University of California
Los Angeles

Gestalt Computing and the Study of Content-oriented User Behavior on the Web

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Roja Bandari

2013

Abstract of the Dissertation

Gestalt Computing and the Study of Content-oriented User Behavior on the Web

by

Roja Bandari

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2013

Professor Vwani P. Roychowdhury, Chair

Elementary actions online establish an individual's existence on the web and her/his orientation toward different issues. In this sense, actions truly define a user in spaces like online forums and communities and the aggregate of elementary actions shape the atmosphere of these online spaces. This observation, coupled with the unprecedented scale and detail of data on user actions on the web, compels us to utilize them in understanding collective human behavior. Despite large investments by industry to capture this data and the expanding body of research on *big data* in academia, gaining insight into collective user behavior online has been elusive. If one is indeed able to overcome the considerable computational challenges posed by both the scale and the inevitable noisiness of the associated data sets, one could provide new automated frameworks to extract insights into evolving behavior at different scales, and to form an altogether different perspective of aggregated elementary user actions.

This thesis addresses this fundamental and pressing problem and offers a *gestalt computing* approach when studying complex social phenomena in large datasets. This approach involves extracting macro structures from aggregated user actions, finding their possible meanings, and arranging data in layers so that it is iteratively explorable. The dissertation includes three major sections; first modeling

and prediction of diffusion of information by users on the social web; next, detection of topics promoted by user communities; finally, presentation of the gestalt computing framework through a methodology that uses graph theory, language processing, and information theory to provide a top-down map of group dynamics on social news websites. What we find is not only statistical significance in the extracted structure, but also that the results are meaningful to human understanding. The efficacy of the proposed methodologies is established via multiple real-world data sets.

The dissertation of Roja Bandari is approved.

Lieven Vandenberghe

Timothy Tangherlini

Stanley Osher

Vwani P. Roychowdhury, Committee Chair

University of California, Los Angeles

2013

We are the essence of joy and the core of sorrow,

We are the soul of compassion and the fount of cruelty,

We are the high and the low, the perfect and the puny,

We are the tarnished mirror, and we are the Cup of Jamshid.

- Attributed to Omar Khayyam

Table of Contents

List of Figures

List of Tables

Acknowledgments

I extend my sincere gratitude to my faculty advisor, Professor Vwani Roychowdhury, for his support throughout these years, and for ensuring that Electrical Engineering remains a multi-faceted, exciting field where one can ask bold questions. I continue to look forward to our thought-provoking discussions in the future. I wish to thank the members of my doctoral committee, Professor Stanley Osher, Professor Timothy Tangherlini, and Professor Lieven Vandenberghe for taking the time to provide feedback on this work. I would like to express my sincere appreciation to my collaborator, Professor Hazhir Rahmandad of Virginia Tech, for our numerous discussions and his intelligent feedback throughout the years working on our mutual projects. I also offer my thanks to Professor Gregory Pottie and Professor William Kaiser for their support during my earlier years as a graduate student at UCLA.

I additionally thank Professor Vandenberghe for his immensely enjoyable graduate optimization courses. I was also very pleased to serve as a teaching assistant to his Numerical Computing class. I wish to especially thank Professor Tangherlini for our conversations about digital humanities, narratives, rumor, and distant reading; also, for organizing the Digital Humanities Institute where I was able to hear important interdisciplinary discussions from leading scholars in several fields; and finally for introducing me to the Research in Industrial Projects for Students (RIPS) program at the Institute for Pure and Applied Mathematics (IPAM) where I served twice as an academic mentor. I would like to thank IPAM director, Professor Russel Caflisch, as well as Professor Christian Ratsch and all of the staff at IPAM, and I extend a special thanks to the great guest conductor at the RIPS program, Dr. Michael Raugh. I also wish to thank Professor Bernadro Huberman at HP Labs as well as my wonderful coworkers during my internship in Summer of 2011 at the Social Computing Group. I further wish to acknowledge the sup-

port of the National Science Foundation through a Graduate Research Fellowship which allowed me the freedom to explore and find a compelling research topic.

Three wonderful people in the Electrical Engineering Department deserve special thanks here: Deeona Columbia, Michelle Welch, and Mandy Smith, from the Graduate Student Affairs Office. I thank them for their friendly and invaluable help that graduate students can consistently rely on.

I would like to extend my gratitude to Dr. Mehdi Yahyanejad who was kind enough to offer data from his social news website, Balatarin. I also extend my deep appreciation to Dr. Kazem Alamdari who suggested several reference readings in Sociology that I drew upon for this thesis and to Professor Nayereh Tohidi for her advice on academic choices and encouraging my interests in Gender Studies.

I am forever indebted to all my mentors and teachers throughout all levels of my education, especially during my years at the excellent Farasat high school in Tehran, who inspired and empowered me to face the world. I would not have been here if it were not for your selfless contributions and I hope to pass on what I learn to others as you did.

I am thankful to friends, partners, and companions who have shared my deepest moments of joy and hardship throughout these year, those who made my years at UCLA better with their kindness and humor, and who remembered and reminded me what I do well. I cannot name you all but you know who you are.

I would also like to thank my brilliant sister, Dorna, who has been my companion, my playmate, my roommate, my classmate, and my colleague. You are fearless and kind and have been giving me energy and valuable feedback during the last stretches of this work.

And finally, my deepest appreciation goes to my parents for encouraging me to pursue science and mathematics, to ask questions about our human society, and to love art and literature. Thank you for your unwavering belief in me. You give me

the courage to step into unknown territories and explore what I find interesting, and the strength to persist in them. I love you.

Vita

2005	B.S. Electrical Engineering, University of California Los Angeles.
2007	M.S. Electrical Engineering, University of California Los Angeles.
2008-2012	Teaching Assistant, Applied Numerical Computing, Principles of Feedback Control, University of California Los Angeles.
2011	Research Associate, Summer internship at HP Labs Social Computing Group.
2011-2012	Research Mentor, Complex Networks Group, University of California Los Angeles.
2012	Academic Mentor at the Research in Industrial Projects for Students, the Shoah Foundation Institute project team, Institute for Pure and Applied Math.
2013	Gender Studies Concentration, University of California Los Angeles.
2013	Academic Mentor at the Research in Industrial Projects for Students, the Google project team, Institute for Pure and Applied Math.

Publications

Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, Vwani Roychowdhury:*Information Resonance on Twitter: Watching Iran.* SOMA 2010

Ansuya Ahluwalia, Allen Huang, Roja Bandari, Vwani P. Roychowdhury:*An Automated Multiscale Map of Conversations: Mothers and Matters.* SocInfo 2012.

Roja Bandari, Sitaram Asur, Bernardo A. Huberman:*The Pulse of News in Social Media: Forecasting Popularity.* ICWSM 2012

Roja Bandari, Hazhir Rahmandad, Vwani P. Roychowdhury:*Blind Men and the Elephant: Detecting Evolving Groups In Social News.* ICWSM 2013

CHAPTER 1

Introduction

The past decade has witnessed a sea change in how people interact and find information. As a result of this change, meted out by the steadily growing use of the web, processes of persuasion and opinion formation in society are being redefined. The continuous flood of footprints from human behavior on the web has produced an expanding body of literature in computer science, engineering, physics, and applied mathematics. The implications of these new modes of interaction have also galvanized research in various disciplines from economics, business, and marketing, to linguistics, sociology, political science, public policy, and journalism [Wes98, LHM02, Pap04, VB05, AG05, Dah07, Hin09, Bee09, Fre10, Him10, KSB11, WGG12, MM12, ABT12].

The industry sector has also been abuzz with speculations about the promise of *big data* and has allocated large resources to collect, process, and store increasingly larger data sets of user actions online. However, despite this multiplying literature and significant investments, there is little insight that can be extracted from such large data sets and we simply do not understand collective human behavior online. Jon Kleinberg, a prolific academic in this field, also acknowledges this in a recent blog post titled "What's the question about your field that you dread being asked?"[1]. The term *collective behavior* can take different meanings across many traditions. It is important to emphasize that in the current dissertation, we are not implying any of the meanings associated with the term in other

[1] http://www.edge.org/conversation/whats-the-question-about-your-field-that-you-dread-being-asked

fields (such as Psychology or Sociology). Here we use the term in a general sense as *attitudes and actions of groups of users*. The same is true for terms such as *gestalt* and *behavior* which have established meanings in Psychology but here are not used in that field-specific context.

This dissertation investigates collective user behavior on the web in connection to content[2]. What is meant by *in connection to content* is that we study behaviors that involve users' orientation toward, or creation, and promotion of content. This content can be published articles online, or any text or material produced and posted by users.

This behavior is captured in elementary user actions, which represent users' existence on the web, and the aggregate of these actions creates the atmosphere in various online spaces. We begin by studying diffusion of information on networks using real world datasets, and move on to constructing an automated framework that summarizes evolving groups based on users' actions signaling their preferred content. The framework follows a *gestalt*[3] approach, where macro structures are derived from aggregated user actions to provide context and meaning and allow for further exploration.

1.1 Background

Information diffusion is the detection, modeling, and prediction of spread on networks and in populations. This spread can be that of diseases, behaviors, messages, ideas, opinions, etc. Consequently the rapidly growing literature on information diffusion involves a large range of methods and applications such as studies of topological structure of information cascades, outbreak detection in

[2]We use the words *content* and *information* interchangeably.

[3]From the Merriam-Webster dictionary: Gestalt is a structure, configuration, or pattern of physical, biological, or psychological phenomena so integrated as to constitute a functional unit with properties not derivable by summation of its parts.

networks, influence maximization, epidemic models, and prediction of popularity. [LAH07b, WH07, LBK09, SRM09, LG10b, YL11, RMK11, Wat02, GLG04, LSK06, LMF07, SR08, KLP10, WHA04, KKT03a, KKT03b, LKG07, ALT08, CHK10, LH10, SH10, LMS10, LH10, TLA11, KKC11].

Another area of research related to this thesis and essential in understanding large data sets is summarization. *Network summarization* techniques often focus on user networks and include clustering and community detection methods and their evolution [GN02, CNM04b, CKT06, GSA07, Bar07, PBV07, TBK07, BGL08, WZY09, For10a, GV12]. It is important to note that community detection and evolution on social networks generally does not take content into account. On the other hand, *content summarization* methods use text-based approaches such as topic modeling and language processing to detect and track topics [GHS09, ABD08, BL06], detect and track events [LZM10], summarize content over time [JHL11, LSK09, MZ05], summarize news articles [SGH12, AHE11] or product reviews and opinions [LHC05, GPL06, GZV12]. Most of these content summarization works do not consider the network of users. Finally, there are a number of works that aim at detecting user orientation toward content such as opinions on product reviews or political leanings [FSS12, KSB11, JA08, DLP03, ZRM11]. The bulk of the work in this area is also basd on complex text processing techniques such as sentiment analysis, subjectivity analysis, or topic modeling.

In spite of this flourishing literature, there is a growing need for automated frameworks that extract the underlying structures to further explore data; works that go beyond partial snapshots of structures present in the data, and instead lay out a path to investigate the whole picture. Such frameworks will allow scholars of other fields, those more familiar with theories of human behavior, to find interesting patterns and form further inquiries from the data. The need for such frameworks has been mentioned in literature on digital humanities such as [Tan13].

The current thesis addresses this need by proposing an automated framework

to extract insights into evolving user behavior at different scales, and form a new perspective using elementary user actions. This approach involves extracting macro structures from aggregated user actions, finding their meanings, and arranging data in layers so that it is iteratively explorable. The method has two characteristics: 1) it offers context to individual user behaviors that would have been invisible otherwise 2) it offers a path to exploration.

We term this approach *Gestalt Computing* since the extracted structure adds meaning to its individual parts. In fact, what provides insight is the relationship of different parts of the structure to one another as well as information about what is *not present* in the picture compared to what is. This approach also preserves the real world dynamics of user behavior: individual user actions lead to large-scale dynamics while in turn the overall dynamics affect user behaviors. Thus one cannot view individual users outside the overall macro dynamics.

1.2 Outline and Summary of Results

The dissertation is organized as follows. In chapter 2 we present a study of information diffusion on Twitter. Twitter has increased the speed and scale of reporting on breaking news and eyewitness crisis reporting. The study was one of the first to address the emerging role of social networks in dissemination of time-sensitive information. Our data represents an early example of this process on Twitter during Iran's 2009 post-election uprising–the Green Movement. In the course of this event, when official media entities were purged from the country, users spread eyewitness reports of protests throughout the international community using Twitter messages (called tweets and retweets). We model diffusion of these messages cascading through the network of users. We find that for the case of developing eyewitness news, diffusion occurs more through a public search channel on Twitter compared with other events that spread more through the

friendship network.

In chapter 3 we move to prediction of information diffusion through predicting popularity of news articles, measured through the number of times they are shared on Twitter. We construct a multi-dimensional feature space derived from properties of an article and evaluate the efficacy of these features to serve as predictors of online popularity. We use and compare a number of regression and classification algorithms and discover that one of the most important predictors of popularity is the source that publishes the article. Our study also serves to illustrate the differences between traditionally prominent sources and those immensely popular on the social web.

Chapter 4 offers one of the first studies on a very large and growing demographic online – mothers. The website, called CafeMom[4], hosts a large number of discussion threads where multitudes of topics are discussed by users with varying opinions and interests. By augmenting topic modeling with simpler text analysis and community detection, we establish an automated method that generates valuable insights into the evolution of forum conversations and highlights similarities in attitudes of socially connected users. The result is a multiscale representation of what topics are being discussed, what the users are saying about each topic, how the conversation is evolving over time, and how friendships relate to content.

Our central questions in this dissertation are: How do group dynamics play out online? And how can we detect them at large scales? In Chapters 5 and 6 we propose an automated and unsupervised methodology for summarization of group dynamics in online forums using simple actions by users based on their content preference. We use this methodology to study political group dynamics in a popular social news aggregation website with 4 years of data. A social news website has mechanisms for users to post and rate stories which are then displayed

[4]http://www.cafemom.com/

based on their popularity among the users[5]. The dataset we use is from a social news site popular among Iranians inside and outside Iran. The website, named *Balatarin*[6] (translated *The Highest*), quickly became a prominent venue for seeking and promoting information and discussing opinions in the Persian-speaking population. The recent surge of political change and popular uprisings in several Middle Eastern and Islamic countries (such as Egypt, Syria, and Turkey), make it compelling to study how political group dynamics manifest in this dataset and how major political events affect these dynamics.

We use indicators of user preference for content (called "votes" in the context of this dataset) and demonstrate that they are a meaningful measure for finding communities in multi-issue contexts. Using a graph-theoretic community detection algorithm we extract groups of users with similar interest in content and track these groups' temporal evolution. We then identify representative content for each group and produce summaries of each path and quantify their characteristics.

The result is a temporal map of paths representing evolving groups of users, each characterized through automatically-identified content preferred by their members. We further quantify the paths' attributes and relations between paths to obtain a full picture of the dynamics. In addition to demonstrating that paths are distinct in terms of a statistically significant difference in their preferred content, we also show that these paths are logical and meaningful to human understanding.

Figure 1.1 illustrates the implementation of this methodology. In this visualization time begins on top of the figure and progresses downward. Each oval shape represents a community at a one-month-long period and its size scales with the square root of number of users in the community (largest communities include 3000 users). Evolution paths are alphabetized from A to N and marked in different colors. The results reveal evolving groups with distinct preferences

[5]Some examples are Reddit, Digg, and Slashdot.
[6]balatarin.com

and demonstrate the immense effect of a contentious political event in June 2009. This event was the post-election uprising following which there is a structural rearrangement of groups, an abrupt and enduring shift in the focus of groups, and a near-complete extinction of certain interests.

Furthermore, through a principal component analysis on path memberships, we find a latent political dimension in the website. The results are illustrated in the biplot of Figure 1.2, which projects each path onto the first two principal components. The two components together account for 43% of variance in membership overlaps. The first component (horizontal axis) matches the time dimension of the data closely, placing pre-election paths to the left and post-election paths to the right. The election event emerges automatically as the focal point of this axis. The second component (vertical axis) represents an underlying political dimension in the website. Based on this component, paths A (Reformist), F, G (Foreign Affairs 1 and 2), D (Anti-Ahmadinejad), N (Green Human Rights), I, K, L (Green Protests 1,2, and 3) are placed in the opposite two quadrants from paths C, B (both Conservative), E (Sarcastic opposition), H (Sensationalist), J (Anti-Reformist) and M (Separatist). This division underscores the role of the Green Movement in defining the political dimension of the site, placing those opposed to the main body of the Green Movement, including those opposing the reformists from both sides of the political spectrum, in opposition to those in support of the movement. In addition, the proximity of paths F and G to paths I,K, and L, on the vertical axis, means that the core users in F and G have shifted their attention from Foreign Affairs to the Green Movement following the election unrest.

Our results show that meticulous study of content shared on the forum is not necessary in detecting meaningful evolving groups. In fact the most readily accessible quantities, the actions of users, provide adequate information. The proposed method is widely applicable to different contexts, requires no expert knowledge of the forum under study, and allows for both high-level and fine-

grained inspection of groups over time.

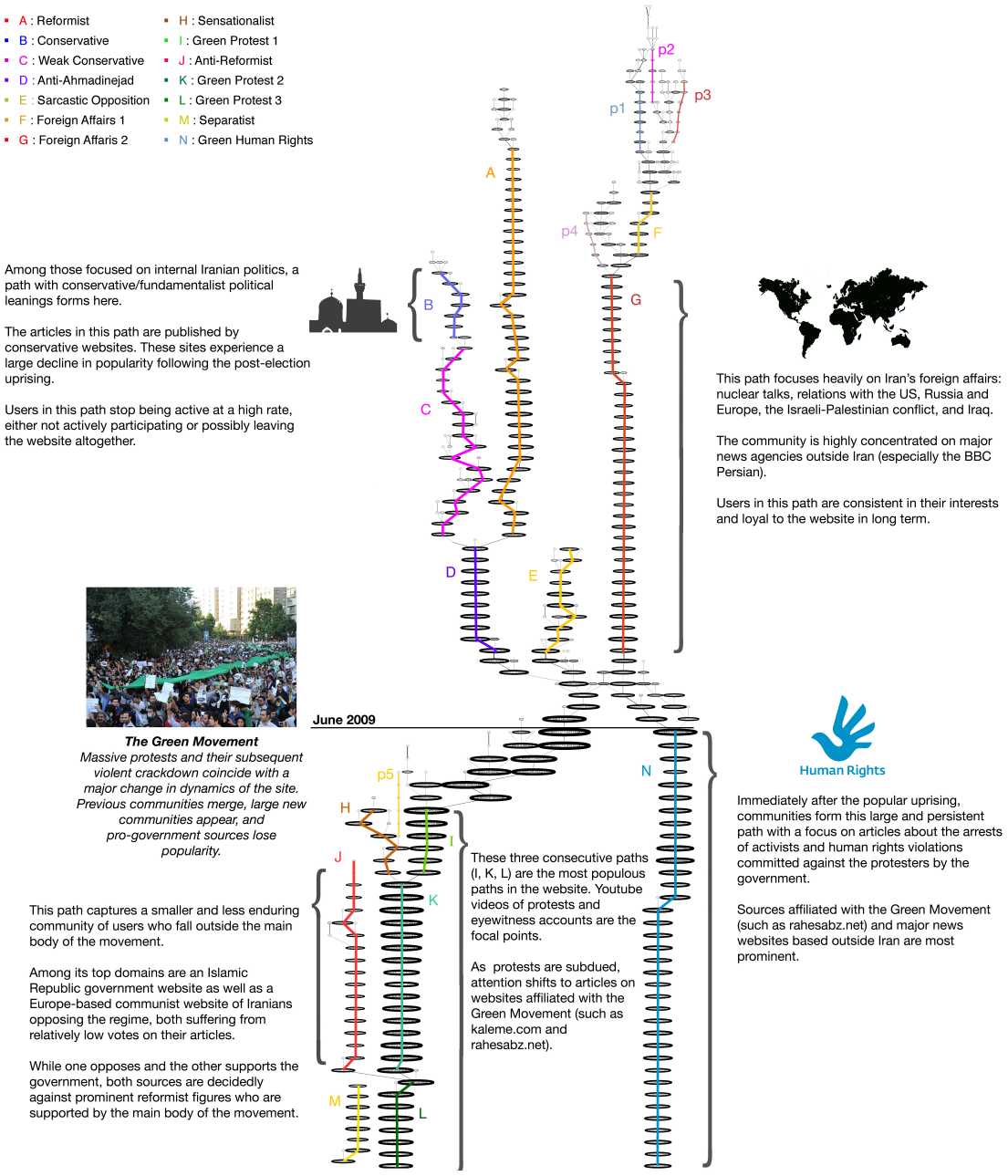


Figure 1.1: A gestalt map summarizing evolving political patterns over four years on a social news site *Balatarin.com*. Content preferences were used to infer implicit user communities and their evolution.

