THE DYNAMICS OF INFORMATION DIFFUSION ON ON-LINE SOCIAL NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Shaomei Wu

August 2012

© 2012 Shaomei Wu

ALL RIGHTS RESERVED

THE DYNAMICS OF INFORMATION DIFFUSION ON ON-LINE SOCIAL NETWORKS

Shaomei Wu, Ph.D.

Cornell University 2012

Although there has been a long history of studying the diffusion of information in various social science fields, existing theories are mostly built on direct observations in small networks or survey responses from large samples. As a result, it is hard to verify or refute these theories empirically on a large scale. In recent years, the abundance of digital records of online interactions has provided us for the first time both the explicit network structure and detailed dynamics, supporting global-scale, quantitative study of diffusion in the real world. Using these large scale datasets collected from social media sites, this thesis is to mainly address the following three questions about the process of information diffusion: "who influence whom?", "how do different types of information spread?", and "how does the network structure impact the diffusion process?"

To understand how influence migrates across mediums, we study user's influence on social media based on their role in the global media ecosystem. By categorizing Twitter accounts into elite (i.e. celebrities, media outlets, organizations, and bloggers) and ordinary users, we find a striking concentration of attention on a minority of elite users, and significant homophily within elite categories. On the other hand, following the definition of "opinion leaders" in the classical "two-step flow" theory, we find a large population of opinion leaders who serve as a layer of intermediaries between the elite users and the masses.

In contrast to previous diffusion research on the virality of information, our

work focuses on the persistence of information, in relation to people and content. Fist, we see a systematical difference, in lifespan, for information broadcase by different categories of users. Second, we find a strong association between the content and the temporal dynamics of information: rapidly-fading information contains significantly more words related to negative emotion, actions, and more complicated cognitive processes, whereas persistent information contains more words related to positive emotion, leisure, and lifestyle.

In the end, we study the local and global structure of a decaying online social network. Although there is a significant correlation in both arrival and departure among friends, we show that the dynamics of departure behave differently from the dynamics of arrival. In particular, for the majority of users with a sufficient number (e.g., greater than 20) of friends, departure is best predicted by the overall fraction of activity within a users neighborhood, independent of size. We also find that active users tend to belong to a core that is densifying and is significantly denser than the inactive users, and the inactive set of users exhibit a higher density and lower conductance than the degree distribution alone can explain. These two aspects suggest that nodes at the fringe are more likely to depart, and therefore induce inactive and subsequent departure of neighboring nodes in tightly-knitcom-munities.

BIOGRAPHICAL SKETCH

ACKNOWLEDGEMENTS

Many many many thanks to my committee!

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vi
List of Tables . . . viii
List of Figures . . . ix

1 Introduction **1**
1.1 The predictability of virality . . . 3
1.2 The Influencer Problem . . . 7
1.3 The role of content . . . 10
1.4 The network effect of disengagement . . . 12

2 Exogenous influence and opinion leaders on Twitter **15**
2.1 Data And Methods . . . 17
2.1.1 Twitter Follower Graph . . . 17
2.1.2 Twitter Firehose . . . 18
2.1.3 Twitter Lists . . . 18
2.1.4 Snowball sample of Twitter Lists . . . 19
2.1.5 Activity Sample of Twitter Lists . . . 22
2.2 Distribution of attention . . . 23
2.2.1 Concentration of attention . . . 23
2.2.2 Homophily of influence . . . 27
2.3 Revisiting two-step flow theory: what are the opinion leaders? . . . 30
2.4 The interaction between people and content . . . 36
2.5 Lifespan of content by category . . . 39
2.6 Conclusion . . . 45

3 The role of content **47**
3.0.1 Data . . . 48
3.0.2 Predicting temporal patterns based on content . . . 50
3.0.3 How temporal patterns vary with types of content . . . 55
3.0.4 The quality and persistence of YouTube videos . . . 63
3.1 Conclusion . . . 66

4 Network structure and the spread of disengagement **67**
4.1 Data . . . 70
4.2 Arrival and departure correlation among friends . . . 71
4.3 Effect of local neighborhoods . . . 75
4.3.1 Dependence on local properties . . . 75
4.3.2 Interaction between local properties . . . 78
4.3.3 Predict the departure of user . . . 79

4.4 Structural trends in network topology 82
4.5 Conclusion 86

5 Conclusion **93**

A Chapter 1 of appendix **94**

Bibliography **95**

LIST OF TABLES

2.1 Snowball Sample 22
2.2 Statistics of crawled lists. The number of users refers only to people who appear in at least one list of the specific category. . . 23
2.3 Top 5 users in each category 26
2.4 # of URLs initiated by category 26
2.5 Information flow among the elite categories 28
2.6 RTs among categories 30
2.7 Re-introductions among categories 31

3.1 Feature size 53
3.2 Results for predicting lastingness of information 54
3.3 LDA Topics 60
3.4 Representative words for two temporal classes 61

4.1 Predict user departure with decision tree 80
4.2 Summary of logistic regression model on $p_{departure}$ 81

LIST OF FIGURES

2.1 Schematic of the Snowball Sampling Method 21
2.2 Average fraction of # following (blue line) and # tweets (red line) for a random user that are accounted for by the top K elites users crawled 25
2.3 Average fraction of # following (blue line) and # tweets (red line) for a random user that are accounted for by the top K most retweeted users in the "Other" category 27
2.4 Share of attention among elite categories 29
2.5 RT behavior among elite categories 31
2.6 Percentage of information that received via an intermediary as a function of total volume of media content to which a user is exposed 34
2.7 Frequency of intermediaries binned by # randomly sampled users to whom they transmit media content 35
2.8 Number of RT's and Reintroductions of New York Times stories by content category 38
2.9 Number of RT's and Reintroductions of most popular URLs originating from media and other 40
2.10 (a) Definition of URL lifespan τ (b) Schematic of lifespan estimation procedure 41
2.11 (a) Count and (b) percentage of URLs initiated by 4 categories, with different lifespans 42
2.12 Top 20 domains for URLs that lived more than 200 days 43
2.13 Average RT rate by lifespan for each of the originating categories 45

3.1 Distribution of URL decay time t_u 50
3.2 URL overall popularity as a funtion of t_u 51
3.3 Normalized time series centroids for two classes 52
3.4 Class distribution in 60 LIWC dimensions, using words from HTML header 56
3.5 Trending LIWC categories 57
3.6 Class distribution in 50 LDA topics, using words from HTML header and body 59
3.7 Distribution of t_u for YouTube videos 64
3.8 The quality of videos as a function of decay time t_u. *Like/dislike rate* is the number of likes divided by the number of dislikes. *Bad videos* are those with the number of dislikes greater than half of the number of likes. There are in total 83 out of 2304 "bad" videos by our definition. 64
3.9 Average number of views and comments as a funtion of t_u 65

4.1 The CDF curve for the difference in arrival and departure time between friends and random pairs of users. 72

4.2 Gap between CDF curves. 73
4.3 Count and Fraction of friends already signed-up when user signs up. . 74
4.4 Probability of departure as function of different local properties. Where (a) is p as f(active friend count) in SN, (b) is p as f(inactive friend count) in SN, (c) is p as f(active friend count) in Dblp, (d) is p as f(inactive friend count) in Dblp, (e) is p as f(active friend fraction) in SN and (f) is p as f(inactive friends who left in the past 6 months) in SN. 88
4.5 Probability of departure as function of local properties, at different levels of active/inactive friend fraction and friend count (snapshot taken at time t =2011/1/1). in SN 89
4.6 Distribution of edges, indicated by the ratio of actual number of edges over the expected number of edges (formed in random process). 90
4.7 Density of the active and inactive subnetworks 91
4.8 Conductance of the active and inactive sets 92

CHAPTER 1

INTRODUCTION

Information diffusion has been a long intriguing topic for scholars from different fields. It is an essential element of many interesting problems, such as the diffusion of innovation[34], formation of public opinion[19, 41], adoption of new products[5]. Historically, most of the research in this area has been done through field observations and/or phone surveys[18, 34]. Confined by the methodology, their results are uaually based on the spread of a specific opinion or product in rather localized networks.

The development of the Internet and the emergence of a whole new set of communication technologies on top of it has largely changed the way people gather and disseminate information, and brought both challenge and opportunity to a better understanding of how information spreads in the society. Historically, most empirical studies in this area were based on data collected through in-person interviews and phone surveys[34, 19], targeted with a specific opinion or product that is under diffusion. Confined by the methodology, theories and conclusions drew from these research suffer at scale and data accuracy, and are challenged in today's massive, hybrid communication environment. Meanwhile, the popularity of eletronic communication tools has offered new possibilities for obtaining much better data on how information and behavior spread over large population. The abundance of digital records of online interactions can provide large amounts of data on both explicit network structure and detailed dynamics, supporting large-scale, quantitative study of diffusion in the real world.

Using these large scale datasets collected from social media sites (i.e. Twitter,

Orkut), we study three major components of information diffusion process:

- *People*, who are the "influentials" and whom are they influencing?
- *Content*, does different types of information spread differently?
- *Network*, what can the macro- and micro-level structure of the network tell us about the underlying mechanism that drives the diffusion process?

Although still a young field, there has been a growing literature that studied these components, mostly focused on modeling and predicting the scale or the spikiness of cascade. Building on top of existing knowledge, our exploration on these three components highlights the following three contributions:

First, we study the "virality" of diffusion with new aspects. Instead of scale, we focus on the temporal patterns of cascades, in particular, the persistence of information. We notice that certain information has a very long lifespan whereas most of the information only exists shortly in social media. We also look at the "negative" virality - the virality of "inactivity", and show how user disengagement, similar to user engagement, has a network effect but operates slightly differently.

Second, we examine the exogenous properties of actors and artifacts beyond the closed, single-medium communication network. For example, by looking in-depth at the textual characteristics, we find that the persistent attention to certain content can be better explained by the innate quality of the content, independent from the mechanism of how it spreads in the network. We also find that, despite the connectivity and demographics, certain users are more influential on social media because of their established role (and fame) from the world outside social media. These results suggest that, diffusion process should be

studied in context, considering the socio-cultural factors that constantly interact with the information cascade.

Third, we study different components of diffusion not only by themselves, but also the interaction between them. As information diffusion is an interactive process between people and content, mediated by the communication networks, different components of diffusoin do not exist in isolation. People with high influence tend to cluster together[1], and the location in the network can tell us a lot about a person's status[21, 2]. In our work, we show that different people have different preference of content and invididual's influence is topic-dependent.

This thesis is an inter-disciplinary effort. Although most of the results are quantitative and are drew from analysis on large online datasets, I have worked closely with experts from various fields including sociologists, computer scientists, phycologists, communication scholars, and political scientists. During this process, I try to conduct my research with not only computational methods but also theories and insights from different fields. As a result, I hope our findings are not limited to online social networks, but can also help uncover the nature of information diffusion process acorss online and offline spaces.

1.1 The predictability of virality

With viral videos and memes outbreak and circulating in the Internet almost everyday, brands and politicians are eager to leverage the power of social networks and run online campaigns that "go viral". The demand from the market, contrasted with our lack of knowledge on the mechanism behind viral phe-

nomenons, has catalyzed many studies focusing on modeling and predicting the virality of diffusion process. In this line of work, virality is usually defined and measured in three aspects:

1. *Scale*: the size of cascade tree originated from the original node. e.g., the cascade Facebook fanning pages;
2. *Range*: the maximal chain length in the diffusion network. e.g., blog[28].
3. *Transmission probability*. As most existing research describe diffusion process with epidemic models, transmission probability is the probability a virus travels through a directional edge in the network and infects the end node. In the context of social networks, transmission probability is mapped to the success rate of interpersonal recommendations in electronic market[24], the probability a tweet is retweeted by the followers of the tweet's author[4], and the likelihood a user mark a Flickr photo as favorite after seeing a friend doing so[10].

Using large scale online social networks data, researchers have employed a whole set of computational tools to asset different aspects of cascades that may lead themselves to viral. However, in practice, it is still very difficult to engineer a success viral campaign[4]. In fact, most marketers and public relation professionals still rely heavily on experience and intuition when launching online campaigns, which mostly received attention in a short timespan[46, 11, 44, 43]. Although the virality is still the holy grail for marketers, there has been an ascending suspicion on the whole idea of "predicting virality".

In their Science article published in 2006, Salganik et al[36] explored the unpredictability of song ratings, argued that, despite the quality of music, the best

predictor for a song's popularity is actually the popularity itself: the songs that received first few thumbs-ups usually rise up to be the hit, even if the first thumbs-ups are given randomly. Some later work showed that the quality of initiators does not predict well with the final scale of cascades[38, 4]. Although some studies showed that the structure of the subgraph induced by existing adopters do have a role in behaviro adoption[2, 35, 28], such role is not consistent in other studies[4], and may depend on the context of diffusion[8].

There are two major factors that might contribute to the difficulty of predicting virality. First, as most studies observed, the power-law (or log-normal) distribution of the popularity of online content results in a skewed training set consisting of many unsuccessful examples and only a few successful examples. While most machine learning algorithms are not robust to skewed training sets[33], the common practice of biased sampling can over-compensate the successful cases and overlook the unsuccessful ones. As a result, most characteristics we discovered in the positive class can also be found in the negative class - the predictive model would fail to tell one from the other. Second, as we mentioned before, given the fact that online communications is an integrated part of the global communication networks, there are many exogenous factors that can not be controlled or observed in the communication channel where the data is collected. The complexity and the fuzziness of real world diffusion process is innately inadequate to be fully captured by epidemic models[4, 9].

Understanding the difficulty of this task does not mean that it is unsolvable. We have seen tremendous progress towards a better understanding of the virality, incorporting theories and methods from multiple fields (e.g., [8, 4]). As part of these efforts, we also study the *transmission probability* of URLs among

different categories of users on Twitter. Our results show that influential users are much more likely to pass on information from those who are similar to them (for more details, see Chapter 2).

Meanwhile, during our study of the spread of URLs on Twitter, we discovered and explored some new metrics for the success of cascade. We first noticed that, although skewed as well, the distribution of the lifespan of information has a much heavier tail towards the end. In other words, there is a substantial amount of information that is consistently tweeted over a long period of time, in the fast-paced medium such as Twitter. When examining the persistent items in depth, we realized that, most of them can not be characterized by any of the traditional metrics for virality: they do not receive many retweets, they do not spread through long-chains, and they are usually not part of super-sized cascades. However, as they generate comparably large interest that lasts for a signficantly long period of time, we believe that persistent content are interesting enough to be studied as "success" examples of diffusion as well.

The study of information longevity is generally missed in the study of information diffusion, even among the string of research on the temporal dynamics of cascades[47, 11, 27]. Although some previous work also spotted the lastiningness of certain content[11, 39], researchers usually concentrate on modeling the speed, intensity, and scale of the peak of attention, while treating the long-last content as corner cases. In our work, we compare the lifespan of information by its initiator and content, and ask what makes certain things so persistently eye-catching in a world where information is overly rich and attention is remarkably scarce[37]. Our major findings include:

- The longevity of information is determined to a large extent by the exoge-

nous qualities of the information, not by social contagion.

- Content picked up by certain group of users (e.g., bloggers) are in general of more persistent interest, as these users actively perform the role of "information filter" on online social media.
- There is a strong correlation between the content and the temporal dynamics of information. Content with cultural/intelligent value are more likely to persist. At the text level, rapidly-fading information cntains signficantly more words related to negative emotion, actions, and more complicated cognitive processes, whereas the persistent information contains more words related to positive emotion, leisure, and lifestyle.
- The decay of attention, although much less visible, also has a network effect that coordinates the disengagement in online social networks among friends.

1.2 The Influencer Problem

The role of the influentials as trend makers has been a center piece of many classic theories about information diffusion process[19, 34]. The existence, and the importance, of the influentials also populated by several best-selling books such as The Tipping Point[15] and The Influentials[20]. Today, challenged by the unbounded opportunity (and efforts) to reach the masses, marketers and PR firms are especially eager to leverage the power of the influentials to "tip" their products. But, who are the influentials?

There are several classic theories about the characteristics of the influentials. Dating back to the 50s, Katz and Lazarsfeld coined the term "opinion lead-

ers". They claimed that, comparing to ordinary people, opinion leaders have more social connections and are more media-savvy[19]. Later, scholars studying the diffusion of innovations also suggested that opinion leaders usually own "greater exposure to mass media than their followers", "are more cosmopolite", "have greater social participation" , "have higher socioeconomic status", and "are more innovative"[34]. Similar ideas are illustrated in best-selling books, in which authors claims that the influentials are "connectors", "mavens", and "salesmen"[14], and that influentials play their role actively and constantly, providing suggestions from what to buy to who to vote for[20].

One critic to the classic theories is their lack of empirical supports. Although intuitively sound, these theories about influentials are too general to be operationalized or examined in practice. It was only since the abundance of online interaction data that we saw a new line of empirical work on measuring and quantifying personal influence in diffusion. Most of these work studied influence in two aspects: personal attributes such as demographics and activities, network attributes such as connectivity and position in the network. Although both aspects are usually considered and studied, most work showed the network attributes more relevant to personal influence[38, 4, 21, 23, 24, 10].

However, we find current knowledge on influence and influentials still limited for several reasons. First, as most work focused on big cascades, the results can be biased towards the "successful" events that are deemed to be rare[4].

In addition, the *context* of influence is usually overlooked. Influence deos not exist in vacumne. It needs to be exercized by people, with certain communication medium. The context of influence is important, because individuals influence can differ by their expertise[9], the subject[24], and the communica-

tion channels[43]. Existing empirical studies measure influence in a variety of ways, from the size of diffusion tree, to the probability of passing the cascade to the next hop. As a result, current definition of influence is not only inconsistent but also ambiguous: different types of influence are usually studied as a whole, regardless differnet mechanisms that operate behind them. As a result, different types of influence are mixed and studied as if driven by a single mechanism. And the exogenous influence migrated from other channels is usually mixed and mistaken as interpersonal influence in social media system studied.

In Chapter 2, we approach the influencer problem by leveraging people's external influence to online social media. Here, we consider online social media as a medium that carries a whole spectum of communications, from personal and private interactions to mass media broadcasting. We thus categorize users based on their role in the global communicaiton system. To eliminate the bias towards successful cascades, we study the influence of each category in terms of visibility and the ability to stimulate and sustain attention from other users. Our work makes three main contributions:

- We introduce a crowd-sourcing method for classifying users into "elite" and "ordinary" users according to their role in the media ecosystem, further classifying elite users into one of four categories of interest łmedia, celebrities, organizations, and bloggers.
- We investigate the distribution of attention among these categories, finding that although audience attention is highly concentrated on a minority of elite users, much of the information they produce reaches the masses indirectly via a large population of intermediaries - local opinion leaders.
- We find that different categories of users emphasize different types of

content, and that content originated by different users exhibit dramatically different characteristic persistency, ranging from less than a day to months.

1.3 The role of content

Although it has been a common belief that different types of information spread differently, the role of content in diffusion process has not been examined thoroughly and systematically. Most empirical work in this area focused on the relationship between information virality and content[6, 17], while some explored the connection between the temporal patterns and the content[16, 11]. However, these studies are limited either by the lack of observations, or – when viewed as content-base predictions – by the relatively weak predictive power.

There are several major challenges here. First, the content itself is difficult to track, especially when it travles and mutates across multiple media channels over a long period time[26]. Second, predicting the virality of information at an individual level is a very hard problem by itself[4, 39]. Third, given the focus of past work on how people interact with information, modeling such dynamics becomes extremely complicated with many unpredictable elements involved[45, 47, 43, 4].

In Chapter 3, we study the relationship between content and temporal dynamics of diffusion on Twitter.

We tackle the information tracking problem by taking advantage of the URL shorteners (e.g. bit.ly, TinyURL, etc) commonly adopted by Twitter users. Con-

sidering each webpage as a unit of information, the URL shortening services tag a page with a unique token that is easily traceable in Twitter communications. As a result, we able able to track the whole lifespan of a specific webpage by the inclusion of the shortened URL in tweets.

In response to the difficulty of predicting the virality (e.g., scale, tranmission probability), we shift our attention to the persistency of information. We compare two extreme temporal patterns in the decay rate of URLs embedded in tweets, defining a predicting task to distinguish between URLs that fade rapidly following their peak of popularity and those that fade more slowly. Our experiments show a strong connection between content and the temporal dynamics of information.

Also, by studying the webpages instead of individual tweets, we have a much richer, and more static corpus of content that allows us to simplify our model by focusing on the textual content instead of the users, while still maintaining sufficient degree of freedom to generate Another advantage of our approach is that if we can predict the temporal pattern of information based on content alone, we will be able to do that at a very stage, presumably when the information is first generated - which can be of interest to practioners.

