more slowly, again proving to be an effective upper bound for the simulated system while accurately describing the behavior for large values of w_A. This behavior is a departure from the previously studied *waning commitment* model, where the critical population is shown to decay exponentially towards the steady value of the *infinite commitment* model. Here, the



(a) (b)

Fig. 3.5: (a) Comparison for the *opinion inertia* model between the values of the critical fraction p_c as a function of minority inertia, w_A, on a complete graph with $N = 1000$ obtained from simulations (averaged over 100 runs) and through the semi-analytic approach (Sec. 3.3.2). In both cases $w_B = 2$. The inset shows the extended numerical results (using Eqs. (6) and (8) to determine p_c for values up to $w_A = 1000$). (b) Log-log plot of the same data as in (a), with power law fits to emphasize the tail behavior.

critical populations following the power-law decay drop more quickly for low values of w_A, before slowing and exhibiting a fat tail as p_c approaches 0.

Note that the full dependence of the tipping point p_c on the interplay between the inertia parameters w_A and w_B is rather complex and non-linear. Our simulation results and scaling suggest that

$$p_c \simeq \frac{\text{const.}}{(w_A^{1/2}/w_B)} \propto w_B w_A^{-1/2} \tag{3.1}$$

in the $1 \ll w_B \ll w_A$ limit, as shown in the inset of Fig. 3.6, but the exact nature of this behavior is left to future work.

3.3.2 Mean-Field Approximation

For convenience, in this subsection the inertia of opinion A is denoted by w and the inertia of opinion B by v. Further, the fraction of nodes holding opinions A and B are denoted by n_A and n_B respectively. Similar to the waning commitment model, the state space of the model must be expanded to account for the fraction of nodes holding opinion A being comprised of distinct sub-populations that hold opinion A and have accrued a certain number

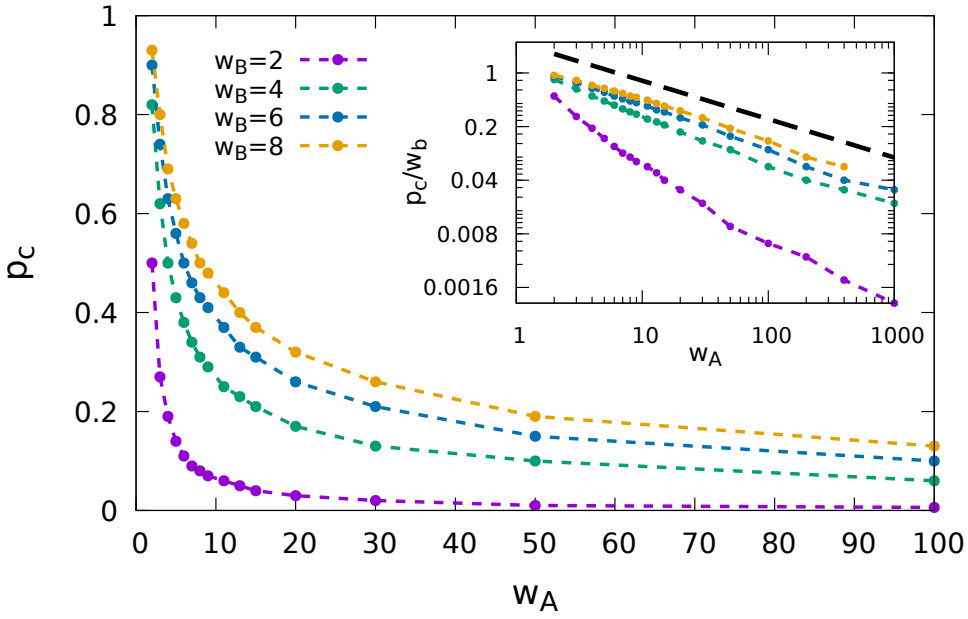


Fig. 3.6: Critical populations in the *opinion inertia model* obtained via simulation for different values of w_B with $N = 1000$ and averaged over 100 runs. The inset shows log-log plot of the scaled data. The dashed line, for reference, corresponds to a power law with exponent $\gamma = 1/2$.

of consecutive interactions with opinion B. The fractional sizes of these sub-populations are denoted by $s_{a,0}$, $s_{a,1}$, ..., $s_{a,w-1}$ for each number of counts towards switching. Analogously, the sub-populations for opinion B are denoted as $s_{b,0}$, $s_{b,1}$, ..., $s_{b,v-1}$ respectively. Thus

$$n_A = \sum_{x=0}^{w-1} s_{a,x} \tag{3.2}$$

$$n_B = \sum_{x=0}^{v-1} s_{b,x} \tag{3.3}$$

The evolution equations for the density of nodes in states A and B can then be written by noting that a change in opinion occurs only when a node whose counter for the alternate opinion is just below the inertia of its current opinion encounters the alternate opinion. Describing the likelihood of this occurring based on the population sizes, the number of nodes with opinion A evolves as

$$\frac{dn_A}{dt} = -n_B s_{a,w-1} + n_A s_{b,v-1} \tag{3.4}$$

where the first term captures the loss of nodes in state A, resulting from nodes represented by the fraction $s_{a,w-1}$ hearing opinion B. The second term analogously captures the gain resulting from nodes represented by the fraction $s_{b,v-1}$ hearing opinion A. Similarly,

$$\frac{dn_B}{dt} = -n_A s_{b,v-1} + n_B s_{a,w-1} \tag{3.5}$$

Due to the necessity of describing the individual rate equations for each possible sub-population, solving these equations can become extraordinarily difficult for arbitrarily large values of w and v. In order to make them more tractable, a quasi-steady state approximation for obtaining the sub-population fractions for each opinion can be employed. Specifically,

$$s_{a,x} = s_{a,0}(n_B)^x \tag{3.6}$$

$$s_{b,x} = s_{b,0}(n_A)^x \tag{3.7}$$

This assumption states that the fraction of nodes in state $\{a, x\}$ at a given time is approximately equal to the probability, given the systems current state, of a node in state $\{a, 0\}$ being chosen as listener for a node in state B on every one of x trials with replacement. This assumption, commonly used in the study of chemical reaction systems with intermediates, is known as the quasi-steady-state assumption, referring to the fact that the intermediate sub-populations arising in the transition from state $\{a, 0\}$ to state $\{b, 0\}$ and vice versa, are assumed to be in steady-state [59]. This can be seen from the evolution equation for a particular sub-population, say $\{a, x\}$:

$$\frac{ds_{a,x}}{dt} = -n_A s_{a,x} - n_B s_{a,x} + n_B s_{a,x-1} \tag{3.8}$$

Since $n_A + n_B = 1$, the steady-state expression (the case where $ds_{a,x}/dt = 0$) for the fraction of nodes in state $\{a, x\}$ is

$$s_{a,x} = n_B s_{a,x-1} \tag{3.9}$$

and thus

$$s_{a,x} = (n_B)^x s_{a,0} \tag{3.10}$$

Using this approximation, Eqs. (3.4) and (3.5) become

$$\frac{dn_A}{dt} = -(n_B)^w s_{a,0} + (n_A)^v s_{b,0} \tag{3.11}$$

$$\frac{dn_B}{dt} = -(n_A)^v s_{b,0} + (n_B)^w s_{a,0} \tag{3.12}$$

Additionally, describing the base case of each state in the same probabilistic manner gives

$$\frac{ds_{a,0}}{dt} = -n_B s_{a,0} + n_A \sum_{x=1}^{w-1} s_{a,x} + n_A s_{b,v-1} \tag{3.13}$$

$$\frac{ds_{b,0}}{dt} = -n_A s_{b,0} + n_B \sum_{x=1}^{v-1} s_{b,x} + n_B s_{a,w-1} \tag{3.14}$$

accounting for the loss from nodes in state $\{a, 0\}$ gaining a count towards switching as well as the gain from nodes with any count towards switching having an affirming interaction or a node switching from state B, and likewise for state $\{b, 0\}$. Then, using the steady state approximations in Eqs. (3.9) and (3.10) these equations can be rewritten as

$$\frac{ds_{a,0}}{dt} = -n_B s_{a,0} + n_A \sum_{x=1}^{w-1} s_{a,x} + s_{b,0}(n_A)^v \tag{3.15}$$

$$\frac{ds_{b,0}}{dt} = -n_A s_{b,0} + n_B \sum_{x=1}^{v-1} s_{b,x} + s_{a,0}(n_B)^w \tag{3.16}$$

and rearranging Eqs. (3.2) and (3.3) to be

$$\sum_{x=1}^{w-1} s_{a,x} = n_A - s_{a,0} \tag{3.17}$$

$$\sum_{x=1}^{v-1} s_{b,x} = n_B - s_{b,0} \tag{3.18}$$

the final equations for the rate of change of the individuals in the base states $s_{a,0}$ and $s_{b,0}$ can be written independent of any other sub-state populations, simplifying the equations

such that they can be solved using only the total populations and the base cases

$$\frac{ds_{a,0}}{dt} = -n_B s_{a,0} + n_A(n_A - s_{a,0}) + s_{b,0}(n_A)^v \tag{3.19}$$

$$\frac{ds_{b,0}}{dt} = -n_A s_{b,0} + n_B(n_B - s_{b,0}) + s_{a,0}(n_B)^w \tag{3.20}$$

Numerically solving the coupled rate of change equations (Eqs. (3.11), (3.12), (3.19), and (3.20)) for different initial populations of n_A^{init} and n_B^{init} (and with $n_A^{init} < n_B^{init}$) yields the steady state values of n_A and n_B respectively. From there, the smallest value of n_A at which the steady state value of n_A becomes greater than 0.99 determines the critical initial minority fraction p_c required to tip the system over. Fig. 3.5 shows a comparison for the tipping point p_c obtained through this semi-analytic approach and that obtained from simulation for different inertia values of opinion A (while the inertia of the other opinion is held fixed at $w_B = 2$). The cause of the higher p_c values yielded by the semi-analytic approach is the overestimation of subspecies densities (s_a, x and s_b, x for $x > 0$) in the initial phase of the dynamics; in reality the subspecies densities take some length of time to attain non-zero values. This overestimation favors the sustenance of nodes in state B, since they are initially in the majority. As a result, the fraction of nodes in state A required to tip the system over as estimated by the quasi steady state approximation is larger. Thus, the semi-analytical estimate of p_c consistently represents an upper-bound to the value observed in simulations. Furthermore, in the event that the inertia of opinion A diverges, Eqs. (3.11) and (3.12) show that for any non-zero initial density of A opinions, n_A grows monotonically while n_B decays monotonically, showing that the true critical fraction p c is bounded above by a value that vanishes in the asymptotic limit of inertia values.

3.3.3 Erdős-Rényi Random Graph

A similar asymptotic dependence of p_c on w_A with $w_B = 2$ is observed for Erdős-Rényi random graphs of size $N = 1000$, as shown in Fig. 3.7. Lowering the average degree $\langle k \rangle$ of the graph tends to lower the critical value slightly, and allows it to reach its lower steady state far quicker. For comparison, we also show the critical values obtained for the corresponding complete graph with 1000 nodes.

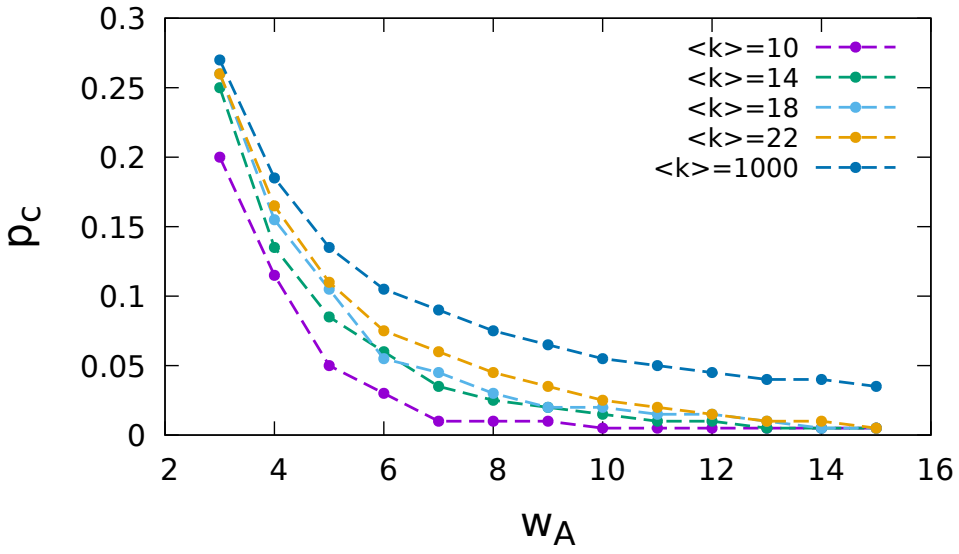


Fig. 3.7: Simulated results of the *opinion inertia* model to find the critical fraction p_c with relation to the inertia values of opinion A on Erdős-Rényi graphs with $N = 1000$, $w_B = 2$, and for various average degrees $\langle k \rangle$.

3.3.4 Curvature-Driven Coarsening with Opinion Inertia

Next, the investigation into the system on other network structures can be extended to include a lattices to determine how the introduction of stickiness into the rules of opinion change affects the coarsening behavior of the system. To facilitate comparison with previous studies, this evolution is implemented as a circular droplet of nodes in state A immersed in a sea of nodes holding opinion B in two dimensions. First, these results can be examined via visual inspection of the evolution for various combinations of values for w_A and w_B. For this system, the nodes are the sites of a square lattice (where each node connected to its 4 nearest neighbors) with sides $L = 250$ and without periodic boundary conditions. The droplet is initially given a radius of $R_0 = 35$. For the case of $w_A = w_B = 1$, there is no stickiness in either opinion and the dynamics reduce to exactly that of the voter model discussed in Sec. 2.2, where a single interaction with the alternative opinion is sufficient to cause a node to change its opinion. Fig. 3.8(a) shows the evolution of the droplet in this case. As has been demonstrated in prior work, the noise-driven roughening of the interface is clearly visible as the droplet evolves [32]. Next, the higher level inertia values are implemented, examining a system with $w_A = w_B = 2$. The initial conditions are identical to those for the case shown in Fig. 3.8(a). Fig. 3.8(b) shows a markedly different picture;

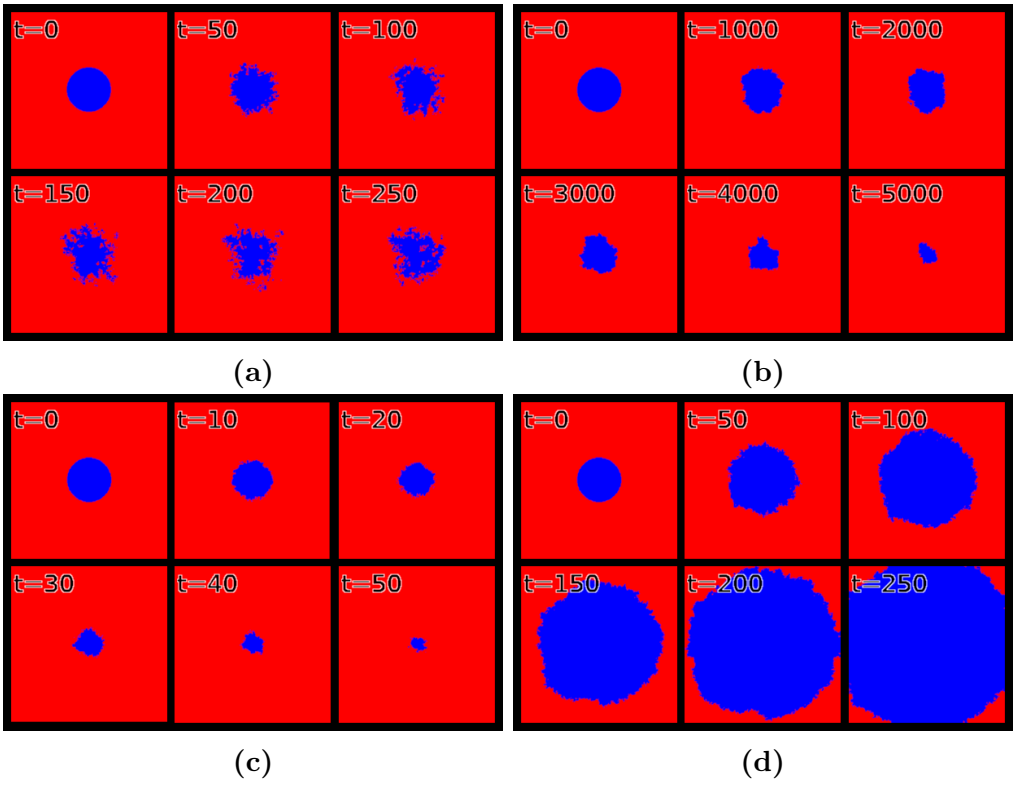
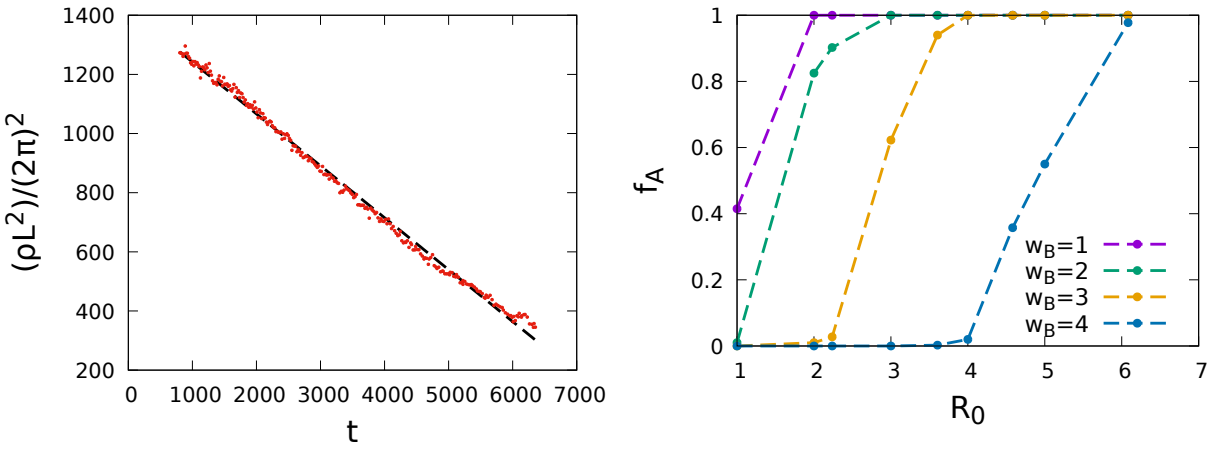


Fig. 3.8: Snapshots of the evolution of the *opinion inertia* model for a droplet of opinion A nodes in a sea of B nodes under different combinations of stickiness parameters. The nodes occupy the sites of a 250×250 two-dimensional square lattice without periodic boundaries. Opinion A is in the minority in every case and represented in blue. Nodes with opinion B are colored red. (a) Without stickiness i.e. $w_A = w_B = 1$, the model becomes identical to the voter model, and consistent with observations for the latter, the interface roughens diffusively, without any perceivable surface tension. With the introduction of stickiness in at least one of the two opinions, (b) $w_A = w_B = 2$, (c) $w_A = 1, w_B = 2$, (d) $w_A = 2, w_B = 1$, the interface evolution becomes curvature driven, and the droplet retains its roughly circular shape as it grows or decays.

the presence of an effective surface tension in the model is evident from the preservation of interface smoothness over time. This curvature-driven evolution is consistent with behavior observed in prior studies on voter-like models with intermediate states or memory, since the effect of inertia (or memory) is similar to that of intermediate states that intercede the transition between two opinions [58], [60], [34], [61]. Finally, it is clear that an inertia greater than one in only one of the two opinions is sufficient (see Figs. 3.8(c) and 3.8(d)) to keep curvature-driven behavior intact.

Next, this investigation can be extended to a quantitative consideration of the coarsening behavior. By track the evolution of the density of interfaces, $\rho(t)$, i.e. the fraction of nearest-neighbor pairs which differ in their opinion, a clear view of the boundary firmness is obtained. This is a commonly used order parameter that characterizes the coarsening behavior [33], [62]. For curvature-driven coarsening systems, the radius of the droplet changes linearly with time [63]. In two dimensions, it follows that the interface density also grows or decays linearly in time, i.e. $\rho(t) \sim c_1 \pm c_2 t$, where c_1 and c_2 are constants. Whether the droplet grows or decays depends on both the initial size of the droplet, as well as the values of stickiness for the two opinions. As shown in Fig. 3.9(a), the decay in interface density is indeed linear, as predicted by theory. Here, the initial radius of the droplet is $R_0 = 35$, the lattice size is $L = 250$, and the stickiness parameters are $w_A = w_B = 2$. Fig. 3.9(b) shows the fraction of simulations run (over a total of 400 runs) for which the droplet grows and spreads over the entire lattice (with $L = 35$, since that is the size of the initial droplet in the previous results where the favored opinion always dominates) as a function of the initial droplet radius for various combinations of stickiness. The results indicate the existence of a critical initial droplet radius for every combination similar to the critical populations seen on a complete graph, such that the probability of droplet growth sharply rises for initial radii above this critical value. Of course, these critical radii are extremely small for most values of opinion inertias, since in this model the system can be effectively reduced to only the nodes that are on the interface or neighbors to the interface (other nodes may speak to each other, but cannot effect change since they will always agree). As such, the system is inherently well balanced in effective populations and the side with the inertia advantage will quickly dominate (Fig. 3.8). However, this interface balance is not complete; the majority opinion still has a slight advantage by virtue of being the "outer ring" in the system. In other words, if the droplet is a perfect circle, the minority opinion would be expected to have $2\pi R$ nodes on the interface, while the minority opinion, by enveloping the minority, has $2\pi(R + 1)$ nodes. For most values of R the extra 2π nodes is not sufficient to tip the scales, but when R becomes sufficiently small this extra contribution is enough to quickly extinguish even higher inertia opinions before they begin to spread, creating the critical radii seen in Fig. 3.9(b).

In diffusive systems like the voter model, it has been theoretically demonstrated that in the asymptotic long-time limit, the interface density decays logarithmically in two dimensions



(a) (b)

Fig. 3.9: Quantitative description *opinion inertia* model droplet growth. (a) The radius of a circular droplet of opinion A nodes in a sea of B nodes as a function of time for $w_A = w_B = 2$. The radius is expressed in terms of the interface density ρ and the lattice size (linear dimension of the two-dimensional square lattice) $L = 250$, and shows a linear decrease with time. (b) The growth or decay of the circular droplet depends on its initial radius; for each combination a critical radii emerges defined by the fraction of simulation runs where the initial droplet grows and takes over all sites on the lattice. In these simulations, the inertia for opinion A is held fixed at $w_A = 5$ and the square lattice has dimensions 35×35.

under random initial conditions (see details below) [64]. Fig. 3.10 shows two snapshots of the coarsening process for the case $w_A = w_B = 1$, on a 100×100 two-dimensional square lattice at time $t = 0$ (with random initial conditions) and at $t = 25$, respectively. The diffusive nature of interface evolution, characteristic of the voter model, is clearly visible and is consistent with the behavior observed in the evolution of the circular droplet shown in Fig. 3.8(a). Fig. 3.11(a) shows the slow decay of the interface density as a function of time. One must be careful, however, as the exact asymptotic inverse logarithmic dependence of the interface density on time has long been known to be challenging to demonstrate numerically [64], [65]. Specifically, for the voter model, the leading-order asymptotic behavior for the interface density is $\rho \simeq \pi/[2\ln(t) + \ln(256)]$ [64]. As indicated by the results of the simulations in Fig. 3.11(b) the *opinion inertia* model with $w_A = w_B = 1$ approaches (albeit slowly) precisely this type of long-time asymptotic behavior, as expected.

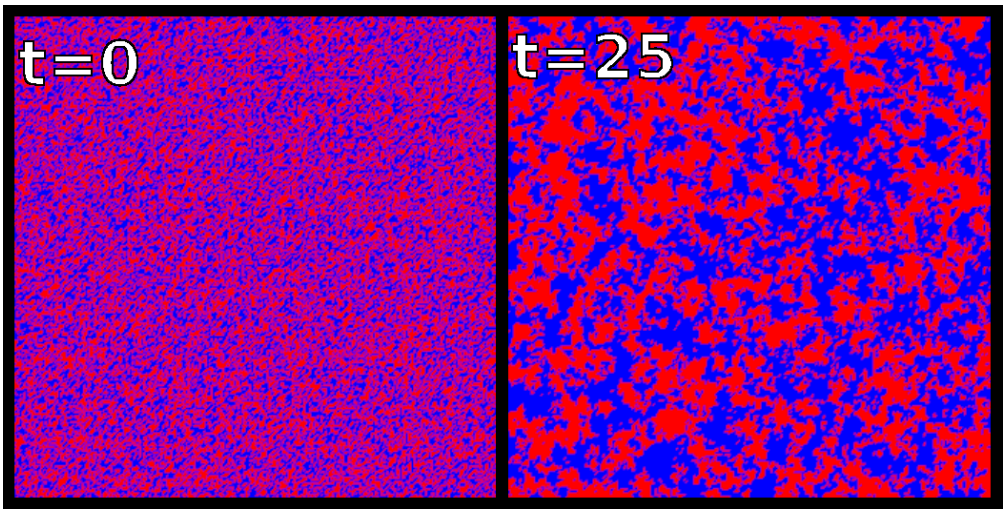


Fig. 3.10: Snapshot of the evolution of a system at times $t = 0$ (left) and $t = 25$ (right) under the *opinion inertia* model. The system uses random initial conditions and $w_A = w_B = 1$ (becoming equivalent to the voter model). The color code is the same as in Fig. 3.8. The lattice is a 100×100 two-dimensional square lattice with open boundary conditions.

3.4 Conclusions

This chapter has modeled a scenario where two competing opinions, ideas or behaviors vie for adoption in a social network where each opinion is endowed with an inherent inertia that impedes an individual holding that opinion from switching to the alternative opinion. From this model it has been demonstrated that the inertia of the dominant opinion on a social network determines how large the fraction of minority opinion holders needs to be in order to tip over the population to the initially minority opinion, and further that increasing the inertia of the minority opinion lowers the critical fraction required for its mass adoption dramatically as shown in Fig. 3.5. In practical contexts, the inertia of an opinion or behavior is related to the costs incurred, or incentives provided by its adoption, in comparison with the alternative. On two-dimensional lattices, it is also shown that the presence of stickiness in just one of the two opinions causes the systems behavior to belong to the universality class of models where coarsening is curvature driven. In contrast, in the absence of opinion inertia, the system belongs to the universality class of the voter model, where coarsening is noise driven.

Of course, as with all research, there are numerous avenues available for future work. For instance, it would be worthwhile to investigate more deeply the relationship between the

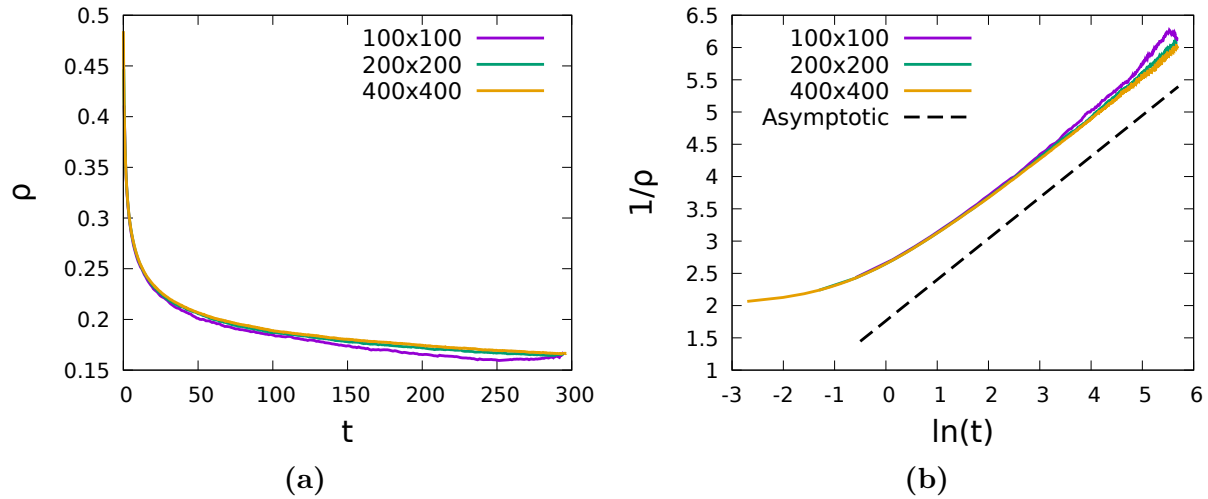


Fig. 3.11: **(a) The interface density ρ as a function of time t on a two-dimensional lattice with $w_A = w_B = 1$ for various system sizes. (b) The same simulation data as in (a) but plotted as the inverse interface density vs. logarithmic time in order to compare to the exact asymptotic limit of the voter model, $1/\rho \simeq (2/\pi)\ln(t) + \ln(256)/\pi$ (the dashed line) [64].**

ratio of the stickiness values and the critical value corresponding to the tipping point. Furthermore, empirical data from venues like massively multi-player online role playing games could be used as a test bed for validating the model and estimating the parameters which govern inertial opinion change [66]. Lastly, controlled experiments with incentives on online labor markets could further narrow down the conditions under which stickiness becomes a discernible feature of opinion dynamics [67].

