Studying Network Structural Changes Using Information Event Signatures

Jeffery Hemsley

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Robert M. Mason

Kirsten Foot

Emma Spiro

Malcolm Parks

Program Authorized to Offer Degree:

The Information School

University of Washington

Abstract

Studying network structural changes using information event signatures

Jeff Hemsley

Chair of the Supervisory Committee:

Professor Robert M. Mason

Information School

Co-chair:

Professor Karine Nahon

Information School

The diffusion of information is important: It impacts commerce. It influences government. It connects people in new ways. The distributed nature of our digital social networks means that traditional gatekeepers (newspapers, radio, television, and governments) lose some control over the flow of information, while new gatekeepers emerge quickly in networks of individuals who share interests or grievances. Using exploratory data analysis and confirmatory statistics to analyze over 64 million Occupy movement tweets, this dissertation makes four essential contributions that enhance our understanding of the relationship between the flow of information and the dynamics of social networks. First, based on a large set of Twitter data related to the Occupy Wall Street movement, it introduces a parameterized signature model of individual information flows. Second, it demonstrates that both the path of the information flow and the changes in the structure of the network, as measured by the growth of network gatekeepers

within the Occupy movement, are related to parameters of the model. Third, the analysis suggests that the Occupy gatekeepers recursively extend their reach by repeatedly promoting information that users shared deep into Twitter's social network of followers. Fourth, the model provides the initial steps towards a theory explaining the process by which social network dynamics and information flows interact. This model's capacity to identify information flows deep in networks and to predict trends that potentially alter those networks will prove useful to individuals, to organizations, and to governments.

Acknowledgements

This research would not have been possible without support from the National Science Foundation (#1243170, #1342251), Microsoft Fuse Labs, and Amazon Web Services.

I am grateful to Robert M. Mason, Karine Nahon, Kirsten Foot, Malcolm Parks and Emma Spiro for providing direction, support and encouragement.

Special thanks to Shawn Walker and Josef Eckert for insight and critical input, to Shawn Walker for his deep knowledge about Twitter data, to Lucas Koepke for help with statistical modeling, and to Vince Guaraldi for respite.

I am most deeply indebted to Arlene Pritzker, whose steadfast love and confidence in me kept me focused and moving forward when it was needed.

Table of Contents

Acknowledgements 5
Table of Contents 6
Table of Figures 7
Preface 10
1 Introduction 11
1.1 Genesis and motivation 11
1.2 Motivating Theory 14
1.2.1 Weak ties 14
1.2.2 Network Gatekeeping 16
1.3 Focus of the study 17
1.4 Delimitations and Limitations 19
1.4.1 Delimitations 19
1.4.2 Limitations 19
1.5 Significance of the Study 20
2 Phase 1: Exploration and Model Building 21
2.1 Related literature 21
2.2 Methodology 28
2.2.1 Approach 28
2.2.2 Procedures 30
2.3 Findings 46
2.3.1 Parameterized Signature Model 46
2.3.2 Operationalization of Network Changes 50
3 Phase 2: Network Dynamics and Information Flows 52
3.1 Related Literature 53
3.1.1 Network structure and information flows 53
3.1.2 Network structural changes and information flows 59
3.2 Approach and Methodology 62
3.2.1 Approach 62
3.2.2 Procedures 67
3.3 Findings 82
4 Discussion 86
4.1 The interaction of flow and network dynamics 86
4.2 Theory Development 97
4.3 The role of scripting in this work 99
5 Conclusions 101
5.1 Summary 101
5.1.1 Findings and Implications 102
5.1.2 Limitations 103
5.2 Contributions 107
5.2.1 Making explicit the exploratory phase 107
5.2.2 The Signature Model 107
5.2.3 Operationalization of network dynamics 108

5.2.4 The interaction of network dynamics and information flows 108
5.2.5 Theory Development 109
5.3 Future directions 109
Appendix A: Glossary 111
Appendix B: Twitter & Social Media Lab data details 118
B.i: Twitter Streaming API process overview 118
B.ii: Social Media Lab collection system 118
B.iii: Social Media Lab search term list 119
Appendix C: Example JSON tweet 122
Appendix D: Inferring and measuring the flow path 129
Appendix E: Example Comparison Retweet Network 136
References 138

Table of Figures

Figure 0.1: Overview of phases of work where the findings from Phase 1 used in Phase 2, but model development in Phase 2 led to further exploration. 11
Figure 1.1. Occupy websites discovered by IssueCrawler. The more incoming links from other sites in the network, the larger and redder is the circle. 12
Figure 2.1. Sigmoid curve, rate of infection, and campaign video views. A, the sigmoid or "S" curve, shows percent (%) exposure or infection. B shows the rate of infection corresponding to sigmoid % of infection. C shows the example rate of campaign video views, which is dissimilar to figure B; therefore, Boynton suggests these are not viral. Plots are based on Boynton's figures. 22
Figure 2.2. Example power-law distribution with various alpha (shape parameter) values. 23
Figure 2.3. Normalized YouTube video views. Reprinted from Broxton, Interian, Vaver, and Wattenhofer, 2010. 25
Figure 2.4: Signature shapes driven by promotional or social forces. Reprinted from (Karine Nahon and Hemsley 2013) 26
Figure 2.5. Daily rate of tweets (black) and retweets (red) in the Social Media Lab's corpus of Occupy tweets. 31
Figure 2.6. Size of RTEs in the Social Media Lab's corpus of tweets. This displays a heavily tailed distribution. 33
Figure 2.7. Graphical overview of the selection of data used in Phase 1 of this study. 34
Figure 2.8. Example of a network data visualization of a retweet network using the Social Media Lab data. 36
Figure 2.9. Example power-law distribution plots. Uses Social Media Lab data. 38
Figure 2.10. Using Zoner Photo Lab to explore network data visualizations 39
Figure 2.11. Plot of the rate of the retweets of a specific tweet over time in minutes. The blue line and large number 50 are easily seen when the plot is zoomed out to compare with other signatures. as Social Media Lab data. 40

Figure 2.12. Plot relating rate of retweets to changes in followers. Bold features (red stars and blue line) allow comparison across a large number of plots. The details are still present as smaller text but only visible when zooming in. Social Media Lab data. 41
Figure 2.13. Example of zooming out to compare many RTE signatures at the same time. 43
Figure 2.14. Example of RTE overlap. Social Media Lab data. 44
Figure 2.15. An RTE initiated by the account OccupyWallSt with a long delay retweet. Social Media Lab data. 45
Figure 2.16. A sample of 1000 information flows from the Occupy corpus of tweets. Social Media Lab data. 47
Figure 2.17. Information flow signature model showing the phases of its life cycle (ramp-up, peak, and decay phase) as well as quantifiable characteristics (peak time, peak rate, and its shape given by alpha). 48
Figure 2.18. The rise in the number of followers of the OccupyWallSt account. Social Media Lab data. 52
Figure 3.1. Information flow topologies in the blogosphere. Blog A initiates a flow that may result in different topologies. This figure depicts star and tree topologies as well as sharing chains in a tree. 55
Figure 3.2. Example information flow topologies (above) with related signatures (below), where the wait time from receipt of a message to sharing it is held constant. 57
Figure 3.3. Example Pareto distribution for modeling human wait times instead of assuming constant wait time. 58
Figure 3.4. Overview of variables used in the models developed for Phase 2. 63
Figure 3.5: Example information flow paths. Given a hypothetical actual path of a flow (left), the apparent path suggested by the metadata will always be a star topology (right), thus the need to infer the path network. 64
Figure 3.6. Plot of a sample of 1000 RTEs, reprinted for convenience. Social Media Lab data.. 66
Figure 3.7. Graphical overview of the selection of data used in Phase 2 of this study. 68
Figure 3.8. Frequency of sizes of retweet events, where size is the number of retweets plus the initial tweet. Social Media Lab data. 69
Figure 3.9. Graphical version of model 1 with variable groups and the specific variables. 72
Figure 3.10. Distribution of alpha and QQ plot showing a roughly normal distribution with some deviation at the high end 73
Figure 3.11. Graphical version of model 2 with variable groups and the specific variables. 75
Figure 3.12. Distribution of change in followers and the QQ plot for the transformed variable. Note that the transformed variable is roughly normal except for deviation at the end high end.. 76
Figure 3.13. Model 1 diagnostic plots: residuals versus fitted and QQ Plot for normality of residuals. 80
Figure 3.14. Model 2 diagnostic plots: residuals versus fitted and QQ Plot for normality of residuals. 81
Figure 4.1. Signature model from phase 1, included here for convenience. 87
Figure 4.2. Regression models from phase 2, included here for convience. 88
Figure 4.3. Conceptual relationship of the regression models. 89
Figure 4.4. Signature model parameter relationships found in model 1. Lower and later peaks are related sharper decay phases. 91
Figure 4.5. Rogers' graph showing the rate of diffusion of innovations. 93
Figure 5.1: Social Media Lab collection, processing, storage, and analysis system. 119

Figure D.1: Example RTE with a pronounced second spike, likely due to a retweet by a celebrity with many followers. 131
Figure D.2. Example networks with closeness centralization measurements. 134
Figure D.3. Relationships between closeness centralization and both path length (left) and number of paths (right). 135

Preface

> Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. (Tukey 1980) (emphasis original)

This dissertation explores the relationship between the flow of information and social network dynamics. The research took place in two overlapping phases. The first phase employed an interpretive exploratory data analysis approach, while the second phase utilized confirmatory statistical analysis. The first phase proceeded in an iterative fashion, with exploration and discovery followed by more exploration and discovery, often leading to dead ends, but ultimately honing in on the ways in which a large set of unstructured data could be used to address my research interests. More specifically, Phase 1 results in an information flow model and an understanding of the ways in which the parameters of the model and the Twitter network can be explored to understand the relationship between information flows and network dynamics. Phase two utilizes inferential statistics to examine this relationship. The findings from Phase 2 illuminate how the parameters of the information flow model are related the network dynamics. Thus the work of the first phase informed and made possible the work of the second phase (see figure 0.1). The phases overlapped because fitting the inferential models in Phase 2 resulted in new discoveries about the data, which led to further exploration.

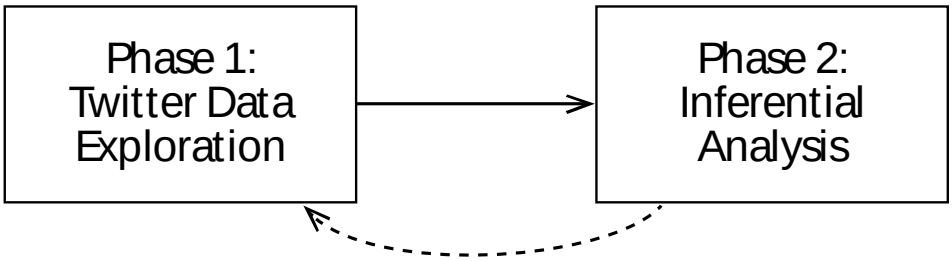


Figure 0.1: Overview of phases of work where the findings from Phase 1 used in Phase 2, but model development in Phase 2 led to further exploration.

The presentation of the research is also broken into two phases. The literature, methodology and findings of Phase 1 are covered in Chapter 2, with Chapter 3 covering the literature, methodology and findings of the work done in Phase 2. This approach simplifies the presentation of the second phase, which requires concepts and language developed in Phase 1.

Finally, a first person narrative is in two sections of the document: the motivation section of the introduction and the data exploration section in Phase 1's methodology. These are intended to record the thought processes, assumptions, decisions, choices, and discoveries made during the research.

1 Introduction

1.1 Genesis and motivation

Like many individuals, I learned about the Occupy Wall Street (OWS) movement in the autumn of 2011, after reading Facebook posts from a friend. I use the word friend here in the same way it is used on Facebook, that is, Josef Eckert (Joe) was more of an acquaintance;

someone I'd not interacted with since the class in which we'd met during the previous year. But the content Joe was posting was interesting to me. It framed OWS as offended by Wall Street's excesses and influence in government at a time when America was barely recovering from the 2008 housing collapse. Joe wasn't the origin of those posts. He'd shared it into his own feed only after seeing it in one of his friend's feed, who'd run across and shared it in a similar way. The post came to my attention through a *chain of social sharing* that traced a specific *path* through a social network. Two things happened as a result of Joe's posts: (1) I realized we had common interests and initiated a closer friendship; and (2) I became informed about a topic of interest to me, one I resonated with. Once I became aware of the content, I sought out the origin, the OWS Facebook Page, and began to *follow* it. This meant that I subscribed directly to the OWS feed so that I would receive posts directly instead of through a sharing chain. It also meant that I created a new link in the network that dramatically shortened the path between OWS and my *followers*, or the people who subscribed to my social media posts.

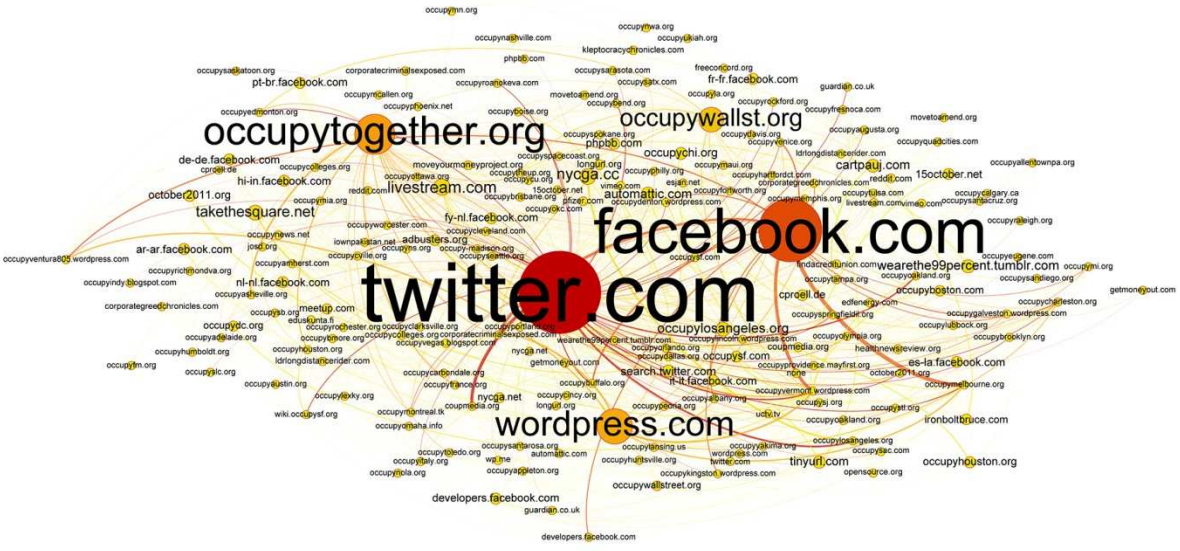


Figure 1.1. Occupy websites discovered by IssueCrawler. The more incoming links from other sites in the network, the larger and redder is the circle.

As I watched the growth of the OWS I decided that there was a great research opportunity, one where I might be able to tie together my interest in networks with information flows. Together with Shawn Walker and Josef Eckert, we began to collect data. First we collected MeetUp data to track the growth of the movement. On MeetUp, people started or joined Occupy groups in cities across America and the world. This allowed us to get an idea of the geographic growth of OWS. I also wanted to see OWS' online presence, so I used IssueCrawler[1] to discover Occupy-related sites. What I discovered was that the vast majority of sites linked to Twitter (see figure 1.1). Based on the number of sites linking to it, Twitter appeared to be a hub for the movement.

As a team, we began collecting Twitter data, and I began exploring the data in an effort to understand how Occupy Twitter could help me understand more about how the flow of information is related to network dynamics.

As this example shows, social media, as a set of tools, is changing how we become informed, connect, and interact. In our globally connected social networks, we can become informed about, and get involved with, social movements like Occupy even when NPR, Fox News or CNN do not provide coverage. The example also shows that individuals, groups, or organizations can gain new followers when their posts are socially shared by interested parties. Indeed, as of October 19, 2011, the Occupy Wall Street user account on Twitter had 83,264 followers, gaining another 1,748 followers in the next day and over 40,000 more over the next month. By selecting to follow OWS, each of those 40,000 people created a new link in the Twitter social network. Thus, as information flows through our social networks, it has the

[1] IssueCrawler starts with the URLs of a few seed sites and crawls the links of those sites to discover related sites. When crawling, if a new site is discovered and shows up on three other sites, it is added to the list of seeds and also crawled. https://www.issuecrawler.net/

potential to alter the linking structure (topological linking patterns) of those networks, which in turn alters the possible paths for future information flows.

In this study, I provide a model that quantifies characteristics of information flows and shows that these characteristics are related to (1) the path through which the information flowed and (2) changes in the social network. Together these results provide us with an understanding of how the path through which information flows in social networks relates to changes in those social networks and how we can detect network altering information flows. Understanding the coevolution of information flows and network structures[2] is critical for understanding social networks in general and, more specifically, the growth of social movements like OWS in our emerging digitally mediated social landscape.

1.2 Motivating Theory

1.2.1 Weak ties

Granovetter (1973) argues that the amount of overlap, or the number of shared relationships in the social networks of two individuals (A and B), is directly related to the strength of their tie, or relationship. If A and B are strongly connected, it is likely that they have a large number of mutual friends/contacts. In fact Granovetter argues that if a strong tie exists between A and B, and a strong tie exists between B and C, this will lead to a tie between A and C. Given Granovetter's theory, if no tie exists between A and C then the tie between A and B, (and likewise B and C) must be a weak tie. Granovetter shows that job referrals came more frequently from weak ties than from strong ties and suggests that in tightly knit groups, or clusters, people tend to know the same people and have the same information.

[2] Unless otherwise noted, network structure refers to topological link patterns in networks.

The theory of weak ties has been used by numerous authors in a variety of contexts. For example, Burt (2004) finds that good ideas in an organization come from weak ties that span what he calls *structural holes* in networks. These holes are spaces between tightly knit groups or in a cluster. Weak ties *bridge* the holes and connect different clusters. In other examples, Ellison et al. (2010) show that users on Facebook maintain larger sets of *weak-tie* connections in their social networks then those not on Facebook, and Bakshy (2012) shows that strong ties are more influential on Facebook, but that it is the abundance of weak ties that are responsible for the propagation of novel information.

When novel information flows across weak ties and bridges network holes, it has a greater potential of reaching new audiences, who then share it with their own friends and followers. This is supported by Petrovic et al. (2011). They show that tweets identified as novel are more likely to be retweeted. In related work, Teng et al. (2012) show that novelty is related to new ties in Twitter discussion networks.[3] Thus novel information flowing over weak tie bridges may also be related to the creation of new ties in Twitter's social network of followers[4].

Granovetter's theory and the empirical work mentioned above provide the theoretical motivation for this work, which posits that models can be developed that show the relationship between information flows and networks. This is because information that crosses weak-tie bridges may precipitate increased flows in new clusters, which should be detectable as higher levels of sharing over time.

[3] In Twitter discussion networks, ties are constituted between users (nodes) who mention each other.
[4] This study employs the phrase Twitter's social network to indicate the follower relationships between Twitter users. Some researchers suggest that Twitter is not a social network, but a kind of news media where users tend to share information from traditional media sources (Kwak et al. 2010). For this study, it shall be considered a social network for conceptual and analytical convenience.

1.2.2 Network Gatekeeping

This study conceptualizes the accounts associated with the Occupy movement as *network gatekeepers* (Barzilai-Nahon 2009; Barzilai-Nahon 2008). Gatekeepers are "people, collectives, companies, or governments that, as a result of their location in a network, can promote or suppress the movement of information from one part of a network to another" (Nahon and Hemsley 2013, 7). One of the key ideas in the conceptualization of gatekeepers is that through their position, they may connect networks that would otherwise not be connected. At the beginning of this study, more than 85,000 Twitter users followed the OccupyWallSt account, giving it the ability to selectively promote information to its followers in disparate parts of the Twitter network.

The individuals who follow the Occupy gatekeepers are the *gated*, individuals who choose to follow the gatekeeper in a media environment where they have many other options for information. The gated may simply receive information from the gatekeeper or there may be a two-way communicative relationship between them. This is because the gated are assumed to have the potential to produce information that the gatekeeper may choose to promote. Because of the plethora of media choices, the gated may choose to follow alternate sources of information. Thus the position of network gatekeepers is dynamic and rises and falls based on the preferences of the gated (Nahon 2011). Over the course of this study, the Occupy accounts generally gained large numbers of followers, indicating a sustained and growing gated audience.

Network Gatekeeping Theory (NGT) is often employed as lens through which to study issues of social power related to the flow of information in networks. In this study the concept is used to understand the observed pattern of a sustained and growing audience of users following the Occupy accounts. Borrowing from social network analysis (Wasserman and Faust 1994), a

gatekeeper could be conceptualized as a hub, a single node with many nodes connected to it. The formation of the gatekeeper hub, the nodes that make up the gated, and the links from each gated node to the hub is a *star* (Wasserman and Faust 1994). These stars structures are assumed to be embedded in larger networks where the gated may or may not share links with each other.

1.3 Focus of the study

In their introduction to *Understanding Robust and Exploratory Data Analysis,* Hoaglin et al. (1983) note that "practical data analysis identifies two broad phases: exploratory and confirmatory" (Hoaglin, Mosteller, and Tukey 1983, 3:1). Exploratory analysis emphasizes flexibility in searching for patterns and evidence in the data, while confirmatory analysis focuses on evaluating the evidence given the patterns found in the exploratory phase. This study employs this breakdown and proceeds in two phases. The first phase addresses the overarching question, *How can we measure and study the process by which information flows are related to social network dynamics?* The second phase examines a second overarching question, *In what ways do information flows and social networks interact?* These questions are informed by prior studies in the area of network science and information flows[5].

The work of the first phase employs *exploratory data analysis (EDA)*, a data analysis approach encouraged by John Tukey (1980; Tukey 1977). EDA is not a formalized method of analysis; rather it is a way of thinking, a curiosity, and a willingness to look at data in different ways that provide the practitioner with both a broad and deep understanding of what is in the data, what kinds of questions can be answered with the data, and the quality of those answers. It is an intuitive approach driven by curiosity.

[5] Study areas also known as diffusion of information and citation analysis.

The second phase employs the findings from the first phase. The first phase findings are (1) a general model of temporal information flows, defined by a set of parameters[6] and referred to as an *information flow signature*, or *signature model*; and (2) a method for measuring changes in the topological linking structure of the Twitter social network, specifically, by tracking changes in counting a user's followers[7].

The second phase is also exploratory in that the goal is to create parsimonious models that confirm (or refute) the relationships between the parameters of the signature model and social network dynamics with which the information flow interacts. Specifically, the second phase employs variance models to (1) understand how the path of sharing through which the information flowed is related to the parameters of the signature model[8]; and (2) understand how the parameters of the signatures of information flows are related to changes in the social network. The findings from Phase 2 show that measureable parameters of the signature model are related both to the path of sharing through which the information flows and to changes in the social network during the flow.

The context within which this study takes place is the Occupy Wall Street movement. The information flows under investigation are Tweets (and their subsequent associated retweets) related to the Occupy movement from October 19, 2011 to June 30, 2012. The data used in this research was drawn from the Social Media Lab's corpus of 64,298,061 tweets collected from Twitter's streaming application programming interface (API).

[6] These parameters include measures like peak rate of flow and time to peak as well as others.
[7] Note that the structure under consideration is local to the user whose number of followers is being tracked. See section 2.3.2 for a detailed discussion.
[8] The path of the flow is inferred using a method developed by Gomez-Rodriguez, Leskovec, and Krause (2010). This method is introduced in section 3.2.1 and detailed in Appendix D.

1.4 Delimitations and Limitations

1.4.1 Delimitations

This study limits the operationalization of an *information flow* to a single unit of content, or *message*, that is shared intact within a social network. Certainly any flow of information takes place within a complex information ecosystem, and related comments, tweets, response videos, mash-ups, memes or any of a myriad of ways that a discussion could be carried on could all be part of the same information flow. Information that is reframed (for example, a news item reported by Fox versus CNN), re-contextualized, de-contextualized, shared across platforms, discussed over lunch, or lectured about could also be considered part of the same flow. These are all beyond the scope of this project. For this study, a tweet and the retweets of that tweet are all considered a single unit and operationalize the concept of an *information flow* or *information flow event*.

1.4.2 Limitations

The data for this study includes tweets and retweets related to the Occupy Movement. This should be considered both an asset and a limitation of this work. First, being constrained by the Occupy movement limits the generalizability of this work because the interaction between OWS information flows and social networks may be different than what might be found in different contexts, such as a new movie, TV show, political scandal, or even other social movements. Second, OWS information flows likely occurred across the myriad of communication platforms available to people: phones, radio, Facebook, Reddit and snail-mail, just to name a few. Therefore, focusing only on Twitter data may not reflect linking patterns in other mediums or linking patterns in Twitter that are due to communications in other platforms.

On the other hand, the data set includes tweets from October 19, 2011 to June 6, 2012, capturing a period of substantial growth in the emergence of the Occupy network on Twitter. Studying this early stage of network emergence should yield richer examples of change than on preexisting networks.

Additional limitations are discussed in sections for which they are relevant as well as in the Conclusions section of this document.

1.5 Significance of the Study

The distributed nature of our digital social media networks means that mainstream media (Fox, CNN, NPR) and governments have less control over the flow of information. Gatekeepers can quickly emerge in networks of resistance, such as the Occupy Wall Street (OWS) movement (Hemsley and Eckert 2014). This study provides new insights into how information flows interact with the network dynamics that power the growth of these gatekeepers. It does so in three ways: (1) by proving a general signature model that parameterizes temporal aspects of information flows; (2) by demonstrating that the parameters of the model are related to both the path of the information flow and to the growth in the number of users following these gatekeepers; and (3) by arguing that the Occupy gatekeeper recursively extend their reach by repeatedly promoting information that users shared deep into Twitter's social network of followers.

Organizations seeking new markets and activists seeking new audiences can employ the signature model as a tool to optimize future messages for these goals. By comparing the content and signatures of their previous messages, organizations and activists identify messages that reached deep into networks. One of the advantages, then, of the model is that it provides feedback that can identify successful content for future replication. Understanding how the

signatures of information flows are related to changes in the network and the emergences of gatekeepers enables researchers and governments to identify early movements of interest. Use of the signature model, in conjunction with methods such as topic modeling, may yield insight into how different discussion topics diffuse in and influence network structures within social media platforms.

This work contributes to our understanding of information flows in social media networks, specifically, about the growth of resistance networks and emergent gatekeepers. In addition, the signature model's ability to compare the reach and longevity of emerging information flows may be applicable in areas of study such as citation analysis used in bibliometrics, where scholarly citation networks can be thought of as "exposing the underlying socio-cognitive structure of science" (Cronin 2001, 27).

2 Phase 1: Exploration and Model Building

This chapter covers the literature, methodology and findings of Phase 1, an exploratory analysis of Twitter data with the aim of answering the overarching question, *How can we measure and study the process by which information flows are related to social network dynamics?*

2.1 Related literature

Numerous authors have examined temporal patterns of information flows. Boynton (2009) examined the patterns of attention to the campaign videos of the presidential candidates of the U.S. 2008 Presidential election by studying the number of daily views of the videos.

Relying on models of virality from epidemiology, where the infection rate of biological viruses follows a sigmoid curve, he argues that the concept of virality does not fit these videos.

An example of a sigmoid curve can be seen in figure 2.1a. The initial point of an epidemic (temporal or spatial) starts with only one or a few number of nodes infected. The curve starts out flat. Because individuals often interact with, and potentially infect, many others, the rate of the spread of the contagion grows exponentially until the network becomes saturated. Once saturation begins, there are fewer and fewer uninfected individuals available, so the rate of growth of infection slows, and eventually reaches zero.

Figure 2.1b shows what the corresponding rate of new infections, or video views, would look like. The videos in Boyton's sample had daily view rates closer to that seen in figure 2.1c, and he declared the videos non-viral.

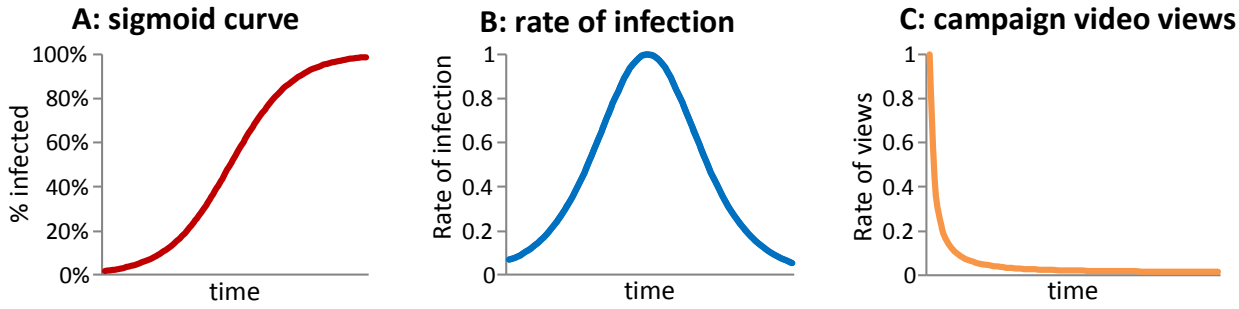


Figure 2.1. Sigmoid curve, rate of infection, and campaign video views. A, the sigmoid or "S" curve, shows percent (%) exposure or infection. B shows the rate of infection corresponding to sigmoid % of infection. C shows the example rate of campaign video views, which is dissimilar to figure B; therefore, Boynton suggests these are not viral. Plots are based on Boynton's figures.

Boynton's work conceptually links the sharing of content in social networks to temporally plotting the rate of attention to that content. He does not touch on the idea that the networks may change as a result of the flows. He only looks at the rate of views, without exploring any other temporal patterns of the campaign videos, such as how the rate at the peak might be related to the changing rate of views over time.

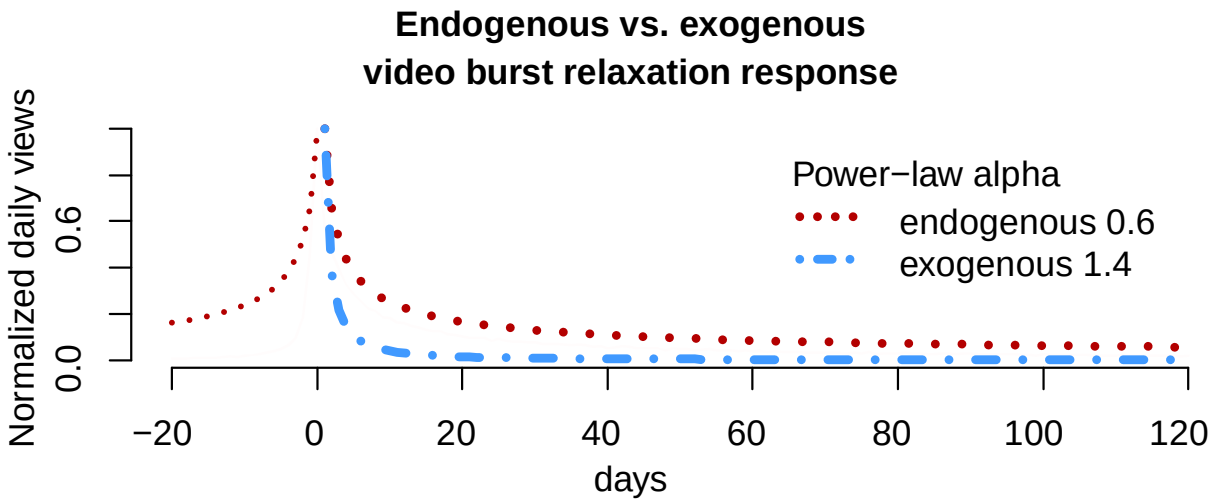


Figure 2.2. Example power-law distribution with various alpha (shape parameter) values.

Crane and Sornette (2008), however, do study patterns of views in more detail. They examine the rate of views of five million YouTube videos, collected over eight months prior to August, 2008, and find that while roughly 90% of the videos receive an insignificant amount of views or can be described statistically as a Poisson distribution, the reminder display a burst of viewing activity followed by *power-law relaxation response*. That is, a peak is followed by a rapid decay in the number of views (see figure 2.2). Key in their research is that they fit a power-law ($y = 1/x^{\alpha}$) to the shape of the relaxation response. Interestingly, Bak (1999) notes that phenomena that exhibit power-law distributions are indicative of complex systems, which are systems with variables that follow a given set of rules and outcomes with large variance. He claims that "explaining the observed statistical features of complex systems can be phrased mathematically as the problem of explaining the underlying power laws, and more specifically the values of the exponents" (Bak 1999, 27).

By clustering the estimated alpha values, Crane and Sornette find that lower values of alpha, or more gradual, relaxed decays are associated with *endogenous events*, videos that originate within the YouTube user community and are spread by users sharing links to the

videos, while higher values of alpha are associated with *exogenous events*, videos that originate outside of the community. In the context of this study, the importance of their work is not whether events are endogenous or exogenous. Rather, their work provides a foundation for considering ways to measure different parameters of information flows. Their exogenous videos tend to peak very quickly, with little or no period of build-up, while their endogenous videos often display a more pronounced build-up to the peak. For Crane and Sornette, these endogenous videos represent the concept of virality, where videos are shared by an increasing number of users with other users. They don't go beyond this to explore how the variation in these shapes might be related to aspects of the networks they pass through, nor do they discuss how those networks might be altered by the social sharing.