In summary, we study intrinsic qualities of the content that may effectly determine the dissemination process, especially, the persistence of information. Our two main contributions include:

- We build a classifier that predicts the decay/persistence of information with textual features, providing one of the first empirical studies of the connection between content and temporal variations of information diffu-

sion processes in online social media.

- We investigate the properties of the text that are associated with different temporal patterns, finding significant differences in word usage and sentiment between rapidly-fading and long-lasting information.

1.4 The network effect of disengagement

The structure of network is another interesting component of diffusion research.

At a local level, many studies have shown the correlation of activity among friends in online communities, and tried to understand the effect of neighborhood structure on the spread of information, or behavior. Classic theory on product adoption suggests an "S-shape" curve[5] in which the probability of adopting a new product grows slowly with a small number of adopted friends k, rapidly as k increases, and quickly saturates once k reaches a certain point. But recent empirical work based on online data showed that the adoption probability follows a "deminishing return" curve in which it first grows rapidly at small values k, then gradually stops to grow as k gets large[2, 35]. When studying the local structure in depth, people found that, the structural closeness of local community (measured by triadic closures or cluster coefficient) has a siginificatn effect at the probability of adoption[2], however, such effect may differ by the types of content under diffusion[35].

In addition to the research efforts on local structure, a number of studies looked at global structure of networks, in the understanding of macro-level dynamics of diffusion. Fixing the diffusion model for local dynamics (e.g., epidemic models, independent cascade models, threshold models), the global

structure of network can determine (and explain) certain characteristics of cascades that run on top of it. For example, in [31], the authors argues that the persistence of low-contagious virus in computer network systems is a result of the scale-free structure of the network. In many other work, the global structure of the diffusion network is used as benchmark for generative models that simulate the mechnism of diffusion[13, 16, 29].

While most existing work focuses on the growth of networks and the increase of activity, our work differs by shifting efforts to the dynamics of user departure from social networks, and the decline of activity. What leads people to depart from their social networks? Is inactivity also contagious? One could argue that since inactivity is by definition less visible than activity, it should have less effect on influencing an individual's behavior. However, the extreme case in which all friends of a user depart suggests that, eventually, there must be an effect.

In Chapter 4, we study these questions in the context of the Dblp co-autorship network and a large online social network. We show that the network effect of departure operates differently from the network effect of formation. In particular, the departure of a user with few friends, say less than 20, may be understood most accurately as a function of the raw number of friends who are active. For the majority of users with larger numbers of friends, however, departure is best predicted by the overall fraction of activity within a user's neighborhood, independent of size. We then study global properties of the subgraphs induced by active and inactive users, and show that active users tend to belong to a core that is densifying and is significantly denser than the inactive users. Further, the inactive set of users exhibit a higher density and lower conduc-

tance than the degree distribution alone can explain. These two aspects suggest that nodes at the fringe are more likely to depart, and therefore induce inactive and subsequent departure of neighboring nodes in tightly-knit communities.

Although our study starts from the inactivity of users in onlie communities, the results are verified in other social networks such as DBLP, thus should be also generalizable to the decay of attention in information diffusion process.

CHAPTER 2

EXOGENOUS INFLUENCE AND OPINION LEADERS ON TWITTER

Several work studied user influence on Twitter with different topological and activity measures[23, 9, 42]. Most of them found these measures inadequate at describing the notion of influence on Twitter, only capturing part of the dynamics observed. In addition, people noticed a significant presence of "mass-media" communications on Twitter[23], and top users are influential not only by "information value" but by the "name value" they hold outside of Twitter[9]. Our work explored the mass-media aspect of Twitter, and studied the influence of different categories of "elite" users, from "mass-media", to "mass-personal".

There is a distinct difference in the mechanisms that drive "mass" influence and "interpersonal" influence. According to classic communication theories, mass influence is exerted through "oneway message transmissions from one source to a large, relativelyundifferentiated and anonymous audience", whereas, "interpersonal" influence plays out through "two-way message exchange between two or more individuals."

Recent changes in technology, however, have increasingly undermined the validity of the mass vs. interpersonal distinction. On the one hand, over the past few decades mass communication has experienced a proliferation of new channels, including cable television, satellite radio, specialist book and magazine publishers, and of course an array of web-based media such as sponsored blogs, online communities, and social news sites. Correspondingly, the traditional mass audience once associated with, say, network television has fragmented into many smaller audiences, each of which increasingly selects the information to which it is exposed, and in some cases generates the information

itself. Meanwhile, interpersonal influence has become amplified through personal blogs, email lists, and social networking sites to afford individuals ever larger audiences. Together, these two trends have greatly obscured the historical distinction between mass and interpersonal influence, leading some scholars to refer instead to "masspersonal" influence[40].

Twitter has showcased this erosion of traditional categories of influence. To illustrate, the top ten most followed users on Twitter are not corporations or media organizations, but individual people, mostly celebrities. Moreover, these individuals communicate directly with their followers, often managing their accounts themselves, thus bypassing the traditional intermediation of the mass media between celebrities and fans. In addition to conventional celebrities, a new class of "semi-public" individuals like bloggers, authors, journalists, and subject matter experts, have come to occupy an interesting role on Twitter, in some cases becoming more prominent than traditional public figures such as celebrities and elected officials. Finally, in spite of these shifts towards masspersonal communication on Twitter, media organizations, along with corporations, governments, and NGO's all remain represented among highly followed users, and are often extremely active.

Twitter, therefore, provides an interesting context in which to study the effect of exogenous factors on social media, especially as Twitter—unlike television, radio, and print media—enables one to easily observe information flows among the members of its ecosystem. However, the effects (e.g. changes in behavior, attitudes, etc) remain difficult to measure on Twitter, and so our study of influence is limited to the distribution of attention and the internsity of interaction within Twitter platform.

To this end, our paper makes three main contributions:

- We introduce a method for classifying users with their external status, using Twitter Lists, into "elite" and "ordinary" users. We further classify elite users into one of four categories of interest—media, celebrities, organizations, and bloggers.
- We investigate the potential influence of these categories, based on how their opinion reach and stimulate the public. We find that although audience attention is highly concentrated on a minority of elite users, much of the information they produce reaches the masses indirectly via a large population of intermediaries.
- We show that different categories of users place slightly different emphasis on different types of content, and that different content types exhibit dramatically different characteristic lifespans, ranging from less than a day to months.

2.1 Data And Methods

2.1.1 Twitter Follower Graph

In order to understand how information is flowing in the Twitter system, we need to know the channels by which it flows; that is, who is following whom on Twitter. To this end, we used the data shared[1] by Kwak et al. [23], which

[1] At the time of this study, the data was free to download from http://an.kaist.ac.kr/traces/WWW2010.html

included 42M users and 1.5B edges. This data represents a crawl of the follower graph seeded with all users on Twitter as observed by July 31st, 2009.

2.1.2 Twitter Firehose

In addition, we were interested in the content that was being shared—particularly bit.ly URLs—so that we could trave the flow of information through the Twitter graph. We examined all tweets over a 223 day period from July 28, 2009 to March 8, 2010 using the data from the Twitter "firehose". From these 5B tweets we observed 260M bit.ly URLs.

2.1.3 Twitter Lists

Our method for classifying users exploits a relatively recent feature of Twitter: Twitter Lists. Since its launch on November 2, 2009, Twitter Lists have been welcomed by the community as a way to group people and organize one's incoming stream of tweets by specific sets of users. To create a Twitter List, a user needs to provide a name (required) and description (optional) for the list, and decide whether the new list is public (anyone can view and subscribe to this list) or private (only the list creator can view or subscribe to this list). Once a list is created, the user can add/edit/delete people in the list. As the purpose of Twitter Lists is to help users organize people they follow, the name of the list can be considered a meaningful label for the listed users. List creation therefore effectively applies the "wisdom of crowds" to the task of classifying users, both in terms of their importance to the community (number of lists on which they

appear), and also how they are perceived (e.g. news organization vs. celebrity, etc.).

There is not yet a standard way to classify users by lists, or even a central portal to obtain lists for all users. In order to capture the variety of users invloved in mass media, masspersonal, and interpersonal commmunication described previously in a reasonably parsimonious manner, we restrict our attention to four classes of what we call "elite" users: media, celebrities, organizations (including both public and private), and bloggers. In addition to these elite users, we also study the much larger population of "ordinary" users, as well as the relationships between elite and ordinary users. [2].

Given the rate limits established by Twitter's API, moreover, crawling all lists for all Twitter users (reportedly over 100M, where some users are included on tens of thousands of lists) would be prohibitively time consuming. Thus we instead devised two different sampling schemes—a snowball sample and an activity sample—each with some advantages and disadvantages, discussed below.

2.1.4 Snowball sample of Twitter Lists

The first method for identifying elite users employed snowball sampling. For each category, we chose a number of seed users that were highly representative of the desired category and appeared on many category-related lists. For each of the four categories above, the following seeds were chosen:

[2]Some third-party sites such as Listorious (http://listorious.com/) now maintain categorized directories of Twitter Lists; however, their methodology is not sufficiently transparent for our purposes. We also found their data largely not-up-to-date.

- Celebrities: Barack Obama, Lady Gaga, Paris Hilton
- Media: CNN, New York Times
- Organizations: Amnesty International, World Wildlife Foundation, Yahoo! Inc., Whole Foods
- Blogs[3]: BoingBoing, FamousBloggers, problogger, mashable. Chrisbrogan, virtuosoblogger, Gizmodo, Ileane, dragonblogger, bbrian017, hishaman, copyblogger, engadget, danielscocco, BlazingMinds, bloggersblog, TycoonBlogger, shoemoney, wchingya, extremejohn, GrowMap, kikolani, smartbloggerz, Element321, brandonacox, remarkablogger, jsinkeywest, seosmarty, NotAProBlog, kbloemendaal, JimiJones, ditesco

After reviewing the lists associated with these seeds, the following keywords were hand-selected as representative of the desired categories:

- Celebrities: star, stars, hollywood, celebs, celebrity, celebrities-on-twitter, celebrity-tweets, celebrity-list, celebrities, celebsverified
- Media: news, media, news-media
- Organizations: company, companies, organization, organisation, organizations, organisations, corporation, brands, products, charity, charities, causes, cause, ngo
- Blogs: blog, blogs, blogger, bloggers

Having selected the seeds and the keywords for each category, we then did a snowball sample of the bipartite graph of users and lists (see Figure 2.1). For

[3]The blogger category required many more seeds because bloggers are in general lower profile than the seeds for the other categories

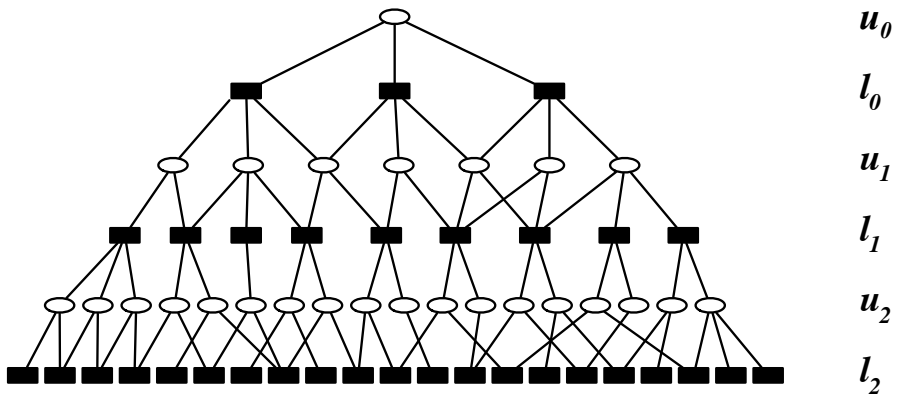


Figure 2.1: Schematic of the Snowball Sampling Method

each seed, we crawled all lists on which that seed appeared. The resulting "list of lists" was then pruned to contain only lists whose names matched at least one of the chosen keywords for that category. We then crawled all users appearing in the pruned "list of lists". We then repeated these last two steps.

Table 2.1 shows how many (a) users and (b) lists were obtained at each level of the snowball sample. In total, $495,000$ users were obtained, who appeared on $7,000,000$ lists. Because users can be listed in multiple categories (e.g., Oprah Winfrey is frequently included in lists of "celebrity" and "media"), we next compute a user u's membership score in category c:

$$w_{uc} = \frac{n_{uc}}{N_c}, \tag{2.1}$$

where n_{uc} is the number of lists in category c that contain user u and N_c is the total number of lists in category c. We then assign each user to the category in which he or she has the highest membership score. Users that appear in the follower graph but not in the snowball sample are assigned to the "ordinary" category.

Table 2.1: Snowball Sample

Level	celeb	media	org	blog
u_0	3	2	4	32
l_0	2342	11403	1170	1347
u_1	3607	5025	20122	16317
l_1	30490	71605	4970	9546
u_2	108836	309056	115034	140251
l_2	91873	171912	22518	19946

2.1.5 Activity Sample of Twitter Lists

Although the snowball sampling method is convenient and is easily interpretable with respect to our theoretical motivation, it is also potentially biased by our particular choice of seeds. To address this concern, we also generate a sample of users based on their activity. Specifically, we crawl all lists associated with all users who tweet at least once every week for the entire observation period.

This "activity-based" sampling method, which yields $750,000$ users and $5,000,000$ lists (see Table 2.2 for comparison to the snowball method), is also clearly biased towards users who are consistently active. Importantly, however, the bias is likely to be quite different from any introduced by the snowball sample; thus obtaining similar results from the two samples should give us confidence that our findings are not artifacts of the sampling procedure.

	Snowball Sample		Activity Sample	
category	# of users	# of lists	# of users	# of lists
celeb	108,836	91,873	22,803	68,810
media	309,056	171,912	66,300	145,176
org	115,034	22,518	19,726	16,532
blog	140,251	19,946	49,987	17,259

Table 2.2: Statistics of crawled lists. The number of users refers only to people who appear in at least one list of the specific category.

2.2 Distribution of attention

After categorizing people into categories, we can calculate the amount of attention sent and received by each category, at a global level. The approach we take is to measure the reach of the "elite" categories, which can be considered as the influence of each category, as well as an estimate of the impact of the information introduced by each category. In other words, it is the maximal reach of the information produced by each category.

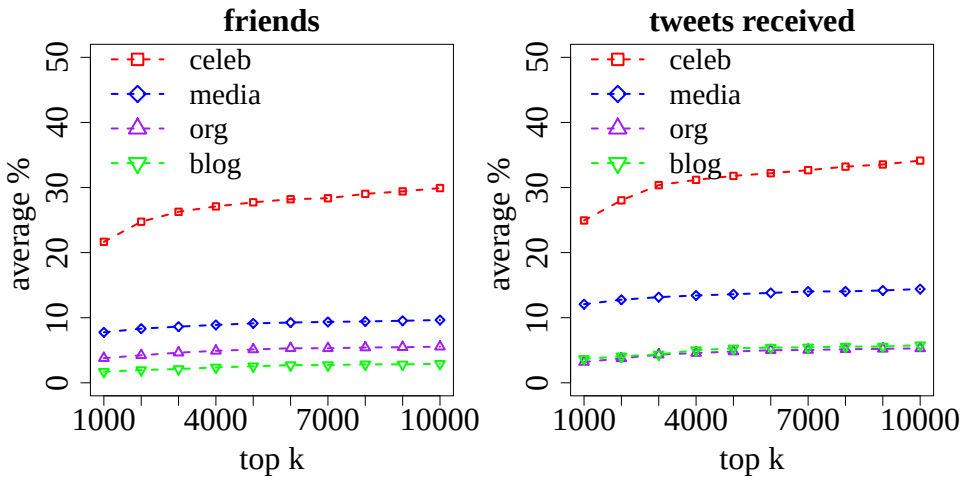
2.2.1 Concentration of attention

With either sampling method, the initial categorization of users is quite coarse and noisy as a result of the arbitrary labeling allowed in Twitter Lists. To filter categories to the most representative users, we further rank the users in each of the 4 elite categories by how frequently they are listed in each category, and take only the top k users in each category, relabeling the remainder as "ordinary" users. To determine the appropriate k, we measure the flow of information from the four elite categories to an average "ordinary" user in two ways: the proportion of people the user follows in each category, and the proportion

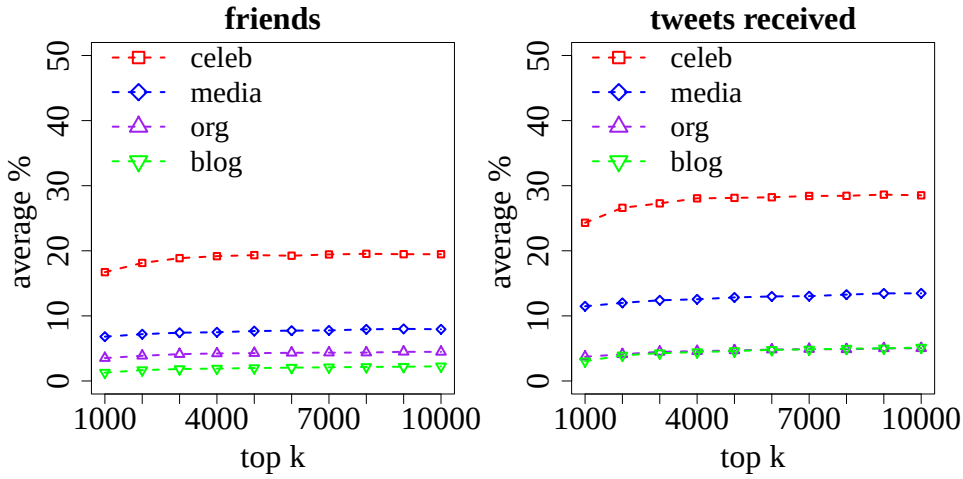
of tweets the user received from everyone the user follows in each category. We sampled 100K random "ordinary" users and calculated the average information flow from the "elite" users using these two measures.

Figure 2.2(a) shows that each category accounts for a significant share of both the following links and also the tweets received by an average user, where celebrities outrank all other categories, followed by the media, organizations, and bloggers. Also of note is that the bulk of the attention is accounted for by a relatively small number of users within each category, as evidenced by the relatively flat slope of the attention curves in Figure 2.2(a). In order to define which users should be classified as "elites", we seek a tradeoff between (a) keeping each category relatively small, so as not to include users who are not distinguishable from ordinary users, while (b) maximizing the volume of attention that is accounted for by each category. In addition, it is also desirable to make the four categories the same size, so as to facilitate comparisons. Balancing these requirements, we therefore choose 5K as a cut-off for the elite categories.

Consistent with this view, we find that the population of users identified by the activity sample is somewhat different from the snowball sample: the intersection of the two populations is only 20% (100,000 accounts). However, the intersection of the top k users in each population increases as k decreases: for the top $5,000$ users in each category, the intersection is 41%, and for the top $1,000$ users it is 51%. Thus, although the population of consistently active users is somewhat different from those reached with the snowball sample, the most frequently listed users in both populations tend to be similar. In addition, Figure 2.2(b) shows that the attention paid to the top k users in the four categories is essentially the same as for the snowball sample. Thus in the rest of this paper,



(a) Snowball sample



(b) Activity sample

Figure 2.2: Average fraction of # following (blue line) and # tweets (red line) for a random user that are accounted for by the top K elites users crawled

when we talk about "celebrity", "media", "organization", "blog", we mean the top 5K users listed as "celebrity", "media", "organization", "blog", respectively, drawn from the snowball sample. Table 2.3 shows the top 5 users in each of the four categories.

To confirm the validity of these categories, we now consider the number of URLs introduced by various categories. As Table 2.4 (left column) shows, the

Table 2.3: Top 5 users in each category

Celebrity	*Media*	*Org*	*Blog*
aplusk	cnnbrk	google	mashable
ladygaga	nytimes	Starbucks	problogger
TheEllenShow	asahi	twitter	kibeloco
taylorswift13	BreakingNews	joinred	naosalvo
Oprah	TIME	ollehkt	dooce

Table 2.4: # of URLs initiated by category

category	# of URLs	per-capita # of URLs
celeb	139,058	27.81
media	5,119,739	1023.94
org	523,698	104.74
blog	1,360,131	272.03
other	244,228,364	6.10

vast majority of URLs are initiated by ordinary users, not by any of the elite categories. This result, however, is deceptive: as we have just determined, our elite categories number only $20K$ users in total, whereas we classify over $40M$ users in the "ordinary" category. A more calibrated view is presented in the right hand column of Table 2.4, which shows the per-capita number of URLs originating from various categories. Here it is clear that users classified as "media" far outproduce all other categories, followed by bloggers, organizations, and celebrities. In contrast to the previous result, ordinary users originate on average only about 6 URLs each—far fewer than any category of elite users.