CHAPTER 4
BURSTY SPEAKING PATTERNS

4.1 Motivation and Related Work

While changing interaction rules of pairwise models such as the naming game allows for a better capturing of the nuances of some social situations, it is also a rather narrow way to look at updating computational modeling. These sorts of updates are specific to the phenomenon and situations they attempt to model, and the broader applicability therefore suffers. For instance, committed agent models work extremely well at furthering the understanding of how activists can adjust their planning to affect change, but when considering opinion spread more generally or how activists may alter their *behavior* to boost their message the models fall somewhat short. In many ways, the nodes in these systems still behave like the autonomous robots they were described as in the original introduction of the model, ignoring some important and divergent aspects of real human behavior [23]. To alter this basic behavior, some of the more basic, deeper assumptions of the naming game must be altered. For instance, the random selection of speakers means that nodes are selected for events via a Poisson process, leading to an exponential distribution of the wait times they experience between speaking events [68]. This Poisson communication pattern, however, lacks the richness of realistic communication dynamics [69], [70], [71]. In fact, recent works show that human interaction occurs in a far more bursty manner [72], [73]; people tend to speak very frequently for short bursts then go silent for long periods of time, while the exponential distribution leads to fairly regular wait times (each node will speak on average once every N micro time steps).

Incorporating these patterns into computational models is an important step towards making them behave more realistically, and to this end a recent study considered the impact of bursty communications on the time to reach the absorbing state in the voter and in the SI models, where all agents exhibit the same non-Poisson communication characteristics [74]. Other past work on the effects of more bursty communication patterns have been shown to

Portions of this chapter previously appeared as: C. Doyle, B. K. Szymanski, and G. Korniss, "Effects of communication burstiness on consensus formation and tipping points in social dynamics," *Physical Review E*, vol. 95, no. 6, p. 062303, June 2017.

cause great changes in the properties of models meant to capture the spread of information and diseases, sometimes facilitating and sometimes slowing spreading depending on the base system and type of network used [74], [75], [76], [77], [78], [79], [80]. In addition, related studies have shown that increasing the propensity for committed nodes to speak lowers the number of them needed to achieve consensus [81]. In social dynamics, however, there is little understanding of how the specifics of the waiting-time distributions may affect the spreading process or what the critical features of the distributions are.

In contrast to these prior works [74], [75], [76], [77], [78], [79], [80], in this chapter bursty communication is investigated by focusing on studying the effects and impact of agents exhibiting bursty communication delivery *competing* with those with Poisson characteristics within the same network. In effect, the framework built here moves the altered behavioral patterns from the listening node (as seen in the prior *opinion inertia* model) to the speaking node.This model also attempts to retain some level of balance in the system, as while all individuals have identical speaker-event frequency in the long-time limit, the difference in their burstiness can have profound impact on the opinion competition and consensus formation.

After introducing the models and methods (Sec. 4.2), this idea is extended to focus on three scenarios: (*i*) Opinion dynamics with competing populations (one with Poisson, the other with bursty communication features) in the binary Naming Game (Sec. 4.3.1.1) [and for comparison, in the voter model Sec. 4.3.1.2]; (*ii*) The impact of committed individuals [31], [41], [47], [48], [49], [50], [51], [54], [81] [82], [83], [84], [85], [86], [87], [88], [89] with bursty communication features in the binary NG (Sec. 4.3.1.3); and (*iii*) an analytic approximation for the expected small-time activations of the waiting-time distributions used in conducted in (Sec. 4.3.2). Additionally, to broaden the applicability of the work, it is shown that that the main findings hold for other network structures by examining the opinion competition and the impact of committed agents in sparse random graphs (Sec. 4.3.3). And finally, for completeness, the impact of committed agents in the baseline scenario where all agents exhibit the same type of bursty communication features is examined in Sec. 4.3.4.

Using these methods to study direct competition between two different inter-event time distributions allows for more clarity in exactly what features of the distributions have the greatest impact on the outcome. By combining this competition with the committed agent variant of the naming game, the different activity patters also serve to further illuminate which conditions are most favorable to real world spreading phenomenon. For instance, these

results inform on what inter-activity time distribution a group of activists should choose to influence a campaign to the largest extent (assuming that the rest of the population uses the exponential distribution by default).

4.2 Description of Model

4.2.1 Model

In order to create competition between nodes following the standard Poisson selection process and those that do not, a set of non-exponential waiting-time distributions can be designed so that each has a mean of one (the same as the exponential distribution generated from the Poisson selection process). Doing so makes the groups identical in speaking frequency over long times, but different in when they speak. A mean wait time of one between speaking events also allows for the definition of a single system time step to be such that, on average, there will be N speaking events per unit time. Simulations are performed using standard naming game/voter model interaction rules with the initial condition that half of the nodes have one opinion (B) and follow the standard Poisson speaker selection, while the other half hold another opinion (A) and use one of the non-exponential waiting-time distributions. To simulate certain communication patterns as a property specific to individuals, the nodes keep their communication patterns as the system evolves, but their opinions still change in accordance with the binary NG or voter rules, respectively (i.e., in this model, speaker's inter-event time distribution is a characteristic of a the nodes, not that of the opinion).

4.2.2 Non-Exponential Speakers' Waiting-Time Distributions

The specific non-exponential distributions chosen for study here can be seen in Table 4.1. The distributions were chosen largely to reflect the power-law nature observed in human communication patterns [73], [71], with the Weibull [90], [91] and the uniform distributions used as a controls.

In Fig. 4.1, the different probability density functions (PDFs) for each of the distributions can be seen along with the PDF of the exponential distribution. These plots provide the basis for a qualitative understanding of why each distribution is used as well as providing an intuition for how each distribution behaves, as both are needed to explain results going forward. First, Fig. 4.1(a) shows the power law with a lower cutoff at $a = (\gamma - 1)/\gamma$. This

Table 4.1: Description of the probability density functions for the different non-exponential speakers' waiting-time distributions for modeling burstiness in communication. The parameters γ, α, b are used to control the burstiness of the distributions.

Name	PDF	Definitions	Restrictions
Lower cutoff power law	$p(x) = \gamma a^{\gamma} x^{-(\gamma+1)}$	$a = \frac{\gamma - 1}{\gamma}$	$\gamma > 1,\ x > a$
Shifted power law	$p(x) = \gamma a^{\gamma} (x + a)^{-(\gamma+1)}$	$a = \gamma - 1$	$\gamma > 1,\ x > 0$
Weibull	$p(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1} \exp(-(\frac{x}{\beta})^{\alpha})$	$\beta = \frac{1}{\Gamma(1+1/\alpha)}$	$x > 0$
Uniform	$p(x) = 1/b$		$1 - b/2 < x < 1 + b/2$ $b < 2$

distribution was chosen for its propensity towards burstiness, but also the regularity caused by the short time dead period. The cutoff means that there is a minimum time a each node must wait between speaking events, and as the system gets burstier (small values of γ) the cutoff grows. Second, the shifted power-law distribution represents an unrestricted bursty nature; Fig. 4.1(b) shows the behavior of a power law translated to the left by the value $a = \gamma - 1$. This system always maintains a higher head density and is thus always burstier than the exponential distribution (though it behaves similar to the system with exponentially distributed waiting-times for large values of γ). Third, Fig. 4.1(c) displays the Weibull function. This function has some behavior derived from both the power-law and exponential distributions, and in the special case of $\alpha = 1$, it is exactly the exponential function [80]. This distribution is a perfect control in this system since for $\alpha < 1$ it is always more bursty and for $\alpha > 1$ it is always less bursty than the exponential distribution. Lastly, Fig. 4.1(d) is a uniform distribution centered around $x = 1$ with a range of b, a function that is always clearly less bursty than the exponential one.

4.3 Results

4.3.1 Complete Graph

4.3.1.1 Opinion Competition in the Binary Naming Game

This section studies simulations of the competition outlined above by running the system to consensus and comparing the fraction of wins for the non-Poisson nodes with

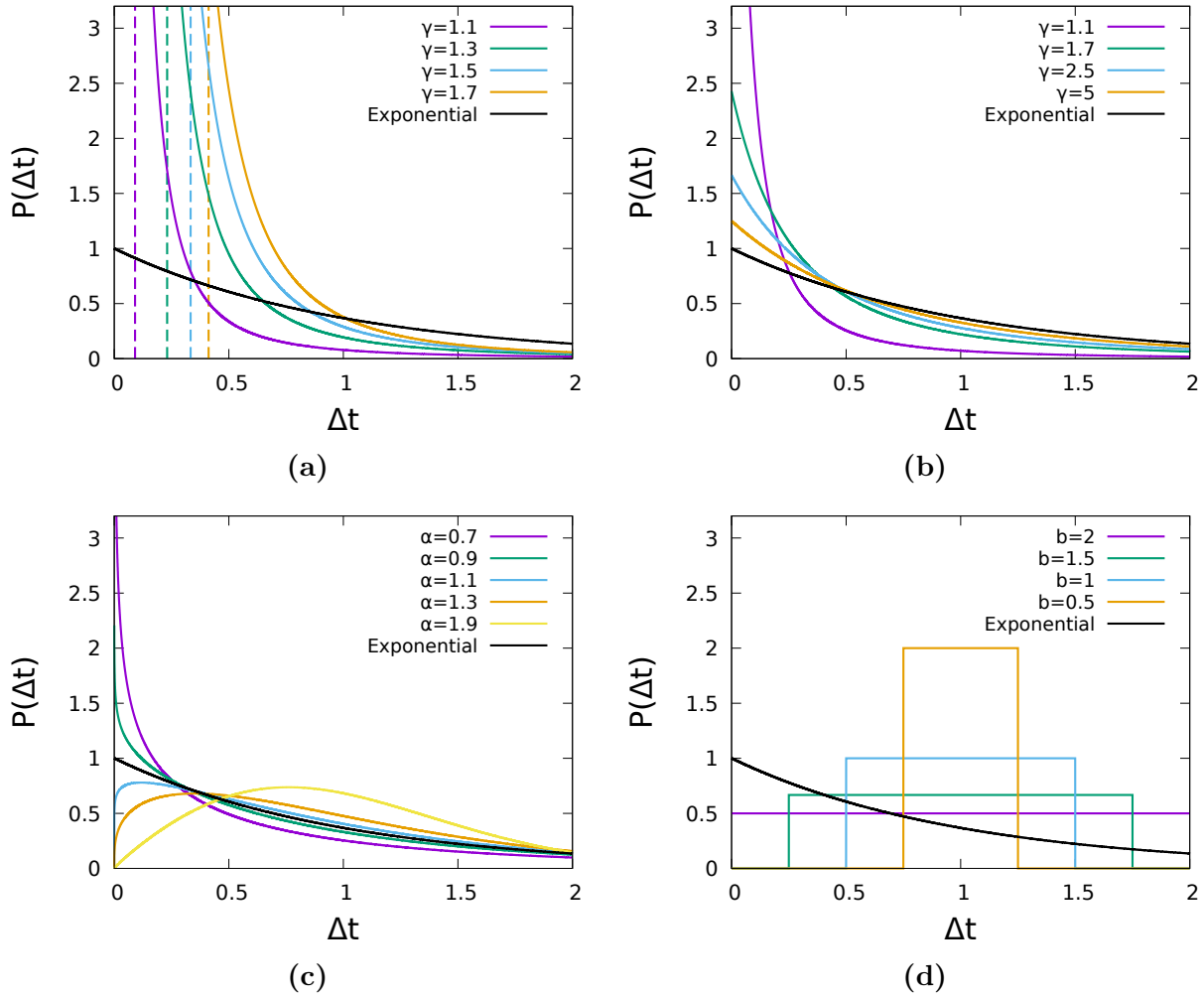


Fig. 4.1: PDFs used to vary burstiness for the non-exponential speakers' waiting-time distributions with various chosen parameters compared to the exponential one. (a) the power law with lower cutoff, (b) the shifted power law, (c) the Weibull distribution, and (d) the uniform distribution.

initial opinion (A) for a given system size and set of control parameters. This analysis is limited to simulations on a complete graph, Sec. 4.3.3 extends the scope and shows that the results found do not change when the system is run on a sparse random network instead. These simulations show that, as seen in Fig. 4.2, opinions corresponding to the burstier waiting-time distribution are favored to create consensus for their opinion, an effect that becomes more pronounced with increasing system size. In fact, in the case of the power law with lower cutoff and the Weibull distributions, there is a clear transition at large system sizes, where parameter values of $\gamma \approx 1.7$ and $\alpha = 1$ mark critical points that determine eventual system consensus. These parameters allow their corresponding distributions to

have either a higher or lower head density than the exponential distribution, and thus mark a transition of which opinion is initially propagated by the burstier nodes. This transition is particularly interesting for the power law with a lower cutoff, since the burstiness transition of the Weibull distribution is well understood (the critical value $\alpha = 1$ is simply the point at which the Weibull distribution is exactly the exponential, while any value $\alpha > 1$ simply shifts the distribution's peak to greater values of Δt). For the power law with lower cutoff, however, the transition is the result of the interplay between the regularity of the forced waiting period between speaking events and the inherent burstiness of the power-law head balancing out around $\gamma = 1.7$.

In both cases, the dominance of the burstier distribution becomes more pronounced at large system sizes, causing the side with the higher head density to win with near certainty in large systems. This is further supported by the results of the simulations with the shifted power-law and uniform distributions. Since there is no transition of head density in these cases (the shifted power law is always burstier than exponential while the uniform is always less bursty), they are always more and less likely to win, respectively. These results imply that despite the efforts to preserve the symmetry of the system by keeping the mean wait times the same across all distributions, simply changing the way these wait times are distributed carries sufficient impact to entirely break the symmetry (in the infinite system size limit, see Fig. 4.2). This is in contrast to the voter model, studied in Sec. 4.3.1.2, where the bias towards the burstier opinion remains constant with increasing system size. In that case, the randomness inherent in the voter model works to mitigate the effect of the early-time dominance of the burstier opinion and allows the system to revert to an even competition more easily.

The question of why these distributions behave this way (and why the head of the distributions matters more in this context than the tail) can be answered by studying the different time regimes of the system. By looking at the average time to consensus for the systems conditioned on which opinion eventually won (as seen in Figs. 4.3 and 4.4), the time scales on which the distributions operate can be seen more clearly. Specifically, the system takes a much longer time to reach consensus for the less bursty opinion than for the more bursty one. This can be explained by dividing the simulations into two time regimes: early-time and late-time. In the early-time regime, the burstier nodes dominate since they are likely to activate (often multiple times) before the less bursty nodes activate at all. They

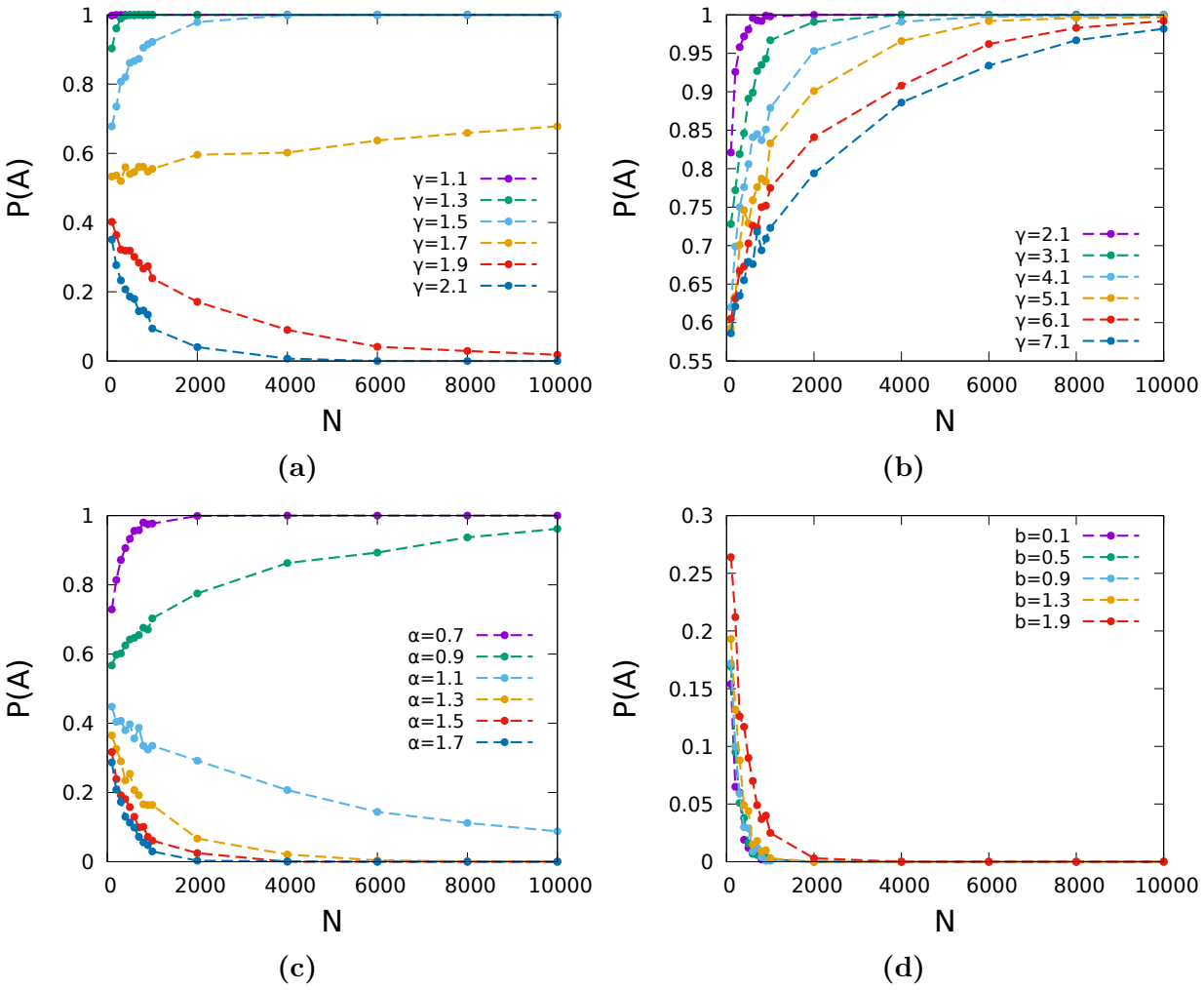


Fig. 4.2: The fraction of runs (out of 1000 trials) vs system size that the non-Poisson (*A*-opinion) nodes won the opinion competition against the Poisson (*B*-opinion) nodes in the binary NG on a complete graph. As before, the speakers' waiting-time distribution for the non-Poisson nodes is (a) power law with lower cutoff, (b) shifted power law, (c) Weibull, and (d) uniform distribution.

are then likely to go dormant for some extended amount of time, beginning the later time regime where the less bursty nodes become far more active. In most cases, however, the early-time dominance of the burstier side switches the opinion for a sufficient number of the less bursty nodes to create a heavy majority for the burstier side before the later time regime is entered. When this happens, the system quickly reaches consensus before many of the nodes even have their opportunity to speak, leading to the heavily unbalanced average activations per time step seen in Fig. 4.5. This result indicates that a high head density

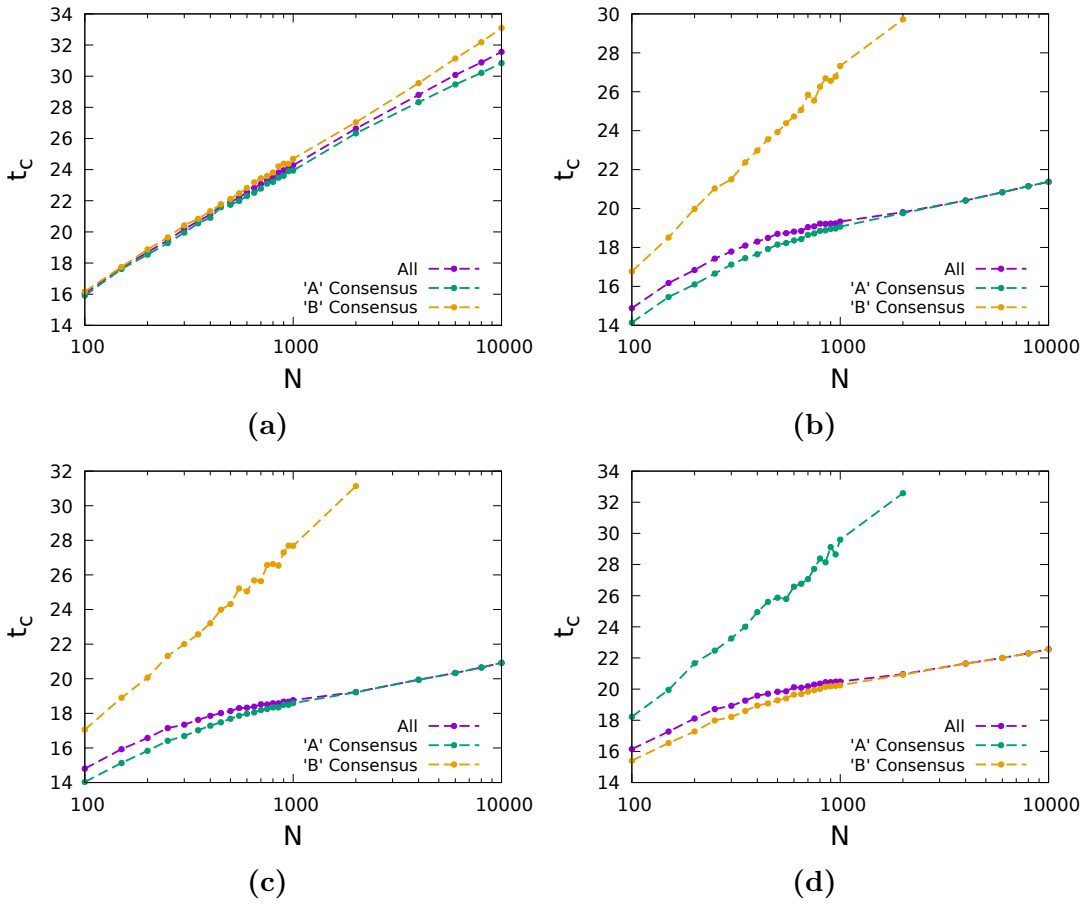


Fig. 4.3: The time to consensus conditioned on each side wining in the binary NG on a complete graph. Initially, half the nodes have non-exponential speakers' waiting-time distribution and hold opinion A, while another half follows an exponential distribution and hold opinion B. Part (a) displays the fairly even results for the power law with a lower cutoff and $\gamma = 1.7$. Parts (b) and (c) shows the case of a more bursty non-exponential distribution (the shifted power law with $\gamma = 2.9$ and the Weibull distribution with $\alpha = 0.7$ respectively) while part (d) shows the less bursty case (uniform distribution with $b = 1.9$). All simulations were run 10000 times.

(correlating to a strong initial push of opinions) is critical to achieving consensus, even if the nodes that initially caused the push go silent for long periods afterwards. Even so, however, occasionally the less bursty opinion still has enough of a presence to push back during the later time regime to allow for the long time victories of the less bursty opinion. A more in depth look at the activation rates of nodes in extreme bursty cases and how they react to different time scales is provided in Sec. 4.3.4.

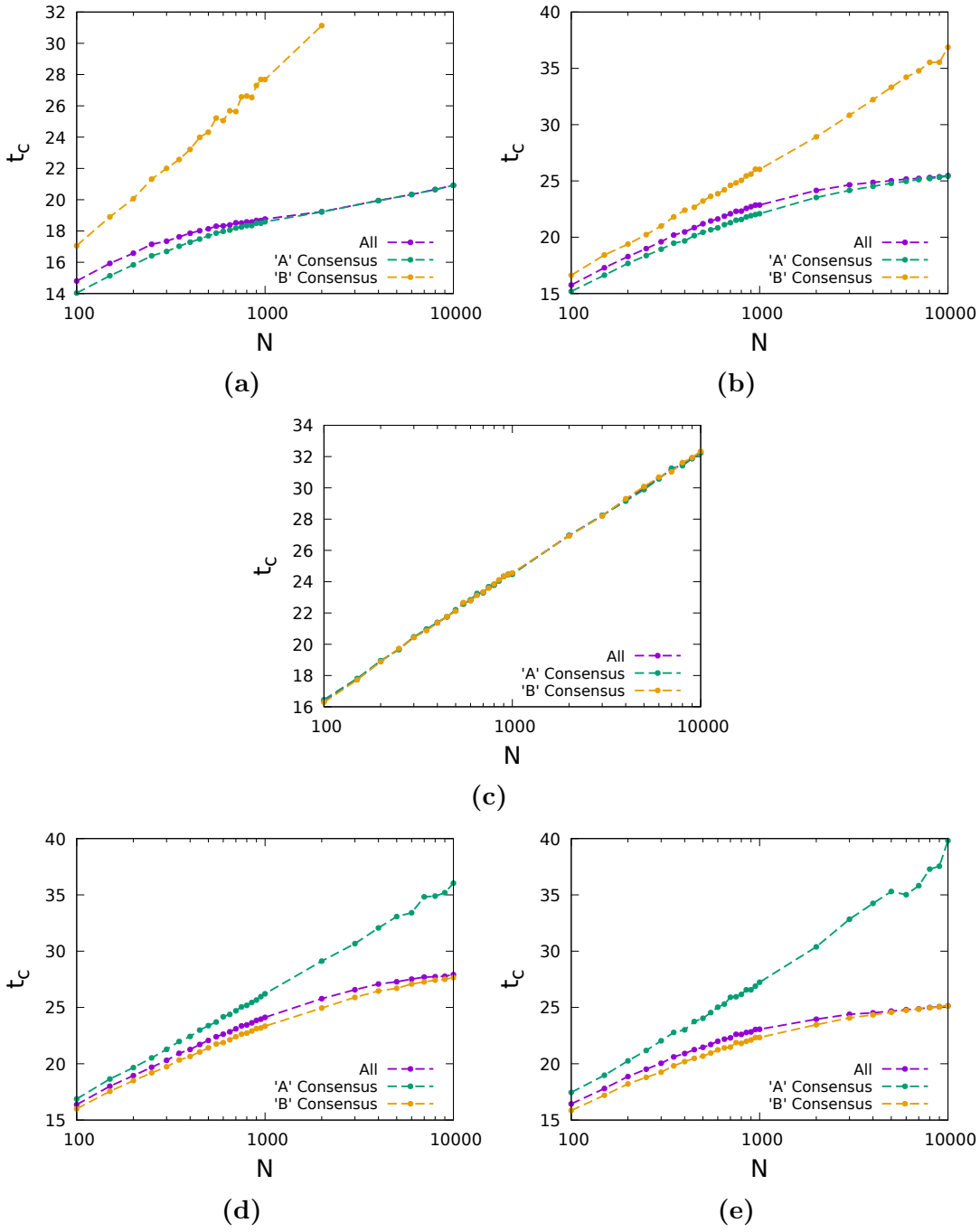


Fig. 4.4: The time to consensus conditioned on each side winning in the binary NG on a complete graph, where half of the nodes are in opinion A initially and follow a Weibull waiting-time distribution and the other half of the nodes initially are in opinion B following an exponential waiting-time distribution. Part (a) shows the results of the conditioned simulations with $\alpha = 0.7$ for the Weibull distribution, (b) with $\alpha = 0.85$, (c) with $\alpha = 1$, (d) with $\alpha = 1.15$, and (e) with $\alpha = 1.3$.