Broxton et al. (2010) go further to relate different levels of social sharing to rates of views of YouTube videos. They look at data from 1.5 million videos, posted between April 2009 and March 2010, excluding videos that received less than 100 views in their first 30 days. They study the relationship between the rate of video views and the fraction of those views that resulted from *social sharing*, cases where viewers of videos came to the video as a result of someone sharing a link with them. As can be seen in figure 2.3, videos with a higher percentile of social sharing tend to ramp up to, and decay off of, their peaks more gradually.

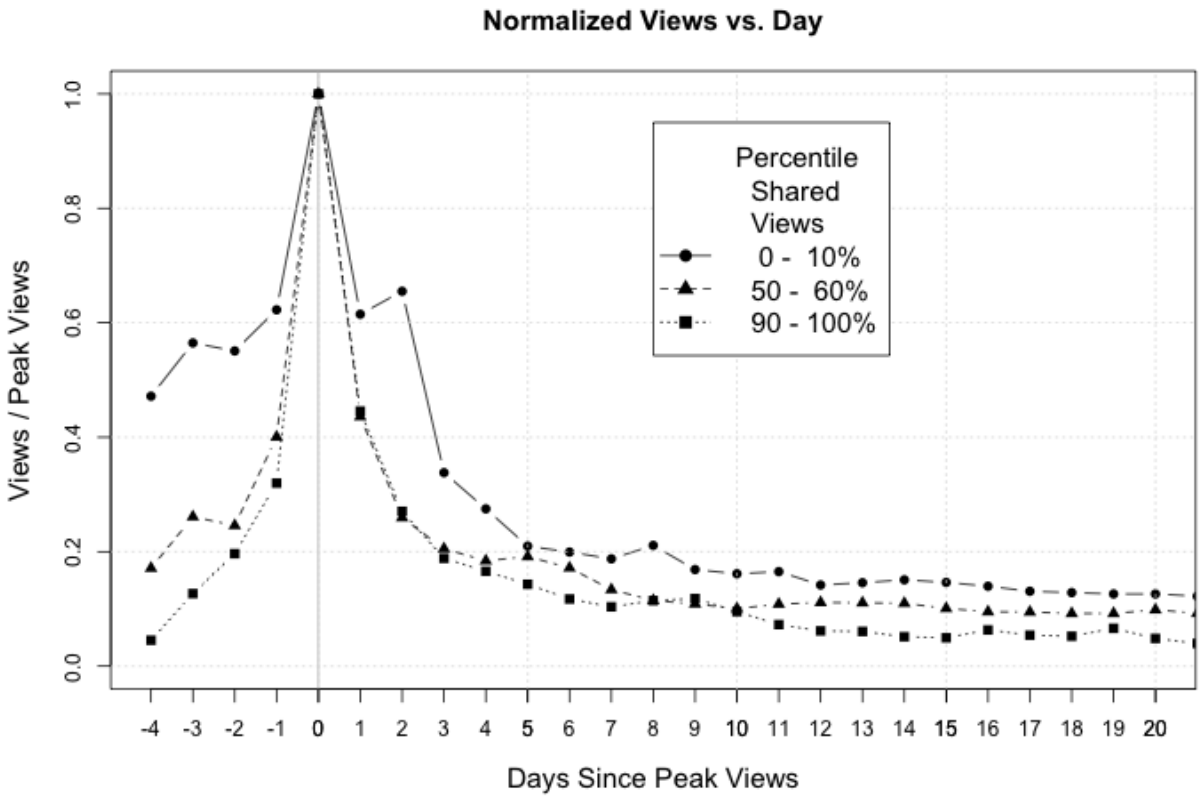


Figure 2.3. Normalized YouTube video views. Reprinted from Broxton, Interian, Vaver, and Wattenhofer, 2010.

Broxton et al. do not specifically identify different parameters of this normalized graph (that is, the lead up time to the peak, the period of decay, the height of the peak, or the shape of lead-up or decay period). And while they do link the rates of social sharing to the rate of daily views, their data does not support exploring the social network of users who shared content, how that network might be related to the rate of views, or how changes in that network were related to rates of social sharing.

Nahon and Hemsley (2013), drawing on previous work (Nahon and Hemsley 2014; Nahon et al. 2011), use the term *signature* to capture the generally observed pattern of the number of people in a system who engage with viral events (view a video or forward content). Unlike Boynton, they do not require that virality be defined by a strict sigmoid curve. They argue

that “top-down promotional and bottom-up social processes work together to create different signatures” (2013, 25) (see figure 2.4). Their argument relies on Crane and Sornette’s (2008) findings that fitted power-laws with higher-estimated alpha values (the shape parameter of the power-law) are related to exogenous events.

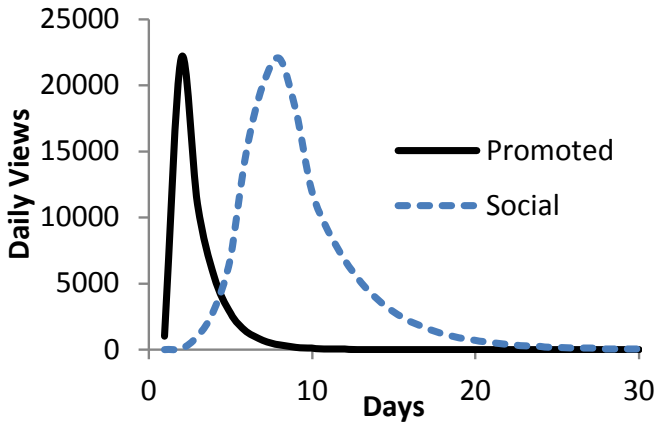


Figure 2.4: Signature shapes driven by promotional or social forces. Reprinted from (Karine Nahon and Hemsley 2013).

However, Nahon and Hemsley do not provide other empirical evidence and, in fact, suggest that the same ranges of alpha that Crane & Sornette found to be related to endogenous events (where $0.2 < \alpha < 0.6$) could be used to identify information flows that were *candidates* for inclusion in research projects studying virality. This is unsatisfying and problematic in a number of ways. For example, it is reasonable to expect that different platforms that support social sharing will show different rates of decay. Also, with the growth of mobile technology and sites like Twitter, researchers today might find alpha values that reflect a faster rate of sharing and network saturation. Beyond this, Nahon and Helmsley do not offer any insight or empirical work that relates the shape of their signatures to changes in network structures. However, they do

suggest that studying signatures "is an area for future work because the pace and shape of the decay phase may hold important clues about the viral process" (Nahon and Hemsley 2013, 101).

Clearly this body of work suggests that generalizing and formalizing a model of temporal information flows will be useful in studying the relationship between the flow of information and network dynamics. *Signature model* and *signature* are the terms used in the remainder of this document for a general model of temporal information flows.

Recall that the overarching question for this phase is about developing an understanding of how to measure and study the interaction of information flows and social network dynamics. The signature model represents only the information flow, not the networks that the information flows through or alters. What we need is ways to measure both. We need an empirical dataset that supports the creation of a signature model as well as ways to measure both the path through which information flows *and* resulting changes to the social network. As noted in the discussion of the motivation for this study, 40,000 people chose to become new followers of the Occupy Wall Street user account in the month leading up to November 20, 2011. Notably, each created a new link in the Twitter social network. Thus the Social Media Lab's corpus of Occupy-related tweets, described in the next section, is well suited for such an exploration. However, at the onset of this work it was unclear how, or even if, this data could be used to study the relationship between the flow of information and social network dynamics. Therefore, the guiding research question for Phase 1 of this dissertation is as follows:

> *Q1: How can Twitter data be used to explore the relationship between the flow of information and social network dynamics?*

2.2 Methodology

2.2.1 Approach

Exploratory data analysis (EDA), encouraged by John Tukey (1962; 1980; Tukey 1977), is an approach to gain insight about data, a kind of quantitative detective work that emphasizes flexibility of viewpoint and a willingness to find what we expect as well as what we don't. The tools of EDA include descriptive statistics, tables, and data visualizations. In this work, other means, such as writing software and socially interpreting data visualization with peers, are also included because they prove useful in developing a deeper insight into the data. Tukey and others (Hoaglin, Mosteller, and Tukey 1983) focused on graphical representations of data and summary statistics; thus a contribution of this work is the inclusion of software development and social sensemaking as means of exploring data.

As employed in this work, EDA is a hermeneutic approach (Benton and Craib 2001) in that the analysis of the data is accomplished through the interpretation of iteratively examining different levels and different representations of the data. In other words, understanding a part of the data requires an "understanding of the whole of which it is a part" (Benton and Craib 2001, 104). It follows that interpreting the visualization of the signature of one specific information flow requires interpreting visualizations of aggregations of many signatures, often across multiple dimensions. These dimensions could be distributions of the number of shares, the participating users, the changing rate of the flow, or any other parameters of the signatures. Often the interpretation requires moving among visualizations of many individual signatures and visualizations of distributions in order to understand how one signature is situated in the context of many information flows. The process is not linear. The exploration from one step may be repeated, may lead to other explorations, or be a dead end. Dead ends later may prove useful or

informative in unexpected ways, and iterations of a given step may involve fine tuning after social sensemaking indicates value in doing so. In an effort to provide context around, and insight into the exploration processes, the underlying assumptions, and the interpretation of discoveries, an account of the analysis is presented in first person (see section 2.2.2.2).

This approach to exploration fits in well with Tukey's (1980) forceful call for both a broad general inquiry (exploration) and confirmatory statistical analysis. He emphasizes exploring the data to gain a general understanding of its context (which he sees as informing question development and hypothesis testing) and then exploring the data in more detail as a means of assuring it met the assumptions necessary for future confirmatory tests.

The strengths of EDA are that it can uncover patterns and structures in the data; assist in identifying variables, verifying the assumptions required for later statistical analysis; and suggest transformations to make the data better fit the assumptions (Mueller and Tukey 1980). EDA is also useful, and possibly necessary, in cases where clear and detailed documentation about the data is unavailable, as was true of the Twitter data used in this project. The role of EDA in this research includes (1) assisting in data cleaning; (2) identifying key variables; (3) determining the structure of the relationships between variables; (4) verifying model assumptions; (5) performing model diagnostics; (6) finding temporal patterns in information flows that lead to a generalized signature; and (7) identifying a means to measure changes in network structure.

The results of the analyses from 6 and 7 address the first research question: *How can we learn about the interaction of information flows and changes in network structures from Twitter data?*

There are cautions to heed when employing EDA. Bolker (2008) notes that data-dredging, or specifically looking for patterns in an effort to devise a testable hypothesis, risks

biasing an analysis. He also notes that humans are quite apt at seeing patterns in data that are not actually present. To ensure a systematic approach, he suggests that researchers need a guiding research question, clear reporting, and common sense in the application of EDA. For this work, a clear guiding research question existed from the start: to understand the interaction between information flows and social networks.

2.2.2 Procedures

2.2.2.1 Data

This study employs Twitter data related to the Occupy Wall Street movement. The Social Media Lab gathered the data from Twitter's Streaming Application Programming Interface[9] (API). The Streaming API allows researchers, application developers and members of the public to collect a stream of public *status updates*. A status update is a collection of data that contains the tweet text as well as a rich set of metadata that provides context for the tweet. Twitter streams tweets back if any hashtags, @-mentions, URLs or text within the tweet match any of the search terms sent in the HTTP request to the API. Members of the lab loosely curated the search term list as OccupyTogether.org updated their spreadsheet with new sites and accounts. The search term list, details about collection and the Social Media Lab's collection system can be found in appendix B. An example of a status update in its raw JSON (JavaScript Object Notation) format can be found in appendix C. For convenience, this document shall generally refer to a status update as a tweet or a retweet.

The Social Media Lab's corpus of Occupy data includes 64,298,061 tweets, sent from 12,159,856 users, collected from October 19, 2011 to June 7, 2012. This study uses a subset of

[9] https://dev.twitter.com/docs/api/streaming

these tweets. Specifically, only *retweets* are selected from the data under the assumption that a retweet represents a flow of information from the user who sent the initial tweet to the user who retweeted. For this work, retweets are grouped together as *retweet events* (RTEs). An RTE is constituted from the initial tweet and all of the retweets of that tweet. Thus the concept of an information flow is operationalized as an RTE.

When Twitter users retweet a tweet using the Retweet button, the entire initial tweet is embedded in the retweet's metadata. The metadata is critical for the construction of variables in this study, so only tweets created through Twitter's retweet functionality are included in this study.[10] There are 12,349,759 retweets in the database with the needed metadata. Figure 2.5 shows the daily rate of tweets and retweets that the Social Media Lab collected and table 2.1 provides summary statistics for the daily rates.

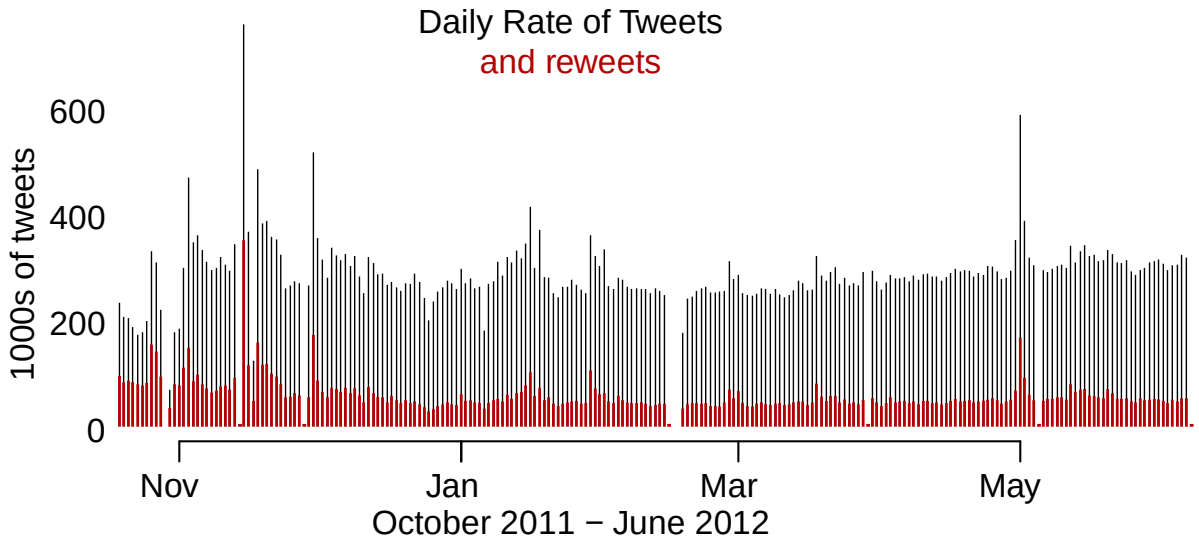


Figure 2.5. Daily rate of tweets (black) and retweets (red) in the Social Media Lab's corpus of Occupy tweets.

[10] While manual retweets, modified tweets (MT) and vias can all be conceptualized as the same information flow, they are excluded from this study because the metadata of these tweets do not contain the initial tweets data within them. See section 5.1.2 for a detailed discussion about the effects of this limitation.

Table 2.1. Summary statistics of daily tweet and retweet rates.

	Min.	1st Qtr.	Median	Mean	3rd. Qtr.	Max	Total
Tweets	8	257,600	280,100	279,600	306,800	755,100	64,298,061
Retweets	1	39,940	45,020	53,690	59,300	346,200	12,349,759

Note that certain dates show sharp drops in the tweet rate. For example, on February 15, we collected 632 tweets, and then collected no tweets on February 16 or 17 before resuming collection on the morning of the February 18. There are similar dropouts on October 29 and 30, 2011; on November 14 and 28, 2011; on March 29, 2012; and on May 5, 2012. System outages, periods when no data were collected or the data were corrupt and never loaded in the Social Media Lab data repository, account for the missing data. The impact of these dropouts, as well as the method for dealing with them, is discussed in the data section of Phase 2.

There are 4,002,284 distinct RTEs. The size of RTEs, as measured by the initial tweet and all of the retweets of that tweet, follows a heavy tailed distribution, such that 2,777,447 RTEs are of size 2, or have only a single retweet. There are 531,144 of size 3, 219,512 of size 4 and so on. The largest RTE in the set has 23,360 retweets. The distribution of the sizes of RTEs is illustrated in figure 2.6[11].

[11] Note that the y-axis is the base 10 log of the frequency of RTEs at the size indicated by the x-axis, also transformed with a base 10 log.

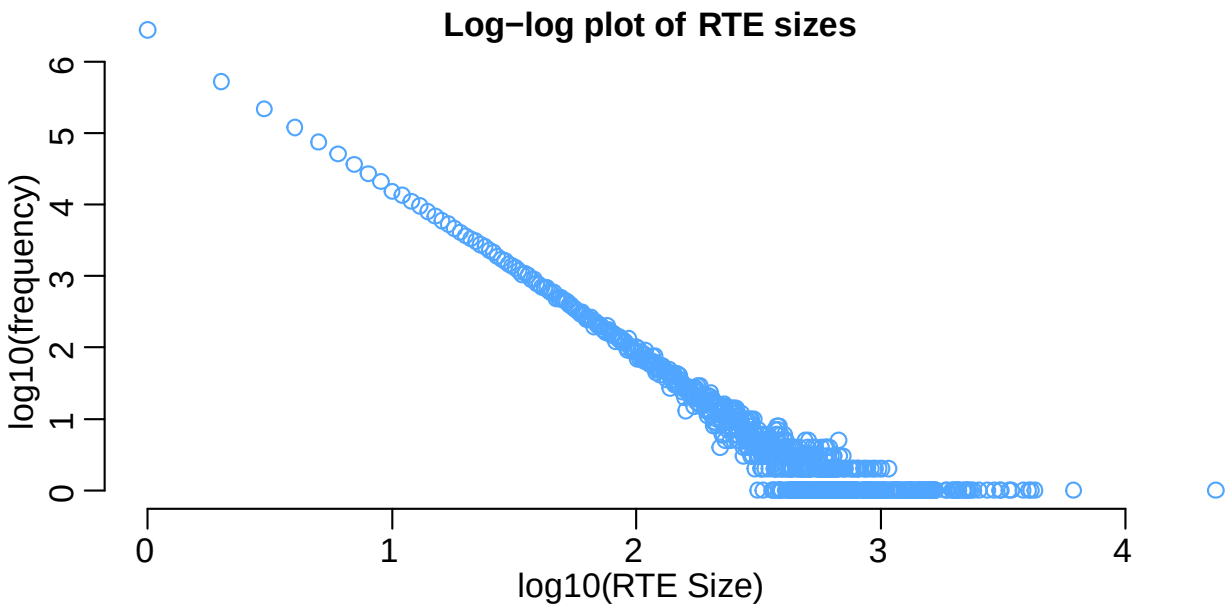


Figure 2.6. Size of RTEs in the Social Media Lab's corpus of tweets. This displays a heavily tailed distribution.

This study excludes RTEs with fewer than 100 retweets. While the cut-off of 100 appears arbitrary, events with fewer retweets are too small to exhibit a pattern[12] or are better approximated using a Poisson distribution.[13] For these smaller events, fitting a power-law to the decay phase of each signature is not useful. This cut-off leaves 5,758 RTEs, made up of a sum total of 1,393,607 retweets.

[12] For example, early EDA for this study showed that about 70% of RTEs consist of a tweet and a single retweet and about 97% of RTEs have 10 or fewer retweets.
[13] When the RTE fits a Poisson distribution the peak, if one is clearly present, is not significantly different from the mean rate of retweets.

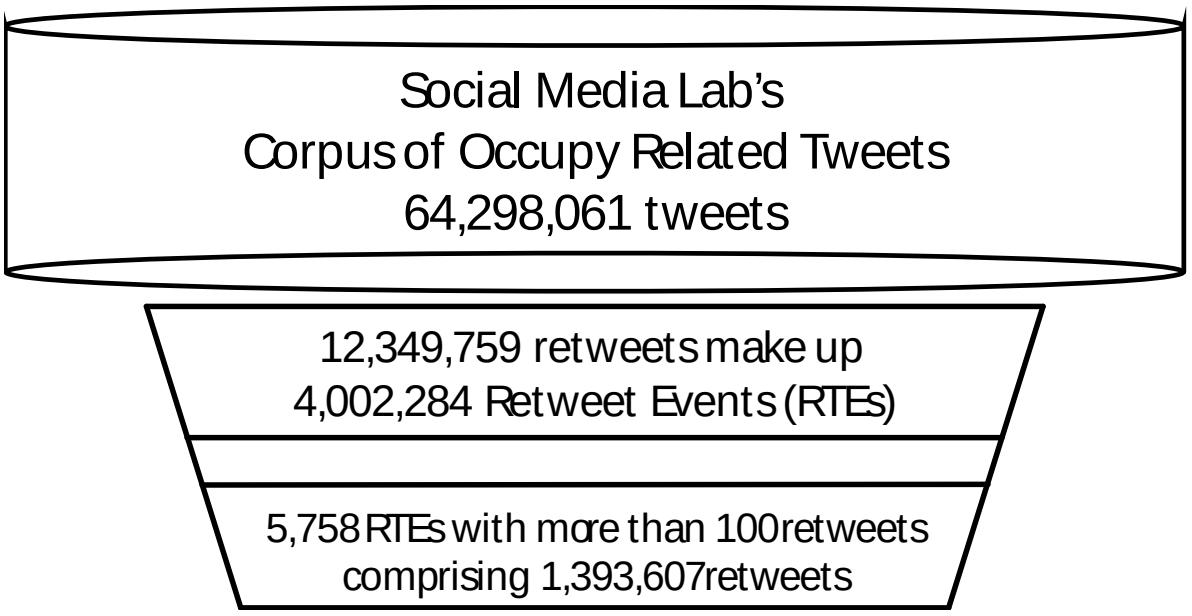


Figure 2.7. Graphical overview of the selection of data used in Phase 1 of this study.

Figure 2.7 graphically represents the narrowing down to the data set as covered in this section. This final dataset is used in the exploratory analysis of data discussed in the next section.

2.2.2.2 Exploration of Twitter Data (Plots that move the plot forward)

> Science … DOES NOT BEGIN WITH A TIDY QUESTION. Nor does it end with a tidy answer. (Tukey 1980, 24) (emphasis original)

In my early attempts to understand how the flow of information and networks were related, I explored *network plots*, a kind of data visualization that typically shows individuals as circles and the relationships between them as lines. I started this work by examining a few examples of raw status updates. The data were in JSON format, so I could open and view a tweet or retweet in a simple text file program. Twitter's API documentation page provided some information on most of the fields, but in general the documentation was vague and often

incomplete. However, it was clear that I could construct *retweet networks*[14] from the metadata.[15] The nodes in a retweet network are the users who tweet and those who retweet them. Each link in the network is a connection constituted by one user retweeting another. Since Twitter users often include hashtags in their tweets, it is easy to bound these networks for comparison. For example, I might compare the retweet network for tweets containing #OccupyOakland with a network of tweet containing #OccupyHouston.

Figure 2.8 illustrates such a network (additional examples are in Appendix E). The purple dots represent users who tweeted and were retweeted or just users who tweeted. Each green line represents a retweet relationship. Summary information about data used for the network, the network, or the layout algorithm is printed on the visualization. The visualization layout creates distance between nodes (users) based on the number of links exist between them. A byproduct of using this layout with this data is that retweets in RTEs are generally grouped together; notice the group of nodes on the middle left.

At first it seemed like studying retweet networks made sense as a way to understand how information flows and networks interact. Certainly sets of retweets could be conceptualized as an information flow. Groupings of retweets can easily been seen in figure 2.8. In fact, this is where the idea came from to use RTEs as a unit of analysis. By comparing different hashtag and time-bound networks some patterns did emerge. For example, the vast majority of RTEs are trivial,

[14] A retweet network is similar to a citation network used in bibliometrics (Borgman and Furner 2002) in that a retweet can be thought of as a kind of influence (Bakshy et al. 2011). However, there are many differences as well. For example, a retweet is an exact restatement of a message, whereas a citation is a pointer to another work. In addition, a citation network can develop over years, but a retweet network can form in minutes or hours.

[15] The metadata fields are *user.screen_name*, which is counterintuitively the name of the user retweeting, and *retweeted_status.user.screen_name*, which is the name of the user being retweeted. If the second field is not present in the metadata, then the status update is a tweet, not a retweet, and as such, *user.screen_name* is the name of the user who posted the tweet. Because the focus is on retweets as links, this type of network is referred to as an *arc sampling design* (Butts 2008), where arc is a term for a directed edge (A links to B, but B may or may not link to A). In the Arc sample design the network is instantiated from the observed directed links, thus the focus is on the connections or flows rather than the nodes or actors.

containing a single retweet. Additionally, a small set of users initiate the majority of the larger RTEs; meanwhile, different city hashtags at different times display greater or lesser degrees of interconnectedness among users retweeting each other.

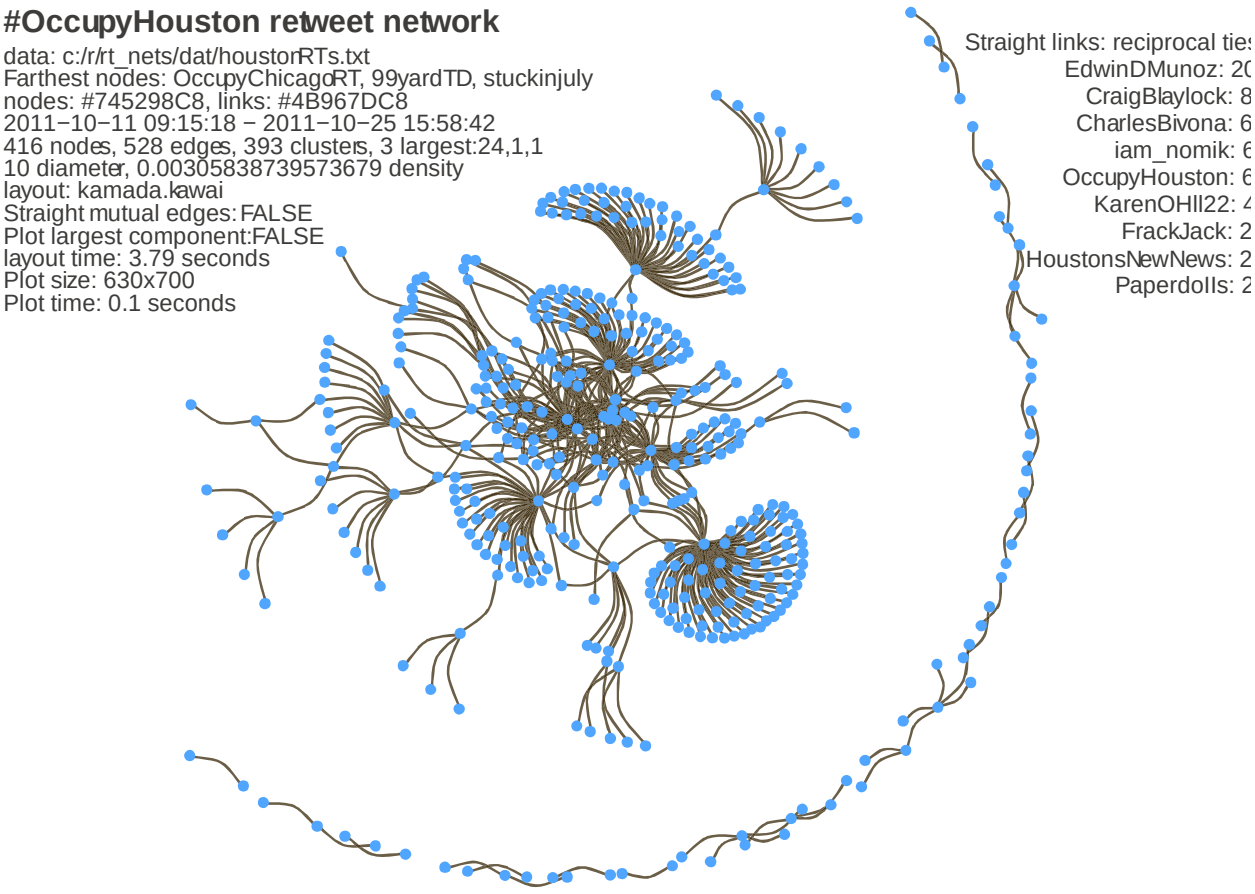


Figure 2.8. Example of a network data visualization of a retweet network using the Social Media Lab data.

Ultimately, these visualizations failed to provide me with an understanding of how the flow of information interacted with networks[16]. However, the process of using these visualizations to explore the data yielded surprising insights into the data itself and, importantly, into how writing software to create the visualizations and how discussing them with others also

[16] For additional information see *Algorithmically supported sense-making of network visualizations: a call for reflexivity and transparency, by Hemsley*, in *Algorithmically supported sense-making of network visualizations: a call for reflexivity and transparency* (Markham et al. 2013).

resulted in a deeper understanding of the data. There are two avenues of understanding the data I want to highlight.

First, whenever I printed one of these network plots and shared it with a colleague, they would inevitably ask what it meant, and a discussion would ensue. In the process of my explaining the plot, we would interpret the meaning based on what data was represented, and then come up with ideas about how to make the visualization easier to understand. We would also question what else might be in the data that could be helpful in understanding how the networks and flows interacted. This social sensemaking of data visualization prompted me to explore aspects of the data that I may never have thought to explore on my own. After each of these sensemaking sessions, I would return to the data and try alternate data fields, or network constraints for the next iteration of plots. Repeatedly going back to the data meant learning more about Twitter's rich metadata and how it might—or, more often, might not—be used to answer my questions.

The second avenue to understanding involves writing software. Over the course of 2013, I wrote software scripts to plot thousands of network visualizations. I used different data; developed different algorithms to layout the networks; and tried assigning different data fields to the visual attributes of the plot, such as the color and size of nodes or links. The process of writing software to work with the data confers knowledge about the data. Each new field prompted me to revisit the data, to explore a field's range, to aggregate and visualize it in different ways. Doing this over and over with minor changes required that I write software to automate the processes of extracting sets of tweets from the database, importing the data into R, processing the data to create a network and process any additional fields assigned to the visual attributes of the network plot. But software often has *bugs*. These defects may be typos or logic

errors, or they may reflect the programmer's *assumptions* about the data being processed. Suppose I make a network plot and size the nodes by how many followers someone has or how often they tweet or are retweeted in the data. Unless I account for the fact that the distribution of these variables can be skewed in unexpected ways (see figure 2.9), some nodes will be so tiny they will be unobservable, while one or two others will overwhelm the plot. Finding and fixing these assumption-based bugs actually provides insight about, for example, the distribution of the data.

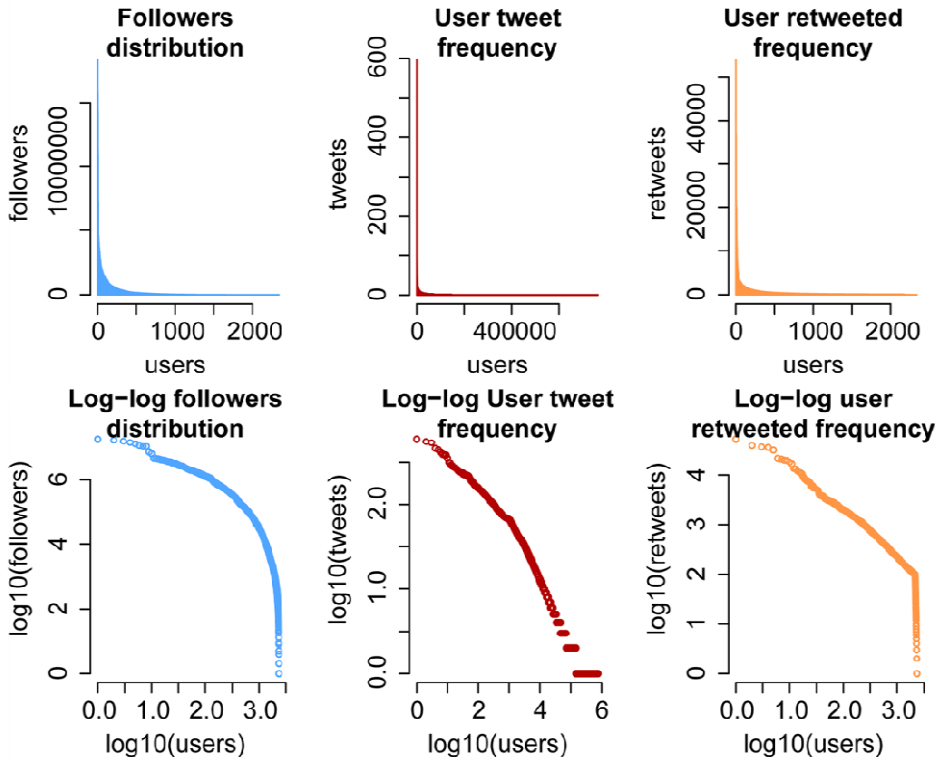


Figure 2.9. Example power-law distribution plots. Uses Social Media Lab data.