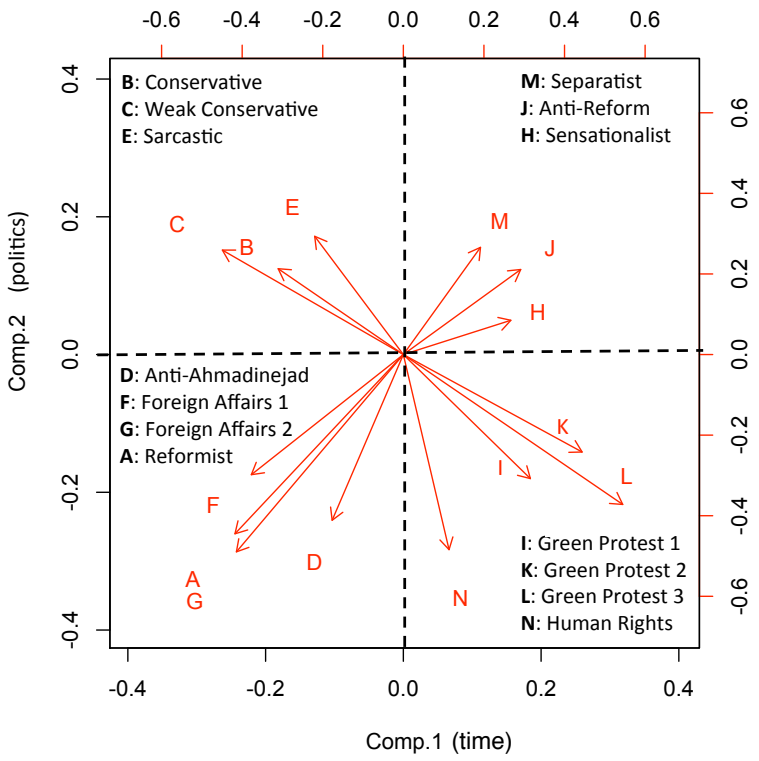


Figure 1.2: Biplot of paths using the first two principal components.

CHAPTER 2

Event-specific Information Diffusion

With an estimated 200 million users posting 65 million messages (tweets) a day, Twitter ranks high in social networking sites. In recent years, Twitter has played a significant role in live crisis reporting on the web. From natural disasters in Haiti and Japan, to popular uprisings such as those in Iran, Tunisia, and Egypt, people from all over the world used twitter to post and spread live eyewitness reports and time-sensitive information.

Twitter's 140 character limit on each user's tweet influences the kinds of messages that are spread; its directed social network of friends and followers provides an avenue for viral information dissemination via retweets (re-posting of another user's message). Finally, a public search capability and direct links to popular topics of the moment create a broadcast channel that can be accessed by everyone. These unique characteristics can lead to information dissemination dynamics that are different from other networks studied in the literature such as those of the general web or the blogosphere.

There have been a number of papers studying different network phenomena on twitter. Java et al. [JSF07] study the topological and geographical properties of Twitter's social network, and show that users with similar intentions connect with each other. Huberman et al. [HRW09] assert that the use of @ replies in twitter indicates a sparse and hidden network of connections underlying the explicit network of friends and followers. Boyd et al. [BGL10] examine the emergence of retweeting as a conversational practice. Lerman et al. [LG10a] track how

interest in news stories spreads among social networks of active users and show that social networks play a crucial role in the spread of information. Kwak et al. [KLP10] crawl the entire Twittersphere to study its topological characteristics and its power as a new medium of information sharing.

The above papers focus on the whole twitter network irrespective of content of tweets, however, one can see that there are differences in user behavior in different contexts; for example, user behavior in reporting crisis is likely to be different from user behavior when discussing a newly released film. Studying the aggregate of these behaviors leads to losing useful nuances. We therefore begin by filtering the user network based on a specific context prior to analysis and believe the results will be more meaningful. This work so far demonstrates that hyperlinks on the web and retweets on twitter might not follow the same model. Additionally, we observe that the relative roles of a public broadcast channel in comparison to the friendship network channel are context-dependant.

In crisis reporting, reliability of information is of great importance and ultimately part of this research is aimed at finding measures and modeling behaviors that constitute a culture in a network with respect to different content types. Such models will help determine which online communities demonstrate a culture resilient to the spread of rumors.

2.0.1 Dataset: IranElection

On June 12th 2009, Iran held its presidential election between incumbent Mahmoud Ahmadinejad and three other candidates, including a popular challenger named Mir Hossein Mousavi. The result, announced as a landslide for Ahmadinejad, led to charges of election rigging, and massive protests. With international news reporters purged from the country shortly after the election, eyewitness citizen reporting became the only means of documenting the events and Twitter

became a window for the world to witness the mass protest movement and its violent crackdown by the authorities. We focused on a network of over 3 million twitter messages (tweets and retweets) posted in June and July of 2009 and analyzed diffusion of information about Iran's post-election protests through cascades of retweets posted by 500K users. According to the official Twitter blog[1],

> Among all the keywords, hashtags, and phrases that proliferated throughout the year [2009], one topic surfaced repeatedly. Twitter users found the Iranian elections the most engaging topic of the year. The terms #iranelection, Iran and Tehran were all in the top-21 of Trending Topics, and #iranelection finished in a close second behind the regular weekly favorite #musicmonday.

The keyword *IranElection* remains in the top all time trending topics on twitter to this day. The timeline of tweets about the events in Iran are shown in Figure 2.1.

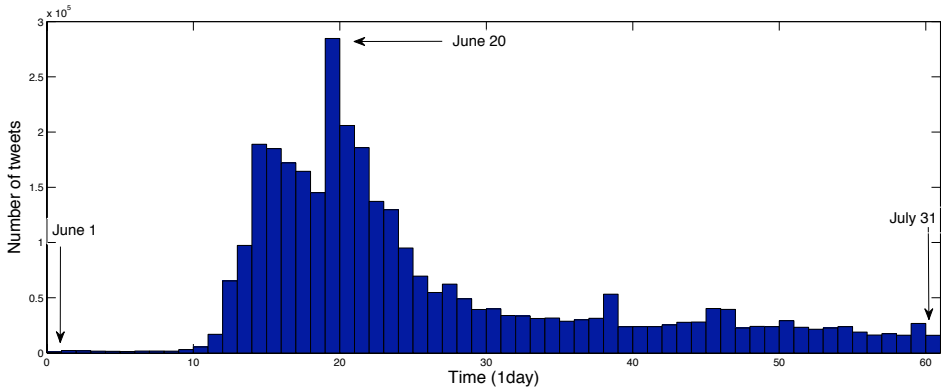


Figure 2.1: Timeline of tweets about Iran's post-election protest. June 20th marks a day of violent crackdown by the government.

We collected the tweets and friendship network of over 20 million users by making parallel calls to the Twitter API, beginning with a list of 100 most active

[1]http://blog.twitter.com/

users on the topic of Iranian Election as reported by the Web Ecology Project [2]. Using these users as seeds, we traversed their directed FF (friends and followers) network and reached 126,000 valid users who were one step away from the seed users. We continued to traverse the FF network of these depth-1 users, and arrived at 23 million distinct depth-2 users.

We used ten widely used keywords related to our desired content to eliminate irrelevant tweets. As a result, we narrowed down the data to a total of more than 3 million tweets posted by 500,000 users connected with 40 million edges. Through random sampling of twitter user IDs, we estimated our coverage to be near 97% of tweets related to Iran's events.

2.0.1.1 Network Topology

Investigating the topology of the directed network of friends and followers is the first step in understanding this dataset. Table 2.1 shows some measured parameters for this topology. The directed network of users who participated in the conversation has power-law ($p_k = k^{-\alpha}$) in and out degree distributions with coefficients of 2.85 and 2.42 respectively.

The value of clustering coefficient for this network suggests the presence of strong local clustering, meaning that many users have mutual friends. Moreover, social networks have been shown to have positive assortativity, indicating that high-degree nodes connect to other nodes with high degree. But for this dataset, the F-F network has negative assortativity, which suggests that nodes are likely connect to nodes with degrees very different from their own. We also found that the correlation between the number of tweets and number of followers is very low (0.04) suggesting that posting more tweets does not lead to gaining more followers. On the other hand, having more followers and more tweets was more correlated to the number of retweets of a user's messages.

[2] http://www.webecologyproject.org/

Table 2.1: Network Topology Measurements

Metric	Value
Total Nodes	470,040
Average Degree	87.10
In-degree Distribution α	2.85
Out-degree Distribution α	2.42
In-degree Distribution D	0.0167
Out-degree Distribution D	0.0087
Correlation of in-degree and out-degree	0.6936
Reciprocity	0.4813
Clustering Coefficient	0.1052
Assortativity	-0.2633

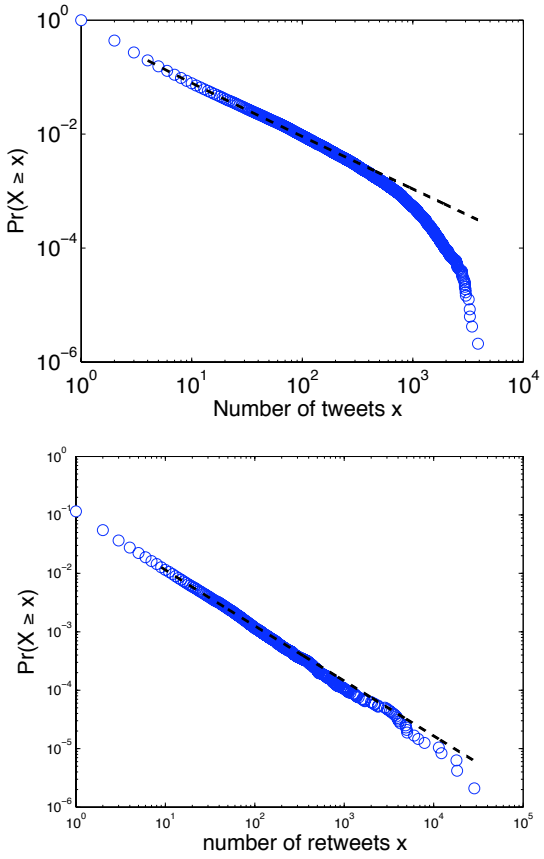


Figure 2.2: CCDF of number of tweets and retweets per person. Tweets have a power law distribution with an exponent of -1.94.

As demonstrated in Figure 2.2 user activity (number of tweets per user) also follows a power law distribution with an exponent of -1.94 and a Kolmogorov-Smirnov goodness-of-fit metric equal to 0.0110.

2.0.1.2 Information Cascades

Consider a network where each node is connected to another node if the second node is a retweet of the first. We call the set of nodes that are connected in this fashion, an *information cascade*. More than 2 million tweets from our set of over 3 million tweets were never retweeted and are thus isolate nodes, leading to a sparse tweet network with 450,000 edges.

Degree distribution of the tweet network was found to be power law with exponent -2.33 as visualized in Figure 2.4; the corresponding Kolmogorov-Smirnov goodness-of-fit metric is 0.0045. Cascade sizes also follow a power-law distribution with exponent of -2.51 (Figure 2.5). Cascades tend to be wide rather than deep, most having a central hub, and more than 99% of the cascades have depth less than 3. The exponent of -2.51 corresponding to the cascade size distribution, is different from what one expects from a branching process (-1.5) usually used to model information cascades, and so is the shallow depth, implying that Twitter information cascades might have different underlying dynamics.

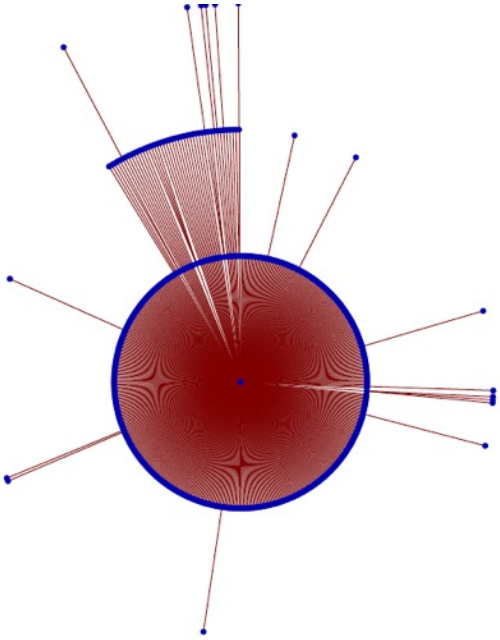


Figure 2.3: Example of a retweet cascade

We found that at most 63.7% of all retweets in our dataset were reposts of someone a user was following directly and the rest were tweets accessed through the public timeline. A similar dataset with keywords related to Michael Jackson consisted of 78.5% retweets within the F-F network, and a dataset relating to the swine flu consisted of 77.3% of retweets of direct friends (both keywords were also trending topics). This shows that keyword search and trending topics on

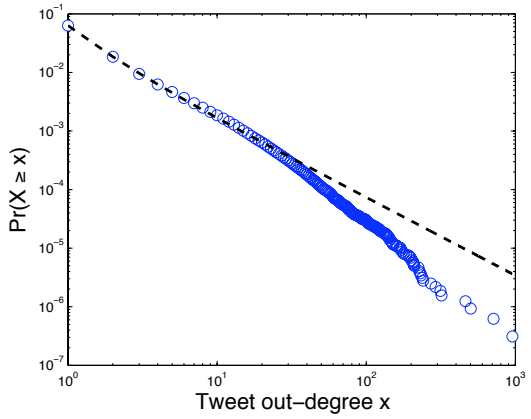


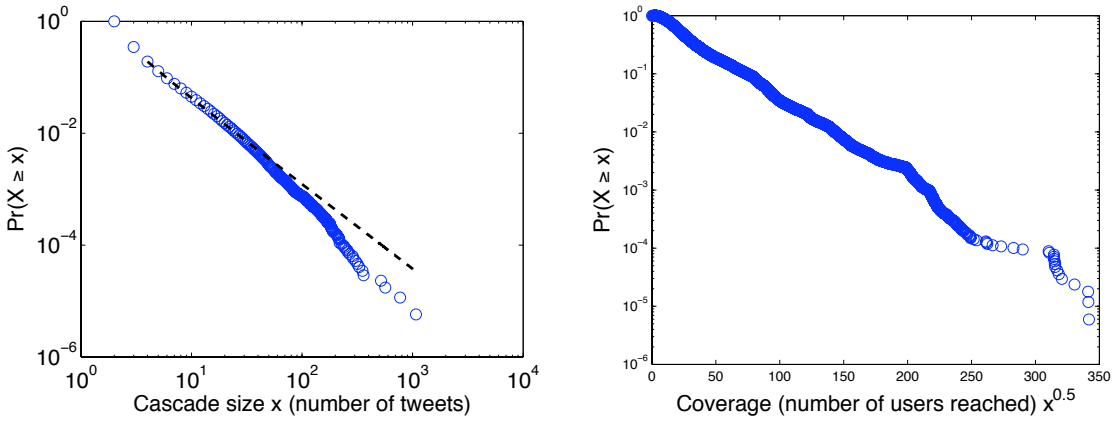
Figure 2.4: Tweet out-degree distribution is power-law with exponent of -2.33

Twitter's front page played a more significant role in the spread of information regarding Iran's protests compared to other topics. Figure 2.6 shows that as the number of tweets grow, there is a drop in the percentage of retweets posted by direct followers. This work has been presented in [ZBK10].

2.0.1.3 Content of Cascades

Study of contents of collected data in its context can be a compelling aspect of data analysis. We looked at the contents of medium and large cascades (with over 30 nodes) in our data set and observed several noteworthy characteristics. The contents of tweets can be categorized as follows:

Breaking news: An important characteristic of the twitter network is the real-time nature of much of the information in tweets. For the dataset studied in this paper, real-time reports of events in Iran were important to individuals following the post-election unrest and so a large number of tweets include breaking news. These tweets were sometimes sent by official news media in the form of links to the news piece on their website. In some other cases tweets were either updates by Iranian people in Iran, or individuals who had direct contact with eyewitnesses in Iran. Some of these tweets kept spreading long after the incident had passed.



(a) (b)

Figure 2.5: (a) Cascade size distributions is power-law with exponent -2.51. (b) Coverage size distribution is a stretched exponential distribution.

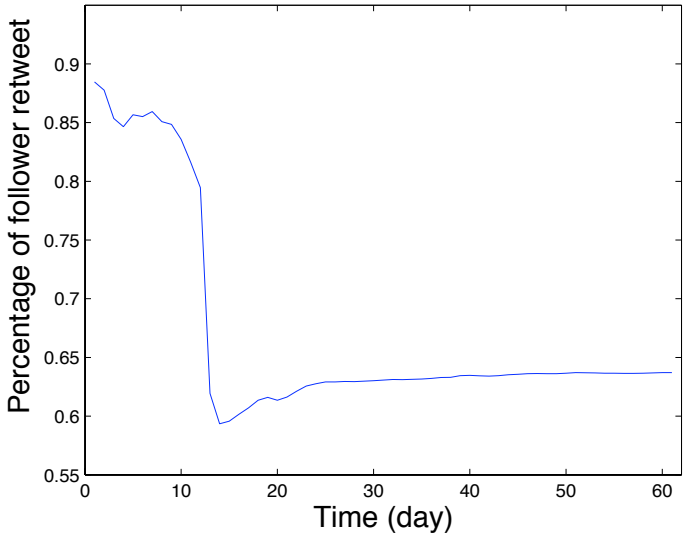


Figure 2.6: Timeline of percentage of retweets by followers.

non-time-sensitive material: Sharing photos and videos, political analysis, personal accounts of protests in blogs, and instructions for the twitter community on how to get involved, were among other types of content in tweets. These tweets commonly included links to websites that contain the information. The two largest cascades in the dataset are about spreading proxies that help Iranians bypass censorship that blocks many websites. Other popular tweets include instructions on engagement of twitter community in support of protests, directions on how to conduct Denial of Service attacks on Iranian government websites, first aid information for people in Iran, and instructions on how to avoid spreading rumors and detect reliable information. Other tweets shared plans for future actions on the ground in Iran, such as time and locations of future protests or plans for a national strike.

In our dataset, 487,005 distinct URLs were used 1,582,537 times. Frequency distribution of URLs was power-law with an exponent equal to 2.14, which suggests the rich-get-richer phenomenon (with Kolmogorov-Smirnov goodness-of-fit metric of 0.0047). The most popular URL[3] found in our dataset is one that adds a green overlay or a green ribbon to a user's Twitter avatar in support of the protesters in Iran who also used the color green.

Rumors and misinformation : Unverified information from unknown sources can lead to spread of rumors and misinformation on twitter. It appears that the twitter community was relatively successful in recognizing reliable users as sources of information. Nevertheless there were rumors that spread during the period of our study. Specifically one rumor that tanks had appeared on the streets in Tehran spread widely on twitter. On a few occasions rumors about the arrest of opposition leader Mir Hussein Mousavi were spread either intentionally or due to some level of fear and hyper-sensitivity to the possibility of such an event.

Spam: We find some irrelevant hash-tags came with our tweets, for example

[3]http://helpiranelection.com/ (appearing about 200K times).

#jobs and #loan which appear more than 5000 times in our dataset. Spammers tried to use the hashtag #IranElection in order to use its popular public timeline to advertise their own websites. It has been confirmed that furniture chain Habitat took advantage of the protests in Iran to market its spring collection on Twitter.