Additionally it is shown that the consensus time increases *logarithmically* with the system size, $t_c \sim \ln(N)$, in the asymptotic large system-size limit (Figs. 4.3 and 4.4). This logarithmic scaling holds for all cases of competing non-Poisson communication dynamics, and regardless of the outcome of the competition. The rate of the logarithmic increase (i.e., the slope of the lines in the log-normal plots in Figs. 4.3 and 4.4), however, is sensitive to the details of the non-exponential waiting-time distributions and to the condition of the outcome of the competition. One should also note that the standard binary NG with only Poisson communication dynamics also exhibits logarithmic scaling of the consensus times with the system size [31], [38], [92].

4.3.1.2 Opinion Competition in the Voter Model

To broaden the scope of this investigation, analogous simulations were run employing the voter model [22], [93] on complete graphs. As mentioned in Sec. 2, at the microscopic level the voter model is very similar to the binary NG, the only difference being that it has *no intermediate opinion state* [94], [95], [61], [82]. Instead, at each time step the listener automatically accepts the speaker's opinion as its own. As before, for this study the system is set up so that half of the nodes follow a non-Poisson update pattern and are initialized to state A, while the other half follow the standard Poisson pattern and are initialized to state B. As shown in Fig. 4.6, the results are quite similar to the standard naming game model. The system remains biased towards the burstier waiting-time distribution, and the critical values of the parameters that switch the bias from the non-exponential distribution to the exponential distribution remain approximately the same. In this case, however, unlike in the binary NG, there is *no* symmetry breaking in the infinite system-size limit (Fig. 4.6). Instead, there is a flat (system-size independent) bias towards the burstier distribution that remains the same as the system size approaches infinity. This behavior is likely the result of the lack of history-sensitivity (i.e., the lack of bi-stability and hysteresis) in the voter model. In other words, in the binary NG simulations it is much harder for nodes to switch opinions, allowing for opinion shifts within the system to gain a sort of momentum as the system moves towards a consensus. In the voter model, however, the ease with which nodes change opinions means that random fluctuations are much more likely to counter all progress towards a single consensus. Thus, there is enough random noise inherent in the system that the symmetry breaking effect of heterogeneous waiting-time distributions is not as strong as

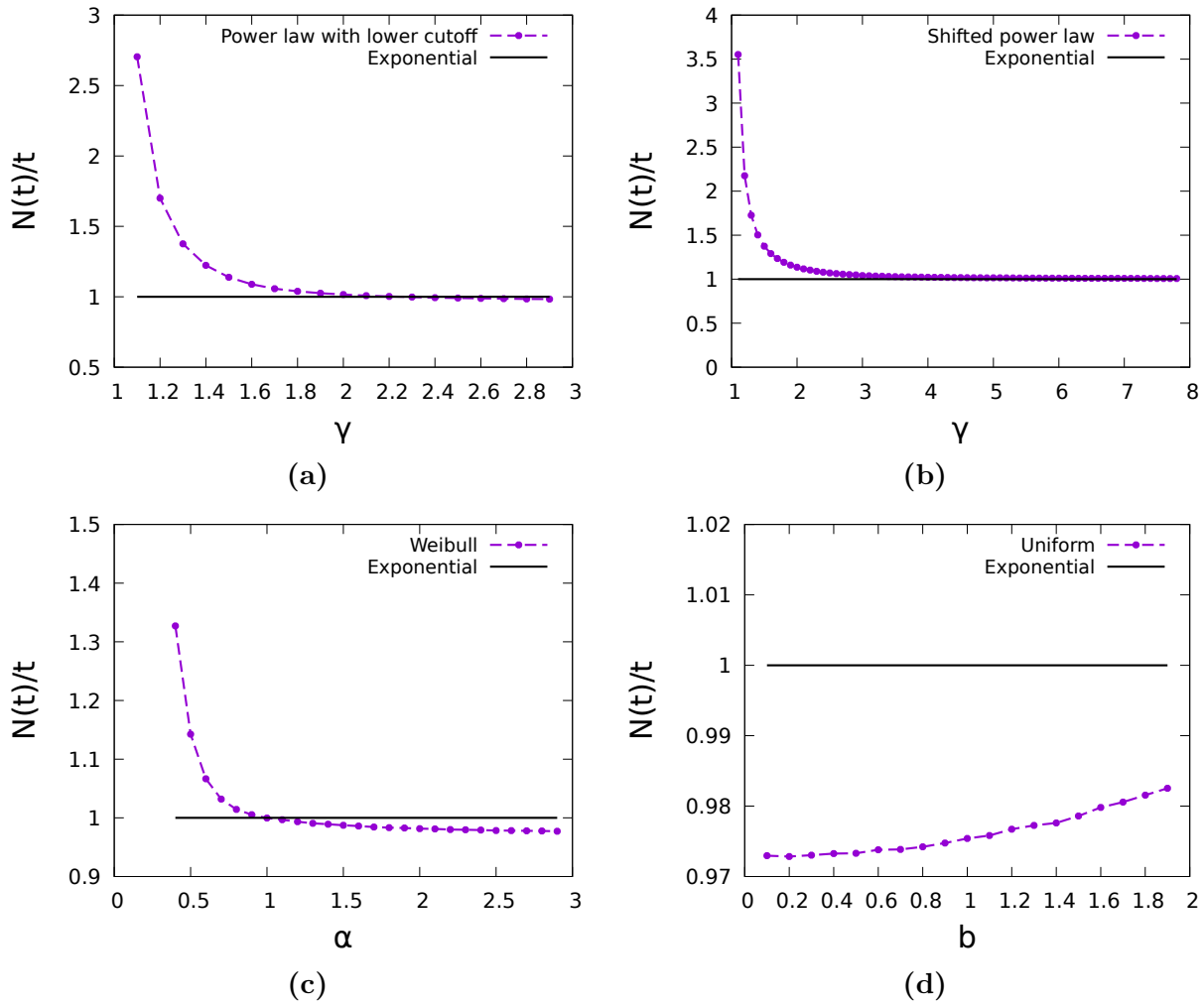


Fig. 4.5: The average number of speaking events in the binary NG for each type of nodes per one system time-step until consensus is reached. Each simulation is averaged over 1000 runs with $N = 1000$ on a complete graph. As before, (a) is the power law with lower cutoff, (b) is the shifted power law, (c) is the Weibull, and (d) is the uniform distribution.

in the binary naming game.

The random progress towards consensus can be partially seen in the consensus times, shown in Fig. 4.7, where while the opinion propagated by the burstier nodes still tends to reach consensus faster, the difference in consensus times is far less drastic. Additionally, the system operates on a much larger time scale in general, meaning that it will nearly always reach the long time limit and the overall activation rates of the two groups will be fairly balanced by the end. Thus, the only advantage given to the burstier nodes is a basic early-time advantage, rather than the total speaking dominance seen in the NG simulations.

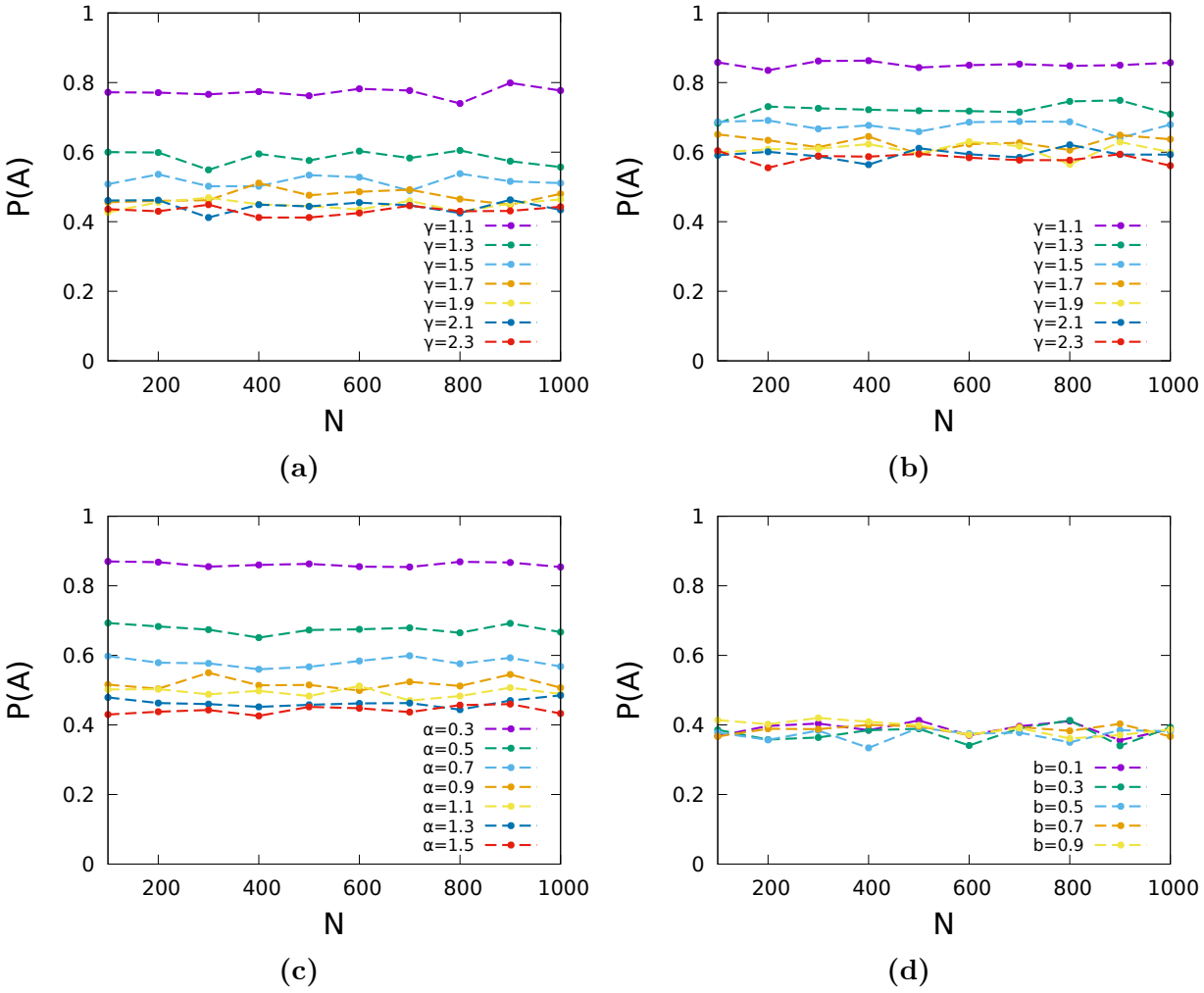


Fig. 4.6: The fraction of runs (out of 1000 trials) vs network size that the non-Poisson (*A*-opinion) nodes won the opinion competition against the Poisson (*B*-opinion) nodes in the voter model on a complete graph. As before, the speakers' waiting-time distribution for the non-Poisson nodes is (a) power law with lower cutoff, (b) shifted power law, (c) Weibull, and (d) uniform distribution.

Made even weaker by the lack of 'momentum' in the voter model in general, this advantage is not strong enough to entirely dominate the system in the infinite system size limit.

4.3.1.3 Consensus Formation and Tipping Points with Committed Agents in the Binary NG

As discussed in depth in the preceding chapters, models with committed agents (or zealots) have often been employed to simulate opinion spread driven by individuals who

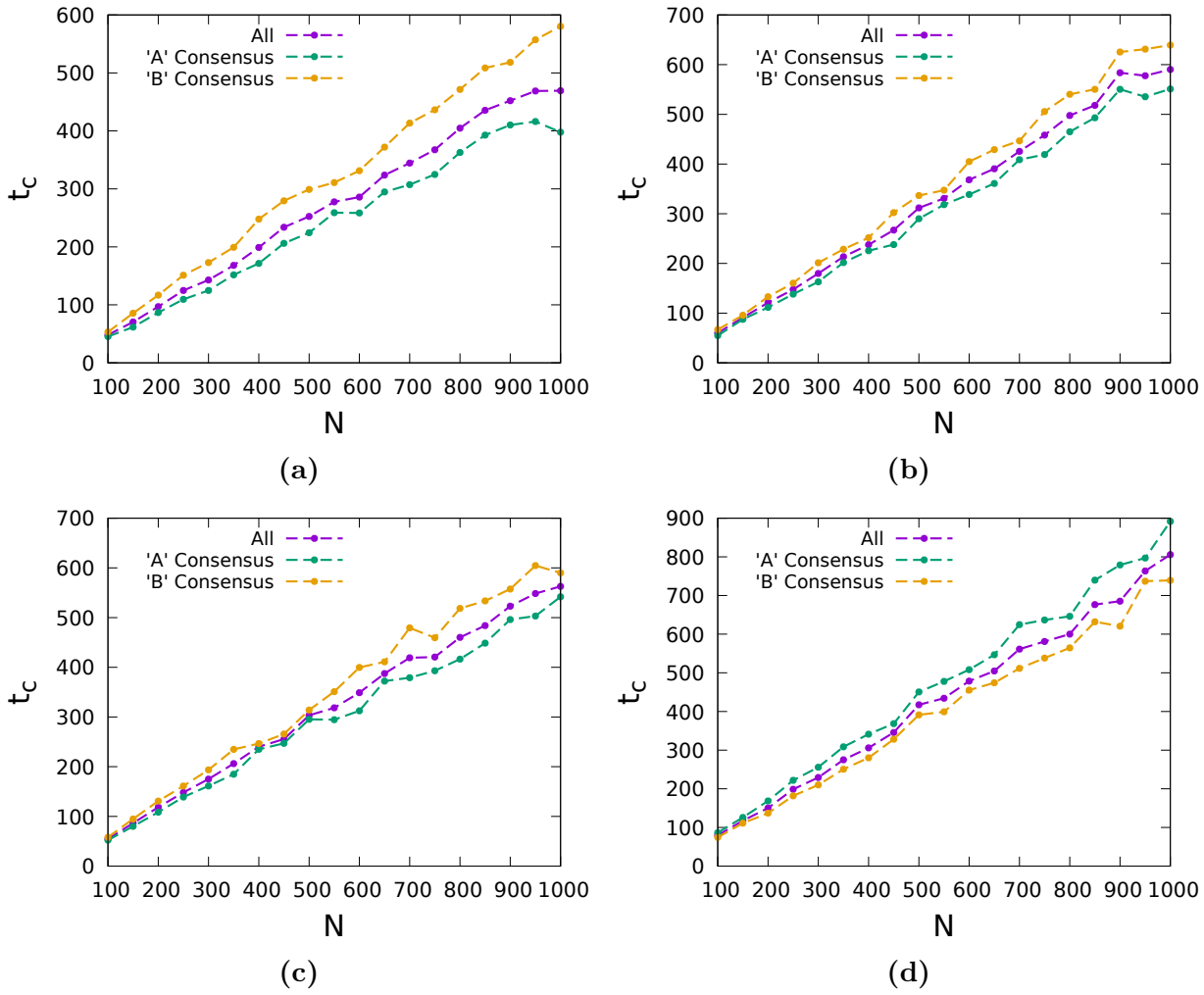


Fig. 4.7: Consensus times for the voter model with the presence of non-exponential speakers, separated by which opinion eventually attained consensus. Opinion A was initially propagated by the non-exponential speakers, with waiting-time distributions characterized by (a) power law with lower cutoff ($\gamma = 1.7$), (b) shifted power law ($\gamma = 2.9$), (c) Weibull ($\alpha = 0.7$), and (d) uniform ($b = 1.9$) distributions.

never change their opinion [31], [41], [47], [48], [49], [50], [51], [54], [82], [83], [84], [85], [86], [87], [88], [89]. In most of these models, simulations with committed agents are set up so that a small population of nodes (p) within the system are designated committed agents and given a single opinion (A) while all other nodes in the system follow the rules of the binary NG as usual and are initialized with the other opinion (B). The the effects of different p values are then investigated until the critical population (p_c) that causes a sharp phase change in the system is revealed. In this section, we repeat these experiments in the presence of

non-exponential committed individuals to model activists that spread opinions via methods other than normal interaction between individuals (such as a political campaign setting up a call bank).

In general, the value of p_c is somewhat sensitive to small alterations within the system rules or the average node degree, and heterogeneous waiting-time distributions no different, presenting the ability to lower p_c considerably as seen in Fig. 4.8(a) [84]. Here, p_c is defined as the committed population at which half of 1000 simulations reaches consensus before $t = 150$, shown in Fig. 4.8(b). The system size in Fig. 4.8(a) is 1000, but as Fig. 4.8(c) shows, there is no shift in p_c at higher values of N. In these simulations, only the committed agents are designated as being non-Poisson nodes, while all other nodes follow the standard Poisson selection process. The results clearly show that when the committed agents are burstier than the surrounding population, they are able to work far more efficiently and lower the critical fraction of the population considerably. Interestingly, the opposite is not true. When the non-committed nodes are burstier, the critical fraction remains steady at $p_c \approx 0.098$ (the general value shown for the standard naming game with committed nodes) [49]. This is due to the same time regime dynamics discussed earlier; the burstier nodes speak frequently in the early-time regime and give a heavy advantage to their side. If those burstier nodes are the committed agents, they establish a strong minority presence in the simulation and gain an advantage. If they are not the committed nodes, however, no advantage is gained because the committed nodes cannot change their opinion and thus with these initial conditions the majority nodes have no ability to affect change on the system until the committed nodes activate. Instead, the committed agents simply ignore the repeated interactions from the surrounding nodes until the system enters the long time regime where the identical mean wait times take over. Once this occurs, the system reverts to the value of p_c that occurs in a simulation with all speakers being Poisson selected, because in this time regime the systems behave very similarly.

To gain further insight in the impact of bursty communication patterns on the tipping point p_c, the analysis extends to the base-line scenario where all individuals in the system exhibit the same type of non-exponential waiting-time distribution. These results are shown in 4.3.4.

Note that in contrast to the naming game, the voter model with committed agents (on a fully-connected network) does *not* exhibit a tipping point. Instead, any non-zero fraction

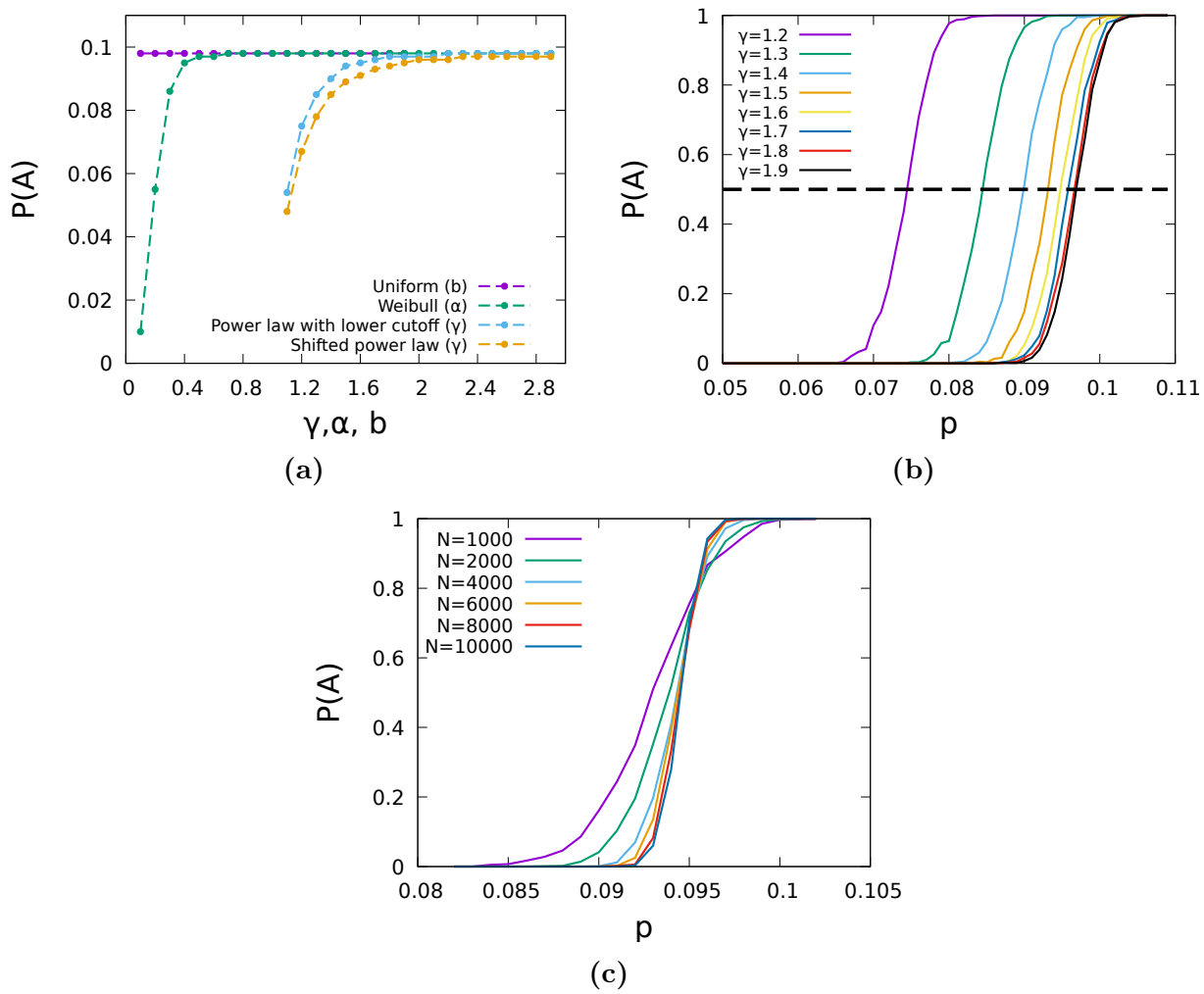


Fig. 4.8: The effects committed agents with non-Poisson speakers' communication patterns in the binary NG on a complete graph. (a) The critical fraction of committed agents (tipping point) necessary to create consensus for the minority opinion with respect to the various parameters that control their burstiness. Averaged over 1000 runs on systems with $N = 1000$. Note that the parameters γ, α, and b are specific to the distribution in which they are used and their impact on the burstiness varies from one distribution to another; they should not be compared directly. (b) Fraction of runs reaching consensus for the committed minority by time $t = 150$. Committed agents follow power law with lower cutoff waiting-time distribution. The critical fractions p_c [shown in (a)] were defined as the population at which the system reaches the minority consensus in over half of the runs. (c) Finite-size effects of the tipping point for nodes following the power-law with lower cutoff distribution and $\gamma = 1.5$, indicating no significant shift in the value of p_c as $N \to \infty$.

of zealots leads to fast (exponential) relaxation to consensus [86], [87]. Therefore, this case is not studied here.

4.3.2 Approximation of the Expected Small-Time Activations

Throughout the preceding section it is demonstrated via direct simulation that competition between two opinions spread by groups with different levels of burstiness will favor the opinion of the group with the higher burstiness. Further, it is shown that the relevant quantity of the waiting-time distribution is the head density rather than the tail due to the importance of dominating the initial stages of the simulation. An analytic description of this phenomenon proves difficult, however, because direct comparisons of the head density via the CDF fail to accurately describe the dynamics of this system. These simple comparisons do not sufficiently account for the probability that a bursty node can activate multiple times before a less bursty node activates once, and thus greatly underestimate the effect that burstiness can have on a system.

To remedy this, this section uses the *expected small-time activations*, D, to characterize the burstiness of a given node. The expected small-time activations is an approximation of how many times a node following a given waiting-time distribution is expected to activate before the mean activation time is reached. This value allows for a direct comparison of the influence that different distributions have over the early-time period of a simulation by sampling the head of the distribution multiple times within different ranges to account for a node's repetitive activity. Using the notation from Table 4.1 where $p(x)$ is the PDF of the waiting-time distribution, and $P(t) = \int_0^t p(x)dx$ is the CDF of the same function, the expected small-time activations, D, can be calculated. First, it is given that the probability that a node will activate exactly m times before t is

$$P_m = \int_0^t p(x)P_{m-1}(t-x)dx \tag{4.1}$$

with the special case of no activations before t being $P_0(t) = 1 - P(t)$. From there D is approximated by summing the probabilities that a node will speak m times before t multiplied by m. This is continued for all values of m up to a maximum value considered (n) after which it is assumed that if a node has not activated n or less times then it must activate exactly $n + 1$ times. Thus, we say that the order of the approximation is n and

an approximation of order n will consider a maximum number of activations $n+1$. The definition of D of order n is given by

$$D_n(t) = (n+1)\Big(P(t) - \sum_{m=1}^{n} P_m(t)\Big) + \sum_{m=1}^{n} mP_m(t) \tag{4.2}$$

Of course, as n goes to infinity this approximation becomes an exact description of the expected number of activations before t. The probability of having $n > 3$ activations, however, vanishes rapidly with increasing n, so the expected value of D for a distribution comparable to the exponential function (where $D = 1$) can be reasonably well approximated by just D_2. Hence, for simplicity this work considers only up to this case, and by following the procedure outlined above the approximate values can be calculated via Eq. (4.3).

$$D_2(t) = 3P(t) - 2P_1(t) - P_2(t) \tag{4.3}$$

For the exponential and uniform distributions, D can be solved exactly up to higher orders. In fact, the specific values of P_n are well known for the exponential distribution as

$$P_n^{\exp}(t) = \frac{t^n}{n!}e^{-t} \tag{4.4}$$

Similarly, for the uniform distribution the values of P_1 and P_2 (and beyond) can be obtained analytically in a closed form,

$$P_1^{\text{uni}}(t) = \Theta\Big(t-(1-b/2)\Big)\left(\frac{t-1+b/2}{b}\right) - \Theta\Big(t-2(1-b/2)\Big)\left(\frac{(t-2+b)^2}{2b^2}\right) \tag{4.5}$$

$$\begin{aligned} P_2^{\text{uni}}(t) = \Theta\Big(t-2(1-b/2)\Big)\left(\frac{(t-2+b)^2}{2b^2}\right) - \Theta\Big(t-3(1-b/2)\Big)\Bigg(\frac{27b^3+54b^2t}{48b^3} \\ + \frac{36bt^2+8t^3-162b^2-216bt-72t^2}{48b^3} + \frac{324b+216t-216}{48b^3}\Bigg) \end{aligned} \tag{4.6}$$

where $\Theta(t)$ represents the Heaviside step function (see [96, Eq. (1.16.13)]). Note that Eqs. (4.5) and (4.6) are only valid for $t \leq 1 + b/2$ (which includes the range of interest here; $t \leq 1$).

Unfortunately, for the other waiting-time distribution functions, the complexity of the

integrals limits to which order the approximation can be taken analytically. For instance, both P_1 and P_2 for the Weibull distribution must be computed numerically, while the values of P_2 for both of the power-law distributions also require numeric integration. The first order (using just P_1) approximation for each of the power laws can be computed analytically as

$$P_1^{\text{shifted}} = -\gamma\left(\frac{a}{2a+t}\right)^{2\gamma}\left(B\left(\frac{a}{2a+t};-\gamma,1-\gamma\right)-B\left(\frac{a+t}{2a+t};-\gamma,1-\gamma\right)\right) \tag{4.7}$$

and

$$P_1^{\text{cutoff}} = \Theta(t-a)\left(1-(a/t)^{\gamma}\right)-\Theta(t-2a)\left[1-\left(\frac{a}{a-t}\right)^{\gamma}-\gamma\left(\frac{a}{t}\right)^{2\gamma}\left(B\left(\frac{t-a}{t};-\gamma,1-\gamma\right)-B\left(\frac{a}{t};-\gamma,1-\gamma\right)\right)\right], \tag{4.8}$$

where $B(x;p,q)$ denotes the incomplete beta function $B(x;p,q)=\int_0^x t^{p-1}(1-t)^{q-1}dt$ (see [97, Eq. (8.17.1)]). The remaining cases of P_2 for the power-law distributions and both P_1 and P_2 for the Weibull distribution require numeric integration as mentioned above, and thus no explicit solution is presented here.

