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

THE DYNAMICS OF INFORMATION DIFFUSION ON ON-LINE SOCIAL NETWORKS

Shaomei Wu, Ph.D.

Cornell University 2012

Your abstract goes here. Make sure it sits inside the brackets. If not, your biosketch page may not be roman numeral iii, as required by the graduate school.

BIOGRAPHICAL SKETCH

Your biosketch goes here. Make sure it sits inside the brackets.

This document is dedicated to all Cornell graduate students.

ACKNOWLEDGEMENTS

Your acknowledgements go here. Make sure it sits inside the brackets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Part I

Introduction and Background

CHAPTER 1
INTRODUCTION

Understanding how information spreads in the society has attracted a growing interest from both practitioners and scholars. It is an essential element for many interesting problems, such as the diffusion of innovation, formation of public opinion, product adoption, and viral marketing.

Historically, one of the biggest challenges in the study of information diffusion is data collection. Given the difficulty of observing and tracing the diffusion process over all possible channels, most empirical studies in this field have been conducted manually by sociologists and communication researchers, through in-person interviews or standardized surveys of population samples. Therefore, these studies are largely confined by the scale and the data accuracy. As a result, existing theoretical models in the literature have relied on many assumptions about the underlying diffusion mechanism instead of empirical evidence. Being widely applied, these models are, however, hard to be verified or rebutted empirically on a large scale.

In recent years, the abundance of digital records of online interactions has provided us 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 (i.e. Twitter, Orkut), this thesis addresses the following three questions: "who influences whom?", "how do different types of information spread?", and "how does the network structure impact the diffusion process?"

Viewing the diffusion of information as an organic process, we study each stage in its lifecycle, from production, dissemination, to comsuption (deceased). Each part is motivated by social science problems and theories, and aim at contribute deeper insights about how information flow in the society and its impact.

our work mainly contributed to the aspectes in the lifecycle of information diffusion that are overlooked by previous studies:

1. Production: people's external influence, especially, the role of mass media
2. Dissemination: temporal dynamics and the role of content
3. End-of-life: how does disengagement happen?

In the end, we will examine a diffusion process that closely tight the online and offline world together - Arab spring. Contributions:

1. Trace the evolution, and unveal the dynamicss of social movement using digital logs;
2. Understand the role and impact of social media in societal changes;

1.1 The Influencer Problem

Can we estimate person influence and pick out the influencers? Think of influence in context:

1. general influence: mass media
2. domain influence: masspersonal (celebrities, bloggers, organizations)

3. personal influence: tie strength, content, language, timing (accidental influencers)

1.1.1 Life of information in Social Media

How long should we expect a piece of information to live in the social media space [?]? Who are involved in each stage (production, circulation, and consumption) of lifecycle of information?

1.1.2 Modeling personal influence

1.1.3 The intersection between content and people

1.2 The role of content in diffusion process

Section 2 text.

1.2.1 Subsection heading goes here

Subsection 1 text

Subsubsection 1 heading goes here

Subsubsection 1 text

Subsubsection 2 heading goes here

Subsubsection 2 text

1.3 Diffusion and network structure

Section 3 text. The dielectric constant at the air-metal interface determines the resonance shift as absorption or capture occurs.

$$k_1 = \frac{\omega}{c(1/\varepsilon_m + 1/\varepsilon_i)^{1/2}} = k_2 = \frac{\omega sin(\theta)\varepsilon_{air}^{1/2}}{c} \tag{1.1}$$

where ω is the frequency of the plasmon, c is the speed of light, ε_m is the dielectric constant of the metal, ε_i is the dielectric constant of neighboring insulator, and ε_{air} is the dielectric constant of air.

CHAPTER 2

BACKGROUND

A variety of methods have been applied by scholars to study the diffusion of information. From the theoretic perspective, sociologists and economists developed agent-based modeling (ABM) to explore the dynamics of diffusion in different networks and interactions [Centola and Macy 2007, Watts and Dodds 2007]. Computer scientists designed the algorithm to maximize the extent of diffusion by seeding the diffusion with specially-picked individuals [Liben-Nowell et al 2008].

On the other hand, modeling and predicting the propensity of real world diffusion going viral through WOM process is of central interest in recent empirical studies. Built on top of the cascade model, Gruhl et al [2004] tracked the diffusion of topics in blogspace to estimate the transmission probability with an expectation-maximization(EM)-like algorithm. Leskovec et al [2007] studied interpersonal recommendations on an e-commerce site to infer the adoption probability based on the category of products and invitation history. Cha et al [2008] studied the viral process of photos being marked as favorite on the Flickr social network. By applying the SIR model with infinite recovery time, they estimated the reproduction number with the mean degree of nodes at each step of contagion. Backstrom et al. [2006] modeled the probability of a user joining a community and the growth of communities on LiveJournal using decision trees with network structure features.

2.1 Social theories

2.2 Diffusion models

2.3 Temporal analysis

2.3.1 Temporal pattern detection

2.3.2 Trend detection

2.4 Lingusitic analysis

2.5 On-line social network as the lab for studying information diffusion

2.5.1 Twitter

In my work, I frequently study the diffusion of information in the context of Twitter. As a highly popular micro-blogging service, Twitter provides a natural environment for the study of diffusion process. Unlike other online social networks (e.g. Facebook), Twitter is expressly devoted to disseminating information in that users subscribe to the information broadcasted by other users; thus the network of potential adoption can be reconstructed by crawling the corresponding ″follower graph″. In addition, because of the need of users to share web content, with the restriction of maximal tweet length as 140 characters, URL shortening services (e.g., bit.ly, TinyURL, etc) are used widely, and effectively tag numerous pieces of information with unique and easily-identifiable tokens. Our study takes advantage of these features in following aspects:

(a) We are able to observe essentially everything that is spreading on Twitter. Thus although we have only one type of social media, we have a very high level of resolution and coverage regarding what is being diffused.

(b) Although it may be non-representative in some respects, Twitter is very representative in at least one respect–namely it includes essentially all actors of any consequence in the information diffusion process in the society: media outlets and formal organizations of all sizes, bloggers, and public figures like celebrities, as well as tens of millions of ordinary individuals. In this sense, it really is a complete sample of one (admittedly narrow) slice through the diffusion landscape.

(c) Because Twitter users themselves classify other users by including them on lists, Twitter effectively provides a ready-made, crowd-sourced classification scheme of users. Thus even though we do have content information for many of the Tweets we observe, we can reliably classify the source of the content being circulated over Twitter.

(d) For certain subsets of URL/Tweets (e.g. those originating from certain news sources) we can automatically classify content into topical domains (“international news”, “entertainment”, “business”, “science” etc.); and for other subsets (e.g. persistent URL's - URL's that are shared repeatedly by different users over a big timespan) we can identify certain content-related attributes (e.g. video, music, etc.). So for certain restricted domains we can make some statements about how content matters in the diffusion process.

(e) Even though our view of “effects” is limited (i.e., URL persistence and retweeting rate), we have very high resolution temporal information over the lifespan of any piece of information in diffusion.

Nevertheless, we are aware that our analysis is limited in some important respects:

(a) It is limited to Twitter, which is not only just one diffusion channel among many, but may well be unlike other channels in a number of ways; (b) We can not observe the kind of offline ”effects” that viral marketers and social scientists are most interested in, such as, adoption of products, change in attitude;

(c) We have only limited information about the content that is being shared (although this constraint could be relaxed with some effort).

2.5.2 Other on-line social media

2.6 Web as a source for knowledge

2.7 MapReduce and parallel network algorithms

Introduction to MapReduce.

MapReduce is good with easy-parallelible tasks, hard for tasks that need certain global information (e.g., shortest path). However, many network problems are the second case.

Methods to convert a global problem to a series of mapreduce jobs [**?**].

Part II

The life of information: the interaction between people and content in online networks

We study the lifespan of information, from production, flow, to consumption. We can also extend our findings to general diffusion process, the spread of behavior.

We have seen different temporal patterns in the life of information - very small amount got picked up, but substantial amount exist for a long time. (although note that exist does not necessarily means spread). Factors that lead to differnet temporal patterns: interactions between people and content.

Information diffusion like virus spread in the sense that it is produced by some people, consumed by some people, and can be further spread by the consumer. However, it is much more complex in the sense that people can be infected by multiple channels, including the environmental factors, but classic epedimic models can not fully describe (ZZZZZ: need more research to make this claim).