Others: Some of the largest cascades are about Twitter itself. The twitter community was very aware of its own activism and role in the Iranian struggle, although sometimes their perception of this role was exaggerated. A number of largest cascades are about the US government, such as president Obama's statements about the unrest. In fact the most retweeted Persian-language tweet was by the White House with a link to Obama's press conference on Iran (247 retweets). Another interesting observation is that some of the cascades -including the fourth largest cascade- are jokes, e.g. by The Onion. There were a lot of jokes, encouraging words, and funny slogans on the ground in Iran during the protests, which helped release tension and diffuse fear among protesters. Funny tweets might serve a similar function for twitter users who were following the stressful developments on Iran around the clock.

Sources of tweets in medium and large cascades can be categorized as follows:

Official news media: Much of breaking news was tweeted by official news media. @breakingnews (breaking news from MSNBC), @cnnbrk (breaking news from CNN), @anncurry (NBC journalist), and @laraabsnews (ABC News) consistently appear in medium and large cascades.

Alternative media: Alternative media such as weblogs also have a presence in our dataset. Mashable, a popular social media news blog, has a significant presence in large cascades. Tehranbureau, a news blog with accurate information on Iran, also has a presence as the source of several information cascades, although it has a much less prominent presence than Mashable.

Iranian tweeters: A significant number of cascades were originated by Iranian

Tweeters, some of these users were tweeting inside Iran and some others were tweeting from other countries (@oxfordgirl,@persiankiwi). These users were the source of many medium-size cascades (between 30 and 150 retweets).

Celebrities: The two largest cascades (1074 and 771 retweets) were originated by a British actor named Stephen Fry. A British author, Neil Gaiman, was also the source of some of the large cascades. These celebrities have a substantial number of followers which helped generate huge cascades.

2.0.1.4 An Observation of Rumor Propagation

Studying spread of news online directly leads to another important practical question: How reliable are the reports that spread widely on social networks and social news websites? Consequences of reliability can be very different depending on the context; for example, rumor about a new movie release has different ramifications than rumors about the condition of survivors in a tsunami stricken town. We were able to track a few rumors that spread through our dataset. One of the most widely spread rumors during the election protests was a report of tanks appearing in the streets of Tehran (Figure 2.7). Figure 2.8 shows the timeline of these tweets as well as the timeline of tweets that refute it. One can think of rumors as expressions of hopes and fears of users and speculate that there is more to them than a simple truth or falsehood value. Rumors also spread on twitter about the 2010 Chilean earthquake, 2011 Egyption revolution, and several other similar events. Kostka et al. [KOW08] studied the spread of competing rumors in social networks using a game theoretic framework, however the two rumors were competing for attention, rather than contradicting one another, and it that sense they were merely two messages competing regardless of validity.

Figure 2.7: Example of a cascade spreading a rumor which spread widely on Twitter

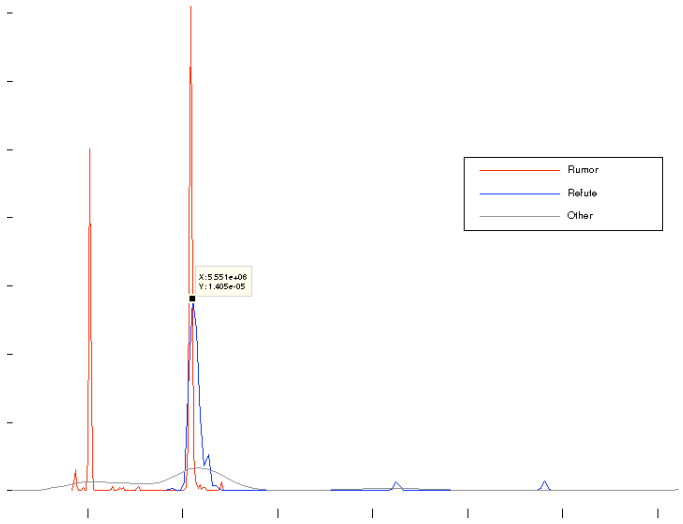


Figure 2.8: Timeline of "Tanks" rumor retweets (red) and refute tweets(blue). The grey curve is chatter that is not specifically about the rumor but indirectly relates to it (e.g. mention of tanks in Tiananmen square)

2.0.2 Modeling User Interaction with Content

Information propagation has been studied both empirically and theoretically for many years by sociologists concerned with diffusion of innovation [Rog95]. Watts [Wat02] theoretically analyzes cascades on random graphs using a threshold model. Wu et al. [WHA04] present an epidemic model to study global properties of the spread of email messages. Leskovec et al. [LSK06] empirically analyze the topological patterns of cascades in the context of a large product recommend network and study efficacy of viral product recommendation [LAH07a]. In another paper, Leskovec et al. [LMF07] examine information propagation structure and build a model that generates realistic cascades on blogosphere. Algorithms for identifying influential nodes for spreading or detecting information dynamic are presented on collaboration networks and blogospheres [KKT03a, LKG07]. To model the web traffic, Simkin et al [SR08] propose a branching process intertwined with fitness factors associated with each website. In most of these studies, the structure of the underlying networks were not defined and network structures have to be inferred

from the information flow.

CHAPTER 3

Information Diffusion Prediction

3.1 Introduction

News articles are very dynamic due to their relation to continuously developing events that typically have short lifespans. For a news article to be popular, it is essential for it to propagate to a large number of readers within a short time. Hence there exists a competition among different sources to generate content which is relevant to a large subset of the population and becomes virally popular.

Traditionally, news reporting and broadcasting has been costly, which meant that large news agencies dominated the competition. But the ease and low cost of online content creation and sharing has recently changed the traditional rules of competition for public attention. News sources now concentrate a large portion of their attention on online mediums where they can disseminate their news effectively and to a large population. It is therefore common for almost all major news sources to have active accounts in social media services like Twitter to take advantage of the enormous reach these services provide.

Due to the time-sensitive aspect and the intense competition for attention, accurately estimating the extent to which a news article will spread on the web is extremely valuable to journalists, content providers, advertisers, and news recommendation systems. This is also important for activists and politicians who are using the web increasingly more to influence public opinion.

However, predicting online popularity of news articles is a challenging task.

First, *context* outside the web is often not readily accessible and elements such as local and geographical conditions and various circumstances that affect the population make this prediction difficult. Furthermore, *network properties* such as the structure of social networks that are propagating the news, influence variations among members, and interplay between different sections of the web add other layers of complexity to this problem. Most significantly, intuition suggests that the *content* of an article must play a crucial role in its popularity. Content that resonates with a majority of the readers such as a major world-wide event can be expected to garner wide attention while specific content relevant only to a few may not be as successful.

Given the complexity of the problem due to the above mentioned factors, a growing number of recent studies [SH10], [LMS10], [TLA11], [KKC11], [LH10] make use of early measurements of an item's popularity to predict its future success. In the present work we investigate a more difficult problem, which is prediction of social popularity without using early popularity measurements, by instead solely considering features of a news article *prior* to its publication. We focus this work on observable features in the content of an article as well as its source of publication. Our goal is to discover if any predictors relevant only to the content exist and if it is possible to make a reasonable forecast of the spread of an article based on content features.

The news data for our study was collected from Feedzilla [1] –a news feed aggregator– and measurements of the spread are performed on Twitter [2], an immensely popular microblogging social network. Social popularity for the news articles are measured as the number of times a news URL is posted and shared on Twitter.

To generate features for the articles, we consider four different characteristics

[1] www.feedzilla.com
[2] www.twitter.com

of a given article. Namely:

- The news source that generates and posts the article
- The category of news this article falls under
- The subjectivity of the language in the article
- Named entities mentioned in the article

We quantify each of these characteristics by a score making use of different scoring functions. We then use these scores to generate predictions of the spread of the news articles using regression and classification methods. Our experiments show that it is possible to estimate ranges of popularity with an overall accuracy of 84% considering only content features. Additionally, by comparing with an independent rating of news sources, we demonstrate that there exists a sharp contrast between traditionally popular news sources and the top news propagators on the social web.

In the next section we provide a survey of recent literature related to this work. Section 3 describes the dataset characteristics and the process of feature score assignment. In Section 4 we will present the results of prediction methods. Finally, in Section 5 we will conclude the paper and discuss future possibilities for this research.

3.2 Related Work

Stochastic models of information diffusion as well as deterministic epidemic models have been studied extensively in an array of papers, reaffirming theories developed in sociology such as diffusion of innovations [Rog95]. Among these are models of viral marketing [LAH07b], models of attention on the web [WH07], cascading behavior in propagation of information [GLG04] [LMF07] and models that describe

heavy tails in human dynamics [VOD06]. While some studies incorporate factors for content *fitness* into their model [SR08], they only capture this in general terms and do not include detailed consideration of content features.

[SDW06] performed a controlled experiment on music, comparing quality of songs versus the effects of social influence[SDW06]. They found that song quality did not play a role in popularity of highly rated songs and it was social influence that shaped the outcome. The effect of user influence on information diffusion motivates another set of investigations [KKT03b], [CHK10],[ALT08], [LH10].

On the subject of news dissemination, [LBK09] and [YL11] study temporal aspects of spread of news memes online, with [LG10b] empirically studying spread of news on the social networks of digg and twitter and [SRM09] studying facebook news feeds.

A growing number of recent studies predict spread of information based on early measurements (using early votes on digg, likes on facebook, click-throughs, and comments on forums and sites). [SH10] found that eventual popularity of items posted on youtube and digg has a strong correlation with their early popularity; [LMS10], [JR09] and [TLA11] predict the popularity of a thread using features based on early measurements of user votes and comments. [KKC11] propose the notion of virtual temperature of weblogs using early measurements. [LH10] predict digg counts using stochastic models that combine design elements of the site -that in turn lead to collective user behavior- with information from early votes.

Finally, recent work on variation in the spread of content has been carried out by [RMK11] with a focus on categories of twitter hashtags (similar to keywords). This work is aligned with ours in its attention to importance of content in variations among popularity, however they consider categories only, with news being one of the hashtag categories. [YCK11] conduct similar work on social marketing messages.

3.3 Data and Features

This section describes the data, the feature space, and feature score assignment in detail.

3.3.1 Dataset Description

Data is comprised of a set of news articles published on the web within a defined time period and the number of times each article was shared by a user on Twitter after publication. This data was collected in two steps: first, a set of articles were collected via a news feed aggregator, then the number of times each article was linked to on twitter was found. In addition, for some of the feature scores, we used a 50-day history of posts on twitter. The latter will be explained in section 3.3.2 on feature scoring.

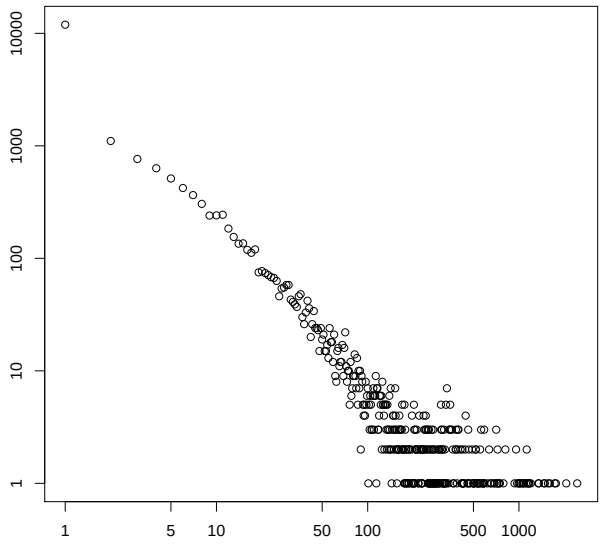


Figure 3.1: Log-log distribution of tweets.

Online news feed aggregators are services that collect and deliver news articles as they are published online. Using the API for a news feed aggregator named Feedzilla, we collected news feeds belonging to all news articles published online

during one week (August 8th to 16th, 2011) which comprised 44,000 articles in total. The feed for an article includes a title, a short summary of the article, its url, and a time-stamp. In addition, each article is pre-tagged with a category either provided by the publisher or in some manner determined by Feedzilla. A fair amount of cleaning was performed to remove redundancies, resolve naming variations, and eliminate spam through the use of automated methods as well as manual inspection. As a result over 2000 out of a total of 44,000 items in the data were discarded.

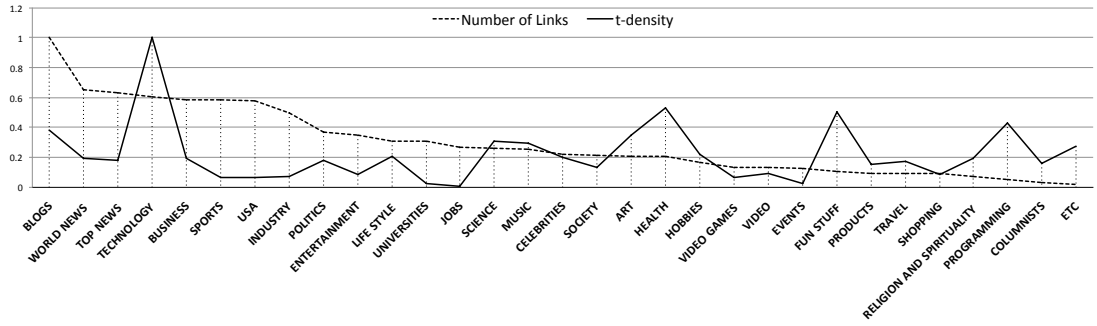


Figure 3.2: Normalized values for t-density per category and links per category

The next phase of data collection was performed using Topsy[3] , a Twitter search engine that searches all messages posted on Twitter. We queried for the number of times each news link was posted or reshared on Twitter (tweeted or retweeted). Earlier research [LBK09] on news meme buildup and decay suggest that popular news threads take about 4 days until their popularity starts to plateau. Therefore, we allowed 4 days for each link to fully propagate before querying for the number of times it has been shared.

The first half of the data was used in category score assignment (explained in the next section). The rest we partitioned equally into 10,000 samples each for training and test data for the classification and regression algorithms. Figure 3.1 shows the log distribution of total tweets over all data, demonstrating a long

[3]http://topsy.com

tail shape which is in agreement with other findings on distribution of Twitter information cascades [ZBK10]. The graph also shows that articles with zero tweets lie outside of the general linear trend of the graph because they did not propagate on the Twitter social network.

Our objective is to design features based on content to predict the number of tweets for a given article. In the next section we will describe these features and the methods used to assign values or scores to features.

3.3.2 Feature Description and Scoring

Choice of features is motivated by the following questions: Does the category of news affect its popularity within a social network? Do readers prefer factual statements or do they favor personal tone and emotionally charged language? Does it make a difference whether famous names are mentioned in the article? Does it make a difference who publishes a news article?

These questions motivate the choice of the following characteristics of an article as the feature space: the category that the news belongs to (e.g. politics, sports, etc.), whether the language of the text is objective or subjective, whether (and what) named entities are mentioned, and what is the source that published the news. These four features are chosen based on their availability and relevance, and although it is possible to add any other available features in a similar manner, we believe the four features chosen in this paper to be the most relevant.

We would like to point out that we use the terms article and link interchangeably since each article is represented by its URL link.

3.3.2.1 Category Score

News feeds provided by Feedzilla are pre-tagged with category labels describing the content. We adopted these category labels and designed a score for them

which essentially represents a prior disribution on the popularity of categories. Figure 3.2 illustrates the prominence of each category in the dataset. It shows the number of links published in each category as well as its success on Twitter represented by the average tweet per link for each category. We call the average tweet per link the *t-density* and we will use this measure in score assignments for some other features as well.

$$\text{t-density} = \frac{\text{Number of Tweets}}{\text{Number of Links}}$$

Observe in Figure 3.2 that news related to Technology receives more tweets on average and thus has a more prominent presence in our dataset and most probably on twitter as a whole. Furthermore, we can see categories (such as Health) with low number of published links but higher rates of t-density (tweet per link). These categories perhaps have a niche following and loyal readers who are intent on posting and retweeting its links.

We use t-density to represent the prior popularity for a category. In order to assign a t-density value (i.e. score) to each category, we use the first 22,000 points in the dataset to compute the average tweet per article link in that category.

3.3.2.2 Subjectivity

Another feature of an article that can affect the amount of online sharing is its language. We want to examine if an article written in a more emotional, more personal, and more subjective voice can resonate stronger with the readers. Accordingly, we design a binary feature for subjectivity where we assign a zero or one value based on whether the news article or commentary is written in a more subjective voice, rather than using factual and objective language. We make use of a subjectivity classifier from LingPipe [Ali08] a natural language toolkit using machine learning algorithms devised by [PL04]. Since this requires training data, we use transcripts from well-known tv and radio shows belonging to Rush Lim-

baugh[4] and Keith Olberman[5] as the corpus for subjective language. On the other hand, transcripts from CSPAN[6] as well as the parsed text of a number of articles from the website FirstMonday[7] are used as the training corpus for objective language. The above two training sets provide a very high training accuracy of 99% and manual inspection of final results confirmed that the classification was satisfactory. Figure 3.3 illustrates the distribution of average subjectivity per source, showing that some sources consistently publish news in a more objective language and a somwhat lower number in a more subjective language.

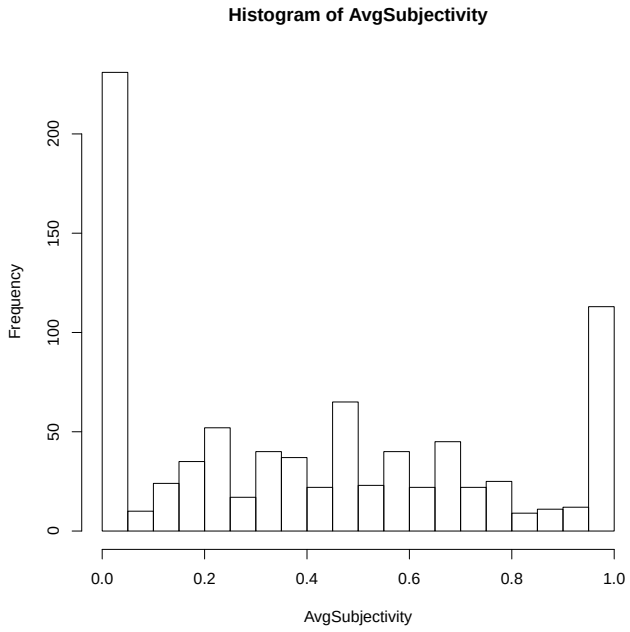


Figure 3.3: Distribution of average subjectivity of sources.

3.3.2.3 Named Entities

In this paper, a named entity refers to a known place, person, or organization. Intuition suggests that mentioning well-known entities can affect the spread of an

[4]http://www.rushlimbaugh.com
[5]http://www.msnbc.msn.com/id/32390086
[6]http://www.c-span.org
[7]http://firstmonday.org

article, increasing its chances of success. For instance, one might expect articles on Obama to achieve a larger spread than those on a minor celebrity. And it has been well documented that fans are likely to share almost any content on celebrities like Justin Bieber, Oprah Winfrey or Ashton Kutcher. We made use of the Stanford-NER[8] entity extraction tool to extract all the named entities present in the title and summary of each article. We then assign scores to over 40,000 named entities by studying historical prominence of each entity on twitter over the timeframe of a month. The assigned score is the average t-density (as defined in section 3.3.2.1) of each named entity. To assign a score for a given article we use three different values: the number of named entities in an article, the highest score among all the named entities in an article, and the average score among the entities.

3.3.2.4 Source Score

The data includes articles from 1350 unique sources on the web. We assign scores to each source based on the historical success of each source on Twitter. For this purpose, we collected the number of times articles from each source were shared on Twitter in the past. We used two different scores, first the aggregate number of times articles from a source were shared, and second the t-density of each source which as defined in 3.3.2.1 is computed as the number of tweets per links belonging to a source. The latter proved to be a better score assignment compared to the aggregate.