Conceivably, our classification scheme above has omitted an important cate-

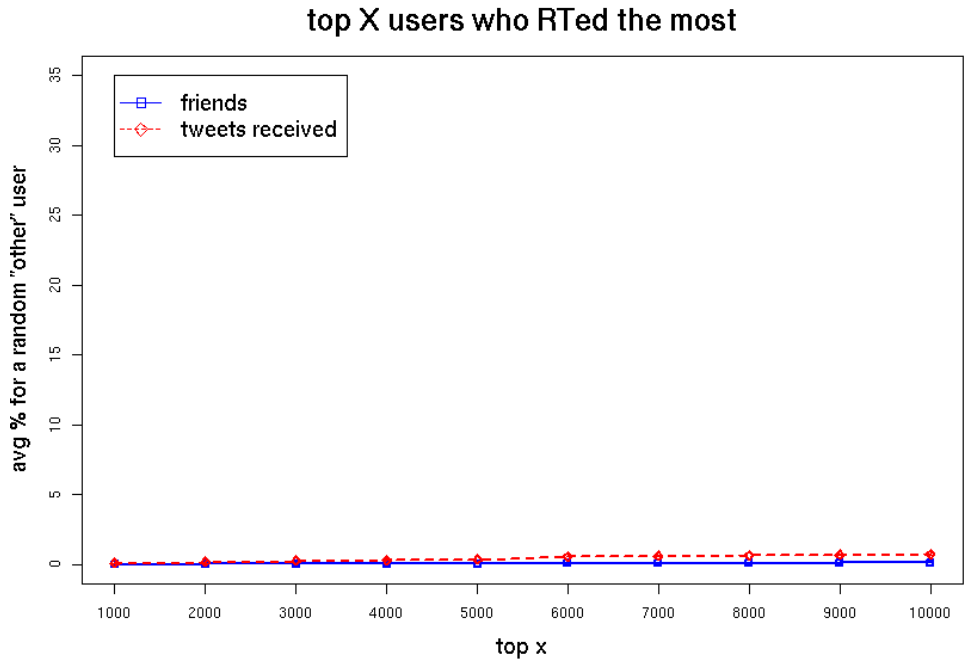


Figure 2.3: Average fraction of # following (blue line) and # tweets (red line) for a random user that are accounted for by the top K most retweeted users in the "Other" category

gory; that is, within the current "other" category may be hidden additional categories of opinions. As Figure 2.2.1 shows, however, even the top 10,000 most followed of these users accounts for a negligible fraction of attention among the remaining population.

2.2.2 Homophily of influence

As indicated above, the top $20K$ elite users account for almost 50% of all attention within Twitter; yet this population of users comprises less than 0.05% of the population. In other words, although Twitter clearly reflects the conventional wisdom that audiences have become increasingly fragmented, it nevertheless shows remarkable concentration of information production and received atten-

Table 2.5: Information flow among the elite categories

% of friends	in celeb	in media	in org	in blog
celeb	**30.56**	3.63	1.99	1.64
media	3.59	**16.67**	2.07	2.15
org	3.62	3.33	**7.38**	2.65
blog	4.41	2.27	2.03	**10.25**
% of tweets	from celeb	from media	from org	from blog
celeb	**38.27**	6.23	1.55	3.98
media	3.91	**26.22**	1.66	5.69
org	4.64	6.41	8.05	**8.70**
blog	4.94	3.89	1.58	**22.55**

tion among a relatively small number of actors. Even if the media has lost attention relative to other elites, information flows have not become egalitarian by any means.

The prominence of elite users raises the question of how these different categories listen to each other. To address this issue, we compute the percentage of following links and received tweets among elite categories. Specifically, Table 2.5 shows the average percentage of friends/tweets category i get from category j. Table 2.5 shows striking homophily with respect to attention: celebrities overwhelmingly pay attention to other celebrities, media actors pay attention to other media actors, and so on. The one slight exception to this rule is that organizations pay more attention to bloggers than to themselves. In general, in fact, attention paid by organizations is more evenly distributed across categories than for any other category.

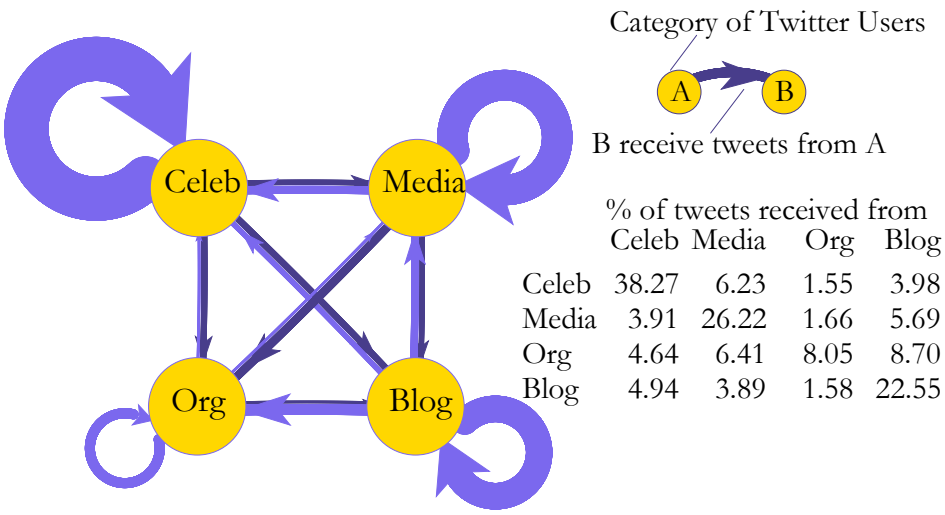


Figure 2.4: Share of attention among elite categories

Figure 2.2.1, it should be noted, shows only how many URLs are received by category i from category j, a particularly weak measure of attention for the simple reason that many tweets go unread. A stronger measure of attention, therefore, is to consider instead only those URLs introduced by category i that are subsequently retweeted by category j.

Before proceeding, it is helpful to differentiate between two mechanisms by which information can diffuse in Twitter. The first is via retweeting, when a user, having received a tweet, subsequently rebroadcasts it to his or her own followers. In some instances, users retweet each other using the official retweet function provided by Twitter, but in other cases they credit the retweet with an informal convention, most commonly either "RT @user" or "via @user." The second mechanism is what we label reintroduction, where a user independently tweets a URL that has previously been introduced by another user.

In addition to attention, Table 2.6 shows how much information originating

Table 2.6: RTs among categories

	by celeb	by media	by org	by blog	by other	TOTAL
celeb	4,334	1,489	1,543	5,039	1,070,318	1,082,723
media	4,624	40,263	7,628	32,027	5,204,719	5,289,261
org	1,570	2,539	18,937	11,175	1,479,017	1,513,238
blog	3,710	6,382	5,762	99,818	3,457,631	3,573,303
other	34,455	93,934	86,630	318,537	34,814,456	35,348,012

from each category is retweeted by other categories, while Table 2.7 shows how much is subsequently reintroduced. As with attention, both retweeting and reintroduction activities are strongly homophilous among elite categories; however, bloggers are disproportionately responsible for retweeting and reintroducing URLs originated by all categories. This result reflects the characterization of bloggers as recyclers and filters of information; however, Table 2.6 and 2.7 also show that the total number of URLs either RT'd or reintroduced by bloggers is vastly outweighed by the number retweeted or reintroduced by ordinary users. Even though on a per-capita basis, therefore, bloggers disproportionately occupy the role of information recyclers, their actual impact is relatively minimal (see Figure 2.4).

2.3 Revisiting two-step flow theory: what are the opinion leaders?

The two-step flow theory, first proposed in the 50's, is still one of the most successful theories that captured the dueling importance of mass media and inter-

Table 2.7: Re-introductions among categories

	by celeb	by media	by org	by blog	by other	TOTAL
celeb	2,868	1,239	522	1,664	488,229	494,522
media	1,678	205,165	2,439	9,681	2,006,888	2,225,851
org	816	1,511	8,628	3,711	610,373	625,039
blog	1,415	5,644	1,416	52,909	1,148,137	1,209,521
other	45,547	793,741	69,441	335,690	86,853,224	88,097,643

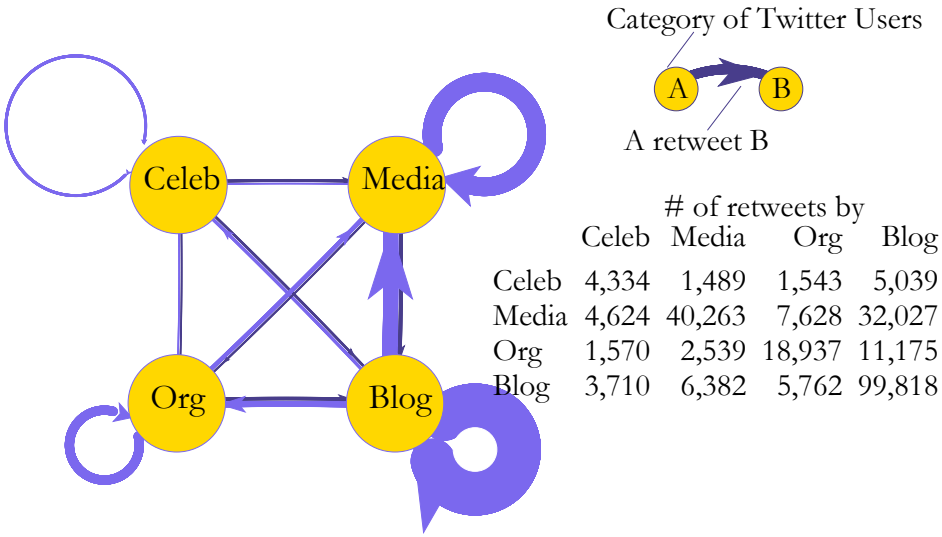


Figure 2.5: RT behavior among elite categories

personal influence. The essence of the two-step flow is that information passes from the media to the masses not directly, as supposed by early theories of mass communication, but rather via an intermediary layer of *opinion leaders*, who act as filters and interpreters for their followers. Although deeply ackowledged by marketers, it has been difficult to identify opinion leaders at a large scale, or quantify their impact to the public. As we have already gathered a confident list of mass media accounts, it becomes a natural problem for us to verify the two-step flow theory on Twitter, and ask, what are the opinion leaders, and

what proportion of the information originating from media sources is broadcast directly to the masses, and what proportion is transmitted indirectly via some population of intermediaries. In addition, we may inquire whether these intermediaries, to the extent they exist, are drawn from other elite categories or from ordinary users, as claimed by the two-step flow theory; and if the latter, in what respects they differ from other ordinary users.

Before proceeding with this analysis, we note that there are two ways information can pass through an intermediary in Twitter. The first is via retweeting, which occurs when a users explicitly rebroadcasts a URL that he or she has received from a friend, along with an explicit acknowledgement of the source—either using the official retweet functionality provided by Twitter or by making use of an informal convention such as "RT @user" or "via @user." Alternatively, a user may tweet a URL that has previously been posted, but without acknowledgement of a source; in this case we assume the information was independently rediscovered and label this a "reintroduction" of content. For the purposes of studying when a user receives information directly from the media or indirectly through an intermediary, we treat retweets and reintroductions equivalently. If the first occurrence of a URL in Twitter came from a media user, but a user received the URL from another source, then that source can be considered an intermediary, whether they are citing the source within Twitter by retweeting the URL, or reintroducing it, having discovered the URL outside of Twitter.

To quantify the extent to which ordinary users get their information indirectly versus directly from the media, we sampled 1M random ordinary users[4],

[4]As before, performing this analysis for the entire population of over 40M ordinary users proved to be computationally unfeasible.

and for each user, counted the number n of bit.ly URLs they had received that had originated from one of our 5K media users, where of the 1M total, 600K had received at least one such URL. For each member of this 600K subset we then counted the number n_2 of these URLs that they received via non-media friends; that is, via a two-step flow. The average fraction $n_2/n = 0.46$ therefore represents the proportion of media-originated content that reaches the masses via an intermediary rather than directly. As Figure 2.6 shows, however, this average is somewhat misleading. In reality, the population comprises two types—those who receive essentially all of their media-originating information via two-step flows and those who receive virtually all of it directly from the media. Unsurprisingly, the former type is exposed to less total media than the latter. What is surprising, however, is that even users who received up to 100 media URLs during our observation period received all of them via intermediaries.

Who are these intermediaries, and how many of them are there? In total, the population of intermediaries is smaller than that of the users who rely on them, but still surprisingly large, roughly 490K, the vast majority of which (484K, or 99%) are classified as ordinary users, not elites. To illustrate the difference, we note that whereas the top 20K elite users collectively account for nearly 50% of attention, the top 10K most-followed ordinary users account for only 5%. Moreover, Figure 2.6c also shows that at least some intermediaries also receive the bulk of their media content indirectly, just like other ordinary users.

Comparing Figure 2.6a and 2.6c, however, we note that intermediaries are not like other ordinary users in that they are exposed to considerably more media than randomly selected users (9165 media-originated URLs on average vs. 1377), hence the number of intermediaries who rely on two-step flows is smaller

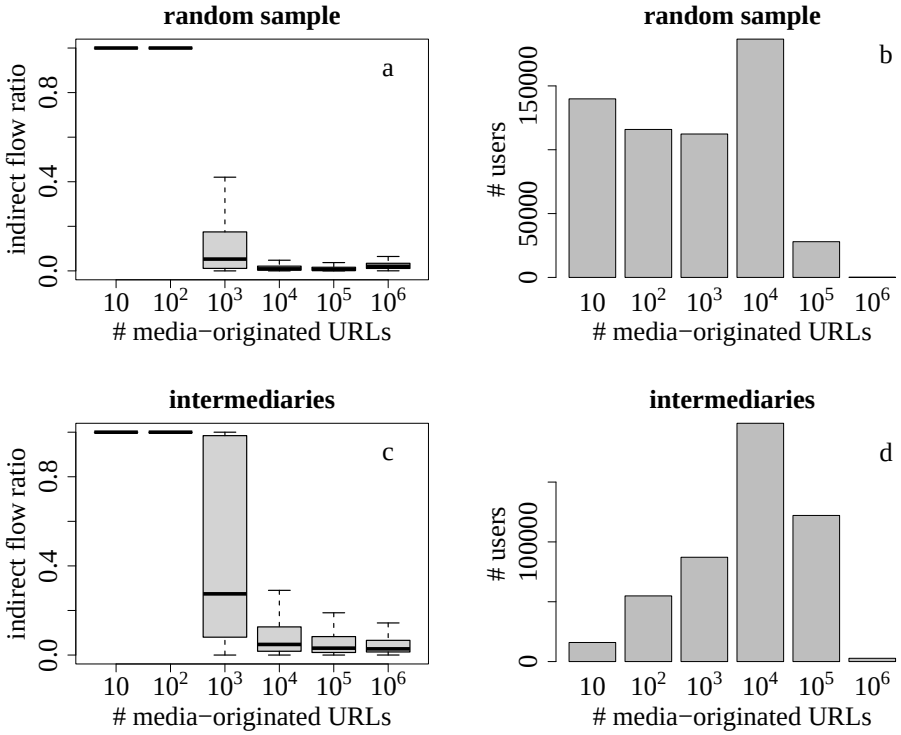


Figure 2.6: Percentage of information that received via an intermediary as a function of total volume of media content to which a user is exposed

than for random users. In addition, we find that on average intermediaries have more followers than randomly sampled users (543 followers versus 34) and are also more active (180 tweets on average, versus 7). Finally, Figure 2.7 shows that although all intermediaries, by definition, pass along media content to at least one other user, a minority satisfies this function for multiple users, where we note that the most prominent intermediaries are disproportionately drawn from the 4% of elite users—Ashton Kucher (aplusk), for example, acts as an intermediary for over 100,000 users.

Interestingly, these results are all broadly consistent with the original conception of the two-step flow, advanced over 50 years ago, which emphasized

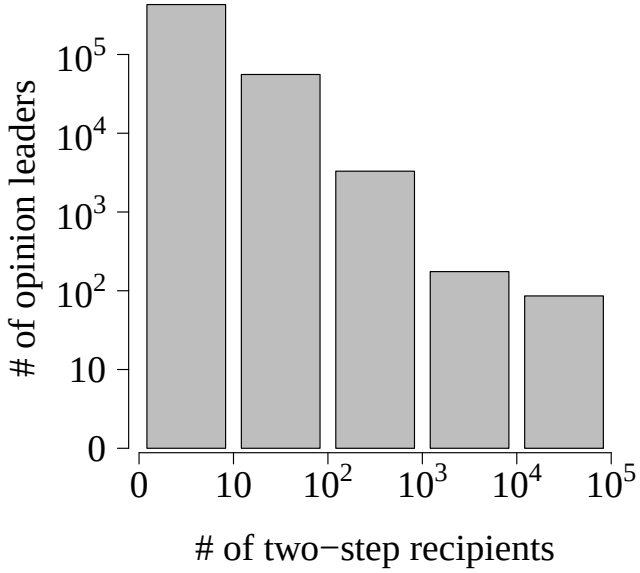


Figure 2.7: Frequency of intermediaries binned by # randomly sampled users to whom they transmit media content

that opinion leaders were "distributed in all occupational groups, and on every social and economic level," corresponding to our classification of most intermediaries as ordinary [19]. The original theory also emphasized that opinion leaders, like their followers, also received at least some of their information via two-step flows, but that in general they were more exposed to the media than their followers—just as we find here. Finally, the theory predicted that opinion leadership was not a binary attribute, but rather a continuously varying one, corresponding to our finding that intermediaries vary widely in the number of users for whom they act as filters and transmitters of media content. Given the length of time that has elapsed since the theory of the two-step flow was articulated, and the transformational changes that have taken place in communications technology in the interim—given, in fact, that a service like Twitter was likely unimaginable at the time—it is remarkable how well the theory agrees

with our observations.

2.4 The interaction between people and content

The results in Section 2.2.1 demonstrate the "elite" users account for a substantial portion of all of the attention on Twitter, but also show clear differences in how the attention is allocated to the different elite categories. As illustrated previously, users have natural preferences on content, and influence on Twitter is topic-dependent[24, 9]. It is therefore interesting to consider what kinds of content is being shared by these categories.

Given the large size of the URL population in our observation period ($260M$), and the large number of ways in which one can classify content (video vs. text, news vs. entertainment, political news vs. sports news, etc.), classifying even a small fraction of URLs according to content is an onerous task. Bakshy et al [3], for example, used Amazon's Mechanical Turk to classify a stratified sample of 1,000 URLs along a variety of dimensions; however, this method does not scale well to larger sample sizes.

Instead, we restrict attention to URLs originated by the New York Times which, with over 2.5M followers, is the second-most followed news organization on Twitter after CNN Breaking News. NY Times, however, is roughly ten times as active as CNN Breaking News, so is a better source of data. To classify NY Times content, we exploit a convenient feature of their format—namely that all NY Times URLs are classified in a consistent way by the section in which they appear (e.g. US, World, Sports, Science, Arts, etc) [5]. Of the 6398 New

[5] http://www.nytimes.com/year/month/day/category/

York Times bit.ly URLs observed, 6370 could be successfully unshortened and assigned to one of 21 categories. Of these, however, only 9 categories had more than 100 URLs over the observation period, one of which—"NY region"—was highly specific to the New York metropolitan area; thus we focused our attention on the remaining 8 topical categories. Figure 2.8 shows the overall RT and reintroduction rates by category. World news is the most popular category, followed by US news, business, and sports, where increasingly niche categories like Health, Arts, Science, and Technology are less popular still. In general, the overall pattern is replicated for all categories of users, but there are some minor deviations: In particular, organizations show disproportionately little interest in business and arts-related stories, and disproportionately high interest in science, technology, and possibly world news. Celebrities, by contrast, show greater interest in sports and less interest in health, while the media shows somewhat greater interest in US news stories.