My contributions:

1. Study one-way, one-hop flow of information, which is although the majority but largely overlooked. We did it by showing the distribution of attention among different groups.
2. Study the temporal pattern of information, especially, the persistence. Most previous work focused on the spikes but not the persistence.
3. (ZZZZZ: possible remove) Diffusion as an organic process integrating people, content, and time.

CHAPTER 3

THE PRODUCTION AND CONSUMPTION OF INFORMATION: DISTRIBUTION OF ATTENTION

Typical lifespan of a piece of information on social media is one-way, zero or one hop - depending how a hop is define: from production directly to (potentially) consumption, without any additional node in the chain.

Why is it interesting to study the direction production-consumption flow of information? 1. Majority of information; 2. Interesting social science problem; 3. very similar to mass communication - debates on the role of social media; 4. significant effect of environmental influence that will largely determine public opinion formation;

In this seciton, we focus on the production and consumption of information. Our study is led by classic communication theories: debate on two modes of communications. Different people play different roles, and which role they are playing is largely determined by two factors: (1) internally, their own goal, taste, and agenda; (2) externally, the amount of attention they enjoyed from the public.

Our contributions:

1. introduce a low-cost, crowd-sourcing method to classify users into categories that parallel to media/communication research;
2. investigate one-way, one-hop flow of information among these categories, present a high level picture of the distribution of attention;
3. show the interaction between people and content.

ZZZZZ Put two-step flow in next chapter?

3.1 Data And Methods

3.1.1 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. [?], which 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.

3.1.2 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.

3.1.3 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

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

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 in section **??** 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] 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.

3.1.4 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:

- 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

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

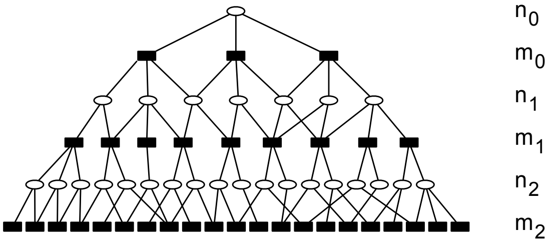


Figure 3.1: Schematic of the Snowball Sampling Method

- 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 3.1). For 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 3.1 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:

$$w_{uc} = \frac{n_{uc}}{N_c}, \tag{3.1}$$

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

Table 3.1: 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

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.

3.1.5 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 3.2 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 3.2: Statistics of crawled lists. The number of users refers only to people who appear in at least one list of the specific category.

3.2 Distribution of attention

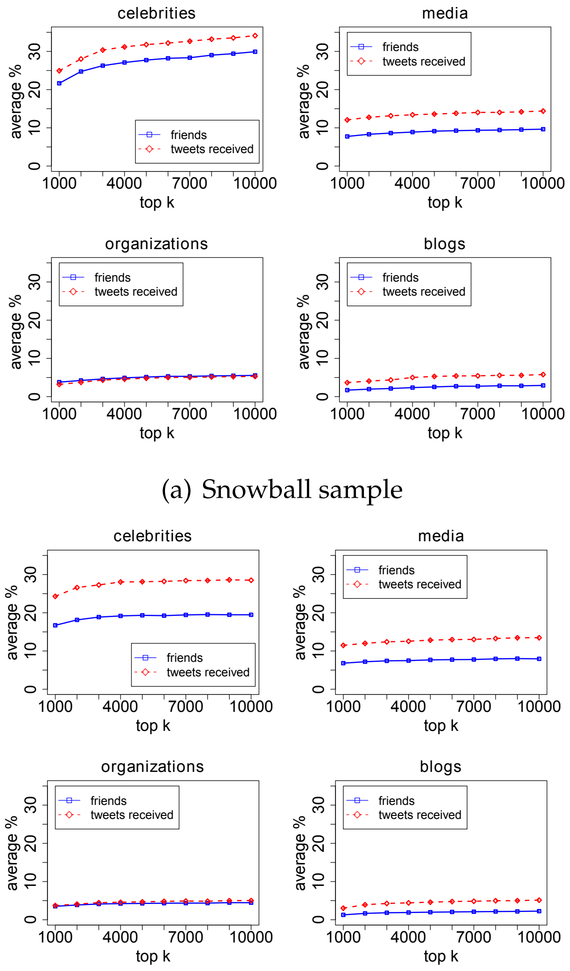
After categorizing people into categories, we can calculate the amount of attention sent and received by each category, at a global level. The way we do it is to show the reach of the “elite” categories. It can be considered as the influence of each category, as well as an estimate of the impact of the information introducted by each category. In other words, it is the maximal reach of the information produced by each category.

3.2.1 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 3.2(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 3.2(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 3.2(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, when we talk about "celebrity", "media", "organization", "blog", we mean the top 5K users listed as "celebrity", "media", "organization", "blog", respectively,



(a) Snowball sample

(b) Activity sample

Figure 3.2: 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

drawn from the snowball sample. Table 3.3 shows the top 5 users in each of the four categories.

ZZZZZ: Move the paragraph and tables below to content part.

To confirm the validity of these categories, we now consider the number of URLs introduced by various categories. As Table 3.4 (left column) shows, the vast majority of URLs are initiated by ordinary users, not by any of the elite

Table 3.3: 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 3.4: # 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

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 3.4, 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 category; that is, within the current "other" category may be hidden additional categories of opinions. As Figure 3.2.1 shows, however, even the top 10,000 most

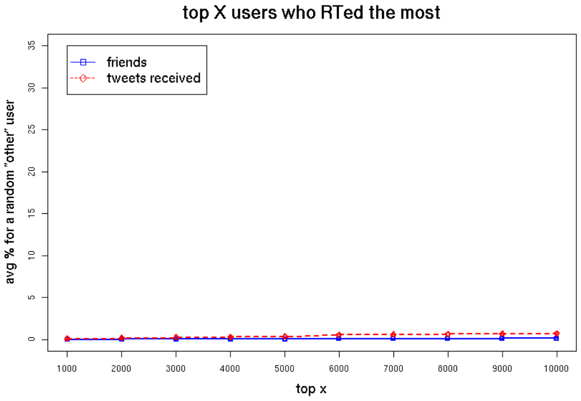


Figure 3.3: Average fraction of # following (blue line) and # tweets (red line) for a random user that are accounted for by the top K most retweeted users in the "Other" category

followed of these users accounts for a negligible fraction of attention among the remaining population.

3.2.2 Homophily of attention

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 attention 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

Table 3.5: 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**

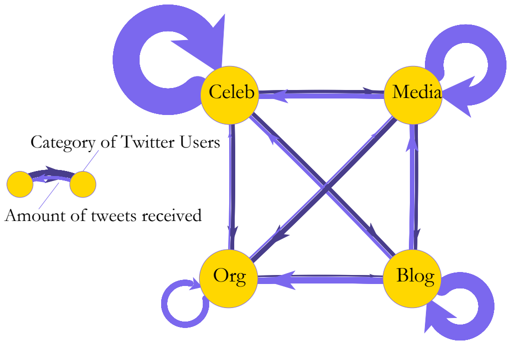


Figure 3.4: Share of attention among elite categories

3.5 shows the average percentage of friends/tweets category *i* get from category *j*. Table 3.5 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.

CHAPTER 4

ZZZZZ: TRANSMISSIVE PROBABILITY

ZZZZZ: should we combine this section with previous or next section?

For information that do spread, most previous works study the factors that contribute to the spread at each hop, independently. (ZZZZZ related work).

The categorization of actors, as introduced previously, also helped shed some light on the one-hop diffusion probability, depending on the type of users in the diffusion edge, and people's interest at different types of content.

My contribution:

1. Show homophily at diffusion;
2. Show difference in attention and influence (as measured by RTs)
3. Show people's interest at different content.

4.1 People

The origin of information will influence how it will be RTed.

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

Table 4.1: 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

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 4.1 shows how much information originating from each category is retweeted by other categories, while Table 4.2 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 4.1 and 4.2 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 3.4).