To investigate whether it is better to use a smaller portion of more recent history, or a larger portion going back farther in time and possibly collecting outdated information, we start with the two most recent weeks prior to our data collection and increase the number of days, going back in time. Figure 3.5 shows the trend of correlation between the t-density of sources in historical data and

[8]http://nlp.stanford.edu/software/CRF-NER.shtml

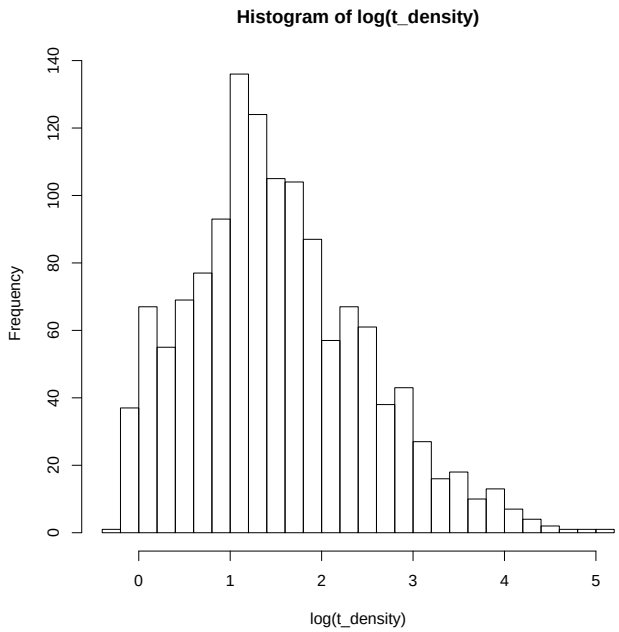


Figure 3.4: Distribution of log of source t-density scores over collected data. Log transformation was used to normalize the score further.

their t-density in our dataset. We observe that the correlation increases with more datapoints from the history until it begins to plateau near 50 days. Using this result, we take 54 days of history prior to the first date in our dataset. We find that the correlation of the assigned score found in the above manner has a correlation of 0.7 with the t-density of the dataset. Meanwhile, the correlation between the source score and number of tweets of any given article is 0.35, suggesting that information about the source of publication alone is not sufficient in predicting popularity. Figure 3.4 shows the distribution of log of source scores (t-density score). Taking the log of source scores produces a more normal shape, leading to improvements in regression algorithms.

We plot the timeline of t-densities for a few sources and find that t-density of a source can vary greatly over time. Figure 3.6 shows the t-density values belonging to the technology blog *Mashable* and *Blog Maverick*, a weblog of prominent entrepreneur, Mark Cuban. The t-density scores corresponding to each of these

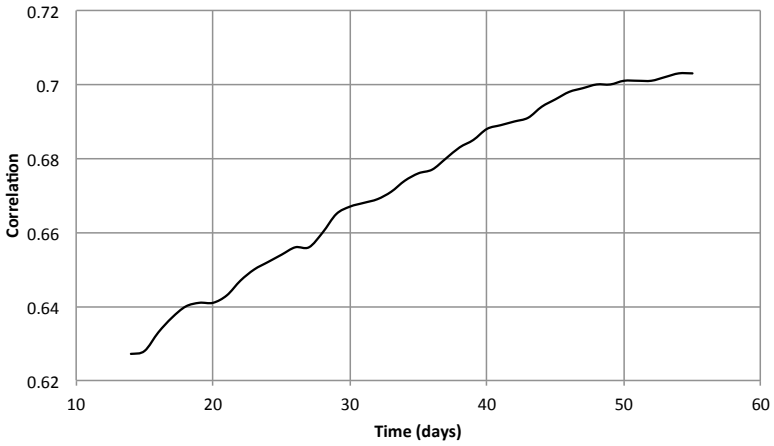


Figure 3.5: Correlation trend of source scores with t-density in data. Correlation increases with more days of historical data until it plateaus after 50 days.

sources are 74 and 178 respectively. However, one can see that *Mashable* has a more consistent t-density compared to *Blog Maverick*.

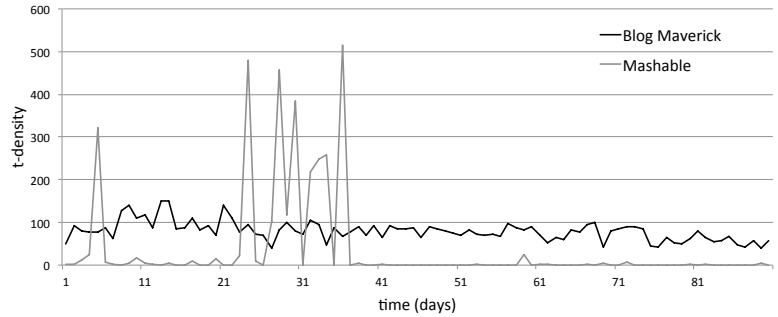


Figure 3.6: Timeline of t-density (tweet per link) of two sources.

In order to improve the score to reflect consistency we devise two methods; the first method is to smooth the measurements for each source by passing them through a low-pass filter. Second is to weight the score by the percentage of times a source's t-density is above the mean t-density over all sources, penalizing sources that drop low too often. The mean value of t-densities over all sources is 6.4. Figure 3.8 shows the temporal variations of tweets and links over all sources. Notice that while both tweets and links have a weekly cycle due to

natural variations in web activity, the t-density score is robust to periodic weekly variations. The non-periodic nature of t-density indicates that the reason we see less tweets during down time is mainly due to the fact that less links are posted.

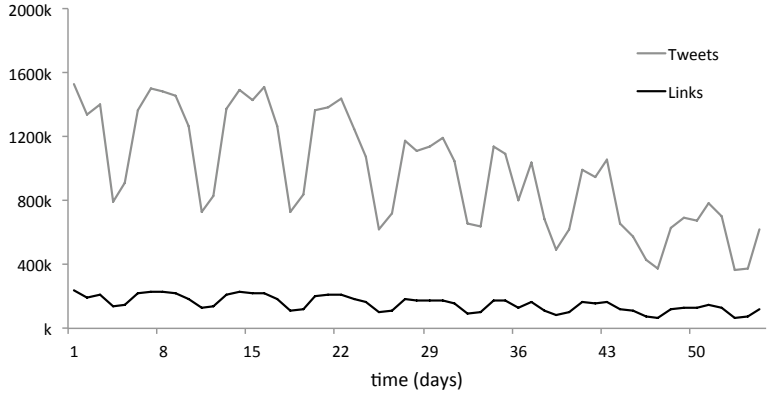


Figure 3.7: Temporal variations of tweets and links over all sources

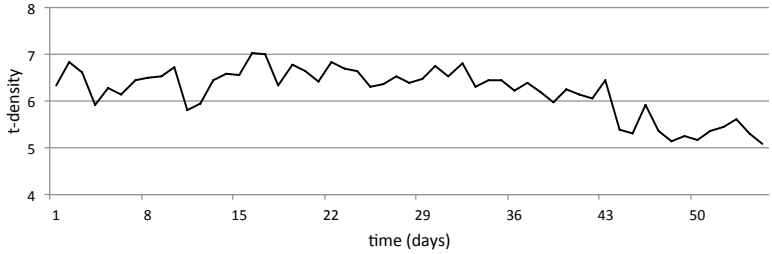


Figure 3.8: Temporal variations of t-density over all sources

3.3.2.5 Are top traditional news sources the most propagated?

As we assign scores to sources in our dataset, we are interested to know whether sources successful in this dataset are those that are conventionally considered prominent.

Google News[9] is one of the major aggregators and providers of news on the web. While inclusion in Google news results is free, Google uses its own criteria

[9]http://news.google.com/

to rank the content and place some articles on its homepage, giving them more exposure. Freshness, diversity, and rich textual content are listed as the factors used by Google News to automatically rank each article as it is published. Because Google does not provide overall rankings for news sources, to get a rating of sources we use NewsKnife[10]. NewsKnife is a service that rates top news sites and journalists based on analysis of article's positions on the Google news homepage and sub-pages internationally. We would like to know whether the sources that are featured more often on Google news (and thus deemed more prominent by Google and rated more highy by NewsKnife) are also those that become most popular on our dataset.

	Total Links	Total Tweets	t-density
Correlation	0.57	0.35	-0.05

Table 3.1: Correlation values between NewsKnife source scores and their performance on twitter dataset.

Accordingly we measure the correlation values for the 90 top NewsKnife sources that are also present in our dataset. The values are shown in Table 3.1. It can be observed that the ratings correlate positively with the number of links published by a source (and thus the sum of their tweets), but have no correlation (-0.05) with t-density which reflects the number of tweets that each of their links receives. For our source scoring scheme this correlation was about 0.7.

Table 3.2 shows a list of top sources according to NewsKnife, as well as those most popular sources in our dataset. While NewsKnife rates more traditionally prominent news agencies such as Reuters and the Wall Street Journal higher, in our dataset the top ten sources (with highest t-densities) include sites such as Mashable, AllFacebook (the unofficial facebook blog), the Google blog, marketing

[10]http://www.newsknife.com

blogs, as well as weblogs of well-known people such as Seth Godin's weblog and Mark Cuban's blog (BlogMaverick). It is also worth noting that there is a bias toward news and opinion on web marketing, indicating that these sites actively use their own techniques to increase their visibility on Twitter.

While traditional sources publish many articles, those more successful on the social web garner more tweets. A comparison shows that a NewsKnife top source such as The Christian Science Monitor received an average of 16 tweets in our dataset with several of its articles not getting any tweets. On the other hand, Mashable gained an average of nearly 1000 tweets with its least popular article still receiving 360 tweets. Highly ranked news blogs such as The Huffington Post perform relatively well in Twitter, possibly due to their active twitter accounts which share any article published on the site.

NewsKnife	*Reuters, Los Angeles Times, New York Times, Wall Street Journal, USA Today, Washington Post, ABC News, Bloomberg, Christian Science Monitor, BBC News*
Twitter Dataset	*Blog Maverick, Search Engine Land, Duct-tape Marketing, Seth's Blog, Google Blog, Allfacebook, Mashable, Search Engine Watch*

Table 3.2: Highly rated sources on NewsKnife versus those popular on the Twitter dataset

3.4 Prediction

In this work, we evaluate the performance of both regression and classification methods to this problem. First, we apply regression to produce exact values of

tweet counts, evaluating the results by the R-squared measure. Next we define popularity classes and predict which class a given article will belong to. The following two sections describe these methods and their results.

Variable	Description
S	Source t-density score
C	Category t-density score
$Subj$	Subjectivity (0 or 1)
Ent_{ct}	Number of named entities
Ent_{max}	Highest score among named entities
Ent_{avg}	Average score of named entities

Table 3.3: Feature set (prediction inputs). *t-density* refers to average tweet per link.

3.4.1 Regression

Once score assignment is complete, each point in the data (i.e. a given news article) will correspond to a point in the feature space defined by its category, subjectivity, named entity, and source scores. As described in the previous section, category, source, and named entity scores take real values while the subjectivity score takes a binary value of 0 or 1. Table 3.3 lists the features used as inputs of regression algorithms. We apply three different regression algorithms - linear regression, k-nearest neighbors (KNN) regression and support vector machine (SVM) regression.

Since the number of tweets per article has a long-tail distribution (as discussed previously in Figure 3.1), we performed a logarithmic transformation on the number of tweets prior to carrying out the regression. We also used the log of source and category scores to normalize these scores further. Based on this transforma-

	Linear Regression	SVM Regression
All Data	0.34	0.32
Tech Category	0.43	0.36
Within Twitter	0.33	0.25

Table 3.4: Regression Results (R^2 values)

tion, we reached the following relationship between the final number of tweets and feature scores.

$$ln(T) = 1.24ln(S) + 0.45ln(C) + 0.1Ent_{max} - 3$$

where T is the number of tweets, S is the source t-density score, C is the category t-density score, and Ent_{max} is the maximum t-density of all entities found in the article. Equivalently,

$$T = S^{1.24}C^{0.45}e^{-(0.1Ent_{max}+3)}$$

with coefficient of determination $R^2 = 0.258$. All three predictors in the above regression were found to be significant. Note that the R^2 is the coefficient of determination and relates to the mean squared error and variance:

$$R^2 = 1 - \frac{MSE}{VAR}$$

Alternatively, the following model provided improved results:

$$T^{0.45} = (0.2S - 0.1Ent_{ct} - 0.1Ent_{avg} + 0.2Ent_{max})^2$$

with an improved $R^2 = 0.34$. Using support vector machine (SVM) regression [CL11], we reached similar values for R^2 as listed in Table 3.4.

In K-Nearest Neighbor Regression, we predict the tweets of a given article using values from it's nearest neighbors. We measure the Euclidean distance between two articles based on their position in the feature space [HTF08]. Parameter K

specifies the number of nearest neighbors to be considered for a given article. Results with $K = 7$ and $K = 3$ for a 10k test set are R-sq= 0.05, with mean squared error of 5101.695. We observe that KNN performs increasingly more poorly as the dataset becomes larger.

3.4.1.1 Category-specific prediction

One of the weakest predictors in regression was the Category score. One of the reasons for this is that there seems to be a lot of overlap across categories. For example, one would expect *World News* and *Top News* to have some overlap, or the category *USA* would feature articles that overlap with others as well. So the categories provided by Feedzilla are not necessarily disjoint and this is the reason we observe a low prediction accuracy.

To evaluate this hypothesis, we repeated the prediction algorithm for particular categories of content. Using only the articles in the Technology category, we reached an R^2 value of 0.43, indicating that when employing regression we can predict the popularity of articles within one category (i.e. Technology) with better results.

3.4.2 Classification

Feature scores derived from historical data on Twitter are based on articles that have been tweeted and not those articles which do not make it to Twitter (which make up about half of the articles). As discussed in Section 3.3.1 this is evident in how the zero-tweet articles do not follow the linear trend of the rest of datapoints in Figure 3.1. Consequently, we do not include a zero-tweet class in our classification scheme and perform the classification by only considering those articles that were posted on twitter.

Table 3.5 shows three popularity classes A (1 to 20 tweets), B (20 to 100

tweets), C (more than 100) and the number of articles in each class in the set of 10,000 articles. Table 3.6 lists the results of support vector machine (SVM) classification, decision tree, and bagging [HFH09] for classifying the articles. All methods were performed with 10-fold cross-validation. We can see that classification can perform with an overall accuracy of 84% in determining whether an article will belong to a low-tweet, medium-tweet, or high-tweet class.

In order to determine which features play a more significant role in prediction, we repeat SVM classification leaving one of the features out at each step. We found that publication source plays a more important role compared to other predictors, while subjectivity, categories, and named entities do not provide much improvement in prediction of news popularity on Twitter.

3.4.2.1 Predicting Zero-tweet Articles

We perform binary classification to predict which articles will be at all mentioned on Twitter (zero tweet versus nonzero tweet articles). Using SVM classification we can predict –with 66% accuracy– whether an article will be linked to on twitter or whether it will receive zero tweets. We repeat this operation by leaving out one feature at a time to see a change in accuracy. We find that the most significant feature is the source, followed by its category. Named entities and subjectivity did not provide more information for this prediction. So despite one might expect, we find that readers overall favor neither subjectivity nor objectivity of language in a news article. It is interesting to note that while category score does not contribute in prediction of popularity within Twitter, it does help us determine whether an article will be at all mentioned on this social network or not. This could be due to a large bias toward sharing technology-related articles on Twitter.

Class name	Range of tweets	Number of articles
A	1–20	7,600
B	20–100	1,800
C	100–2400	600

Table 3.5: Article Classes

Method	Accuracy
Bagging	83.96%
J48 Decision Trees	83.75%
SVM	81.54%
Naive Bayes	77.79%

Table 3.6: Classification Results

3.5 Discussion and Conclusion

This work falls within the larger vision of studying how attention is allocated on the web. There exists an intense and fast paced competition for attention among news items published online and we examined factors within the content of articles that can lead to success in this competition. We predicted the popularity of news items on Twitter using features extracted from the content of news articles. We have taken into account four features that cover the spectrum of the information that can be gleaned from the content - the source of the article, the category, subjectivity in the language and the named entities mentioned. Our results show that while these features may not be sufficient to predict the exact number of tweets that an article will garner, they can be effective in providing a range of popularity for the article on Twitter. More precisely, while regression results were not adequate, we achieved an overall accuracy of 84% using classifiers. It is

important to bear in mind that while it is intriguing to pay attention to the most popular articles –those that become viral on the web– a great number of articles spread in medium numbers. These medium levels can target highly interested and informed readers and thus the mid-ranges of popularity should not be dismissed.

Interestingly we have found that in terms of number of retweets, the top news sources on twitter are not necessarily the conventionally popular news agencies and various technology blogs such as Mashable and the Google Blog are very widely shared in social media. Overall, we discovered that one of the most important predictors of popularity was the source of the article. This is in agreement with the intuition that readers are likely to be influenced by the news source that disseminates the article. On the other hand, the category feature did not perform well. One reason for this is that we are relying on categories provided by Feedzilla, many of which overlap in content. Thus a future task is to extract categories independently and ensure little overlap.

Combining other layers of complexity described in the introduction opens up the possibility of better prediction. It would be interesting to further incorporate interaction between offline and online media sources, different modes of information dissemination on the web, and network factors such as the influence of individual propagators.

CHAPTER 4

Network and Content Summaries

4.1 Introduction

Fostering spaces for discussion and exchange of ideas is one of the central functions of the web. Discussion forums, social network messages, youtube comments, and social news services are examples of these spaces and the study of characteristics and dynamics of these environments has fueled an increasing amount of research in recent years.

While there is a common mental and emotional layer driving users to interact with content in various ways (user-content relations), online spaces also foster connection and interpersonal relationships (user-user relations). These two processes work together to create a dynamic environment of conversations where different topics become prominent at different times, are talked about in different ways, and where user friendships may resonate with emergence of some themes in the topic domain. Therefore, a fundamental need in the study of such spaces is to have at our disposal a fully automated method that provides a comprehensive summary of the dynamics of conversations without being cumbersome. In this chapter we achieve this goal using a multi-layered approach, considering both the temporal variations in content as well as friendship connections in the network. Since these dynamics can be observed at different granularities, we will also use a multi-scale approach utilizing different tools at different scales to reach a meaningful picture of these dynamics.

4.1.1 Related Work

A number of studies relevant to the current work center on mining and tracking opinions in product review websites.The goal of this family of literature is to extract summaries of online reviews, track user sentiment, or compare products (some examples are [LHC05] [GPL06] and [DLP03]). In contrast with the data in the current chapter, product reviews are often more structured, and there are known features of a product (such as the resolution of a camera) which users express positive or negative sentiment with respect to, so extracting and tracking feature-based opinion and sentiment is the focus of this family of studies.

Tracking topics, detecting events, and creating summaries of news content is the subject of another set of studies (e.g. [AHE11]). News datasets are often curated and tagged, and are usually created by experts. In contrast, the current work takes user-generated content in a public forum. Consequently, the data is extremely noisy and users are quite loosely self-organized around certain topics. Therefore the task of indexing and creating a granular summary of content and users becomes more challenging.

Finally, a number of papers aim at tracking changes in content across time by finding topics at consecutive time slots in the data and mapping them together [GHS09] [MZ05] [ABD08] [BL06]. In the current chapter, we instead detect topics over the whole corpus and use these topics to separate all posted content into topic categories. We then dial in to consecutive time-slots and get a more fine-grained perspective using unigram analysis in each of the topically separated categories of content. Although there are recent papers that propose more sophisticated topic evolution methods (e.g. incorporating temporal evolution in the definition of a topic [JHL11]), in the end the current work produces a simpler and more comprehensible summary and thus we believe is more readily usable. Our method demonstrates that simple tools used in proper succession can create a comprehen-

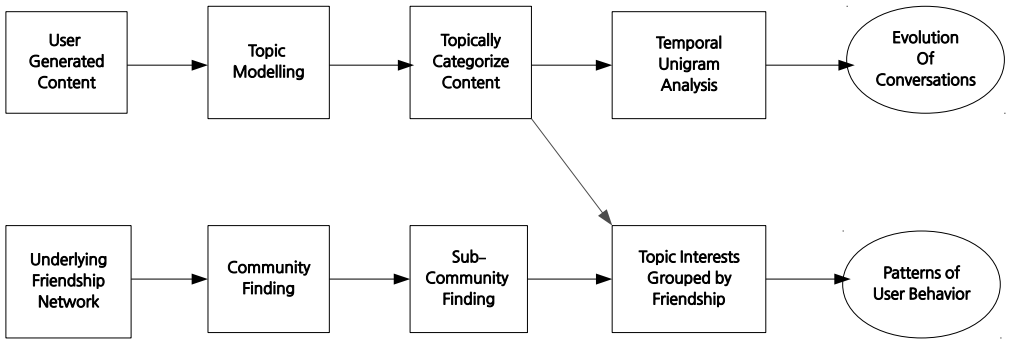


Figure 4.1: Flowchart for general methodology used. Figure shows the interplay of content-driven and socially driven approaches employed to study the evolution of conversations and the patterns of user behavior.

sive multiscale overview of a large forum with noisy data and that one can index this data at a granular level, indexing temporally, socially and content-wise.