Using these formulas to find approximate values for the expected small-time activations via Eq. (4.3) (and using $t = 1$) yields the the results in Fig. 4.9, giving accurate representation of the approximate burstiness of each distribution with respect to its controlling parameter and thus categorizing their dominance within the early-time period of opinion spread modeling. In the most simple cases of the shifted power law and the uniform distribution, this just means accurately displaying that they are always more or less bursty (respectively) than the exponential, with trends towards the exponential for higher values of γ and b respectively. For the distributions with a transition point, however, this means accurately defining that point using only the approximation. For the Weibull distribution this is trivial as the Weibull becomes exactly the exponential when $\alpha = 1$, thus the approximation reduces to exactly that of the exponential as well. For the power law with lower cutoff, however, this prediction is more telling. In this case, the approximation predicts the transition point to be $\gamma \approx 1.64$, a value in very close agreement with the simulated results shown in Sec. 4.3.1.1. This agreement coming from a value produced using only information from the head of the waiting-time distributions further strengthens the assertion that the

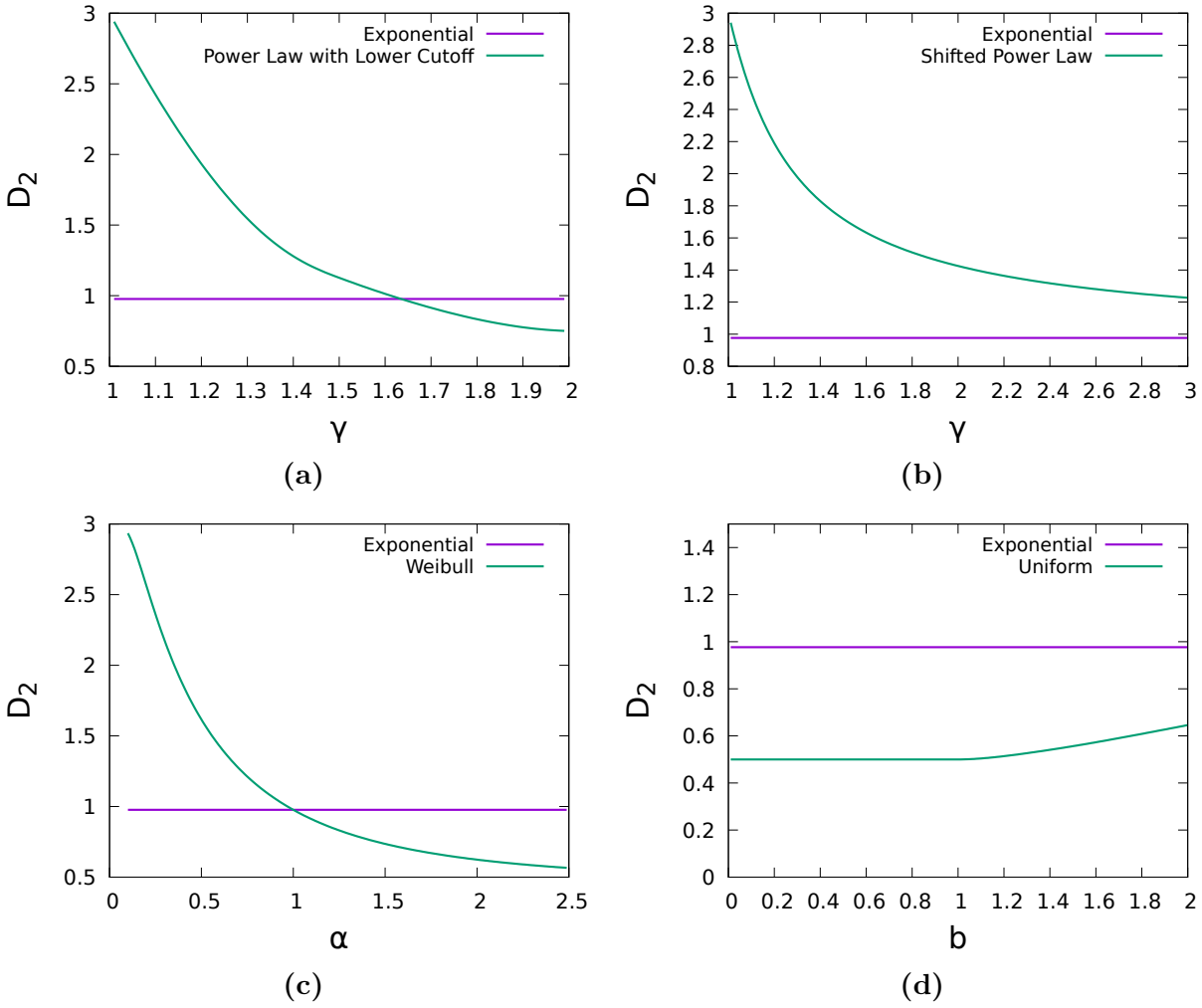


Fig. 4.9: Comparison of the second order approximation of the small-time activation densities for each of the non-exponential distributions vs the exponential. (a) shows the power law with lower cutoff, (b) shows the shifted power law, (c) shows the Weibull distribution, and (d) shows the uniform distribution.

dominant region of the distribution for the outcome of social simulations is the head density rather than the tail, as contributions from any other regions must be small and make up at most the $< 5\%$ difference between the values.

4.3.3 Erdős-Rényi Random Graphs

In the prior sections all analysis is focused on the dynamics of the competition on complete graphs, but it has been mentioned that qualitatively similar effects hold in the binary NG on Erdős-Rényi (ER) random graphs [98]. In the direct competition case (seen

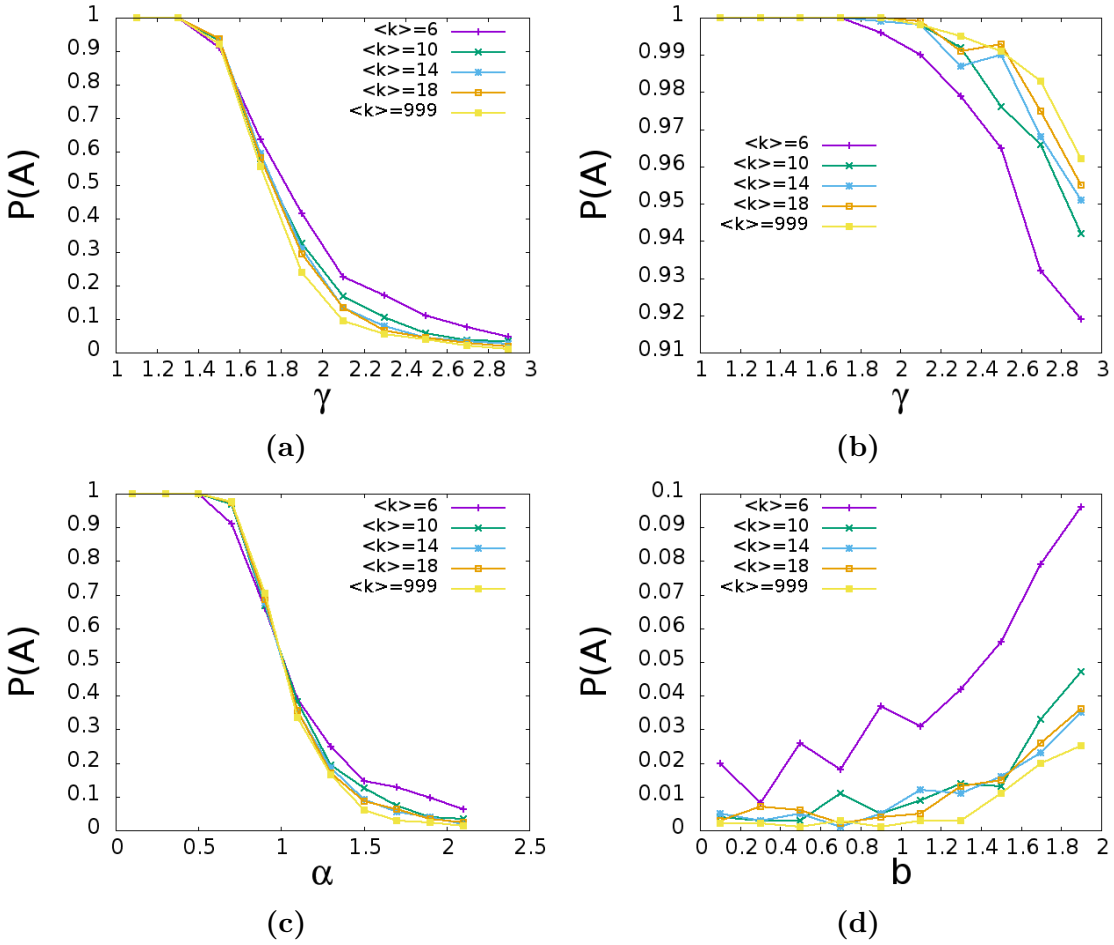


Fig. 4.10: The fraction of runs (out of 1000 trials) that reached consensus on opinion A in ER networks with $N = 1000$ nodes and various values of the average degree $\langle k \rangle$. Half of the nodes follow a non-exponential waiting-time distribution and initially have opinion A. The other half follow the exponential waiting-time distribution and initially have opinion B. The non-exponential distributions in each figure are (a) the power law with lower cutoff, (b) the shifted power law, (c) the Weibull distribution, and (d) the uniform distribution.

in Fig. 4.10), where the simulations are initialized in the same way as in Sec. 4.3.1.1, the average degree can be seen to have minimal effect on the outcome. Having a higher average degree corresponds to a slightly more well defined transition point, but the effect is extremely small in all cases. In general, the relative burstiness at which one group can dominate the simulation is unaffected by the average degree of the network on which the system is run.

Fig. 4.11 shows, however, that the average degree does affect the critical population of committed agents required for fast consensus in the system. This is to be expected, as prior

works have shown that lower average degree lowers the critical population necessary for a fast consensus of the system [49], [50], [54], [82], a result repeated here on systems where the nodes with non-exponential wait times are not very bursty and the system is similar to a normal naming game simulation. When high levels of burstiness are present, though, the consequent effect dominates over the average degree of the network, leading to similar critical populations for many different values of $\langle k \rangle$. When taken even further into the extreme cases of burstiness (such as the Weibull distribution with $\alpha = 0.1$), a lower average degree in fact raises the critical population, mitigating the effect of the extreme bursty nature of the nodes.

4.3.4 Individuals with Identical Burstiness

In prior sections, the analysis focuses on simulations with differing levels of burstiness among the competing groups, and no attention is given to the case where all nodes in the system has the same non-Poisson characteristics. In some cases this is due to the results being trivial; for instance for competition between two equal groups with no committed agents, the non-Poisson characteristic has no effect on the outcome. If the groups are of equal size at the start of the simulation they will each win approximately half of the simulations, and if one group is larger, it will win a larger number of the simulations just the same as if they followed Poisson selection patterns. In the presence of committed agents, however, the system is far less simple.

In this section, the case of committed agents in the naming game is considered where all agents (including the committed ones) exhibit *identical* bursty communication characteristics. Under these conditions, the critical fraction of the total population (tipping points) required for fast consensus on the system exhibits some small drift with regards to the burstiness of the waiting-time distributions used, as seen in Fig. 4.12(a), but still shows no drift with increased system sizes (Fig. 4.12(b)). In general, the critical fraction has very little dependence on the burstiness except for cases of extreme burstiness, such as a Weibull waiting-time distribution with $\alpha = 0.1$. In these cases, however, the effect is extreme as a result of the setup of the simulation. Each of these simulations with committed agents is set up so that there is some small fraction of individuals p that is committed and in state A, while the rest of the network is uncommitted and in state B. The simulation is then run either until consensus, or until $t = 150$ is reached, at which point the system is deemed as having not reached consensus. The critical population is then chosen to be the one where half

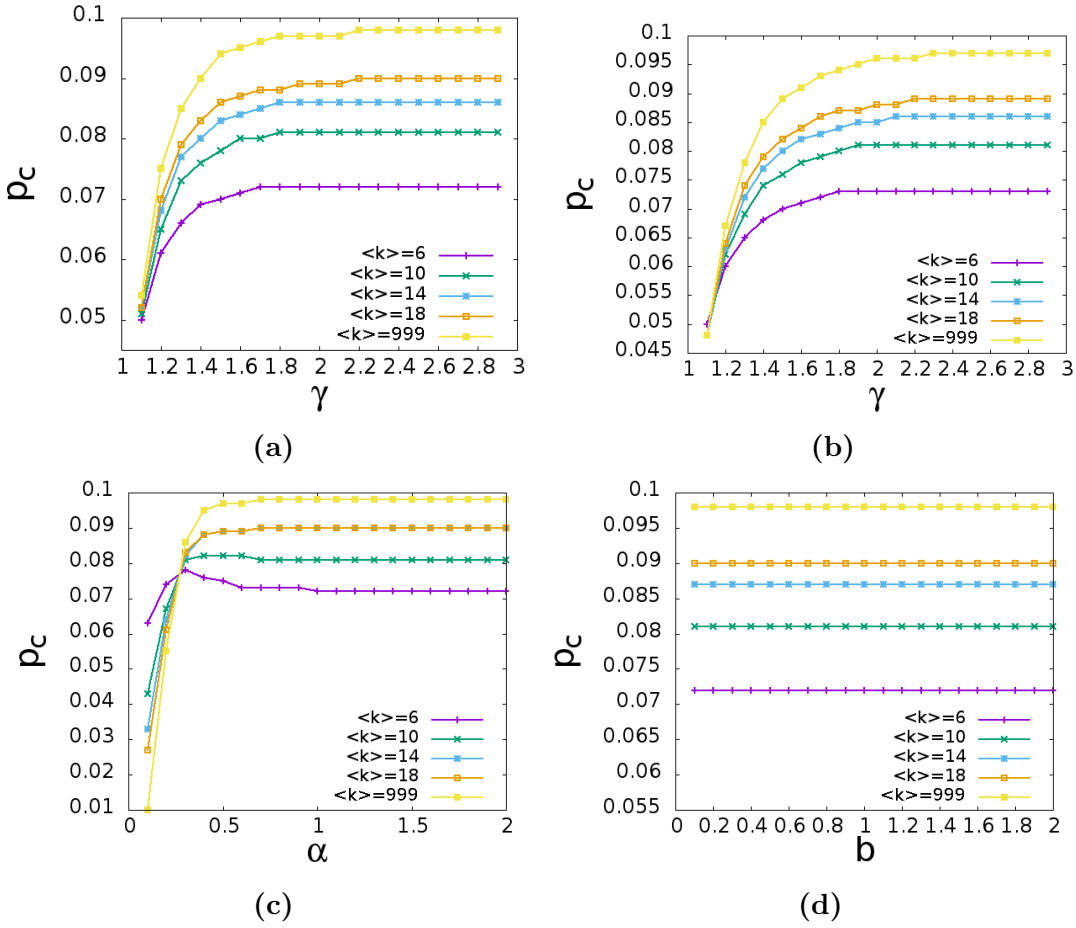


Fig. 4.11: The critical population p_c of committed nodes following a non-exponential waiting-time distribution that resulted in half of 1000 trials reaching minority consensus on ER graphs. $N = 1000$ with average degree $\langle k \rangle$. A minority fraction of the population p is committed to opinion A and follows a non-exponential waiting-time distribution. The rest of the nodes have opinion B and follow the exponential distribution. The non-exponential distributions in each figure are (a) the power law with lower cutoff, (b) the shifted power law, (c) the Weibull distribution, and (d) the uniform distribution.

of the simulations run reached consensus. This means that the system is somewhat sensitive to the value chosen for the long-time cutoff. For instance, a system left to run until $t = 1000$ will return a lower value for p_c because it is far more likely that somewhere in that time frame a large fluctuation will have pushed the system into consensus. The same effect can be achieved by increasing the number of speaking events per unit time t, yet again increasing the number of chances for a large fluctuation to occur. This is exactly what happens in this scenario, as evidenced by Fig. 4.13.

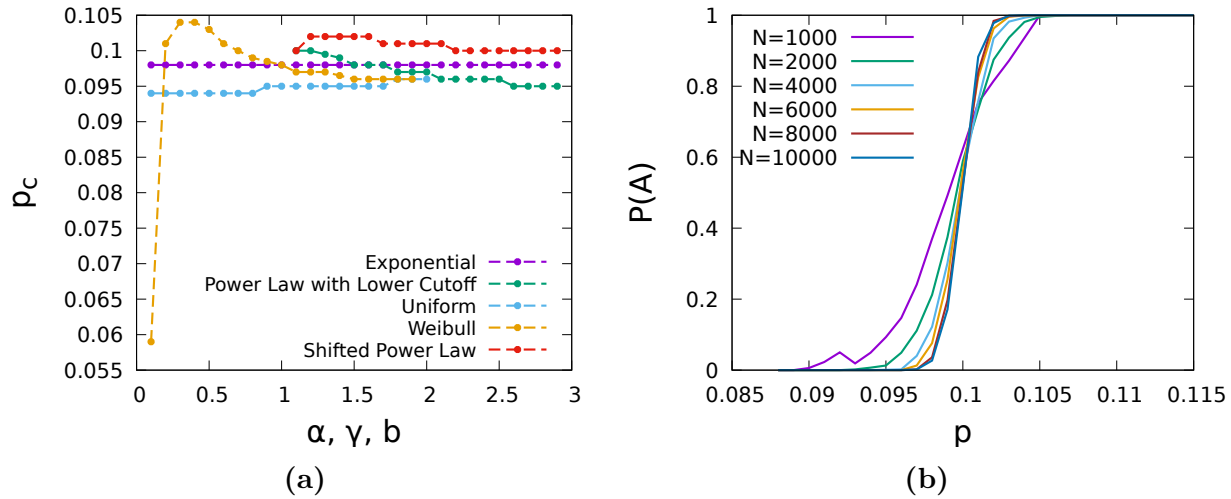


Fig. 4.12: (a) Critical populations of committed nodes (tipping points) in the binary NG on a complete graph when each node in the network has identical waiting-time distributions and the system size is $N = 1000$. (b) The Fraction of runs reaching consensus in 1000 simulations (by time $t = 150$) vs the fraction of committed individuals for various system sizes. In this plot, each node has the Weibull waiting-time distribution with $\alpha = 1.3$.

Fig. 4.13(a) shows that the number of speaking events per unit time in these simulations with committed agents is extremely high for the very bursty case of a Weibull waiting-time distribution with $\alpha = 0.1$, but levels out quickly for more reasonable parameter values. This is in line with what is seen in Fig. 4.12(a), where the only large deviation based on burstiness is from the simulation using $\alpha = 0.1$. At first glance, it is not clear why the rate should be so much higher in this case than others, considering the construction of the waiting-time distribution to have $\langle \Delta t \rangle = 1$, but Fig. 4.13(b) shows that for these extreme values of α, the rate does not begin to normalize down to one until an extreme long time limit is reached. Similar results were obtained for the two power-law distributions, however reaching such a ill-behaved parameter set for those distributions required values of γ much close to the limit of $\gamma = 1$ than were present in the tests in Fig. 4.12. In fact, most simulations with committed agents complete in around $t \approx 50$, making the max allowed time of $t = 150$ reasonable for nearly all of the distributions used. For the most extreme cases, however, this creates an abnormally high activation rate that can skew the results as shown.

The high rates of activation in the short times effectively explain the single large deviation in Fig. 4.12(a), but also explain some of the other irregularities contained within.

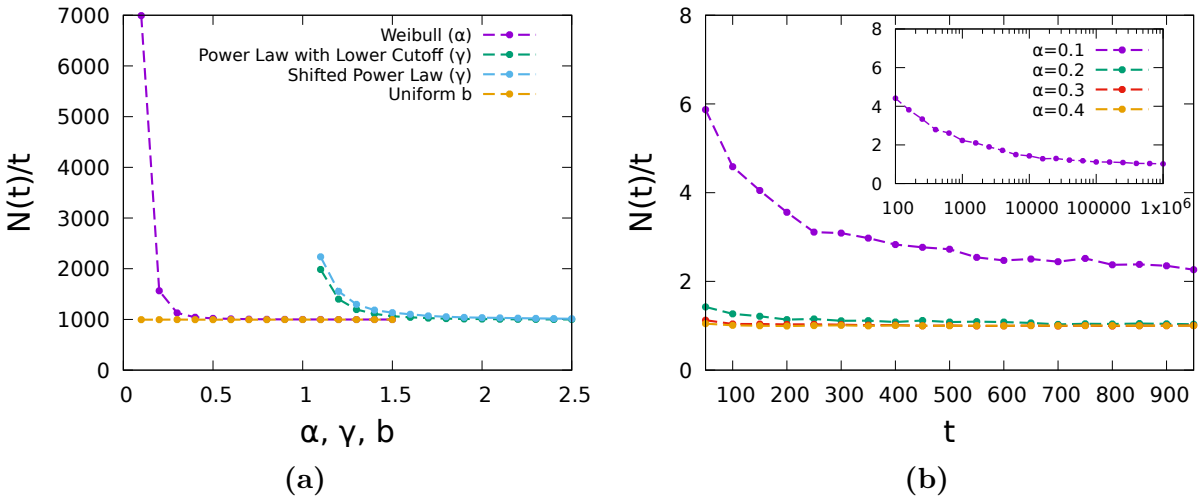


Fig. 4.13: (a) Number of speaking events per unit system time relative to the type of waiting-time distribution used in the system for the same simulations as in Fig. 4.12(a). As such, the simulations are still done on system of $N = 1000$, over 1000 runs on a complete graph. (b) Average number of speakers' events per time step for a single node updating with a Weibull distributed waiting-time over different time intervals. The values are for the updates of a single node averaged over 1000 simulations. The inset shows the data for $\alpha = 0.1$ on extended (logarithmic) time scales.

Upon close inspection, the distributions all either monotonically increase or decrease a very small amount, except for the Weibull and shifted power-law distributions. These are the two distributions with the highest propensity for burstiness, and each has an inflection point where they change from concave to convex for values of p_c. This inflection point is in the same spot as the normalization point for each distribution's speaking events per unit system in Fig. 4.13(a). From this it can be gathered that outside of the effects of an abnormally increased speaking rate, increased burstiness works to *increase* the critical fraction of the population in non-Poisson update systems by a small amount (the same pattern can be seen in the systems with uniform and power law with lower cutoff distributions), indicating that increased burstiness over the entire system hinders spreading in these systems.

4.4 Conclusion

Attempts to bring more realistic human communication patterns to social dynamics models are often difficult, but understanding the effects of different changes helps to further

bring the abstract models closer to reality. Using a Poisson process to select speakers in pairwise interaction models is popular and extremely attractive for its simplicity, yet it is quite different from empirical descriptions of people behave. Not only do people tend to have a much burstier communication patterns than are present in a Poisson selection process, but they tend to act individually and thus heterogeneously. When implemented on common pairwise interaction models, these processes can add up to creating great effect on the behavior of the system, giving a powerful advantage of the community with burstier communication patterns. Further, in the presence of intermediate states such as those seen in the binary naming game, this effect is stronger as the system size increases, demonstrating that the symmetry of the system is broken and allowing even a very small difference in the waiting-time distribution to have a large effect on the final outcome of a simulation for a sufficiently large system. Without the intermediate state (such as in voter model simulations), however, this scaling effect is lost, though the overall bias towards the burstier community remains.

Further, when committed agents are introduced to the models, prior work indicates that there are many factors that can impact the value of the tipping point; including the number (or fraction) of committed agents, the level of commitment of those agents, their eagerness to leave an intermediate opinion state, the average node degree, and their rate of activation relative to other nodes in the system [31], [49], [50], [54], [82], [77], [78], [81]. The results presented here indicate another factor: the waiting-time distribution of the committed agents relative to that of the surrounding nodes, as the general bias towards a more bursty community remains with the presence of committed agents. In fact, the waiting-time distribution effect is particularly interesting in situations where it is desirable to minimize the size of the committed fraction because the heterogeneous waiting-time distributions can have only a positive impact on the efficiency of the committed agents. If the committed agents are less bursty, the system simply enters the long time regime and reverts to the critical fraction for a system of homogeneous nodes. This effect is important in the study of facilitating the growth of a single opinion in a society, as it implies a new strategy consisting of multiple strong pushes for the new opinion even if they are separated by long periods of inactivity. Such a pattern can heavily decrease the cost of spreading an opinion throughout a society by increasing the efficiency of any concerted effort by activists to aid the spread. Additionally, it is shown that burstier communications patterns shared among every node

in the system work to *inhibit* spreading (outside of the notable exceptions mentioned in Sec. 4.3.4), meaning not only is this advantage lost if the entire population picks up the committed agent's activity patterns, but the effect is actually reversed and *more* committed nodes are required to create consensus.

Finally, despite the difficulties of analytic considerations of this type of system due to the the non-Markovian nature of the various selection processes, the phenomenon can be accurately approximated by calculating the expected number of activations before the mean wait time. This allows for a relatively simple method for comparing the burstiness of different distributions in terms of their impact on the early-time period of a simulation. Using this information from only the head of the nodes' waiting-time distributions, accurate predictions on the outcomes of simulations can be obtained, further clarify the mechanisms through which systems that are otherwise entirely symmetric can be heavily biased towards one opinion or another via changing the waiting-time distribution of a portion of the nodes.

CHAPTER 5
EMPIRICAL BEHAVIOR

To this point, this work has focused on only one side of the problem of realistic social modeling: bringing the theoretical models closer to the realm of empirical understanding. Progress in this sphere focuses on making more complex models and frameworks with which to simulate and describe real behavior, but it relies on a certain sense of generality and abstraction to allow for the unpredictability of human behavior. For instance, in Chap. 3 (and the related waning commitment model), the commitment levels of individuals to ideas are left to the arbitrary value w. Since it is not clear how staunchly people stick to their ideas or what it takes to make them change given a certain situation, these values are left to be fitted to specific applications and solved in the most general case in the meantime. Similarly, Chap. 4 uses arbitrary probability density functions to describe the inter-event times for speaking nodes to account for people behaving differently in different situations (or designing behavior to best fit a situation).

There is another side of social modeling, though, that focuses on bringing the empirical understanding of human behavior closer to modeling. In many ways, this is the driving force behind how the models move forward. There would be no waning commitment or inertia without an intuitive understanding of the role of stubbornness in opinion dynamics, and there would be no studies on the effects of heterogeneous communication wait-times without prior work showing the inaccuracy of the Poisson selection process [72], [73]. In this chapter, further efforts are made to understand social behavior in a way that can feed into these computational models. First, Sec. 5.1 informs on the specific cases of waning commitment, inertia, and other similar social contagion models by investigating individual response thresholds to form or change opinions. Then, for more general application Sec. 5.2 analyzes a large scale social network to gain a better description of its structural properties. While this is not directly applicable to the mostly dynamics based work in the prior chapters,

Portions of this chapter are to appear in: C. Doyle, A. Meandzija, G. Korniss, B. K. Szymanski, D. Asher, and E. Bowman, "Mining personal media thresholds for opinion dynamics and social influence," in *IEEE/ACM ASONAM 2018*, Barcelona, Spain, 2018.

Portions of this chapter previously appeared as: Z. Herga, C. Doyle, S. Dipple, C. Nasman, G. Korniss, B. Szymanski, J. Brank, J. Rupnik, and D. Mladenic, "Building Clients Risk Profile Based on Call Detail Records," in *SiKDD*, Ljubljana, Slovenia, 2017.

deeper structural knowledge of social networks provides another alternative to the standard random graph models that are often used lieu of real networks.

5.1 Opinion Thresholds

5.1.1 Motivation and Related Work

The importance of understanding how individuals interact with and are influenced by online media has been growing as people continue to get increasingly large amounts of their news via various social networking websites and media exchange platforms [99], [100], [101]. The problem of knowing exactly how an individual takes in and processes this information is difficult to accurately study, and further the variation among different people makes it even worse when scaling up to understand how the preferences and tendencies of an individual inform on the behavior of the larger population as a whole. In this section, some of these difficulties are bridged, extending the work done in [102], [103]. In those studies, a group of individuals was given a survey on the number of media items they would need to consume given various parameters such as the media type (format), source (general like mindedness) and context (general level of controversy of the subject). With this information, general thresholds were established for various media types as well as their interaction with the source and context of the media. Yet, each individual was asked a very limited set of questions to establish the general behavior of a large population. Here the scope of the study is broadened such that each individual is asked a larger number questions covering many of the combinations between media types, sources, and contexts in order to establish a better profile for how the individual responses change for each person. Additionally, a new field is added for some participants where they are asked about the number of media items required for "shifting" their opinion instead of "forming" it. Using this data, various frequent patterns are mined, attempting to pick up on different types of people whose behavior deviates from the average while still being common enough to not be considered outliers.