Table 4.2: 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

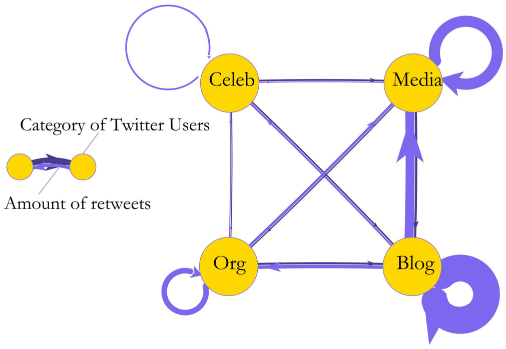


Figure 4.1: RT behavior among elite categories

4.2 Content

People have different interest at different content. So the type of content can also determine what a person will RT.

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 [**?**], 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) [1]. Of the 6398 New 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 4.2 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

[1] http://www.nytimes.com/year/month/day/category/
title.html?ref=category

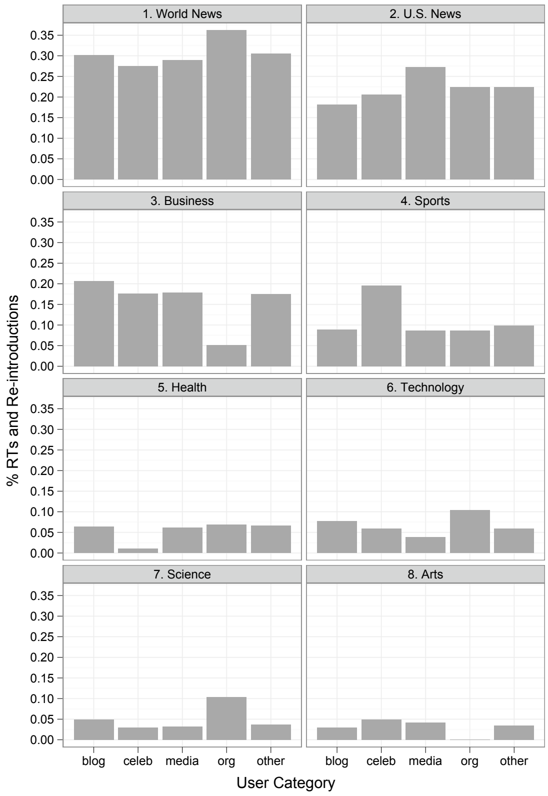


Figure 4.2: Number of RT's and Reintroductions of New York Times stories by content category

for a small selection of the most popular URLs. As Figure 4.3 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 4.3 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 and coupons for all types of stores," was popular among ordinary users, but al-

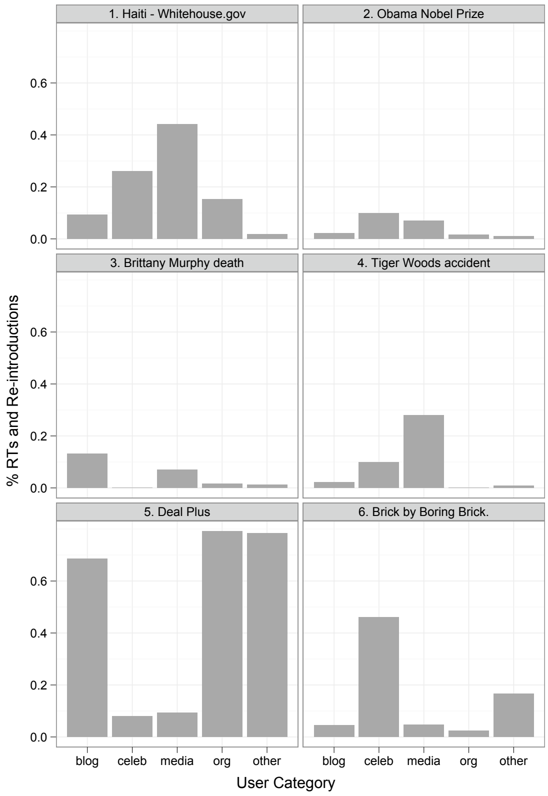


Figure 4.3: Number of RT's and Reintroductions of most popular URLs originating from media and other

most 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.

CHAPTER 5

PERSISTENCE OF INFORMATION

In our work, we noticed that although a small amount of information spread and travels, among them, a substantial portion has a very long lifespan. We think this is very interesting phenomenon that has not been well investigated yet.

5.1 Lifespan by the category of originator

ZZZZZ: Content produced by different people have different persistence. Blogger's role of information filter.

By lifetime, we mean the time lag between the first and last appearance of a given URL on Twitter. Naively, measuring lifetime seems a trivial matter; however, it is complicated by the finite observation window, which results in "censoring" of our data. 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 lifetime of a URL, in other words, is in reality a lower bound on the lifetime. 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 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.

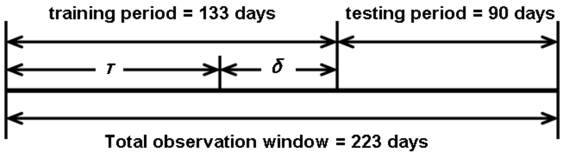

training period = 133 days
testing period = 90 days
τ
δ
Total observation window = 223 days

Figure 5.1: Schematic of window 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 lifetime 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 Figure 5.1 shows, to determine δ we first split the 223 day period into two segments - the first 133 day training segment and the last 90 day testing segment, and then ask: if we (a) observe a URL first appear in the first $163-\delta$ days and (b) do not see it in the δ days prior to the splitting point, how likely are we see it in the last 90 days? Clearly this depends on the actual lifetime of the URL, where initially we know for each URL that it persists for at least τ days. As the longer a URL lives, the more likely it will re-appear in the future, Figure 5.2 shows the upper-bound on lifetime for which we can determine the actual lifetime 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 lifetime τ if it appears at least τ days before the end of our window, we need to pick τ and δ such that $\tau + 2\delta < 223$. From Figure 5.2, we determined that $\tau = 60$ and $\delta = 48$ sufficiently satisfy our constraints.

ZZZZZ: some of the numbers here should be move above!!!

Figure 5.3 is the histogram of the lifespan of URLs, grouped by the category of users who introduced the URLs[1]. URLs initiated by the elite categories ex-

[1] This figure only shows URLs that appeared in our dataset more than once. The majority of

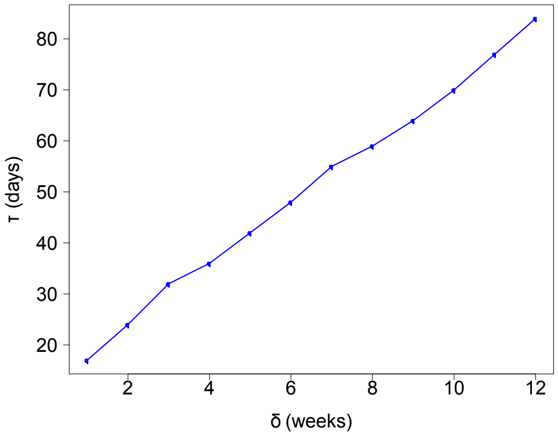


Figure 5.2: Upperbound of τ with confidence level ¿ 0.95, as a function of δ.

hibit a similar distribution over lifespan to those initiated by ordinary users. As Figure 5.4 shows, 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 lifespan 0, which are URLs that only appeared once); and second, URLs originated by bloggers are overrepresented among the longer-lived content. Both these results can be accounted for 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 stories that are picked up by bloggers are of more persistent interest, and so are more likely to be RT'd or reintroduced months or even years after their initial introduction.

the URLs (220M) appeared only once, which is 10 times as many URLs as had a lifespan of only a day.