Of course the process of organizing, exploring and comparing many network visualizations was challenging. Printing out an initial network visualization that I would almost certainly throw away did not make sense. To help me manage the process I purchased a software

program[17] that allows photographers to sort through and compare large amounts of photographs. Figure 2.10 illustrates the use of this program to compare visualizations. A byproduct of working with a large set of data visualizations is an understanding of how to quickly make and explore very large collections of visualizations.

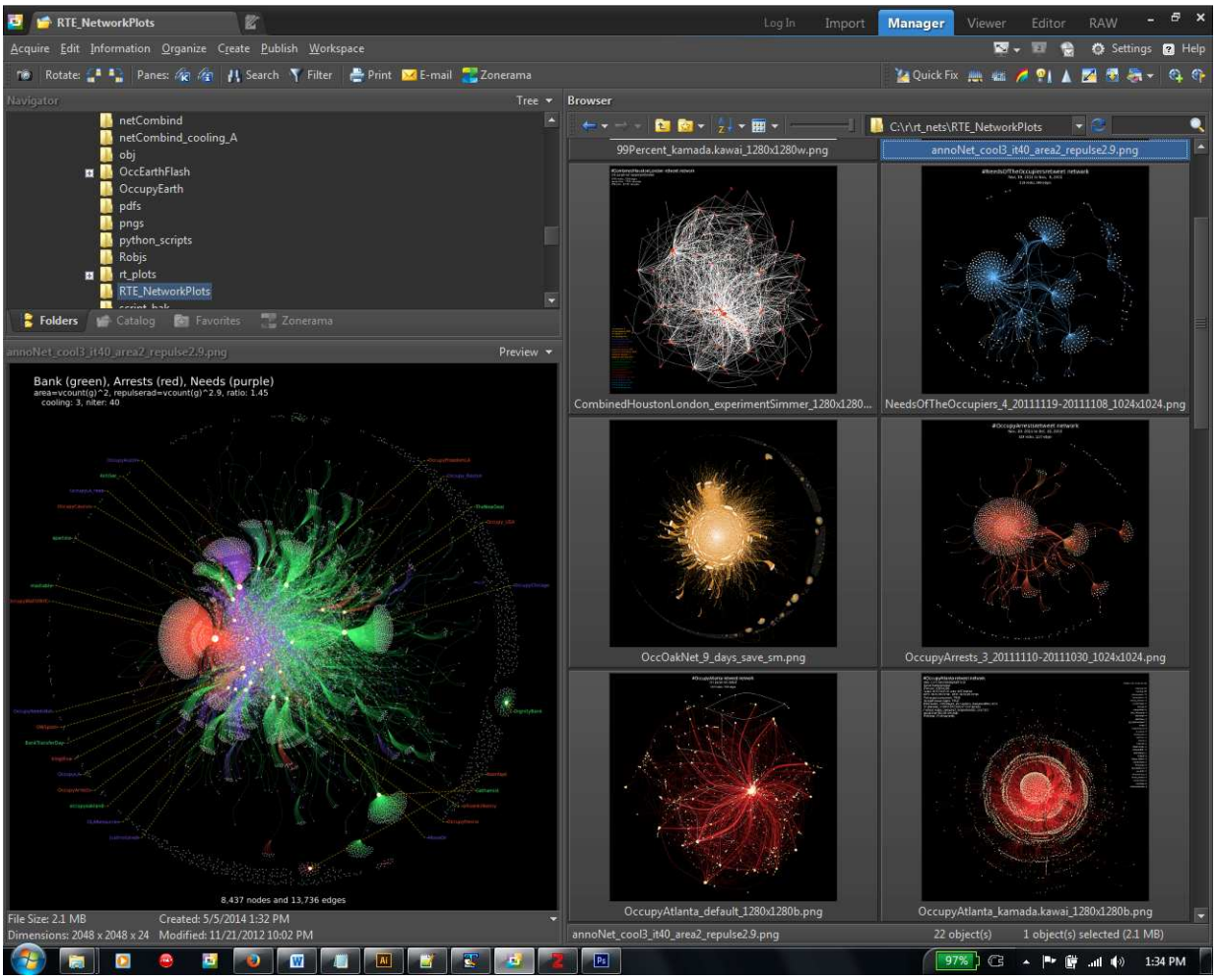


Figure 2.10. Using Zoner Photo Lab to explore network data visualizations.

When I felt I'd exhausted the possibilities for network visualization I began exploring ways to look at the data from the perspective of time. Fortunately, I already had developed the software and processes needed to explore a large set of visualizations. For example, I plotted out

[17] Zone Photo Lab https://www.zoner.com/

the rate of retweets per minute, or the RTE signature, for all 5,758 RTEs, and used Zoner Photo Lab to look for patterns. Figure 2.11 contains an example of such a plot. The blue line is the rate per minute with the peak volume in blue text at the top right. These features are bolder so that they are visible and comparable even when visually scanning sets of thumbnails of the signature plots. Additional details, like who initiated the RTE and when it started, are in smaller text. These smaller features are not visible on thumbnails, so only the salient information is compared. In this case, I was looking for relationships between the height of the peak and the shape of the decay phase of signatures.

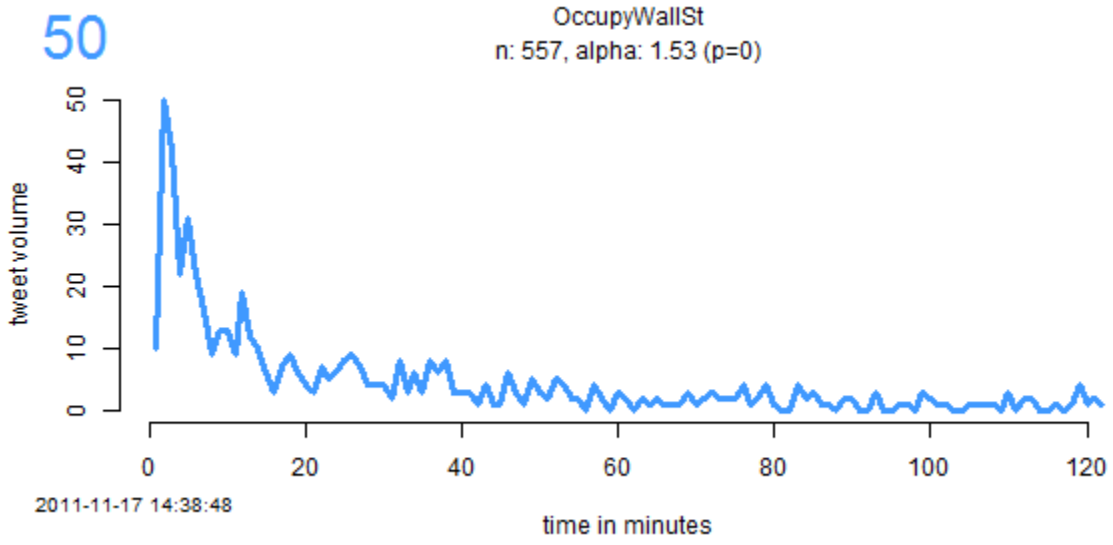


Figure 2.11. Plot of the rate of the retweets of a specific tweet over time in minutes. The blue line and large number 50 are easily seen when the plot is zoomed out to compare with other signatures. as Social Media Lab data.

An iteration of the signature plot, seen in figure 2.12, shows the rate of tweets per minute for an RTE and includes additional information from the RTE metadata. As before, the rate of tweets per minute is given by the blue line, but the red stars plotted over that represent additional information for each retweet. The position of the star on the x-axis indicates when the retweet

happened in time, while the position on the y-axis shows the change in the number of followers *for the user who initiated the RTE*. In other words, the height of the star indicates the number of followers gained or lost from the first tweet to the retweet. The star is sized by the number of followers of the person who posted the retweet, and the black line is the estimated power-law fit to the decay phase of the RTE.

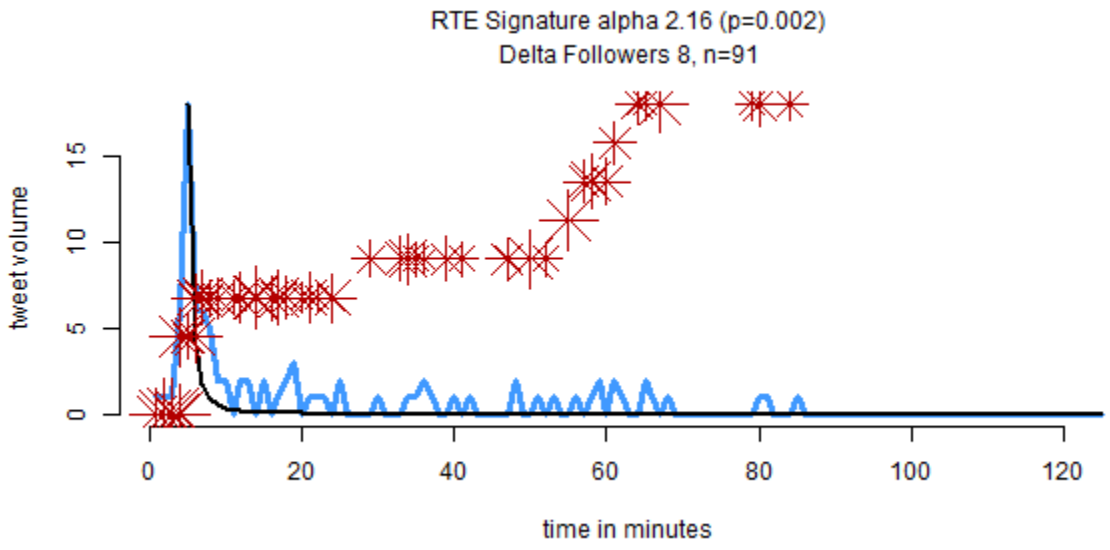


Figure 2.12. Plot relating rate of retweets to changes in followers. Bold features (red stars and blue line) allow comparison across a large number of plots. The details are still present as smaller text but only visible when zooming in. Social Media Lab data.

Both of the previous two illustrations provide an example of how I made plots for large-scale exploratory work. Features that I want to compare across many plots are bold or use heavier line weights, while details are small. The plots are otherwise free of clutter that could interfere with the analysis. The bold features stand out when scanning many thumbnail sized plots, allowing comparison of many plots at the same time (figure 2.13). If I see interesting

features in a particular plot, I can zoom in to see the detail text. Since these are exploratory plots, the audience is assumed to be me, the researcher. Legends or other information that might be present for presentation plots clutter the plot and make comparison more difficult. Given the audience (me), some short cuts can be made. For example, in figure 2.12, the change in followers is represented by the y-axis position, but it is not to scale. The caption at the top of the plot indicates that the initial user gained eight followers, but the highest star is over 15. I was looking for patterns in the change in followers for the initial user, and I didn't care as much about how great the change was. So the trend is bold (the position of the stars) but the detail (the number in the caption) is small. This exploration led to the observation that the initial user's followers almost always go up during RTEs.

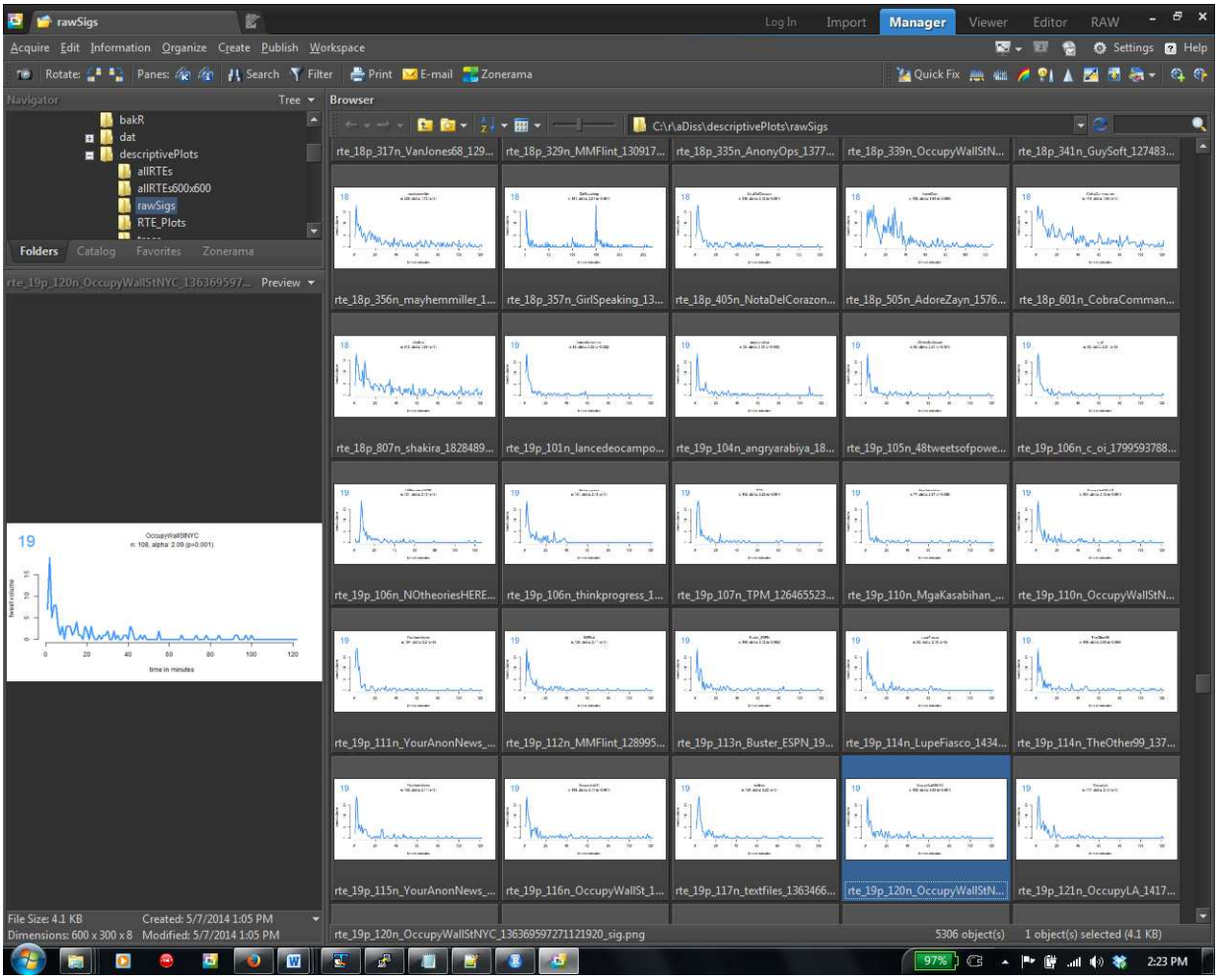


Figure 2.13. Example of zooming out to compare many RTE signatures at the same time.

Figure 2.12 illustrates another important and interesting observation. The pattern of stars in this particular plot shows a rise in followers that is not apparently associated with an increase in the rate of tweets per minute. This led me to go back to the data to probe a bit deeper. The result of this was the recognition that RTEs initiated from a single user can overlap in time, and that this overlap is salient to what I was seeing. For example, the later rise in the stars in figure 2.12 is related to a second RTE that was initiated by the same user about 25 minutes later.

Figure 2.14 and figure 2.12 together demonstrate how interpreting the visualization of the signature of one specific information flow sometimes requires interpreting visualizations of

aggregations of many signatures. Figure 2.14 shows the overlap RTEs for the user OccupyWallSt, one of the most prolific tweeters in the data set. The x-axis displays time in minutes, starting from midnight, September 17, 2011. The y-axis shows the number of the RTE for the OccupyWallSt user, sorted from bottom to top by when the RTE was initiated. Each line starts on the x-axis at the time the RTE was initiated and ends when the last retweet was sent. Each line is a semi-transparent red, so when they are stacked on top of each other they are darker. The darker red areas show times when RTEs overlapped. This observation made it clear that I would need to control for this overlap effect in confirmatory models.

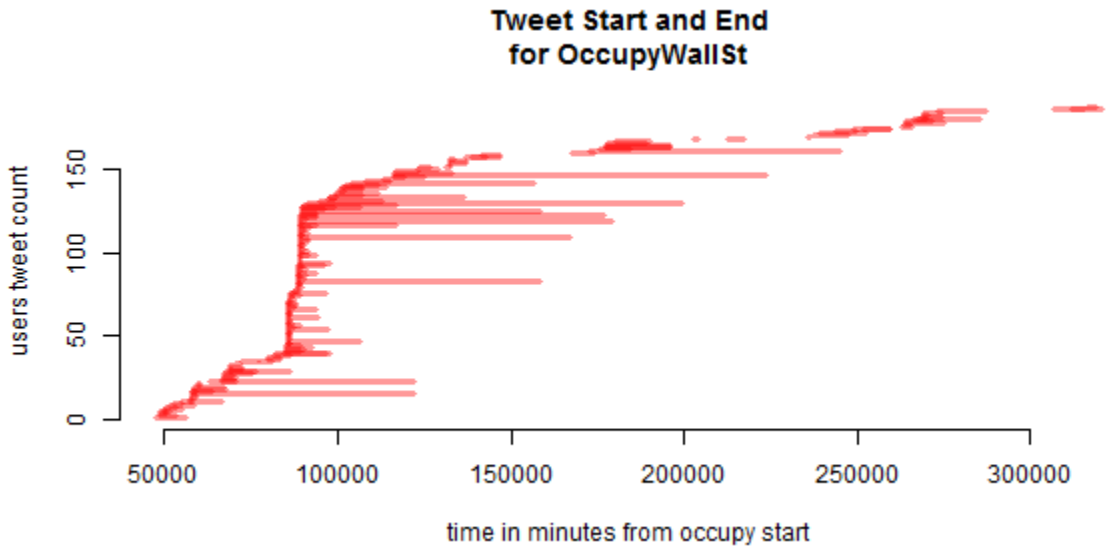


Figure 2.14. Example of RTE overlap. Social Media Lab data.

Another observation about this plot is that some RTEs can have very long delays to the last retweet. Figure 2.15 provides an example of one such RTE in which the final retweet in the dataset was posted 26,617 minutes, or 18 days, after the initiating tweet. This observation led to the bounding of RTEs to a uniform window in time around the peak of the event for the confirmatory analysis in Phase 2. Section 3.2.2 provides more detail on this.

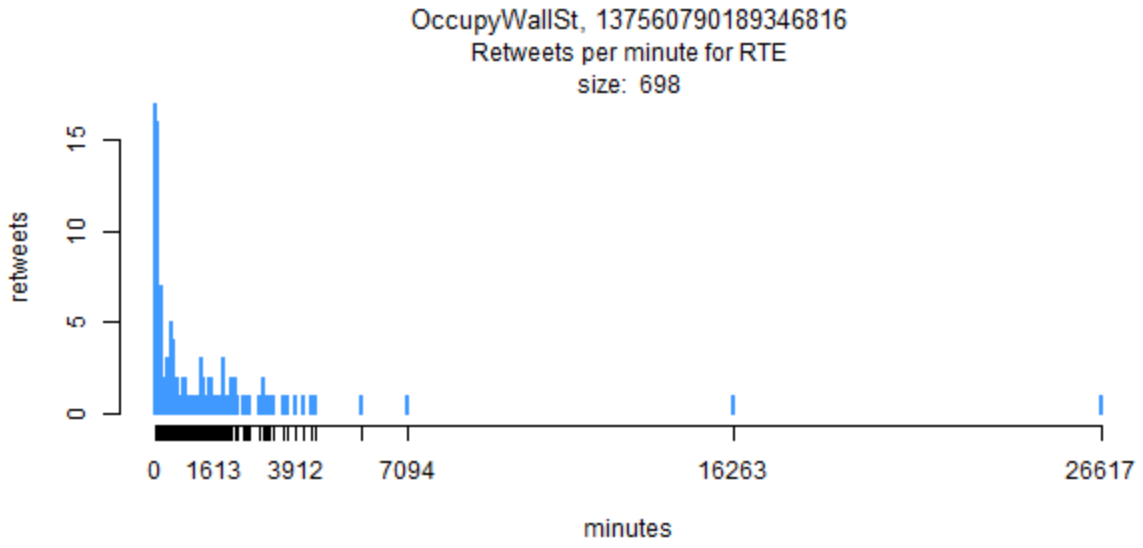


Figure 2.15. An RTE initiated by the account OccupyWallSt with a long delay retweet. Social Media Lab data.

I want to leave you with a sense of scale and scope of this work. Iteratively plotting out thousands of images means I have explored many thousands of visualizations. As of this writing, I have 130,040 .png plot files, taking up about 2.3 gigabytes of disk space on my hard drive. In many cases the exploration of a batch of plots results in a new set that supersedes the previous batch. As such, I have deleted many thousands of files. I have begun to think of my visualizations as my data (or part of it, or a representation of it); deleting them seems almost sacrilegious. But because the iterative process is often about finding the right lens to explore the data, deleting plots makes sense in this context. Finding the right lens might mean matching it to the right time scale. For example, figure 2.15 isn't as informative as figure 2.11 for scanning for patterns around the peak when looking at screen full of plots. However, the former is better for understanding the context of a specific RTE and how it might overlap with other RTEs. Finding the right lens might also mean finding a set of colors and symbols and font sizes that supports scanning many plots and drilling down for detail.

In this work I have explored a subset of the Social Media Lab's corpus of 64,298,061 tweets. In addition to observations about the data and the findings (discussed in the next section), I have developed a set of tools and approaches to visualizing and interacting with data visualizations that supports the exploration of the data from which the visualizations derive. Social sensemaking, the process of working iteratively with others to create and interpret data visualizations, proved to be an effective way to navigate a large unstructured dataset and gain insight about data. Attentiveness to the assumptions built into the software also provided insight into the data. These insights included the ways in which the data in different fields are distributed, what potential relationships exist in the data, what kinds of answers the data supports, and the quality of those answers.

2.3 Findings

The exploratory phase yielded two findings. The first of these is a stylized model of information flows defined by parameters identified in empirical data in the form of RTEs. The second result is an operationalization of changes in networks that provides a way to quantify structural changes in networks without specific link information. These two findings provide the means for statistically testing relationships between structural changes and the parameters of information flows.

2.3.1 Parameterized Signature Model

As discussed in the methodology section, an important avenue of data exploration for this project involved looking for patterns in the rate of retweets for RTEs (see figure 2.11). Instead of reproducing thousands of plots here, figure 2.16 shows a normalized sample of 1,000 RTE

signatures.[18] The figure shows all 1,000 normalized signatures plotted as a semi-transparent red line, one on top of another. The result shows the general pattern of a fast ramp-up in the rate, followed a wide variation in the rate of decay. The range between the first and third quartiles for the rate per minute is highlighted in orange, with the median represented as a dashed line and a thick blue line showing the mean.

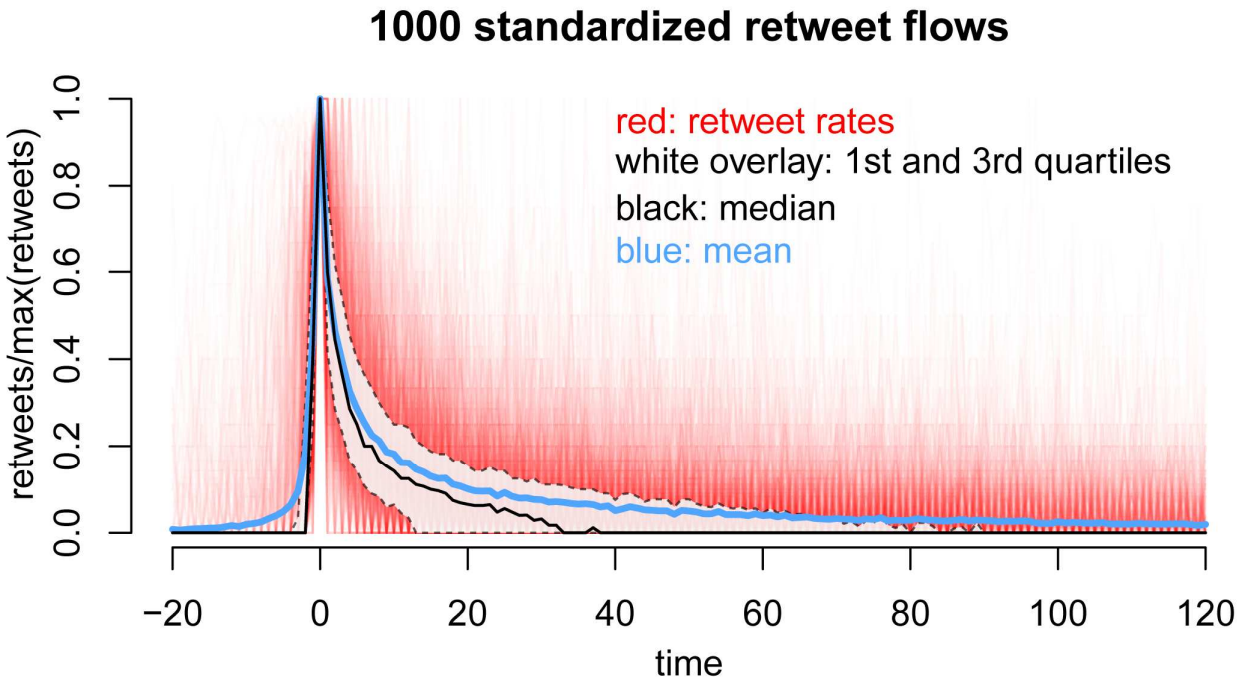


Figure 2.16. A sample of 1000 information flows from the Occupy corpus of tweets. Social Media Lab data.

The patterns seen in figure 2.16 are captured in the stylized model of an information flow seen in figure 2.17. The model provides nomenclature for parameters related to rate of flow for an RTE signature. The model also specifies two main temporal phases, ramp-up and decay, which correspond respectively to the period of time from the first tweet to the peak rate of retweets, and from the peak rate to the last retweet. These phases can be measured or used

[18] The sample was drawn from the set of 5,758 RTEs with over 100 retweets. Each RTE in the sample has been normalized by finding the maximum rate of tweets per minute, scaling it to 1, and dropping out tweets and retweets occurring more than 20 minutes before or 120 minutes after the peak.

generically in discussion about an information flow[19]. Table 2.2 identifies the parameters of the model and provides nomenclature and the type of measurement.

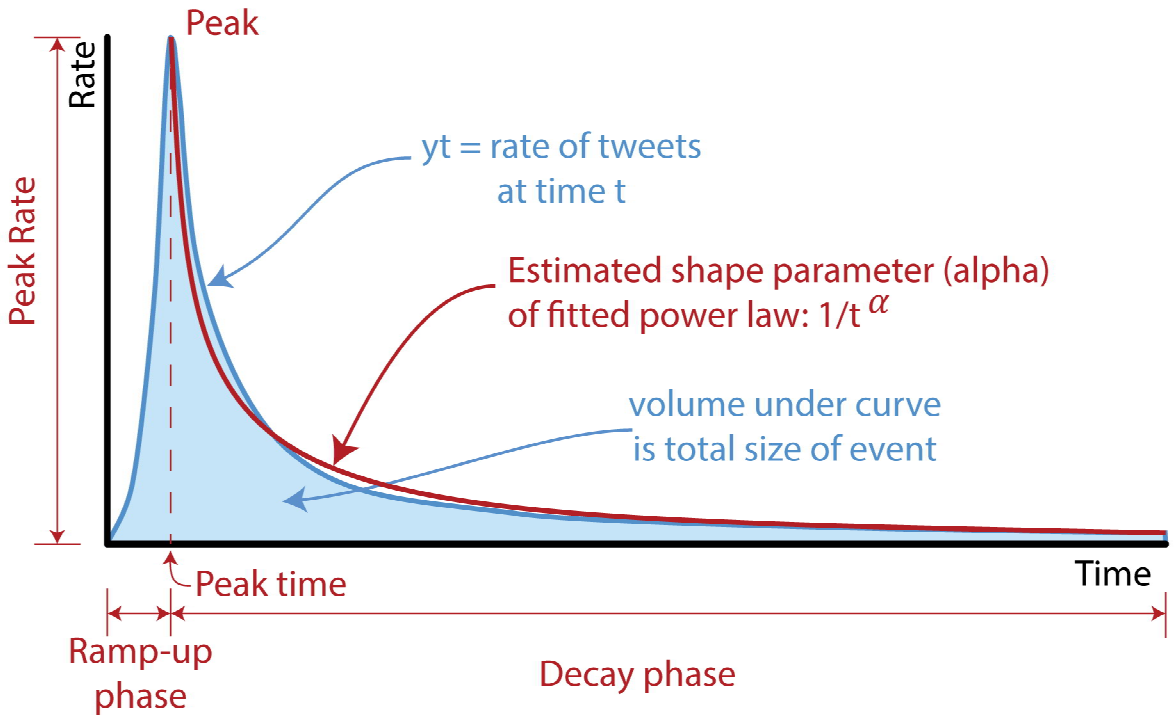


Figure 2.17. Information flow signature model showing the phases of its life cycle (ramp-up, peak, and decay phase) as well as quantifiable characteristics (peak time, peak rate, and its shape given by alpha).

Table 2.2. Quantifiable parameters of an information flow.

Parameter	**Description**	**Type**
Peak Time	The period of time from the original tweet to the point in time when a peak rate of retweets is reached.	Time in minutes. Integer / real
Peak Rate	The maximum rate of retweets for a given time period (typically in minutes).	Integer
Decay Shape (or *shape*)	The shape parameter alpha given by fitting a power law to the decaying rate of retweets over time.	Real number
Size	Total number of retweets.	Integer
Decay phase	A measurement of time from the peak to the last retweet.	Time in minutes. Integer / real

[19] A number of limitations of the model are discussed below.

This is not the first examination of temporal aspects of information flows. Both Crane & Sornette (2008) and Broxton et al. (2010) studied the daily rate of views of YouTube videos, but in both cases they normalized the peak rate of views to 1. While this made the ramp-up and decay phases easily comparable across many videos, the scaling makes exploring the relationships among the peak rate, size, and shapes of the decay and ramp up phase impossible. Since the values of the parameters of the signature model remain intact, they are each individually well-suited for inclusion in statistical models. This allows the exploration of the relationships among the parameters of the model. In addition, while Nahon and Hemsley (2013) used the term *signature*, they do not formalize a model defined by parameters identified through the exploration of empirical data.

Note that the model is limited in a number of ways. Key among these limitations is that the model only represents the changing rate of observable information flows. It does not represent information flows that cannot be detected. For example, an individual may receive information from a tweet and then pass that information to another individual during a phone conversation. The model does not represent other dimensions of the flow of information, such as its path through a network, its flow from one geographic location to another, the type of information, or the impact of the information. However, the model could be used to discuss how different categories of information, within the dimensions just listed, behave temporally differently.

The model does not account for information flows with multiple peaks or flows where no clear peak exists. In this study, very few RTEs had more than a single peak. In those cases the analysis focused on the highest peak, but it would be reasonable to treat these cases differently. For example, one might break multiple-peak RTEs into separate cases for analysis, each with its

own peak. Alternately, one might consider them a different category of event to be studied separately. The vast majority of RTEs in the Social Media Lab corpus of Occupy tweets are small enough that a peak would not be discernable. Using the model to examine these flows would not be appropriate.

2.3.2 Operationalization of Network Changes

Another important finding resulting from the EDA is that it is possible to quantify, and thus measure, changes in the linking structure of Twitter's social network by tracking the change in the number of followers a user has from tweet to tweet or retweet to retweet. The number of followers a user has is the number of other users who link to them. Each time someone tweets or is retweeted, the tweet metadata records not only the number of followers they have, but, crucially, how many followers they have at the time of the tweet/retweet. With a set of a user's tweets, and the retweets of their tweets, it is possible to track the change in the number of other users who link to them over time. As an example of this, figure 2.18 shows the rise in the number of followers of the OccupyWallSt during the course of time the dataset covers. Each data point represents the number of followers (y-axis) the OccupyWallSt account had when the tweet/retweet was sent (x-axis).[20] Figure 2.12 illustrates an example of tracking the change in followers of the initiating user over the course of a single RTE.

Of course, there are limitations to this method. A user can gain or lose followers at any time. When there are long temporal gaps between data points, all that can be observed is the *net change*, not the exact number (or point in time) of users who linked or unlinked over the time

[20] Note this figure represents the data in this set, that is, the tweets and retweets contained within RTEs with over 100 retweets. The graph would be much denser, and likely be without gaps, if it had been created from all tweets/retweets in the Social Media Lab Occupy corpus of tweets.

interval.[21] This means that given two retweets, if five users added links and three users removed links, all that is observed is a net change of two added links. In addition, this approach provides no information about which users are adding or removing links, only the number. Different actors in a network can have dramatically different numbers of connections, locations in the network relative to the core, or any other factors. Thus, the links that added or removed are not necessarily equal. Tracking the change in followers of users who are retweeted provides information about the ego network for that user. Of course, when a user on Twitter has over 85,000 followers, such as the OccupyWallST account did at the beginning of this study, it is assumed that changes in their ego network can impact the structure of the wider network by increasing or reducing the distance between actors who follow the account. Certainly it would be ideal to know the exact link structure of the network before the flow, at each step of sharing during the flow, and after the flow has stopped. However, the status updates returned from Twitter's streaming API do not contain lists of a user's followers, only the number.