In addition, we also consider the accumulated RT/Reintroduction behavior for a small selection of the most popular URLs. As Figure 2.9 shows, the link to the official White House blog, which expressed the administration's initial response to the Haiti earthquake, was rebroadcast in largely the same manner by all categories of users, as was the announcement of President Obama winning the Nobel Peace Prize. By contrast, the news story announcing the unexpected death of the actress Brittany Murphy was rebroadcast largely by bloggers, while the breaking news about Tiger Woods' accident and affair was picked up mostly by the news media and other celebrities. Finally, Figure 2.9 shows two examples of URLs that exhibit very different patterns from news stories. First, the URL for DealPlus, a website for "finding, discussing, and sharing thousands of deals

title.html?ref=category

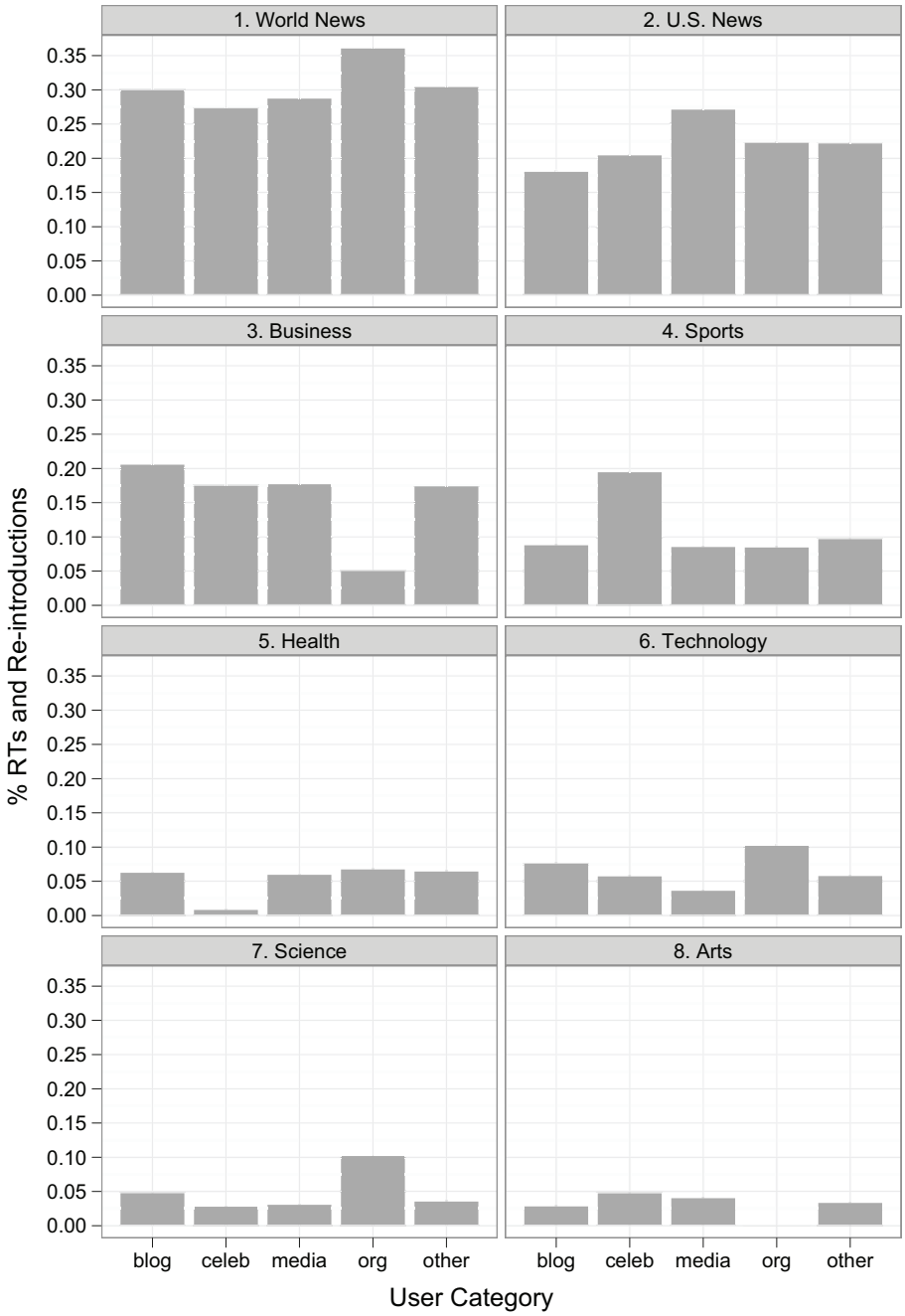


Figure 2.8: Number of RT's and Reintroductions of New York Times stories by content category

and coupons for all types of stores," was popular among ordinary users, but almost completely ignored by all categories of elite users. And second, the video for the song "Brick by Boring Brick," by the band Paramore, was again reposted mostly by ordinary users, but in this case celebrities also reposted it. Although this analysis is far from systematic, it suggests that different categories of users respond to different sorts of content in ways that are consistent with our classification scheme.

2.5 Lifespan of content by category

In addition to different types of content, URLs introduced by different types of elite users or ordinary users may exhibit different lifespans, by which we mean the time lag between the first and last appearance of a given URL on Twitter.

Naively, measuring lifespan seems a trivial matter; however, a finite observation period—which results in censoring of our data—complicates this task. In other words, a URL that is last observed towards the end of the observation period may be retweeted or reintroduced after the period ends, while correspondingly, a URL that is first observed toward the beginning of the observation window may in fact have been introduced before the window began. What we observe as the lifespan of a URL, therefore, is in reality a lower bound on the lifespan. Although this limitation does not create much of a problem for short-lived URLs—which account for the vast majority of our observations—it does potentially create large biases for long lived URLs. In particular, URLs that appear towards the end of our observation period will be systematically classified as shorter-lived than URLs that appear towards the beginning.

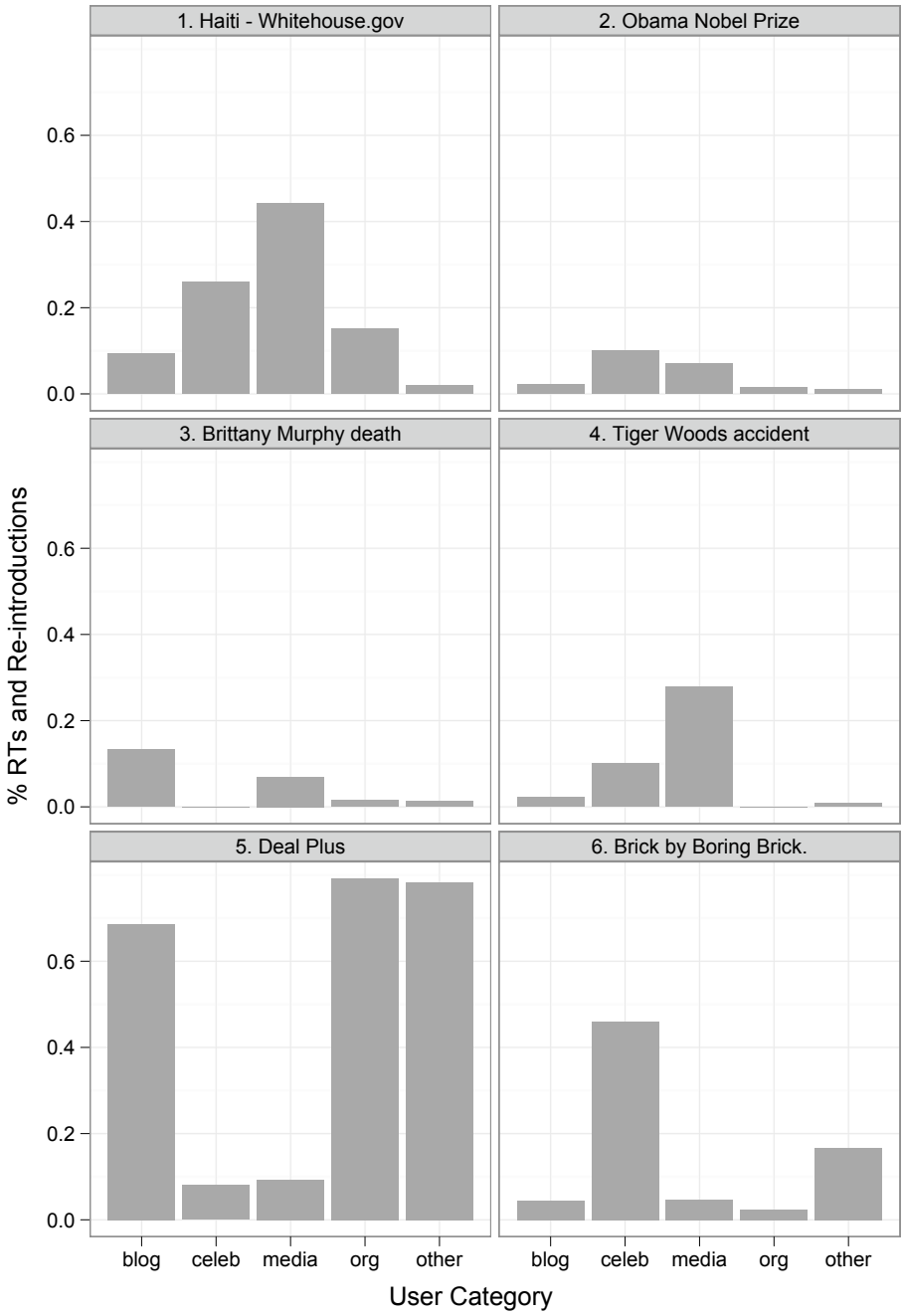


Figure 2.9: Number of RT's and Reintroductions of most popular URLs originating from media and other

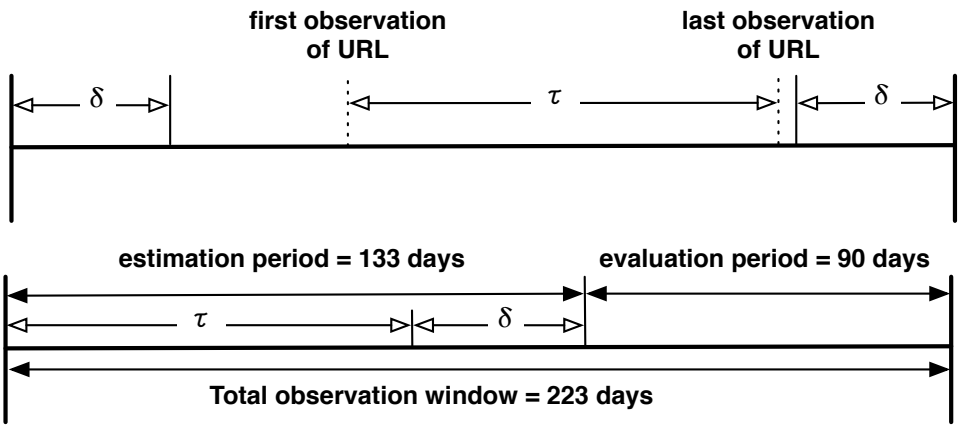
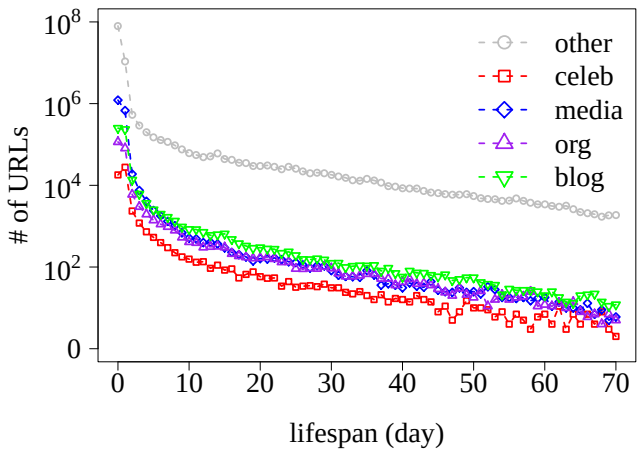


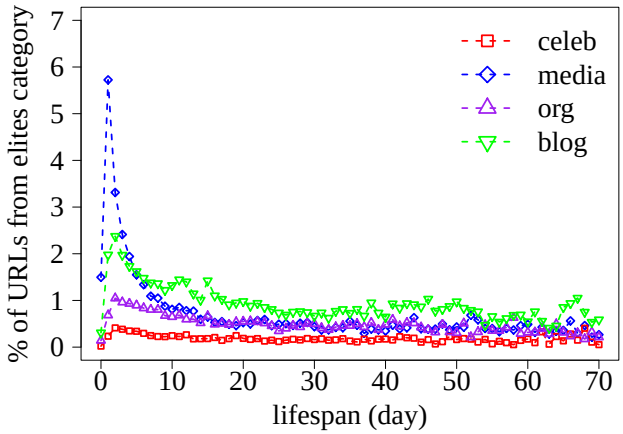
Figure 2.10: (a) Definition of URL lifespan τ (b) Schematic of lifespan estimation procedure

To address the censoring problem, we seek to determine a buffer δ at both the beginning and the end of our 223-day period, and only count URLs as having a lifespan of τ if (a) they do not appear in the first δ days, (b) they first appear in the interval between the buffers, and (c) they do not appear in the last δ days, as illustrated in Figure 2.10(a). To determine δ we first split the 223 day period into two segments—the first 133 day estimation period and the last 90 day evaluation period (see Figure 2.10(b))—and then ask: if we (a) observe a URL first appear in the first $(133 - \delta)$ days and (b) do not see it in the δ days prior to the onset of the evaluation period, how likely are we see it in the last 90 days? Clearly this depends on the actual lifespan of the URL, as the longer a URL lives, the more likely it will re-appear in the future. Using this estimation/evaluation split, we find an upper-bound on lifespan for which we can determine the actual lifespan with 95% accuracy as a function of δ. Finally, because we require a beginning and ending buffer, and because we can only classify a URL as having lifespan τ if it appears at least τ days before the end of our window, we need to pick τ and δ such that $\tau + 2\delta \leq 223$. We determined that $\tau = 70$ and $\delta = 70$ sufficiently satisfied our constraints; thus for the following analysis, we consider

only URLs that have a lifespan $\tau \leq 70$ [6].



(a) Count



(b) Percent

Figure 2.11: (a) Count and (b) percentage of URLs initiated by 4 categories, with different lifespans

Having established a method for estimating URL lifespan, we now explore the lifespan of URLs introduced by different categories of users, as shown in Figure 2.11(a). URLs initiated by the elite categories exhibit a similar distribution over lifespan to those initiated by ordinary users. As Figure 2.11(b) shows,

[6]We also performed our analysis with different values of τ, finding very similar results; thus our conclusions are robust with respect to the details of our estimation procedure.

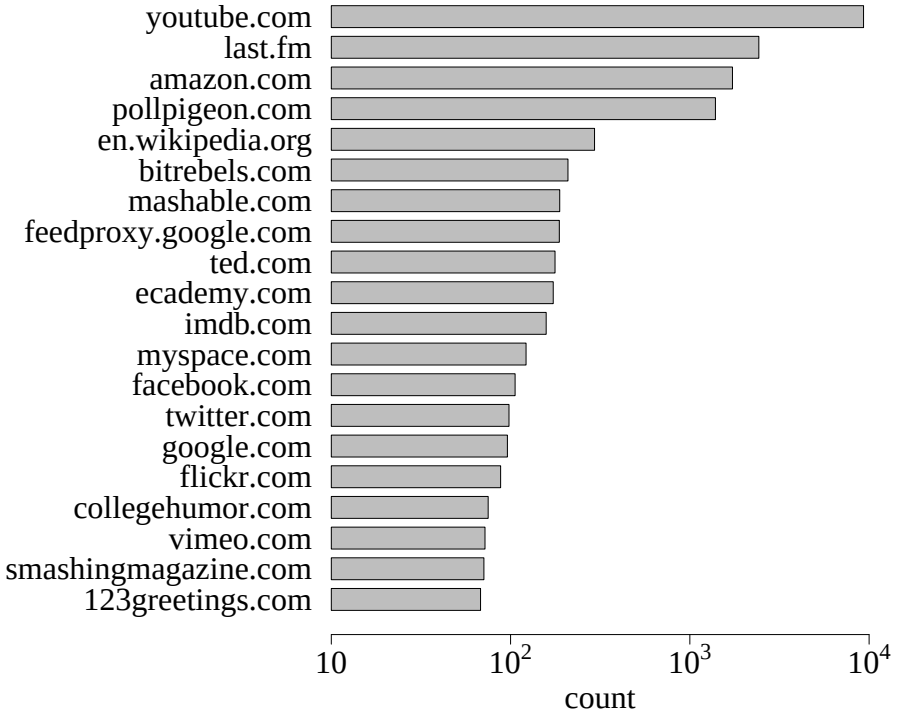


Figure 2.12: Top 20 domains for URLs that lived more than 200 days

however, when looking at the percentage of URLs of different lifespans initiated by each category, we see two additional results: first, URLs originated by media actors generate a large portion of short-lived URLs (especially URLs with $\tau = 0$, those that only appeared once); and second, URLs originated by bloggers are overrepresented among the longer-lived content. Both of these results can be explained by the type of content that originates from different sources: whereas news stories tend to be replaced by updates on a daily or more frequent basis, the sorts of URLs that are picked up by bloggers are of more persistent interest, and so are more likely to be retweeted or reintroduced months or even years after their initial introduction. Twitter, in other words, should be viewed as a subset of a much larger media ecosystem in which content exists and is repeatedly rediscovered by Twitter users. Some of this content—such as daily news

stories—has a relatively short period of relevance, after which a given story is unlikely to be reintroduced or rebroadcast. At the other extreme, classic music videos, movie clips, and long-format magazine articles have lifespans that are effectively unbounded, and can seemingly be rediscovered by Twitter users indefinitely without losing relevance.

To shed more light on the nature of long-lived content on Twitter, we used the bit.ly API service to unshorten 35K of the most long-lived URLs (URLs that lived at least 200 days), and mapped them into 21034 web domains. As Figure 2.12 shows, the population of long-lived URLs is dominated by videos, music, and consumer goods. Two related points are illustrated by Figure 2.13, which shows the average RT rate (the proportion of tweets containing the URL that are retweets of another tweet) of URLs with different lifespans, grouped by the categories that introduced the URL[7]. First, for ordinary users, the majority of appearances of URLs after the initial introduction derives not from retweeting, but rather from reintroduction, where this result is especially pronounced for long-lived URLs. For the vast majority of URLs on Twitter, in other words, longevity is determined not by diffusion, but by many different users independently rediscovering the same content, consistent with our interpretation above. Second, however, for URLs introduced by elite users, the result is somewhat the opposite—that is, they are more likely to be retweeted than reintroduced, even for URLs that persist for weeks. Although it is unsurprising that elite users generate more retweets than ordinary users, the size of the difference is nevertheless striking, and suggests that in spite of the dominant result above that content lifespan is determined to a large extent by the type of content, the source of its origin also impacts its persistence, at least on average—a result that is consistent

[7]Note here that URLs with lifespan = 0 are those URLs that only appeared once in our dataset, thus the RT rate is zero.

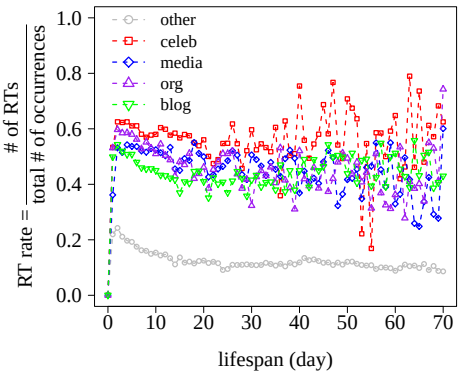


Figure 2.13: Average RT rate by lifespan for each of the originating categories

with previous findings [3].

2.6 Conclusion

In this chapter, we investigated the influencer problem by incorporating exogenous influence into the context of Twitter. In particular, we find that although audience attention has indeed fragmented among a wider pool of content producers than classical models of mass media, attention remains highly concentrated, where roughly 0.05% of the population accounts for almost half of all posted URLs. Within this population of elite users, moreover, we find that attention is highly homophilous, with celebrities following celebrities, media following media, and bloggers following bloggers. Second, we find considerable support for the two-step flow of information—almost half the information that originates from the media passes to the masses indirectly via a diffuse intermediate layer of opinion leaders, who although classified as ordinary users, are more connected and more exposed to the media than their followers. Third, we

find that although all categories devote a roughly similar fraction of their attention to different categories of news (World, U.S., Business, etc), there are some differences—organizations, for example, devote a surprisingly small fraction of their attention to business-related news. We also find that different types of content exhibit very different lifespans: media-originated URLs are disproportionately represented among short-lived URLs while those originated by bloggers tend to be overrepresented among long-lived URLs. Finally, we find that the longest-lived URLs are dominated by content such as videos and music, which are continually being rediscovered by Twitter users and appear to persist indefinitely. We will further investigate the role of content in the diffusion process in next chapter.

By restricting our attention to URLs shared on Twitter, our conclusions are necessarily limited to one narrow cross-section of the media landscape. An interesting direction for future work would therefore be to apply similar methods to quantifying influence via more traditional channels, such as TV and radio on the one hand, and interpersonal interactions on the other hand. Moreover, although our approach of defining a limited set of predetermined user-categories allowed for relatively convenient analysis and straightforward interpretation, it would be interesting to explore automatic classification schemes from which additional user categories could emerge.