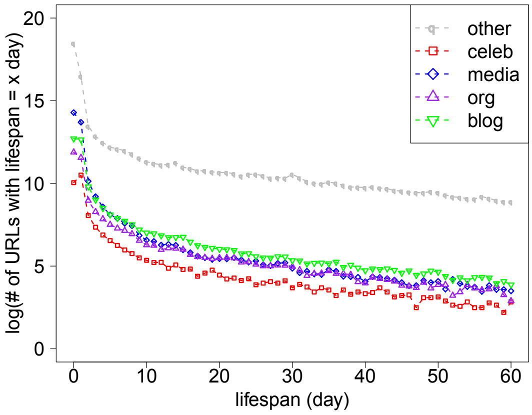


Figure 5.3: Histogram of lifespan of URLs originating from different categories

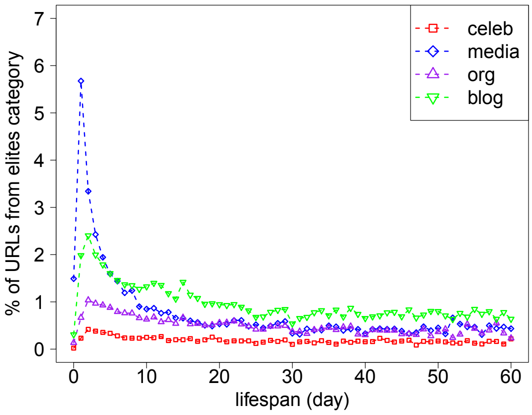


Figure 5.4: Percentage of URL initiated by 5 categories, with different lifespan

5.2 Content

As shown previously that the content originated by different people exhibit different lifespan, but it is very hard to predict the lifetime. The persistence of long-last content can not be fully contributed to contagion process - the role of content.

Evidence: Low RT-rate for long-lasting content.

A second related point, is illustrated by Figure 5.5, which shows the average RT rate = (# of retweets) / (total # of occurrences) of URLs with different lifespan, grouped by categories[2]. Unsurprisingly, URLs introduced by elite users are much more likely than those introduced by ordinary users to be RT'd—a result that is likely driven by the higher-than-average number of followers for elite users. Somewhat less expected, however, is that for all categories the majority of appearances of URLs after their initial introduction derives not from rebroadcasting, hence diffusion within Twitter, but rather from reintroduction. As large and diverse as Twitter is, in other words, it is nevertheless a subset of a much larger media ecosystem; that is, content "lives" outside of Twitter, where users can rediscover it repeatedly. 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 be rediscovered and reintroduced 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 5.6 shows, the population of long-lived URLs is dominated by videos, music, and books, consistent with our interpretation above that certain types of online content retain their relevance indefinitely, and their persistence on Twitter is driven mostly by users rediscovering content outside of the Twitter ecosystem.

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

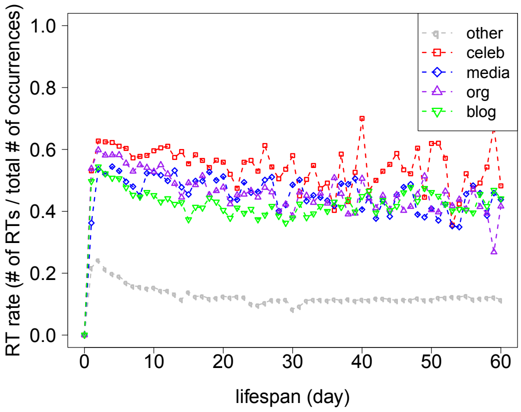


Figure 5.5: Lifetime avg RT rate, by categories

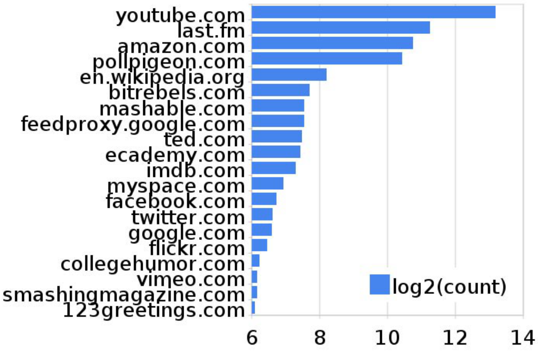


Figure 5.6: Top 20 domains for URLs that lived more than 200 days

Following up these findings, we start to look for intrinsic qualities of the content that effectively determine the dissemination process, especially, the persistence of information.

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.

5.2.1 Data

To study content in more details, we introduce a smaller Twitter dataset with richer HTML content.

Summary

In this study, we used the dataset publicly shared by the authors of [**?**][3], 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 [**?**]. 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

[3] http://snap.stanford.edu/data/twitter7.html

times in retweets [4], 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.

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 [**?**], 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 approximately follows a power-law (see Figure 15.2.1), as found previously in the distribution of URL lifespan [**?**]. Among all URLs we studied,

[4] We recognize a post as retweet when it contains “RT @” or “via @”.

the mean t_u is 217.3 hour and the median t_u is 19 hours.

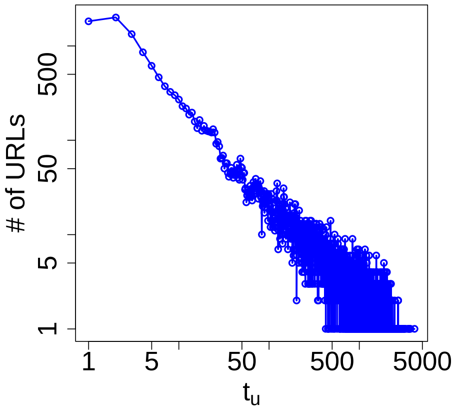


Figure 5.7: Distribution of URL decay time t_u

We further examine the relationship between t_u and the overall popularity of URLs. Figure 5.2.1 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 [?].

5.2.2 Predicting temporal patterns based on content

Strong correlation between content and the persistence of information.

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

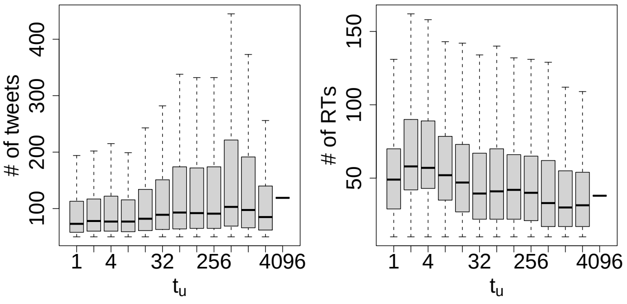


Figure 5.8: URL overall popularity as a funtion of t_u

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 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 [**?**] and calculate the centroid of time series for each class, as shown in Figure 35.2.2. 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.

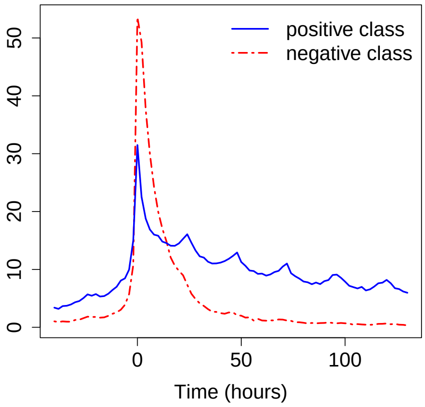


Figure 5.9: Normalized time series centroids for two classes

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>".
- 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.

Table 5.1: Feature size

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

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 15.2.2 gives a summary of the number of features in each set.

Classifier performance

To predict the persistence of webpages, we employ a Support Vector Machine (SVM)[5] 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 5.5 gives the performance of classifiers with different sets of features using 10-fold cross validation.

Table 5.5 shows that in general, the simple linear-kernel SVM classifier can predict the persistent/rapidly-fading category of URLs with impressively high

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

Table 5.2: 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

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

5.2.3 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) [?] 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.

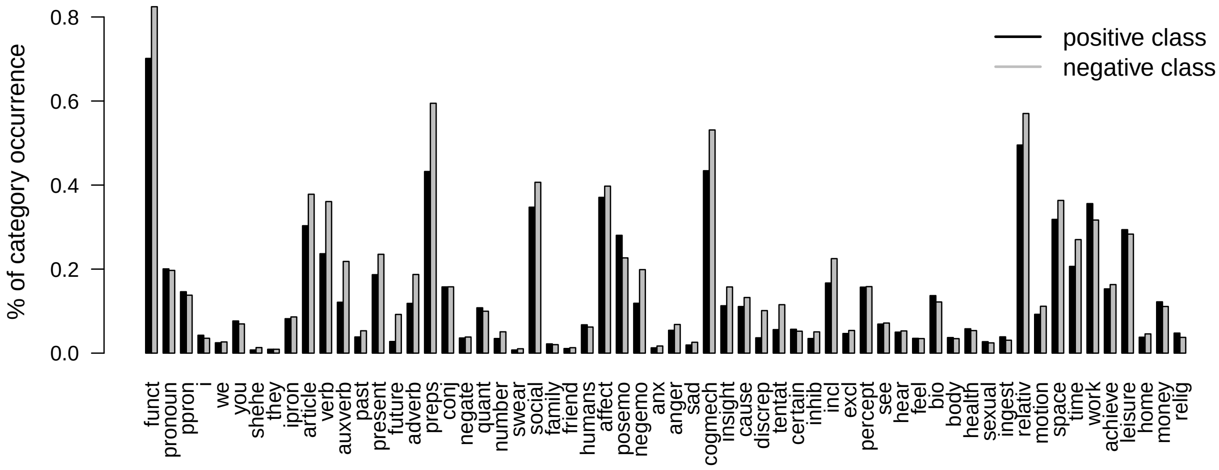


Figure 5.10: Class distribution in 60 LIWC dimensions, using words from HTML header

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.[6] Figure 45.2.3 shows the percentage of occurrence for all LIWC categories in webpages from two classes. As illustrated by Figure 45.2.3, the two classes differ the most in the following three groups of LIWC categories,

- Emotion: *posemo* (positive emotion), *negemo* (negative emotion).