[21] Note that I also found that I could create visual animations of retweet networks, which suggests that there are many ways to visualize and measure changes in networks. However, these explorations did not provide a means for using confirmatory statistics to explore the relationships among the signature model parameters and social networks. Examples of these network data animations can be found at https://www.youtube.com/watch?v=QHi8Bn6UTxQ and here: https://www.youtube.com/watch?v=XX9he5lkNSo

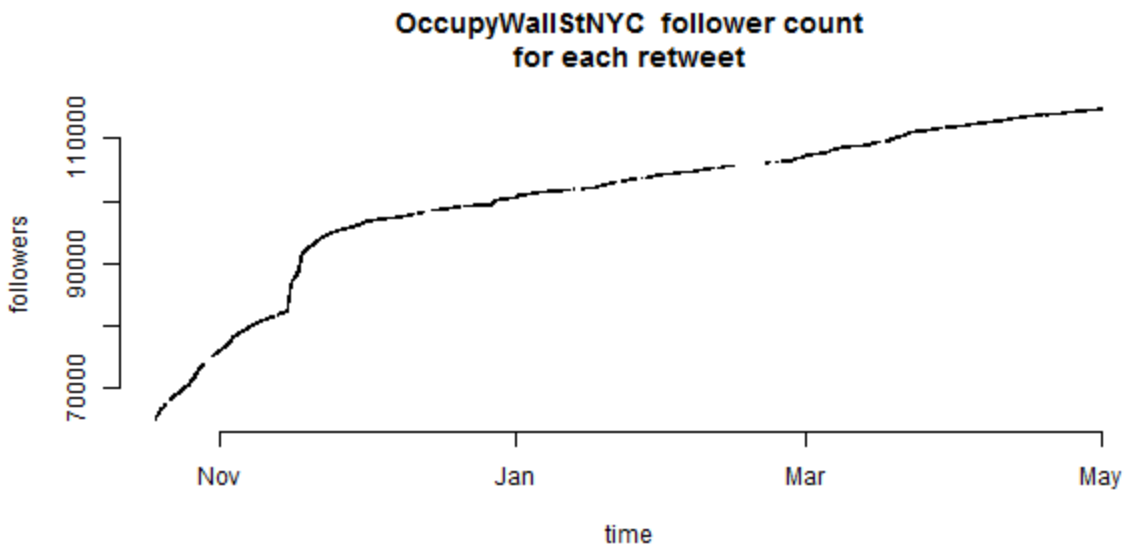


Figure 2.18. The rise in the number of followers of the OccupyWallSt account. Social Media Lab data.

The research question for this phase was, *How can Twitter data be used to explore the relationship between the flow of information and social network dynamics?* The findings from this phase include the signature model and an understanding of how it is possible to measure changes in the Twitter social network using the Social Media Lab corpus of Occupy tweets. Since the signature model includes a number of parameters that can be quantitatively measured, and since tracking the change in a user's followers is an approach to quantitatively measure change in Twitter's social network, confirmatory statistical methods, such as regression, can be used to explore the relationships among the changing network and the measure of the signature model. Such an analysis is the topic of the next chapter and is the phase of this study.

3 Phase 2: Network Dynamics and Information Flows

This chapter covers the literature, methodology and findings of Phase 2, which focuses on answering the overarching question, I*n what ways do information flows and social networks interact?*

The chapter first examines prior research along two lines: (a) the relationship between the *structure* (topology) of social networks and *information flows* through these networks; and (b) the *dynamic nature* of this relationship—how the structure may change as a result of these flows. The analysis of this phase focuses on creating parsimonious models that confirm (or refute) the relationships between the parameters of the signature model and social network dynamics with which the information flow interacts. Specifically, the second phase employs variance models to (1) understand how the path of sharing through which the information flowed is related to the parameters of the signature model; and (2) understand how the parameters of the signatures of information flows are related to changes in the social network.

3.1 Related Literature

3.1.1 Network structure and information flows

This section summarizes research that examines how the structure of networks is related to information flows within those networks. This section also addresses important gaps in the literature and introduces a research question intended to address these gaps. The second half of this section discusses the expectations for the answer to be based on related literature.

Bampo et al. (2008) examined the propagation of marketing messages in static network models. They found that a scale-free network structure, where the distribution of links follows a power-law (Barabási 2002), outperformed other models (specifically, the Erdos random link model or Watts' 2004 small world model) where the average number of links between nodes is low. In another example, Kitsak et al. (2010) examined core and periphery structures, where core refers to a central cluster of densely linked nodes that is central in connecting many other, often

outlaying (periphery), clusters to the network. This kind of work has focused on network-wide structures on static networks without looking at smaller scale, or local, structures.

Leskovec et al. (2007) looked at the flow of information in a blog citation network made up of 2.2 million blog posts, generated by 45,000 blogs, from July to September 2005. They operationalized an information flow as an initial blog post followed by the citations of other bloggers linking back to the first post or to an intermediary post. In other words, if a story originated at blog A and then blog B cited the post on A, that would constitute an information flow. Further, if C cited either A or B, then C, along with the link to A or B, would be part of the same information flow.

For each information flow, Leskovec et al. mapped the topological linking structure, finding a number of patterns in the way information flows in the blogosphere. The trivial case, where a post received no citation, made up 97% of the flows; the next most common, making up 1.8% of the flows, was a dyad, in which a blog post received a single citation link. The remainder consisted of *stars*[22] or *trees*, where trees are made up of *sharing chains*, or *chains*. Figure 3.1 illustrates these topologies. A star is comprised of an initial post and other blog posts, all of which cite the initial post. A tree will be a star with one more additional posts that cite an intermediary post, or a post between themselves and the origin post. A chain exists as a specific path within a tree. The red links in figure 3.1 are chains of sharing comprised of A → C → F and A → E → H → I. Note that these topologies represent a single flow and are a special case of hierarchical network topologies because the links rarely, if at all, link across or backwards in the

[22] This study adopts the usage of the term star from Leskovec et al. (2007). That is, a star is a structure where a single central node is connected to all other nodes, and each of those other nodes only connect to the central node. This is consistent with the definition put forth by Wasserman and Faust (1994), see page 178. The central node in a star formation is called a hub. In the information flows of Leskovec et al.'s work, a star is a directed network where the hub initiates a flow that others cite in their blog post, but that goes no further. In other words, no one cites the bloggers who cited the hub. However, Leskovec et al. note that most of their cascades fall somewhere between this concept of a perfect star formation and a perfect chain.

hierarchy. In addition, the observed flow does not capture cases where H received a message from both D and E, only that H cited E.

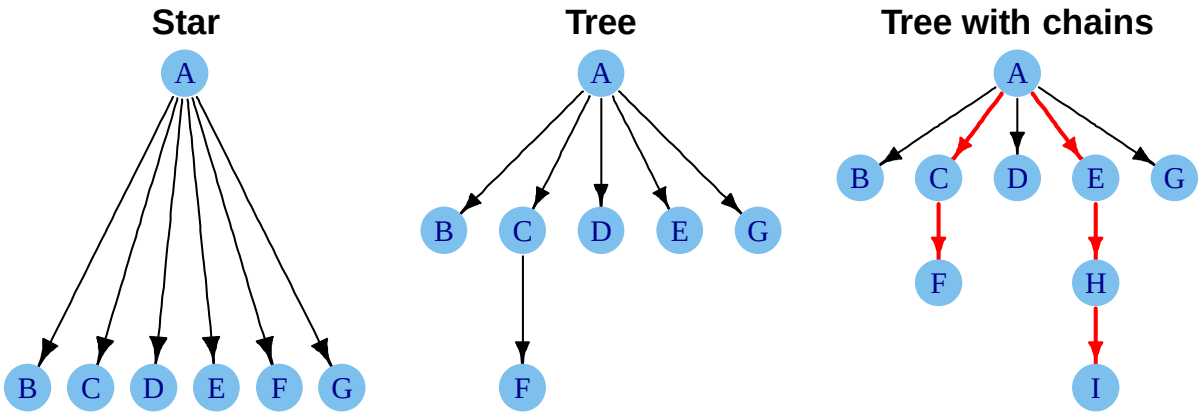


Figure 3.1. Information flow topologies in the blogosphere. Blog A initiates a flow that may result in different topologies. This figure depicts star and tree topologies as well as sharing chains in a tree.

Leskovec et al. (2007) found that star topologies were more common than trees and that the length of chains followed a power-law: the majority of chains were short (A → C → F) with few longer sharing chains. The predominance of short chains has been found in other research as well; for instance, Leskovec et al. (2006) studied the user recommendation network of what they termed a “large on-line retailer” and found that information cascades tended to be shallow, with only a few events reaching deep into networks. In 2009 Kwak et al. (2010) crawled the entire Twitter follower network of 41.7 million users and 1.47 billion links, matching the data up to over 4,000 trending topics, finding that the height of trees, of the length of the longest chain in trees, and the size of cascades all followed a power-law. In an alternate approach, Bakshy et al. (2011) studied influence on Twitter by tracking the diffusion of 74 million URLs; again, the finding was that the length of chains, as well as the size of cascades, followed a power-law.

The work discussed above highlights that large sharing events and events that reach far from their source are rare. However, none of these studies looked for relationships between the lengths or frequencies of chains and temporal parameters of information flows, like those in the signature model. In order to measure how star-like an information flow is, this study uses *closeness centralization* (Wasserman and Faust 1994), a network level measurement of how close actors in a network are to each other (see Appendix D for more detail). Note that the path of an information flow is a network, or, more precisely, a sub-network[23] in a pre-existing network comprised of who reads, who follows, and who is friends with whom. The path of the flow is but one sub-network in a great number of possible sub-networks, where that path is instantiated by the information flow. The path of the flow is assumed to be a snap-shot of part of the larger network at the point in time the information flow began.[24] Calculating the closeness centralization for the information path of each RTE provides a way to statistically examine the relationship between the path of the flow and the parameters of the signature model.

[23] Technically this is a *sub-graph* of the larger *graph*. For ease of reading, sub-network and network are used throughout this document.

[24] Technically the network can change during an information flow, meaning that the potential paths can also be altered during the flow. However, such a change would not result from the information flow; for the information flow to change the network, the information would need to arrive before the change in the network, not afterward as is the case here. Alternately, we could assume that the path represents the state of the network at the end of the flow. This risks a greater misrepresentation, because the flow itself could cause un-linking partway through the flow and break apart the path. Both conceptions present challenges, but the first is preferable for its simplicity.

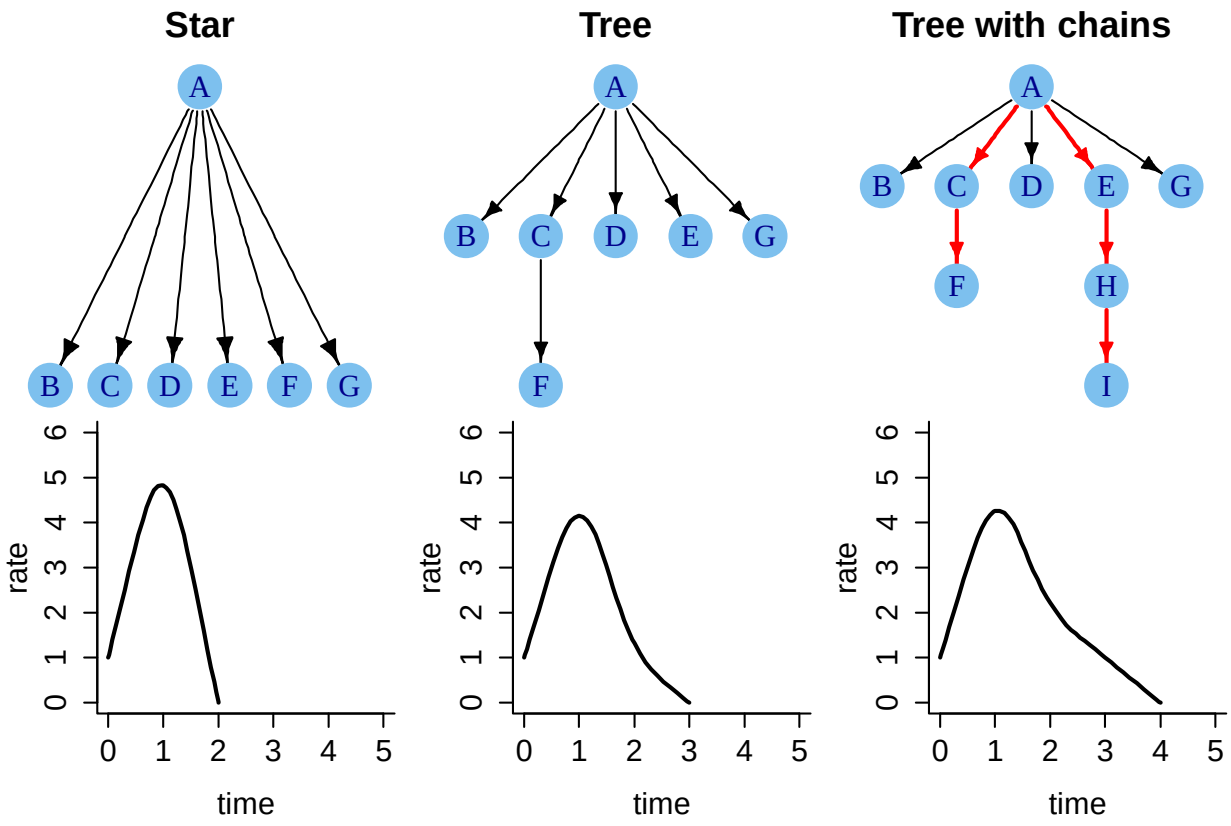


Figure 3.2. Example information flow topologies (above) with related signatures (below), where the wait time from receipt of a message to sharing it is held constant.

The concept of information flow sub-networks in the forms of stars, trees and chains gives some clues as to how the path through which the information flows might be related to parameters of the temporal flow models. Consider a single information flow with a sender (*A*) and many receivers (*B*, *C*, *D*, *etc.*). For this exercise, let's assume that there is a constant delay of one minute between when a message is sent and when a receiver resends it. When the sub-network is in the form of a star, there will be a one-minute wait period from when the initial message is sent until all of the retweets are sent. In figure 3.2, I have re-plotted the basic sub-network topologies and added the related signatures based on a one-minute wait time, below.

Note that the decay is more gradual for the information flows in the form of trees; this is particularly true in the last example with the longer chains.

Certainly a consistent one-minute wait time is not realistic. Indeed, Barabasi (2005) found that wait times for responses to e-mail, online game messages, and instant messages could be approximated by a heavy-tailed, or Pareto, distribution, with the majority of messages responded to very quickly and only a few with very long wait times (see figure 3.3).

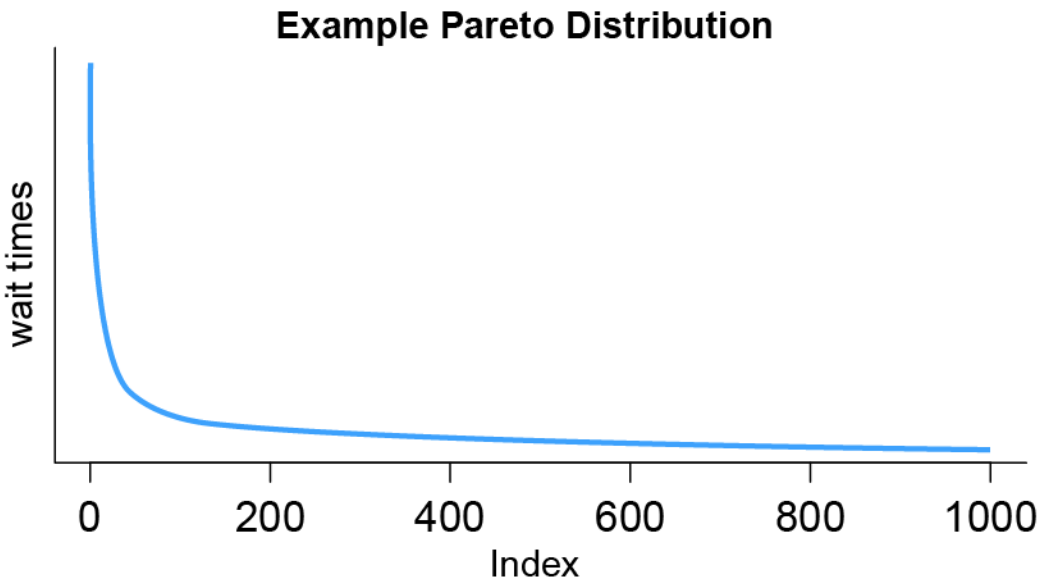


Figure 3.3. Example Pareto distribution for modeling human wait times instead of assuming constant wait time.

So while the wait times are not a constant number, they can be modeled with a Pareto distribution. As such, it is reasonable to assume that over many RTEs, the variance seen in the shape of the decay phase (see figure 2.16 and figure 2.17) is related to the length of chains in the sub-networks created during the RTEs. Specifically, more gradual decay phases (lower alpha values) are related to more and longer chains (lower values of closeness centralization) in the sub-networks of RTEs. Conversely, sharper decays (higher alpha values) ought to be related to sub-networks with fewer and shorter chains, or sub-networks that resemble star formations

(higher values of closeness centralization). Thus the specific question this leads to, and that this work addresses, is:

Q1: How is closeness centralization related to the shape of the decay phase of a signature?

3.1.2 Network structural changes and information flows

Literature that relates to the characteristics of content to sharing and linking behavior provides the background needed to understand the relationship between the flow of information and changes in the linking structure of networks. This section discussed content characteristics such as novelty, interestingness, emotional appeal, and richness of content.

Petrovic et al. (2011) asked human subjects to select, from a set of 200 tweets, the ones most likely to be retweeted. With this information, they trained a machine-based learning algorithm to duplicate human choices in selecting and retweeting tweets. Using their model on 21 million tweets gathered in October 2010, they find that tweets rated as having *novel* content are significantly more likely to be retweeted. Novelty in a message means the content is new to the user; they haven't seen it before. Suh et al. (2010) finds that tweets with hashtags, urls and @mentions are significantly more likely to be retweeted than those without such *rich content.* In addition, Bakshy et al. (2011), asking Mechanical Turkers to rate content on Twitter along different measures, find that tweets ranked as possessing *emotional appeal* or *interestingness* correlated with tweets that received more retweets. In addition, Hemsley and Mason (2013) theorize about the formation of *interest networks*, which are instantiated as a result of the flow of information and can become durable through repeated instantiations.

The above body of literature makes it clear that the content of the message is certainly a key driver in driving information flows. But none of this work examines how these flows might

be related to changes in the networks through which the information flows. Hemsley and Mason do discuss the relationship between information flows and changes in network structure, but they offer no empirical evidence for their theoretical stance.

In addition to driving diffusion, novelty has also been shown to be related to link creation in networks. Teng et al. (2012) examines the relationship between the novelty of content and discussion networks on Twitter. By using cosine similarity on the text of messages over different time periods, they find that the creation of new links was related to novel messages, that is, messages that have sharply lower similarity between time periods. They also look at the relationship between novelty and a changing network-level centralization measurement to determine to what degree the network topology is *star-like*.[25] They find that stable levels of centralization are related to lower novelty scores between time *t* and *t -1*, but that higher novelty between t and *t - 1* was correlated with higher centralization in *t + 1*. In other words, star formations would tend to grow after (not before) novel information was shared in the network but would tend to stay stable in the absence of novel information.

Note that Teng et al.'s work focuses on direct communication between users; it does not relate novel information to changes in Twitter's social network, that is, the network of follower relationships. And while their work does look at changing topic-based communication between users, Teng et al.'s models do not capture information flows and so cannot relate changes in their networks to temporal parameters of an information flow. However, this work certainly suggests that novel information flows could precipitate new links in Twitter's follower network.

From the above work, it is clear that novel content is more likely to be retweeted and is related to new network links. But where does novel information come from in networks? In other

[25] Teng et al. use a Gini coefficient to measure the inequality of links among nodes. More information on this is in appendix D.

words, what kind of links are the most likely carriers of that kind of information? In his landmark work, Granovetter's (1973) introduces the concepts of *strong* and *weak ties*. He argues that the strength of the tie, or relationship, between two individuals is directly related to the amount of overlap, or mutual relations, in their networks. Thus if A and B all know the same people, they are more likely to be strongly connected than if A and B know each other but share no similar connections. Granovetter refers to this second case, where A and B are linked but where none of A's associates are connected to B and none of B's associates are connected to A, as a *weak tie*. Granovetter shows that job referrals come more frequently from weak ties than from strong ties and suggested that in tightly knit groups, or clusters, people tend to know the same people and tend have the same information.

The theory of weak ties has been used by numerous authors in a variety of contexts. For example, Burt (2004) shows that good ideas in an organization come from weak ties that span what he called *structural holes* in networks. These holes are spaces between tightly knit groups (*clusters*). Weak ties *bridge* these holes and connect different clusters. In other examples, Ellison et al. (2010) shows that users on Facebook maintain larger sets of *weak-tie* connections in their social networks than those not on Facebook, and Bakshy (2012) demonstrates that on Facebook, people are more likely to share novel information from a strong tie than from a weak tie, but that since users have far more weak ties than strong ties, novel information, is more likely to come from a weak tie. Thus the preponderance of novel information on Facebook comes from weak ties. Given the breadth of work linking novel information to weak ties, it is reasonable to assume novel information is related to link creation in Twitter's social network of follower relationships.

Putting this together suggests two things: new links in Twitter's follower network could be related to the flow of novel information through weak ties and over bridges; and, from the

previous section, when information flows through chains deep into networks (that is, bridges to new clusters), we expect the decay phase of signatures to be more gradual as reflected by lower alpha values. In short, it is reasonable to expect a negative correlation between the change in followers during an RTE and the fitted alpha of the signature of the RTE.

In addition, Teng et al.'s (2012) work suggests that while star formations in networks grow after novel information, they are otherwise stable. And, again from the previous section, when the path through which an RTE flows is more star-like, a reasonable expectation is for sharper signatures and lower increases in followers. Thus the specific research question for this section is as follows:

Q2: How is the shape of the decay phase of a signature related to changes in the number of followers of users who initiate RTEs?

3.2 Approach and Methodology

3.2.1 Approach

The research addresses the questions for Phase 2 using variance regression models, one model for each question (see figure 3.4). Each model explores the relationships between and among the parameters of the signature model from Phase 1. In model 1, this takes the form of tracing the path through which the information flows; in model 2, changes in the social network are traced. Both models also control for context around the information flow. This section provides an overview of each of these models and justifies the methodological choice of utilizing regression for this analysis.

The first question is, *How is closeness centralization related to the shape of the decay phase of a signature?* Model 1, illustrated in figure 3.4, answers this question by examining the

relationship between the path of the information flow and the shape of the decay phase of the flow's signature. Also explored or held constant are the parameters of the RTEs in relationship to the temporal context of the Occupy movement (including controls for overlapping RTEs), and when, within in the Occupy movement, the RTE took place. The specific variables and their construction are detailed in section 3.2.2.2.

Model 1

$$\text{Decay Phase Shape} = f(\text{Signature Parameters}, \text{Network Path}, \text{Temporal Context})$$

Model 2

$$\text{Changes in Social Network} = f(\text{Signature Parameters}, \text{RTE User Context}, \text{Temporal Context})$$

Figure 3.4. Overview of variables used in the models developed for Phase 2.

It is important to note that the Twitter data used in this study does not contain the sharing chain of users needed to reconstruct the path of the RTE. A maximum likelihood model is used to infer (rather than determine) the path of the information flow, labeled *Network Path* in figure 3.4. Gomez-Rodriguez et al. (2010) developed this method for situations where the *path* of the diffusion of information is unobserved but the times when nodes are "infected" are known. They validate the method on both synthetic and real datasets, finding that more than 90% of the network edges are discovered. A brief overview of the data used in this study will demonstrate the necessity of this step.

As noted in the Phase 1 discussion of the data, the Twitter data used in this study contains a rich set of metadata. The data includes information specific to the tweet (the tweet text and data, and the time) as well as some information about the account of the user who sent the

tweet/retweet, such as the user name and the number of followers they have when the tweet/retweet is sent. However, there is a great deal of information that researchers might like to have that is not available. Especially relevant here is that status updates of retweets do not contain the sharing chain of users. In other words, the retweet chain given by {A → E → H → I} is not captured in the metadata. Rather, the impression from the metadata will be that {E, H, I} all directly retweeted {A} (see figure 3.5). But the research by Leskovec et al. (2007) mentioned in Chapter 2 make it clear that information does flow through chains, and thus a model that presents the path of information flows needs to represent flows through chains. I use the maximum likelihood method developed by Gomez-Rodriguez et al. (2010) for this purpose; Appendix D discusses this method, and the ways the path is quantified, in more detail.

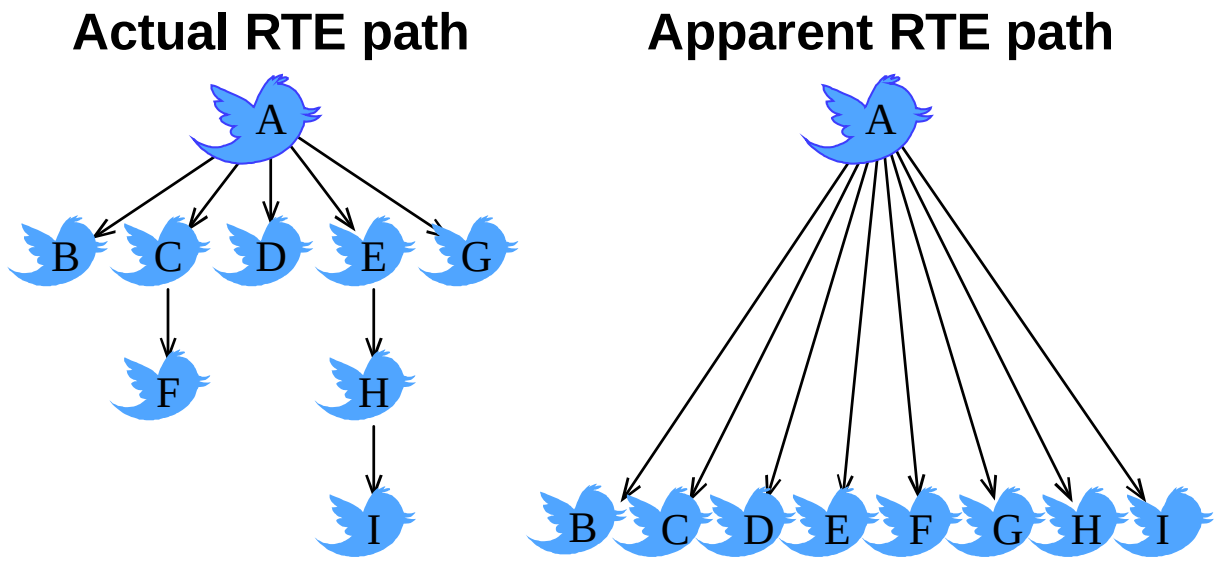


Figure 3.5: Example information flow paths. Given a hypothetical actual path of a flow (left), the apparent path suggested by the metadata will always be a star topology (right), thus the need to infer the path network.

The second research question for Phase 2 is, *How is the shape of the decay phase of a signature related to changes in the number of followers of users who initiate RTEs?* Again, a

multivariate regression model, illustrated in figure 3.4, answers this question by examining the relationship between the change in the structure of the social network, as measured by number of links created or eliminated over the course of the information flow, and the shape of the decay phase and other parameters of the signature model. Variables control for the user and temporal context of the RTE as well. The specific variables and their construction are detailed in section 3.2.2.2.

Multivariate variance regression models have several characteristics that make them well suited to addressing the research questions in Phase 2 of this project. Note that regression has long been used in studies of diffusion (Bass 2004), and more recently used to study the relationship of network centrality measures and information flows (Susarla, Oh, and Tan 2012), and life cycles of viral events (Nahon et al. 2011). This class of models is well suited to explaining or predicting the relationship between a dependent variable and an independent or explanatory variable while controlling for other effects (Faraway 2004; Kahane 2001). The control variables for both models in figure 3.4 are represented by “RTE User Context” and “Temporal Context”.

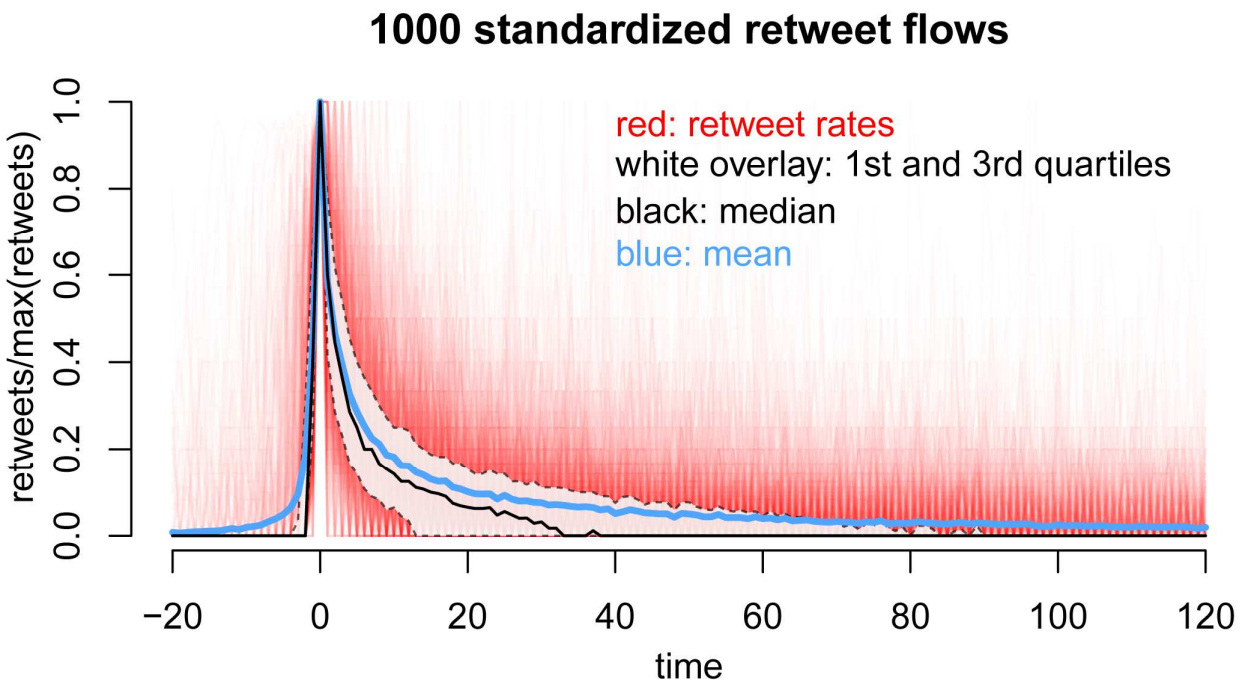


Figure 3.6. Plot of a sample of 1000 RTEs, reprinted for convenience. Social Media Lab data.

In addition, a variance model is a good choice for exploring the relationship between the variability of the shape of the signature (see figure 3.6) and both the path through which the information flows (research question 1) and the change in the network structure (research question 1). In the context of this work, another important strength of a variance model is that although it assumes that variable relationships are linear, non-linear relationships can still be assessed through variable transformations (Kahane 2001). The use of transformation also means that variables are not required to be continuous or to follow a normal distribution (Faraway 2005). These are important considerations because, as will be seen in the section describing the data, many of the variables do not follow a normal distribution.

Note that process analysis (Langley 1999; Pettigrew 1997; Pettigrew, Woodman, and Cameron 2001; van de Ven 2007) could be used to understand how, or the *process* by which, information flows and social networks interact. Pettigrew (1997) states that the aim of such an

analysis is to produce a case study that (1) identifies patterns in the process and compares them across cases; (2) finds the mechanisms that shape those patterns; and (3) employs not only inductive pattern recognition but the deductive theory that offers a starting point for the analysis. However, such an analysis requires that the researcher be able to observe the process in detail as it progresses from one step or stage to another. As noted above, Twitter metadata does not provide information about the path of the flow, and thus it is not possible to observe the actual process.

3.2.2 Procedures

3.2.2.1 Data

Recall from the previous chapter that the analysis for Phase 1 included 5,758 RTEs, made up of a sum total of 1,393,607 retweets. The findings section for Phase 1 discussed how the Twitter data set could be used in statistical analysis to understand the relationships between information flows, as represented by the signature model, and changes in social networks, as measured by a change in followers. This section provides additional, specific details about how the data are cleaned and prepared for analysis.