CHAPTER 3

THE ROLE OF CONTENT

As shown in previous chapter, the content originated by different people exhibit different lifespan, however, the connection between content and authors are relatively weak thus does not have much predictive power (for example, bloggers can write about a variety of things). In this chapter, we present a large scale empirical study directly on the textual content in relationship to the perisistence of information. Our goal is to look for intrinsic qualities of the content that effectively affect the dissemination process, especially, resulting in different lifespans of information. We make two main contributions:

- We build a classifier that predicts the decay/persistence of information with textual features, providing one of the first empirical studies of the connection between content and temporal variations of information in online social media.
- We investigate the properties of the text that are associated with different temporal patterns, finding significant differences in word usage and sentiment between rapidly-fading and long-lasting information.

In the following sections of this chapter, we first provide an overview of the data we are using for this study, then present a binary classifier that predicts the persistence of URLs, using texual features from webpages pointed to by the URLs. We further examine and discuss different aspects of the content that are correlated with the difference in temporal patterns. In the end, we provide some additional insights about the quality of YouTube videos in relationship to the decay time.

3.0.1 Data

Summary

In this study, we used the dataset publicly shared by the authors of [47][1], consisting of approximately 20%-30% of all the tweets generated between June 1, 2009 and December 31, 2009. We only study the temporal patterns of bit.ly URLs for two reasons, following the arguments of [43]. First, shortened URLs have a unique token that is easily traceable in individual tweets. Second, the associated webpages provide a much richer source of content beyond the 140-character limit of tweets. From the total 476M tweets contained in the dataset, we find 118M distinct URLs embedded in 186M tweets. Among all the URLs, nearly half of them (56M) are bit.ly URLs (i.e., start with http://bit.ly/). For simplicity, we only extract the time series of bit.ly URLs and use them as a representative sample of all temporal patterns. Considering that a large portion of URLs mentioned in Twitter are spam and may not be able to provide meaningful content, we restrict our study to the bit.ly URLs that appeared more than 10 times in retweets [2], which gives us 131K bit.ly URLs. We are able to crawl 117K webpages pointed to by these bit.ly URLs, the remaining 14K URLs that we fail to crawl are mostly misspelled or linked to webpages that no longer exist.

We further restrict our study to URLs that are mentioned more than 50 times in order to remove spam and have sufficient observations to measure temporal dynamics, which leaves us with 21K URLs. In the rest of this paper, when we talk about URLs and temporal patterns, we mean these 21K bit.ly URLs and the temporal pattern in their time series.

[1] http://snap.stanford.edu/data/twitter7.html
[2] We recognize a post as retweet when it contains "RT @" or "via @".

Persistence of URLs

After extracting the data of interest, we first propose a quantitative metric of persistence and present some insights on the overall temporal pattern of the URLs we study.

As the focus of this study is how fast URLs fade, we measure decay rates following peak attention. For each URL u, let the hour of maximum attention (also called the peak of attention) be hour 0. Then the *decay time* t_u is defined as the hour after the peak when the number of mentions first reaches 75% of the total. Instead of measuring the time lag between the first and last mention of a given URL [43], we intentionally choose to measure the time lag from the peak of attention to the point when the URL fades away, as given the limited observation window when the dataset was collected, it is not obvious to determine when exactly a URL was first introduced or last appeared on Twitter. The distribution of t_u shows heavy tail(see Figure 3.0.1), as found previously in the distribution of URL lifespan[43]. Among all URLs we studied, the mean t_u is 217.3 hour and the median t_u is 19 hours.

We further examine the relationship between t_u and the overall popularity of URLs. Figure 3.0.1 shows the average number of tweets and retweets accumulated by each URL as a function of t_u. Given the power-law distribution of t_u, we bin URLs by the integer part of $\log_2(t_u)$, and calculate the mean for each bin. Although the persistent URLs are mentioned in slightly more tweets, the rapidly-fading URLs do better at attracting retweets. This result is consistent with previous findings that the longevity of information is determined not by diffusion, but by independent generation of tweets of the same content over time [43].

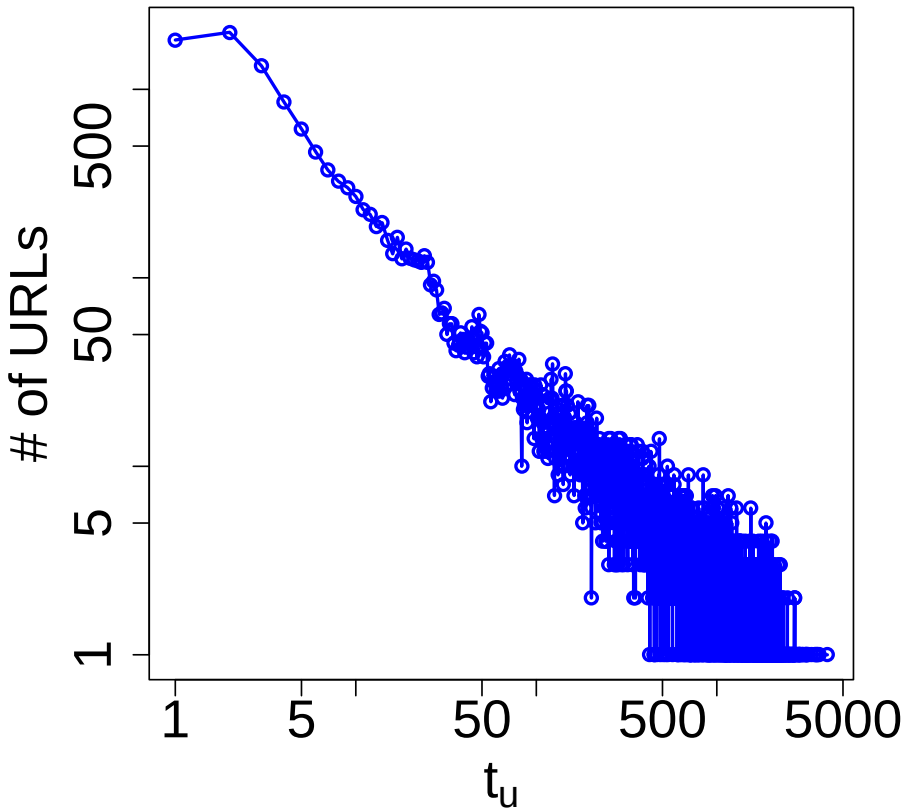


Figure 3.1: Distribution of URL decay time t_u

3.0.2 Predicting temporal patterns based on content

In this section, we formally define the temporal pattern classification task and present our findings.

Identifying information with two distinct temporal patterns

We start by casting our question into a binary classification problem in which class 1 is defined as consisting of those URLs with $t_u < 6$ and class 0 is defined

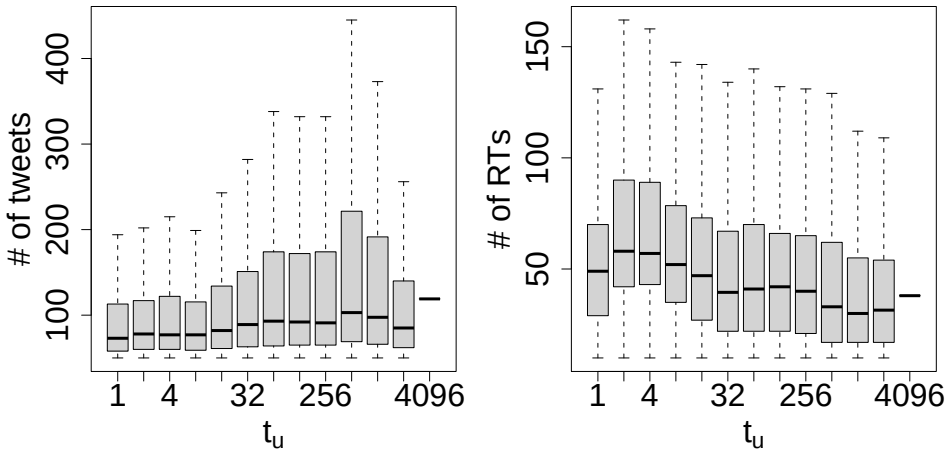


Figure 3.2: URL overall popularity as a funtion of t_u

defined as consisting of those URLs with $t_u > 24$. In this way we get a positive class with 7042 examples and a negative class with 6185 examples. We exclude the 7K examples in the middle, as the data is much noisier and the persistence of these URLs is ambiguous — our goal in this first exploration of persistence prediction is to construct a well-defined and tractable task from which we can understand whether there are features that meaningfully separate rapidly-fading URLs from long-lasting ones.

To better illustrate our classification scheme, we apply the time series normalization method introduced in [47] and calculate the centroid of time series for each class, as shown in Figure 3.0.2. The two classes we define do in fact collectively exhibit very different temporal patterns: URLs of the positive class fade away slowly, with periodic, multiple peaks of attention; URLs of the negative class have a single spike and a rapid decay afterwards.

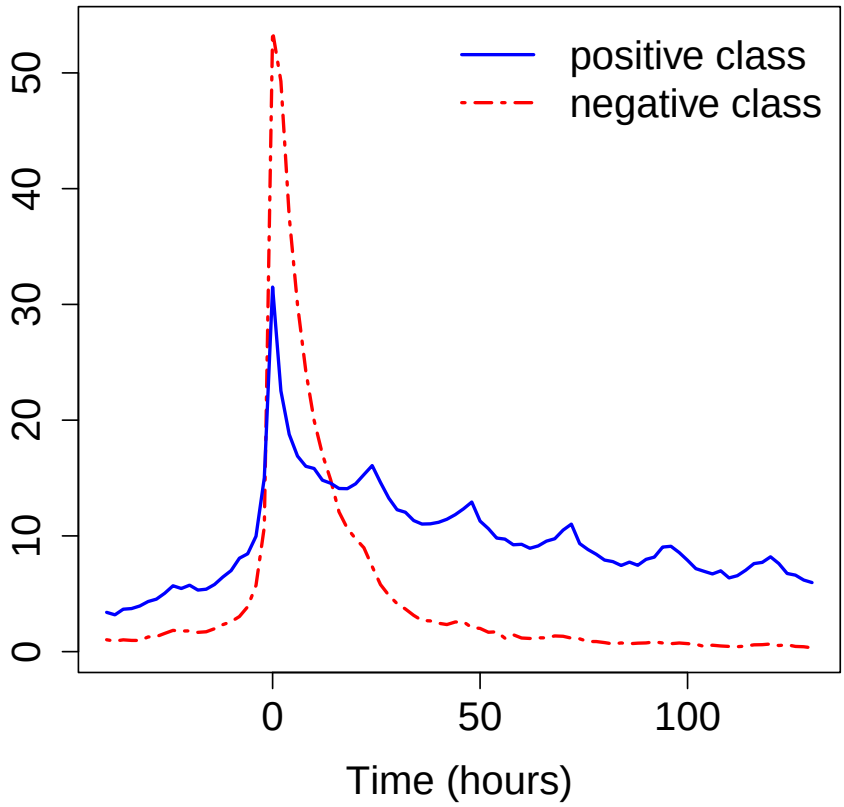


Figure 3.3: Normalized time series centroids for two classes

Features

To predict the temporal class of URLs, we extract and experiment with the following four incremental sets of unigram features from the HTML webpages linked by the URLs (one-character tokens and those that consist only of numbers are filtered out):

- Header. The text in the header of HTML, within tags "<title>","<description>", and "<keywords>".

Table 3.1: Feature size

Feature	*# of unique unigram terms*
Header	18471
Header + URL	27433
Header + Body	59475
Header + Body + URL	76487

- Header + URL. In addition to Header, this feature set also uses the terms tokenized from the URL links embedded in the HTML (i.e.,within "<href>").
- Header + Body. In addition to Header, this feature set includes all the text in the body of HTML.
- Header + URL + Body. This feature set combines all the features mentioned above.

As mentioned above, to get more meaningful unigram features, after tokenizing all the textual content into word terms, we filter the terms with length 1 (e.g., "s", "t") and the terms consisting of only numbers. As the dimension increases tremendousely in the last 3 sets of features, we also filter the infrequent terms (i.e., terms with total frequency less than 20). Table 3.0.2 gives a summary of the number of features in each set.

Table 3.2: Results for predicting lastingness of information

Feature	*Accuracy*	*Pos F1*	*Neg F1*
Header	0.6909	0.7399	0.6186
Header + URL	0.7177	0.7666	0.6423
Header + Body	0.7136	0.7664	0.6296
Header + Body + URL	0.7224	0.7708	0.6478

Classifier performance

To predict the persistence of webpages, we employ a Support Vector Machine (SVM)[3] classifier with a binary representation of unigram features (if a term appears in a webpage, the corresponding coordinate has value 1, and value 0 otherwise). To work with high-dimensional features, we use the linear SVM kernel for efficiency. We also apply the default parameters for SVM classifier for a fair comparison among different sets of features. Table 3.2 gives the performance of classifiers with different sets of features using 10-fold cross validation.

Table 3.2 shows that in general, the simple linear-kernel SVM classifier can predict the persistent/rapidly-fading category of URLs with impressively high accuracy (around 70%), as comparedd to 53% for always predicting positive. Also, the F1 score for positive class is around 75%, which shows a remarkable balance of precision and recall at identifying the persistent content. This result provides strong evidences for the connection between the content of HTML pages and the persistence of the associated URLs. Moreover, comparing across 4 feature sets, we see that the more information we have about the content, the better the classifier performs. This finding further confirms the relationship be-

[3]The SVM package we use is SVMLight, `http://svmlight.joachims.org/`

tween textual content and the persistence of attention of the information.

3.0.3 How temporal patterns vary with types of content

The SVM classifier shows that the content provides enough information to predict persistence reasonably well. However, SVMs are not as effective at providing a readily comprehensible sense for which properies of the text are the most related to the variations in temporal patterns. Here we address this question, by looking more closely at the textual content and identifying the aspects that exhibit the most significant difference across temporal classes.

LIWC analysis

Linguistic Inquiry and Word Count (LIWC) [32] is a widely used text analysis tool that maps words onto 60 pre-defined categories, covering linguistic, psychological, and social dimensions. Using LIWC categories, we start by comparing the distribution of words across two classes.

We say a LIWC category occurs in a URL when we find at least one word under that category from the header of the associated HTML page.[4] Figure 3.0.3 shows the percentage of occurrence for all LIWC categories in webpages from two classes. As illustrated by Figure 3.0.3, the two classes differ the most in the following three groups of LIWC categories,

[4]We also conduct the same analysis with text from the other 3 feature sets, however, since the number of words increases markedly in these feature sets, and LIWC dictionary many times maps a word into multiple categories, the binary vector for each URL is easily saturated and the $f_w(t)$ curve becomes too flat to show interesting difference.

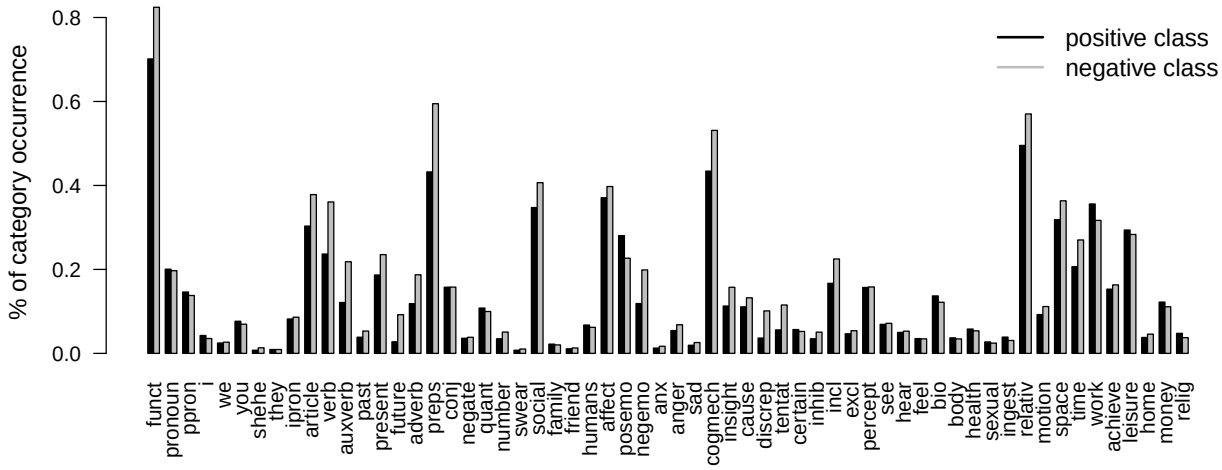


Figure 3.4: Class distribution in 60 LIWC dimensions, using words from HTML header

- Emotion: *posemo* (positive emotion), *negemo* (negative emotion).
- Cognitive process: *cogmech* (cognitive process), *insight* (words like *think*, *know*, *consider*), *incl* (inclusive, words like *and*, *with*, *include*), *discrep* (discrepancy, words like *should*, *would*, *count*).
- Part of speech: *verb* (common verbs), *auxverb* (auxiliary verbs), *preps* (prepositions), *present* (present tense, words like *is*,*does*, *hear*), *future* (future tense, words like *will*,*gonna*).

To better see the trend in the frequency of specific categories as a function of t_u, for each category w, we define $f_w(t)$ as the fraction of occurrences of w in all URLs u for which $t_u = t$, and plot $y = f_w(t)$ for different groups of LIWC categories in Figure 3.0.3.

Again, to balance the power-law distribution of t_u, we bin t_u by integer part of $\log_2(t_u)$, and plot the value $f_x(w)$ for each bin x (instead of hour x). In this way, the later bins would still contain a substantial number of URLs so that the prob-

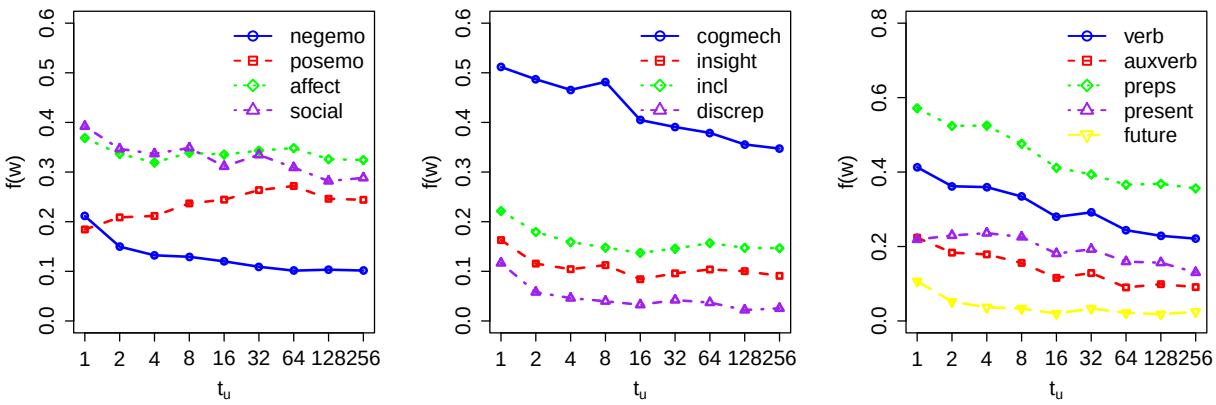


Figure 3.5: Trending LIWC categories

abilistic curve is smoother. Similar as in [6, 17] we find the sentiment of content plays an important role in its dynamics: there is a clear trend of words with positive emotion rising in the persistent content, and the opposite for words with negative emotion. However, the amount of words related to affect stays more or less constant across t_u. We also see a drop of words related to cogntive process when t_u increases, suggesting that, content associated with more complicated cognitive process can be more viral[6], yet not so persistent. Not surprisingly, we find that rapidly-fading content with more words related to actions (verb, auxverb, preps) and tense (present, future), presumably because these webpages contain more action-demanding, time-ciritical information that expires after a certain event or time.

Topic analysis

Although LIWC offers the most straightforward insights from the text, as a manually-generated, pre-defined category system, it is limited by the underlying psychololinguistic concepts. To extend the dimensions of text described

in LIWC, we also build topic models that represent mixtures of words, and see how these topics vary across our temporally-defined classes. For this we use Latent Dirichlet Allocation (LDA)[5], a flexible generative model for collections of discrete data[7]. Here, we use it to find proper underlying generative probabilistic semantics from content. We use the corpus consisting of the unigrams in the two classes. With the topic distribution for each document, we try to study whether the temporal patterns are correlated with "topics". First, we will show the probability of topics in the two classes and find those topics with significant differences across different topics. Then we interpret these topics to find some differences between persistent webpages and rapidly-fading webpages. As for the details of running LDA, we use the features in "header+body" because we find that when using features from URLs, the results will include some irregular words, while with only "header", it cannot include enough words in detail.

First, since the output of LDA provides a continuous value of topic weight for each document, we cast it into binary by assigning 1 when the weight is above the default value. For each topic, we compute the probability that one document contains this topic in the positive class and in the negative class respectively. More specifically, we conduct a paired t-test between the two classes on each topic and find that, on 39 topics, the two classes are different at significance level $\alpha = 5\%$. 24 of them are with p-value 0. This shows that these two classes differ significantly in the space of topics. Figure 3.0.3 shows topics distribution in all 50 topics. We notice that the most significant differences occur at topics 18, 25 (with a high probability in rapidly-fading webpages), and topics 32, 37 (with a high probability in persistent webpages).