[6] 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.

- 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*).

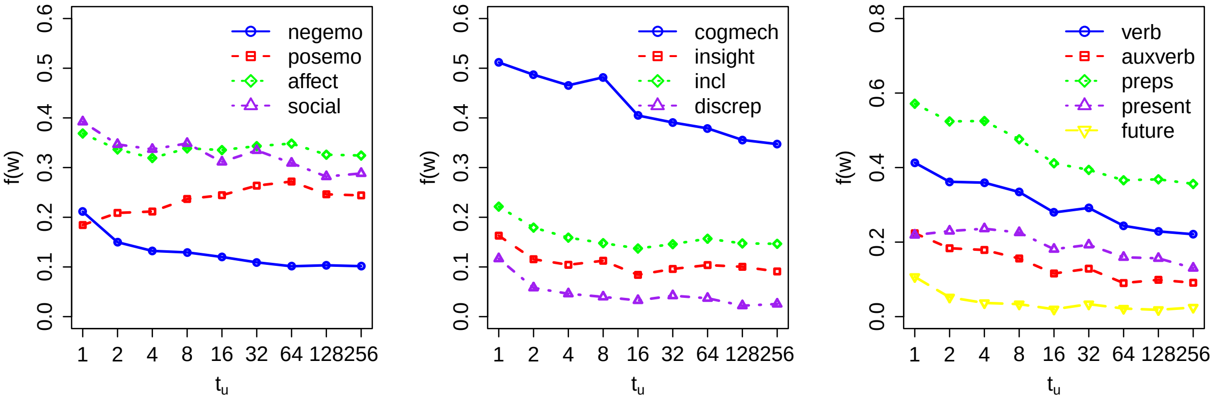


Figure 5.11: Trending LIWC categories

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 55.2.3.

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 probabilistic curve is smoother. Similar as in [**?**, **?**] 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[?], 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)[7], a flexible generative model for collections of discrete data[?]. 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

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

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 65.2.3 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).

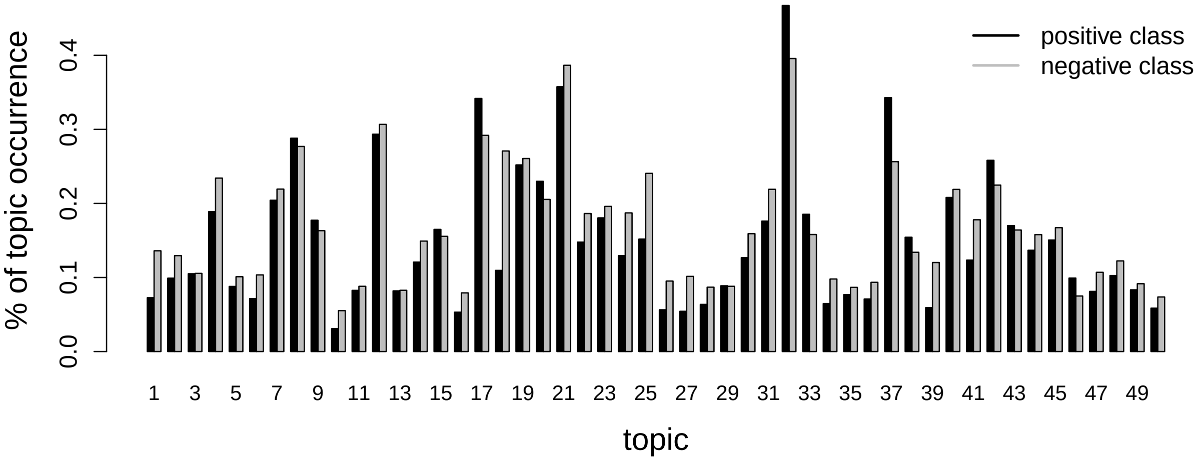


Figure 5.12: Class distribution in 50 LDA topics, using words from HTML header and body

Providing a closer look at those topics, Table 35.2.3 shows top 20 words given by the topic model. We see some similar phenomena as in previous section:

Table 5.3: 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

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 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 [**?**] on the Header feature set, and generate the top 20 trending words for each class (see Table 45.2.3)[8].

To get the words that are most meaningful, we filter all the numbers, and the words with frequency less than 20 (mostly specific names) or greater than 400 (mostly stopwords and website names). As discussed in [**?**], 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 [**?**] 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

[8] 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.

Table 5.4: 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

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 [**?**] that the persistent content - although not as viral as news - exhibits more association with art.

5.2.4 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 75.2.4 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 15.2.1), YouTube videos in general receive a longer span of attention.

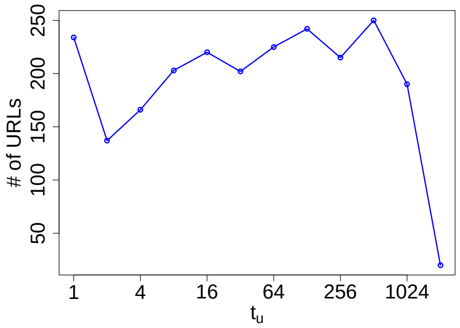
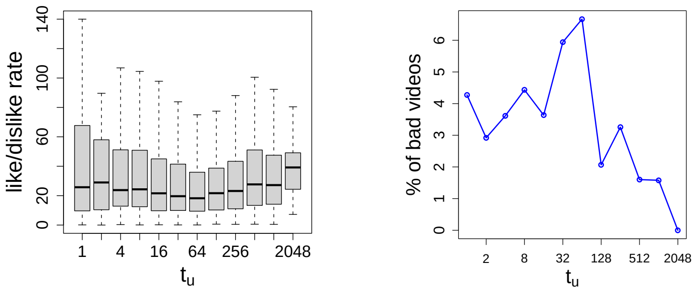


Figure 5.13: Distribution of t_u for YouTube videos

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 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 25.2.1, content that fades away quickly might not have lasting value, but in general is more viral.

Finally, in Figure 95.2.4, we consider the number of views and comments on



(a) Average likes/dislike rate (b) Bad video rate

Figure 5.14: 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.

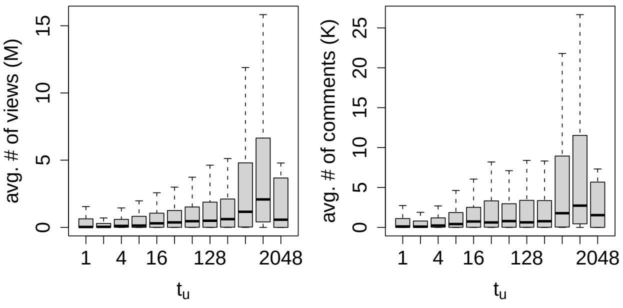


Figure 5.15: Average number of views and comments as a funtion of t_u

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 25.2.1, 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.

5.3 Negative influence

After talking about the origination and dissemination of information, now we will study the end-of-life of diffusion.

Here, the diffusion process is the engagement on an online social network. It can be analogized to how people engagement with certain topic in social media.

Disengagment can be considered as a negative diffusion. To better see the trend, we compare the natural arrival and departure of users in several communities, and ask whether the dynamics of arrival, which have been studied in some depth, also explain the dynamics of departure, which are not as well studied.

Through study of the Dblp co-autorship network and a large online social network, we show that the dynamics of departure behave differently from the dynamics 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 conductance than the degree distribution alone can explain. These two aspects suggest that nodes at the fringe are more likely to depart and additionally induce inactive and subsequent departure of neighboring nodes in tightly-knit communities.