One of the most significant steps in cleaning the data arises out of the understanding gained in EDA that RTEs can overlap in time (see figure 2.14). If a user initiates two or more RTEs that overlap in time, and if they gain followers from each one, then the measured gain in followers for each RTE could be inflated, and thus this needs to be controlled for. The only users that can be effectively controlled for are those for whom the dataset contains all of the RTEs they initiated. For the Occupy account users, the only applicable set of users is those whose names match the search terms used to collect the Twitter data. There are 796 RTEs made up of 142,878 retweets and initiated by 86 distinct Occupy user accounts.

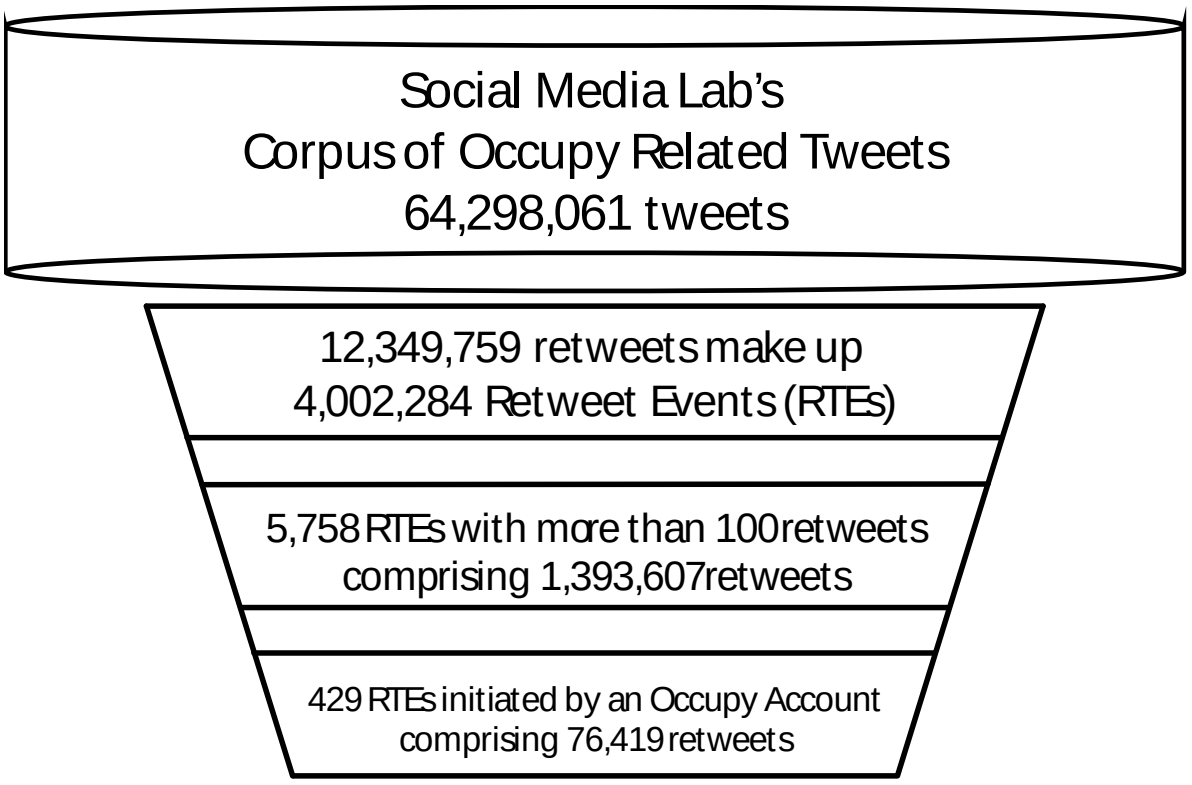


Figure 3.7. Graphical overview of the selection of data used in Phase 2 of this study.

The exploration phase also showed that some RTEs can have very long lifespans (see figure 2.15) which could compound the problem of overlap. For this reason, a uniform *window* is constructed around the peak rate of all RTEs. Since Kwak et al. (2010) found that 90% of retweets happen within an hour of the initial tweet, the window size has been set to twice that: 120 minutes. A shorter window could work, but Clauset et al. (2009) note that roughly 100 or more observations are needed to fit data to a power-law. To estimate the shape parameter of the decay phase of a signature, at least 100 minutes past the peak are needed.

Thus retweets with timestamps more than 120 minutes after the peak are dropped from the analysis. As a result, the size of some RTEs falls below 100, and these, too, are dropped. In

addition, RTEs are dropped if they did not have a clear peak[26] or if they had long periods of quiescence (no retweets) between the initial tweet and the first retweet in the set. This last criterion deals with the data missing because of the Social Media Lab's system outages (mentioned above in the data section of the first phase). The final data set used in the statistical analysis discussed below contains 429 RTEs made up of a total of 76,419 tweets/retweets. An illustration of the narrowing of the data can be found in figure 3.7, and the distribution of sizes of RTEs can be seen in figure 3.8.

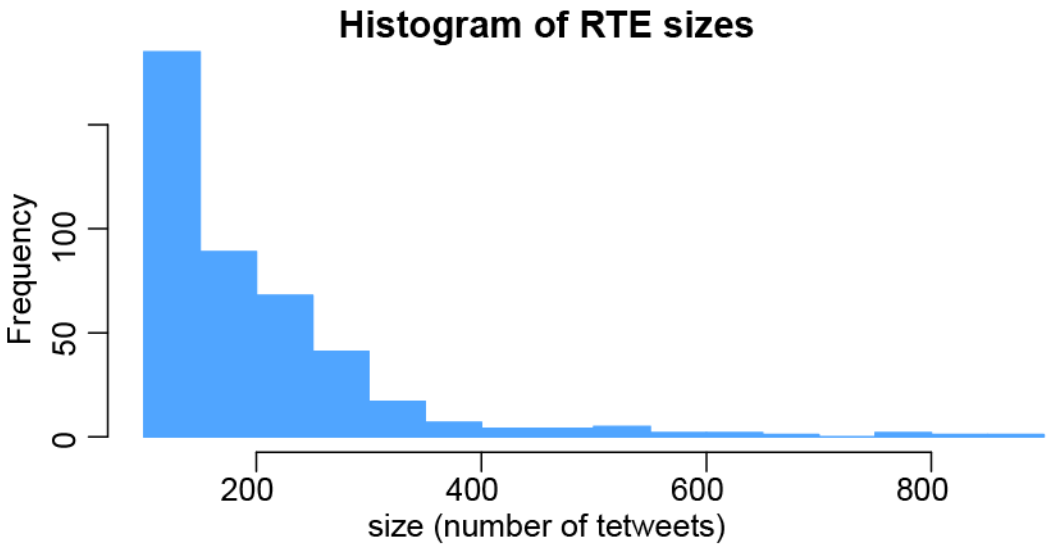


Figure 3.8. Frequency of sizes of retweet events, where size is the number of retweets plus the initial tweet. Social Media Lab data.

From a data centric point of view, an RTE is a complex data object. Once the retweets that form the RTE are grouped together, the initial tweet is constructed from the embedded metadata, saving the step of executing large numbers of database queries. Only a subset of the

[26] In addition to fitting a power-law to the decay phase, a Poisson goodness-of-fit test is run for the rate of retweets for the entire window. Signatures that fit a Poisson were flagged, inspected, and excluded.

metadata fields from Twitter status updates are required for the initial steps of creating an RTE. These fields are listed in table 3.1 and can be compared to an example of a full status update in appendix C. These metadata fields are used to either identify the RTE in cases where deeper probing is warranted or to calculate the variable groups used in the variance analysis.

Table 3.1. Status update metadata fields used in the construction of a RTE.

Field Name	**Description**
`id`	The ID of the tweet.
`created_ts`	The time stamp. When the tweet was sent/posted
`screen_name`	The user account name.
`followers`	The number of followers the user has at the time when the tweet is sent.
`retweeted_status.id`	*Only present when the tweet is a retweet.* Contains the ID of the tweet being retweeted.
`retweeted_status.created_ts`	*Only present when the tweet is a retweet.* Time stamp. The date the initial tweet was sent/posted
`retweeted_status.screen_name`	*Only present when the tweet is a retweet.* The screen name of the user whose tweet is being retweeted.
`retweeted_status.followers`	*Only present when the tweet is a retweet.* The number of followers of the user who sent the initial tweet *at the time the tweet is retweeted.*

The variable groups listed in model 1 and model 2 (figure 3.4), and corresponding to Phase 2 research questions 1 and 2, respectively, are all calculated or inferred from the metadata fields in table 3.1. All of the parameters of the signature model are derived from the timestamps by finding the number, or rate, of retweets for each minute over the course of the signature. Table 3.2 illustrates the result: a vector of numbers where each cell contains the rate of retweets for the minute that corresponds to the index of the vector. This facilitates estimating alpha and finding the peak rate and other parameters of a signature (see figure 2.17).

Table 3.2. Example vector of aggregated time stamps over time for a RTE signature.

Rate	1	20	50	40	15	7	3	9	2	0	2	…	0

Index	1	2	3	4	5	6	7	8	9	10	11	…	t

The timestamps are also used to create the temporal context of the RTE. This is done by finding the difference in days and in minutes from the start date of the Zuccotti Park protests that kicked off the Occupy movement at midnight GMT on September 17, 2011.[27] The change in network is calculated by finding the difference between the number of followers the initial user has at the end and the start of the RTE's window. Finally, the RTE's timestamps are also used in the maximum likelihood model that infers the network path. More details about this and the other variables for the models are discussed in the next section.

3.2.2.2 Variables

This section provides details about the variables in the models. For both models, the mathematical equations are provided, followed by a discussion of each variable that includes its purpose in the model, how it has been constructed, and, where appropriate, how it has been transformed to fit the model assumptions. As will be discussed in the next section, both models are multi-variant regression models; as such, the model equations will be presented in traditional form, but a graphical representation will be included as well (figures 3.9 and 3.11).

Model 1

[27] Note that picking this date is somewhat arbitrary since technically the movement started before the Zuccotti Park protests. But the Social Media Lab's data collection didn't start until October 19, and the point of the variables are to provide general context. Thus a more specific date and time is unnecessary for this analysis.

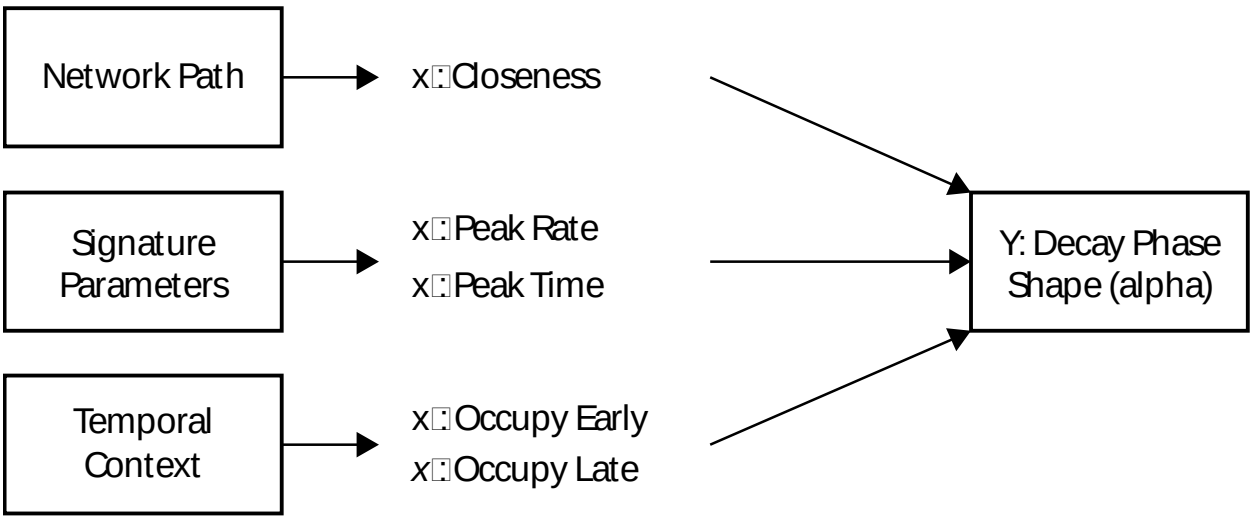


Figure 3.9. Graphical version of model 1 with variable groups and the specific variables.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 +$$

alpha (dependent variable). Alpha is the estimated shape parameter of the power-law fitted to the decay phase of the signature. As outlined by Clauset et al. (2009), the estimation employs a maximum likelihood algorithm that attempts to optimize for the best p-value for a Kolmogorov-Smirnov test (Marsaglia, Tsang, and Wang 2003) between the fitted distribution and the original sample. Note that what is important is having a systematic way of measuring the shape of the decay of a signature. In other words, a way that allows for the comparison of decay shapes across signatures. The p-value for the estimated alpha is not used to determine if an RTE event ought to be included or not. Note that according to the QQ-plot in figure 3.10, alpha is roughly normally distributed, though the high end does show some deviation.

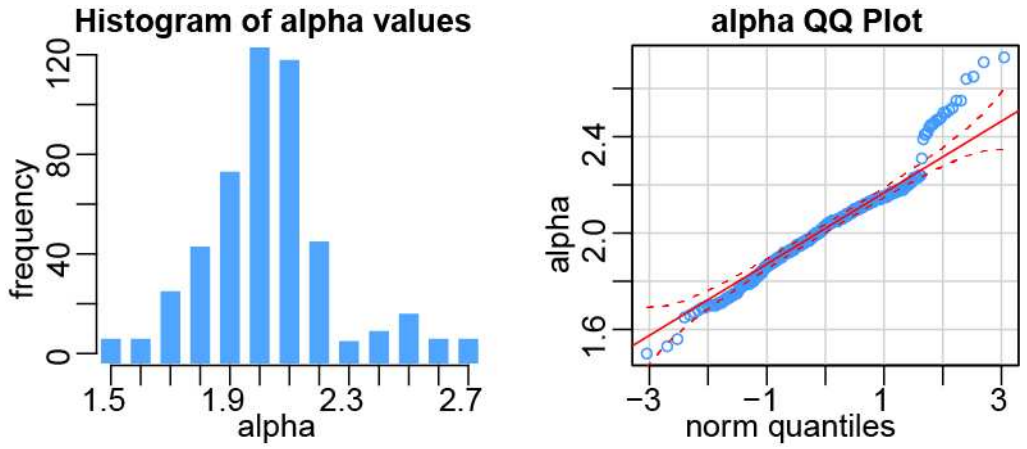


Figure 3.10. Distribution of alpha and QQ plot showing a roughly normal distribution with some deviation at the high end

closeness centralization. The network path is established and measured in two steps. First, Gomez-Rodriguez et al.'s (2010) maximum likelihood method is used to infer the path of the flow. The result of this process is a directed tree network with the initiating user of the RTE as the *root*, or starting node. Information flows from the root node to (and through) other nodes, each of which represents a user who retweeted the initial user's tweet. This produces a directed network for each RTE. This network is characterized by *closeness centralization*, a surrogate for the number and length of chains comprising the path of the flow of the RTE. Note that this is not equivalent to closeness *centrality*, which measures the average distance, measured in links, between a node and all other nodes in the network. Closeness *centralization*, used in this study, is a network-level measurement found by averaging the closeness centrality for all of the nodes in the network (Wasserman and Faust 1994). When closeness centralization is 1, the network is a perfect star with the center being the initiator of the RTE. As the number and length of chains in the network increase, the closeness centralization measure decreases. For a detailed

description of the model used to infer the RTE path network, and more details about closeness centralization, see appendix D.

peak (max) rate. This variable equals the rate of retweets at the peak. In the signature model, it is the height of the peak. A negative relationship is expected between max peak and alpha because a higher peak means more people simultaneously retweeting, so, holding all else equal, there is a greater potential for more retweets in the next minute and the following minutes. More retweets in the window after the peak implies a higher volume under the curve, and thus a more relaxed decay phase (lower alpha). The distribution of the variable is skewed, so it is log-transformed to fit the model better.

peak time. Peak time measures the time in minutes from the initial tweet to the peak of the RTE. This variable has a skewed distribution, so the variable is transformed with a logarithm. The variable is included to provide an understanding of how the parameters of the signature model are related to the shape of the decay, or alpha.

occupy day. This variable provides temporal context for the RTE within the Occupy movement. For each RTE, the value of the variable is the number of days from the start of the Zuccotti Park protest on September 17, 2011, to the peak of the event. This controls for trends within the Occupy movement not captured in individual RTEs. The relationship between occupy day and alpha is not linear. Exploration of various transformations and early model fits revealed that a *broken stick* or *piecewise* regression method (Faraway 2004) resolves the issue. In a broken stick regression the variable is split into two or more variables, which are then included in the regression. The first contains all of the original values up to a given cut off, there after the values are zero. The second variable contains

zero up to the cut off, and the original values thereafter. Thus *occupy early* and *occupy late* are derived from occupy day and included in the regression[28].

Selection of the cutoff occurred after early attempts to fit the regression. A Bonferonni test for outliers (Faraway 2005) identified 20 observations as outliers. Investigation showed that these observations have two qualities in common: they tend to have high alpha values, and the RTEs all peak during or later than March 1, 2012, that is, *late* in the Occupy Movement. Faraway (2004) notes that a structural modification of the model is required in the presence of an observable pattern in a set of outliers. A cut off of 150 days, February 17, 2011, situates all of the outliers in the occupy late variable. This cutoff date happens during a period of relative quiescence and corresponds to a brief collection outage in mid-February (see figure 2.5).

Model 2

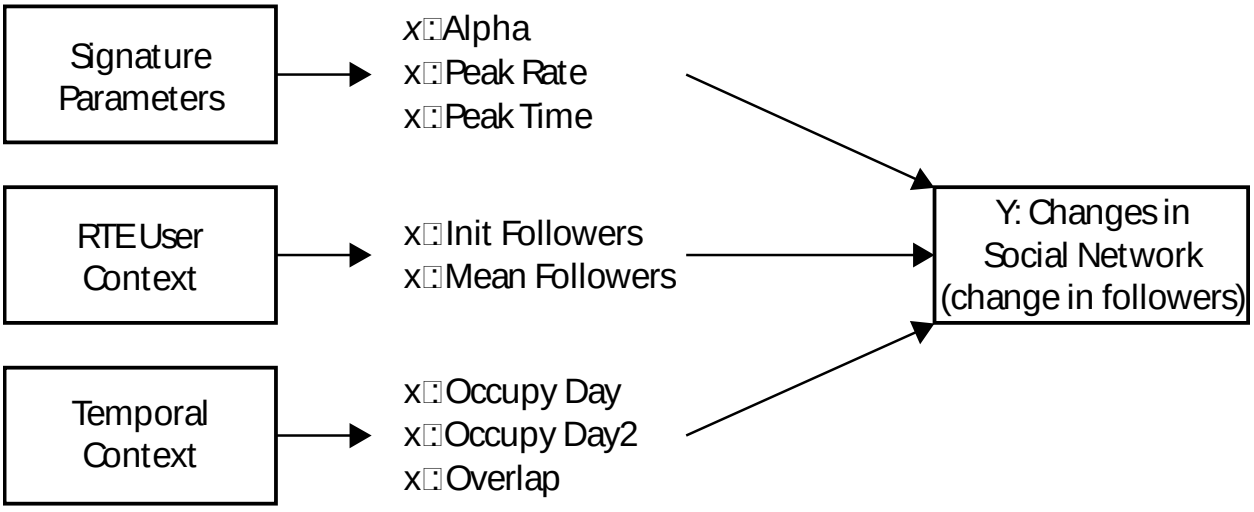


Figure 3.11. Graphical version of model 2 with variable groups and the specific variables.

[28] Exploration of alternative functional forms and transformations included second- and third-degree polynomials, square root transformation, and logarithmic. The broken stick method and a second degree polynomial (used in the second model) resulted in similarly adjusted R-squares and, based on an F test, were not significantly different. With the broken stick method, residuals appeared more normal in the diagnostic plots.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7^2 + \beta_8 x_8 +$$

changes in followers (dependent variable). This is the dependent variable, the net change in the number of followers of user who initiated the RTE. It is measured during the window of the RTE. The distribution is skewed, and a number of transformations were attempted. Ultimately the Box-Cox method (Faraway 2004) best estimates the optimal power transformation, 0.19 for this model. The transformation puts the variable more in line with a normal distribution (see figure 3.12). Also, the transformation resolves a problem with non-constancy of variance in the predictor and improves the linearity of the relationships with the independent variables.

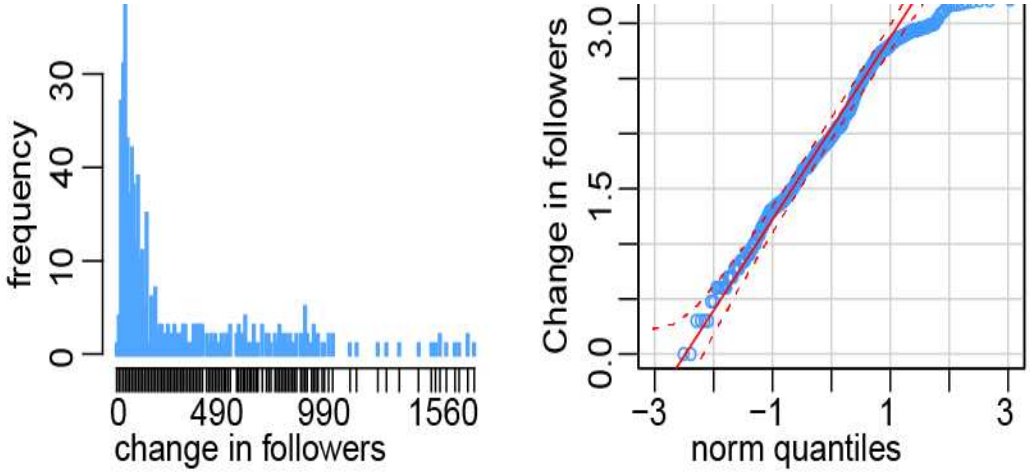


Figure 3.12. Distribution of change in followers and the QQ plot for the transformed variable. Note that the transformed variable is roughly normal except for deviation at the end high end.

alpha. See the description for alpha for model 1. A negative value is expected for this variable. As the value of alpha decreases, the decay phase of the signature is more relaxes, meaning higher levels of retweets are sustained. All else being equal, higher levels of

sustained retweets should be accompanied by greater increases in the number of followers that the initiator RTE receives. See section 3.1.2. for a more detailed discussion.

peak (max) rate. See the description for this variable for model 1. A negative relationship is expected because a higher peak means that more people are simultaneously retweeting. Holding all else equal, there is a greater potential for more retweets in the next minute, and thus a greater chance for the user of the initial tweet to gain more followers.

peak time. See the description for this variable for model 1. The variable is included to provide an understanding of how the parameters of an RTE signature are related to the change in the number of followers of the RTE initiator.

initial follower count (init followers). This variable is the number of followers that the RTE initiator has at the start. The assumption is that users with large numbers of followers have greater reach than those who do not, and thus have a greater potential to be retweeted and reach new users who might become followers. This variable is log-transformed.

mean followers. This variable is the mean number of followers of all the users who retweet in an RTE. When a user with a very large number of followers participates in an RTE by retweeting the initial tweet, then we can expect, all else being equal, that more people will be exposed to the initial tweet than if such a user had not participated. This exposure ought to create more potential for the initial user to gain new followers. Mean followers controls for this follower effect. The distribution of this variable is skewed. Note that in many cases the median would be used as the central measurement for a skewed

distribution, but since the intent is to capture the effects of very large users, the mean is used. Therefore, the variable is log-transformed.

overlap. As was discussed in section 2.2.2.2, when a user initiates more than one RTE in a short period of time, the RTEs can overlap. Unless overlapping RTEs are controlled for, the change in followers for each of the overlapping RTEs will be over-represented. The overlap variable, an integer count of the number of overlaps of one RTE with other RTEs initiated by the same user, controls for this effect. This variable has been transformed using a square root to improve the imparity of the relationship with the dependent variable.

occupy day. As noted for model 1, this variable measures the number of days into the Occupy movement and controls for the temporal context of the RTEs. Unlike model 1, where a broken stick regression approach was used, for model 2 a quadratic relationship between Occupy Day and the response variable provides the best fit.

3.2.2.3 Inferential Analysis

All models are wrong, but some are useful. - George Box

The purpose of this section is to discuss the model specification, the interpretation of variables, and the diagnostics employed to ensure the model assumptions are met. As mentioned before, multivariate regression models are employed to answer both Phase 2 research questions. Because the project aims to develop an understanding of how parameters of the signature model are related to network dynamics, hypothesis testing of specific individual variables was more appropriate than was model development. In addition, the focus is the significance and direction

of relationships, not necessarily the magnitude of those relationships. Thus interpretation consists of noting the significant variables and interpreting the direction of the relationship. For both models, the findings section reports the coefficients and their p-values, along with the confidence intervals for these estimates. The findings also report the R-squared values and F statistics for both models.

Development of the models proceeded iteratively using stepwise regression for variable selection. The goal of stepwise regression in this context is to find the smallest model that fits the data best (Faraway 2004). Thus a model was fit with a given set of variables, followed by another fit with a variable added or removed. When adding variables, an F-test was used to determine if the variable added significantly to the model. For each promising mode, diagnostics were run to provide feedback for further improvement of the models.

The stepwise approach to variable selection supports the exploratory nature of the work. Indeed, Faraway states that "variable selection is a process that should not be separated from the rest of the analysis" (2004, 63:121), indicating that it is the process of variable selection that can provide insight into the data and variable relationships.

Of course, as noted before, the tweet metadata is quite rich, meaning that a wide range of variables could be included in model development. The criterion for variable inclusion is as follows:

1. The variable represents one of the measurable parameters of the signature model
2. The variable measures some potentially important characteristic of the users who participate in an RTE, for example, the number of followers they have, or whether or not the user account has been identified as a bot (an automated account)[29]

[29] Early modeling efforts included a variable for the number of bots that participated in a given RTE. These variables failed to be significant. For a list of Occupy bots, see: https://twitter.com/cdubs/occupybot/members

3. The variable controls for important context within which the RTE is situated, for example, overlapping RTEs or when the RTE took place over the course of Occupy Movement[30]

The assumptions of a linear regression model are that the parameters enter the model linearly, that the model errors are independent, have constant variance, and are normally distributed. In the context of these models, linearity means that the relationship between the dependent variable and the independent variables should be linear. These relationships are explored graphically using scatter plots and, where necessary, variables are transformed to improve the linearity of the relationship. Note that variable transformation can make the interpretation of the coefficients difficult. This is particularly true when the dependent variable is transformed, as is the case for model 2. However, as noted at the beginning of this section, the focus for this analysis is primarily on the direction of the relationship for significant variables. Thus the sign of the coefficients shall be interpreted, but not the magnitude of the relationship.

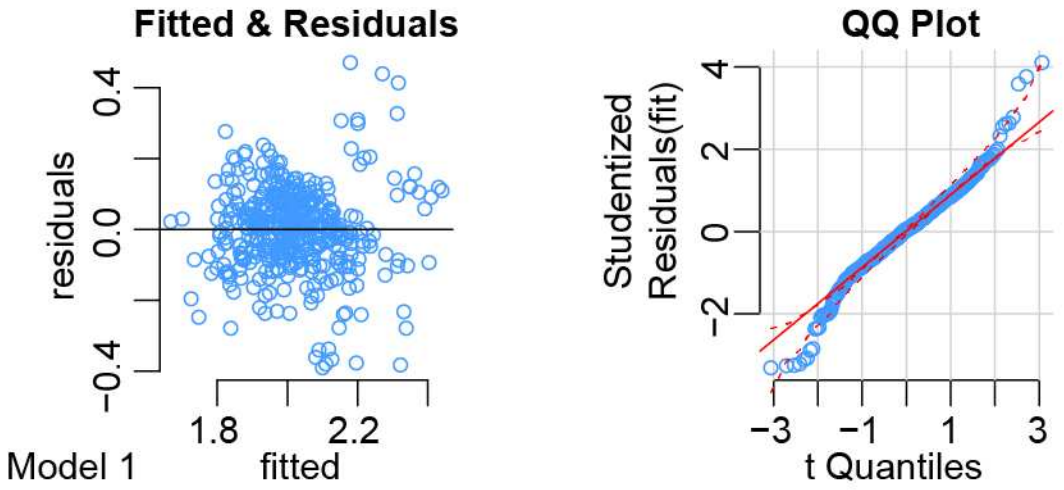


Figure 3.13. Model 1 diagnostic plots: residuals versus fitted and QQ Plot for normality of residuals.

[30] Butts and Cross (Butts and Cross 2009) found season effects to be a factor in volatility in blog citation networks. Variables for the time of day are not included in the models for this study because the time of day is not significantly correlated with either alpha (Pearson's correlation: 0.080, d.f. 427, 95% CI: -0.015, 0.175) or change in followers (Pearson's correlation: -0.024, d.f. 427, 95% CI: -0.118, 0.071).

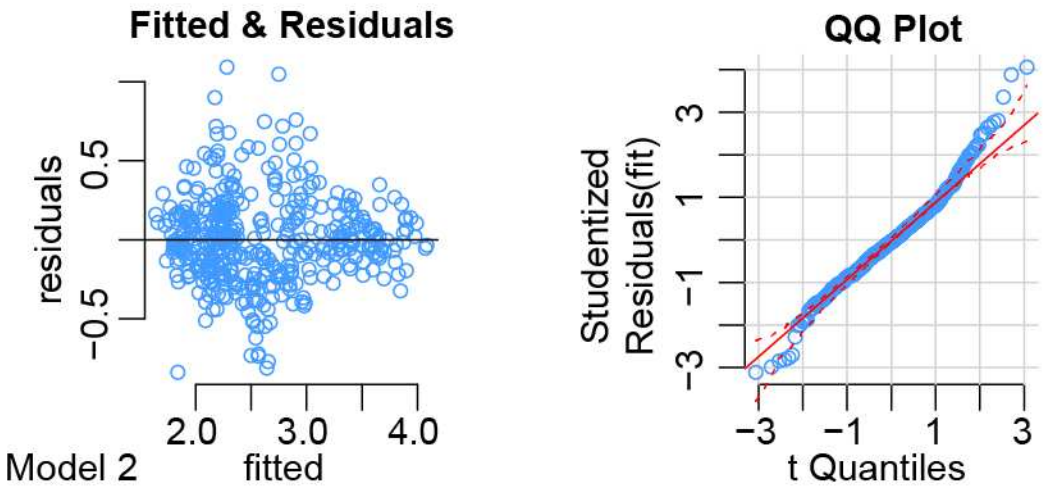


Figure 3.14. Model 2 diagnostic plots: residuals versus fitted and QQ Plot for normality of residuals.

The following paragraphs examine the validity of the regression model as an adequate statistical representation of the data. This involves checking the residuals for constant variance, and testing for multicollinearity and autocorrelation. For both models, the fitted vs. residual scatter plots demonstrate constant variance, and the QQ plots note that the residuals are roughly normal except for a mild drift at the ends. Figures 3.13 and 3.14, respectively, show model 1 and 2 diagnostic plots.

A Durbin Watson test is run on both models to test for autocorrelation, which is when the residuals are correlated. Low p-values, less than 0.05, would suggest the presence of autocorrelation (Ott and Longnecker 1993). For model 1 the D-W test statistic was 1.85, with a p-value of 0.114. For model 2 the test statistic was 2.10, and the p-value was 0.264. These values indicate that autocorrelation is not present these models.

Multicollinearity is the case where the independent variables in a model are strongly correlated, which can result in poor estimates of the model coefficients (Kahane 2001). A

variance inflation factor (VIF) test was run for both models. This test calculates how much multicollinarity increases the variance of a coefficient. Values greater than 4 indicate the presence of multicollinearity. Table 3.3 and table 3.4 provide the VIF values for the coefficients of models 1 and 2, respectively.

Table 3.3. Variance inflation factors for model 1 coefficients.

VIF values for model 1				
closeness centralization	Peak Rate	Peak Time	Occupy Early	Occupy Late
1.26	1.3	1.8	2.45	3.27

Table 3.4. Variance inflation factors for model 2 coefficients.

VIF values for model 2							
Alpha	Peak Rate	Peak Time	Init Followers	Mean Followers	Occupy Day	Occupy Day^2	Overlap
1.35	1.75	2.11	1.50	1.37	2.60	1.27	1.73

3.3 Findings

The first regression model has been designed to answer the research question, *How is closeness centralization related to the shape of the decay phase of a signature?* As noted before, *closeness centralization* is used to measure the degree to which the path of the information flow included chains, and *alpha* quantifies the shape of the decay phase and is the dependent variable. The R^2 of the regression model was 0.5448, indicating that the model explained 54% of the variance seen in alpha for the RTEs included in the study. The F statistic (5, 423 df) for the model was 101.3 (p-value=0), indicating the model is *significant*, or that all coefficient estimates are non-zero. Table 3.5 contains the coefficient estimates, the confidence interval (2.5% CI and 97.5 CI), standard error, t-value, and p-value for those estimates.

Table 3.5. Model 1 results.