4.1.2 Overview and Approach

We propose an automated methodology that provides a granular representation of content over time and reveals patterns of user behavior. The steps of this process are demonstrated in Figure 4.1. Using Latent Dirichlet Allocation topic modeling [BNJ03] on the text of forum posts, we generate a set of distinct themes prevalent in user discussions. These topics establish an initial framework by which we can classify conversations. Within the context of these topics we observe how conversations evolve over time. We find that subsequent sub-topic modeling of each time segment produces an insufficient characterization of conversations. However, unigram analysis of the segments used in conjunction with topic modeling provides the depth and granularity needed to extract meaningful information. This method provides the right amount of detail without becoming too convoluted.

Concurrently, we perform community detection on the friendship network of users and find that there are clear ties between friendship communities and topics, implying that user connections are highly related to common topics of interest. Finally, we find similarities between topics in terms of user communities that participate in them and we find that some topics are highly correlated.

We implemented this methodology on an online platform called *Cafemom*, a forum for mothers to connect and discuss their views on a variety of issues. In this chapter we focus our attention on conversations around vaccination and immunization. Vaccination has become an increasingly prominent topic in the public sphere and speculation about its adverse effects and concerns about safety have been on the rise[BRB10, WSL02, ZWF05]. These concerns range from short-term vaccine side effects to more serious ones such as the much discussed link between immunizations and autism[KLN11, FCB10]. Consequently, public health officials are worried about public opinion leading to a drop in vaccination rates, exposing the population to dangerous epidemics.

Applying topic modeling to this dataset, we found areas such as *Religion, Autism, Government, Birth*, and *Food* with different levels of prominence at different times. Further unigram analysis within each topic created the next level of granularity; for example within the topic of *Government* the method was capable of capturing external events such as the 2008 presidential election as well as the 2011 tsunami in Japan and the resulting nuclear crisis. We then used the Cafemom friendship network to detect communities and sub-communities and found that a heatmap of communities and topics (Fig. 4.4) shows strong correlations among the two. Furthermore, a comparison among topics showed that some topics have positive or negative correlations based on the user communities active in them. For example, *Birth* and *Religion* correlate, whereas *Birth* and *Autism* are inversely correlated (more details in Section 4.5.2).

4.1.3 Outline

The rest of this chapter is organized as follows: in Section 4.2, we describe the dataset, in Section 4.3 we discuss topic modeling and illustrate the temporal variation of topics. In Section 4.4, we compare sub-topic detection with unigram analysis in progressive time windows and show that simple unigram analysis provides more meaningful results at this granularity. In Section 4.5, we find communities in the friendship network and show that there is a high correlation among communities and topics and that some topics are highly correlated based on the communities of users who generate them. Section 4.6 discusses the findings and concludes the chapter.

4.2 Data Characteristics

The dataset for this chapter is obtained from forum posts in cafemom.com, a US-based online space where mothers discuss and exchange ideas on a variety of issues. Cafemom's discussion boards are divided into groups (which are in turn divided into forums containing threads of individual posts), and while some portions are open to the public, a majority of the groups are private. Therefore, to access the complete data we create a membership profile and crawl all data from the discussion groups that appear in a keyword search for the relevant issue, i.e vaccination. We obtain a corpus consisting of 139,457 threads spanning 18 discussion groups with a total of 1,700,086 posts from 27,790 users over a span of around 5 years –Feb 6th 2007 to Apr 24th 2012. During this time, there were a total of 18,498,306 thread views (by users and non-users).

4.3 Topic Generation

We employ Latent Dirichlet Allocation (LDA), an unsupervised method of topic discovery[BNJ03], to generate topics for this dataset . These topics help categorize the threads of the forum into distinct themes, and are the basis by which we study the evolution of user interests and concerns over time. In LDA, each document (in this chapter we use threads as documents) is comprised of a mixture of topics and each word in a document can be ascribed to one of the topics generated. Listed in Table 4.1 are the top words for each of the ten sets of topics in the corpus. Note that topic names are assigned by the authors for the purpose of understandability and they are based on the inspection of the words in each topic. In the next section we will describe the levels of prominence of each topic over time [GS04, ZJZ06].

4.3.1 Topic Characteristics

LDA topic modeling using Mallet [McC02] produces a set of proportions associated with each topic for every document (i.e each thread)[BNJ03]. In other words, for each thread, we have a list of all ten topics in Table 4.1 along with the proportion or strength of each topic in that thread. Using these values, we categorized each thread under a topic in the following manner: In a thread, if the topic with the highest proportion has a proportion greater than 0.3, then the thread is categorized under that topic. The threshold is chosen as 0.3 because all such threads were found to have relatively low proportions for the other topics associated with that thread. 65.63% of the total threads fall under this criteria and are clearly associated most with one topic, and thus are used for further analyses.[1] By only considering the threads that have a high topic proportion, we can map each thread to exactly one of the 10 topics.

[1]In the rest of the chapter , we use only those 91,528 threads containing 1,339,250 posts that have a topic proportion greater than 0.3 for further findings.

Table 4.1: Top 15 words generated for 10 topics found using LDA Topic Modeling.

Topic	Top Words	Topic
0	people god post group life make person agree read things time point understand women good	**Religion and Ethics**
1	love girl watch dog fun good show hair year pretty day life funny thought mom	**Love and Fun**
2	time day son back things kids night put room good sleep thing bed home house	**Day to Day**
3	vaccine vaccines children health autism flu disease research mercury study medical vaccination vaccinate risk cancer	**Vaccination**
4	child kids children parents people life husband family time make feel things mom mother care	**Family**
5	baby group months birth doctor babies time give mother child born hospital moms mom good	**Birth and Babies**
6	food eat water make milk good eating diet foods oil organic drink free buy made	**Food**
7	money home work free pay people make business job time company insurance month team working	**Money and Work**
8	state government people law country public states news case obama american court america world rights health military police president	**Government**
9	autism school son child kids children year good autistic special admin teacher things pdd group great daughter asperger spectrum	**Autism**

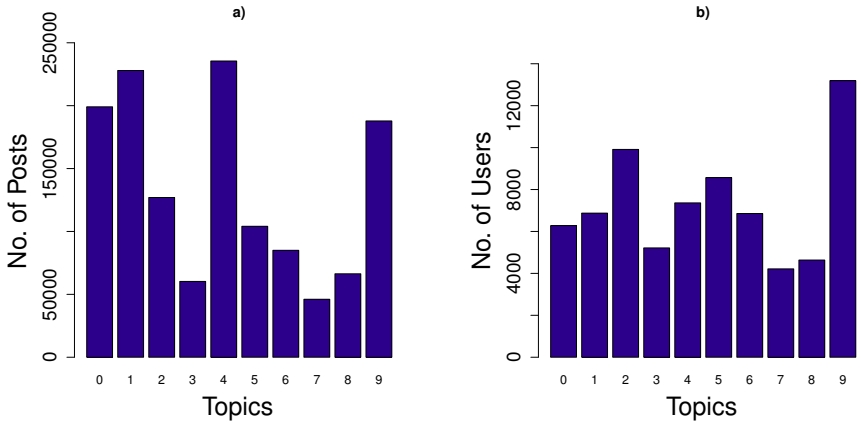


Figure 4.2: a) Number of posts in Cafemom about each topic found in Table 4.1. b) Number of unique Cafemom users who posted on each topic.

Figure 4.2 shows the histograms of number of users and number of posts per topic. It can be observed that topics such as *Love and Fun* and *Family* in general have a greater volume of posts consistent with our intuition about these topics. On the other hand, a larger number of users post in topics such as *Autism*. Examining the growth of conversations over time as shown in Fig. 4.3, we find that topics such as Autism and Vaccination started receiving more attention from early 2007 lasting till 2009. From 2010 onwards, the activity levels declined and remained relatively constant. For *Autism*, there are peaks from around July to October 2007 and peaks around early 2008 for *Vaccination*.

4.4 Evolution of Conversations

Unigram analysis can be used to create meaningful representations of the flow of information over time, especially in discovering the effect of external events on forum conversations. In the following sections we show how unigram analysis of posts categorized under each topic provides a more comprehensive depiction of user conversations than sub-topic modeling over time.

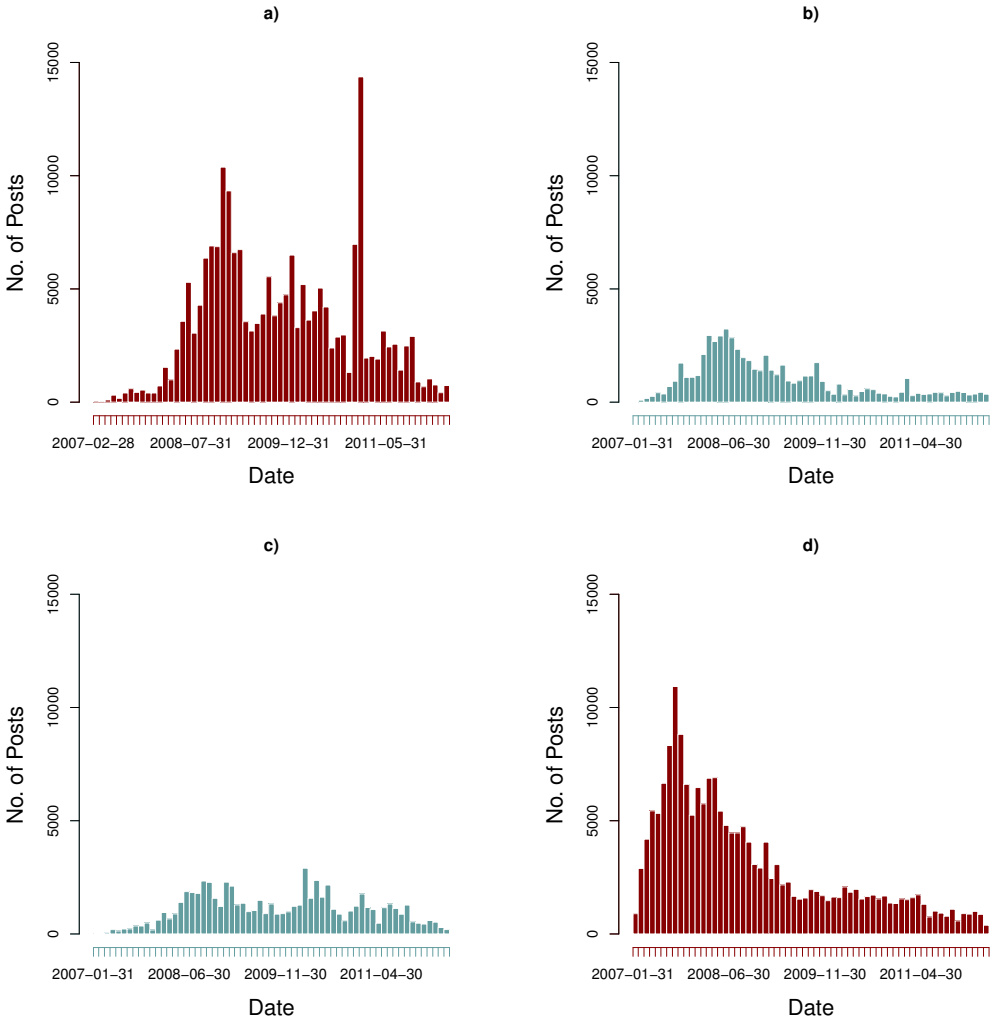


Figure 4.3: Select topics exhibiting variable and stable posting activity. a)Religion. b)Vaccination. c)Government. d)Autism.

4.4.1 Unigram Processing

Tables 4.2 and 4.4 demonstrate the results of unigram analysis on topics of *Government*, and *Money* respectively. As described previously, forum posts are categorized under different topics and divided into 6-month time slots beginning from Feb 6th 2007. For all posts within a time slot, we perform tokenization using appropriate regular expressions, filter out the stop words, and create a bag of words. The term weight $w(t, d)$ for each unigram (or term) t in a time slot d is defined as

$$w(t,d) = \frac{tf(t,d)}{\max_t tf(t,d)} - \frac{tf(t)}{\max_t tf(t)} \quad . \tag{4.1}$$

where $tf(t, d)$ is the term frequency of term t in time slot d, $\max_t tf(t, d)$ is the maximum term frequency of all terms in time slot d, $tf(t)$ is the term frequency of term t in all time slots, and $\max_t tf(t)$ is the maximum term frequency of all terms in all time slots. We then sort the unigrams in the order of decreasing term weight, filter out words that contribute as noise and select the top 20 unigrams for each time slot.

Looking at the results across the entire time span for two of the topics in Tables 4.2 and 4.4, we see many references to major external entities and events. Beginning in the August of 2008 for *Government* (Table 4.2), names of political candidates appear, capturing the Presidential Elections of 2008. Then in the first half of 2009, the discussion shifts to the topic of swine flu epidemic and the health issues relevant to the pandemic at that time. In Table 4.4, terms related to numerous major corporations and organizations such as Verizon, Walmart, and the Food and Drug Administration (FDA) are cited. Furthermore, concerns among moms about finance reflect the economic downturn when words such as poor and bankruptcy gain strength around the end of 2011, lasting till early 2012.

One can see that this simple yet fully automated approach provides a picture of the prominence of issues during different time periods while also establishing the context and showing how different topics are talked about.

4.4.2 Sub-Topic Modeling Vs Unigram Analysis

Here we demonstrate the advantages of using unigrams over sub-topics to study the evolution of user generated content. We perform further topic modeling on content categorized under each topic at every time period and find several drawbacks. Table 4.3 compares these two methods during the time period –February to August 2011– for the topic of *Government*. The table lists the top 5 words in each of the five sub-topics (amounting to a total of 25 words). We can see that these 25 words not only have a great deal of overlap, but also bear no value in providing a concrete sense of what is being discussed. In contrast, the top 20 unigrams provide a much more detailed and diverse account of discussions during that period. Therefore, if we wish to create an efficient summary of the forums with as little human involvement as possible, the unigram approach is superior. We immediately see that within the topic of *Government*, the users were discussing issues of sex, abortion, and Japan's nuclear crisis (a significant external event that happened during that time period).

Producing more sub-topics (e.g 10 sub-topics instead of 5) in each time window and considering more words in each sub-topic (e.g top 20 words instead of top 5) will produce more reasonable results for sub-topic modeling method. However, this would require the study of an order of magnitude greater number of words (e.g 200 words) per time slot in order to extract any meaningful results. In contrast, considering even the top 10 words of the unigram analysis provides a picture that correlate well with external events.

Our methodology provides a simple, yet descriptive view of matters important to users. There are two main inferences drawn through these granular findings: (1) Study of temporal trends of references to external entities and the study of their recurrence and prominence, highlight the importance of latent administrative and governing bodies. (2) The interplay and intersection of topics of interest such as

health, education, finance, politics, and law as evident from Table(s) 4.2 and 4.4 are indicators of the complexity with which certain topics behave on discussion forums.

4.5 Friendship Network Communities

In addition to the online discussion boards, Cafemom has an underlying friendship network. Out of 27,790 users, 16,731 (60% of all users) have friends on the site, which forms the underlying friendship network in our dataset. We use a greedy agglomerative community detection approach to cluster users in our network dataset. [CNM04a, For10b].

The method used for community detection (described in [CNM04a]) optimizes the *modularity* –a measure of the distinctness of communities– across the entire network. The vertices (users) are clustered dendrogrammatically, with each vertex initially categorized as its own community. The communities are then iteratively joined until modularity is maximized[CNM04a]. The algorithm uses data structures suited specifically for sparse networks, making it an efficient clustering algorithm for a friendship network of this size. We wish to see how the community structure relates to user content in the context of topic modeling. The method of partitioning networks into sub-networks and the identification of themes based on unique characteristics have also been employed in other fields such those of neural networks [BRB05].

4.5.1 Network Characteristics

After performing community detection on 16,731 users, we eliminate users belonging to communities having sizes less than 100, leaving us with 15,332 users. Community detection on this set of users produces 88 communities with a modularity of 0.5. We perform sub-community detection on the 5 biggest communities

Table 4.2: Timeline of unigrams of content categorized under topic- 'Government'.

Feb '07–Aug '07	Autism, Bill, Brigid, Children, Education, Vaccines, Immunization, Conference, California, Mercury, Alert
Aug '07–Feb '08	Autism, School, Press, State, Health, Medical, Vaccines, Services, Special, Education, Law
Feb '08–Aug '08	Autism, Drugs, Medical, Vaccine, Savage, Health, Marijuana, Hemp, Legal, California, FDA
Aug '08–Feb '09	Obama, McCain, Palin, Bush, President, Act, Vaccine, War, Iraq, Health, Vote, Tax, Campaign, FDA
Feb '09–Aug '09	Autism, Swine Flu, Mexico, Health, Rights, States, North, Public, Illegal, Virus, Gun, Military, Ticker
Aug '09–Feb '10	CPS, Health, Swine Flu, Emergency, H1N1, Altamira, School, Pandemic, Prison, Haiti, Nascar, Pot
Feb '10–Aug '10	Israel, Oil, System, Land, Einstein, Fetus, Ronald Reagan, Immigration, Palestinians, Abortion
Aug '10–Feb '11	CPS, School, CMSD, Political, Slavery, County, Smoke, CCDCFS, Black, Court, Book, South, Separation
Feb '11–Aug '11	Gun, Home, Women, Japan, Abortion, Radiation, Death, Police, Sex, Nuclear, Reactor, Scientology, Water, Jail
Aug '11–Feb '12	Police, Ticket, Speed, Religious, Limit, Student, Traffic, Sticker, Afraid, File
Feb '12–May '12	Exemption, Religious, Immunization, School, Gov, Santorum, State, Law, Hospital, Zimmerman, Board, Medical

Table 4.3: Comparison of sub-topics with unigrams for the time period Feb 2011 to Aug 2011 for content categorized under topic- Government. Sub-topic modeling exhibits less granularity and clarity of important information aspects due to formation of overlapping topic clusters.

Sub-Topic	Top 5 words
0	case home child gun court
1	news found time water students
2	state public government system school
3	people states slavery food south
4	people country time things war

Unigrams

Gun, Home, Women, Japan, Abortion, Radiation, Death, Police, Sex, Nuclear, Reactor, Scientology, Water, Jail

to break them down into smaller communities having sizes less than 1000 to make all communities comparable in size. The top 5 biggest communities have sizes 4030, 3508, 2572, 2546 and 1314 respectively. Further community detections on these 5 large communities yield modularities 0.53, 0.47, 0.65, 0.39 and 0.77 respectively. Our aim was to break down larger communities into smaller chunks in order to find more meaningful groups.

4.5.2 Communities and Topics

We choose communities with sizes greater than 100 for topic tagging. There are 33 such communities comprising of 11,365 users. To tag communities based on the topic most discussed by that group of users, we calculate a weight for each topic belonging to a community. Every community has users who post in different topics. We assign each topic a count 1 if a user from that community posted for

Table 4.4: Unigrams timeline for content categorized under topic- 'Money and Work'.

Feb '07–Aug '07	SSI, Autism, Income, Medicaid, Disability, Qualify, Insurance, Applied, Family, Job
Aug '07–Feb '08	Walmart, Autism, SSI, United, Tupperware, Family, Account, Medicaid, Middot, Therapy, PCP, Medical
Feb '08–Aug '08	Business, GBG, Ameriplan, Prosperity, Product, Downline, Training, Wellness, Opportunity, Vitamins
Aug '08–Feb '09	Tally, Free, Secret, Parties, Work, Candles, Ebay, Risk, Inventory, Vacci, Trial, Consultant, Woomer, United, Shopper
Feb '09–Aug '09	Pay, Income, Food, Job, Bills, Free, Check, Insurance, Avon, Account, Tax
Aug '09–Feb '10	Insurance, Pay, Medicaid, Welfare, Health, Bill, Private, Services, Credit, Taxes, Food, Money, Afford, Cover, Tip, EIC
Feb '10–Aug '10	Pay, Job, School, Property, Necessity, Grocery, Arbonne, Tip, Products, House, Unemployment, Food
Aug '10–Feb '11	Baskets, Gift, Moms, Sales, Card, Money, Internet, Free, Buy, Gold, Cards, Training, Holiday, Debt
Feb '11–Aug '11	Tax, Money, House, Food, Job, Tip, Stamps, Kids, Welfare, Credit, Service, Loans, Car
Aug '11–Feb '12	Tax, School, Poor, Job, Wealth, Email, Country, Walmart, Military, Rich, Facebook, Mortgage
Feb '12–May '12	Exemption, Religious, Immunization, School, Gov, Santorum, State, Law, Hospital, Zimmerman, Board, Medical

that topic and 0 if not. Topic weight $w(t, g)$ is defined as

$$w(t, g) = \frac{tc(t, g)}{\max_t tc(t, g)} - \frac{tc(t)}{\max_t tc(t)} \quad . \tag{4.2}$$

where $tc(t, g)$ is the topic count for topic t in community g, $\max_t tc(t, g)$ is the maximum topic count of all topics in community g, $tc(t)$ is the topic count for topic t in all communities, and $\max_t tc(t)$ is the maximum topic count of all topics in all communities.