5.3.1 Data

Orkut: social network with detailed structure, spread of behavior.

In this section, we study the dynamics of arrival and departure 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.

5.3.2 Arrival and departure correlation among friends

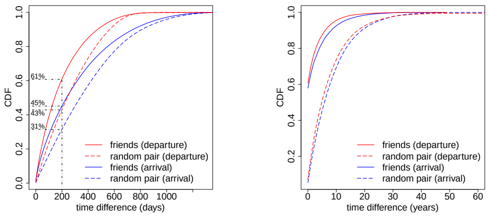
In this section, we study the basic properties of arrivals and departures. We wish to understand whether users typically arrive and/or depart together in social networks. However, we cannot directly compare gaps between arrivals and departures of friends, as networks are not stationary—consider for example 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[9]. 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 5.16.

The CDF for both arrivals and departures of friends lies significantly above

[9] 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.

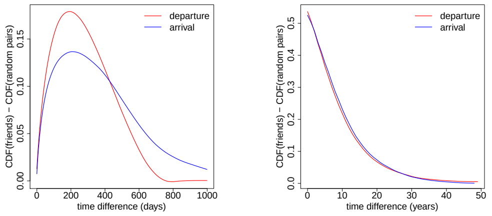


(a) SN (b) Dblp

Figure 5.16: The CDF curve for the difference in arrival and departure time between friends and random pairs of users.

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 5.17 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.

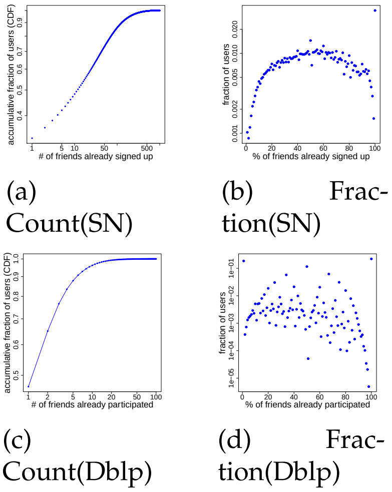


(a) SN (b) Dblp

Figure 5.17: Gap between CDF curves.

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[?], the number of friends 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 5.18(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 5.18. True social networks are of course non-regular, and while the distribution of plot (Figure 5.18(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.

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



(a) Count(SN)

(b) Fraction(SN)

(c) Count(Dblp)

(d) Fraction(Dblp)

Figure 5.18: Count and Fraction of friends already signed-up when user signs up.

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, [**?**] 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.

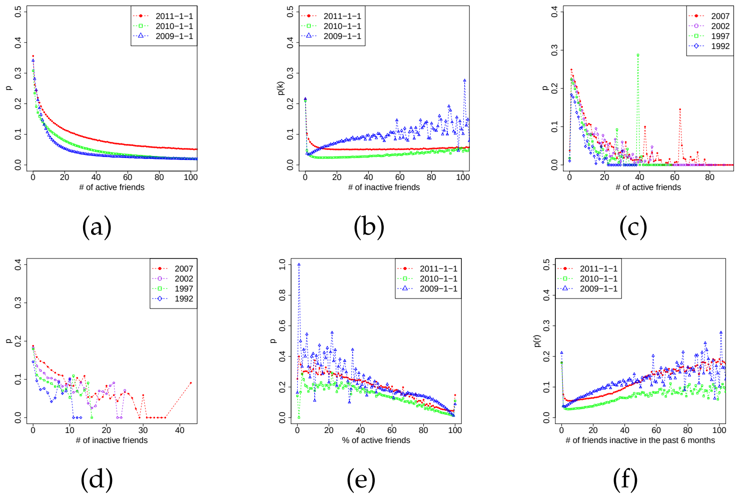
5.3.3 Dynamics of local neighborhoods

As shown in previous section (Figure 5.16), 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.

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;



(a) (b) (c)

(d) (e) (f)

Figure 5.19: 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.

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 [**?**]: 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 5.19(a) and Figure 5.19(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 5.19(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.

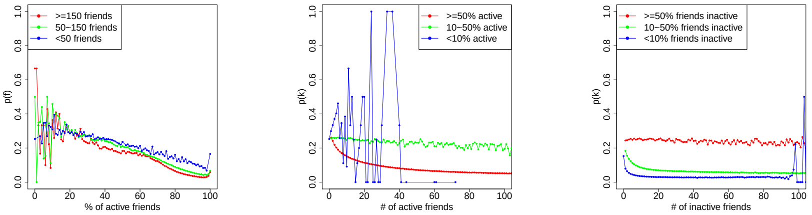
Not surprisingly, Figure 5.19(a), Figure 5.19(c) and Figure 5.19(e) show that as more and more friends stay active, a user is less and less likely to be inactive. The curve of $p(k)$ (see Figure 5.19(a)) also matches very well with what has been seen in other domains [**?**], exhibiting the "diminishing 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 5.19(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 5.19(b) and Figure 5.19(d). The curves in Figure 5.19(b) and Figure 5.19(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 5.19(f) confirms this idea, showing that the curves we see in Figure 5.19(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).

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 5.20(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 5.20(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 5.20). There are two things we note from Figure 5.20: First, for users with different fractions of inactive friends, there is a big gap between their probabilities of departure - for example, compared to users with less than 10% friends left (blue line in Figure 5.20(c)), users who have more than 50% friends left (red line in Figure 5.20(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 5.20(b)). Note that the blue line in Figure 5.20(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 5.20(c).



(a) (b) (c)

Figure 5.20: 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

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

Table 5.5: 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

distribution of examples of both classes.

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 5.5 gives the performance of the classifier with different sets of features under 10-fold cross validation.

Table 5.5 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

Table 5.6: 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	$< 2e - 16$ ***

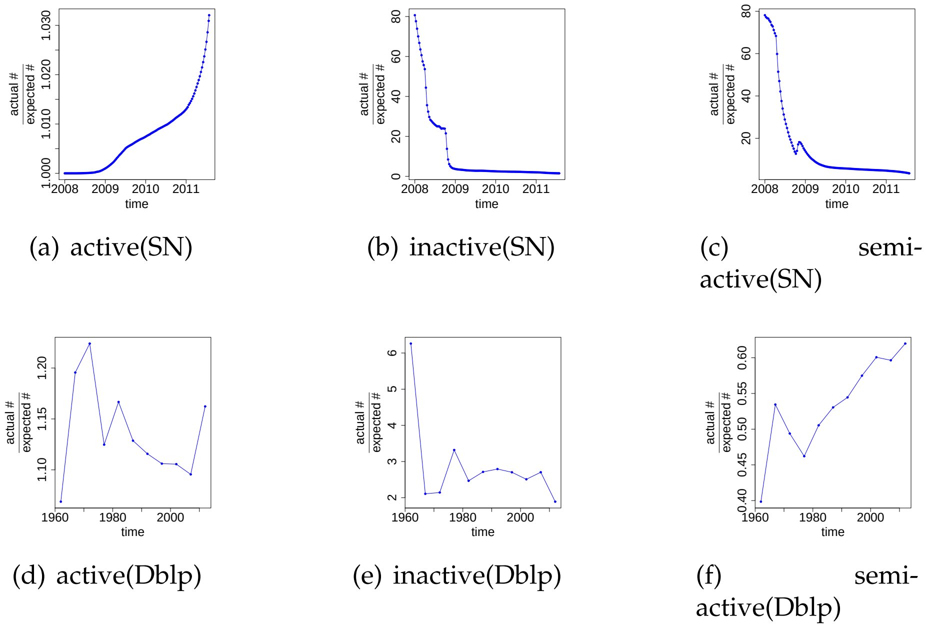
that although the activity features are more powerful, neighborhood features can also provide rather accurate insights on the departure of users.

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 5.6

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 5.6, 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.

5.3.4 Structural trends in network topology



(a) active(SN)

(b) inactive(SN)

(c) semi-active(SN)

(d) active(Dblp)

(e) inactive(Dblp)

(f) semi-active(Dblp)

Figure 5.21: Distribution of edges, indicated by the ratio of actual number of edges over the expected number of edges (formed in random process).