Model 1	Estimate	2.5% CI	97.5% CI	Std. Err.	t-value	p-value
Intercept	2.502	2.401	2.603	0.051	48.782	0***
Closeness centralization	0.226	0.199	0.253	0.014	16.382	0***
Peak Rate	-0.177	-0.230	-0.124	0.027	-6.568	0***
Peak Time	0.227	0.189	0.266	0.019	11.666	0***
Occupy Early	0.0012	0.0006	0.0018	0.0002	4.197	0***
Occupy Late	0.0008	0.0006	0.0011	0.0001	6.676	0***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1						

Note that the relationship between closeness centralization and alpha is significant and positive. A lower alpha means a more relaxed decay phase, and a lower closeness centralization means the information flow traveled through more and longer chains. Thus, all else being equal, more relaxed decay phases are related to information that flowed through more and longer chains. Because this indicates how chains are related to the shape of the signature, this addresses the first research question.

This finding is consistent with Granovetter's (1973) theory of weak ties and other authors who note that messages spread faster and easier in strong tie networks, but need weak links to span structural holes and disseminate novel information to more distant parts of networks (A. L. Barabasi 2003; Burt 2004; Walther et al. 2010; Rogers 1995). When novel information flows over weak ties to new clusters it may spread quickly among these strong ties and be detectable as higher sustained rates of retweets. Each time the users in the cluster retweet, individuals they are weakly linked to may pick up and spread the message in their own strong tie clusters, creating the possibility that another cluster will again quickly share and spread the information.

Note that a limitation of this work is that bridges and weak ties are not directly detectable in the inferred paths of the RTEs. Detecting weak ties would require having knowledge of both the sub-network and the larger social network's linking structure. Thus, even though longer

sharing chains are related to more relaxed signatures, this study cannot confirm that the path flowed through weak ties.

The peak rate and peak time, other parameters of the signature model, are also significantly related to alpha. Higher peaks, which measures the number of retweets during the peak, are related with lower alphas (negative relationship), and earlier peak times are also related with lower alphas (positive relationship). Thus for more relaxed decay phases we would expect higher peaks that occur earlier in the event.

With respect to the temporal context of the RTE, both broken stick regression variables are positive and significant. It follows, then that as time goes on the decay phases tend to become sharper, all else being equal. However, the rate that the decay phase became steeper was significantly different before and after day 150 of the Occupy Movement (February 14, 2012). Before the cut-off, the change was faster; after, it was slower.

The second regression model has been designed to address the research question, *How is the shape of the decay phase of a signature (alpha) related to changes in the network structure?* As noted before, alpha quantifies the shape of the decay phase—and the change in followers of the user who initiated the RTE—is used to measure the change in the social network. The change in followers is the dependent variable in model 2. The R^2 of the regression model was 0.8218, indicating that the model explained 82% of the variance seen in the change in followers. The F statistic (8, 420 df) for the model was 242.1 (p-value=0), indicating the model is significant or that all coefficient estimates are non-zero. Table 3.6 contains the coefficient estimates, the confidence interval (2.5% CI and 97.5 CI), standard error, t-value, and p-value for those estimates.

Table 3.6. Model 2 results.

Model 2	Estimate	2.5% CI	97.5% CI	Std. Err.	t-value	p-value
Intercept	1.773	1.063	2.408	0.378	4.686	0.000***
Alpha	-0.309	-0.453	-0.109	0.097	-3.195	0.002 **
Peak Rate	0.382	0.200	0.482	0.079	4.824	0.000***
Peak Time	-0.008	-0.101	0.089	0.053	-0.146	0.884
Init Followers	-0.128	-0.203	-0.027	0.049	-2.587	0.010 *
Mean Followers	0.408	0.238	0.498	0.073	5.572	0.000***
Occupy Day	-4.079	-4.577	-2.833	0.491	-8.311	0.000***
Occupy Day^2	1.361	0.628	1.847	0.343	3.970	0.000***
Overlap	0.360	0.300	0.347	0.013	26.920	0.000***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1						

To answer the second research question, note that the relationship between alpha and the change in followers is negative and significant, as expected. More relaxed decay phases, all else being equal, correspond with higher increases in followers or with more new links in networks. This finding is also consistent with Granovetter's (1973) theory of weak ties. Holding all else equal, higher levels of sustained retweets ought to provide more opportunities for the message to cross weak tie bridges and expose the novel information to new users who might find it interesting enough to create a new follower relationship in the network.

Of course, people retweet messages for many reasons, not only because they are novel. For example, section 3.1.2 noted that tweets with emotional appeal or interestingness are more likely to be retweeted. Even with this caveat, the findings of this study are consistent with the expectations set by Granovetter's theory. Indeed, the findings imply that more relaxed decay phases of signatures may be related to novel information. This is supported the finding that more relaxed decay phases are related greater increases in followers, particularly when we recall that Teng et. al (2012) linked novel information to new links in discussion networks.

With the exception of Peak Time, other parameters of RTEs are also significant. Peak Rate and mean followers were both positively related to the change in followers, whereas Initial Followers was negative. This means that RTEs with higher peaks and RTEs where participants had higher numbers of followers correspond with greater increases in followers (new links). Interestingly, when the user who initiates the RTE has more followers, they tend to gain fewer new followers.

With respect to the temporal context of the RTE, both variables for Occupy day (that is, how many days the RTE is into the Occupy Movement) were significant but of opposite signs. The sign of the highest order provides information about the shape of the curve of the relationship, so in this case a convex, downward sloping relationship exists. Thus as time passes, Occupy users who initiate RTEs gain fewer followers. Finally, the variable that controls for overlap of RTEs is also significant and positive, as expected. This means that when a user initiates multiple RTEs with overlapping windows, we expect to detect higher increases in the follower count.

4 Discussion

4.1 The interaction of flow and network dynamics

This study examines the relationship between temporal parameters of an information flow and network dynamics. The methods include both exploratory data analysis (EDA) and confirmatory statistical analysis. Previous sections split the work into distinct phases; this section ties together and discusses the implications of the findings for both phases. Findings from the exploratory analysis include (1) a signature model defined by parameters as illustrated in figure 4.1, and (2) an operationalization of Twitter social network changes (as a change in follower

count). The findings from the statistical analysis include the ways the parameters of the signature model are related to the changes in the Twitter social network as well as how the path through which the information flows is related to the parameters of the signature model.

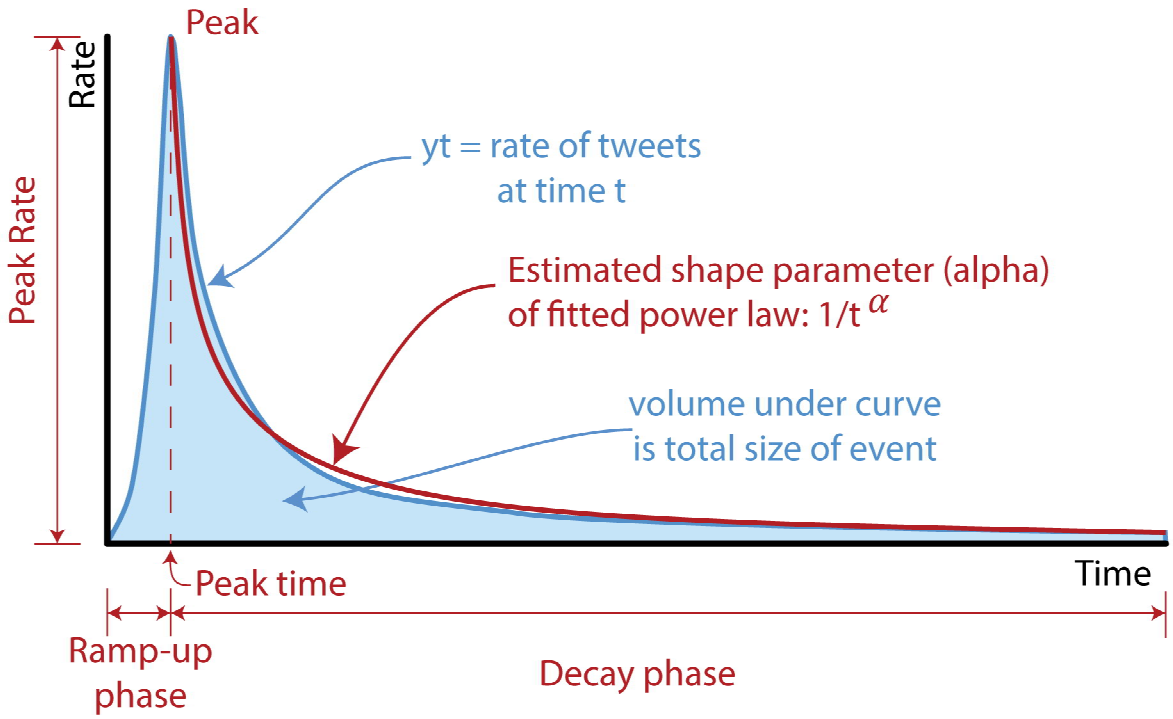


Figure 4.1. Signature model from phase 1, included here for convenience.

The dependent variable for model 1 measures the shape of the decay phase of the signature, and the independent variables includes the network path (as measured by closeness centralization) and other parameters of the signature model, while holding constant the temporal context of the RTE. The dependent variable for model 2 measures the change in social networks as operationalized by the change in the number of followers of the initiating user of the RTE. The independent variables includes the parameters of the signature model, with a particular focus on the shape of the decay phase, and hold the context of the RTE constant by including variables

related to the users who participated in the RTE, as well as temporal variables. Figure 4.2 graphically illustrates both models.

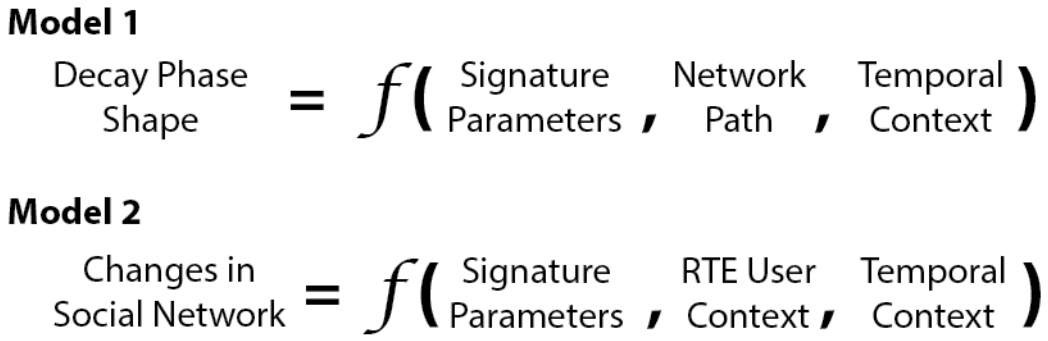


Figure 4.2. Regression models from phase 2, included here for convience.

One way to think about the relationship between these models is to think of the first as relating the state of the network before and during the information flow to the signature model, and the second as relating the network state during and after the flow to the signature model (see figure 4.3). In the first model, closeness centralization measures the path of the information flow. The information is assumed to flow through a pre-existing network. As the information flows, it instantiates a sub-network (the path of the flow) made up of the users who share the information and of the links connecting them. This path is assumed to exist as a potential path within the larger social network before the information flows. The dependent variable of the second model is the net change in followers of the user who initiated the RTE. This variable provides information about the state of the network at the end of the information flow in terms of how much the network changed.

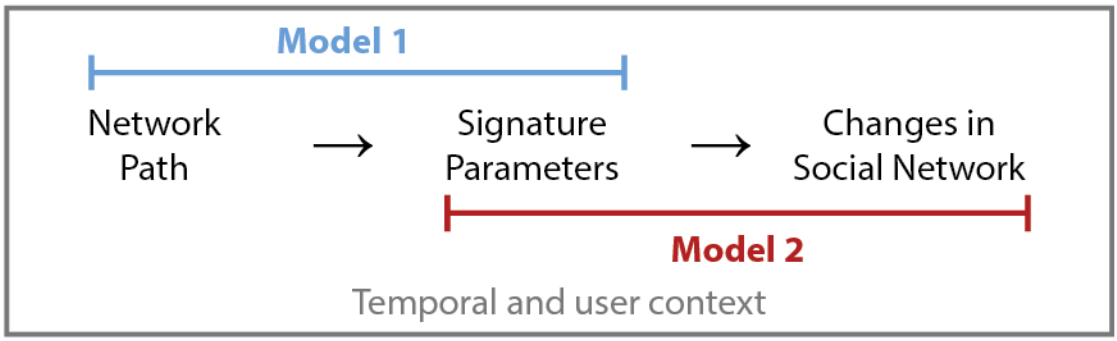


Figure 4.3. Conceptual relationship of the regression models.

Of course, the signature model is a temporal representation of the information flow. But the assumption of a pre-existing path, coupled with the findings of model 1, suggest that the signature model also reflects the path of the flow of information, and thus the state of the network before the information flow. Why? Recall that a positive and significant relationship exists between closeness centralization and alpha. Closeness centralization measures the presences of star-like structures in the sub-network. Alpha is the shape of the decay phase of the signature. If closeness centralization is not significant in the model, then no relationship between the path of the flow and the shape of the decay phase will be detected.

In the second model, alpha is negatively related to the change in followers, which reveals the state of the network at the end of the flow. The negative relationship indicates that more relaxed decay phases correspond to greater increases in followers for the initiating user. In addition to the signature reflecting more sharing chains and the state of the network before the flow, then, it also reflects the changes in the network and state of the network at the end of the flow. Indeed, closeness centralization and change in followers are correlated. The Pearson's product-moment correlation estimate is 0.217, with the 95% confidence interval between 0.125 and 0.306 (df = 427). The implications of the direction of this relationship are explored below,

but it means that information flows that follow more star-like paths relate to higher increases in follower counts. This relationship implies, possibly intuitively, that the initial conditions of the network (captured by closeness centralization) are a factor in the final state of the network (reflected in the change in followers).

Not so intuitive, perhaps, is the implication that the process of information flows and the dynamics of social networks is recursive. At least in the case of the Occupy movement, any individual information flow is situated in the larger context of the discussions taking place within the network of participants. So the state of the network at the end of a given information flow is the initial condition network state for the start of a second information flow. The final network state for the second flow becomes the initial condition network state for the next, and so on. This process is termed an *iterative flow structuring model* for the remainder of this document. This iterative flow structuring model is a simplification of a more complicated process. One such complication is that information flows can temporally overlap. An information flow, F_0, starting at t_0, has an initial network state of G and a set of sub-networks that make up the potential paths for the flow. An overlapping flow, F_1, that alters G between t_0 and t_1, alters the potential paths for F_0 for all subsequent times, creating interim network states.

The results from model 1 also indicate that the parameters of the signature model are related. A negative relationship exists between the peak rate and alpha, and a positive relationship exists between the peak time and alpha (see figure 4.4). In the former, the finding makes sense because the height of the signature at each minute corresponds to the number of retweets in that minute, and so a high peak at time t_0 means more users are exposed to, and may retweet, the message at time t_1. Successive rounds of higher retweets per minute will increase the volume under the decay phase curve; these rounds relate to a more gradual decay.

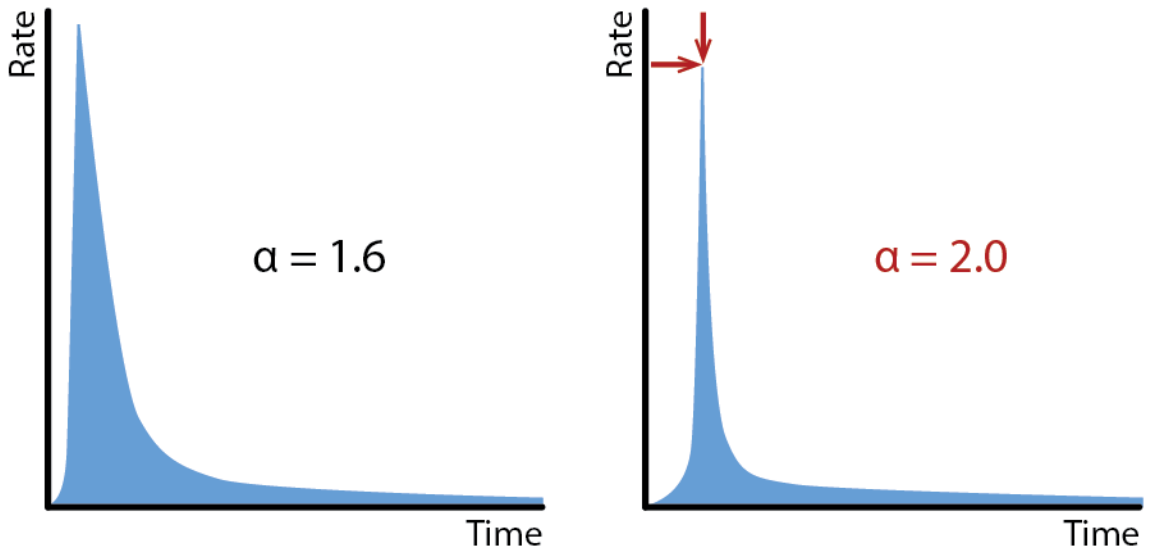


Figure 4.4. Signature model parameter relationships found in model 1. Lower and later peaks are related sharper decay phases.

In the latter relationship case, the finding indicates that shorter delays from the initial tweet to the peak are related to lower alpha values, and thus to more gradual decay phases. This finding runs counter to an earlier study, Crane and Sornette's (2008) research on YouTube videos. They find that more gradual decay phases in the rate of video views tend to be associated with longer and more gradual build-ups. One could account for the discrepancy by considering the differences in the unit of analysis for both studies. In this study, the unit of analysis is an RTE, an initial tweet and the retweets of that tweet identified from Twitter's metadata. The information flows are thus confined to the Twitter environment. YouTube is a content hosting site; the views videos receive come from referrals originating from many platforms, including e-mail, social networking sites, blogs, tweets, and YouTube itself (Broxton et al. 2010). To study the daily views of videos on YouTube is to study the aggregation of information flows across many platforms. The analogous process on Twitter might be a trending or viral topic (Huang, Thornton, and Efthimiadis 2010; Nahon and Hemsley 2013), where many tweets and RTEs

together form a topically bound conversation. Each of the RTEs in this study are all part of the Occupy movement conversation. As such, the discrepancy between the findings of this study and of Crane and Sornette's study could reflect the difference between individual flows and the aggregation of many individual flows, respectively.

The second model focuses on the relationship between the change in followers of the initiating user and parameters of the signature. The model found alpha and Peak Rate significantly related to the change in followers, with the former being negative and the latter being positively related. Recall that a higher peak means more users are exposed to the message in the next round of retweets, holding other factors constant. Thus, more users who are not already following the initiator of the RTE could be exposed to the message and opt to start following the RTE initiator. Model one found the relationship between alpha and Peak Rate to be significant. It would be reasonable to wonder if the alpha and Peak Rate are doing the same work in the model, but the VIF scores for model 2 indicate otherwise, so we can assume that the relationships are independent and that the interpretation above is valid.

Recall that the second model also includes variables to control for the user and temporal context of the RTE. The negative coefficient for Initial Followers indicates that users with fewer followers tend to gain followers more quickly. The inclusion of the Occupy Day variable in the model holds time constant, so the relationship holds early in the Occupy Movement as well as later. Also, since the change in followers is a net-change variable, the proportional effects are even more significant: a user who starts off with 500 followers and gains 50 experiences a far larger proportional gain than a user who starts off with 500,000 and gains 50.

Note that the dataset includes 429 RTEs initiated by only 17 users. Since users show up more than once in this dataset we have some insight into how the relationship between initial

followers and change in followers might change as users gain more followers. Users with large followings tended to experience greater follower growth when they had fewer followers and slowed down as they had more followers. This is consistent with figure 2.18, which showed a faster rate of follower growth for the OccupyWallSt account earlier than later. Indeed the relationship between the OccupyDay variable and the change in followers in model 2 is represented by a second degree polynomial; it is significant, and the squared term is positive, indicating the Occupy users in this study experienced a faster rate of follower growth earlier in the movement than later. Again, the VIF scores for the model indicate the effects are independent.

The growth of followers for the Occupy users in this study calls to mind Rogers' (1995) theory concerning the rate that new ideas and inventions diffuse through cultures. Figure 4.5 illustrates Rogers' S-curve, with various stages of when different types of people would adopt the new idea or invention. Note that the Social Media Lab's corpus of tweets starts on October 19, 2011, about a month after the start of the Occupy movement. Thus, while the data appear to roughly fit the s-curve from between the early adopters to early majority, the data does not provide any insight into the rate of follower growth before October 19, 2011.

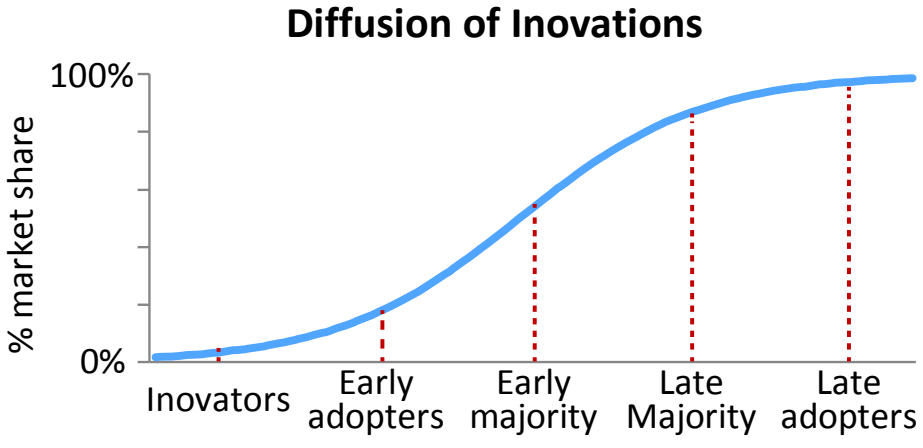


Figure 4.5. Rogers' graph showing the rate of diffusion of innovations.

Of course, other factors not included in the model play a role as well. For instance, it is reasonable to expect that the number and duration of stories related to Occupy from the major media outlets would play a role in who would become aware of, and follow, the Occupy accounts. Occupy activity on other social media sites would be a factor as well. However, there are two reasons to suspect the growth in followers did start out relatively slowly. First, a timeline of the Occupy movement on Wikipedia[31] suggests that media coverage of the movement was sparse in the first week or so. Thus new followers would have become aware of the movement through other sources, such as word of mouth, blogs, Facebook, Twitter and other social media. The second reason for potentially slower growth at the start is related to the iterative flow structuring model briefly discussed above.

Initially, when the Occupy accounts had few followers, the number of other Twitter users exposed to their tweets, and thus retweet them, would have been small. Each information flow that resulted in new followers would increase the size of the follower base for the next information flow, which could then reach an even greater audience. Users with large numbers of followers can be identified as the hubs in star formations in social networks. They can also be thought of as *gatekeepers*, "people, collectives, companies, or governments that, as a result of their location in a network, can promote or suppress the movement of information from one part of a network to another" (Nahon and Hemsley 2013, 7). The Occupy gatekeepers grew in *waves*. Each wave extended the Occupy gatekeeper's reach deeper and deeper into the network. Certainly different content will have influenced the process in different ways, but whether the tweet included a link to a Wall Street Journal article, a YouTube video, a blog post, or contained

[31] http://en.wikipedia.org/wiki/Timeline_of_Occupy_Wall_Street

information about a raid at an Occupy encampment, each of these information flows contributed to growth of followers over the period of this study.

The rate of followers would tend to grow slowly at first. At some point an epidemic threshold[32] of followers would be reached in the network and growth would become exponential (Watts, Peretti, and Frumin 2007). Over time the number of people in the network available to become followers, those both interested and previously unexposed to the messages, would shrink and growth would slow (the "late majority" section of the curve in figure 4.5). Note that this explanation doesn't take into account that users on Twitter can discover users and their messages by searching for tweets with key words.

Over time the increase in number of users following the Occupy accounts will result in network structural changes. This is because a net increase in the number of links in a network results in a *densification* of the network. Density of a network is a measure of the total number of links divided by the total *possible* number of links (Wasserman and Faust 1994). More links mean a denser network. Leskovec et al. (Leskovec, Kleinberg, and Faloutsos 2005) found the process of densification in networks to be fairly stable over time. Leskovec et al. also found that with densification comes a decrease in the *diameter* of the network, which measures the maximum shortest path between any two nodes in the network. Thus, as the Occupy accounts extended their reach deep into networks, they effectively brought the network closer and closer to themselves.

The last variable to discuss for model 2 is mean followers, which is significant and positively related to the change in followers (see figure 3.10). The positive relationship means that RTE initiators gain more followers when the users retweeting them have more followers.

[32] An epidemic threshold is the point at which growth becomes self-propelled and takes off exponentially. Other terms that mean the same thing are *phase transition*, *critical mass*, and Gladwell's *Tipping Point* (2002).

The selection of mean, instead of median, is intended to control for the presence of users with large numbers of followers within the set of users who retweeted the initial tweet. In other words, this variable will be larger when a gatekeeper retweets the initial tweet, suggesting a gatekeeper effect from users with a large number of followers is an important driver in users gaining new followers. This isn't surprising since gatekeepers tie together different parts of the network, which would tend to extend the potential reach of the message to parts of the network that would be otherwise inaccessible.

This gatekeeper effect supports the correlation earlier in this section: closeness centralization and change in followers have a positive relationship. Thus, as the path through which the RTE flowed is comprised of more star-like structures, indicating the presence of gatekeepers, greater gains in followers are observed. A negative relationship between closeness centralization and change in followers would imply that information flowing through more and longer chains would lead to more followers. The positive relationship seems to run contrary to what model 1 and model 2 indicate. Model 1 indicates that more and longer chains are related to more relaxed decay phases, and model 2 indicated that more relaxed decay phases are related to greater increases in followers. From this we might expect that closeness centralization would be negatively related to change in followers instead of having a positive relationship.

The explanation is that the signature is a function of the rate of retweets, not changes in followers. More relaxed signatures indicate that the rate of retweeting decreases slowly, or that retweets sustain their level over time. More and longer chains are related to sustained levels of retweets over time. However, holding all else equal, sustained levels of retweets over time (more relaxed signatures) are related to higher increases in followers. Note that closeness centralization

is anchored on the high side by a perfect star and on the low side by a single long chain.[33] Thus, the positive relationship between closeness centralization and change in followers indicates that stars (gatekeepers) have the greater effect in driving increases in followers. It seems reasonable to assume that more or larger stars will precipitate more and longer chains than smaller stars would. Both chains and stars drive increases in followers, gatekeepers, who connect dispirit networks, have a larger effect. More and longer chains have the larger effect in sustained levels of retweeting over time.

4.2 Theory Development

This study provides initial steps towards a theory that explains the interactions of dynamic social networks and flows of information. A critical initial step is conceptualizing the information flow by defining a *retweet event* (RTE) and precisely characterizing this RTE through its *signature*. A second step is to characterize the network dynamics in terms of changes in follower count of the initiator of the RTE and in terms of network measures of *density*, *centrality*, and *diameter*. Collectively, these concepts and their operationalization by measureable event and network parameters provide a foundation for developing a theory of dynamic social networks and information flows.

García-Murillo and Gozen (2012) state that "in general terms, a theory is an explanation of a phenomenon" (García-Murillo and Gozen 2012, 128), and that a *process theory* is one in which the order and relationship of events in time is explained. They state that that such theories focus on answering *how* questions. Pettigrew defines a *process* as "a sequence of individual and collective events, actions, and activities unfolding over time in context," (Pettigrew 1997, 338).

[33] Technically, a circle, where all nodes have the same number of links, will result in a closeness centralization of zero.

This study set out to understand *how* information flows and dynamic social networks interact. In other words, this study focuses on understanding how individual (retweet) and collective (RTE) events, situated in time and within the context of the participating users (mean followers), the path of the flow (closeness centralization) and the Occupy Movement (overlap of events and days into the event), are related to changes in the social network within which those events take place. The discussion above explains how the initial network state is related to the observed temporal signature of an RTE and how RTEs are related to both short term and long term changes in social networks.

Section 3.2.1 noted that Twitter data does not support the use of process analysis in this work because it is not possible to observe the actual process by which information flows and social networks interact. And yet Pettigrew's (1997) statement that the aim of a process analysis is to (1) identify patterns in the process and compare them across cases, (2) find the mechanisms that shape those patterns, and (3) and employ both inductive pattern recognition and deductive theory that offers a starting point for the analysis. Each of these is shown in this work: the EDA identifies patterns that resulted in the signature model; the confirmatory statistical analysis links the paths of flows to the observed patterns and changes in the networks; and the study employs inductive pattern recognition in the form of EDA and utilizes existing theory and empirical work to suggest the direction for the development of the signature model. So while the outcome has not been the case study that Pettigrew (1997) envisions for process analysis, the other criteria have been met. Indeed, Sabherwal and Robey (1995) suggest that using process and variance methods can be profitably combined.

For this work to develop into a more formal theory, additional work needs to be done that shows the generalizability of the explanations in different contexts. This could be done by

studying information flows on different platforms, or by performing detailed case study on a large information flow, or on a set of moderately sized flows, situated within a network. In addition, more work can be done to make the explanations more parsimonious so that others might test the theory or employ it in explaining processes they are observing.

4.3 The role of scripting in this work

O'Neil and Schutt (2013) surveyed self-identified data scientists and found that they spend between 70% and 90% of their time *data wrangling*. Data wrangling is the process of getting data ready for analysis. O'Neil and Schutt also discuss the importance of data exploration, noting that data scientists at Google include it in their list of best practices for making sense of and analyzing vast amounts of data. Data wrangling certainly makes up a large portion of the work reported in this dissertation, but it is difficult to untangle the process of exploring the data from wrangling it. Together, exploring and wrangling easily account for 85% of this research project.

In this research, data wrangling and EDA required custom software scripts. As of this writing, about 6,000 lines of code[34] can reproduce this research.[35] The amount of code written and abandoned is far greater than this.[36] The effort of writing the abandoned code, though, has not been wasted. Section 2.2.2.2 notes that the process of writing software to explore the data confers knowledge about the data, and that fixing defects in the code can lay bare erroneous assumptions build into the code about the data.

A one-sided discourse with the data exists when software is iteratively developed to explore the data using visualizations. The researcher asks questions of the data and interprets the responses—or rather, the researcher interprets a visual representation of the data, which is itself

[34] All of the analysis code is written for R (http://www.r-project.org/). Scripts written in both Python and JavaScript queried and extracted data from the SoMe 'Lab's Mongo database.
[35] https://github.com/jhemsley/Dissertation
[36] This can also be made available upon request, but is currently not checked into a public repository.

an interpretation of the data performed by the algorithm. It is tempting to claim that the visualization is a reflection of the data – that it represents the data in an objective way. It does not. The researcher-programmer imbues the algorithm that creates the visualization with "a particular knowledge logic, one built on specific presumptions about what knowledge is and how one should identify its most relevant components " (Gillespie 2014, 168). When the code is written well, the algorithm encapsulates the researcher-programmer's interpretive intention.

Certainly there is danger in this. Poorly developed software may not encapsulate the researcher-programmer's interpretive intention but can nonetheless perform some interpretation. Confidence in the algorithm is gained through its systematic development and rigorous testing. Systematic development includes practices such as developing and testing small units of code before integrating them into the system, maintaining version control that supports examining previous versions and maintaining a record of change, and including comments in the code. Rigorous testing practices include writing software to test the software under development, stepping through the code as it is running to verify the expected behavior matches the actual behavior, and verifying small samples of the output before executing large automated runs. Further, reproducing the findings of this work would lend credibility to the software and findings. However, this would best be done by creating new scripts based on the text of this document rather than re-running the software used in this work. If the results match, then the interpretive intention embedded in the software matches the text of this document.

The interpretive layer of software and visualization mediates the researcher's interactions with the data. The visualization becomes a data point that is interpreted by the researcher-programmer. Alternately, the visualization is interpreted socially with research peers through a sensemaking process. The mediating effect of the visualization has value. Just as documents

have been shown to extend the human voice through time and space (Levy 2001), the visualization of data extends the reach of researcher-programmers through vast amounts of human trace data. Utopian and dystopian rhetoric about 'big data' (boyd and Crawford 2011) aside, researcher-programmers do have an unprecedented ability to observe (as this study does) emergent signals from aggregated individual human interactions at large scale.

A final note on software: it makes for excellent field notes. The software makes explicit and reproducible the exact steps employed in the data wrangling for this study. Writing the data sections of this document essentially involved translating R scripting language into English. The sequential narrowing of the data presented in this document reflects the fact that R executes code sequentially. As a researcher-programmer, I can go back to the code at any time and demonstrate exactly the chain of evidence used in this study.

5 Conclusions

5.1 Summary

This study employs exploratory data analysis (EDA) and confirmatory statistics to study the relationship between information flows and dynamic social networks. The research is in two phases, with the EDA preceding the confirmatory analysis. The EDA consists of inductive pattern recognition of large numbers of data visualizations, more direct exploration through the development of software, and social sensemaking of the visualizations. The findings, summarized below, include the ways in which the relationship between information flows and dynamic social networks can be studied. Variance regression models examine and confirm the existence and nature of these relationships.