[5]We use the software from `http://www.cs.princeton.edu/~blei/lda-c/index.html` with the number of topics set to 50.

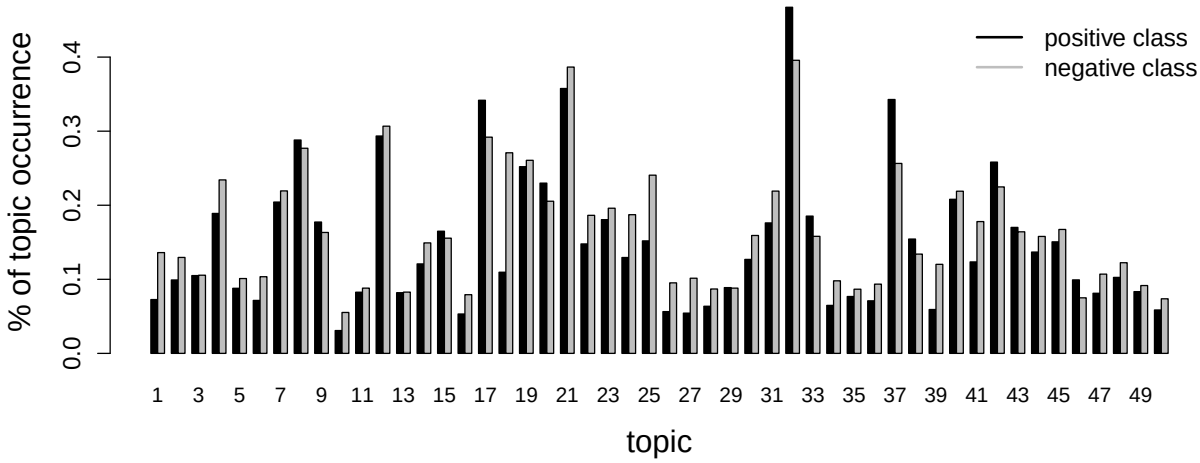


Figure 3.6: Class distribution in 50 LDA topics, using words from HTML header and body

Providing a closer look at those topics, Table 3.0.3 shows top 20 words given by the topic model. We see some similar phenomena as in previous section: words related to strong - and mostly negative - emotions tend to appear more in the topics highly weighted in rapidly-fading webpages. For example, negative words, such as “die”, “freaking”, “incredibly”, “incredible” and “destroy”, show up in topic 18 and 25. In the topics associated with persistent webpages, interestingly, we notice an increase of nouns.

Trending words analysis

After measuring the content in LIWC categories and latent topics, in this part, we examine the content with more details, trying to discover the nuance between classes at the word level. We calculate and compare the most representative words in the two classes. Picking the words to describe a collection of documents can be turned into a trend detection problem: let the webpages of

Table 3.3: LDA Topics

Topic 32	*Topic 37*	*Topic 18*	*Topic 25*
fred	incident	net	die
net	website	dan	gov
care	subscriber	fred	fields
produce	clean	pack	static
incident	rates	gov	say
mas	net	impressed	expensive
office	considering	read	read
hello	potentially	native	york
julian	die	worm	freaking
teen	gov	user	seek
red	money	attempts	destroy
democratic	donation	treatment	dear
boy	dennis	august	supporters
tagging	seek	incredibly	tagged
ways	read	incident	office
opinion	dislike	potentially	microwave
read	il	talented	challenges
different	challenges	die	fred
british	posted	placed	british
heads	kind	busy	august

negative class be the corpus of early period and the webpages of positive class be the later period, negative class can thus be described by the most significant "falling words" whereas the positive class can be described by the most significant "rising words". To do so, we apply the methods as presented in [22] on the Header feature set, and generate the top 20 trending words for each class (see

Table 3.4: Representative words for two temporal classes

Absolute change		*Relative change*		*Prob. change*	
pos	*neg*	*pos*	*neg*	*pos*	*neg*
twibbon	cnn	twibbon	cnn	small	plan
marketing	google	marketing	blogs	mp3	net
support	iphone	contest	source	creative	better
giveaway	blogs	trailer	finest	open	girl
quot	america	review	onion	view	file
free	source	support	apple	vs	touch
best	apple	vote	house	story	smashing
contest	onion	giveaway	iphone	kids	pictures
win	finest	big	white	ipod	using
review	app	movie	guardian	american	organizing
design	house	design	google	know	cancer
trailer	white	quot	users	party	game
vote	jackson	win	app	dj	technology
big	live	good	download	use	want
amp	official	best	america	star	page
movie	uk	love	jackson	things	single
good	obama	green	public	daily	don
home	iran	week	myspace	care	action
music	michael	funny	today	life	watch
love	guardian	version	uk	song	need

Table 3.0.3)[6].

To get the words that are most meaningful, we filter all the numbers, and the

[6]We also tried the same method on the other three feature sets, but as the number of terms largely increases, the data becomes too noisy to be described with a few words, and the results are difficult to interprete.

words with frequency less than 20 (mostly specific names) or greater than 400 (mostly stopwords and website names). As discussed in [22], trending words identified by the three metrics have different bias. Words based on *normalized absolute change* are biased towards words that are frequent in both classes. Words selected by *relative change* are biased towards words frequent in one class but not the other. Words selected by *probablistic change* are the ones that based on the frequency of occurrence in one class, most unlikely to be seen in the other class. Although [22] recommends the probabilistic change as a metric that gives the cleanest results, we find the selected words in all three categories highlight interesting points that reinforce, and provide some intuitive basis for, the results to emerge from the LIWC analysis earlier in this section.

- normalized absolute/relative change. First of all, we again find the persistent content most represented by positive words (e.g. *good*, *best*, *love*). In terms of the semantics of content, the persistent webpages are more related to art (e.g. *music*, *movie*), advertisement, and online marketing (e.g. *twibbon*, *marketing*, *givaway*, *free*, *win*, *review*), whereas the rapidly-fading webpages contain more news (e.g. cnn, google, onion, guardian, blogs), and names (e.g. *michael jackson*, *white house*, *obama*, *iran*, *america*, *uk*).
- probabilistic change. By this metric, we find the trending words for persistent content are more associated with lifestyle (e.g. *party*, *dj*, *care*, *life*, *song*) and family (e.g. *kids*, *care*, *life*), whereas the short-lived content again has a higher portion of words related to time critical concepts (e.g. *technology*, *game*), or action (e.g., *plan*, *touch*, *using*, *want*, *action*, *watch*, *need*).

These results are mostly consistent with the findngs from the previous parts, confirming the prominence of positive emotion in the persistent content, and the

fleetingness of content with many action and time-critical terms. The distinct existence of news and art content of two classes supports the claim by authors of [43] that the persistent content - although not as viral as news - exhibits more association with art.

3.0.4 The quality and persistence of YouTube videos

In our dataset of 20K bit.ly URLs, there is a significant portion (15%) of them linked to YouTube videos. Among these linked videos, 707 are already removed by the user and 2304 are still available online. Noting that the *content* of videos may not be accurately represented by the text of the YouTube page, we conduct a separate study of the persistence of YouTube videos, leveraging the user rating feature YouTube provides - namely, *likes* and *dislikes* - to assess the content from the quality perspective.

First, Figure 3.0.4 shows the distribution of decay time t_u for the 2304 available YouTube videos. In contrast to the overall distribution of t_u for all URLs (see Figure 3.0.1), YouTube videos in general receive a longer span of attention.

We also study the user-rated quality of these 2304 videos as a function of t_u. Figure 8 shows two indicators of the quality (a) the average likes/dislikes rate, (b) the ratio of *bad* videos, for videos in each bin of t_u(the binning method is the same as in previous sections). Interestingly, we find that although the quality of video overall increases with t_u, there is a drop of quality in the middle - videos with medium persistence seem to be of the worst quality.

Sampling videos with different t_u values suggests a further way to break

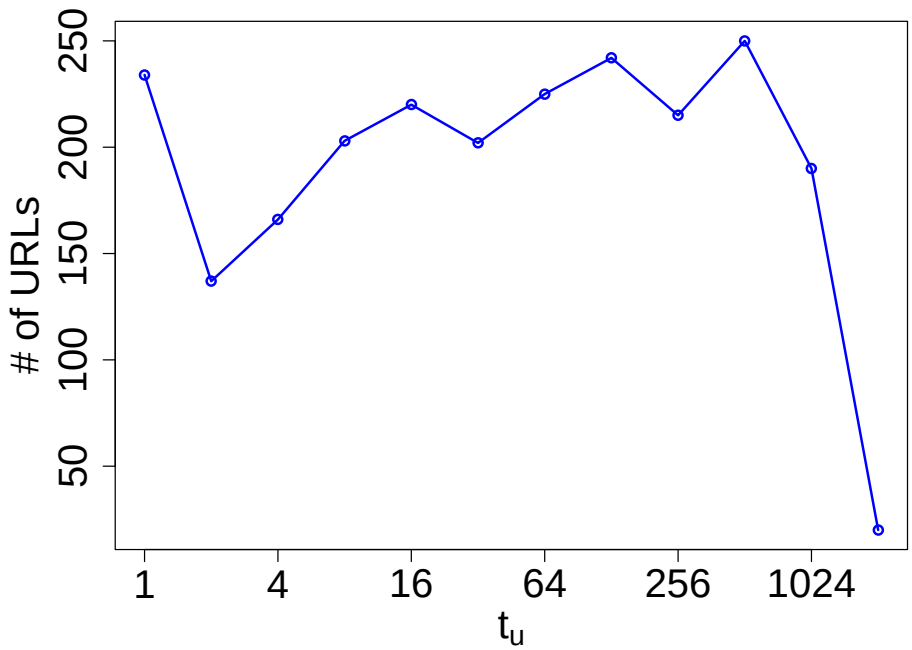
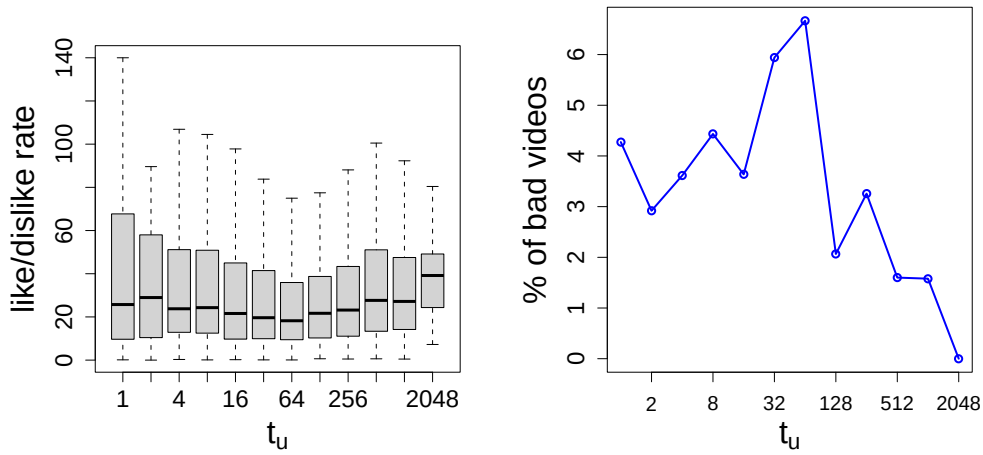


Figure 3.7: Distribution of t_u for YouTube videos



(a) Average likes/dislike rate

(b) Bad video rate

Figure 3.8: The quality of videos as a function of decay time t_u. *Like/dislike rate* is the number of likes divided by the number of dislikes. *Bad videos* are those with the number of dislikes greater than half of the number of likes. There are in total 83 out of 2304 "bad" videos by our definition.

the YouTube videos in our set into categories. We find that the most persistent videos are mostly music videos, again underscoring the increasing appearance of art-related topics in this class. On the other hand, many home-recorded video clips have very small value of t_u; as seen in Figure 3.0.1, content that fades away quickly might not have lasting value, but in general is more viral.

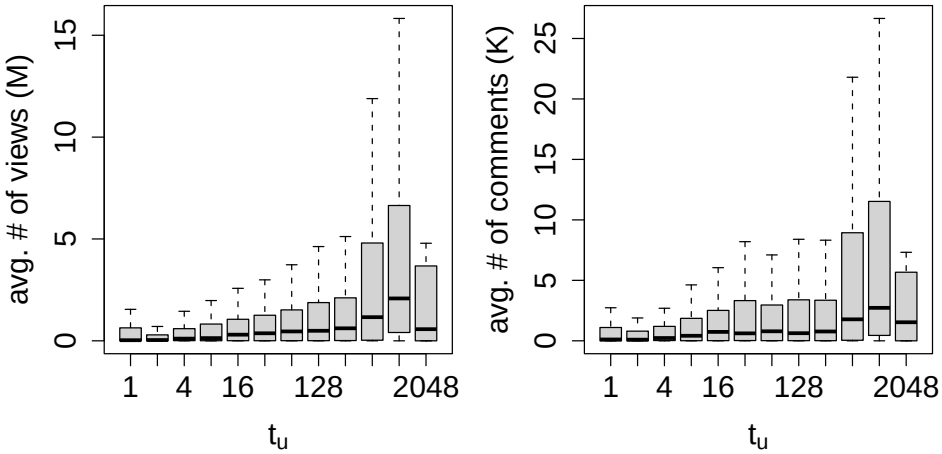


Figure 3.9: Average number of views and comments as a funtion of t_u

Finally, in Figure 3.0.4, we consider the number of views and comments on the videos in our set. We find an increase in views and comments particularly for very large values of t_u, in a way that is more extreme than the variation in the number of tweets from Figure 3.0.1, and that also forms an intriguing contrast with the trend in the number of RTs from that figure. Understanding how persistence translates into these secondary popularity measures such as view count is an interesting question.

3.1 Conclusion

In this chapter, we explore the relationship between content and the temporal dynamics of information in the context of TWitter. In particular, we find that using the textual features extracted from the content, we can predict the persistence of information with high accuracy. Second, we employ different text analysiss techniques to understand the nature of content that contributes to persistence. We examin and compare the fleeting and lasting content on three aspects, including the psychologuistic characteristics, trending words, and latent topics distribution. We find that the persistent information is more related to content with long-term value, such as positive emotions, life, family, and art. On the other hand, the rapidly-fading content contains mostly time-critical information that carries relatively more negative sentiments, demanding more cognitive effort, or is associated with quick action.

By restricting our scope of study to the time series of bit.ly URLs mentioned on Twitter, our findings can be limited to the types of social media and the dynamics of information they support. An interesting direction for future work could be studying the content and temporal patterns of information across systems. One possibility is to use the transcribed content of TV, radio, together with materials from online social madia. Also, since we only predicted the persistence of content for two extreme cases, it would be interesting to further investigate the connection between content and persistence as a continuous variable.

CHAPTER 4

NETWORK STRUCTURE AND THE SPREAD OF DISENGAGEMENT

There has been significant focus on the dynamics of propagation in social networks, especially, with the local and global structure involved[30, 12, 4, 25, 2, 35, 29, 16]. In many cases, such as the decision to become a member or use a product, the story does not end at adoption. Instead, the user may decide at any point to cease using the product, or to depart the community. It is not clear that a decision of this type, to "reverse" a prior socially-mediated decision to adopt, will follow the same dynamics as the original decision to adopt. In this paper, we study this question in the context of arrivals and departures within online social networks.

A natural place to look for models of arrivals and departures is the existing literature on the spread of infectious physical disease. These models often include a recovery component[30, 12], which is akin to a reversal of the decision to become infected. Typically, however, this component assumes that an infected user recovers based on properties of the immune system, without reference to any social process. In our case, we are motivated by the metaphor of a user at a party with friends. The user is more likely to attend upon discovering that some number of friends will also attend. If some or all of the friends then opt to depart, whether for a new party or to curl up with a good book, then the original user is much more likely to follow suit. Hence, we anticipate that social forces play a significant role in both arrivals and departures.

We begin to address this question with a basic study of the temporal correlation of arrivals and departures, and show that both processes introduce significant correlation among friends; in fact, we show that time intervals between

the departure of friends are more tightly distributed than the equivalent distribution of gaps between arrival of friends.

In this sense, we might consider arrival as the propagation of a "join" virus, and departure as simply the propagation of a new virus, in this case representing the decision to cease usage. However, this formulation is at odds with our conception of the underlying process. It is plausible that seeing one friend join a social network, then two, then three, might impel a user to join, as we see in prior work [2]. However, once a user has two hundred friends, will the departure of one, then two, then three friends have a qualitatively different impact on the user's likelihood to depart? Perhaps like the decision to join, the decision to depart depends more on the number of active friends than the number of inactive friends. Or perhaps departure is a fundamentally different decision that depends on an assessment of the pulse of the neighborhood, captured more accurately by the fraction of friends who remain active.

We study this question in the context of a large social network, and argue that in fact a hybrid of these models provides the most accurate characterization. While number of active neighbors is known to be a strong predictor of joining a group, for users with twenty or more friends, overall neighborhood activity, measured by the fraction of friends who remain active, is by far the best predictor of likelihood to depart. Surprisingly, this likelihood is linear in the fraction of active friends throughout almost its entire range, and the linear form is identical in both slope and intercept for several different buckets of neighborhood size. Raw counts of inactive friends have low predictive power, and raw counts of active friends, while stronger, remain weak compared to the overall fraction of active friends. On the other hand, for users with fewer than

twenty friends, the actual count of active friends remains a strong predictor of likelihood to depart. From these findings we reach a picture that users with few friends rely heavily on their continued presence, while users with more friends are pushed one unit closer to departure by each successive fraction of existing friends observed to depart.

From this emerging local picture of behavior, we may then ask how arrival and departure dynamics interact with the global structure of the graph. In particular, we seek to understand where departures happen in the graph. It is possible, for example, that departures tend to occur as high-status users in the core of the graph choose to depart in search of the next big thing. Alternately, it is possible that departure happens at the "fringes" of the graph, and then spreads inwards from there. We study this problem by computing the density and conductance of the subgraphs of active and inactive users through time, and comparing these results to thought experiments in which each node decides independently whether to remain active. These experiments allow us to conclude that a core of active nodes remains at much higher internal density than the set of inactive nodes. We also compare the densities observed against the expected density and conductance under a planted degree constraint model. The results suggest that although the inactive set of nodes densifies, its densification is not just a consequence of the degree distribution, but really a consequence of well-connected cluster of nodes from the fringes departing. We are led to believe that departures happen from the fringes and heavily influence their immediate neighborhoods, while an internal dense core of active nodes survives.

4.1 Data

In this chapter, we study the structure properties in relation to the arrival and departure of users, using a snapshot of the DBLP co-authorship graph and a well-known social network. The DBLP snapshot that we consider contains 1072718 nodes and 1839605 edges, for each author we store his/her co-authors and the year of the last publication. Furthermore for each author to author edge we also store the year of the first publication. in the rest of the paper we will refer to it as DBLP. The network we study contains millions of users and over a billion edges. For each user, we have the timestamp of signup and last login, and for each edge, we have the timestamp of edge creation. In the rest of the paper we will refer to this network as SN.

To study the pattern of user arrivals and departures, we first describe each user at each timestamp as either active or inactive, based on his most recent activity time. Given a snapshot of the SN network at time t, we consider a user *inactive* if his last login time is earlier than two months prior to t, and consider a user *active* otherwise. Given a snapshot of the DBLP network at time t, we consider a user *inactive* if he/she has not published any paper in the earlier than five year prior to t, and consider a user *active* otherwise. Note that our results do not depend on the time frame that we used. In fact, they hold for two quite different networks and time frames.

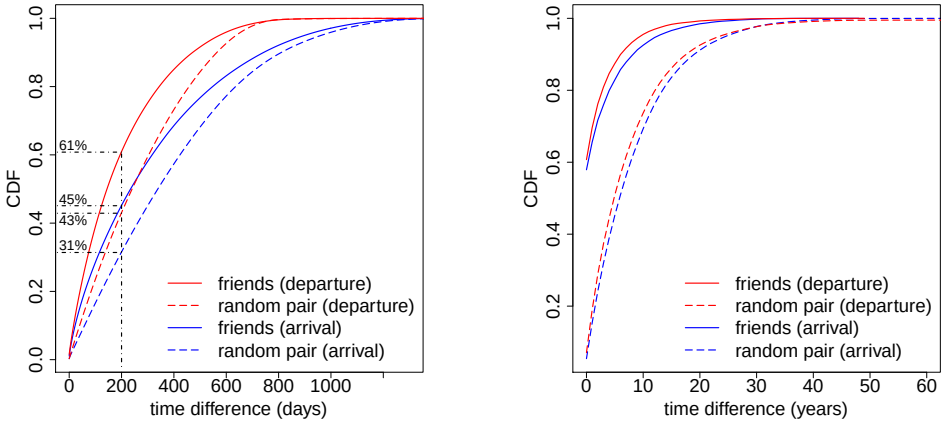
4.2 Arrival and departure correlation among friends

In this section, we study the temporal patterns of arrivals and departures, for both local dynamics and global trends. We wish to understand whether users typically arrive and/or depart together in social networks. However, we cannot directly compare gaps between arrivals and departures of friends, as networks are not stationary—consider for example the case of a network that grows very rapidly during a brief period, resulting in a flurry of temporally-proximate arrivals, leading to a mistaken conclusion that arrivals tend to be tightly clustered in time.