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 5.21(a) and Figure 5.21(d)), edges between inactive nodes (Figure 5.21(b) and Figure 5.21(e)), and the edges across active and inactive nodes (Figure 5.21(c) and Figure 5.21(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 5.22 and plots the overall density of the active (5.23(a) and 5.22(c)) and inactive (5.22(b) and 5.23(b)) 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.

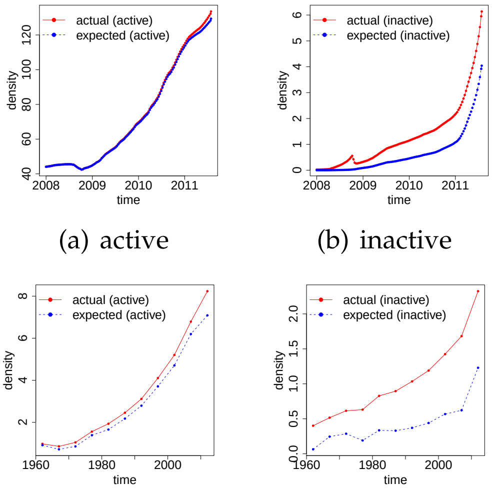
We here define density of a set of nodes(or average induced degree) as the number of edges between them divided by the number of nodes; i.e. for a set of nodes S, $density(S) = \frac{|E(S,S)|}{|S|}$ (here $E(S,S)$ contains all edges (u,v) such that

$u, v \in S$). 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 5.22 shows that the density of the active set $density(S_a)$ increases rapidly with increase in time. Comparing this with



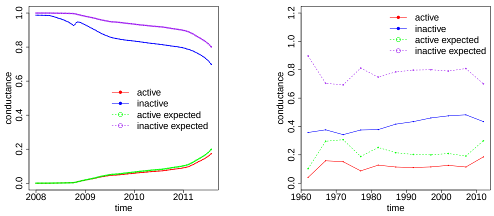
(a) active (b) inactive

(c) active (d) inactive

Figure 5.22: Density of the active and inactive subnetworks

the plot on distribution of edges in Figures 5.21, 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.



(a) SN (b) Dblp

Figure 5.23: Conductance of the active and inactive sets

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 5.23). 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 5.23 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.

Part III

Answer social science questions with social media data

Information diffusion is a long term intersting problem in social science. Problems and small lit-review.

Online social media data provide us for the first time a more or less complete corpus to examine these social phenomenons in details.

To handle data at this scale, we apply mixed methods and develop new tools (such as MR).

By finding pattern in these large-scale dataset, we shed light on some long standing social science questions.

Need to bear in mind that what we have found might be characteristics for social media, and might not be general truth in the offline environment.

ZZZZZ: future work, compare results with other social media or offline data.

A better understanding of diffusion dynamics can contribute to some interesting long-term social science questions, such as the pattern of communications, diffusion of innovation, unfolding of social movements. Todays user behavior data in social media can help us examine and possibly answer these questions quantitatively at a very large scale.

CHAPTER 6

LASSWELL’S MAXIM

CHAPTER 7

TWO-STEP FLOW

CHAPTER 8

SOCIAL MOVEMENT

Part IV

ZZZZZ: Previous structure...possibly move to intro

CHAPTER 9

THE INTERACTION BETWEEN PEOPLE AND INFORMATION

In [**?**], we studied the production, flow, and consumption of information in Twitter. As suggested in previous research in public communications, we classified users into 5 categories (celebrities, bloggers, mass media, organizations, and others) and found a striking concentration of attention on a small number of "elite" users on Twitter, as well as a significant homophily within categories. We also applied the classical "two-step flow" theory of communications in the context of social media sites such as Twitter. Our results confirmed that there are a large number of intermediary users who actively filter and disseminate information from media to the masses, and the composition of intermediaries is highly diverse. We also examined the lifespan and content of URLs broadcasted by different categories of users. We found that although content picked up by bloggers tends to stimulate a more persistent interest, the longevity of information is determined not by diffusion process, but by many different users independently rediscovering the same content.

CHAPTER 10

THE ROLE OF CONTENT

Following up our previous work, in [**?**], we studied the relationship between content and the temporal dynamics of information on Twitter, focusing on the persistence of information. Our results demonstrated a strong association between the content and the temporal dynamics of information. For example, 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.

CHAPTER 11

NETWORK EFFECT AND BEHAVIORAL CONTAGION

CHAPTER 12

DIFFUSION WITHOUT ACTIVE DISSEMINATION

Arrival and Departure Dynamics in Online Social Networks (submitted to the International Conference on Web Wide Web, 2012).

In this paper, we compared the dynamics of user arrival and departure in online social networks. We showed although user disengagement has been considered less viral than engagement, there is a substantial network effect of the departure of friends on a user's tendency to leave a network. Taking into account such network effect, we are able to build machine learning models to predict the departure of users based on their local network properties.

CHAPTER 13

SOCIAL MEDIA IN ARAB SPRING MOVEMENT

Joining social media data with real world events, we are able to study one of the most interesting (and also the most difficult) parts in media communication research (Lasswell's maxim): the effect of information. One of my ongoing projects is to study the role of social media in social movements, in order to understand how the propagation of information is leading or reflecting societal changes. We have collected a large number of tweets and twitter networks related to big social movements (i.e., Middle East Revolution, Occupy Wall-Street Movement). Using effective algorithms for community detection, hub detection, trend detection, and opinion mining, we will be able to identify the informal structure of massive communication networks for social movements and study the diffusion of ideology and behaviors within and across organizational/geographical boundaries.

13.1 Background

ZZZZZ: Summary of Arab Spring movement, major event timeline.

ZZZZZ: We observed a sharp increase of the social media usage during the arab spring period[**?**]. Such increase is considered not an accident but a driving force of the social movement [].

However, there is a debate about the role of social media in the revolution.

Previous research in this domain:

Questions to answer:

1. Whether Twitter is leading or responding to the social events?
2. How were the protest organized and spread (on the social media)?

We collected data and conducted one of the largest empirical studies of these questions. Our contributions:

1. show a landscape of Arab Spring in social media space in Middle East countries;
2. show the top-down spread of protest on Twitter;
3. show the outside-in spread of protest on Twitter.

In a 2010 New Yorker essay written just prior to the onset of the Arab Spring, Malcolm Gladwell1 famously proclaimed that "the revolution will not be tweeted", based on his assessment of the weak social ties in the Twitter user network. This sparked a lively debate with critics who pointed to the growing use of social media in the 2008 South Korean candlelight protests in opposition to the KOR-US Free Trade Agreement, in the civil unrest in Moldova and Iran protesting election results in 2009, and more recently the Arab Spring and London riots. The Tunisian and Egyptian revolutions, especially, have been dubbed as the "Twitter Revolutions"[] leading to numerous newspaper articles and government officials praising the "wired and shrewd" local activists.[]

Most of this debate and proclamations, however, are based on anecdotal evidence and are not empirically grounded. An increasing use of social media during periods of protest could be a response to the events, no different than conventional reportage by news media. Therefore, the case for Twitter as a protest medium requires evidence that an upsurge in protest content occurred prior to the outbreak of protest. A recent article by Howard et. al. [] attempts to accomplish this goal, but the data is not detailed enough to establish whether protest content occurred before or after protest onset.

13.2 Data

We did two-step crawl to get large amount of geographical constrained data. First find users from interested areas, then use them as seeds to crawl their neighbors (based on the homophily idea, their neighbors should be more likely to be physically near them).

To collect a substantial set of users and their tweets from the Middle East area in the period of the recent social movements, we first identified a set of countries of interest, including Tunisia, Egypt, Libya, Bahrain, Iran, Iraq, Israel, Algeria, Morocco, Saudi Arabia, Kuwait, Yemen, United Arab Emirates, Palenstine, Quatar, Oman, Jordan, Cyprus, Syria, and Lebanon. For each country, we used Yahoo Maps APIs to get the list of cities and towns in that country, together with the geographical centroid point for each city/town, in the form of (latitude, longitude).

After we had the centroid points of cities/towns within a country, we used Twitter search APIs to retrieve all the recent tweets generated within 100 miles from every centroid point in that country. We then parsed these tweets and extracted the authors of these tweets.

Using these authors as "seeds", we crawled one degree out from the seeds, and retrieved the profiles of all the seeds, and all the neighbors (friends/followers) of the seeds. The size of graph grows rapidly in one-degree distance. In fact, the network induced by the seeds and their one-degree neighbors already cover over 3 millions distinct Twitter users.