5.1.1 Findings and Implications

Phase 1 yields two findings. The first of these is a stylized model of information flows defined by a set of parameters that describe the temporal shape of the retweet event. The model is referred to as a *Signature Model*. The second result is an operationalization of changes in networks that provides a way to quantify structural changes in networks without specific link information. These two findings provide a means for statistically testing relationships between structural changes and the parameters of information flows. It answers the research question for Phase 1, *How can Twitter data be used to explore the relationship between the flow of information and dynamic social networks?*

Phase 2 confirms that the parameters of the Signature Model relate to the path of the flow of information. Specifically, the decay phase of the signature model tends to be more relaxed (extend longer) when chains of social sharing are longer and more frequent. The decay phase tends to sharpen (drop off more quickly) when the path of the flow, as represented by a network topology, is more star-like. Closeness centralization quantifies the path of flow as low (more and longer chains) or high (more star-like). This result answers the first research question for Phase 2, *How is closeness centralization related to the shape of the decay phase of a signature?*

Phase 2 also confirms that the parameters of the Signature Model demonstrate an association with changes in the topological linking structure of the social network through which the information flowed. Specifically, more relaxed decay phases, higher peaks, and earlier peaks each is associated with greater increases in the number of followers of the user who initiated the information flow. Changes in the number of followers during the information flow operationalize changes in the linking structure of the social network. This finding answers the second research

question for Phase 2, *How is the shape of the decay phase of a signature related to changes in the number of followers of users who initiate RTEs?*

The findings from Phase 2 imply that the network at the end of a given information flow is the initial condition network state for the start of future information flows. The findings also imply that hub structures, or gatekeepers in the Occupy movement, gain followers and extended their reach recursively. In other words, when an information flow brings a new set of followers, those users increase the potential reach of the Occupy gatekeeper for the next information flow. Over the period of time studied, the rate at which these gatekeepers gained followers slows, possibly reflecting network saturation in terms of the number of interested users who are likely to create new follower ties in the network.

This study provides the initial steps towards a theory that explains the processes by which dynamic social networks and information flows, as modeled using signatures, interact. Additional work that confirms the generalizability on different platforms and different contexts could lead to a more formal theory.

5.1.2 Limitations

This section highlights four important areas of limitations. These are (1) limitations in the signature model, (2) limitations in the method used to infer the path of the RTEs, (3) limitations due to potential inflation of the change in followers for RTEs, and (4) the generalizability of this study.

5.1.2.1 Signature Model

This study operationalizes an information flow as an RTE. Each RTE is derived from tweet metadata. When users retweet using Twitter's Retweet button, the interaction is captured in the metadata of the retweet and included in the RTE. Users can manually retweet by appending

the text of the retweet with "RT: @USERNAME." These *manual retweets* are not included in an RTE because there is no metadata trace of the relationship. In a blog post, Hemsley (2012) noted that manual retweets make up about 8% of the retweets in a dataset of #OccupyOakland tweets drawn from the SoMe Lab corpus of tweets. The effect on the signature for each affected RTE should be to underestimate the rate, and thus the height, at the point in time when the manual retweet was posted. Unless manual retweets consistently occur at the same time relative to the peak, this limitation should not change the pattern found in the shape of the signatures.

The signature model is also limited: Because all of the RTEs analyzed have a minimum of 100 retweets, the model represents relatively large flows of information. It does not capture patterns that might exist in smaller-scale events. It also does not represent events with multiple peaks (see section 2.3.1 for more discussion). The analysis focused on the maximum rate as the peak and truncated retweets more than two hours before or after the peak. This means that retweets that occur days or weeks after the initial tweet are not represented in RTEs and thus not captured in the model. Note that 80% of retweets in the dataset occur between the time of the initial post and the end of the two hour window. Thus the model represents the majority of cases.

5.1.2.2 Inferred Path Method

Recall that the actual path of RTEs is unobservable and so a maximum likelihood model (Gomez-Rodriguez, Leskovec, and Krause 2010) is used to infer the path of the information flow. Gomez-Rodriguez et al. (2010) show that the method discovered over 90% of the edges in a network. An important limitation of the method is that it assumes that there are no gaps in the flow, that everyone in the path network is connected to everyone else. Thus it fails to capture cases where someone found a tweet during a search and retweeted it. Users who retweeted from a search will be treated identically to those who retweeted from others in the RTE. This study

assumes that these users make up a small fraction of the retweeters in an RTE. This study also assumes a random distribution for the time when, relative to the peak time of the RTE, users find and retweet messages. As such, the model treats these found and retweeted tweets as noise in the model. Finally, an inferential model represents one explanation of the observed data. Other models could also explain the observed data.

5.1.2.3 Change in Followers

There are three ways that the change in followers for RTE may be inflated, compared to the number of retweets in the RTEs. The first is due to the exclusion of manual retweets. The increase in followers detected during the RTE will include any new followers related to manual retweets. As mentioned above, these manual retweets will not be present in the RTE and so the signature will not reflect them. Unless manual retweets consistently occur at the same time relative to the peak, which seems unlikely, the significance and nature of the relationship among the model parameters and the change in followers should not be affected; the only thing affected would be the strength of the relationship relative to the other variables in the model. As a special case, a manual retweet could spawn a related RTE that will not be captured because users would be retweeting the manual retweeter. The size of RTEs follows a power-law; because these related RTEs should be relatively small, it is reasonable to expect, even in these cases, that the analysis ought not be strongly biased.

Another way that the change in followers for RTEs could be inflated is related to *rate limiting*. Twitter allows users to collect up to 1% of the total volume of tweets when using the Streaming API. If Twitter imposed a rate limit during an RTE, those tweets would not exist in the Social Media Lab corpus of tweets. Additionally, Driscoll and Walker (2014) note that 2.5%

of the tweets were missing from the Streaming API[37] when compared to data collected from Twitter's fire hose.[38] Thus some RTEs may have missing retweets due to the limitations of the API. Rate limiting likely occurred on November 15th, 2011, the day the New York City Police Department cleared Zuccotti Park of protestors. If this were a serious issue, then the affected RTEs would have been detected as outliers in the regression models.

Finally, Twitter users can gain followers during RTEs that they initiated for reasons unrelated to the RTE. For example, a person may become part of the OccupyWallSt account from the news, a blog post, or a Facebook post; or they could sign into their Twitter account and follow the OccupyWallSt account directly. This is particularly possible during times of heavy news coverage, such as November 15, 2011, the day the New York City Police Department cleared Zuccotti Park of protestors. This effect may inflate the gain in the number of followers associated with each RTE. However, the fact that RTEs are limited to a 2-hour window after the peak of the event, meaning that inflation to users must follow the Occupy accounts during the window, should mitigate this effect.

The generalizability of this study is limited in two important ways. First, the exploratory analysis included only Tweets related to the Occupy Movement. The signature model is thus limited to Occupy or, possibly, to similar political and social movements. More work will need to be done to extend this model to different contexts. Second, because the regressions include all of the available observations, the findings are generalizable only to the Occupy dataset, and specifically, to the Occupy user accounts.

[37] It should be noted that the streaming API collected nearly 4 times as many tweets as was collected from the Search API when using the same search terms.
[38] The fire hose returns all tweets in real-time.

5.2 Contributions

The following are the main contributions of this research:

5.2.1 Making explicit the exploratory phase

This dissertation contributes to methods of studying large datasets by providing details about the use of EDA in this work. The EDA employed Twitter data, but the exploratory techniques are suitable for making sense of a wide range of unstructured data. This could include data from other social media sites to data from non-human sources, often referred to as the Internet of Things. Future researcher-programmers may draw on the details included in this work to determine if existing datasets could be useful in answering their questions, understanding the limitations and quality of the data, and discovering structures and patterns in the data. Details expected to be helpful include the use of professional photography software to explore, organize and find patterns in large quantities data visualization; the use of both bold features and fine details in plots that facilitate exploring features across plots and within them; and the use of social sensemaking as a means for identifying new avenues of exploration. This work also calls attention to the use of software in both exploring the data and as field notes that make explicit the assumptions embedded in the work of data wrangling and exploration. The dissertation discusses risks associated with using multiple interpretive layers and suggests best practices for lowering these risks and ensuring rigor.

5.2.2 The Signature Model

This study introduces the signature model as a temporal model of information flows. The parameters of the model are the time to peak, the peak rate, alpha (the shape of the decay phase), and the size of the event. The model is a useful lens for examining the relationship between

information flows and network dynamics. Researchers can use the signature model to make comparisons of information flows across different platforms and content types. The model may also be useful in examining the temporal behavior of citations in knowledge networks. For example, it might help in early identification of emerging fields that have widespread impact.

5.2.3 Operationalization of network dynamics

This dissertation demonstrated the usefulness of three methods of studying network dynamics. First, the study extended the method for inferring the path of information diffusion (Gomez-Rodriguez, Leskovec, and Krause 2010) by including a parameter for the followers of users in the network. This is assumed to more accurately model Twitter networks than simply using retweet time stamps. Second, this study operationalized the structure of the path through which information flows using closeness centralization, which allows for the differentiation of paths that are star-like vs. those that flow through more and longer chains. Finally, this study shows that by tracking the change in the number of followers of users who initiated RTEs, researchers can measure changes in the linking structure of social networks when explicit linking information is not available. These approaches can be used together or individually in studies seeking to understand aspects of network dynamics.

5.2.4 The interaction of network dynamics and information flows

This study provides new insight into how protest networks like Occupy grow and how gatekeepers within those networks gain followers. More generally, this works provides a set of methods for exploring how information flows and network dynamics interact in other contexts. Further, by examining individual flows in an interest or topic network, the use signatures may be useful in identifying (1) network altering information flows, and (2) flows that reach deep into networks through long sharing chains. Individuals and organizations interested in identifying

information flows with these kinds of impacts can construct and examine signatures within two hours of the peak of an information flow. Thus this work has applications in research, industry, activism and government agencies.

5.2.5 Theory Development

This study provides initial steps towards a theory that explains the process by which social network dynamics and information flows interact. The work introduces two important conceptualizations. First, an information flow is operationalized by defining a *retweet event* (RTE), which is precisely defined by parameters that make up a *signature*. Second, network dynamics are characterized in terms of changes in follower count of the initiator of the RTE and in terms of network measures of density, centrality, and diameter. Analysis of empirical data demonstrates how these conceptualizations are related and the sequence in which events are related. More work can extend the utility of this theory by employing the signature model in future research.

5.3 Future directions

The signature model should prove useful in studying information flows in a number of situations. Literature discussed in section 3.1.2. highlights that rich content, content with emotional appeal, or content rated as interesting, are all factors in tweets being retweeted. Future studies can explore how content type may be associated with specific parameters of the signature model. For example, interesting content might have longer ramp-up times than content with emotional appeal, or the relationship between peak height and the shape of the decay phase may be weaker for emotional content than for interesting content. Likewise, studies that examine how different platforms propagate different kinds of information (see Nahon et al. (2013) for an

example) might also find differences in the signatures of the events they study. Because the signature can provide clues as to both the path of information flow and how much the network changes, understanding these relationships can provide insight into how different kinds of content interact with networks.

Almost all of the RTEs in this study resulted in net follower gains for the initiator of the event. In what cases might we see spreads with losses or zero gains, and how might these be related to the shape of the decay phase? Content analysis and topic modeling both appear to be useful methods to mix with the variance models that relate content to the parameters of the signature model. If different types of content result in significantly different signatures, then watching signatures may be one way of identifying the type of content flowing through networks.

Another fruitful approach to studying the interaction of information flows and network dynamics would be to employ a detailed process analysis on a few specific events. Such an examination could provide more depth to the context surrounding the flows and could explore the motives behind users' decisions to propagate information.

This dissertation provides a foundation for a series of studies that can provide greater understanding of the processes of information flows in social networks. The accelerated pace of information exchange afforded by new social media platforms, suggests that the methods and models employed in this work may find utility a wide variety of applications in the social media ecosphere.

Appendix A: Glossary

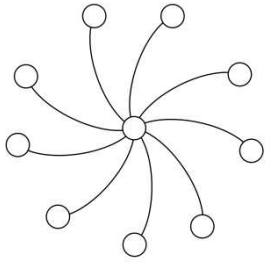
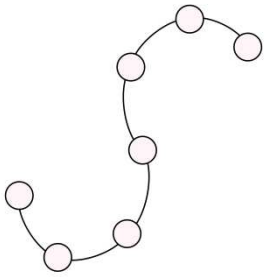
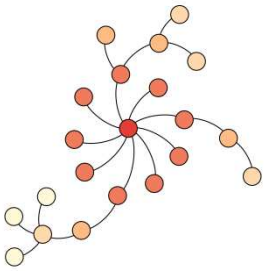
alpha. The shape parameter of a power-law distribution. The formal equation is $1/t^{\alpha}$, where *t* is time.

Application Programming Interface (API). An interface to a software system or platform that requires utilizing a programming language to access.

bots. Software agents that operate on social media to fulfill some purpose. For example, a bot can be used to send spam, aggregate content, or otherwise automate a task for a human, organization, or other entity. An example collection of OccupyBots can be found at https://twitter.com/cdubs/occupybot/members

bridge. A bridge is a link that spans a structural hole (see below) in a network. Note that bridge can also refer to an actor at either end of the bridge.

chain. Network topology where information flows from some node A to B and then from B to C (see figures below).

		
Star network formation	Chain network formation	Combination of chains and a star

cluster. A collection of nodes in a network that are more densely connected to each other than to other nodes.

core. Represents the central part of the network, typically as measured by number of links (in, out or both) of the actors; however, other concepts of centrality can be applied.

decay phase. A period of time after a peak of activity in an information flow event characterized by a decline in activity.

degree centrality. At the individual node level, this is a measurement of the number of links (in, out or both) that a given node has. At a network wide level this is typically a measurement of the average or median of the degree measurement for all nodes in the network.

densify, densification. An increase in the density (see below) of a network over time.

density. Refers to number of links in a network or cluster. Networks with more links are considered denser. Formally, it is the total number of existing links divided by the total number of possible links.

diameter (network). The diameter of a network is maximum minimum path between any two nodes.

endogenous (event). A shock to a system from inside of that system that precipitates events within the system. In the context of studying information flows on social media this may be a video or tweet that is spread around social media.

exogenous (event). A shock to a system from outside of that system that precipitates events within the system. In the context of studying information flows on social media, this may be a news story that starts on, for example, at the San Francisco Chronicle and spreads within social media.

***information flow*.** For this study an information flow is a single unit of content or links to a specific unit of content: a tweet or retweet, a video or links to it, a news story or links to it, a blog post or links to it, are all considered a single unit that, when shared, are the objects of an information flow or information flow event. Certainly an information flow could include comments, memes, related tweets, response videos, mash-ups, or any of a myriad of ways that a discussion could be carried on. Indeed, information that is reframed, re-contextualized, shared across platforms, discussed over lunch, or lectured about could all be considered part of the same flow. These are all beyond the scope of this project.

***interest network*.** A temporally bound, self-organized network in which membership is based on an interest in the information content or in belonging to the interest network of others.

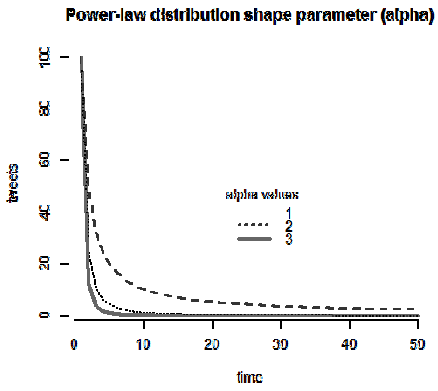
***network*.** Generally, a collection of entities and the relationships between them. For this study, the entities are thought to be social actors, which are generally people but can be organizations, institutions, or *bots*.

***network holes*.** Conceptually, holes are open spaces between clusters with no or few links (Burt 2004). Links that span holes are referred to as *bridges*.

***network rewiring*.** When a node, user, or person in the network adds a new link or deletes a link to another member of the network.

***network structure*.** The patterns created by the links of a network; the topology (for example, star or chain).

***power-law*.** A probability distribution mathematically expressed as $f(x) = 1/x^{\alpha}$, where α is called the *shape parameter*. When plotted, a power-law is characterized by a tall peak that slopes sharply downward and then levels off. The shape parameter effects how quickly the distribution levels off with higher values leading to sharper slopes and lower values leading to more relaxed declines.



retweet event (RTE). An information flow event (see above) on a specific social media platform, Twitter. A collection of tweets made up of a tweet and all of the subsequent retweets of it. The "origin tweet" or simply "origin" is the first tweet in the event that other people retweeted. In general set terms a RTE is expressed as `RTE ∈ {OriginTweet, RT1, RT2, … RTn}`

RTE path. A directional, ephemeral network created when one or more users forwards (retweets) a tweet; includes all retweets of the origin tweet.

share. Generic term for posting content on a social media regardless of the platform or of whether the content is being posted the first time or as a result of a larger information flow. Thus when someone tweets, they share, and when someone else retweets that tweet, they are also sharing (though the term can apply to other types of content on other platforms as well). When it is necessary to distinguish between an origin post (first post) and posts later in the information flow, share and re-share (or sharing and re-sharing) shall be used.

***sharing chains*.** A flow of information including at least 2 links and 3 nodes such that: A → B → C, where A, B and C are users who share information.

signature. A time series data structure that captures the volume of flow for a given time frame. For example, viral videos are often graphed as views per day.

size of a retweet event. Numeric; the total number of retweets for a given origin tweet.

social media (SM). The set of Internet platforms that enable "people to connect, communicate, and collaborate" (Jue, Marr, and Kassotakis 2009, 44), which includes blogs, wikis, content sharing sites, and social network sites (Avram 2006; Jue, Marr, and Kassotakis 2009).

***star*.** Network topology where a single central node *A* is connected to all other nodes, and where information from one node to another must first pass through the central node (see figures above).

***user*.** The entity—person or persons, organization, or software agent (see bot) – associated with an account on a social media site. For this study it is assumed that users have a one-to-one mapping to accounts, but certainly this is not always the case. The term *user*, rather than person or individual, is used because not all accounts are associated with a *person*.

weak ties. Granovetter suggested that "the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (1973, 1361).

Appendix B: Twitter & Social Media Lab data details

B.i: Twitter Streaming API process overview

Twitter returns status updates in JSON (JavaScript Object Notation) format, which is a simple text-based key-value pair format. For example, the text of a tweet is indicated in the data with the *key*, "text", followed by a colon and then the *value*, or the tweet text (see appendix C for a full example):

```
"text" : "Jeff sent this tweet text as an example"
```

When using Twitter's Streaming API, an HTTP request is sent to Twitter's severs that includes a list of search terms. A tweet is returned (streamed back) if any tweet text, hashtags, @-mentions, or URLs within the tweet match any of the search terms sent in the HTTP request. Twitter streams the tweet back over the original HTTP connection. As long as the connection is left open, Twitter will continue streaming back matching tweets. The Social Media Lab based its initial search term list on a spreadsheet of Occupy-related accounts and hashtags maintained by the movement's OccupyTogether website (http://www.occupytogether.org/). The list, loosely curated by members of the lab, is included below.

B.ii: Social Media Lab collection system

Figure B.1 is an illustration of the Social Media Lab collections system (Walker et al. 2013). Amazon Web Services (AWS) hosted the system for the majority of the project. Except for the some of the analysis (and data wrangling), most of the data flow steps occurred on AWS Ubuntu serves. A lightweight application called cURL (written in 'c,' its primary function is to make and deal with HTTP requests) established the actual HTTP connection. A BASH shell

script monitored the cURL application and reestablished the connection to the API as needed. Together, the cURL application the BASH script stored the raw JSON data objects in text files on a daily basis; that is, one file contained a single day's tweets, where each line was a single JSON tweet object. These files were compressed and saved as raw data. Python scripts read the data, transformed date strings into date objects and inserted them into a Mongo ("MongoDB" 2013) database, a noSQL database that stored each JSON object as a single document and provided fast query functionality on key value-pairs for large non-structured data. As data was needed for analysis, queries extracted sets of data into tab-separated files, which were downloaded for analysis in R.

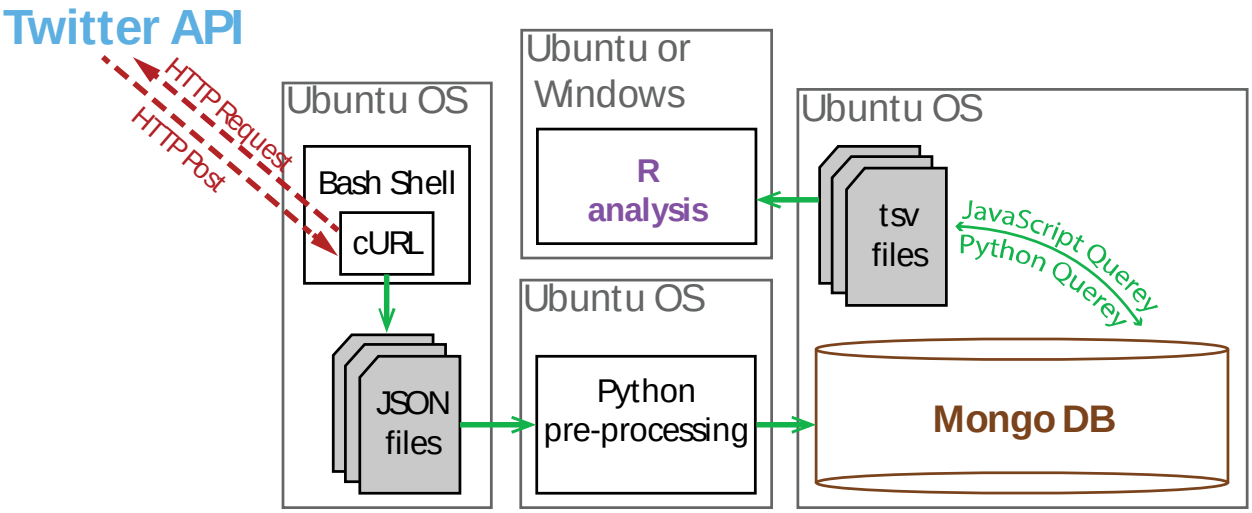


Figure 5.1: Social Media Lab collection, processing, storage, and analysis system.

B.iii: Social Media Lab search term list

12d	17nov	30n
15o	1percent	53percent
15oct	29oct	5nov
17n	2nov	99percent

acampadamataro
acampvalladolid
acampvalladolid
acorn
banktransferday
bofa
cabincr3w
d21
frankietease
generalstrike
ioccupy
ioccupyoccupyashland
k8_revolution
lakajo97
louderthanwords
moveyourmoney
n17
n30
needsoftheoccupiers
nov17
nov2
nov30
nov5
occopywmpt
occuponsmontrea
occupy
occupy_albanyny
occupy_boston
occupy_okc
occupy_ottawa
occupyaarhus
occupyabilene
occupyadelaide
occupyafrica
occupyafrica1
occupyakron
occupyalbany
occupyalbanyny1
occupyallentown
occupyamsterdam
occupyanchorage
occupyannarbor
occupyappleton
occupyarcata
occupyarizona
occupyarkansas
occupyarrests
occupyashland
occupyashlandky
occupyaspen
occupyastoria
occupyathens
occupyathensga
occupyatl
occupyatlanta
occupyatlanticcity
occupyatlcity
occupyauburn
occupyaugusta
occupyaurora
occupyaustin
occupyb0ulder
occupybaltimore
occupybhgrove
occupybkny
occupyboise
occupyboston
occupyboulder
occupyboulderco
occupybrisbane
occupybrussels
occupybucharest
occupybuffalo
occupycarsoncty
occupycc
occupycha
occupychi
occupychicago
occupychucktown
occupycincinnati
occupycincy
occupyclarksvil
occupycleveland
occupycolumbia
occupycosprings
occupycu
occupycville

occupydallas
occupydc
occupydelaware
occupydenhaag
occupydenmark
occupydenver
occupyearth
occupyeugene
occupyflorida
occupyfm
occupyfortmyers
occupyftcollins
occupygtown
occupyhardford
occupyhartford
occupyhouston
occupyhsv
occupyhumboldt
occupyindy
occupyisu
occupyitaly
occupyjax
occupykeene
occupykelowna
occupykingston
occupyla
occupylansing
occupylasvegas
occupylausd
occupylondon
occupylsx
occupymadison99
occupymartnsbrg
occupymemphis
occupymia
occupymilwaukee
occupymn
occupymontrea
occupynashville
occupynewportor
occupynj
occupyns
occupyoakland
occupyobise
occupyokc
occupyomaha
occupyorlando
occupyorlandofl
occupyottawa
occupypei
occupyphoenix
occupypics
occupyportland
occupyprov
occupyquebec
occupyraleigh
occupyredlands
occupyrichmond
occupyroanokeva
occupyrockford
occupysacto
occupysalem
occupysananto
occupysanjose
occupysantacruz
occupysarasota
occupysarasotaoccupysanjose
occupysaskatoon
occupysb
occupysd
occupyseattle
occupysenhaag
occupyslc
occupysr
occupystaugust
occupystl
occupytampa
occupythemedia
occupytoronto
occupyueg
occupyukiah
occupyvermont
occupyvictoria
occupywallst
occupywallstnyc
occupywallstreet

occupywinnipeg	oct29	rebuilddream
occupywmpt	oo	storydoula
occupywv	ows	strike
occupyyakima	owslosangeles	tokumtorgin
occupyyeg	owsspacecoast	usdor
occupyyork	perversmas	wearethe99
occypyftcollins	quimbanda	

Appendix C: Example JSON tweet

Twitter's Streaming API returns tweet data in Java Script Notation Object (JSON) format. Below is an example of one entire tweet, or *status update*, collected from the streaming API. This tweet is also a retweet, which means that the entire origin tweet is embedded in it. This example tweet is a modified tweet: the contents of certain fields anonymize the tweet.

For convenience, the following formatting has been added to highlight certain parts of the data:

- Fields used in the analysis are in bold and underlined.
- The text of the metadata associated with a user is colored red.
- The embedded retweet is set on a grey background.


```
{
	“_id” : ObjectId(“50f6616854b450b3ce1”),
	“contributors” : null,
	“truncated” : false,
	“text” : “RT @someperson: some tweet text #aHashTag http://t.co/someurl”,
	“hashtags” : [“aHashTag”],
```

```
“text_hash” : “9a04f77e5883ed69cd0e1ae67a7d1”,
“in_reply_to_status_id” : null,
“counts” : {
       “user_mentions” : 1,
       “hashtags” : 1,
       “coded_urls” : 0,
       “urls” : 1
},
“id” : NumberLong(“131050009612078080”),
“codes” : [ ],
“source” : “web”,
“retweeted” : false,
“coordinates” : null,
“entities” : {
       “user_mentions” : [
              {
                     “id_str” : “197999360”,
                     “id” : 107010360,
                     “name” : “Almost Home”,
                     “screen_name” : “Almost_Home”
              }
       “hashtags” :
              {“text” : “aHashTag”}
       “urls” :
              {
              “url” : “http://youtu.be/imkOCwgU2ms”,
              “expanded_url” : “http://youtu.be/imkOCwgU2ms”,
              “display_url” : “http://youtu.be/imkOCwgU2ms”
              }
},
“in_reply_to_screen_name” : null,
“id_str” : “73705797967207878”,
```

```
        “track_kw” : {
                “mentions” : [ ],
                “hashtags” : [“aHashTag”],
                “text” : [ ]
        },
        “retweet_count” : 1,
        “in_reply_to_user_id” : null,
        “favorited” : false,
        “retweeted_status” : {
                “favorited” : false,
                “entities” : {
                        “user_mentions” : [ ],
                        “hashtags” :
                                {“text” : “aHashTag”}
                        “urls” :
                                {
                                        “url” : “http://t.co/someurl”,
                                        “expanded_url” : “http://youtu.be/imkOCwgU2ms”,
                                        “display_url” : “http://youtu.be/imkOCwgU2ms”
                                }
                },
                “contributors” : null,
                “truncated” : false,
                “text” : “ some tweet text #aHashTag http://t.co/someurl “,
                “created_at” : “Mon Oct 31 17:20:31 +0000 2011”,
                “retweeted” : false,
                “in_reply_to_status_id_str” : null,
                “coordinates” : null,
                “in_reply_to_user_id_str” : null,
                “source” : “web”,
                “in_reply_to_status_id” : null,
                “id_str” : “131057978525757440”,
```

```
            “place” : null,
            “user” : {
                    “follow_request_sent” : null,
                    “profile_use_background_image” : true,
                    “id” : 197910360,
                    “verified” : false,
                    “profile_image_url_https” : “<<clipped for space jh>>“,
                    “profile_sidebar_fill_color” : “252429”,
                    “is_translator” : false,
                    “geo_enabled” : false,
                    “profile_text_color” : “666666”,
                    “followers_count” : 397,
                    “protected” : false,
                    “location” : “Chicago, IL”,
                    “default_profile_image” : false,
                    “id_str” : “197910360”,
                    “utc_offset” : -21600,
                    “statuses_count” : 1209,
                    “description” : “<< user account description>> “,
                    “friends_count” : 900,
                    “profile_link_color” : “2FC2EF”,
                    “profile_image_url” : “<<clipped for space jh>>“,
                    “notifications” : null,
                    “show_all_inline_media” : true,
                    “profile_background_image_url_https” : “<<clipped for space jh>>“,
                    “profile_background_color” : “1A1B1F”,
                    “profile_background_image_url” : “<<clipped for space jh>>“,
                    “screen_name” : “someperson”,
                    “lang” : “en”,
                    “profile_background_tile” : true,
                    “favourites_count” : 0,
                    “name” : “Almost Home”,
```

```
                    “url” : “http://www.almosthomeproject.com”,
                    “created_at” : “Sat Oct 02 20:16:53 +0000 2010”,
                    “contributors_enabled” : false,
                    “time_zone” : “Central Time (US & Canada)”,
                    “profile_sidebar_border_color” : “181A1E”,
                    “default_profile” : false,
                    “following” : null,
                    “listed_count” : 4
                },
                “in_reply_to_screen_name” : null,
                “retweet_count” : 1,
                “geo” : null,
                “id” : NumberLong(“131057978525757440”),
                “possibly_sensitive” : false,
                “in_reply_to_user_id” : null
        },
        “user” : {
                “follow_request_sent” : null,
                “profile_use_background_image” : true,
                “id” : 3000300010,
                “verified” : false,
                “profile_image_url_https” : “<<cut for space>>“,
                “profile_sidebar_fill_color” : “DDEEF6”,
                “is_translator” : false,
                “geo_enabled” : true,
                “profile_text_color” : “333333”,
                “followers_count” : 195,
                “protected” : false,
                “location” : “home”,
                “default_profile_image” : false,
                “id_str” : “387369210”,
                “utc_offset” : -21600,
```

```
        “statuses_count” : 12318,
        “description” : “<<cut for space>>“,
        “friends_count” : 0,
        “profile_link_color” : “0084B4”,
        “created_ts” : ISODate(“2011-10-08T23:02:29Z”),
        “profile_image_url” : “<<cut for space>>“,
        “notifications” : null,
        “show_all_inline_media” : true,
        “profile_background_image_url_https” : “<<cut for space>>“,
        “profile_background_color” : “C0DEED”,
        “profile_background_image_url” : “<<cut for space>>“,
        “screen_name” : “someperson”,
        “lang” : “en”,
        “profile_background_tile” : false,
        “favourites_count” : 0,
        “name” : “some name”,
        “url” : “”,
        “created_at” : “Sat Oct 08 23:02:29 +0000 2011”,
        “contributors_enabled” : false,
        “time_zone” : “Central Time (US & Canada)”,
        “profile_sidebar_border_color” : “C0DEED”,
        “default_profile” : true,
        “following” : null,
        “listed_count” : 13
},
“geo” : null,
“in_reply_to_user_id_str” : null,
“possibly_sensitive” : false,
“created_ts” : ISODate(“2011-10-31T17:20:31Z”),
“created_at” : “Mon Oct 31 17:20:31 +0000 2011”,
“in_reply_to_status_id_str” : null,
“place” : null,
```

```
        “mentions” : [ “someperson” ]
}
```

Appendix D: Inferring and measuring the flow path

This appendix provides details on the use of the Gomez-Rodriguez et al. (2010) method for inferring the path of information diffusion and the use of social network analysis centrality measurement, referred to as closeness centralization, to measure the path of the flow. For each RTE, the method infers the path, and then closeness centralization captures the degree to which chains exist in the flow. In other words, closeness centralization is used to capture both the number of chains and the height of the tallest sharing tree.

Gomez-Rodriguez et al. (2010) note that while information diffusion typically happens on a network, researchers are often unable to observe the actual path of the flow. However, attributes of the node are observable, such as the time a node has been infected (shared information), or, in the case of this study, the user name and the number of each user's followers. The purpose of their work is to discover the underlying social network over which the diffusion happens. To accomplish this, they model a large number of *cascades*, or individual information flows, and then find the most likely underlying network based on the modeled cascades. They model the cascades themselves using a maximum likelihood method, which this study also employs to model the paths of flows in the RTEs.