We must therefore normalize in some way against global rates of arrival and departure, which we do by the following technique. Given a snapshot of the network at time t, we consider two samples of user-pairs, one in which the pair of users are friends, and another in which the pair of users is chosen uniformly from all possible pairs[1]. We then consider the distribution of the gap in arrival time between pairs in the two cases. Differences in these distributions will then highlight temporal correlation of arrivals of friends compared to strangers.

To study departures, we adopt the same technique. We consider only inactive users, and generate again a set of pairs of friends, and another set of pairs chosen uniformly at random. In this section, we fix t then define the last login time of inactive users as their departure time. We pick 1M pairs for each of these four sample groups, and shows the Cumulative Distribution Function (CDF) for these distributions in Figure 4.1.

[1]Note that although technically, it is possible for a random pair to be a pair of friends, given the service policy that each user has a rather small upperbound for the number of friends, the chance of a random pair being friends is negligible.



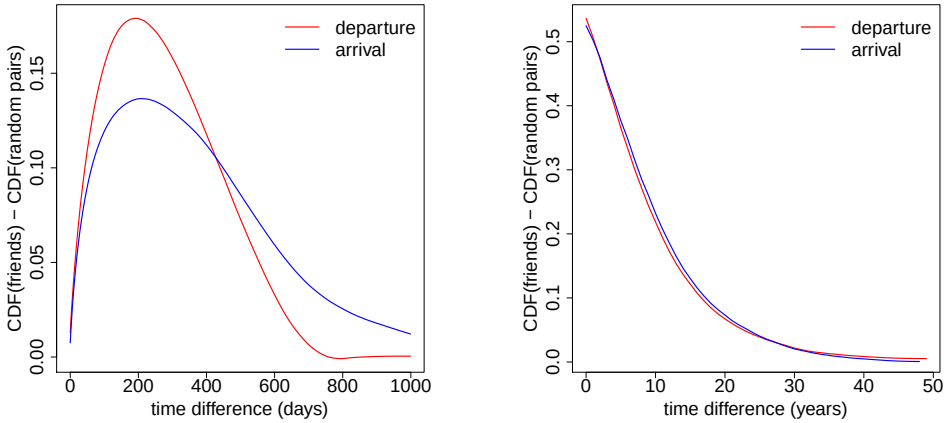
(a) SN (b) Dblp

Figure 4.1: The CDF curve for the difference in arrival and departure time between friends and random pairs of users.

The CDF for both arrivals and departures of friends lies significantly above the CDF for random pairs, indicating that friends both arrive and depart together, in comparison to the control group of random pairs. As the figure shows, in the case of SN 43% of random pairs depart within 200 days of one another, while 61% of friends depart within the same period, a large relative increase of 41%. We find similar pattern in the time interval of arrival - only 31% of random pairs arrive within 200 days, but 45% of friends arrive within the same period. This observation is even more evident in Dblp where the lines are clearly apart.

To quantify the differences, we plot in Figure 4.2 the distribution of absolute difference in the CDF values at each time, for arrivals and departures(at least in SN). The correlation of departures is seen to be stronger than the correlation of arrivals, although the two gaps peak around roughly the same value.

We also consider the eventual set of friends acquired by a user at the snapshot time t, and ask whether those friends join before or after the user. First, in contrast to the observations in previous research[2], the number of friends

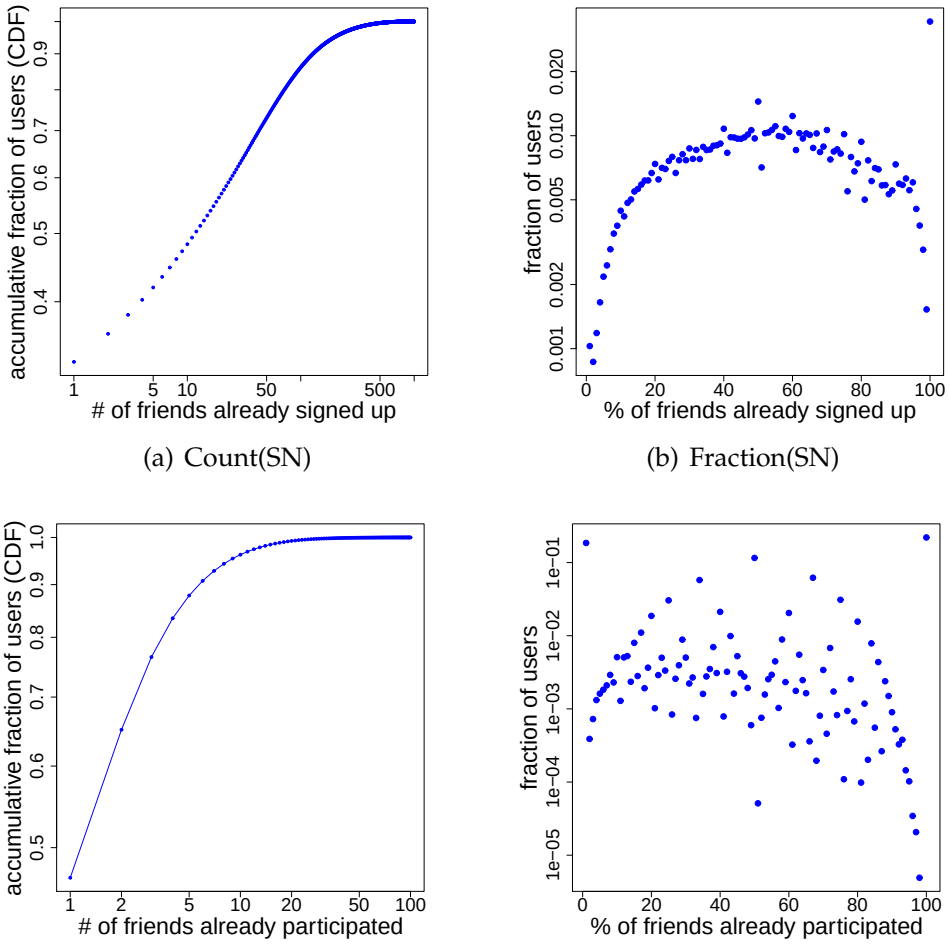


(a) SN (b) Dblp

Figure 4.2: Gap between CDF curves.

who already signed up seem to have a "diminishing effect" only on the case of the co-authorship graph and not of SN. In Figure 4.3(a), we see that in SN as the number of adopted friends increases, the probability of a user signup increases, but rather linearly, throughout almost the entire range x-axis. On the other hand, the expected fraction of friend pairs joining before the user will always be 0.5, as the friend network is undirected and each edge contributes one pair in which a joins before b, and the opposite pair in which the reverse happens. Thus, for regular graphs (of constant degree), the mean of the distribution of fraction of friends already signed up will be at 0.5. The results are shown in Figure 4.3. True social networks are of course non-regular, and while the distribution of plot (Figure 4.3(b)) appears largely symmetrical, there are some outliers. In particular, both in SN there are more than 20 times as many users users for whom, at signup, 100% of their friends have already signed up, compared to users for whom 0% of their eventual friends have already signed up. These can be explained by many low-degree nodes who are attracted to the network by a friendship invitation but never really engage in the network afterwards. In Dblp the situation is a bit different, this is probably explainable by the fact that

several papers are written by community of student that after the master or the PhD do not publish any more. Overall, we think that there is certain network effect towards the arrival of users, however, this effect is quite weak in the formation of the network, and may not be enough to actively engage users after they sign up.



(a) Count(SN) (b) Fraction(SN)

(c) Count(Dblp) (d) Fraction(Dblp)

Figure 4.3: Count and Fraction of friends already signed-up when user signs up.

Our conclusion from this set of graphs is that friends tend to arrive and depart together, but departures are more tightly clustered than arrivals. This observation relates only to individual friends, while we expect that the effects are

better understood in terms of the entire "neighborhood" of friends in the graph.

Arrivals are difficult to study in this model, as the nature of the neighborhood is largely unknown to a new user until after the decision to join. Other authors consider the related problem of joining a particular group as the adoption of an innovation within the substrate of an existing social network. For example, [2] consider joining an interest group within the LiveJournal network. We may therefore employ our longitudinal data to flesh out the picture given by this earlier, by looking more closely at the impact of the neighborhood structure on departure. Subsequently, we then relate the results back to known literature on adoption of innovations around group membership, to compare what is known about arrivals and departures.

4.3 Effect of local neighborhoods

As shown in previous section (Figure 4.1), friends are more likely than strangers to have logged in within around half year of one another. This dramatic difference causes us next to look beyond single-edge correlations to properties of the entire neighborhood of a node.

4.3.1 Dependence on local properties

The correlation in the timing of last login among friends suggests the effect of friends' inactivity on the decrease of activities on an individual. To better understand how a user's departure is influenced by his local community, in this

section, we look at the probability of a user's departure in relation to the following four properties related to the user's neighborhood.

- number of active friends;
- fraction of active friends;
- number of inactive friends;
- number of inactive friends who left in the past 6 months;

To study how the probability of a user becoming inactive depends on the number of friends who are active, we use a similar method as in [2]: we first take two snapshots (t_0, t_1) of the network, three months apart in SN and three years apart in Dblp; we then find all pairs (u, k) such that u is active at the time of first snapshot t_0, and has k friends who are also active at t_0; $p(k)$ is calculated as the fraction of such pairs (u, k) for a given k such that u had left the network at the time of second snapshot t_1. In other words, $p(k)$ is the fraction of active users who left the network in the next three months, given that k friends are active at the first snapshot time. Figure 4.4(a) and Figure 4.4(c) shows the curves of $p(k)$ at three different sample points of t_0. In a similar way, we can fix f as the fraction of friends who are active at time t_0, and calculate the probability $p(f)$ of an active user leaving the network as function of f (see Figure 4.4(e)). Note that for this figure and all the following figures involving the fraction of active/inactive friends, we exclude all the nodes with no friends in SN, which are around 10% of all active users as of 2011/1/1, and among those users, 35% of them left within three months.

Not surprisingly, Figure 4.4(a), Figure 4.4(c) and Figure 4.4(e) show that as more and more friends stay active, a user is less and less likely to be inactive. The curve of $p(k)$ (see Figure 4.4(a)) also matches very well with what has been seen in other domains [2], exhibiting the "diminishing returns" property - as the

number of active friends k increases, the probability of departure continues to decline, but more and more slowly, eventually converging to a constant for very large values of k. This observation indicates that the marginal gain of having each additional active friend is quite significant for users with a small number of active friends, but rather negligible when a user has already more than 50 active friends. In contrast, in Figure 4.4(e), we do not see such a "diminishing returns" trend, but a steeper, and almost constant rate of decrease in the probability of departure throughout the course when the fraction of active friends increases. This is an interesting observation that has not been previously seen (specifically in various positive influence studies).

To see how the inactivity of the neighborhood influences the departure of a user, we also plot the probability of departure as a function of number of inactive friends, in Figure 4.4(b) and Figure 4.4(d). The curves in Figure 4.4(b) and Figure 4.4(d) show an interesting trend of decreasing slope through time: while the probability of a user departing increases with the growth in the number of inactive friends in initially, it becomes more and more insensitive to the value of k in the later curves. This phenomenon is quite intriguing to us: if the departure of friends do have a clear effect on the departure of the user, as shown in the earlier curves, why is this effect diminished so much in the latest years? To answer this question, we note that we are counting the number of inactive friends as prior to the time of each snapshot, but many of them could have been inactive for a long time thus could hardly influence the user's experience in the network at the snapshot time. Figure 4.4(f) confirms this idea, showing that the curves we see in Figure 4.4(b) are somewhat misleading - in general, the probability of user's departure constantly grows with the number of friends r who *recently* became inactive (when r is not too small).

4.3.2 Interaction between local properties

The results of the previous section provide qualitative evidence that an individual's probability of departure is related to the activeness of his neighborhood. Intuitively, as more and more friends leave a social network, a user will start to feel desolated and will be more likely to leave as well. Our previous results also suggest that the fraction of friends who are active/inactive contributes to the overall atmosphere of the neighborhood, and that this matters more than the raw number of active/inactive friends. However, does that apply to all users? Do the highly connected users act differently than the more marginally connected ones? Can people really notice and act on the degeneration of their neighborhood, or will they stay as long as there are a few active friends, in spite of a large fraction of their friends having left? To address these issues, we compute the probability of user's departure in SN in relation to the interaction between local properties. Specifically, in Figure 4.5(a), we divide users into three groups based on their degrees, and plot the probability of departure as a function of the number/fraction of active friends, for each group separately. We note that for users with different levels of connectivity in the network, the curves of $p(f)$ (Figure 4.5(a)) are qualitatively identical. This result demonstrates again that the fraction of friends who are active has a stronger effect on the probability of an individual's departure, regardless of the size of the user's neighborhood.

In addition, we aggregate users by the fraction of active/inactive friends, and look at how the probability of departure depends on the number of active/inactive friends for each group (see Figure 4.5). There are two things we note from Figure 4.5: First, for users with different fractions of inactive friends, there is a big gap between their probabilities of departure - for example, com-

pared to users with less than 10% friends left (blue line in Figure 4.5(c)), users who have more than 50% friends left (red line in Figure 4.5(c)) are 10 times more likely to leave as well. Second, once the user is in an inactive part of the neighborhood, the raw count of inactive friends has little effect in determining the probability of the user's departure (green line in Figure 4.5(b)). Note that the blue line in Figure 4.5(b) is very noisy because there are very few people in a highly obsolete neighborhood but still with a substantial amount of active friends. We still plot it just to be symmetric with Figure 4.5(c).

4.3.3 Predict the departure of user

Given a strong correlation between the probability of a user becoming inactive and the inactivity of his friends, the next question is, can we actually predict individuals's departures based on local properties? In this section, we explore the problem of modeling the departure of users using simple linear regression models and decision tree classifiers. In particular in this subsection we will focus exclusively on SN because we have a richer set of feature available.

To start, we formalize our problem as a binary classification task in which class 1 is defined as consisting of those users who were active as of Jan 1st, 2011 (t_0) and departed within two months after t_0, and class 0 is defined as consisting of those who stayed active for two months after t_0. We then randomly sample 500K positive examples and negative examples separately, from all the users who were active at t. Note that among all examples, there are 90% negative and only 10% positive examples; our sampling scheme provides a more balanced distribution of examples of both classes.

Table 4.1: Predict user departure with decision tree

Feature	*Accuracy*	*F1 pos*	*ROC area*
Neighborhood	0.694	0.694	0.755
Activity	0.730	0.735	0.801
All	0.755	0.761	0.833

We extract two sets of local features for each user:

- Neighborhood features. The local structural properties of the user's direct neighborhood, including the number of friends who already departed, the number of friends who are active, the number of friends who departed recently (six months prior to t_0), and the fraction of friends who departed recently.
- Activity features. The properties reflecting user's participation to activities in the network, including the number of contents he received, the number of contents he sent, and the number of status updates.

To predict the departure of users, we train a simple decision tree (REPTree) classifier on our examples. Table 4.1 gives the performance of the classifier with different sets of features under 10-fold cross validation.

Table 4.1 shows that relying on only local features of individuals, the simple decision tree classifier can predict the departure of user with high accuracy (75% with all features, as compared to 50% for always predicting one class). This result demonstrates a strong connection between user's local properties and the propensity of departure. Moreover, comparing across 3 sets of features, we see that although the activity features are more powerful, neighborhood features can also provide rather accurate insights on the departure of users.

Table 4.2: Summary of logistic regression model on $p_{departure}$

Feature	*Coefficient*	*p value*
1/(number of active friends)	0.0579	$< 2e-16$ ***
fraction of active friends	-1.5340	$< 2e-16$ ***
number of friends left recently	0.0067	$< 2e-16$ ***
fraction of friends left recently	-0.0020	0.0737 .
number of contents received	-0.0012	$< 2e-16$ ***
number of contents sent	0.0000	$1.28e-06$ ***
number of status updates	-0.0017	$< 2e-16$ ***

The decision tree classifier demonstrates that local properties provide strong evidence to predict the departure of user. It also suggests that the activity features are more effective at predicting user departure. However, the decision trees we trained contain over a thousand nodes and thus is too complicated to illustrate how the local properties influence the probability of user departure. To better understand the effect of different features, we also fit the data with a logistic regression model that predicts the probability of departure. The model is constructed on 7 independent variable covering both neighborhood features and activity features, the results of the model is summarized in Table 4.2

We evaluate the logistic regression model using 10-fold cross validation as well, and it only slightly under-performs the decision tree classifier, with the ROC area as 0.774.

The results of the regression model nicely confirm the descriptive results we showed previously, and quantify the effect of different variables on the departure probability. In particular, from Table 4.2, we see that the existence of active friends and continued activities, both decrease a user's tendency to depart while

the number of friend who departed recently contribute to this tendency. We also notice that although most of the activity variables and the neighborhood variables have very high significance (very low p-value) in the estimated model, each unit of the fraction of active friends has the most substantial effect on the probability of user departure.

4.4 Structural trends in network topology

We explore the overall structural changes that occur in the network as a result of the departure of several users, as well as the steady arrival of new users. Topological changes have been studied in the context of new nodes arriving but here we pay specific attention to how the global structure changes as a result of the departure or decline of user activities based on their local neighborhoods.

To get a sense of the how the structure of the network evolves over time, we first study the distribution of edges among active and inactive nodes. Specially, we look at the edges between active nodes (Figure 4.6(a) and Figure 4.6(d)), edges between inactive nodes (Figure 4.6(b) and Figure 4.6(e)), and the edges across active and inactive nodes (Figure 4.6(c) and Figure 4.6(f)), and plot the ratio between the actual number of edges over the expected value over time.

Here, the expected number of edges is computed based on the total number of edges, $|E|$, in the network and the number of nodes in each of the active and inactive sets. The expected number of edges of any type is the expected number of edges if the the total $|E|$ edges are placed between randomly chosen pairs of nodes.

To understand the overall structure among the sets of active and inactive nodes, we study the density and conductance of these two sub-networks in the rest of this section.

Figure 4.7 and plots the overall density of the active (4.7(a) and 4.7(c)) and inactive (4.7(b) and 4.7(d)) set of nodes, as a function of time. For comparison, we also plot the *expected* densities of the respective sets, as determined by the number of active and inactive nodes and edges and the degree distributions.

We here define density of a set of nodes(or average induced degree) as the number of edges between them divided by the number of nodes; i.e. for a set of nodes S, $density(S) = \frac{|E(S,S)|}{|S|}$ (here $E(S,S)$ contains all edges (u,v) such that $u,v \in S$). Therefore, the density of set S is half of the average induced degree of the set of nodes in S. In order to compare the the density we observe for the set of active nodes and the set of inactive nodes, we define an *expected* density for each of these components. The expected density of the inactive set of nodes could be computed simply as the density of the entire graph times the fraction of inactive nodes.

However, we even use a stronger baseline to see if the trends we observe are a result of a trend more than just that of degrees. Therefore, we compute expected density subject to the overall degree constraints on active and inactive nodes as follows.

Consider each edge as occupying two slots (end points), each slot being in either S_a (the active set of nodes), or S_i (the inactive set of nodes); therefore $S_a \cup S_i = V(G)$. Let the fraction of all these slots that are in S_i be P_i (which is the number of edges going across the active and inactive component plus

twice the number of edges in the inactive component); therefore the number of such slots occupied in S_a is $P_a = (1 - P_i)$. Suppose that all the $|E|$ edges were randomly placed in two slots each, with probabilities determined such that in expectation we respect P_i and P_a, then we consider the induced density of this process as the expected density (for respective components). Notice that this is a more stringent baseline for our comparison. Therefore, an edge is contained in the inactive component with probability P_i^2 and so the expected density of the inactive set is $(|E|P_i^2)/|S_i|$. Similarly the expected density of the active component can be computed.