We crawled the profiles of these 3M users, and tried to identify their country of origin in three ways:

1. look for the country name in their self-reported location in their profile;
2. if the time-zone city is specified in their profile, map the city to the corresponding country;
3. if the location-tracking service is turned on, get the tracked location in Twitter meta data.

After parsing the profiles for all 5M users, we were able to identify the country of origin for 260K of them.

In the end, we crawled the maximal available history of tweets generated by these $260K$ users, which is, up to 3200 tweets per user. In the end we collected in total $96,350,865$ tweets in this way. Among them, 36,857,387 were generated between Dec 1st, 2010 and March 31st, 2011, by 112,661 users from the countries listed above.

Here is a breakdown of the amount of data we collected from each country.

We noticed the data sparseness issue in our dataset. However, compared with other data source, we have a good coverage of users from Egypt and .... []. As a result, in most of the studies we present below, we use only those countries that have good data density.

Table 13.1: Summary of Twitter data collected from Middle East countries

Country	number of users	number of tweets
Egypt	30, 270	6, 451, 149
Kuwait	23, 475	9, 526, 023
Saudi Arabia	16, 147	7, 202, 700
Israel	11, 292	3, 063, 959
United Arab Emirates	9, 536	4, 147, 190
Oman	3, 617	1, 348, 999
Bahrain	3, 018	793, 595
Jordan	2, 535	375, 352
Morocco	2, 114	520, 618
Iran	1, 983	1, 111, 373
Lebanon	1, 901	395, 609
Tunisia	1, 680	410, 628
Iraq	1, 132	277, 080
Qatar	1, 083	502, 199
Syria	631	123, 907
Cyprus	556	180, 394
Turkey	528	158, 832
Algeria	507	109, 864
Libya	337	91, 032
Yemen	319	66, 884
total	112,661	36,857,387

13.3 Method and Results

13.3.1 Identify Protest content on Twitter

To compare the diffusion of protest and non-protest content on Twitter, we first identify protest-related tweets.

Two major approaches in previous studies: first, tracing only specific keywords or URLs that known are related to protest[]; second, tracing the flow of tweets by well-known activist users[].

ZZZZZ: (TODO) add the mobilization time for the identified activist users (maybe in the mobilization part)!

Similarly to previous studies [], our first attempt is to have experts read through a sample of collected tweets and identify keywords or phrases associated with protest; during this process, our expert coders observed that most protest tweets contained with protest-related hashtags, such as #Jan25 and #Egypt.

Based on this insight, we say a tweet is related to protest if it contains at least one protest hashtag. Now the problem becomes to effectively pick protest-related hashtags from hundreds of thousands of hashtags seen in our dataset. On the trade-off between accuracy and recall, we narrow the pool of hashtags to be examined based on two metrics: (1) the volume of tweets containing the hashtag; and (2) the bursty-ness (as defined by Kleinberg 2004 KDD) of the hashtag occurrence. In this way, we narrow the scope down to only the top 1000 most frequently used hashtags and the top 1000 most bursty hashtags. We then have the experts to only go through those 2000 hashtags, and label them as protest-related or not.

It turns out that it is not trival to label the hashtags even for domain experts, especially in the context of our study. After the first round, the experts agreed on about 500 hashtags to be protest related, which includes most of the names of countries and cities where big protests had taken place (e.g., Egypt, Bahrain, Cairo). Although most of these hashtags are indeed used with protest content during the protest period, they are also used for non-protest content such as Egypt tourism and Bahrain flood (ZZZZZ: need to check the fact), and thus introduced certain amount of false positive, especially, before or inbetween the time when the protests were taking place.

False positive is worrisome because it might amplify the effect of social media in the revolution, and the results of our studies below are sensitive to false positive.

ZZZZZ: why is it better to miss protesters than to mis-identify them?

Knowing that social media only give us a subset of protestres no matter how accurately we identify them, we decide to take a more conservative selection of protst-related hashtags which are exclusively used by protesters, and not worry so much about true negative. We have experts go through the set of 600 protest hashtags again and keep only the ones that are protest dates, names of major players (activists and politicians), and words related to the revolution (jasmine, constitution, free+protesters name). There are around 200 hashtags in this list.

Most frequently used protest hashtags are listed in Table **??**, compared with the top non-protest hashtags in the same dataset.

Most frequently used hashtags and most bursty hashtags are listed in Table **??**, together with their labeling.

ZZZZZ: (TODO) table of most frequent and most bursty hashtags.

(Note here we consider those who sympathetic to the protest also protesters - they do not need to actually go out to the street - will be an interesting future study though, need a lot better NLP stuff to classify tweets as street tweets of sympathy tweets.)

ZZZZZ: further restriction of the conservative set of hashtags.

13.3.2 Global

Temporal Patterns

Temporal patterns of activities.

Temporal patterns of mobilization.

Top-down

As shown in the previous section, there had been a substantial amount of protest content introduced by Egyptian users on Twitter, even before Jan 25, 2011, when the first big protests took place in Tahrir Square, Cairo. Who were those foresighted users? Were they planning and organizing the protests? Were they qualitatively different than other users on Twitter? In this section, we will investigate these questions, focusing on the relationship between the status of users and their earliness at participating in the protest activities on Twitter.

To start, we first represent the earliness of a user by his mobilization day. A user u's mobilization day d(u), is defined as the day when u first used any protest hashtag. We then quantify the status of user u on day t, by the number of Twitter followers u has on day t. In order to show the aggregated status of users who started to participate in the protest at different times, we group users by their mobilization day d, and calculate f(d), the median value of user status, for each group. In Figure 4, we plot f(d) for d between December 10, 2010 and January 25, 2011. Here we can see a clear trend of decreasing status as the mobilization day gets closer to the actual protest day.

Outside-in

13.3.3 Local

13.3.4 Attention Network

To study the spread of content,

13.3.5 Triangle effect

13.4 Conclusion

By analyzing Twitter activity in Middle East area during the Arab Spring movement, we have shown that social media were used to both activate and reflect the on-goings of Middle East social movement. The relative weights of these two roles differed across countries. In particular, Egyptian users actively used Twitter to plan protests and call for a critical mass, and the users from Saudi Arabia or UAE mostly used Twitter to support or comment on on-going events. We also found that protest content travelled directionally from the central to the peripheral of the Twitter network: most protest memes were initiated by hub users and later picked up by the masses. At the individual level, we found that the adoption of protest content can be modeled by the complex contagion process - while the overall adoption rate of protest content is relatively low, people become significantly more likely to start tweeting about the protest when more than 2 friends already doing so.

Although our work is to our best knowledge the largest study of the role of social media in social movements, we have to acknowledge that our dataset is rather disproportionate: 80% of the tweets we studied came from only 5 Middle East countries. Due to technical issues, we were not able to collect an equally large number of tweets from countries such as Libya, Tunisia, and Algeria, when dramatic societal changes were taking places in these countries.

For the future work, we want to extend our study to the diffusion of protest content among countries and communities through social media. Another interesting direction is to understand how mass media (newspaper, TV, radio) and social media interact and influence each other in social movements.

This work presents one of the largest studies on the role of social media in the Arab Spring movement. Using over 2 million tweets generated by 110 thousand users in 11 Middle East countries during early 2011, we depict the landscape of aggregated Twitter usage in those countries as the revolution unfolded. Our results suggest that social media has been used to both lead and reflect real world protest activities. Compared to non-protest-related content on Twitter, we find that protest-related content travels directionally from central users to peripheral users, and the adoption of protest-related content can be modeled by a complex contagion process.

13.5 Future work

Join with Facebook results and news media data.

CHAPTER 14

CONCLUSION

In summary, I am deeply intrigued by the developing characteristics of information diffusion in online social media. Thanks to the Internet and social media technologies, I believe that we are heading towards a more democratic era where revolutions can be started by ordinary people and the power to change is in the hands of the masses. As part of this process, social media sites such as Facebook and Twitter have also evolved from friendship networks to a much broader platform for organizing social/political changes and communicating with various communities. I hope my work can help understand this movement and foster the effective flow of information in the society.

APPENDIX A

CHAPTER 1 OF APPENDIX

Appendix chapter 1 text goes here