While data mining techniques are popular in the realm of opinion formation, they are mostly used to understand the evolution of opinions as they change in empirical networks [104], [105], [106]. Other applications of data mining to the field have been applying mining techniques to extract personality types and behavior from social network and cell phone data [107], [108]. In contrast, this study uses a more direct data set and frequent pattern detection tools to search for categories of people that relate directly to their social

media use and behavior. First, Sec. 5.1.3.3 investigates the basic statistics corresponding to the data, then analyzes the pattern lists to search for strong relationships between items. Finally, Sec. 5.1.3.4 presents the combinations that represent the most interesting relationships as well as general conclusions that can be drawn from the overall shape of the patterns found.

5.1.2 Description of Data

5.1.2.1 Platform and Participant Selection

The data in this experiment was collected using Amazon Mechanical Turk (MTurk), which is a survey hosting platform that connects researchers to diverse pools of participants [109], [110], [111]. In total, the experiment involves 1431 participants. Before beginning the survey, participants were asked if they use social media; only those that responded 'yes' and had not already participated were admitted to the study. In addition, only users aged eighteen years or older and located in the US were accepted for participation in this experiment.

5.1.2.2 Data Collection

The questionnaire used was simply an extension to that used in prior work [102], [103]. After being screened for social media use, the participants were assigned randomly to one of two groups, the "fixed-source" group or the "fixed-context" group. Users were then asked to fill out a brief demographic questionnaire (Fig. 5.1) with general information about the user as well as more detailed questions on their social media usage (favorite websites, amount of social media consumption, and main news sources). The responses to this demographic and usage questionnaire are not covered in this work, but could be the source of future work in identifying different consumption behaviors within demographic groups or social media communities.

After the demographic form, the subjects were asked to report on the number of social media items they would need to see before forming an opinion on the subject of the media. These questions include three parameters that describe the media: *type*, *context*, and *source*. We identify three media types:

1. Images: for still photos and drawings

2. Videos: for any animations or moving pictures

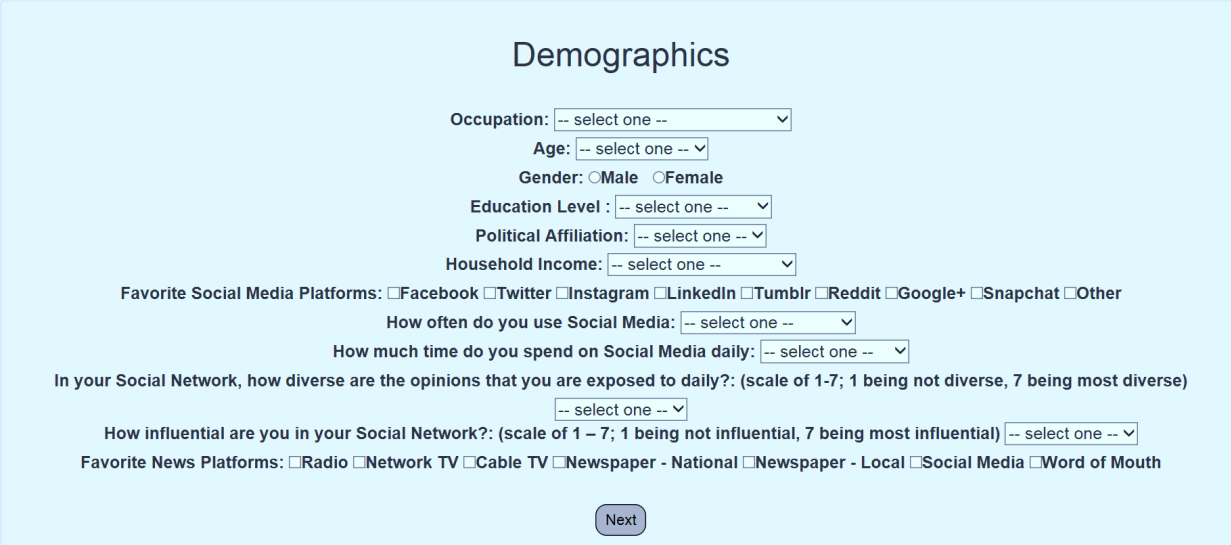


Fig. 5.1: Demographic questions asked of participants in the study. Questions were presented on a single page and all questions were required to be answered before moving onto the media threshold questions.

3. Messages: for text, tweets, and Facebook posts

Four media controversy levels defined:

1. Low: minimal (some people would form an opinion)
2. Medium: generally controversial (most would form an opinion)
3. High: very controversial (most or all would form an opinion)
4. None: no reference to controversy

And three media sources:

1. Unknown: individual has no knowledge of the source
2. Like-minded: the source of the media generally thinks similarly to the recipient
3. Different-minded: the source of the media generally thinks differently from the recipient

A sample question for this portion of the survey is shown in Fig. 5.2. Each question was given its own page, and participants were only asked for one response at a time. The controversy level and media type were listed above and below the question, respectively, and the source was incorporated into the question. Participants were required to answer

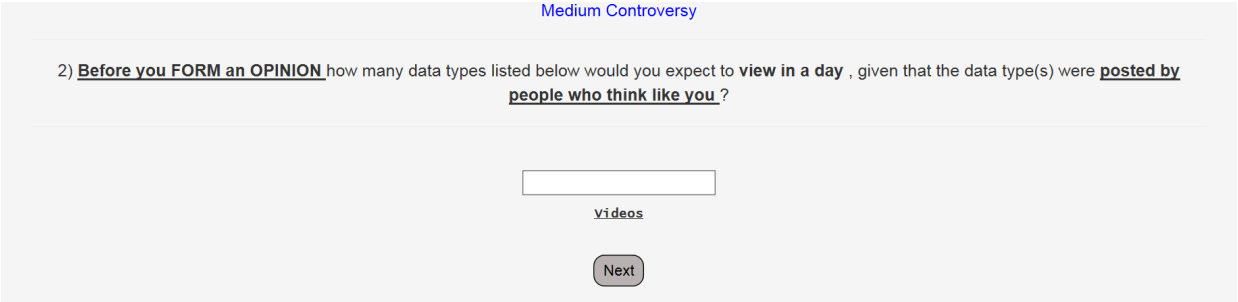


Fig. 5.2: Sample question asked of participants. Each question is presented on its own page, and participants are given the context at the top, then asked a question containing the source, and given the media type in question at the bottom.

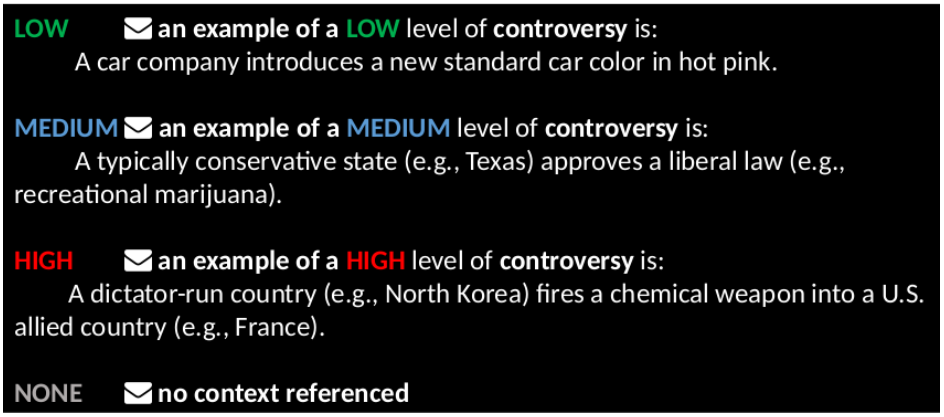


Fig. 5.3: Definitions of the context values used throughout the paper, as given to participants in the study.

each question to move onto the next question in the survey, and their prior responses were not visible to them when answering the following questions. This process continued until the participant was asked every combination of media type, source, and context for their randomly chosen grouping and the survey was complete. Upon completing the survey, the participants were paid a small sum and thanked for their time.

Users were also given the definitions of different controversy levels to be used in the study with examples of each controversy level on the instructions page to better orientate each participant to the same approximate level of controversy for each question (Fig. 5.3). In this work, context is equivalent to controversy, so the examples given for controversy here correspond to the definitions for the different context values.

To limit survey fatigue, the number of questions asked was reduced from the full set of possibilities. Instead of asking each participants all possible combinations of the three

parameters, they were asked a subset depending on the group they were initially assigned to. The participants in each group ("fixed-context" and "fixed-source") were given surveys that corresponded to each possible combination of the above parameters with the exception of their group parameter. For instance, individuals in the "fixed-context" group were assigned a random context at the start of questioning, and asked about every combination of type and source with that one context level. Similarly, the "fixed-source" group was given a single source and asked about all combinations of type and context.r

Finally, the "fixed-source" group was given an extended questionnaire that included questions on how many media items they would need to see to *shift* their opinion as well as the original set that asked about *forming* their opinion. In these cases where the participants were asked about *shifting* their opinion instead of *forming* it, the question was the same as shown in Fig. 5.2 only with 'shift' in place of the word 'form' in the given example. From here on out, responses to the opinion shifting questions will be referred to as their own group: "shifting", and the "fixed-source" group will be implied to mean data only from the opinion formation answers. When discussing the full suite of answers provided by those individuals, the data will be referred to as the "shifting-formation" group. The final sizes of each group are 616 users in the "fixed-source" group (and consequently 616 users for the "shifting" and "shifting-formation" groups) and 815 users in the "fixed-context" group.

5.1.3 Results

5.1.3.1 Prior Analysis and Binning

The prior results on similar data sets revealed some key features of how individuals respond to these types of questions [102], [103]. First, reported thresholds are shown to be distributed log-normally, and thus a log transform can easily be performed to normalize the data for analysis. Second, both the source and context each have a significant effect on the requisite number of media items viewed dependent on the value of other parameters. In general, images are less susceptible to change depending on other parameters, while videos are more sensitive to source and messages are more sensitive to context. Average values for each media type are $4-7$ images, $2-5$ videos, and $3-6$ messages required to form opinions. This work does not repeat much of this analysis as it is beyond the current scope, but the overall average number of media items to form an opinion (4.5 items) is commonly seen throughout this analysis.

Using the prior results to inform on the current analysis, the data was cleaned and pre-processed for mining by coding reported thresholds into bins. Since the data has been shown to be normalized via a log transform, we utilize a logarithmic binning scheme to sort the data (i.e., bin one contains responses of threshold 1, bin two contains responses with thresholds $2-3$, bin three responses of thresholds $4-7$, bin four responses of thresholds $8-15$, etc.). From here on in all mentioned 'response' values will correspond to the log-binned value of the reported thresholds, not the raw threshold given. This allows for better discussion of the generalities of the behavior, as threshold values represent the category instead of specific values that have less meaning in a self-reported data set like the one here. Additionally, the binning allows for a fuzzy look at thresholds for the sake of finding more representative patterns by grouping together similar thresholds instead of attempting to pick out only patterns that contain the exact same values.

5.1.3.2 Association Rule Mining and Processing

In order to identify interesting subsets and trends within the overall population, *frequent pattern* and *association rule* mining were performed on the data set. The frequent patterns are simply itemsets that appear commonly (identified via a minimum support defined as the fraction of all transactions that contain that pattern), while the rules contain directional information on the implications of the frequent itemsets, i.e., individuals that give responses A and B are also likely to give response C [112]. Rules are defined as frequent via both a minimum support (the fraction of all responses that contain the set, Eq. (5.1))

$$sup(X \rightarrow Y) = P(XY) = P(YX) \tag{5.1}$$

and confidence (defined as the probability that a transaction contains the consequent given that it also contains the antecedent, Eq. (5.2))

$$conf(X \rightarrow Y) = P(X|Y) = \frac{P(XY)}{P(X)} = \frac{sup(XY)}{sup(X)} \tag{5.2}$$

The frequent patterns were mined using the Apriori algorithm within the `arules` package in R [113], [114], [115]. For the "fixed-context" group, the minimum support value $sup = 0.01$ was used with a confidence of $conf = 0.6$, yielding 3263 rules with a minimum absolute count of 8. Similarly, the "fixed-source" group was mined with a minimum support

of $sup = 0.015$ and a confidence of $conf = 0.6$, yielding 4716 rules with an absolute minimum count of 9. In order to remove redundancies they were then filtered to remove any rule belonging to an itemset that is not maximal [112]. Maximal itemsets are those for which all super sets of the itemset are not frequent (if Eq. (5.3) holds true)

$$sup(X) \geq sup_{\min} \quad \&\& \quad Y \supset X : sup(Y) \leq sup_{\min} \tag{5.3}$$

Further, all rules were tested for statistical significance via the Fisher Exact Test (using $p > 0.01$, see Ref. [112] for details), and insignificant rules are also removed. This yielded a final rule set of 1484 in the "fixed-context" group and 2212 in the "fixed-source" group.

The mining on opinion shifting was much the same. The "shifting" group was mined with a minimum support of $sup = 0.015$ and minimum confidence of $conf = 0.6$, yielding 2767 rules with a minimum absolute count of 9. Due to its larger size, the "shifting-formation" group was mined with a minimum support of $sup = 0.025$ and confidence $conf = 0.6$, yielding 2346 rules and a minimum absolute count of 15. After pruning for maximal and significant rules, the sets became 1790 and 1957 rules long, respectively.

The effects of these restrictions can be seen in Fig. 5.4, where the majority of high support rules are lost due to not being maximal or significant (and thus being lesser reproductions of longer rules). Among the rules that remain after the pruning, however, include the vast majority of high *lift* scores (described in Sec. 5.1.3.3), indicating that the rules that are most statistically surprising are preserved.

5.1.3.3 Response Statistics

This section looks at a few different cross sections of the data, starting with the general user data to get a view of how the individuals behaved as a whole. Then, this analysis is focused to look at the mined rules, investigating if there are any patterns or statistics among the frequent itemsets that differ from the user statistics. Finally, the rules are ranked based on their lift value to determine those that are most statistically interesting. The lift score is a measure of the surprise of the rule, defined as the ratio of the percentage of times the pattern appeared in the data set divided by the probability that the pattern would arise randomly if the items within were independent, represented as

$$lift(XY) = sup(XY)/sup(X)sup(Y) \tag{5.4}$$

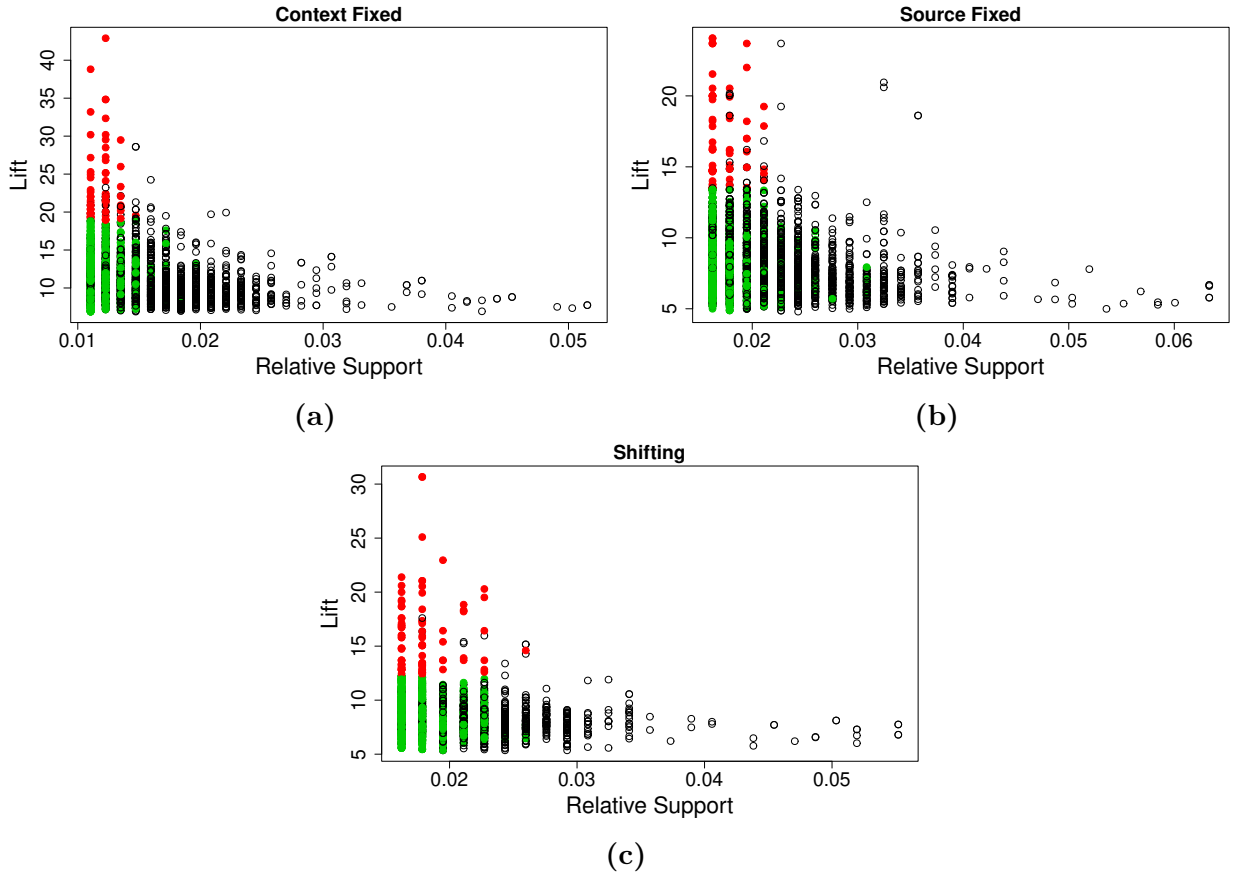


Fig. 5.4: All mined rules for each group with regards to the relative support and lift score of those rules. Highlighted in green are rules that are both maximal and productive (statistically significant). In red are the 'Top 100' rules; corresponding to the highest 100 lift scores among the maximal and productive rules. (a) shows rules from the "context-fixed" group, (b) from the "source-fixed" group, and (c) from the "shifting" group.

After being ranked based on their lift, the top 100 rules are studied more closely to see how the most unlikely patterns behave in relation to the rest.

The main focus of this analysis is to study how different parameters affect the response scores in a more group-focused approach than the prior analysis. By calculating the average and standard deviation for each rule we can observe some initial trends in the rules, shown in Table 5.1. Immediately it is clear that the rules have a generally very low average standard deviation, meaning that most rules contain consistent thresholds for varying parameter values. This effect is strong for all rules when compared to the thresholds for the users as a whole, indicating the rules are picking up on a tendency outside of simply general user

Table 5.1: Average thresholds for rules based on survey results and mined for frequent patterns. "Average SD" corresponds to the standard deviation of the whole rule set, while "Average SD w/ Change" corresponds to only rules that had variation within their responses. "Top 100" rules are those with the 100 highest lift values.

Group	Rules	Average Threshold	Average SD	Average SD w/ Change
Fixed Source	All	3.021	0.217	0.568
	Top 100	2.257	0.061	0.544
Fixed Context	All	2.774	0.254	0.578
	Top 100	2.620	0.063	0.608
Shifting	All	3.064	0.204	0.578
	Top 100	2.717	0.115	0.566
Shifting Formation	All	2.749	0.079	0.586
	Top 100	2.419	0.012	0.577

Table 5.2: Average thresholds for all participants based on survey results. "Average SD" corresponds to the standard deviation of the whole response set, while "Average SD w/ Change" corresponds to only users that had variation within their responses.

Group	Average Threshold	Average SD	Average SD w/ Change
Fixed Source	3.228	0.826	0.884
Fixed Context	3.183	0.870	0.892
Shifting	3.392	0.880	0.909

behavior. Further, the effect is increased for the top 100 rules by lift. In both cases the effect is true even if, to account for the many short rules with no changes at all, we include only rules with changes. Even with that reduced data set the standard deviation of rules is noticeably lower than that for the users as a whole. Another important difference between the lists is the average response for the rules is lower than the responses in general, another effect that is stronger among rules with high lift scores. These findings show two important features of the frequent patterns within the data sets: there are significant groups within the populations that tend to have consistent parameter subsets that lead to lower general thresholds for forming and shifting their opinion.

Some of the details causing this effect can be seen in Fig 5.5, where the distributions show a large difference in the tail end of the responses. The averages for all users tend to drop off after bin four (reported thresholds ranging from eight to fifteen), but there is a size-

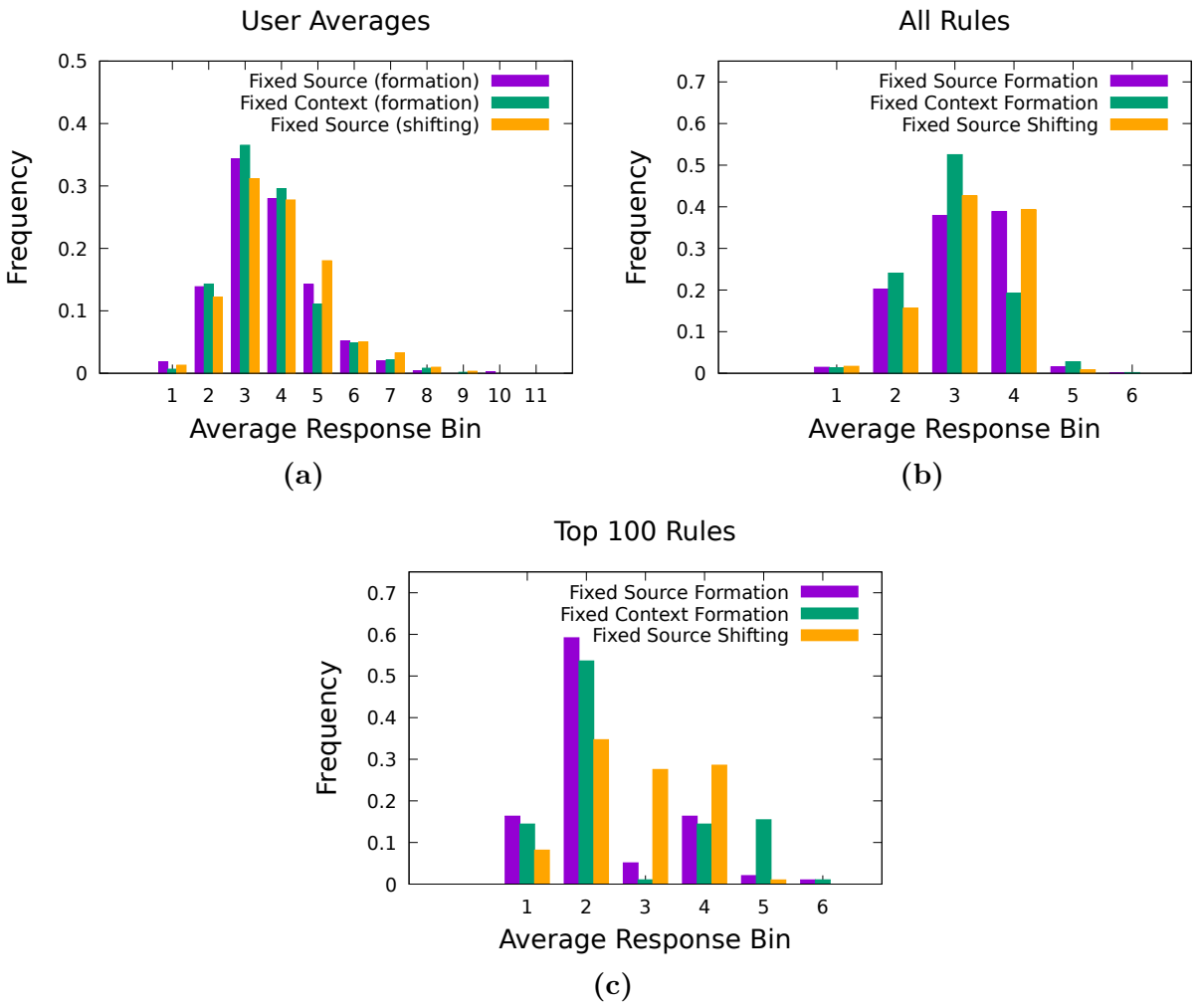
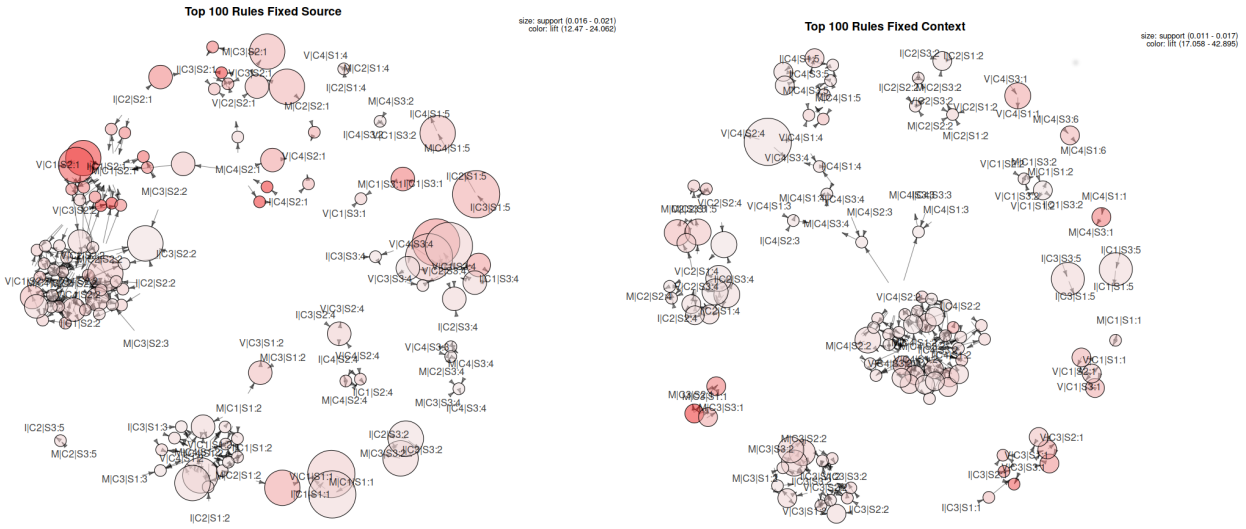


Fig. 5.5: (a) Distribution of the average thresholds reported by each user. Thresholds are binned logarithmically as described in Sec. 5.1.3.1. (b) Distribution of the average thresholds within each calculated rule. (c) Distribution of average thresholds from only the top 100 rules by lift.

able tail to the distribution that contains many more extreme responses. When studying the averages across the rules, the tail disappears, and after the drop-off at bin four there are nearly no responses with thresholds in the higher bins. In this case it is not necessarily that the rules have a higher volume of low values, it's that users that respond with high values don't tend to be consistent enough to form significant groups despite the logarithmic providing more flexible boundaries for group inclusion at those values. The top ranking rules by lift show that this is no mere statistical effect, either, as the rules that contain responses with the lowest thresholds dominate the high lift rules. Interestingly this is not absolute, as despite the head dominance of the high lift values there are still significant groups of



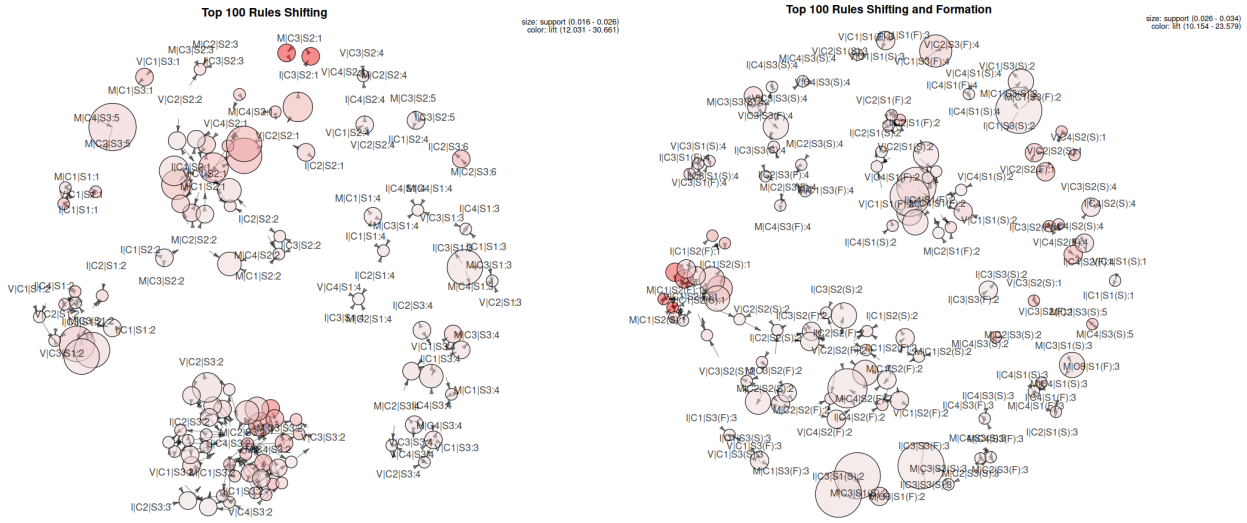
(a) (b)

Fig. 5.6: Network representation of the top 100 rules by lift for the (a)"fixed-source" and (b)"fixed-context" groups. Each circle represents a rule with the connections being the items within those rules. The size of the circle is scales with the support of the rule, while the color represents the lift score (darker red corresponds to higher lift).

individuals that respond with the higher threshold values. So while it is possible that the volume of rules with average thresholds similar to those of the population as a whole could be a relic of the high number of responses with thresholds in that range, the high lift rules show that there are statistically interesting groups that respond mostly with low thresholds.