The plots on these densities in Figure 4.7 shows that the density of the active set $density(S_a)$ increases rapidly with increase in time. Comparing this with the plot on distribution of edges in Figures 4.6, we see that as the number of inactive nodes starts increasing, the number of edges in the active set, and correspondingly its density, becomes much higher than the density of the inactive set of nodes. We notice that the density of the active set is only marginally higher than its expected density. However, for inactive nodes, the density is significantly higher than the expected density, even conditioned on the degree distribution. This is only explainable by the fact that the decision to depart is correlated across edges, as supported by our local analysis; the nodes that are departing are still probably at the periphery of the network (since the inactive set has much lower density than the active set), but these inactive nodes continue to be internally well-connected because of a higher-than-expected density. This strengthens the evidence from previous sections that a node's likelihood to become inactive is influenced by the extent of neighboring inactivity.

After learning about the connectedness of the active/inactive subnetwork

separately, we now switch our gear to look at the connection of each subnetwork to the rest of social graph. We use conductance to measure the amount of possible connections between different sets of nodes in a network.

Conductance of a set of nodes S, $\phi(s)$ is measured as $\phi(S) = \frac{|E(S,V(G)\S)|}{|E(S)|}$. Here $E(S, V(G)\S)$ contains all edges (u, v) such that $u \in S, v \notin S$, and $|E(S)| = 2|E(S,S)| + |E(S, V(G)\S)|$. So notice that conductance is always less than 1, and any set with more than half its edges going across to the complement set has a conductance of more than $\frac{1}{3}$. We again measure the conductance of sets S_a and S_i through time and compare with their expected conductances (see Figure 4.8). The computation of expected conductance is also performed in a similar manner to as described previously for expected density.

We see a similar trend in conductance in Figure 4.8 as seen for densities. The conductance of the active set of nodes S_a, $\phi(S_a)$ remains somewhat less than the conductance expected for this set. This suggests that there are somewhat fewer edges going across from S_a to the inactive set S_i and far more edges within S_a itself, than would be expected. The conductance plots for the set of inactive nodes however is again more contrasting. $\phi(S_i)$ remains far lower than the expected conductance. Nodes that are becoming inactive continue to have many more edges within, than one would expect. This clearly suggests that the inactive set of nodes are influencing neighbors to inactivity. Yet again, the absolute conductance value still suggests that nodes at the periphery of the network are more susceptible to becoming inactive.

The takeaway from these plots are two fold. Firstly, of course, these trends corroborate our findings from the previous sections suggesting that there is a strong influence of inactivity on its neighborhood and that nodes are much

more likely to depart from the network if they are surrounded by inactive nodes. However, these plots on global measures such as density and conductance also suggest a picture of the evolving network. With the active set's density being much higher than the inactive, and the inactive set showing higher than expected density and lower than average conductance, we are led to believe that nodes in the *core* of the network are much more likely to survive, while nodes at the periphery are more susceptible to departure.

4.5 Conclusion

In this chapter, we have studied the dynamics of user departure from online social networks, from the perspectives of local and global network structure. We considered the influence of local neighborhoods on the behavior of nodes as well as studied global changes in the network topology. At the local level, we studied individuals and the dynamics in their local neighborhood, measured the probability of user arrival and departure in relation to the activity of their friends. Our findings are three fold: first, there is a strong clustered effect in the timing of departure among friends while this is not as visible in arrivals; second, although both numbers and fractions of neighborhood (in)activity are correlated to the probability of the individual's departure, the fraction of inactive friends has arguably the strongest effect on the departure probability, providing an interesting complement to literature on arrivals which shows number of active friends as the most predictive of these measures; third, once a significant fraction of friends depart, the overall connectivity of individuals in the entire network does not save the user from leaving the network. At the global level, we looked at the trend of network topological properties over the past few

years, showing that as the network evolves, users at the peripheral region of the network are more likely to depart in groups; yet an internal core of the network survives and densifies over time.

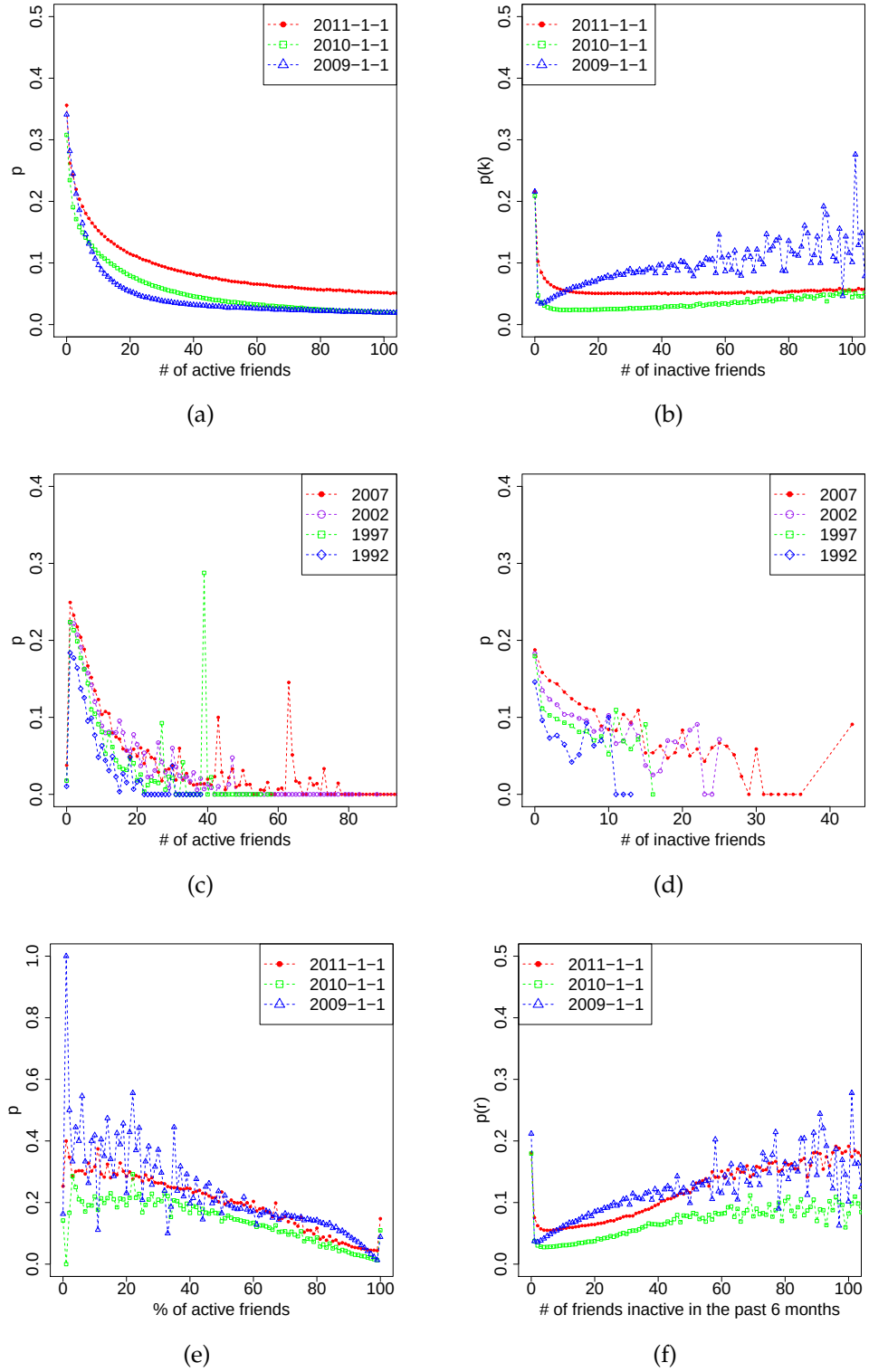
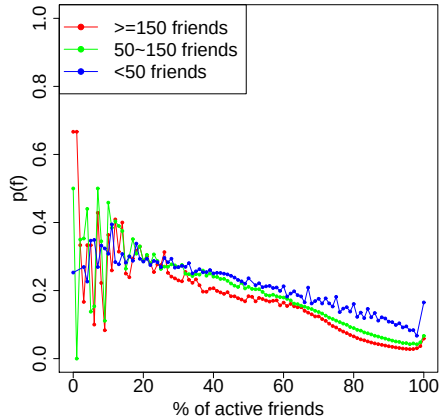
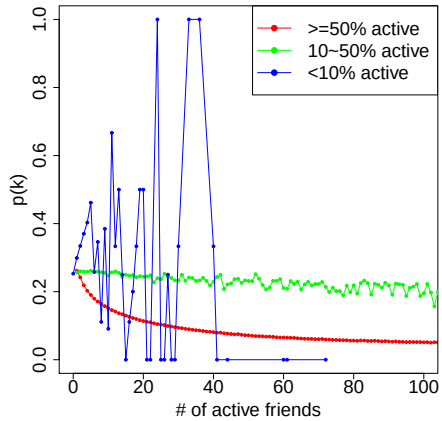


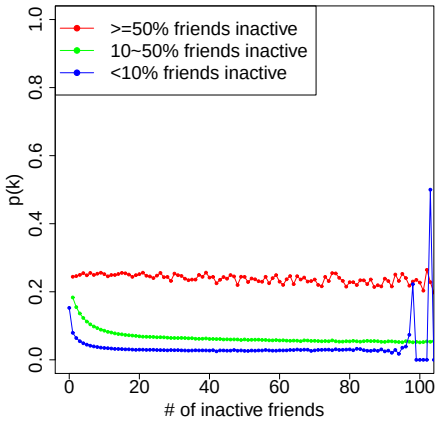
Figure 4.4: Probability of departure as function of different local properties. Where (a) is p as f(active friend count) in SN, (b) is p as f(inactive friend count) in SN, (c) is p as f(active friend count) in Dblp, (d) is p as f(inactive friend count) in Dblp, (e) is p as f(active friend fraction) in SN and (f) is p as f(inactive friends who left in the past 6 months) in SN.



(a)

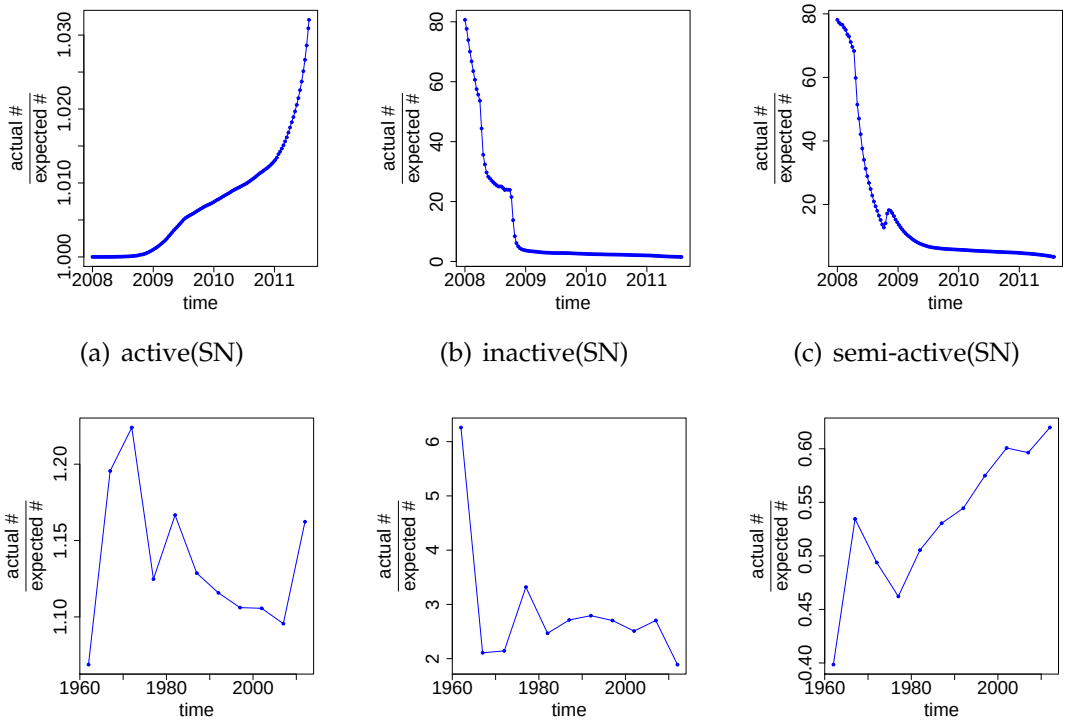


(b)



(c)

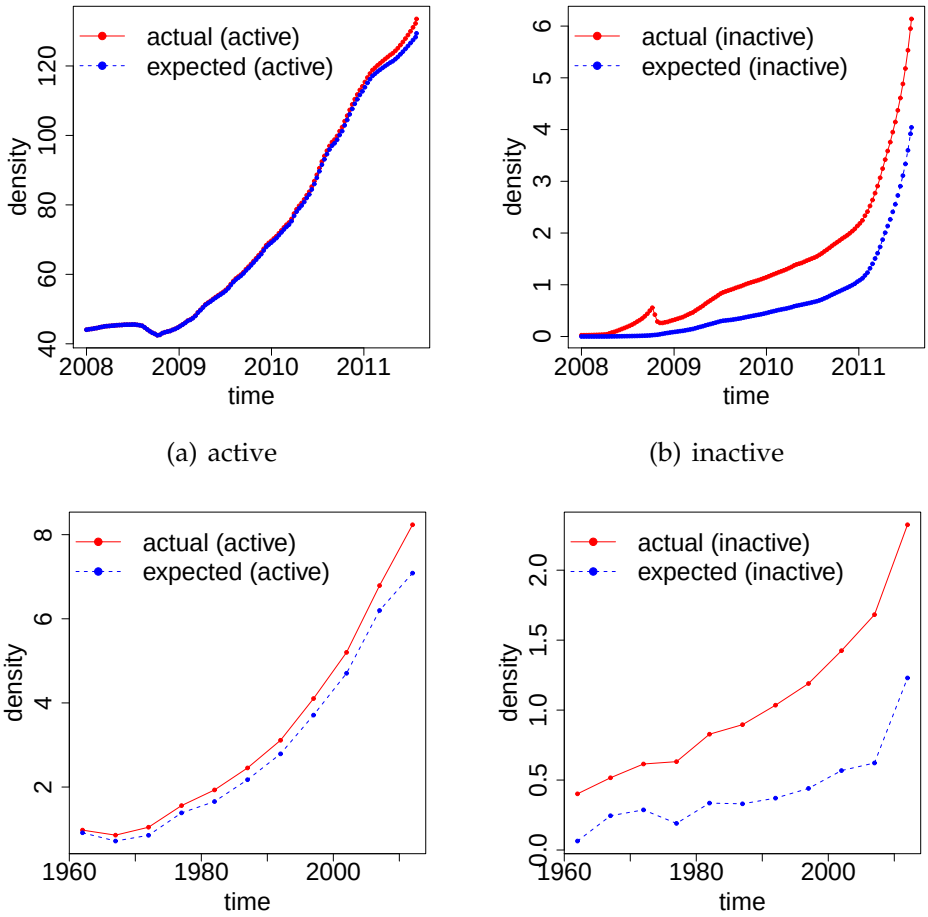
Figure 4.5: Probability of departure as function of local properties, at different levels of active/inactive friend fraction and friend count (snapshot taken at time t =2011/1/1). in SN



(a) active(SN) (b) inactive(SN) (c) semi-active(SN)

(d) active(Dblp) (e) inactive(Dblp) (f) semi-active(Dblp)

Figure 4.6: Distribution of edges, indicated by the ratio of actual number of edges over the expected number of edges (formed in random process).



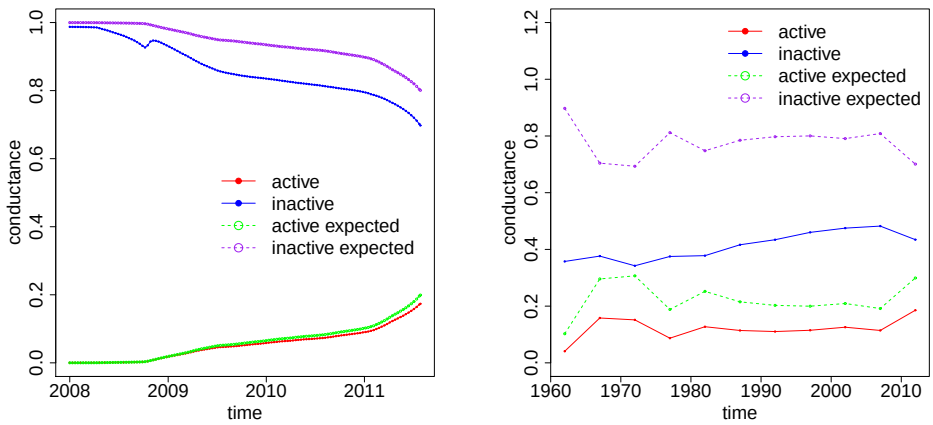
(a) active

(b) inactive

(c) active

(d) inactive

Figure 4.7: Density of the active and inactive subnetworks



(a) SN

(b) Dblp

Figure 4.8: Conductance of the active and inactive sets

CHAPTER 5

CONCLUSION

We study three major aspects of information diffusion process: influencers, content, and network structure. In the future, I would like to study the impact of information in both on-line and off-line environement, and leverage my work to foster the effective flow of information in the society.

APPENDIX A

CHAPTER 1 OF APPENDIX

Appendix chapter 1 text goes here

BIBLIOGRAPHY

[1] Sinan Aral and Dylan Walker. Identifying Influential and Susceptible Members of Social Networks. *Science*, June 2012.

[2] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution.

[3] Eytan Bakshy, Jake. M. Hofman, A. Mason, Winter, and Duncan J. Watts. Identifying 'influencers' on twitter. In *Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*, Hong Kong, 2011. ACM.

[4] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[5] F. M. Bass. A new product growth for model consumer durables. *Management Science*, 1969.

[6] Jonah Berger and Katherine Milkman. Social transmission, emotion, and the virality of online content. *Wharton Research Paper*, 2010.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. 3:993–1022, 2003.

[8] Damon Centola and Michael Macy. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 2007.

[9] Meeyoung Cha, Hamed Haddadi, Fabrłcio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

[10] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[11] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, October 2008.

[12] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 2005.

[13] V. M. Eguiluz and K. Klemm. Epidemic threshold in structured scale-free networks. *Physical Review Letters*, 89, 2002.

[14] Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little Brown, New York, 2000.

[15] Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. 2002.

[16] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.

[17] Lars Kai Hansen, Adam Arvidsson, Finn rup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news - affect and virality in twitter. *CoRR*, abs/1101.0510, 2011.

[18] Elihu Katz. The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly*, 21(1):61–78, 1957.

[19] Elihu Katz and Paul Felix Lazarsfeld. *Personal influence; the part played by people in the flow of mass communications*. Free Press, Glencoe, Ill.,, 1955.

[20] Ed Keller and Jon Berry. *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy*. Free Press, New York, NY, 2003.

[21] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[22] Jon Kleinberg. Temporal dynamics of on-line information streams. In *Data Stream Managemnt: Processing High-speed Data*. Springer, 2004.

[23] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, pages 591–600. ACM, 2010.

[24] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. In *Proceedings of the 7th ACM conference on Electronic commerce*, EC '06, pages 228–237, New York, NY, USA, 2006. ACM.

[25] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1, May 2007.

[26] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[27] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations.

[28] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *7th SIAM International Conference on Data Mining (SDM)*, 4 2007.

[29] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[30] M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 2002.

[31] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, April 2001.

[32] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Lawrence Erlbaum Associates, 2001.

[33] F. Provost. Machine learning from imbalanced data sets 101. *Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets*, 2000.

[34] Everett M. Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 5th edition, August 2003.

[35] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[36] M.J. Salganik, P.S. Dodds, and D.J. Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856, 2006.

[37] Herbert A. Simon. Designing organizations for an information rich world. In Martin Greenberger, editor, *Computers, communications, and the public interest*, pages 37–72. Baltimore, 1971.

[38] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! modeling contagion through facebook news feed. *Proc. ICWSM*, 9, 2009.

[39] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, August 2010.

[40] J. B. Walther, C. T. Carr, S. S. W. Choi, D. C. DeAndrea, J. Kim, S. T. Tong, and B. Van Der Heide. Interaction of interpersonal, peer, and media influence sources online. In Zizi Papacharissi, editor, *A Networked Self: Identity, Community, and Culture on Social Network Sites*, pages 17–38. Routledge, 2010.

[41] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34:441–458, 2007.

[42] J. Weng, E. P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[43] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM.

[44] Shaomei Wu, Chenhao Tan, Jon Kleinberg, and Michael Macy. Does bad news go away faster? 2011.

[45] J. Yang and S. Counts. Predicting the speed, scale, and range of information

diffusion in Twitter. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, May 2010.

[46] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *ACM International Conference on Web Search and Data Minig (WSDM)*. Stanford InfoLab, 2011.

[47] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *ACM International Conference on Web Search and Data Minig (WSDM)*. Stanford InfoLab, 2011.