Comparing the topic scores within the community, we are able to identify the most popular topics for that community. Comparison of the topic scores among different communities (communities and sub-communities) provides a clear picture of the topic prominence for each community. Figure 4.4 is a heatmap generated for these 33 communities and shows which topics are more prevalent in a community. We find 15 communities that discuss *Autism* more than any other topic. Similarly all sub-communities for community 1 (10 communities) discuss *Birth and Babies* more than anything else. From these findings we can speculate that friends on Cafemom share strong similarities around topics of interests. In fact, related work on user similarity suggests that these characteristics also affect user evaluations of each other. Anderson et al. [AHK12] analyze these evaluations in terms of Wikipedia promotions, and votes on user-created content in Epinions and Stack Overflow.

Finally, we investigate the correlation between topics based on preference among different communities to discuss them. We calculate the pairwise correlation for all topics as follows. For each topic, we take a vector of its weights $w(t, g)$ over all communities. We compute the correlation matrix for these topics as $cor(u, v)$ where u and v are topic weight vectors. Table 4.5 shows the computed correlation matrix. Most notably, communities that post most often in *Birth and Babies* also post more in *Religion and Ethics*, with a 0.91 correlation index, and post the least in *Autism* (-0.22). The strong correlation between these two top-

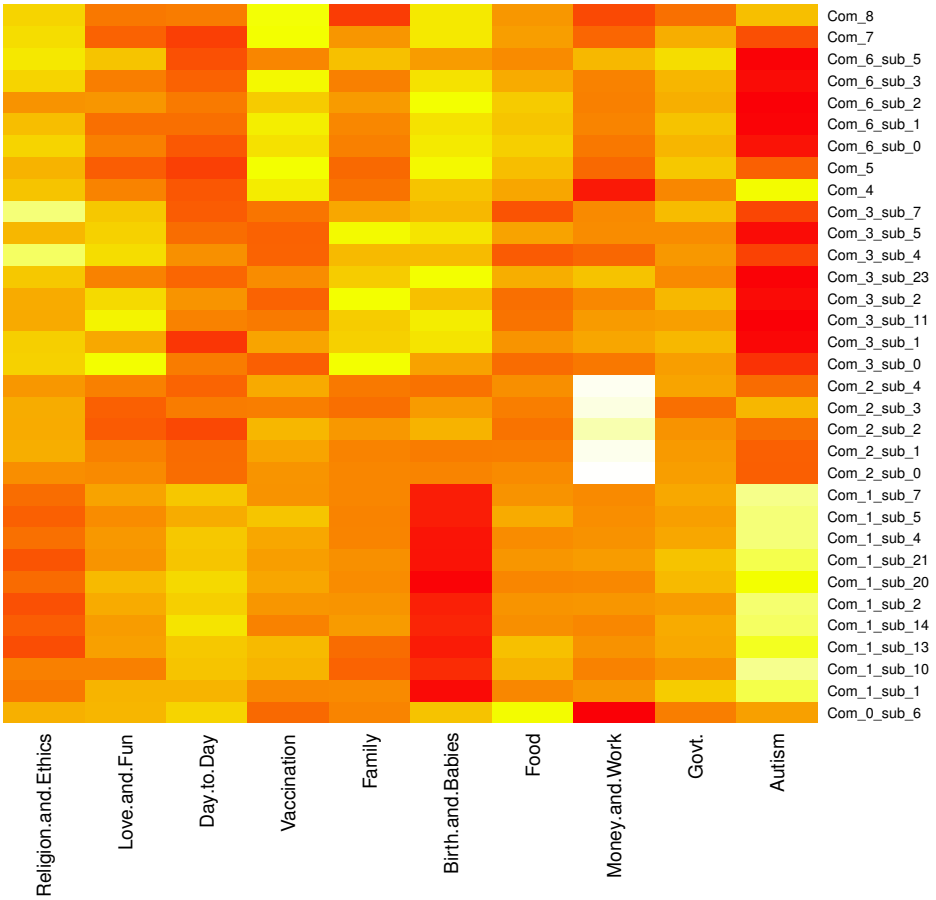


Figure 4.4: Topics of interest among different friendship communities detected by an agglomorative community detection algorithm based on modularity. Heatmap shows how certain communities are topic centric. Communities 6 and 3 along with its sub-communities show high affiliation for topic- Autism whereas community 1 and its sub-communities show high affiliation for topic- Birth and Babies.

ics as dictated by user behavior reflects our intuition about the birthing process. Many issues in the birthing process have ethical or religious implications, ranging from issues of a natural birth to abortion (Refer top words for these topics stated in Table 4.1). It also stands to reason that women who are concerned about issues of birth have just or have yet to give birth and since autism is diagnosed usually only after 2 years of age [FKB11], the topic will be of less importance to women in their gestation period or moms with new born babies. Similarly, communities with a high affiliation with *Autism* also post more frequently in *Day to Day*(0.6). Intuitively, parents with autistic children will have more questions and discussions concerning the daily happenings and challenges of caring for an autistic child. These findings give a qualitative evaluation of interests of friendship communities as well a quantitative evaluation of topic relationships based on user inclinations. Correlation of topics helps reveal patterns of user behavior and commonality of conversation interests shared among friends.

4.6 Concluding Remarks

In order to obtain a better understanding of user content and interactions on online forums, we propose an automated methodology that generates a comprehensive and multi-layered depiction of how forum conversations evolve over time and how friendships within a network highlight particular patterns in user conversations. By integrating unigram analysis and topic modeling temporally, we achieve a degree of detail and granularity of user content that efficiently captures external events such as the 2008 presidential election, the 2011 tsunami and nuclear disaster in Japan as well as references to major corporations and organizations such as Verizon and the Food Drug Administration. The level of specificity enables us to track how conversations progress over time. Furthermore, analysis of topic correlations based on friendship communities reveals how user-user interactions

Table 4.5: Correlation matrix for topics based on topic weights per community i.e. community/sub-community. Matrix shows communities that are affiliated highly with one topic, also correlate with other topics. This correlation can be verified by examining the heatmap in Fig. 4.4. For example communities that post most in topic- Birth and Babies, also post highly in topic- Religion and Ethics and much less in topic- Autism.

Topics	0	1	2	3	4	5	6	7	8	9
0	1	0.75	0.38	0.73	0.81	0.91	0.78	0.19	0.88	-0.22
1	0.75	1	0.7	0.3	0.94	0.65	0.6	-0.12	0.78	0.01
2	0.38	0.7	1	0.17	0.58	0.29	0.52	-0.45	0.48	0.6
3	0.73	0.3	0.17	1	0.4	0.79	0.81	-0.06	0.71	-0.17
4	0.81	0.94	0.58	0.4	1	0.76	0.66	0.01	0.83	-0.11
5	0.91	0.65	0.29	0.79	0.76	1	0.85	0.11	0.81	-0.36
6	0.78	0.6	0.52	0.81	0.66	0.85	1	-0.19	0.82	-0.05
7	0.19	-0.12	-0.45	-0.06	0.01	0.11	-0.19	1	0.01	-0.41
8	0.88	0.78	0.48	0.71	0.83	0.81	0.82	0.01	1	-0.15
9	-0.22	0.01	0.6	-0.17	-0.11	-0.36	-0.05	-0.41	-0.15	1

reflect inclinations of interest. We identify a strong positive correlation between topics of *Birth and Babies* and *Religion and Ethics* as well as between *Autism* and *Day to Day*. Correspondingly, we also see a strong negative correlation between *Birth and Babies* and *Autism*. By employing a methodology that takes into account both the content-driven and socially driven aspects of forum conversations, we are able to efficiently generate a detailed summary of the dynamics of conversations as well as the similarities in interest among socially connected users. These results are exciting and present a path for future work where some of the issues in the current chapter can be improved –for example choosing the number of topics was somewhat arbitrary. While we don't expect major shifts in the results, nevertheless the individual steps taken can be made more rigorous.

CHAPTER 5

A Gestalt Computing Methodology

5.1 Abstract

As an ever-growing volume of human communication moves to the web, fundamental questions about group dynamics persist in these new spaces. Our central questions are: How do group dynamics play out online and how can we detect them at large scales? Here we propose an automated and unsupervised methodology for summarization of group dynamics in online forums using simple actions by users based on their content preference. This methodology produces a temporal map of paths representing evolving groups of users, each characterized through automatically-identified content preferred by their members. We further quantify the paths' attributes and relations between paths to obtain a full picture of the dynamics. We use this methodology to study political group dynamics in a popular social news site with 4 years of data. The results reveal evolving groups with distinct preferences and demonstrate the immense effect of a contentious political event. We discover a structural rearrangement of groups following the event, an abrupt and enduring shift in the focus of groups, and a near-complete extinction of certain interests. Our results show that meticulous study of content shared on the forum through language processing techniques is not necessary in detecting meaningful evolving groups. In fact the most readily accessible quantities, the actions of users, provide adequate information. Furthermore, the proposed method is widely applicable to different contexts, requires no expert knowledge of the forum under study, and allows for both high-level and fine-grained inspection

of groups over time.

5.2 Introduction

With society's growing shift toward online communication, understanding the processes of interaction and opinion formation on the web becomes increasingly important. It is especially compelling to study how people interact in websites that facilitate discussion among large groups of users who typically would not communicate. One principal class of such spaces are social news websites[1], where users with different interests and opinions post, discuss and promote their preferred content. These websites, which are influencing traditional media outlets with increasing frequency, rank articles based on how well they are favored by users and give the higher-ranking articles more prominence and exposure. Thus the rankings are the product of a natural and sometimes implicit competition among various interests and opinions held by users. Meanwhile, interests and opinions of groups of users are not fixed and evolve with time as a result of interaction with others, exposure to new information, and external events. We are aware that the terms *group dynamics* and *collective behavior* take on different meanings across a number of traditions such as psychology, sociology, or management. Here we are not referring to those contexts. Instead, we use the terms in a broad sense as *the actions of, and interactions between and within collections of individuals who demonstrate cohesion in their traits or behavior.*

It goes without saying that group dynamics in online spaces and at such large scales are unprecedented in the real world where social behaviors have been extensively studied. This amplifies the need for comprehensive methodologies that allow scholars to investigate large scale group dynamics on the web. Yet the study of human behavior on the web involves making sense of cluttered masses of data,

[1]Some examples are Reddit, Slashdot and Digg.

often driven by dynamics that further increase complexity and disorganization. The ultimate goal of the growing research in this field is to move toward developing robust, replicable concepts grounded in data. The challenge is that the processes of exploration that lead to formation of concepts are obscured. Just as forming a scientific inquiry is premised on observation, data-driven research is premised on a great deal of data exploration. Yet the resulting manuscripts conventionally do not include this process of exploration that is essential in developing insights into larger structures and mechanisms and arrival at specific inquiries. Scholars dive into the ocean of data, bring out a gem leaving a line that leads to the location of that gem. Yet the process of combing through the vast ocean floor for the next gem, the process of exploration, is often omitted. Furthermore, the ability to zoom in to specific phenomena and then zoom out to observe the big picture is particularly important. It allows one to see context, produce frameworks that explain the regularities, and identify irregularities for further investigation and to navigate the results from high-level and selective, to fine-grained.

To meet this need for exploration at different granularities, we propose an approach that can be best described as *gestalt*[2] *computing*. When studying complex social phenomena on large datasets, we produce a context while laying out a path to investigate data at different granularities and in increments. In other words, we begin by extracting macro structures, finding their meanings, and arranging data in layers so as to make them easy to repeatedly and conveniently explore. In addition, much of human behavior online includes macro phenomena that are not only an aggregation of individual user actions, but that perform a function of their own, in turn affecting the individual actions as a result. Thus gestalt computing can in turn be understood as computation for the purpose of detecting gestalt.

In the current chapter we propose a methodology that provides a top-down

[2]From the Merriam-Webster dictionary: Gestalt is a structure, configuration, or pattern of physical, biological, or psychological phenomena so integrated as to constitute a functional unit with properties not derivable by summation of its parts.

map of group dynamics on social news websites and we demonstrate its implementation on a compelling dataset. We use explicit indicators of user preference for content as the basis for our methodology. Some examples of such indicators are the "Like" button in Facebook, an "up" vote in reddit, a "+1" in Google-plus, or a "digg" on Digg. We will call these indicators *votes* in the context of this chapter and will use them as clear and simple signals that can be used to infer user orientation toward content. For example, intuition suggests that communities of users that prefer and promote the same political articles will have similar political leanings (whereas explicit friendships do not necessarily suggest similar political orientations). Using a graph-theoretic community detection algorithm we extract groups of users with similar interest in content and track these groups temporal evolution. We then identify representative content for each group and produce summaries of each path and quantify their characteristics. The result is a layered representation of evolving groups in the website. Once we produce a summary of the evolving groups, several interesting questions can be raised: Do users form polarized and insular groups? Does one group dominate or drive out other groups? Is there movement between groups? How can we design websites to foster cross-group understanding? How do external events affect these dynamics? What are the evolving interest patterns and what is driving them?

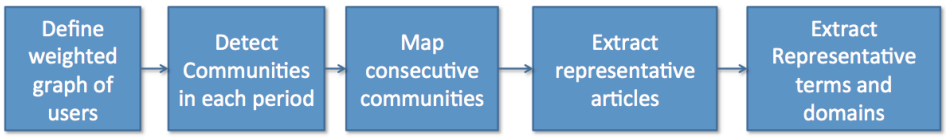


Figure 5.1: Methodology steps.

5.3 Overview and Approach

In the this section we will describe our approach and then demonstrate the results of its implementation on the data set.

5.3.1 Community Detection and Evolution

To group users who vote similarly, we define a bipartite network of users and articles where each edge is a vote cast by a user to an article. Figure 5.2 illustrates this structure. We project this bipartite network onto a weighted unipartite (single-mode) graph consisting of users only, where the weight of an edge between two users reflects how similarly they vote. The weight of an edge between users x and y is assigned using the Jaccard Index $W_{jaccard} = \frac{n(X \cap Y)}{n(X \cup Y)}$ where X and Y are sets of articles voted for by user x and y respectively, and n stands for set cardinality.

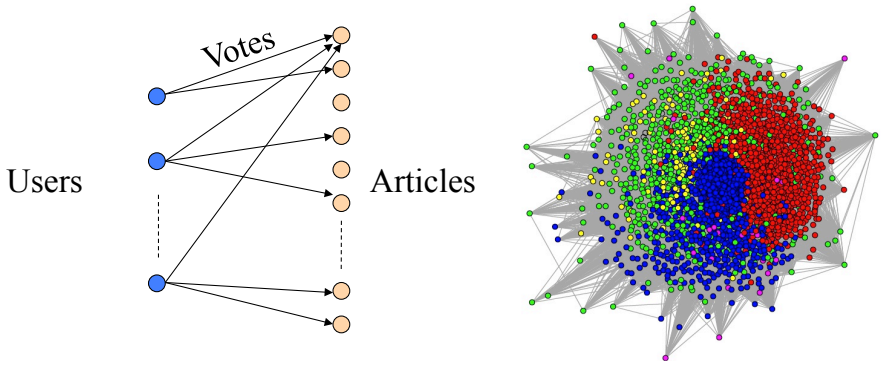


Figure 5.2: (Left) Bipartite graph of users and articles. (Right) Example of projected graph of users and the communities found in a one month time frame of data, each community in a different color.

Groups are detected using an agglomerative hierarchical community detection algorithm that optimizes the modularity [GN02] metric. This metric is one of the most widely used measures of community formation in the study of network

topologies. Modularity compares the connections within each community to that of the same nodes in a null model randomly generated with the same degree sequence. More formally, for an unweighted graph, let A_{xy} be the element of the adjacency matrix corresponding to vertices x and y (such that $A_{xy} = 1$ if they are connected and $A_{xy} = 0$ otherwise) and let the total number of edges in the graph be m. Then the fraction of edges that are in the same community is equal to:

$$\frac{\sum_{xy} A_{xy}\delta(C^x, C^y)}{\sum_{xy} A_{xy}} = \frac{1}{2m}\sum_{xy} A_{xy}\delta(C^x, C^y)$$

let k_x be the degree of a vertex x, $k_x = \sum_y A_{xy}$, then the probability of an edge between x and y in the null model is $P(A_{xy} = 1) = \frac{k_x k_y}{2m}$

$$Q = \frac{1}{2m}\sum_{x,y}(A_{xy} - \frac{k_x k_y}{2m})\delta(C^x, C^y)$$

Similarly, modularity for a weighted graph is defined as [New04]:

$$Q = \frac{1}{2W}\sum_{x,y}(W_{xy} - \frac{s_x s_y}{2W})\delta(C^x, C^y)$$

where W_{xy} is the weight of the edge between vertex x and vertex y, W is the sum of the weights of all edges and s_x is the strength of vertex x, defined as sum of the weights of its adjacent edges $s_x = \sum_y W_{xy}$. C^x is the community that vertex x belongs to and δ is the Kronecker delta, being equal to 1 if x and y are in the same community and zero otherwise. The expression $\frac{s_x s_y}{2W}$ computes the expected number of edges between vertices x and y in the null model.

We find sequences of such vote-based groups, by first constructing bipartite graphs and their single-mode projections for the data in consecutive time frames. Then, using a fast modularity maximization algorithm [CNM04b], we find communities for each time frame. Figure 5.2 shows a visual example of communities found in a one month time frame of our dataset that will be described in detail later in the chapter.

To map the evolution of groups over time, for every pair of successive time frames we compute a two-way transition probabilities between every community pair i and j in times τ and $\tau+1$. Consider the user overlap between communities in consecutive time periods and Let C_i be the set of users in community i. The fraction of users in community i at period τ who move to community j during the next period is $\frac{n(C_i \cap C_j)}{n(C_i)}$. Similarly, the fraction of users in community j at period $\tau+1$ who have come from community i in the previous period is $\frac{n(C_i \cap C_j)}{n(C_j)}$. We map together pairs of communities that maximize the product of these two factions to produce the map in Figure 5.6. We highlight paths lasting at least 6 time periods.

5.3.2 Representative Articles

The evolving communities detected in the previous section will define the skeleton of voting behavior among users. In order to characterize the nature of detected communities and add a layer of meaning, we first find the articles most representative of each community. Representative articles for each community are selected according to their unique appeal to the members of that community. For each community at each time period, articles are assigned a score that quantifies this appeal. Intuitively, articles preferred by a community will receive an improbably high number of votes from that community compared to a case where votes are cast purely at random.

Specifically, calling p_i the global probability of a vote being from community i (i.e. $p_i = \frac{N_i}{N}$; where N_i is the total number of votes by community i and N is the total number of votes from all communities, in that period), the null model predicts a binomial distribution for observing o_{ij} votes given to article j by community i :

$$p(o_{ij}) = \binom{N_j}{o_{ij}} {p_i}^{o_{ij}} (1-p_i)^{(N_j - o_{ij})}$$

Here N_j is the total number of votes article j has received. A representative

article for the community is one that has received a disproportionately large number of votes from that community. That is, the probably of observing o_{ij} votes is very small under the random voting model ($p(oij)$ is very small and $o_{ij} > p_i \times N_j$). Using this expression, we rank (in ascending order) articles for every community belonging to a time frame and create a list of most representative articles for each community. We expect that the articles representing each community will have similarities in their content that signify a difference from other communities, and that this preference within each community will carry over through the whole evolution path.