For each retweet in an RTE, a square matrix contains the number of rows and columns equal to the number of users involved in the RTE. Users, ordered by the timestamp of when they retweeted (or tweeted in the case of the initiator of the event), form the labels of the rows and columns. The label of the first row and column is the initiator of the RTE user name, while the label of the last row and column is the last user to retweet in the RTE window. The cells contain the difference in time between users (see table D.1). This results in negative numbers below the

diagonal and zeroes along the diagonal, though zeroes can also exist in cases where the timestamps are identical.

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

Table D.1: example time difference matrix

Next to be applied is the wait time distribution. Gomez-Rodriguez et al. (2010) find that using either a power-law or an exponential distribution for the wait times performs equally well at discovering links in their simulated network data. This study employs the power-law, with the shape parameter (alpha) set to 2 (shown in the equation below), which is consistent with Gomez-Rodriguez et al. Taking the inverse of the time difference raised to alpha yields the probably for each given time difference in the matrix.

$$P(\Delta) \propto {}^{1}/_{\Delta^{\alpha}}$$

The values on and below the diagonal are zero to indicate that there is zero probability that someone retweets themselves in a RTE. The values on and below the diagonal are zero to indicate that there is zero probability that someone can retweet someone who retweets after them. The resulting matrix contains an upper triangle where the cells indicate the probability that the user labeled by the column retweeted the user indicated by the row. The maximum value in each column is identified, and the row label for that value represents the row-user for whom the column-user is most likely to have retweeted.

This is the extent to which Gomez-Rodriguez et al. (2010) model cascades. They do note, however, that the model can be made arbitrarily complicated to capture attributes of the user of the information. In the context of this study, it is reasonable to assume that if two users retweet the same message, all else being equal, the one with more followers is more likely to trigger additional retweets. This is particularly true because some users have very high numbers of followers. For example, when Russell Brand, an English comedian and actor with the twitter handle *rustyrockets*, retweeted one of OccupyWallSt's tweets, he had 3,485,361 followers. Figure D.1 shows a second small peak directly following the rustyrockets retweet.

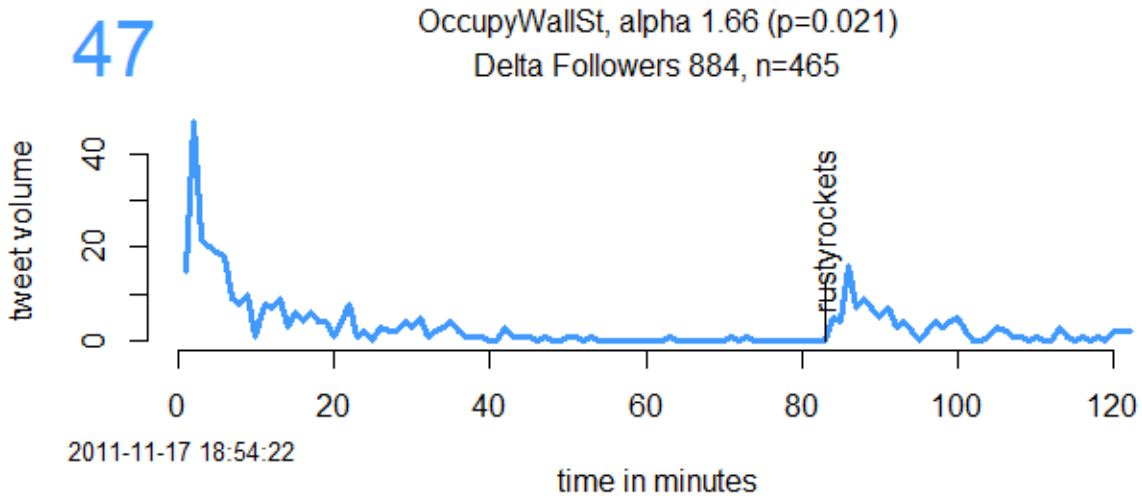


Figure D.1: Example RTE with a pronounced second spike, likely due to a retweet by a celebrity with many followers.

A new matrix captures the large follower effect and includes it in the model. This matrix weights sending users (the rows) by the proportion of followers, and then completes a cell-by-cell multiplication with the time difference matrix. Dividing a user's number of followers by the

sum of all user's followers results in the proportion of followers. Each cell in a user's row contains their proportion of followers (see table D.2).

	A	B	C	D	E
A	.012	.012	.012	.012	.012
B	.00003	.00003	.00003	.00003	.00003
C	.004	.004	.004	.004	.004
D	.00005	.00005	.00005	.00005	.00005
E	.00002	.00002	.00002	.00002	.00002

Table D.2: example follower proportion matrix

When the cell-by-cell multiplication of the follower proportion matrix and wait time matrix is accomplished, each cell in the resulting probability matrix contains the probability that the column user retweeted the row user for that cell. For the maximum value in each column, that value's row label represents the row-user whom the column-user is most likely to have retweeted.

This matrix is transformed so that the cell with the maximum value for each column is set to one, and all other cells in that column are set to zero. One can conceptualize this matrix as a directed network where a cell value of one indicates that the column user retweeted the row user; the matrix, then, is the inferred information flow of the RTE. These networks are all tree topologies with the root node, or start of the tree, being the user who initiated the RTE.

To measure the extent to which the tree is more or less star-like, *closeness centralization* (Wasserman and Faust 1994) of the network is calculated. Closeness centralization is a network-level measurement based on a node-level centrality measurement called *closeness centrality* or *closeness*. Closeness centrality finds the geodesic distance, or shortest path, between a given node and all other nodes. Closeness is the inverse of the average geodesic distance to all other nodes. The maximum value of 1 occurs for an individual node when it is connected to all other nodes.

At the network level, closeness centralization reaches its maximum value of one in a star graph, when one node is connected to all other nodes. The minimum value of zero only occurs in circular networks, were the geodesic is the same for all actors. Such a configuration is not possible for the inferred hierarchical tree graphs because the root node never receives an in-link. In other words, the directed hierarchical nature of the network prohibits a closeness centralization of zero. However, closeness centralization tends to decrease when the tree includes more and longer chains (becomes less star-like) since the distribution of geodesics becomes more varied (see figure d.2).

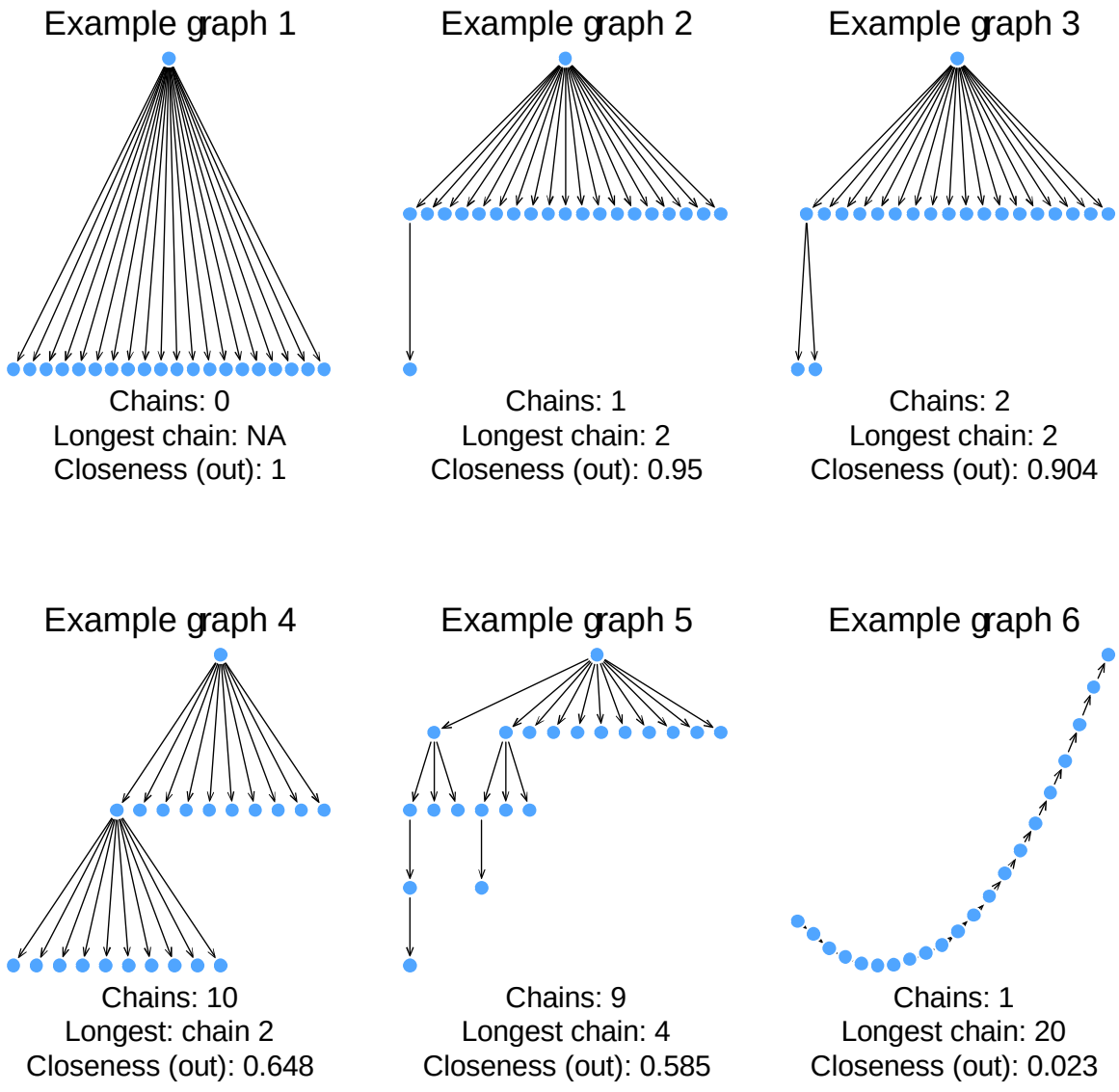


Figure D.2. Example networks with closeness centralization measurements.

How good is closeness centralization at measuring the star-likeness of RTE paths? Figure D.3 shows the scatter plots and correlations of closeness centralization with both the longest path and the number of chains. While the plots indicate curved relationships, they also suggest that for inferred tree networks, closeness centralization is an effective proxy for both the length and number of chains.

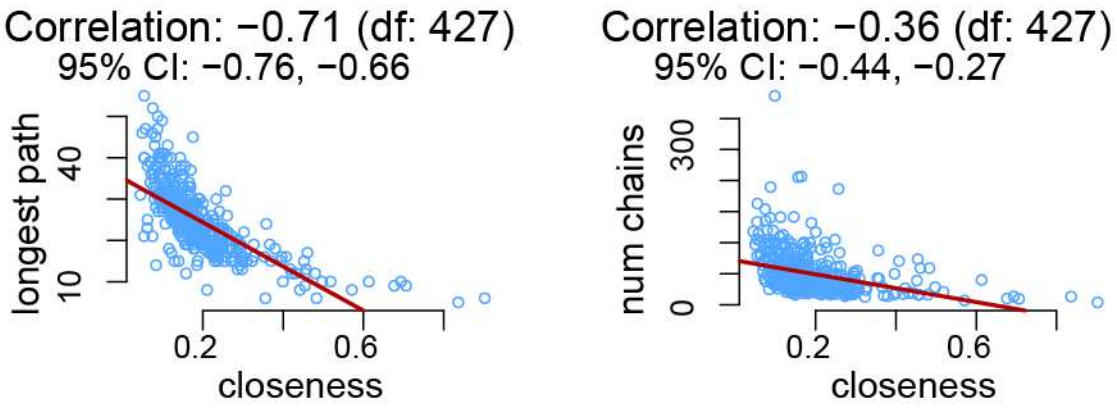
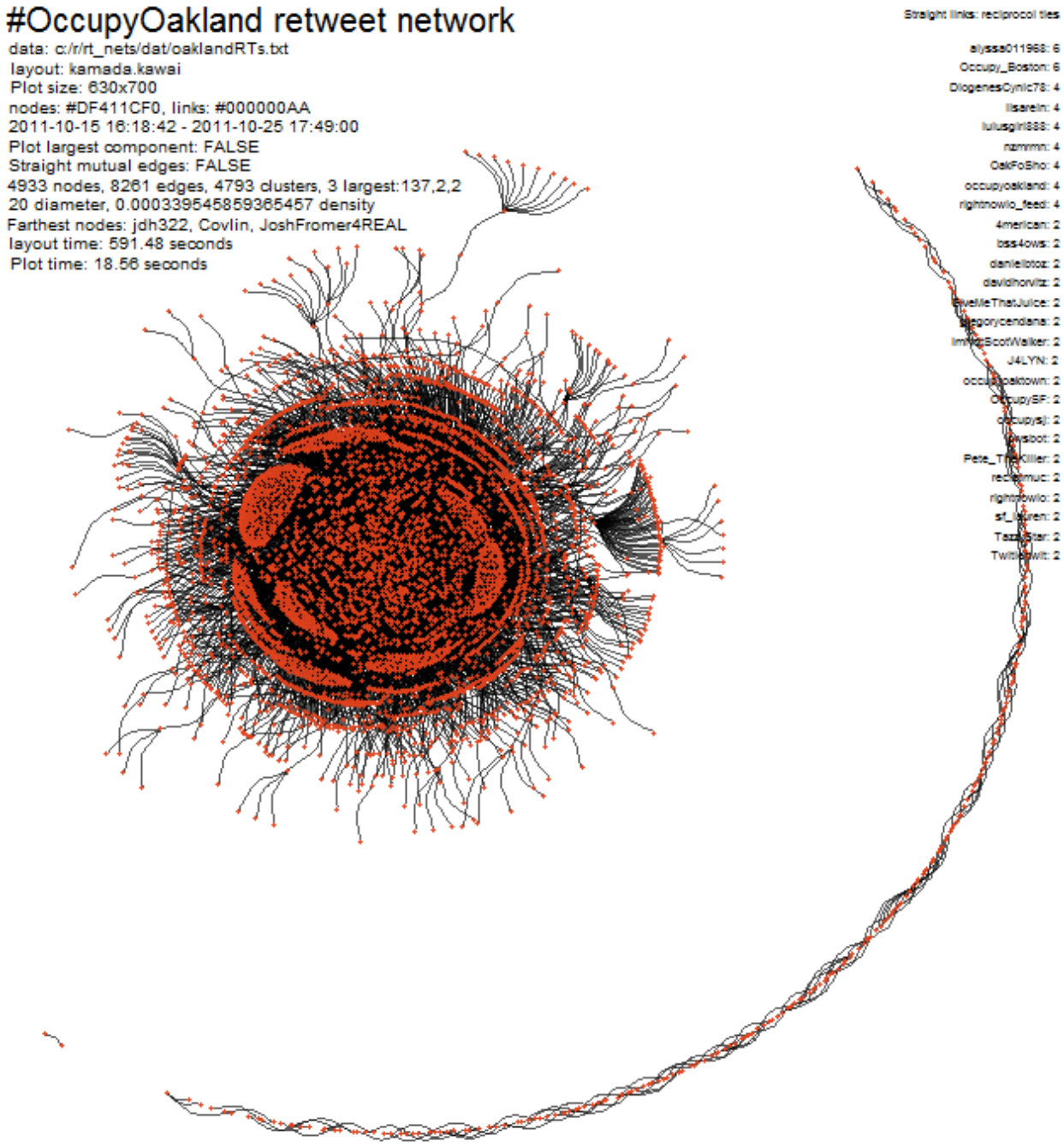


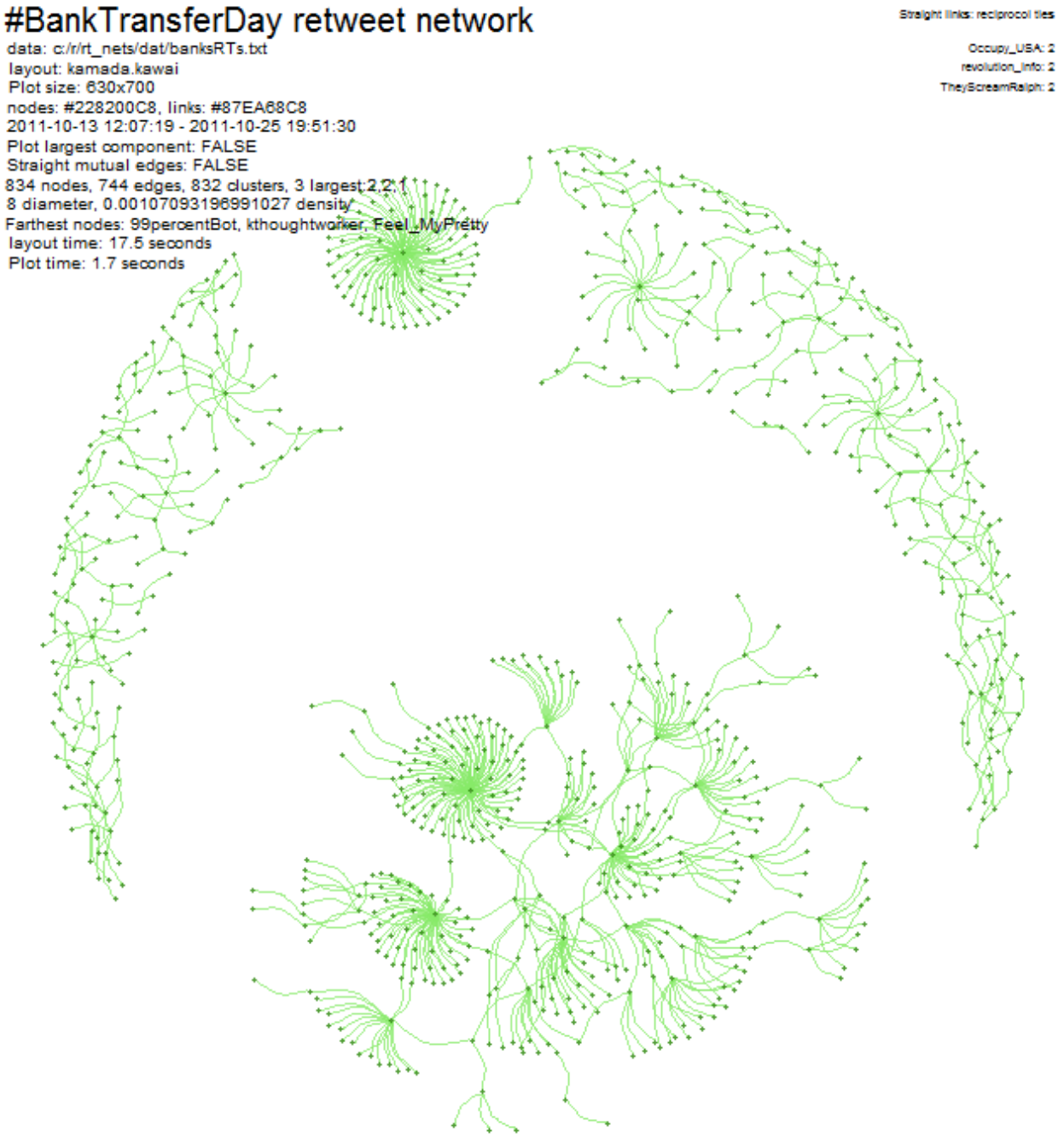
Figure D.3. Relationships between closeness centralization and both path length (left) and number of paths (right).

Teng et al. (2012) use a Gini coefficient for a similar purpose. They measure the changing structure of networks by comparing the Gini coefficient for the undirected degree distribution of a network at different points in time. Since the Gini coefficient of a star network topology is 1, they equate the Gini value as measuring how star-like a network is. Unfortunately, the Gini coefficient used with the inferred RTE networks performs poorly for capturing the length and number of chains. The correlation for the Gini coefficient and the longest paths is -0.22 (%95 CI -0.31, -0.13; df 427), and for the Gini coefficient and the number of chains the correlations is 0.35 (%95 CI 0.27 0.43; df 427). Additionally, the Gini coefficient is related to the number of chains positively and the length negatively, which makes interpretation difficult.

Note that an important limitation of the inferred network model is that it assumes that there no gaps in the flow, that, in effect, everyone in the path network is connected to everyone else. Thus it fails to capture cases where someone searching for a hashtag runs across a tweet and retweets it without being connected to anyone else. In such a case, the user's retweet will be assumed to have been retweeted from a user based on the rules discussed above.

Appendix E: Example Comparison Retweet Network




#BankTransferDay retweet network
data: c:/r/rt_nets/dat/banksRTs.txt
layout: kamada.kawai
Plot size: 630x700
nodes: #228200C8, links: #87EA68C8
2011-10-13 12:07:19 - 2011-10-25 19:51:30
Plot largest component: FALSE
Straight mutual edges: FALSE
834 nodes, 744 edges, 832 clusters, 3 largest 2,2,1
8 diameter, 0.00107093196991027 density
Farthest nodes: 99percentBot, kthoughtworker, Feel_MyPretty
layout time: 17.5 seconds
Plot time: 1.7 seconds
Straight links: reciprocol ties
Occupy_USA: 2
revolution_info: 2
TheyScreamRalph: 2

References

Avram, G. 2006. “At the Crossroads of Knowledge Management and Social Software.” *Electronic Journal of Knowledge Management* 4 (1): 1–10.

Bak, Per. 1999. *How Nature Works: The Science of Self-Organized Criticality*. 1 edition. New York, NY, USA: Copernicus.

Bakshy, E., J. M Hofman, W. A Mason, and D. J Watts. 2011. “Everyone’s an Influencer: Quantifying Influence on Twitter.” In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 65–74. ACM.

Bakshy, E., I. Rosenn, C. Marlow, and L. Adamic. 2012. “The Role of Social Networks in Information Diffusion.” *Arxiv Preprint arXiv:1201.4145*.

Bampo, M., M. T Ewing, D. R Mather, D. Stewart, and M. Wallace. 2008. “The Effects of the Social Structure of Digital Networks on Viral Marketing Performance.” *Information Systems Research* 19 (3): 273–90.

Barabasi, A. L. 2003. *Linked: How Everything Is Connected to Everything Else and What It Means*. Penguin Group New York.

Barabasi, Albert-Laszlo. 2005. “The Origin of Bursts and Heavy Tails in Human Dynamics.” *Nature* 435 (7039): 207–11.

Barabási, Albert-László. 2002. *Linked: The New Science of Networks*. Basic Books.

Barzilai-Nahon, Karine. 2008. “Toward a Theory of Network Gatekeeping: A Framework for Exploring Information Control.” *Journal of the American Society for Information Science and Technology* 59 (9): 1493–1512.

———. 2009. “Gatekeeping: A Critical Review.” *Annual Review of Information Science and Technology* 43: 433–78.

Bass, Frank M. 2004. “A New Product Growth for Model Consumer Durables.” *Management Science* 50 (12): 1825–32. doi:10.2307/30046153.

Benton, T., and I. Craib. 2001. *Philosophy of Social Science: The Philosophical Foundations of Social Thought (Traditions in Social Theory)*. Palgrave Macmillan (18 Jun 2001).

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton University Press.

Borgman, Christine L., and Jonathan Furner. 2002. “Scholarly Communication and Bibliometrics.”

boyd, danah, and K. Crawford. 2011. “Six Provocations for Big Data.”

Boynton, GR. 2009. “Going Viral:The Dynamics of Attention.” In *The Journal of Information Technology and Politics (JITP) Annual Conference*, 11.

Broxton, Tom, Yannet Interian, Jon Vaver, and Mirjam Wattenhofer. 2010. “Catching a Viral Video.” In *Data Mining Workshops, International Conference on*, 296–304. Los Alamitos, CA, USA: IEEE Computer Society. doi:http://doi.ieeecomputersociety.org/10.1109/ICDMW.2010.160.

Burt, R. S. 2004. “Structural Holes and Good Ideas.” *American Journal of Sociology*, 349–99.

Butts, C. 2008. “Social Network Analysis: A Methodological Introduction.” *Asian Journal of Social Psychology* 11 (1): 13.

Butts, C, and R Cross. 2009. “Change and External Events in Computer-Mediated Citation Networks: English Language Weblogs and the 2004 U.S. Electoral Cycle.” *Journal of Social Structure* 10 (3). file:\\C:\Users\Public\Documents\EndNote\20100602_Jeff.Data\PDF\Butts-Cross-3442969393\Butts-Cross.pdf.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. “Power-Law Distributions in Empirical Data.” *SIAM Review* 51 (4): 661–703.

Crane, R., and D. Sornette. 2008. “Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System.” *Proceedings of the National Academy of Sciences* 105 (41): 15649.

Cronin, Blaise. 2001. “Bibliometrics and beyond: Some Thoughts on Web-Based Citation Analysis.” *Journal of Information Science* 27 (1): 1–7.

Driscoll, Kevin, and Shawn Walker. 2014. “Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data.” *International Journal of Communication* 8 (0): 20.

Ellison, N. B, C Lampe, C Steinfield, and J Vitak. 2010. “With a Little Help From My Friends: How Social Network Sites Affect Socail Capital Processes.” In *A Networked Self: Identity, Community, and Culture on Social Network Sites*, edited by Z Papacharissi, 124–45. New York: Routledge.

Faraway, Julian J. 2004. *Linear Models with R*. Vol. 63. Chapman and Hall/CRC.

———. 2005. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. 1st ed. Chapman and Hall/CRC.

García-Murillo, Martha, and Ezgi Nur Gozen. 2012. “Process Theory: Components and Guidelines for Development.” In *Research Methodologies, Innovations and Philosophies in Software Systems Engineering and Information Systems*, 126–48. IGI Global. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-4666-0179-6.ch007.

Gillespie, Tarleton. 2014. “The Relevance of Algorithms.” In *Media Technologies: Essays on Communication, Materiality, and Society*, 167. MIT Press.

Gladwell, Malcolm. 2002. *The Tipping Point: How Little Things Can Make a Big Difference*. Boston: Back Bay Books.

Gomez-Rodriguez, Manuel, Jure Leskovec, and Andreas Krause. 2010. “Inferring Networks of Diffusion and Influence.” In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1019–28. KDD ’10. New York, NY, USA: ACM. doi:10.1145/1835804.1835933.

Granovetter, MS. 1973. “The Strength of Weak Ties.” *American Journal of Sociology* 78 (6): 1360.

Hemsley, Jeff. 2012. “Retweets, Modified Tweets, Vias: What’s in the SoMeLab Dataset | SoMe Lab”. Blog. *SoMe Lab*. http://somelab.net/2012/06/retweets-modified-tweets-vias-whats-in-the-somelab-dataset/.

Hemsley, Jeff, and Josef Eckert. 2014. “Examining the Role of ‘Place’ in Twitter Networks through the Lens of Contentious Politics.” In *Proceedings of the 47th Hawaii International Conference on System Sciences*. Waikoloa, Big Island, HI.

Hemsley, Jeff, and Robert M. Mason. 2013. “Knowledge and Knowledge Management in the Social Media Age.” *Journal of Organizational Computing and Electronic Commerce* 23 (1-2): 138–67. doi:10.1080/10919392.2013.748614.

Hoaglin, David C., Frederick Mosteller, and John Wilder Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. Vol. 3. Wiley New York.

Huang, J., K. M Thornton, and E. N Efthimiadis. 2010. “Conversational Tagging in Twitter.” In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, 173–78. ACM.

Jue, Arthur L., Jackie Alcalde Marr, and Mary Ellen Kassotakis. 2009. *Social Media at Work: How Networking Tools Propel Organizational Performance*. 1st ed. San Francisco, CA: Jossey-Bass.

Kahane, Leo H. 2001. *Regression Basics*. Thousand Oaks, [Calif.]: Sage Publications.

Kitsak, M., L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. 2010. "Identifying Influential Spreaders in Complex Networks." http://arxiv.org/pdf/1001.5285.

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. "What Is Twitter, a Social Network or a News Media?" In *Proceedings of the 19th International Conference on World Wide Web*, 591–600. ACM.

Langley, A. 1999. "Strategies for Theorizing from Process Data." *Academy of Management Review*, 691–710.

Leskovec, J., J. Kleinberg, and C. Faloutsos. 2005. "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations." In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 177–87. ACM.

Leskovec, J., M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. 2007. "Cascading Behavior in Large Blog Graphs." *Arxiv Preprint arXiv:0704.2803*.

Leskovec, J., A. Singh, and J. Kleinberg. 2006. "Patterns of Influence in a Recommendation Network." *Advances in Knowledge Discovery and Data Mining*, 380–89.

Levy, DM. 2001. *Scrolling Forward: Making Sense of Documents in the Digital Age*. Arcade Publishing.

Markham, Annette N., Holly Kruse, Jeff Hemsley, and Molly Steenson. 2013. “Algorithmic Identity: Networks, Data, and the Terrible Beauty of the Black Box.” In *Selected Papers of Internet Research*. Vol. 3. Denver, CO. http://spir.aoir.org/index.php/spir/article/view/891/466.

Marsaglia, George, Wai Wan Tsang, and Jingbo Wang. 2003. “Evaluating Kolmogorov’s Distribution.” *Journal of Statistical Software* 8 (18): 1–4.

“MongoDB.” 2013. Accessed April 23. http://www.mongodb.org/.

Mueller, Charles W., and John W. Tukey. 1980. “Exploratory Data Analysis.” *Administrative Science Quarterly* 25 (4): 700. doi:10.2307/2392291.

Nahon, Karine. 2011. “Fuzziness of Inclusion/exclusion in Networks.” *International Journal of Communication* 5: 756–72.

Nahon, Karine, and Jeff Hemsley. 2013. *Going Viral*. Cambridge, UK: Polity Press Cambridge.

———. 2014. “Homophily in the Guise of Cross-Linking Political Blogs and Content.” *American Behavioral Scientist*, 0002764214527090.

Nahon, Karine, Jeff Hemsley, Robert M. Mason, Shawn Walker, and Josef Eckert. 2013. “Information Flows in Events of Political Unrest.” In *iConference 2013 Proceedings*. Fort Worth, TX.

Nahon, Karine, Jeff Hemsley, Shawn Walker, and Muzammil Hussain. 2011. “Fifteen Minutes of Fame: The Power of Blogs in the Lifecycle of Viral Political Information.” *Policy & Internet* 3 (1): 2.

Ott, R. L., and M. Longnecker. 1993. *An Introduction to Statistical Methods and Data Analysis*. Duxbury press Belmont, CA.

Petrovic, Sasa, Miles Osborne, and Victor Lavrenko. 2011. “RT to Win! Predicting Message Propagation in Twitter.” In *Fifth International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPaper/2754.

Pettigrew, A. M. 1997. “What Is a Processual Analysis.” *Scandinavian Journal of Management* 13 (4): 337–48.

Pettigrew, A. M, R. W Woodman, and K. S Cameron. 2001. “Studying Organizational Change and Development: Challenges for Future Research.” *The Academy of Management Journal* 44 (4): 697–713.

Rogers, E. M. 1995. *Diffusion of Innovations*. Free Pr.

Sabherwal, R., and D. Robey. 1995. “Reconciling Variance and Process Strategies for Studying Information System Development.” *Information Systems Research* 6 (4): 303.

Schutt, Rachel, and Cathy O’Neil. 2013. *Doing Data Science*.

Suh, B., L. Hong, P. Pirolli, and E. H Chi. 2010. “Want to Be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network.” In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 177–84. IEEE.

Susarla, Anjana, Jeong-Ha Oh, and Yong Tan. 2012. “Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube.” *Information Systems Research* 23 (1): 23–41.

Teng, Chun-Yuen, Liuling Gong, Avishay Livne Eecs, Celso Brunetti, and Lada Adamic. 2012. “Coevolution of Network Structure and Content.” In *Proceedings of the 3rd Annual ACM Web Science Conference*, 288–97. ACM.

Tukey, John W. 1962. “The Future of Data Analysis.” *The Annals of Mathematical Statistics*, 1–67.

———. 1977. *Exploratory Data Analysis*. 1 edition. Reading, Mass: Pearson.

———. 1980. “We Need Both Exploratory and Confirmatory.” *The American Statistician* 34 (1): 23–25.

Van de Ven, Andrew. 2007. *Engaged Scholarship: A Guide for Organizational and Social Research*. Oxford University Press, USA.

Walker, Shawn, Jeff Hemsley, Josef Eckert, Robert M. Mason, and Karine Nahon. 2013. “SoMe Tools for Social Media Research.” In *iConference 2013 Proceedings*, 971. Fort Worth, TX. doi:10.9776/13496.

Walther, J. B, C. T Carr, S. S.W Choi, D. C DeAndrea, J. Kim, S. T Tong, and B. Van Der Heide. 2010. “Interaction of Interpersonal, Peer, and Media Influence Sources Online.” *A Networked Self: Identity, Community, and Culture on Social Network Sites*, 17.

Wasserman, S, and K Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge Univ Pr.

Watts, Duncan J. 2004. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company.

Watts, Duncan J., Jonah Peretti, and Michael Frumin. 2007. “Viral Marketing for the Real World.” *Harvard Business Review*, May.