5.1.3.4 Contents of Rules

With a basic understanding of what the general responses are and how they are represented when mining for rules, the next step is to understand the general content patterns within the rules. In Fig. 5.6, a visualization of the top 100 rules is presented for the "source-fixed" and "context-fixed" responses. Inspecting these plots provides a qualitative understanding of how the individual rules tend to interact with each other. In these cases, the rules are highly modular, dominated by a few very large and tight clusters. As would be expected from the prior analysis, these clusters contain many rules that match together similar responses. Similarly, Fig. 5.7 shows a graphical view of the top 100 for the "shifting" and "shifting-formation" groups. These plots show that the "shifting" group still maintains some of this structure, although there are fewer large clusters than with the formation groups. Moreover in the "shifting-formation" group the the large structures are largely gone and re-



(a) **(b)**

Fig. 5.7: Network representation of the top 100 rules by lift for the (a)"shifting" and (b)"shifting-formation" groups. Each circle represents a rule with the connections being the items within those rules. The size of the circle is scales with the support of the rule, while the color represents the lift score (darker red corresponds to higher lift).

placed by many smaller, isolated structures. Again, this mirrors the results in Table 5.1, which shows a much lower standard deviation for this group than for the others arising from the lack of rules bridging the gaps to make larger structures, and instead staying more within the same threshold values.

To look deeper into what makes the rules (and in particular what drives changes in threshold values), the statistics of the parameter makeups of each rule are presented. Unfortunately, a clear understanding of these parameter effects is somewhat difficult to accomplish due to the deep interplay between the three parameters. Each parameter contributes to the user reported thresholds in a different way, and nearly every rule has at least one change in parameter so it is challenging to isolate the effect of a single contribution. For instance, in the "fixed-source" group, 94% of rules have at least one change in media type, while 97% have at least one change in context level. This remains true for each of the other groups as well, with every group having $> 90\%$ of their rules changing in both media type and context (or source for the "fixed-context" group). However, looking more deeply into the makeup of each of these rules reveals some common trends.

Fig. 5.8 shows that for the formation groups, the media types are relatively stable across

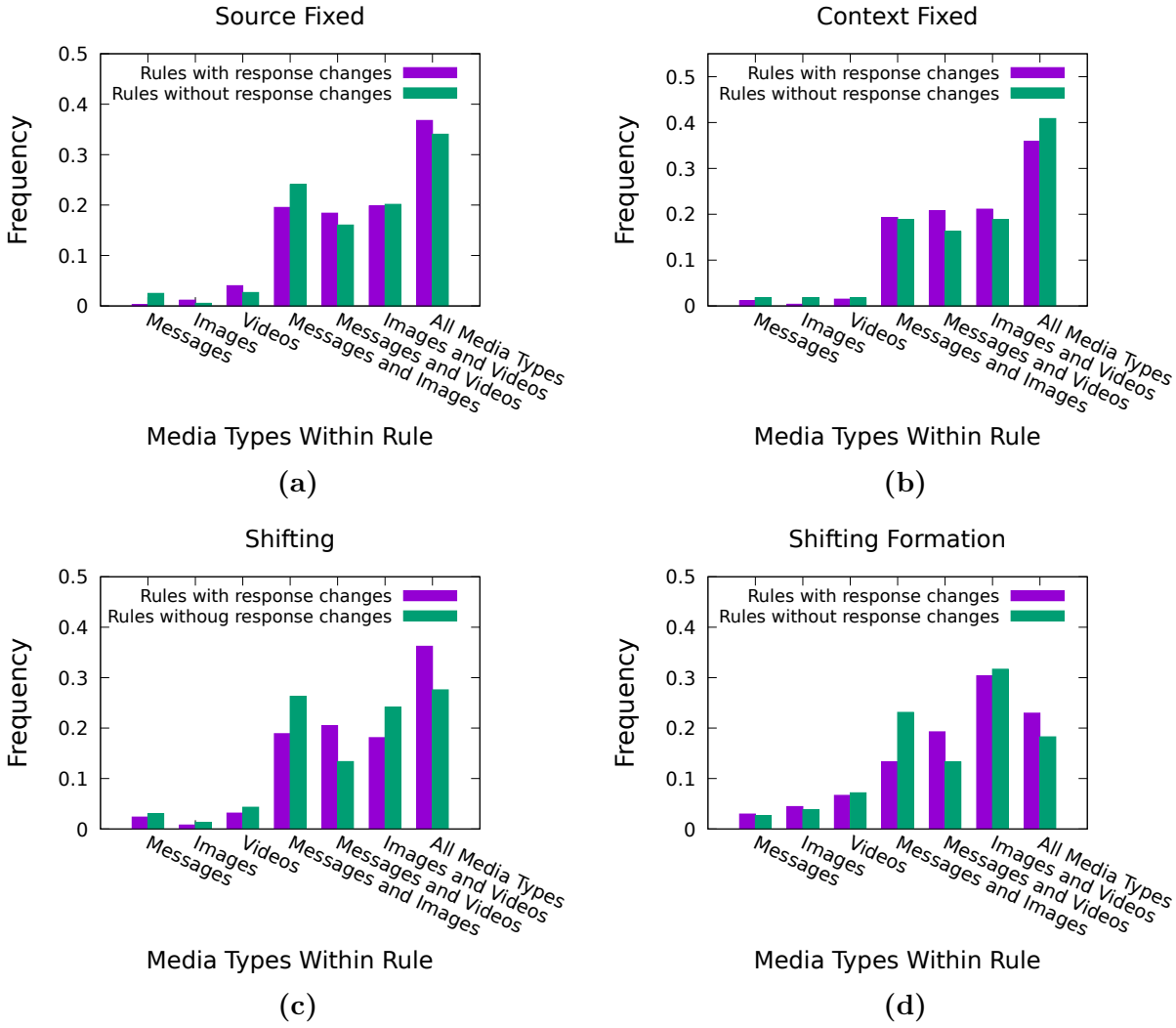


Fig. 5.8: Percentage of rules that contain each possible combination of media types, separated by whether that rule also contains a change in response. Groups corresponding to the plots are (a)"fixed-source", (b)"fixed-context", (c)"shifting", and (d)"shifting-formation".

combinations. There are very few rules with only a single type, and no clear preference for any pairing of two media types. There is, however, a slight difference in the percentage of rules in each category for the "fixed-source" group when looking at rules with messages and images versus those with messages and videos. This effect is even more pronounced in the groups where shifting responses are included, and the percentage of rules that produce no change is much higher for rules that contain messages and images than those that contain messages and videos. Similarly, the percentage of rules that contain a change in response and include both messages and videos in them is higher than that which includes messages

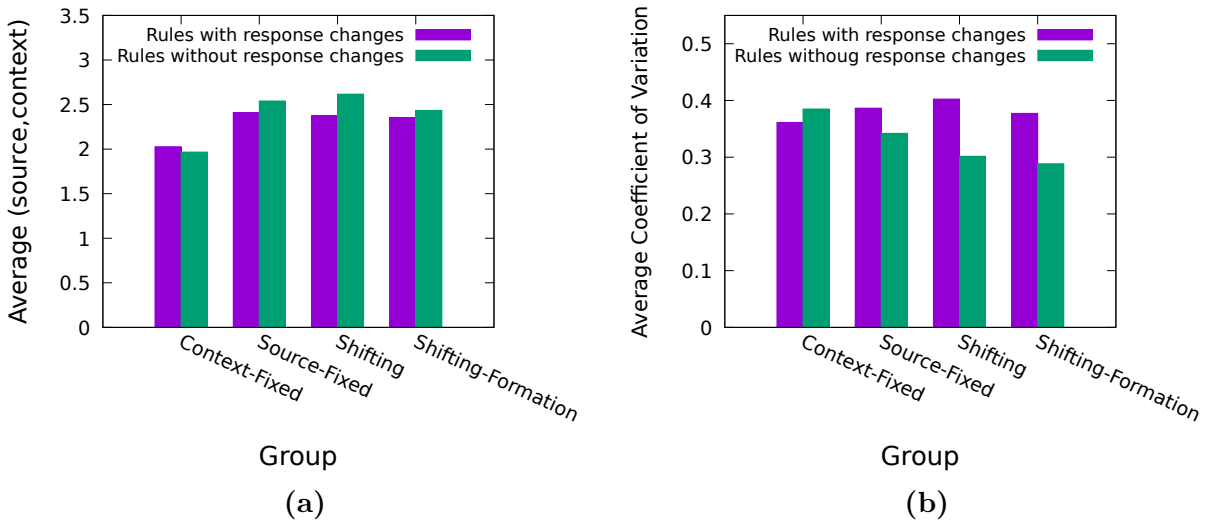


Fig. 5.9: (a) The average source (for "fixed-context" group) or context (for all other groups) level for each rule averaged over all rules within that group, separated by whether the rule produced a change in response. (b) The average coefficient of variance (σ/μ) of the source (for "fixed-context" group) or context (for all other groups) level for all rules within that group, separated by whether the rule produced a change in response.

and videos but no change in response. From this, it can be concluded that users tend to think of messages and images as being more similar in their potential to shift opinions than messages and videos.

In Fig. 5.9, this analysis is extended to rules that include response changes to investigate the source and context levels within those rules. The average values for the source (in the "fixed-context" group) and context (in the "fixed-source" group) are generally very close to the expected mean for a random sample, but when separated based on whether the rules contain a change in response or not some trends begin to emerge. First, the source level is slightly higher in the rules that produce change, indicating more changes happen when the source is differently minded, although the differences are too small to draw any concrete conclusions. A similarly small effect can be seen in the "fixed-source" group indicating that lower context values (less controversy) are slightly more common in rules that produce changes. This effect is also present in the "shifting" and "shifting-formation" groups, with the "shifting" effect being the most pronounced. Similarly, the coefficient of variation (a normalized version of the standard deviation $C_v = \sigma/\mu$ [116]) shows a trend towards a higher variation in context implicating a higher likelihood that the rule also contains a

response change. Conversely, it shows the opposite for the source. This indicates a stronger relationship between the context and response than the source, especially in the case of opinion shifting, where the effect is far more pronounced.

5.1.4 Conclusion

While studies that focus on mining data from large data sets such as those collected by social networking websites can present many interesting and unbiased findings on the nature of how users interact with data online, they are limited in the types of data that can be collected in this way. More detailed questions relating to exactly when users reach conclusions, what pushes them in one direction or another, and how staunchly they stick to the opinions they form are difficult to answer using only these types of data collection. In order solve this issue, this section presents the use of large scale surveys to crowd source a solution, asking individuals in a more direct manner how they interact with media online. From these surveys, not only can the basic thresholds of individual opinion formation begin to be understood [102], [103], but insight into what kinds of popular trends various groups of people exhibit can be gained by mining the survey sets for interesting features and patterns.

The most predominant feature discovered in this way is the high degree of consistency within mined itemsets. Users tend to behave more similarly to each other when describing the pieces of media they feel are equivalent, and thus is far easier to pick up on the aspects of media that make them similar than what makes them different. Additionally, the frequent itemsets have a lower average threshold value than the thresholds of the populations as a whole, meaning that users are far more consistent when describing media that they feel is more convincing, and thus it is easier to group people based on what they are partial to rather than what they dislike. Further, when looking at rules that do contain response changes, some trends in the parameter values become apparent. For instance, media types such as messages and images have very similar swaying power over individuals, while others such as messages and videos tend to be far more different. Additionally, the context (controversy level) of the media appears to be more important in predicting response changes than other parameters, as the variation in the context values within rules is far higher when that rule also contains a response change.

These conclusions are a start towards a deeper understanding of the size and behavior of different groups of individuals, but further analysis is necessary. First, the surveys con-

ducted recorded basic demographic data, social media consumption statistics, and favorite social media websites and news sources for each participant. While not studied in depth in this work, this information has potential for future examination into the links between demographic groups, popular social media communities, and how individuals respond to and consume data. The results presented here could be extended via additional surveys to include more study groups that are asked about opinion shifting as well as formation, in particular in the case of a fixed context and varying source. Further studies could also extend into the differences even among single media types could be valuable in differentiating how users respond to videos or articles of different lengths for example.

Still, as information and opinions spread through societies at ever greater rates, the subject of how social influence occurs becomes increasingly centered around the media with which people interact online. Greater insight into the how different groups behave and how large they are in the first place is necessary in order to understand and model these processes, informing on the stochastic techniques provided in Chaps. 3 and 4. Not only can this work be directly applied to some of the models already presented, but it presents the opportunity to extend that work to account for different communication mediums, personality types, and usage patterns. The ability to create heterogeneous populations within stochastic models represents a large step towards creating more accurate models, as old techniques of utilizing large populations of identical individuals lack the depth to describe the social effects of human diversity.

5.2 Mobile Phone Network

5.2.1 Motivation and Related Work

Underneath each stochastic simulation is a network structure that connects the nodes and determines the pathways through which information can flow. Attempting to understand how different structural patterns influence the dynamics of the system is a field in itself, constantly under as much if not more scrutiny than the rules of the simulations. The details of how different structures influence simulations is well beyond the scope of this work, but their effects still cannot be ignored entirely. For simplicity, in the preceding chapters complex network structures were simulated using synthetic networks such as random graphs and lattices, and indeed some synthetic networks are often a good substitute and capture many of the details present in social networks. Still, it is often necessary to go back to the source,

both to understand how close the synthetic graphs are to real systems and to provide settings for simulations in their own right. For this reason, opportunities to examine real social networks are always valuable, providing yet another avenue through which simulations can bridge the gap towards describing human behavior.

Call Detail Record (CDR) data sets, created from cell phone logs of large groups of people, are rapidly becoming popular for these purposes thanks to the the large amount of detailed data they provide [117]. These data sets typically include both basic information about the users (age, sex, location) as well as event records for calls and text messages that contain information such time, location, and direction. This information provides a bird's-eye view of human interactions without the self reporting bias inherent to many other behavioral studies. It also enables researchers to work with otherwise prohibitively large sample sizes and to study delicate relationships between various aspects of the users behavior.

This section analyzes one such data set, focusing on its network properties and presenting an overview of the underlying social network that can be obtained from the data. To this end, various methods for building the network are examined in an attempt to mitigate the noise inherent within the cell phone records and deal with other issues identified in prior work such as the definitions of links, communities, reciprocity and data types suitable for the study. These issues are exacerbated in this application by the large amounts of noise and potential biases inherent to various schemes [118], [119], [120], [121]. To remedy this, the methods presented here build the network in a way that accounts for many of these various pitfalls and provides insights into the balance between noise reduction and loss of information within the network building schemes. Then, by combining the features extracted from the built social network with the raw usage and geographic features, ties between the mobile phone use and general behavior are investigated, comparing the social and geographic communities individuals belong to.

This sort of study is not unique; early work on CDR data sets relied upon individual location data to investigate the relationship between the individual's distance and basic social properties such as their likelihood to interact [122]. Since then, the idea of comparing the geographic location of individuals to phone usage and interactions has expanded to utilize higher level statistics. Most recently various other details of the users, including their social groups and socio-economic status are also considered [123], [124]. Many of these studies focused on predicting the relative socio-economic status of geographic regions using factors

such as the total volume of calls. Others still have concentrated on predictions for individual users via features such as personal mobility and social network centrality [125], [126], [127], [128]. In this work, many of these techniques are applied to this new data set, attempting to describe a new setting for future work and lay the groundwork for better social simulations.

5.2.2 Description of Data

The CDR data set used here includes details about the call histories of $500,000$ clients of a cell phone company over a three month period. The data set itself contains information about the users in the form of basic demographic data (age, home district, gender, and default status at the end of the three month period) as well as usage information based on how the clients used the cell network in that time (frequency and duration of calls, messages, and movement records based on frequently used cell towers).

This data provides a large amount of basic information to draw upon for the network, and provides a setting to understand social behavior on a very detailed level. First, the data allows a new graph of the relationships to be built and used for future studies moving forwards. Further, analysis of this data provides a better idea of how future networks should be build to simulate this structure, and usage information informs on how people interact and behave to create more accurate dynamics models.

5.2.3 Results

Understanding the network properties of the data set begins with studying how each node fits into the overall scheme of the social network formed between them. The consequent node-level properties including network location and contribution provide the means to establish a high level description of how embedded each node is in the network. This information allows us to not only understand opinion spread more easily, it provides insight into how people interact with the technology each other. This information is vital to any model that contains node removal, as how deeply embedded a node is in the overall structure is crucial to understanding how likely they are leave (and the damage they will cause when they do so).

Unfortunately, while building social networks out of CDR data sets is a common path in analyzing the relationships of the users the information does not come without issues [117]. As mentioned above the high level of detail within these data sets comes with a large amount

of noise, making the initial process of creating the network difficult. Quantifying what level of communication between individuals indicates a connection between them is a challenge, and methods are often chosen to highlight a specific aspect of the network since more general results are not feasible. This issue is made worse considering the bias introduced by emphasizing certain types or patterns of communication, as generational and cultural divides are present in patterns of phone use [119]. Some attempts at a more general solution to the problem include requirements of reciprocity and certain levels of activity before a link is drawn in order to lower the noise of the system. Yet these solutions lose much of the directional and fine details of the system [118]. Others have suggested statistical methods to detect and remove links that are more likely to be random, offering a useful but costly strategy for creating a more reasonable system [121]. Finally, there is ambiguity about the representation of the graph as there are reasonable arguments for the use or disuse of weights and directional edges to represent different facets of the relationships between users.

In this section, many of these questions are answered by defining a relatively narrowly focus of building a social map for how information flows between users. To this end, only directed graphs are used in order to preserve the imbalances that tend to arise even among reciprocal relationships [129]. Further, the activity frequency between individuals is used to define edges using both a weighted and unweighted scheme as appropriate for different metrics. First, to investigate the general structure of the network, a frequency cutoff requirement is used to define an unweighted edge and identify relationships between individuals. By examining the effect this cutoff has on the structure, the cutoff at which the system is stable or best represents a true friendship network between the individuals can be determined. Next, switching to a weighted representation allows for a more general view of the network with greater detail about the closeness of individuals. For this purpose, edge weight is defined as the frequency of communications between individuals, preserving many of the benefits of the unweighted cutoff scheme while providing a more complete view of the data. While this scheme is inherently noisier than the alternate schemes, it also provides a better description of relationship strength.

5.2.3.1 Unweighted Network with Frequency Cutoff

First, by analyzing a weight cutoff for edges a more strict structural view of the true friendship network can be obtained. This scheme also addresses the need for noise reduction

by pruning down the network and shows how robust the network is to increasingly strict friendship requirements. The decay of the network for high values of the edge cutoff can be seen in Fig. 5.10(a), where the giant component of the network shrinks exponentially with an increased cutoff and is generally very sensitive to such an increase. Without a cutoff, the giant component of the network contains 99.1% of the nodes, but drops to only half of the network when the communication cutoff reaches 30. A similar effect can be seen in Fig. 5.10(b), where the total number of edges in the social network is shown to decay as a power law with increased cutoff. This power-law decay of edges implies that the scheme is useful for removing noisy, low frequency communications while leaving intact the dense communications which are more representative of strong social ties.

It is, however, difficult to establish the proper cutoff value for a given situation. Much of this difficulty comes from the fact that so many of the overall edges in the network are low frequency, and thus in this regime small changes in the cutoff can have drastic effects. In fact, Fig. 5.10(b) shows that less than half of the original edges remain when a cutoff of only four is established. Further, though the network tends to stabilize at high cutoff values, the loss of edges means that it is no longer highly connected. Instead, it shows a rough sketch of the community structure as raising the cutoff breaks the very dense graph into a few tightly tied communities that are resistant to the higher cutoffs. This effect is seen in Fig. 5.10(c), where the percentage of isolated nodes outside of the giant component *decreases* with increased cutoff. This somewhat counter-intuitive results is due to small communities being separated from the giant component but remaining intra-connected and thus creating stable communities of their own. This process, of course, yields only a crude approximation of the community structure because the high cutoffs lower the connectivity within the communities as well in addition to separating them from the giant component. This trade-off becomes apparent at cutoffs larger than 43, as the the number of isolated nodes increases and even the tightly bound communities unravel. A more direct and robust detection of the network's community structure is discussed in Sec. 5.2.3.3.

Additionally, the communication cutoff has a significant impact on the overall degree distribution of the network as seen in Fig. 5.11. As expected [120], [130], for all values of the cutoff the degree distribution shows a power-law tail. The rate at which this tail decays, however, changes as the edges are removed. Estimating the exact value of the power-law exponent is extremely difficult due to both low and high degree (k) saturation effects that

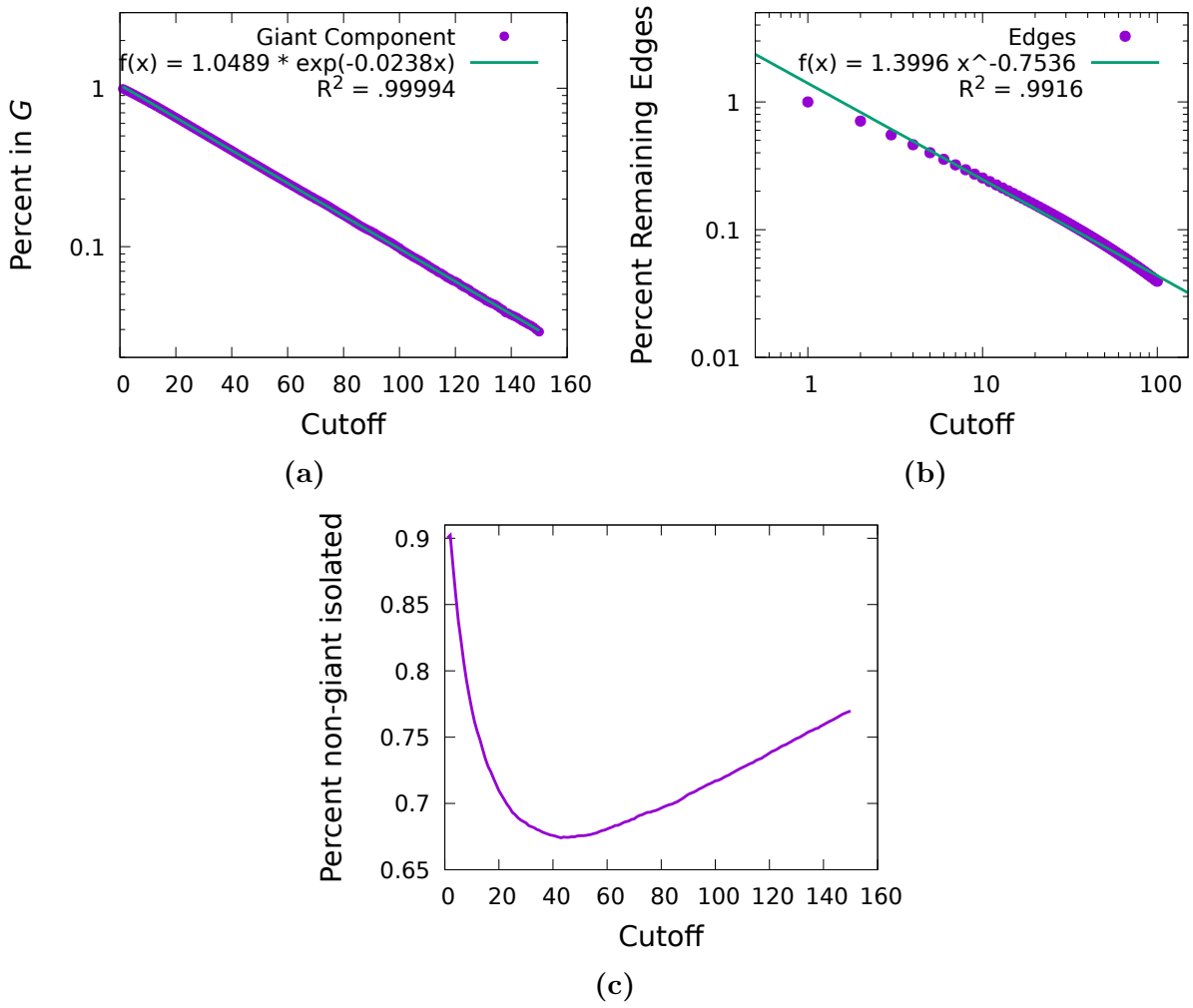


Fig. 5.10: **(a) The giant component of the social network decays exponentially ($\lambda = 0.0238$) with increased minimum number of communications required for an edge to be drawn. (b) The number edges in the network also decays rapidly with increased cutoff, closely fitting a power law with $\gamma = 0.7536$. (c) The percentage of non-giant component nodes that are isolated for a given cutoff. The isolated nodes reach a minimum at a cutoff of 43.**

muddy the data [130]. In the high k regime, this is simply due to the rare occurrence of outlying nodes; the discrete nature of degree binning makes it difficult to show the true probability for extreme values, as seen in Fig. 5.11(a) where the tails tend to flatten out and lose resolution. This problem is easily remedied by performing logarithmic binning or

studying the complementary cumulative distribution function (CCDF), defined as

$$P(k) = \sum_{q=k+1}^{\infty} p(q) \tag{5.5}$$

where $p(q)$ is the PDF. This solution can be seen in Fig. 5.11(b), where the distribution preserves its shape for much longer before losing accuracy due to high-k effects, while the scaling exponent can be shown to simply be

$$P(k) \sim k^{-\gamma+1} \tag{5.6}$$

if the PDF is of the form

$$p(k) \sim k^{-\gamma} \tag{5.7}$$

The low-k saturation effects are more difficult to account for, however, as there is no universally k_{min} for which the power-law scaling should begin. Instead, it is common to iteratively test every value of k_{min} over an estimated range to find the best possible fit [131], [132], [133]. Using these methods, the γ values for different cutoffs are estimated and shown in Fig. 5.11(c). For higher cutoff values, many of the highest degree nodes lose the vast majority of their edges, as individuals with 1000 different contacts are highly unlikely to support each edge with a large number of events. Due to this, the higher cutoffs greatly reduce the maximum degree of the network and increase the power scaling exponent of the degree distribution. As with the prior results on increased cutoff, however, this is only true to a point. The exponent increase appears to saturate at higher values of the cutoff for which the impact of the edge requirement being more strict gets weaker and weaker. These properties allow the unweighted graph to be used with low cutoffs to define various basic network features such as the general in-degree and out-degree of each node, while higher cuff values can be used to reduce noise and to find the stable number of strong contacts a given node has. In this way, the unweighted scheme is ideal for situations of finding common interactions and tightly bound communities, as it quickly and easily provides a sketch of this structure. The large amount of nodes that are lost from the giant component, however, make it difficult to use this sort of network as a setting for opinion spread simulations. Since a large proportion of the nodes do not have access to the other nodes in the network, there is no reasonable method for creating consensus using standard modeling techniques such as

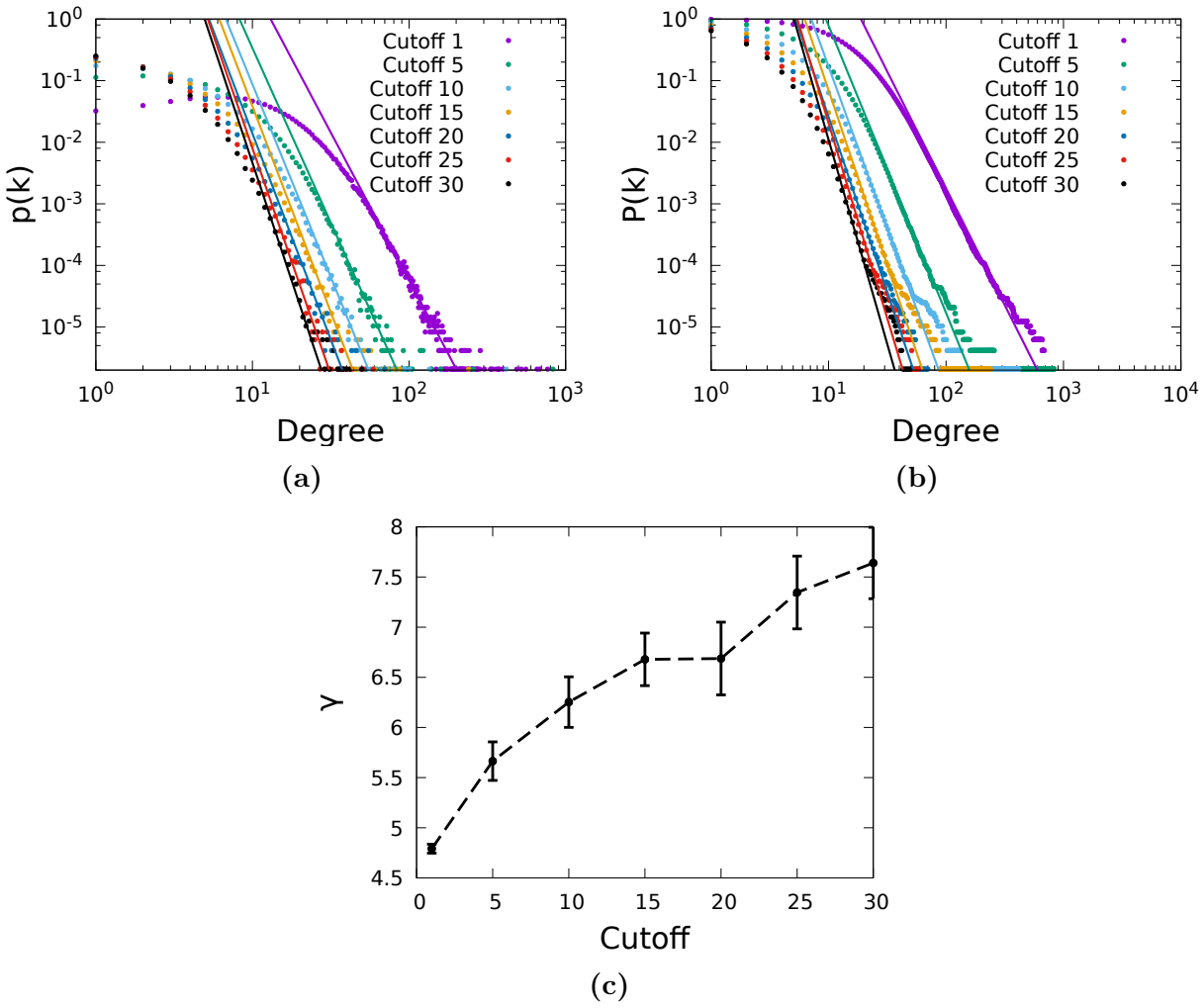


Fig. 5.11: (a) The out-degree PDF of the mobile phone network for various minimum event cutoffs for the edges, with fitted power-law tails. (b) The out-degree CCDF for the same data for improved visual fit. (c) The estimated power-law exponent of the PDF tail— for the various cutoff values to highlight the different scaling rates.

local thresholds or pairwise interactions. Additionally, while this method provides a good approximation of each node's local reach (which other nodes it is likely to spread information to easily), it loses all information on the maximum reach. Once split off from the larger components of the network, all information about the larger possible reach of the node is also lost.