5.3.3 Domain and Word Summaries

Each path is summarized through a list of web domains and words extracted from its representative articles. We arrive at the list of domains by aggregating top representative articles of communities that form the path and noting the domains of host websites where they were published. The domains that appear most frequently demonstrate the aggregated preference of the users in the path.

We will then extract a more granular characterization of content reflected in each evolution path. For this purpose, we consider the ranked list of representative articles within each community. Given that each article that is posted to the site includes a title and a summary of its content (as is almost always the case in online forums and social news sites), we use a bag of words model to find the deviation between the words used in representative articles for a community and the rest of the articles posted in a time frame:

$$\text{Score}(T) = \frac{\text{tf}(T, C)}{\max_t \text{tf}(t, C)} - \frac{\text{tf}(T)}{\max_t \text{tf}(t)}$$

where $tf_{T,C}(t)$ is the term frequency for term T in community C at one time interval[3]. This expression computes the difference between the frequency of a

[3]Although this process of scoring is similar to term frequency–inverse document frequency

term in representative links and the overall frequency of the term in all articles posted in a time window. It will reflect whether certain words are used more often in the representative articles of a community. Based on this score, we find the words with highest score in each community and aggregate them over all the communities in a path. These words differentiate the content preferred by users in each path.

At this point, we will have a summary visualization of the overall dynamics over time, a set of relevant words and domains (i.e. publication sources) most representative of each evolution path, as well as the capability to drill down to any specific time frame and get a list of representative words and publication sources for each community at that time. Finally, for each community at any time frame, a ranked list of specific representative articles and the url to the full article is available for an in-depth examination.

Once we have these we can decide whether the paths are meaningful. we will now describe the dataset and the results of this process.

5.4 Implementation

We apply this methodology to a dataset from a social news site popular among Iranians inside and outside Iran. The website named *Balatarin*[4] (translated *The Highest*), quickly became a prominent venue for seeking and promoting information and discussing opinions in the Persian-speaking population. The recent surge of political change and popular uprisings in several Middle Eastern and Islamic countries (such as Egypt, Syria, and Turkey), make it compelling to study how political group dynamics manifest in this dataset and how major political events affect these dynamics. We will describe this dataset in the next section.

(tf-idf) weighting [MRS08], note that we are not ranking documents and are instead finding a normalized ranking of terms only, so we do not use inverse-document-frequencies.

[4]balatarin.com

5.4.1 Data

The dataset of Balatarin consists of 1.2 million articles, 26,000 users and 31 million votes posted from August 2006 to November 2010. Less than 3% of users are casting more than 55% of the votes. The articles are posted under pre-specified categories where the *Politics* category includes 352,000 of the total articles posted to the site. This is where we concentrated the bulk of the study. A sudden rise in number of articles, observable in Figure 5.3, coincides with the 2009 protests[5].

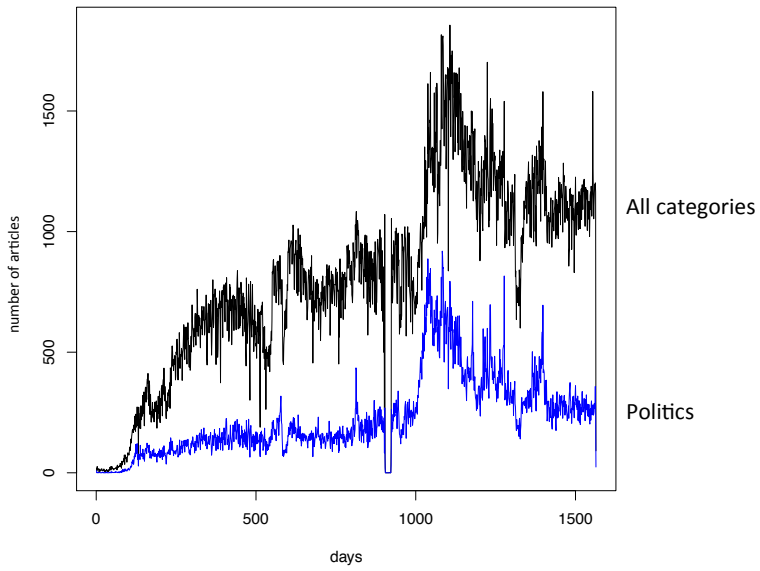


Figure 5.3: Timeline of number of articles posted to the site overall and in the politics category.

5.4.2 Results

We form consecutively overlapping time windows, each 30 days long sliding over the whole duration of the data. Sliding the windows 14 days at a time produces 110 temporally consecutive datasets. The data was analyzed with anonymous user ids and in an aggregated fashion with no identifying or confidential information

[5]The short sharp drop to zero marks a shut-down due to an attack on the site in February 2009.

Figure 5.4: Screenshot of a Balatarin article.

about the users. Figure 5.6 illustrates the implementation of this methodology on 4 years of data from this site. Each oval shape represents a group at a 1-month duration (consecutive groups overlap in time) and the size of an oval scales with the square root of the number of users in the group. Time begins in 2006 on top of the figure and proceeds downward to 2010. Paths highlighted in different colors and named alphabetically represent the evolution of groups in time and the legend on the top left shows the labels we assigned to each path based on the results.

The results show nineteen distinct paths of varying sizes, each lasting between 3 and 18 months, having between tens to three thousand users. Finding representative articles within each group along a path, we extract what each path signifies and find that paths bear distinct meanings.

Observe that around June 2009 several large groups appear and repeatedly shuffle until clear paths form. This date coincides with Iran's contested presidential election, which prompted widespread protests dubbed the *Green Movement*. Some examples of political groupings before the contentious political event include path B which has a strong preference for conservative/ traditionalist politics inside Iran, path A which favors reformist views, and path G which focuses acutely on content about Iran's foreign affairs. Following the election unrest, observe large paths (I, K, L and N) that focus heavily on the Green Movement. While path N

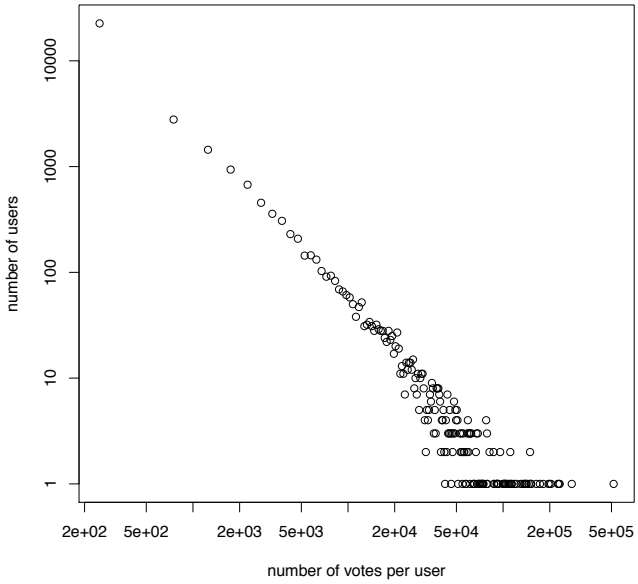


Figure 5.5: Number of votes per user (log-log scale)

centers on human rights violations and arrests of political activists, consecutive paths I, K, and L focus on protest reports, planning, dissemination of news, and eyewitness videos and photos. A third, smaller group forms with content opposing the reformist core of the Green Movement, and contains both pro-government and radical anti-government views. Note that the meanings within each path exist despite the fact that information from the text of the articles was *not used* to produce the paths and thus the coherence in content underscores the effectiveness of our methodology.

Further characterizing each path, we find that paths with reformist and pro-Green Movement politics are more consistent and retain more of their users with time whereas the conservative path suffers a near extinction after the election event.

The results demonstrate the dramatic effect of the post-election uprising on the paths and their theme. As one may expect, path themes shift abruptly to the Green Movement, and they remain so for more than a year. In fact, a principal

component analysis of user overlaps within paths, demonstrates that the popular uprising forms a focal point in the dynamics of the website, and that political orientations are shaped around the Green Movement. The uprising was so significant that it affected non-political forums within this website, such as a forum on sports for which we also performed this methodology. You can read more about this result in the Discussion. Additionally, we find that users who focused on foreign affairs prior to the election in fact are close to reformist path and form a major pro-Green Movement path after the election.

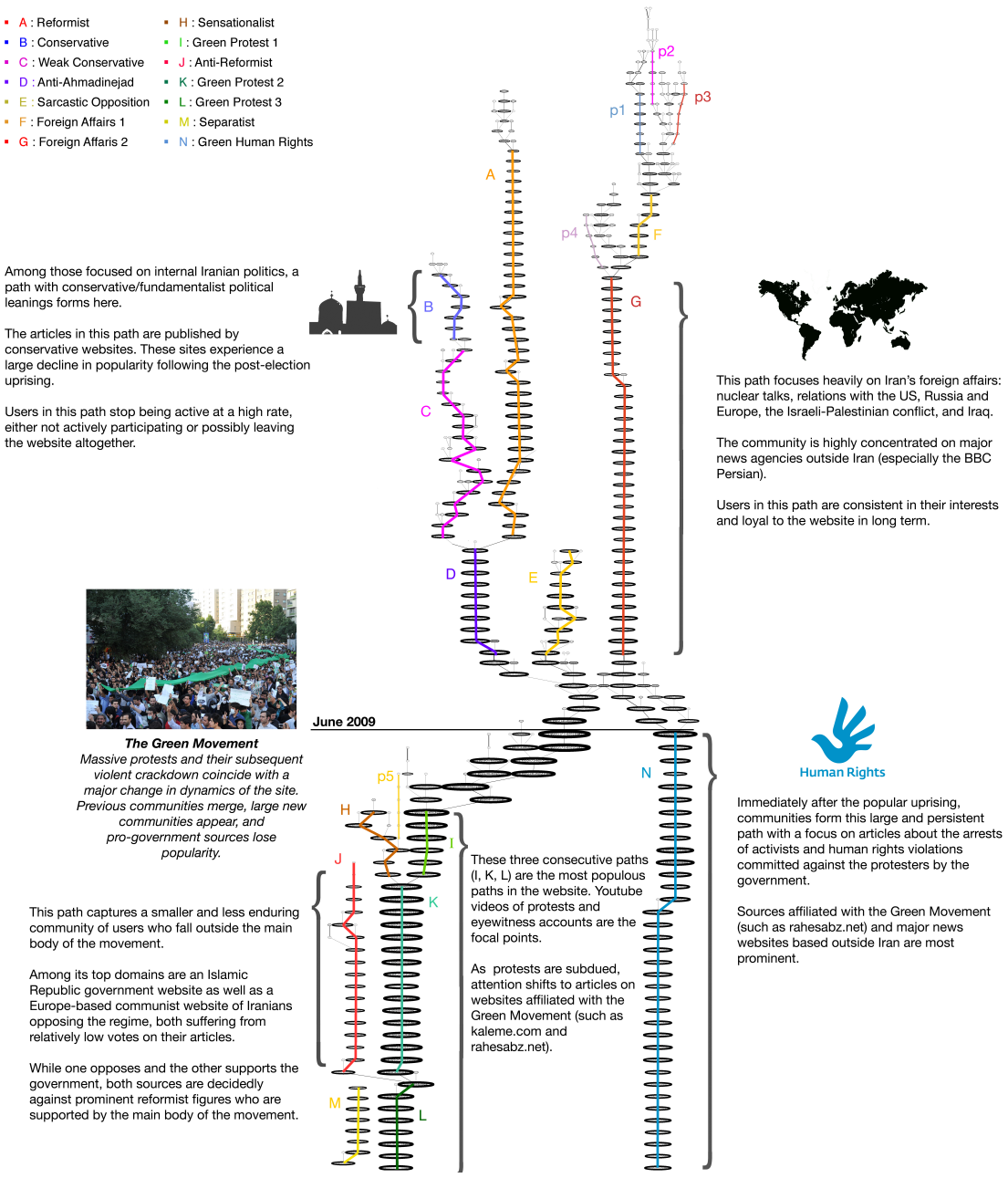


Figure 5.6: A gestalt map summarizing evolving political patterns over four years on a social news site *Balatarin.com*. Time begins on top of the figure and progresses downward. Each oval shape represents a community at a one-month-long period and its size scales with the square root of number of users in the community (largest communities include 3000 users). Evolution paths are alphabetized and marked in different colors.

Table 5.1: Summary of domains and terms associated with paths. Terms have been translated from Persian to English.

Path	Domains	Terms
A	www.youtube.com www.tabnak.ir video.google.com www.roozonline.com www.autnews.info	Ahmadi-Nejad Khamenei President Students Speech Photo Work University Khatami Prison Tehran Distribution Execution Hashemi
B	www.qodsdaily.com www.jahannews.com kaargar.blogfa.com www.persianblog.ir	Site, Office, Sentence, Ayatollah, Germany, Criticism, Student, Information, Karroubi, Published, Resignation, Ninth, Universe, Community, Reporter, Familiar, Deny
C	www.qodsdaily.com nonoghalam.blogfa.com khorshied.blogfa.com www.roozonline.com smto.ir www.rahesevvom.com	Regime Hand Ahmadinejad President War America Obama Site President Decision Recent News Khomeini Change Mosque
D	www.isna.ir www.youtube.com advarnews.info news.bbc.co.uk www.jahannews.com	Hand President Oxford Imam Iranian University Ahmadinejad Rights Representatives Islamic People Khatami Word Oil Impeachment

Table 5.2: Summary of domains and terms associated with paths, continued.

Path	Domains	Terms
E	ghazizade.blogfa.com www.fararu.com www.khabaronline.ir www.rajanews.com freedomvatan.blogspot.com www.inn.ir www.milliharakat.com	Intelligence Activists Arrest Civil Free Azerbaijani Language Law Tabriz Prison America Family Sentence National Demand Exchange Execution Photo Prisoner Imam Capture Yesterday
F	www.bbc.co.uk www.roozonline.com www.radiozamaneh.org www.aftabnews.ir www.roozna.com	Foreign America Negotiation Representative Political Garden Britain Round Nuclear Government Authorities Students Officials Peace Condition Saudi Arabia University Program
G	www.bbc.co.uk www.dw-world.de www.roozonline.com radiozamaaneh.com www.radiofarda.com	Minister, Nuclear, Spokesperson, Russia, Council, Continuation, Israel, Security, Iraq, Arrangement, Agency, Europe, America, Declare

Table 5.3: Summary of domains and terms associated with paths, continued.

Path	Domains	Terms
H	hammihannews.com www.parcham.ir www.rahesabz.net irancpi.net	Language Green Political Headquarters Weblog Movement Green Karroubi Bakeri Mehdi Martyr War
I	www.youtube.com 7tir.info rahesabze-omid.blogspot.com harfehesaaby.blogspot.com www.kaleme.com www.drsoroush.com	Film Islam Ahmadinejad Coup Scene Death Khamenei Student Answer Question Leader Property Now Mahmoud Trample
J	inn.ir rowzane.com www.youtube.com www.asrefarda.com www.tahavolesabz.com khabarnegaran.info	Party Power Azerbaijan Proletarian Khamenei Revolution Regime Forces Declaration Total Communist Month Protests Hamid Activists Report Future Human Move Republic

Table 5.4: Summary of domains and terms associated with paths, continued.

Path	Domains	Terms
K	www.youtube.com www.kaleme.com twitter.com friendfeed.com www.rahesabz.net tinypic.com norooznews.info	People Mousavi Ahmadinejad Khamenei Film Fight Statement Violence Principle Watch Government Exit Hand Open Defend Ayatollah Program
L	www.kaleme.com www.rahesabz.net www.youtube.com iarandoost657.blogspot.com gomnamian.blogspot.com enghelabe-eslami.com	Khamenei Mousavi Attack Society Country Mehdi Mir Hossein People Number Published Photo Lie Defense
M	www.youtube.com rowzane.com www.oyrenci.org www.tribun.com noislamicrepublic.blogspot.com traxturfans.blogsky.com urmiye.blogsky.com	Azerbaijani Tabriz Political Turkish Independence Non-Persian Turkey Iranian Freedom University

Table 5.5: Summary of domains and terms associated with paths, continued.

Path	Domains	Terms
N	www.rahesabz.net	Prison Arrest Status Rights Free Political
	www.radiofarda.com	Court Family Report Prisoners Evin Law
	www.dw-world.de	Human Attorney Continuation Mention
	zamaaneh.com	Execution Revolution Basic Sentence Uni-
	news.gooya.com	versity Emphasis Freedom Spouse Mehdi
	www.roozonline.com	Newspaper Protest Demanding Letter
	www1.voanews.com	
	www.kaleme.com	

Table 5.6: Summary of domains and terms associated with five minor paths.

Path	Domains	Terms
P1	aftabnews.ir www.bbc.co.uk www.farsnews.com www.tik.ir jomhour.org	Ahmadinejad Tehran Security Declared Iran Country Relations Parliament Government Representatives Ratification Report Israel Conservatives Emphasis City Republic Syria Spokesperson
P2	azaadir.blogspot.com www.peiknet.com www.bbc.co.uk www.baztab.com	Mention Time Discussion Presidial University Letter Parvin
P3	www.titronline.com www.peiknet.com ferdoss.blogfa.com iranukyellowpages.com www.baztab.com asriran.com	Muhammad Jebel Akbar Chief Foreign Hassan Patrice Lumumba Faghih Declared Era Ali Kongo Rights Klein Information Summit Introduction Speech Respond Cause
P4	www.irannewsagency.com	America Declared Newspaper Order News Laden Report Minister Fatemeh Islam Away Qods Award Bush Nouri Action Blair Faghihi Iran Government System President Network Negotiations Response
P5	www.dailylink.ir	Convention United Nations Qods Presence Turkey Ahmadinejad Country Representatives Parliament Japan Demonstration New York Israel Obama War World Policy Public

CHAPTER 6

Quantifying the Structural Results

6.1 Evaluation

6.1.1 simulation

We begin by assigning each user a position on a 2-dimensional Cartesian space that will represent the underlying opinion space[1]. Users are randomly placed according to a normal distribution around one of four equidistant center points in the four quadrants. The position of users is considered the ground truth, with each user belonging to one of the four communities specified by the four quadrant centers (Figure 6.1). Given this structure, a k-means algorithm that uses the (otherwise unobserved) user positions can find the four user clusters with relative ease thus serves as an approximate lower bound for error in detecting communities. We then generate a set of articles by randomly selecting users who will each post articles and votes. Each generated article is positioned in the opinion space according to a Gaussian distribution near the user who posts it. Each user will vote for an article with some probability, if that article is positioned closer than a certain threshold to him/her in the political space, thus an article is likely to get a vote if it's close to a reader's opinion. The result of this process is a set of users, articles, and votes which we then use as a simulated graph for a social news platforms. Complete details of the simulation parameters and more detail on results: Individuals post stories with a rate. Stories embed an opinion close to the posters. Individuals

[1]While for clarity this simulation assumes a 2-dimensional opinion space, we make no such assumptions in the general methodology.

read stories with another rate. Stories are read on the first page or new stories page. Probability of reading a story depends on its votes VjY. Story likely to get a vote if close to readers opinion. Stories are removed from the system after a while. Stories with votes above a threshold go to first page. Individuals heterogenous in opinions, going online, and posting.

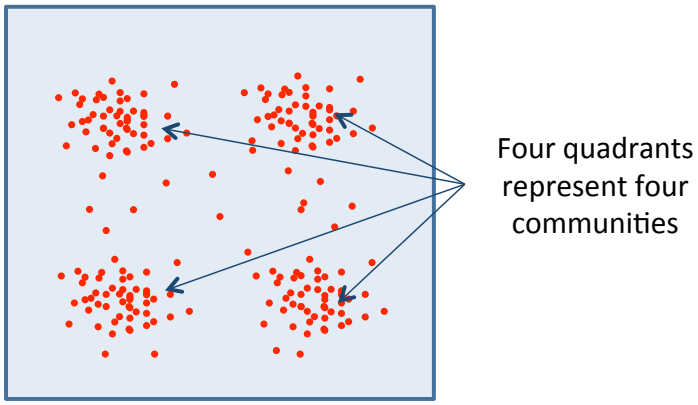


Figure 6.1: Two-dimensional opinion space with users normally distributed around ± 25 or ± 75 (bounded to [0-100]) with a standard deviation s_0.