5.2.3.2 Weighted Network Based on Event Frequency

For more in-depth network descriptions that describe the global reach of each node, more traditional methods of analyzing networks must be utilized to avoid the loss of information inherent in using the above scheme. For instance, many metrics that use distance-based calculations (diameter, centrality, some forms of community detection) are more accurate for social application when they take into account the reduced 'distance' between two heavily connected nodes. For these purposes, a basic weighted edge scheme can be employed such that each edge between two nodes is assigned a weight equal to the total frequency of communications in that direction. Then, for all distance-based applications, this weight can define the normalized distance as

$$d_{ij} = w_{avg}/w_{i,j} \tag{5.8}$$

where w_{avg} is the average weight of all connections in the network and $w_{i,j}$ is the weight of the connection between the source (i) and the target (j) [134]. The higher noise in many of these calculations is not only mitigated by the edge weights, but also by the nature of the distance measure. For the calculations such as finding the shortest paths in the network, it is highly unlikely that a short path will run through an extremely sparse communication given this distancing scheme.

Further, to aid in interpreting this analysis, the overall network can be rewired by swapping the edge destinations to create a pseudo-random weighted graph that maintains the in and out-degree structure of the original network. With swapping, the analysis can be expanded to get an idea of the density of the network and individual positioning of the nodes. For these purposes, the network can first be examined via the harmonic closeness centrality of each node (closeness centrality adapted to non-connected graphs) [135], [136]. This measure indicates how critical a node is to the various pathways through the network and is defined as $C_H(i) = \frac{1}{N-1}\sum_{j\neq i}\frac{1}{l_{ij}}$, where N is the total number of nodes and l_{ij} is the weighted shortest path between nodes i and j. Using this measure shows that the network has a surprising lack of crucial hubs, as throughout the network it is generally fairly high and evenly distributed. The average harmonic centrality is $C_{avg}^{cell} = 4.61$ with a standard deviation of $C_{std}^{cell} = 1.84$, both higher than those for the randomly rewired graph which yields $C_{avg}^{rand} = 4.11$ and $C_{std}^{rand} = 1.20$. Interestingly, the diameter and average shortest path length of the cell network (at $D^{cell} = 6.24$ and $\langle l\rangle^{cell} = .311$, respectively) are also slightly

higher than the corresponding values ($D^{rand} = 4.35$ and $\langle l \rangle^{rand} = .30$) for the randomly rewired network. Of course, these values in both cases are significantly lower than those for non-weighted networks in which a path shorter than one is not possible. The definition and scaling of distance in this network allows the shortest paths to seek out the extremely high density (low distance) connections that will provide multiple jumps at extremely low cost, even with the normalization factor imposed. This further emphasizes the role of the strong ties rather over the noise in the network, indicating that despite the fact that the vast majority of edges are low frequency, the short paths for information to flow through are still extremely efficient and even distant nodes have high frequency channels between them.

5.2.3.3 Community Detection and Geographical Districts

Using the above weighted directed graph further allows the network to be analyzed for social communities based on the communications between individuals. The method for community detection has a great impact on the resultant groups, and community detection in social networks is a vast field of study that contains many different methods to analyze different aspects of the network structure. The effects of various detection strategies are beyond the scope of this work, however, so instead only the GANXiS(SLPA) algorithm is used due to its focus on the detecting even disjointed and overlapping communities in order to fully encapsulate the social structure of the network. The GANXiS(SLPA) algorithm is a speaker-listener label propagation technique that allows membership in multiple communities at once and is optimized for extremely large networks [137], [138].

This method was able to identify a set of over 6450 social communities, allowing for comparison with the 231 geographic districts that the users report living in (information already contained within the CDR data). While the groups are different sizes (the number of individuals in each geographic district ranges from 1 to 63491, while the number of people in the detected social communities ranges from 2 to just 741), the average social community is still large enough at 74.65 individuals to compare with the geographic districts. In particular, it may be intuitively assumes that the social communities would be highly influenced by the geographic district of their members, as people would be likely to engage socially with individuals that are geographically nearby. Instead, Fig. 5.12 shows that on average only 41% of each community comes from the same district, and in fact even the top five districts only account for 78% of each community's makeup. The diversity of the geographic locations

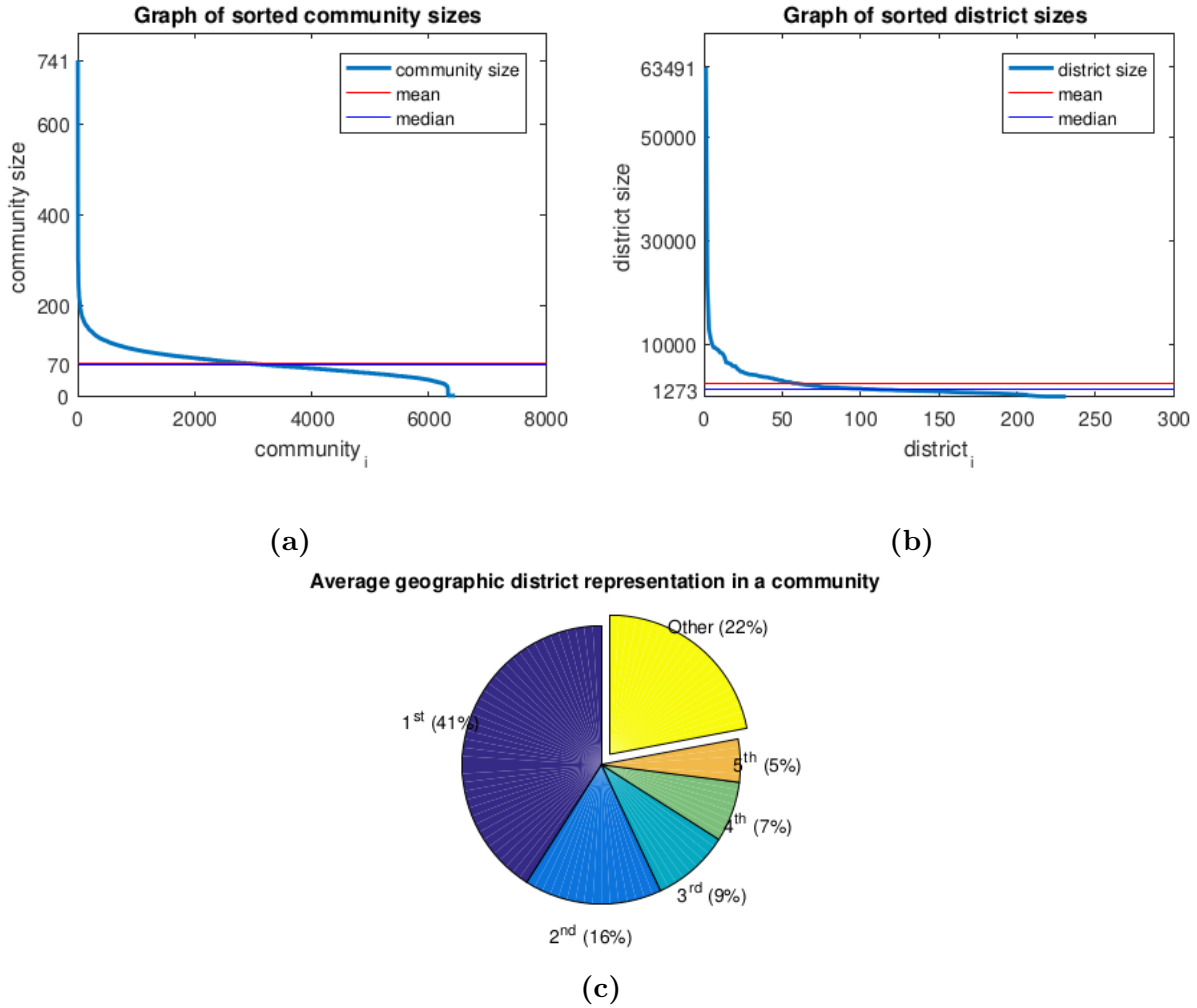


Fig. 5.12: (a) Community sizes sorted in descending order to highlight the tail. The average social community size is 75 with a median of 70. (b) Geographic district sizes in descending order. Districts have an average size of 2, 380 and median of 1273. (c) average proportion of users from a community that belong to the same geographic district - from most represented geographic district by users (1^{st}) to the 5^{th} most represented.

within social groups is especially surprising given the generally small size of the social groups compared to the geographic districts. This result indicates that future models should exercise extreme caution when designing their underlying network to clearly identify whether the process is more likely to spread via physical proximity or social interaction, as it should not be assumed that the two overlap.

5.2.3.4 Wait Times Between Events

Finally, to use this information to directly inform on the models presented in the prior chapters, we can compare the wait times between speaking events for each individual within the mobile network. As expected, the frequency of these events follows a power-law tail as discussed in Chap. 4, but the specifics of the distribution are not trivial to determine. Like many empirical studies into the waiting-time distributions between speaking events, the high noise environment and many confounding factors limit the ability to find a clean fit. Working and sleeping patterns form odd spikes of common wait times, and the question of separating calls and text messages becomes all the more important to determining accurate distributions. The question of finding exact waiting-time distributions to describe behavior has been covered in prior work [69], so this level of complex analysis is not repeated here. Some key values that describe the activity of nodes is valuable to note, however, to give a frame of reference for the speed at which consensus is achieved on prior models in terms of real world time. In this system, the mean time between speaking events is $\langle \Delta t \rangle = 2.802$, whereas for the systems described in Chap. 4 it is held to $\langle \Delta t \rangle = 1$. In terms of actual time scales, and using the definition of a system time step being N micro transactions, there are 6.28 system time steps per day (i.e. each system time step is 3.82 hours). This emphasizes the rapid convergence towards consensus for many of our systems, as systems with 1000 nodes would routinely reach consensus in $20 - 30$ system time steps which correspond to less than a weeks worth of interactions for nodes following the behavioral patterns from this cell phone data.

5.2.4 Conclusion

Understanding the underlying social network of cell phone usage data presents an opportunity for an unbiased, quantitative view at social interaction on a large scale. The resultant complex system is unique due to the individual influences of their members, making for a valuable contribution towards future simulations. Both as a general study of human behavior and as a setting for future simulations, the information contained in such data sets represents a path towards greater realism in the realm of social influence studies and offers another opportunity to add to the overall goal of this work. Due to the high complexity and depth of the field of empirical social network analysis, this section represents only a brief overview with some basic conclusions to motivate and inform on future work, focusing

largely on how different methods of building the network affect its final structure.

For this purpose, two main methods of building the network are studied. The first method uses an event frequency cutoff requirement to define the edges of the system. This method, while effective in removing noise from the system, comes with the drawback of removing information critical to the calculation of some common network analysis metrics. In fact, it is shown that while this method is extremely effective at removing the majority of the noise in the system, it also has far reaching effects on the underlying structure. The edge frequency cutoff tends to remove so many edges from the system that it fairly rapidly degrades the giant component of the system, splitting off many smaller tightly bound communities from the rest of the population. Further, it increases the exponent on the power-law scaling factor of the degree distribution, removing many hubs from the system. As such, this method works well to estimate the basic friendship structure of the system, preserving tightly bound communities, but is not well suited for any application that requires information traveling throughout the system since there is no way for many of these communities to interact.

For cases where detailed considerations of the node's place and reach within the global network are necessary, a second method is proposed that uses a simple weighted scheme to define the edges. This method regains the required information for more complex calculations despite the high noise environment, allowing for a complete view of the network to be presented. Using this edge scheme, it is shown that the network is actually surprisingly wide, with a significantly larger diameter than a randomly rewired version of the same network. Further, it also has a longer average shortest path between nodes and a larger and less defined average centrality. In general, this means that there are more "fringe" nodes in this network than would be expected, and despite the tight community structure there are nodes that are very well isolated from each other that would struggle to share information across the network, even in its most connected state.

Finally, the community structure of the network is studied from the viewpoint of the geographic districts that the nodes live in, showing that the social communities and the geographic districts do not overlap as much as might be expected. Intuitively it would seem that individuals would mostly interact with others that they live around, but the work here shows that on average less than half of every social community lives in the same district. Even with the long range nature of mobile phone communications this is surprising, as it is often assumed that most communication comes from proximity, and indeed prior work has

shown that the probability of interaction between nodes drops a square root of the distance between them [118]. Still, this result shows that tight connections and social groups have a tendency to transcend some of these geographic limitations.

Overall, this analysis gives only a brief overview of what can be gleaned from CDR data sets, yet still makes it clear why their popularity as the subject of social research is on the rise. Since the drawbacks of the high noise environment can be mitigated by clever network building schemes, these systems are prime candidates for use as empirical complex networks in future simulations on social influence. Indeed, there is enough information contained for the data to be more than just the setting of social simulations; the phone records presented here offer many insights on how and when individuals tend to interact, providing critical information to help build model dynamics as well.

CHAPTER 6
CONCLUSIONS

The process of social spreading is incredibly complex, influenced by a variety of factors that are often not fully understood. The volatile tipping point behavior exhibited by social and political systems, however, makes it crucial to build a knowledge base that helps explain these effects. A full understanding of the situations and behaviors that lead to these large scale changes is important, and can provide strategies to either prevent or facilitate consensus depending on the desired outcome. In any case, understanding how the process works allows for further control and better preparation when critical situations arise. The applications of this work have a very wide reach, from increased convenience such as driving more efficient marketing to societal safety measures such as predicting and preventing riots and other large scale unrest from reaching dangerous levels. Fortunately, explanations and predictions for these situations are becoming easier to obtain as people increasingly use online communications to spread their ideas. With many of the these discussions occurring over messaging platforms, the resulting dialogue is discretized and the patterns of communication are more easily accessed and studied. The same communication technology, however, also increases the need for these understandings as the internet and large scale mobile phone networks make for denser connectivity across the world and grow the effect sizes of movements. As such, tipping phenomenon are becoming more common and easier to create as critical masses of individuals can be unified much more simply than has previously been possible.

Despite the field of modeling these effects having been around for a long time, many models still in use have their roots in mathematical simplicity rather than a focus on social realism. As a result many of these original models are very well understood, but contain glaring weaknesses when applied to more specific real world scenarios. Improved modern modeling focuses on fixing some of these issues, keeping in mind that a realistic model does not have to be perfect in reproducing all facets of behavior. Instead, it is worthwhile to simply understanding the scale of the effects that each modification creates, then tailor the model to fit the standards of a given situation.

To this end, this work presents two modifications to the basic naming game model that address areas where the original formulation lacks realism. First, the *opinion inertia*

model is presented. This model is a off-shoot of the popular *committed agent* class of social influence models, treating opinions as a social contagion and requiring individuals to receive a certain "dosage" before adopting a new opinion. In particular *opinion inertia* focuses on opinion tractability, a situation where the competing opinions have different sticking power in individuals. In effect, this model is similar to a partial commitment model [54], [55], [82] where each node recruited to an opinion adopts the level of commitment of its recruiter, creating a large memory dependence and subsequent "momentum" towards consensus. As such, it retains many of the prominent features of *committed agent* models, including the tipping point behavior resulting from critical populations for a phase change in the system. In fact, due to the opinion-based commitment present here, the *opinion inertia* model is able to reach incredibly small critical populations with differences in the inertias between the competing opinions. Further, on a more theoretical note, the model is able to mimic the behavior of both memory-less (voter) and intermediate state (naming game) models. It shows the characteristic diffusion based coarsening with no surface tension in the special case where it reduces to the voter mechanics, but given any higher level of inertia to even one of the two opinions it regains the surface tension and spreading effects of standard models with intermediate states. Finally, though high levels of inertia can make direct simulation slow, the system can be approximated analytically by treating it as a steady state to obtain accurate thresholds for arbitrary inertia values. As a whole, this model extends the prior work on the subject into an opinion centric focus that emphasizes the importance of the way different ideas are presented, showing that an opinion that is framed more favorably is able to spread extremely quickly and heavily disrupt the balance of the system.

The second model considered is a standard naming game model where the basic assumption of how speaking nodes behave is modified. For this model, in order to fit a more realistic speaking pattern, certain nodes are given non-exponential waiting times between speaking events to account for different temporal communication dynamics. In effect, the "burstiness" (propensity to speak rapidly and then go silent), is changed. In particular, to keep with the theme of investigating tipping points and opinion competition, this model uses groups of nodes with different speaking patterns competing to propagate different opinions. It is shown that under otherwise balanced initial conditions the group with the burstier speaking patterns tends to dominate. This effect breaks the symmetry of the system and guarantees consensus in the infinite system size limit. The driving force behind this effect

is the control that bursty nodes are able to exert over the initial stages of the simulation; by propagating their opinion rapidly in the early portion of the simulation it doesn't matter if they silent afterwards, they have likely already altered enough opposing nodes to create an eventual consensus anyway. Further, this bias remains present (albeit in a weaker state) in situations such as the voter model, which are all but guaranteed to reach the long time limit of the system and don't exhibit the momentum towards consensus that the naming game has. Finally, it is shown that due to the early-time dominance playing such a crucial role in these sorts of simulations, the head of the wait-time distribution alone is sufficient to accurately describe the behavior of the system. For this purpose, the critical populations and relative opinion propagation rates are described analytically in terms of only the head densities of their related wait-time distributions. In a practical sense, these results suggest a clear strategy for political and marketing campaigns, where strong early thrusts would be expected to correspond to a high rate of eventual consensus. This is especially true when combined with the use of committed agents, where it is shown that alternate speaking patterns for the committed individuals can serve only to lower the critical population, and will not raise it even if those nodes are less bursty.

Finally, to tie these theoretical discussions back to the central focus of increased realism, two major empirical data sets are presented in order to inform on the computational models. The first data set shows the results of a survey on individual thresholds for forming opinions in the face of various pieces of information. This data is investigated in terms of its clear relation to the *waning commitment* and *opinion inertia* models presented, seeking to describe how groups of people behave relative to their peers. By mining the data for frequent response patterns to identify various groups of individuals, it is shown that by far the most consistent groups are those that are actually *more* easily swayed than their peers. The more stubborn individuals (analogous to some types of committed nodes) tend to be far more varied in their responses, a stark difference to the classical formulation of building a system with a small group of committed individuals attempting to sway a larger population. Further, it is shown that the context of information and media type it is presented in play a role in the resulting response. Certain media types (such as messages and images) are relatively equivalent, while others (messages and videos) tend to elicit response changes. Similarly, different contexts or controversy levels of the information correlate to differing response values, and high controversy topics tend to elicit more consistent responses. This information

presents some major differences that are currently unaccounted for in most computational modeling considerations, as media type and context level are not considered at all in most of the basic modeling schemes. The results, however, suggest that these sorts of information play a sufficiently large role in media consumption that it would be worthwhile to study their effects on spreading models.

Additionally,d a brief overview of a mobile phone network is presented in order to give some context on the structural aspects of social simulations. The system is examined for the purposes of creating an empirical network that can be used as the setting for future modeling, but also presents the opportunity for far more in depth research. Unfortunately, the original data set is highly noisy, and thus is not appropriate for use without some modification. This issue is shown to be largely eliminated by using a frequency cutoff for the edges, but it is also shown that doing so causes large amounts of damage to the overall structure of network and shrinks the giant component rapidly. Thus, this sort of edge scheme is appropriate for obtaining a general view of the basic community structure of the network but not for most spreading scenarios. For spreading, it is more appropriate to utilize a weighted edge scheme, where the network is kept relatively whole and the noise issues can be mitigated by utilizing a shortened "distance" between nodes that have strong connections. Doing so creates a network that has sufficient information to describe the various efficient pathways that information can flow through the network, and provides a setting that would more easily allow for full opinion consensus in spreading models.

In all, the results presented in this work represent a step towards bringing classical stochastic models of opinion spread closer to real human communication. Many of the changes proposed have been previously overlooked due to the mathematical and computational consequences of introducing memory into the system and making the processes non-Markovian, but it is shown that by carefully approximating around these complications the models can still be described analytically. Further, the benefits of adding these sorts of effects to the more basic simulations provides an opportunity to study fully how they change the large scale dynamics of the system, improving strategies and general understanding of common systems that are not conveniently described by the simpler base cases. The contributions of such new models help to address and understand issues before they arise, allowing for better action and more rapid dissemination of information and innovations where appropriate. Further, it is shown that there are far more aspects of information consumption and

personal interaction that are not yet incorporated, indicating the need for continued work in the field to stay up to date. As technology advances, so to do the ways in which people interact and share what they know, constantly creating new aspects of communication that must be understood to accurately describe information flow. As such, the contributions presented here provide not only an extension of prior work to advance the current knowledge base, but also create a foundation for future work to build off of and address newly emerging challenges.

6.1 Future Work

Naturally, modeling such complex processes as opinion spread and large scale human interaction always leaves avenues for future growth. The models created may benefit from further tailoring to be useful in predictive analysis of specific situations, and the relative importance of various behaviors should be weighed to determine where computational effort is best allocated. Additionally, as technology evolves the nature of communication changes, and models that were appropriate for describing spreading processes in the past become outdated in the age of evolving social media and online communication. In particular, new technology allows for many different types of media to be shared using many different methods. Some pieces of information are shared via simple pairwise interactions, but many others are broadcast towards large groups as various social media platforms give users the opportunities to share content more generally. Altering communication dynamics to account for these processes is difficult, but provides a more accurate snapshot of the behavior of online communities as opposed to the personal interactions generally described here. Additionally, different types of media play a role in how heavily people weigh information, making the classical practice of treating each communication identically increasingly obsolete. Finally, the need for heterogeneity can be extended to the individuals themselves, creating more diverse populations that can accurately capture how opinions move through groups with varying levels of stubbornness and incorporating different individual demographics, age, and behaviors. This idea has been touched upon in studies of the *waning commitment* model [55], but a more detailed consideration within the contexts of how individuals interact with different types of information is still needed. Sec. 5.1 shows how the amount of information necessary to form and shift opinions can vary based on many of these factors that are currently unaccounted for in stochastic simulations, indicating the need for a better understanding of how

these alterations may affect the overall consensus. In general, including such heterogeneity in stochastic systems causes a great deal of mathematical difficulty, making it understandable why these more complex systems have been avoided. Still, a fuller understanding of individual behavior is valuable enough to push beyond these limitations. Without further studies bringing these models towards increased realism, the applications and predictive power of the field will continue to be limited to more abstract and general conclusions.

REFERENCES

[1] M. Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference*. New York, NY: Little Brown, 2006.

[2] A. Grübler, "Time for a change: on the patterns of diffusion of innovation," *Daedalus*, vol. 125, no. 3, pp. 19–42, 1996. doi: 10.2307/20027369

[3] B. Tadić, V. Gligorijević, M. Mitrović, and M. Šuvakov, "Co-evolutionary mechanisms of emotional bursts in online social dynamics and networks," *Entropy*, vol. 15, no. 12, pp. 5084–5120, Nov. 2013. doi: 10.3390/e15125084

[4] A. Morris, *Origins of the Civil Rights Movement*, 1st ed. New York, NY: Simon and Schuster, 1984.

[5] E. Dubois, *Feminism and Suffrage: The Emergence of an Independent Women's Movement in America, 1848-1869*. Ithaca, NY: Cornell University Presss, 1999.

[6] W. van de Donk, B. Loader, P. Nixon, and D. Rucht, *Cyberprotest: New Media, Citizens and Social Movements*. London, UK: Routledge, 2004.

[7] P. Gerbaudo, *Tweets and the Streets : Social Media and Contemporary Activism*. London, UK: Pluto Press, 2012.

[8] H. H. Khondker, "Role of the New Media in the Arab Spring," *Globalizations*, vol. 8, no. 5, pp. 675–679, Oct. 2011. doi: 10.1080/14747731.2011.621287

[9] S. Harlow, "Social media and social movements: Facebook and an online Guatemalan justice movement that moved offline," *New Media & Soc.*, vol. 14, no. 2, pp. 225–243, Mar. 2012. doi: 10.1177/1461444811410408

[10] M. Pelling and K. Dill, "Disaster politics: tipping points for change in the adaptation of sociopolitical regimes," *Progress in Human Geography*, vol. 34, no. 1, pp. 21–37, Feb. 2010. doi: 10.1177/0309132509105004

[11] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proc. of the 13th Conf. on World Wide Web - WWW '04*. New York, New York, USA: ACM Press, 2004. doi: 10.1145/988672.988739 p. 491.

[12] D. J. Bartholomew, *Stochastic Models for Social Processes*, 2nd ed. Madison, WI: Wiley, 1973.

[13] A. Diekmann and P. Mitter, *Stochastic Modelling of Social Processes*, A. Diekmann and P. Mitter, Eds. Vienna, Austria: Academic Press, 1984.

[14] L. M. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chávez, "The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models," *Physica A*, vol. 364, pp. 513–536, May 2006. doi: 10.1016/J.PHYSA.2005.08.083

[15] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi, "Characterizing social cascades in Flickr," in *Proc. of the 1st Workshop on Online Social Networks, WOSP '08*. New York, New York, USA: ACM Press, 2008. doi: 10.1145/1397735.1397739 p. 13.

[16] R. Iyengar, C. Van den Bulte, and T. W. Valente, "Opinion leadership and social contagion in new product diffusion," *Marketing Sci.*, vol. 30, no. 2, pp. 195–212, Mar. 2011. doi: 10.1287/mksc.1100.0566

[17] S. Aral and D. Walker, "Creating social contagion through viral product design: a randomized trial of peer influence in networks," *Manage. Sci.*, vol. 57, no. 9, pp. 1623–1639, Sept. 2011. doi: 10.1287/mnsc.1110.1421

[18] A. M. El-Sayed, P. Scarborough, L. Seemann, and S. Galea, "Social network analysis and agent-based modeling in social epidemiology," *Epidemiologic Perspectives & Innovations*, vol. 9, no. 1, p. 1, Feb. 2012. doi: 10.1186/1742-5573-9-1

[19] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Rev. of Modern Physics*, vol. 81, no. 2, pp. 591–646, May 2009. doi: 10.1103/RevModPhys.81.591

[20] S. Galam, "Sociophysics: a review of Galam models," *Int. J. of Modern Physics C*, vol. 19, no. 03, pp. 409–440, Mar. 2008. doi: 10.1142/S0129183108012297

[21] M. Granovetter, "Threshold models of collective behavior," *Amer. J. of Sociology*, vol. 83, no. 6, pp. 1420–1443, May 1978. doi: 10.1086/226707

[22] T. M. Liggett, *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*, ser. Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer Verlag, 1999, vol. 324.

[23] L. Steels, "A self-organizing spatial vocabulary," *Artificial Life*, vol. 2, no. 3, pp. 319–332, Apr. 1995. doi: 10.1162/artl.1995.2.3.319

[24] T. C. Schelling, "Hockey helmets, concealed weapons, and daylight saving," *J. of Conflict Resolution*, vol. 17, no. 3, pp. 381–428, Sept. 1973. doi: 10.1177/002200277301700302

[25] M. Kitsak *et al.*, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, Nov. 2010. doi: 10.1038/nphys1746

[26] D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation," *J. of Consumer Res.*, vol. 34, no. 4, pp. 441–458, Dec. 2007. doi: 10.1086/518527

[27] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Threshold-limited spreading in social networks with multiple initiators," *Scientific Rep.*, vol. 3, no. 1, p. 2330, Dec. 2013. doi: 10.1038/srep02330

[28] P. D. Karampourniotis, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "The impact of heterogeneous thresholds on social contagion with multiple initiators," *PLoS ONE*, vol. 10, no. 11, p. e0143020, Nov. 2015. doi: 10.1371/journal.pone.0143020

[29] P. D. Karampourniotis, B. K. Szymanski, and G. Korniss, "Influence Maximization for Fixed Heterogeneous Thresholds," arXiv preprint, Mar. 2018. arXiv:1803.02961 [cs.SI].

[30] W. Pickering and C. Lim, "Solution of the multistate voter model and application to strong neutrals in the naming game," *Physical Rev. E*, vol. 93, no. 3, p. 032318, Mar. 2016. doi: 10.1103/PhysRevE.93.032318

[31] W. Zhang, C. Lim, S. Sreenivasan, J. Xie, B. K. Szymanski, and G. Korniss, "Social influencing and associated random walk models: asymptotic consensus times on the complete graph," *Chaos: An Interdisciplinary J. of Nonlinear Sci.*, vol. 21, no. 2, p. 025115, June 2011. doi: 10.1063/1.3598450

[32] I. Dornic, H. Chaté, J. Chave, and H. Hinrichsen, "Critical coarsening without surface tension: the universality class of the voter model," *Physical Rev. Lett.*, vol. 87, no. 4, p. 045701, July 2001. doi: 10.1103/PhysRevLett.87.045701

[33] C. Castellano, D. Vilone, and A. Vespignani, "Incomplete ordering of the voter model on small-world networks," *Europhysics Lett.*, vol. 63, no. 1, pp. 153–158, July 2003. doi: 10.1209/epl/i2003-00490-0

[34] F. Vazquez and V. M. Eguíluz, "Analytical solution of the voter model on uncorrelated networks," *New J. of Physics*, vol. 10, no. 6, p. 063011, June 2008. doi: 10.1088/1367-2630/10/6/063011

[35] R. A. Blythe, "Ordering in voter models on networks: exact reduction to a single-coordinate diffusion," *J. of Physics A*, vol. 43, no. 38, p. 385003, Sept. 2010. doi: 10.1088/1751-8113/43/38/385003

[36] W. S.-Y. Wang and J. W. Minett, "The invasion of language: emergence, change and death," *Trends in Ecology & Evolution*, vol. 20, no. 5, pp. 263–269, May 2005. doi: 10.1016/J.TREE.2005.03.001

[37] V. Loreto, A. Baronchelli, A. Mukherjee, A. Puglisi, and F. Tria, "Statistical physics of language dynamics," *J. of Statistical Mechanics*, vol. 2011, no. 04, p. P04006, Apr. 2011. doi: 10.1088/1742-5468/2011/04/P04006

[38] A. Baronchelli, V. Loreto, and L. Steels, "In-depth analysis of the naming game dynamics: the homogeneous mixing case," *Int. J. of Modern Physics C*, vol. 19, no. 05, pp. 785–812, May 2008. doi: 10.1142/S0129183108012522

[39] A. Baronchelli, M. Felici, V. Loreto, E. Caglioti, and L. Steels, "Sharp transition towards shared vocabularies in multi-agent systems," *J. of Statistical Mechanics: Theory and Experiment*, vol. 2006, no. 06, pp. P06 014–P06 014, June 2006. doi: 10.1088/1742-5468/2006/06/P06014

[40] A. Baronchelli, L. Dall'Asta, A. Barrat, and V. Loreto, "Nonequilibrium phase transition in negotiation dynamics," *Physical Rev. E*, vol. 76, no. 5, p. 051102, Nov. 2007. doi: 10.1103/PhysRevE.76.051102

[41] Q. Lu, G. Korniss, and B. K. Szymanski, "The naming game in social networks: community formation and consensus engineering," *J. of Econ. Interaction and Coordination*, vol. 4, no. 2, pp. 221–235, Nov. 2009. doi: 10.1007/s11403-009-0057-7

[42] A. Baronchelli, L. Dall'Asta, A. Barrat, and V. Loreto, "Topology-induced coarsening in language games," *Physical Rev. E*, vol. 73, no. 1, p. 015102, Jan. 2006. doi: 10.1103/PhysRevE.73.015102

[43] L. Dall'Asta, A. Baronchelli, A. Barrat, and V. Loreto, "Agreement dynamics on small-world networks," *Europhysics Lett.*, vol. 73, no. 6, pp. 969–975, Mar. 2006. doi: 10.1209/epl/i2005-10481-7

[44] A. Baronchelli, L. Dall'Asta, A. Barrat, and V. Loreto, "The role of topology on the dynamics of the Naming Game," *The Eur. Physical J. Special Topics*, vol. 143, no. 1, pp. 233–235, Apr. 2007. doi: 10.1140/epjst/e2007-00092-0

[45] Q. Lu, G. Korniss, and B. K. Szymanski, "Naming games in two-dimensional and small-world-connected random geometric networks," *Physical Rev. E*, vol. 77, no. 1, p. 016111, Jan. 2008. doi: 10.1103/PhysRevE.77.016111

[46] L. Dall'Asta, A. Baronchelli, A. Barrat, and V. Loreto, "Nonequilibrium dynamics of language games on complex networks," *Physical Rev. E*, vol. 74, no. 3, p. 036105, Sept. 2006. doi: 10.1103/PhysRevE.74.036105

[47] S. Galam and F. Jacobs, "The role of inflexible minorities in the breaking of democratic opinion dynamics," *Physica A*, vol. 381, pp. 366–376, July 2007. doi: 10.1016/J.PHYSA.2007.03.034

[48] A. Waagen, G. Verma, K. Chan, A. Swami, and R. D'Souza, "Effect of zealotry in high-dimensional opinion dynamics models," *Physical Rev. E*, vol. 91, no. 2, p. 022811, Feb. 2015. doi: 10.1103/PhysRevE.91.022811

[49] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski, "Social consensus through the influence of committed minorities," *Physical Rev. E*, vol. 84, no. 1, p. 011130, July 2011. doi: 10.1103/PhysRevE.84.011130

[50] J. Xie, J. Emenheiser, M. Kirby, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Evolution of opinions on social networks in the presence of competing committed groups," *PLoS ONE*, vol. 7, no. 3, p. e33215, Mar. 2012. doi: 10.1371/journal.pone.0033215

[51] A. M. Thompson, B. K. Szymanski, and C. C. Lim, "Propensity and stickiness in the naming game: tipping fractions of minorities," *Physical Rev. E*, vol. 90, no. 4, p. 042809, Oct. 2014. doi: 10.1103/PhysRevE.90.042809

[52] C. W. Wu, "Can stubbornness or gullibility lead to faster consensus? A study of various strategies for reaching consensus in a model of the naming game," in *2011 IEEE Int. Symp. of Circuits and Syst. (ISCAS)*. IEEE, May 2011. doi: 10.1109/ISCAS.2011.5938015 pp. 2111–2114.

[53] Y. Treitman, C. Lim, W. Zhang, and A. Thompson, "Naming game with greater stubbornness and unilateral zealots," in *2013 IEEE 2nd Network Science Workshop (NSW)*. IEEE, Apr. 2013. doi: 10.1109/NSW.2013.6609208 pp. 126–130.

[54] D. Galehouse, T. Nguyen, and S. Sreenivasan, "Impact of network connectivity and agent commitment on spread of opinions in social networks," in *AHFE 2014*, Kraków, Poland, 2014, pp. 2318–2329.

[55] X. Niu, C. Doyle, G. Korniss, and B. K. Szymanski, "The impact of variable commitment in the Naming Game on consensus formation." *Scientific Rep.*, vol. 7, p. 41750, Feb. 2017. doi: 10.1038/srep41750

[56] R. Nickerson, "Confirmation bias: a ubiquitous phenomenon in many guises," *Rev. of General Psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[57] P. Dodds and D. Watts, "A generalized model of social and biological contagion," *J. of Theoretical Biology*, vol. 232, no. 4, pp. 587–604, Feb. 2005. doi: 10.1016/J.JTBI.2004.09.006

[58] L. Dall'Asta and C. Castellano, "Effective surface-tension in the noise-reduced voter model," *Europhysics Lett.*, vol. 77, no. 6, p. 60005, Mar. 2007. doi: 10.1209/0295-5075/77/60005

[59] G. E. Briggs and J. B. Haldane, "A note on the kinetics of enzyme action." *The Biochemical J.*, vol. 19, no. 2, pp. 338–9, Jan. 1925. doi: 10.1042/BJ0190338

[60] L. Dall'Asta and T. Galla, "Algebraic coarsening in voter models with intermediate states," *J. of Physics A*, vol. 41, no. 43, p. 435003, Oct. 2008. doi: 10.1088/1751-8113/41/43/435003

[61] W. Zhang, C. C. Lim, G. Korniss, and B. K. Szymanski, "Opinion dynamics and influencing on random geometric graphs," *Scientific Rep.*, vol. 4, no. 1, p. 5568, May 2014. doi: 10.1038/srep05568

[62] C. Castellano, V. Loreto, A. Barrat, F. Cecconi, and D. Parisi, "Comparison of voter and Glauber ordering dynamics on networks," *Physical Rev. E*, vol. 71, no. 6, p. 066107, June 2005. doi: 10.1103/PhysRevE.71.066107

[63] A. Bray, "Theory of phase-ordering kinetics," *Advances in Physics*, vol. 43, no. 3, pp. 357–459, June 1994. doi: 10.1080/00018739400101505

[64] L. Frachebourg and P. L. Krapivsky, "Exact results for kinetics of catalytic reactions," *Physical Rev. E*, vol. 53, no. 4, pp. R3009–R3012, Apr. 1996. doi: 10.1103/PhysRevE.53.R3009

[65] J. W. Evans and T. R. Ray, "Kinetics of the monomer-monomer surface reaction model," *Physical Rev. E*, vol. 47, no. 2, pp. 1018–1025, Feb. 1993. doi: 10.1103/PhysRevE.47.1018

[66] A. Grabowski, "Opinion formation in a social network: the role of human activity," *Physica A*, vol. 388, no. 6, pp. 961–966, Mar. 2009. doi: 10.1016/J.PHYSA.2008.11.036

[67] S. Suri and D. J. Watts, "Cooperation and contagion in web-based, networked public goods experiments," *PLoS ONE*, vol. 6, no. 3, p. e16836, Mar 2011. doi: 10.1371/journal.pone.0016836

[68] S. Karlin and H. Taylor, *A Second Course in Stochastic Processes*. San Diego, CA: Academic Press, 1981.

[69] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A. L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *J. of Physics A*, vol. 41, no. 22, p. 224015, June 2008. doi: 10.1088/1751-8113/41/22/224015

[70] A. L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, May 2005. doi: 10.1038/nature03459

[71] A. Vázquez, J. G. Oliveira, Z. Dezsö, K. I. Goh, I. Kondor, and A. L. Barabási, "Modeling bursts and heavy tails in human dynamics," *Physical Rev. E*, vol. 73, no. 3, p. 036127, Mar. 2006. doi: 10.1103/PhysRevE.73.036127

[72] K. I. Goh and A. L. Barabási, "Burstiness and memory in complex systems," *Europhysics Lett.*, vol. 81, no. 4, p. 48002, Feb. 2008. doi: 10.1209/0295-5075/81/48002

[73] M. Karsai, K. Kaski, A. L. Barabási, and J. Kertész, "Universal features of correlated bursty behaviour," *Scientific Rep.*, vol. 2, no. 1, p. 397, Dec. 2012. doi: 10.1038/srep00397

[74] O. Artime, J. J. Ramasco, and M. San Miguel, "Dynamics on networks: competition of temporal and topological correlations," *Scientific Rep.*, vol. 7, Apr. 2017. doi: 10.1038/srep41627

[75] A. Vazquez, B. Rácz, A. Lukács, and A. L. Barabási, "Impact of non-poissonian activity patterns on spreading processes," *Physical Rev. Lett.*, vol. 98, no. 15, p. 158702, Apr. 2007. doi: 10.1103/PhysRevLett.98.158702

[76] M. Karsai *et al.*, "Small but slow world: how network topology and burstiness slow down spreading," *Physical Rev. E*, vol. 83, no. 2, p. 025102, Feb. 2011. doi: 10.1103/PhysRevE.83.025102

[77] J. L. Iribarren and E. Moro, "Impact of human activity patterns on the dynamics of information diffusion," *Physical Rev. Lett.*, vol. 103, no. 3, p. 038702, July 2009. doi: 10.1103/PhysRevLett.103.038702

[78] R. Lambiotte, L. Tabourier, and J.-C. Delvenne, "Burstiness and spreading on temporal networks," *The Eur. Physical J. B*, vol. 86, no. 7, p. 320, July 2013. doi: 10.1140/epjb/e2013-40456-9

[79] T. Takaguchi, N. Masuda, and P. Holme, "Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics," *PLoS ONE*, vol. 8, no. 7, p. e68629, July 2013. doi: 10.1371/journal.pone.0068629

[80] P. Van Mieghem and R. Van De Bovenkamp, "Non-Markovian infection spread dramatically alters the susceptible- infected-susceptible epidemic threshold in networks," *Physical Rev. Lett.*, vol. 110, no. 10, p. 108701, Mar. 2013. doi: 10.1103/PhysRevLett.110.108701

[81] D. Mistry, Q. Zhang, N. Perra, and A. Baronchelli, "Committed activists and the reshaping of status-quo social consensus," *Physical Rev. E*, vol. 92, no. 4, p. 042805, Oct. 2015. doi: 10.1103/PhysRevE.92.042805

[82] C. Doyle, S. Sreenivasan, B. Szymanski, and G. Korniss, "Social consensus and tipping points with opinion inertia," *Physica A*, vol. 443, pp. 316–323, Feb. 2016. doi: 10.1016/j.physa.2015.09.081

[83] X. T. Liu, Z. X. Wu, and L. Zhang, "Impact of committed individuals on vaccination behavior," *Physical Rev. E*, vol. 86, no. 5, p. 051132, Nov. 2012. doi: 10.1103/PhysRevE.86.051132

[84] W. Zhang, C. Lim, and B. K. Szymanski, "Analytic treatment of tipping points for social consensus in large random networks," *Physical Rev. E*, vol. 86, no. 6, p. 061134, Dec. 2012. doi: 10.1103/PhysRevE.86.061134

[85] M. Mobilia, "Does a single zealot affect an infinite group of voters?" *Physical Rev. Lett.*, vol. 91, no. 2, p. 028701, July 2003. doi: 10.1103/PhysRevLett.91.028701

[86] M. Mobilia, A. Petersen, and S. Redner, "On the role of zealotry in the voter model," *J. of Statistical Mechanics*, vol. 2007, no. 08, pp. P08 029–P08 029, Aug. 2007. doi: 10.1088/1742-5468/2007/08/P08029

[87] E. Yildiz, D. Acemoglu, A. E. Ozdaglar, A. Saberi, and A. Scaglione, "Discrete opinion dynamics with stubborn agents," *SSRN Electronic J.*, Jan. 2011. doi: 10.2139/ssrn.1744113

[88] S. A. Marvel, H. Hong, A. Papush, and S. H. Strogatz, "Encouraging moderation: clues from a simple model of ideological conflict," *Physical Rev. Lett.*, vol. 109, no. 11, p. 118702, Sep. 2012. doi: 10.1103/PhysRevLett.109.118702

[89] G. Verma, A. Swami, and K. Chan, "The impact of competing zealots on opinion dynamics," *Physica A*, vol. 395, pp. 310–331, Feb. 2014. doi: 10.1016/j.physa.2013.09.045

[90] G. Trenkler, "Continuous univariate distributions," *Computational Statist. & Data Anal.*, vol. 21, no. 1, p. 119, 1996. doi: 10.1016/0167-9473(96)90015-8

[91] J. S. White, "Weibull renewal analysis," in *SAE Tech. Paper*, Jan. 1964. doi: 10.4271/640624. ISSN 0148-7191 p. 640624.

[92] X. Castelló, A. Baronchelli, and V. Loreto, "Consensus and ordering in language dynamics," *The Eur. Physical J. B*, vol. 71, no. 4, pp. 557–564, Oct. 2009. doi: 10.1140/epjb/e2009-00284-2

[93] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Rev. of Modern Physics*, vol. 81, no. 2, pp. 591–646, May 2009. doi: 10.1103/RevModPhys.81.591

[94] L. Dall'Asta and T. Galla, "Algebraic coarsening in voter models with intermediate states," *J. of Physics A*, vol. 41, no. 43, p. 435003, Oct. 2008. doi: 10.1088/1751-8113/41/43/435003

[95] F. Vazquez and C. López, "Systems with two symmetric absorbing states: Relating the microscopic dynamics with the macroscopic behavior," *Physical Rev. E*, vol. 78, no. 6, p. 061127, Dec. 2008. doi: 10.1103/PhysRevE.78.061127

[96] R. Roy, F. W. J. Olver, R. A. Askey, and R. Wong, "Algebraic and analytic methods," in *Digital Library of Mathematical Functions*, F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. Saunders, Eds. Gaithersberg, MD: National Institute of Standards and Technology, 2018, version 1.0.19. [Online]. Available: https://dlmf.nist.gov/8.17.E1. Accessed on: July 12, 2018.

[97] R. B. Paris, "Incomplete gamma and related functions," in *Digital Library of Mathematical Functions*, F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. Saunders, Eds. Gaithersberg, MD: National Institute of Standards and Technology, 2018, version 1.0.19. [Online]. Available: https://dlmf.nist.gov/8.17.E1. Accessed on: July 12, 2018.

[98] K. Weber, "On the evolution of random graphs in the n-cube." *Public Math. Inst. Hungarian Academy Sci.*, vol. 5, no. 1, pp. 17–60, 1985.

[99] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. of the 19th Int. Conf. on World Wide Web - WWW '10*. New York, New York, USA: ACM Press, 2010. doi: 10.1145/1772690.1772751 p. 591.

[100] H. Gil de Zúñiga, N. Jung, and S. Valenzuela, "Social media use for news and individuals' social capital, civic engagement and political participation," *J. of Comput. Mediated Commun.*, vol. 17, no. 3, pp. 319–336, Apr. 2012. doi: 10.1111/j.1083-6101.2012.01574.x

[101] A. Hermida, F. Fletcher, D. Korell, and D. Logan, "Share, like, recommend," *Journalism Stud.*, vol. 13, no. 5-6, pp. 815–824, Oct. 2012. doi: 10.1080/1461670X.2012.664430

[102] D. Asher *et al.*, "The investigation of social media data thresholds for opinion formation," in *ICCRTS 2017: 22nd Int. Command and Control Res. & Technol. Symp.*, Los Angeles, CA, 2017, p. 27.

[103] A. Derrik, E. Bowman, C. Doyle, G. Korniss, and B. K. Szymanski, "An experimental approach to identify perceived opinion formation thresholds from social media computers in human behavior," *Comput. in Human Behavior*, submitted 2017.

[104] M. Adedoyin-Olowe, M. M. Gaber, and F. Stahl, "A survey of data mining techniques for social media analysis," Dec. 2013.

[105] M. Kaschesky, P. Sobkowicz, and G. Bouchard, "Opinion mining in social media," in *Proc. of the 12th Annu. Int. Digital Government Res. Conf. on Digital Government Innovation in Challenging Times - DGO '11.* New York, New York, USA: ACM Press, 2011. doi: 10.1145/2037556.2037607 p. 317.

[106] P. Sobkowicz, M. Kaschesky, and G. Bouchard, "Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web," *Government Inform. Quarterly*, vol. 29, no. 4, pp. 470–479, Oct. 2012. doi: 10.1016/J.GIQ.2012.06.005

[107] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell, "Mining facebook data for predictive personality modeling," in *Proc. of the 7th Int. AAAI Conf. on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA*, 2013, pp. 23–26.

[108] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Personal and Ubiquitous Computing*, vol. 17, no. 3, pp. 433–450, Mar. 2013. doi: 10.1007/s00779-011-0490-1

[109] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk," *Perspectives on Psychological Sci.*, vol. 6, no. 1, pp. 3–5, Jan. 2011. doi: 10.1177/1745691610393980

[110] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, June 2010.

[111] K. Crowston, "Amazon Mechanical Turk: A research tool for organizations and information systems scholars," in *Shaping the Future of ICT Res. Methods and*

Approaches, ser. IFIP Advances in Inform. and Commun. Technol., A. Bhattacherjee and B. Fitzgerald, Eds. Springer, Berlin, Heidelberg, 2012, vol. 389, pp. 210–221.

[112] M. J. Zaki, W. Meira Jr, and W. Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York, NY: Cambridge University Press, 2014.

[113] M. Hahsler, C. Buchta, B. Gruen, and K. Hornik, *arules: Mining Association Rules and Frequent Itemsets*, 2018, r package version 1.6-1.

[114] M. Hahsler, B. Gruen, and K. Hornik, "arules – computational environment for mining association rules and frequent item sets," *J. of Statistical Software*, vol. 14, no. 15, pp. 1–25, Oct. 2005.

[115] M. Hahsler, S. Chelluboina, K. Hornik, and C. Buchta, "The arules r-package ecosystem: analyzing interesting patterns from large transaction datasets," *J. of Mach. Learning Res.*, vol. 12, pp. 1977–1981, 2011.

[116] B. Everitt and A. Skrondal, *The Cambridge Dictionary of Statist.*, 4th ed. Cambridge, UK: Cambridge University Press, 2010.

[117] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Sci.*, vol. 4, no. 1, p. 10, Dec. 2015. doi: 10.1140/epjds/s13688-015-0046-0

[118] R. Lambiotte *et al.*, "Geographical dispersal of mobile communication networks," *Physica A*, vol. 387, no. 21, pp. 5317–5325, 2008. doi: https://doi.org/10.1016/j.physa.2008.05.014

[119] R. Ling, T. F. Bertel, and P. R. Sundsøy, "The socio-demographics of texting: An analysis of traffic data," *New Media & Soc.*, vol. 14, no. 2, pp. 281–298, Mar. 2012. doi: 10.1177/1461444811412711

[120] J. P. Onnela *et al.*, "Structure and tie strengths in mobile communication networks," *Proc. of the Nat. Academy of Sci. of the USA*, vol. 104, no. 18, pp. 7332–6, May 2007. doi: 10.1073/pnas.0610245104

[121] M.-X. Li *et al.*, "Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data," *New J. of Physics*, vol. 16, no. 8, p. 083038, Aug. 2014. doi: 10.1088/1367-2630/16/8/083038

[122] G. Zipf, *Human Behavior and the Principle of Least Effort*. Hoboken, NJ: John Wiley and Sons, 1950.

[123] H. Mao, X. Shuai, Y.-Y. Ahn, and J. Bollen, "Mobile communications reveal the regional economy in Côte d'Ivoire," in *Proc. of NetMob*, 2013.

[124] C. Smith-Clarke, A. Mashhadi, and L. Capra, "Poverty on the cheap," in *Proc. of the 32nd Annu. ACM Conf. on Human Factors in Computing Systems - CHI '14*. New York, New York, USA: ACM Press, 2014. doi: 10.1145/2556288.2557358 pp. 511–520.

[125] V. K. Singh, L. Freeman, B. Lepri, and A. S. Pentland, "Predicting spending behavior using socio-mobile features," in *2013 Int. Conf. on Social Computing*. IEEE, Sep. 2013. doi: 10.1109/SocialCom.2013.33 pp. 174–179.

[126] S. Luo, F. Morone, C. Sarraute, M. Travizano, and H. A. Makse, "Inferring personal economic status from social network location," *Nature Commun.*, vol. 8, p. 15227, May 2017. doi: 10.1038/ncomms15227

[127] Y. A. de Montjoye, J. Quoidbach, F. Robic, and A. Pentland, "Predicting personality using novel mobile phone-based metrics," in *Int. Conf. on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, Berlin, Heidelberg, 2013. doi: 10.1007/978-3-642-37210-0_6 pp. 48–55.

[128] V. Frias-Martinez, J. Virseda, and E. Frias-Martinez, "Socio-dconomic levels and human mobility," in *Qual Meets Quant Workshop-QMQ*, 2010.

[129] L. Kovanen, J. Saramaki, and K. Kaski, "Reciprocity of mobile phone calls," *Dynamics of Socio-Economic Syst.*, vol. 2, no. 2, pp. 138–151, Feb. 2010.

[130] A.-L. Barabási and M. Posfai, *Network Science*. Cambridge, UK: Cambridge University Press, 2016.

[131] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: a Python package for analysis of heavy-tailed distributions," *PLoS ONE*, vol. 9, no. 1, p. e85777, Jan. 2014. doi: 10.1371/journal.pone.0085777

[132] A. Klaus, S. Yu, and D. Plenz, "Statistical analyses support power law distributionsfFound in neuronal avalanches," *PLoS ONE*, vol. 6, no. 5, p. e19779, May 2011. doi: 10.1371/journal.pone.0019779

[133] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, Nov. 2009. doi: 10.1137/070710111

[134] M. E. J. Newman, "Scientific collaboration networks: shortest paths, weighted networks, and centrality," *Physical Rev. E*, vol. 64, no. 1, p. 016132, June 2001. doi: 10.1103/PhysRevE.64.016132

[135] Y. Rochat, "Closeness centrality extended to unconnected graphs : the harmonic centrality index," in *ASNA*, Zurich, 2009.

[136] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, July 2010. doi: 10.1016/J.SOCNET.2010.03.006

[137] J. Xie and B. K. Szymanski, "Community detection using a neighborhood strength driven Label Propagation Algorithm," in *2011 IEEE Network Sci. Workshop*. IEEE, June 2011. doi: 10.1109/NSW.2011.6004645 pp. 188–195.

[138] J. Xie, B. K. Szymanski, and X. Liu, "SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *2011 IEEE 11th Int. Conf. on Data Mining Workshops*. IEEE, Dec. 2011. doi: 10.1109/ICDMW.2011.154 pp. 344–349.