We simulate this data with different variances for the aforementioned Gaussian distributions. We then run our network-based community detection algorithm on this graph and compute relative error as we change the variance of underlying data generation process (simulation model). Figure 6.2 compares the results of community detection (based on votes) with k-means clustering (based on true positions of users) as the standard deviation of the Gaussian distribution used to generate user positions changes. The algorithm is generally robust and successful in finding true underlying clusters while error increases with the standard deviation of user positions (i.e. as users are more scattered). When the value of standard deviation reaches the mid-point between the two centers, neither k-means nor the network based algorithm can detect clusters correctly. This simply means that users are distributed such that clear clusters do not exist anymore.

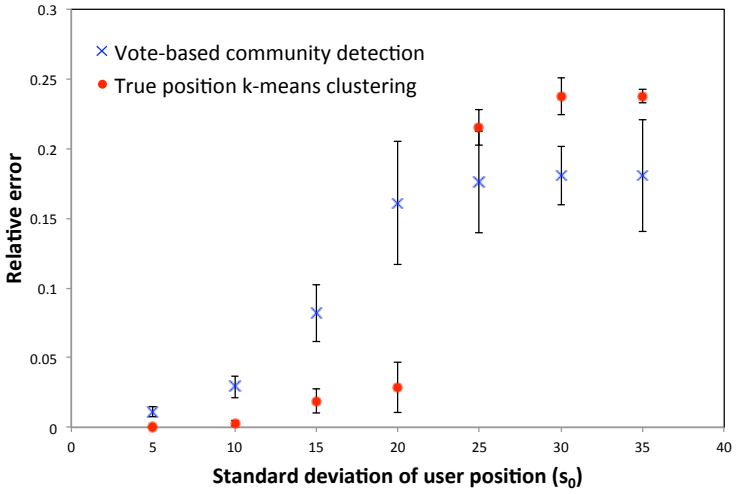


Figure 6.2: Relative error vs. standard deviation of user positions in the opinion space. The jump in the k-means error is due to the fact that true user memberships are no longer recognizable. Relative error is computed based on pairs of users that are classified incorrectly together or separate. 500 users were generated. Results are based on an average of 10 simulations. Error bars mark two standard deviations.

6.1.2 Domain vote p–values

Let us say that we have m domain names appear as top domains in a certain path and that domain i has received N_i votes from communities in this path. Assuming that X_i the random variable for the number of votes received by by domain i, we can write the following:

$$\text{p-value} = P(X_1 > N_1, X_2 > N_2, ..., X_m > N_m)$$

$$\leq \min_{i\epsilon\{1,...,m\}} [P(X_i > N_i)]$$

This minimum can be used as a conservative estimate of p-value. But let us say there is one (or some) extreme cases with very low probabilities. In order to remedy this, instead of using the minimum, we use a geometric mean of the K

smallest probabilities. So our more conservative estimate of the p-value is:

$$\text{p-value} \leq \sqrt[K]{\prod_{k \epsilon K\text{-smallest}} (P(X_k > N_k))}$$

and each $P(X_k > N_k))$ is computed as the binomial probability of document k getting more than N_k of the total N votes cast in the path. Since N is large, this can be approximated as a gaussian with mean of $\mu = n \cdot p$ and variance of $\sigma^2 = n \cdot p \cdot (1-p)$. We performed this over all the paths prior to the election event separately from those after the event in order to provide a more fair condition, since the dynamics changed a lot before and after the election and several domains lost popularity and others gained popularity. We found the p-values by computed the Q-function of a Gaussian distribution for values of $Q(\frac{\mu - x}{\sigma})$. The top 20 p-values for domains along each path were all in the ranges smaller than 10^{-24}, thus the overall p-value for the observed number of votes along each paths is $\leq 10^{-24}$. These results demonstrate that the votes by communities in a path are highly biased toward a set of preferred domains compared with a random set of votes.

6.2 Evolution-Path Characteristics

6.2.1 User Retention

Since we group users solely based on their vote similarity at individual time steps, it is not clear whether paths will remain meaningful and consistent after several time steps. If at each step a number of users leave and new users join the community, will any of the same users remain after several time frames? Will there still be content coherence within the whole path? Will it be reasonable to assume this is the same evolving community after so many time steps? To answer these questions we compute *user retention* by studying membership within a path across several time steps. User retention within a path is computed as the fraction of users who remain in the path after $\Delta\tau$ time windows. For a path P, we compute

user retention after $\Delta\tau$ time windows from time τ_i as:

$$\text{Retention}(P, \Delta\tau) = \frac{n(P(\tau_i) \cap P(\tau_i + \Delta\tau))}{n(P(\tau_i))}$$

where $P(\tau_i)$ is the set of users in path P at time τ_i.

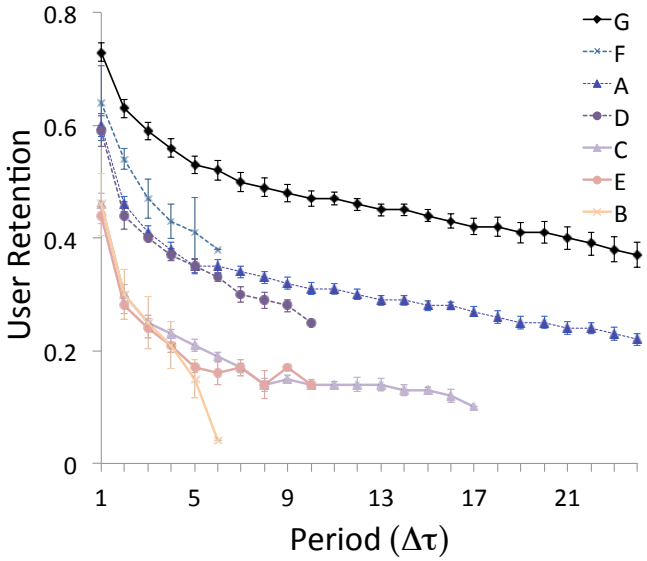


Figure 6.3: User retention (average fraction of users remaining in path) vs. $\Delta\tau$ for paths before the June 2009 event.

Figures 6.3 and 6.4 illustrates retention values vs. $\Delta\tau$ along the length of each path. Individuals who vote similar to one another over a long duration, form paths with high user retention. We find that paths G (Foreign Affairs), N and K (both related to the Green Movement) have higher user retention, containing between 40% and 50% of their original users six months later. A number of other paths demonstrate a sharp drop in user retention, such as paths B and C (both Conservative), path J (Anti-Reformist), and paths E and H (both with more vague political leaning and instead focused on sensational or sarcastic content). A lower user retention signals that users have either migrated to other paths, or have ceased to be active altogether. In the case of path B (Conservative) we find that a large fraction of users in fact completely stop participating after the election

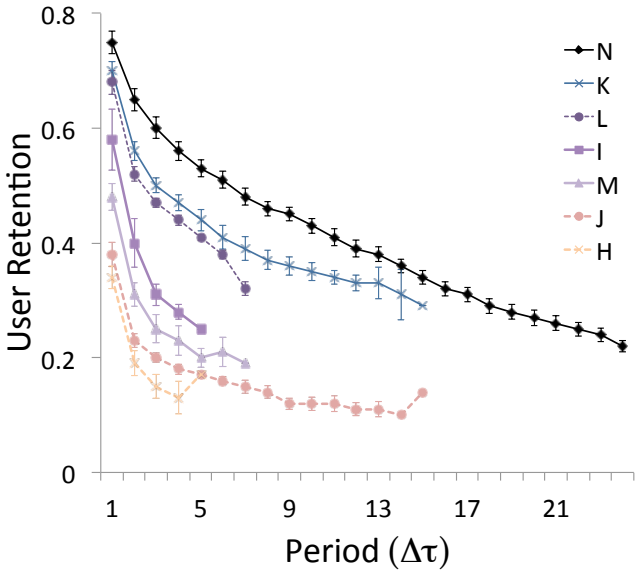


Figure 6.4: User retention (average fraction of users remaining in path) vs. $\Delta\tau$ for paths after the June 2009 event.

event.

6.2.2 Domain Diversity

We also find whether throughout the length of a path, there is a preference for a few sources of publication. Since the votes are cast to completely different articles, there should be no expectation that their sources be the same unless users in each evolution path are favoring certain sources of information over others, an indication of common underlying preferences.

We aggregate the top n representative articles over all the time frames in a community evolution path. We then calculate the *Shannon Entropy* [SWB49] of the source of these articles (as indicated by their domains). This will signify the amount of source variation over top preferred articles for each evolution path:

$$\text{Entropy}(C) = -\sum_i p_i log_2(p_i)$$

where p_i is the probability that an article from source i is in the top n most preferred articles of community C.

A lower entropy value indicates lower variation and higher uniformity in sources of articles. Entropies found for evolving communities are then compared to entropies from sets of articles drawn at random. We generate the random sets by randomly choosing votes, finding which articles the votes were cast for, and then extracting the domain of the article. We randomly choose *votes* rather than randomly choosing *articles* because we want the articles with higher votes to have a higher probability of being chosen. This is important because the list of most preferred articles in each community is also based on the preference of a community's users to vote for that article.

We then compute the *effective number of sources* in an evolution path as $2^{Entropy}$ and compare with that of the randomly selected sets[2] and compute the ratio as:

$$\text{Relative Recurrence} = \frac{2^{Entropy(random)}}{2^{Entropy(path)}}$$

A higher recurrence in sources of information compared with the randomly drawn dataset will strongly suggest that the evolution paths are highly preferential toward certain sources, corroborating that they are meaningful.

Table **??** lists relative recurrence of sources within paths. We observe that all paths have an increase in recurrence of information sources. We see as much as ?? times more recurrence of sources compared to the set drawn randomly (proportionally to an article's votes), demonstrating strong preferences toward some sources of information.

[2]This measure is used in Ecology as the *effective number of species*[Hil73] in an ecosystem. Another metric for diversity is found by comparing the effective number of sources with the number of unique sources in each set. Using this metric we reached similar results.

6.3 A Structural Understanding

6.3.1 Political Dimension via Principal Component Analysis

Our findings so far show that using this methodology, paths derived without the use of content analysis exhibit coherence in their preferred content. Based on these promising results, we ask whether overlaps between paths suggest a latent dimension that reflects underlying political orientations present in the website.

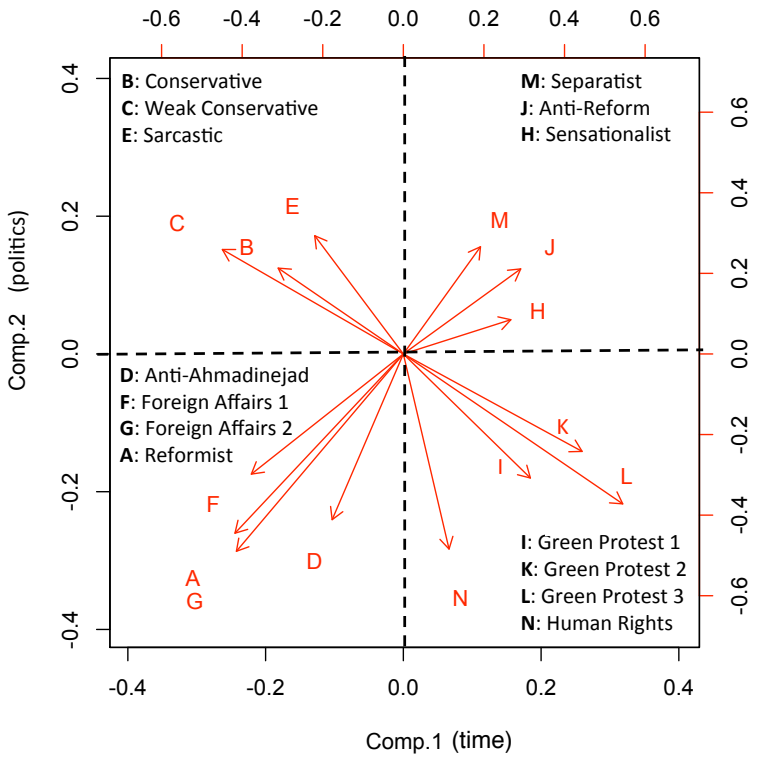


Figure 6.5: Biplot of paths using the first two principal components. The PCA is based on user membership overlaps. The first component (horizontal axis) matches closely with progression of time, with all paths prior to the election appearing on the left and all the paths after the election appearing on the right. The second component (vertical axis) reflects political orientations and the position of paths along this axis is in close agreement with the automatically-identified content for each path.

To detect this latent political dimension we perform principal component analysis (PCA) on the matrix of core-user overlaps between paths. Core users are defined as those with high PageRanks in their respective communities (more specifically, those more than one standard deviation above the average PageRank in the community). The results are illustrated in the biplot of Figure 6.5, which projects each path onto the first two principal components. The two components together account for 43% of variance in membership overlaps. The first component (horizontal axis) matches the time dimension of the data closely, placing pre-election paths to the left and post-election paths to the right. Notice that the election event emerges automatically as the focal point of this map.

The second component (vertical axis) represents an underlying political dimension in the website. Based on this component, paths A (Reformist), F, G (Foreign Affairs 1 and 2), D (Anti-Ahmadinejad), N (Green Human Rights), I, K, L (Green Protests 1,2, and 3) are placed in the opposite two quadrants from paths C, B (both Conservative), E (Sarcastic opposition), H (Sensationalist), J (Anti-Reformist) and M (Separatist). This division underscores the role of the Green Movement in defining the political dimension of the site, placing those opposed to the main body of the Green Movement, including those opposing the reformists from both sides of the political spectrum, in opposition to those in support of the movement. In addition, the proximity of paths F and G to paths I,K, and L, on the vertical axis, means that the core users in F and G have shifted their attention from foreign affairs to the Green Movement following the election unrest.

The fact that a meaningful political dimension emerges should not be taken for granted since the analysis was performed without any use of content or meaning of the paths and merely through user actions.

6.3.2 Path Relationships

To measure the flow between paths, we compute the proportion of core users in each path who move to another path later. Core users are defined as those with high PageRank within the subgraph of their respective communities (as derived through the community detection step). Figure 6.6 provides an overall summary of path sizes, user retention within each path, and inter-path user migration. Each path is signified by a rectangle proportional in width to the number of unique users in each path. Darker color represents higher user retention and arrows represent core user migration, with their thickness representing the proportion of users migrating. Furthermore, the horizontal distances between the paths reflect their position on a political dimension derived using the principal component analysis, which will be described in the next section.

6.4 Discussion

To summarize, the fully automated and unsupervised method described in chapters 5 and 6 infers group evolution paths with distinct and meaningful preferences. The paths are distinct in terms of their difference between preferred articles, urls, and words and were, in addition, were meaningful to human understanding. Meanwhile, deriving the structure requires no expert knowledge of the forum under study. In fact there is no human involvement up to the point where we have to study the results and interpret what the preferred words and URLs mean.

The method incorporates both users and content (rather that just one or the other), while avoiding computationally intensive language models to process the content. The results show that meticulous study of content shared in the forum is not necessary in detecting meaningful evolving groups.

The benefits of using a gestalt approach also become evident once we circle

back to interpret individual users in the context of the global structure. The process began with elementary user actions (votes), from which we obtained the global structure of user groups and interests in the website. In turn the context provided by this structure can be used to give back possible meanings to individual user actions. In order to demonstrated this, we compared two users from this dataset (which we will call user1 and user2) who appear to be quite similar outside the context we derived. In fact, out of the top 20 domains most voted for by these users, only 2 are different. Now we consider a global measure derived from the gestalt result, let us define a consistency score for a user as:

$$\text{Consistency Score} = \frac{N_{active}}{N_{switch} \times N_{paths}}$$

Where N_{active} is the number of of active periods, N_{switch} is the number of times user switches paths and N_{paths} is the total number of unique paths the user has been a member of. This metric will reflect how much a user has moved between paths and the results for the two users are very different. User1 has a consistency score of 3.75, being a core user in paths A, F, G, N (Reformist, Foreign affairs, Human rights). On the other hand, user2 receives a 0.6, and has been a core user in paths A, C, D, G, K, L, N (Reformist, Weakly conservative, Anti-Ahmadinejad, Foreign affairs, Eyewitness, Human rights). This difference can only be observed once we have the gestalt view of the website.

In the coming sections we will discuss some of the considerations and questions that arise with this approach.

6.4.1 Sensitivity Analysis

Figure 5.6 in chapter 6 is not meant to be the only valid summary map of the site's group dynamics. Group dynamics are at play at different granularities and representations of these dynamics at different granularities are equally valid. The question, therefore, is how useful and meaningful the results are. For the current

analysis, different parameter choices lead to variations in the length and number of paths in the summary figure. Finding a small number of long paths produces coherent yet coarse-grained results while deriving a large number of paths gives us fine-grained results but the paths may be fleeting and fail to form a coherent theme. Therefore getting a map of paths like Figure 5.6 includes a tradeoff that can be adjusted using the method's parameters. It is desirable to choose parameters that produce more and longer paths, ensuring both coherence and granularity.

Recall that community detection is performed along sliding time windows. The length of this time window (W) and the sliding window shift size (S) are the first two parameters. The third parameter (Th) is an activity threshold eliminating users who have low participation. Notice that low-vote users whose votes happen to match will be connected with a strong edge while in reality their low number of votes does not carry enough information to assign them to one group or another. To address this, we remove users who vote less than a threshold (Th) during a time window.

We perform the methodology on combinations of these three parameters, producing the map of paths for each set. Figure 6.7 demonstrates the variation in number of paths and average length of paths for four parameter combinations. One can see, for example, that Th=2, W=30, S=14 produces twenty short paths, while choosing W=60, S=30, Th=10 produces very few paths that are very long. The choice that simultaneously produces more paths that are also longer corresponds to W=30, S=14, Th=5, which is what we chose to use for this chapter. The paths produced four combination of parameters are illustrated in figures 6.8 and 6.8.

Table 6.1: Summary of domains and terms associated with two paths using parameters W=30, S=14, Th= 10.

Path	Domains	Terms
1	www.bbc.co.uk www.radiofarda.com www.roozonline.com www.rahesabz.net www.isna.ir www.dw-world.de www.noandish.com radiozamaaneh.com www.farsnews.com www.youtube.com	Arrest, Revolution, Political, Security, Government, Forces, Law, Ahmadinejad, Students, Power, Prison, Past, Meet, University
2	www.qodsdaily.com www.youtube.com www.asrefarda.com nonoghalam.blogfa.com ghazizade.blogfa.com www.irannewsagency.com www.fararu.com twitter.com www.bbc.co.uk khorshied.blogfa.com	Ahmadinejad, Hand, Photo, God, President, Reform, Raid, Meet, People, Mousavi, Freedom, Election, Imam, Language, Khatami, Israel, Iranian

6.4.2 Implementation on Alternative Dataset

The methodology can be widely applied to other contexts wherever users collaboratively promote content. We performed the methodology again on the dataset of sports posts on Balatarin. The process again produces the graph of evolving groups, their representative articles, terms and domains (Figure 6.10).

We find that paths mainly focus on football (soccer) and are shorter and merge regularly compared with the politics paths. We also observe that the paths are created around external events such as the Iranian Football League, the European Football Championships, the Asian Cup, the Olympics and Paralympics, as well as sports scandals.

We notice an unexpected similarity to the politics dataset as the effects of the post-election unrest are also evident in the sports group dynamics. During the months following the election, paths in sports become disrupted and for seven moths following the election event no paths form. In addition, the post-election unrest is followed by a sudden decline of an immensely popular sports page. More specifically, the sports page of the Fars News Agency initially appears as a top preferred source in nearly all the paths, yet after the election it does not emerge in any of the paths. The reason for this decline becomes clear if one considers the political orientation of the new agency as a whole, which was a staunch supporter of the incumbent president, running regular articles against the popular Green Movement.This trend suggests that users intentionally boycotted the sports page of the news agency, demonstrating that the political climate had touched spaces that were previously apolitical. Closer inspection of representative articles shows that in some cases users shared screenshots of the sit. In discussion with one of the regular users of the site we found this was intentional and the screenshots were posted in order to avoid boosting the sites visits. These observations can lead to further investigations about the specifics of such collective activism in unexpected

spaces on the web.

6.4.3 Concerns about a Gestalt Approach

A number of important concerns can be raised when considering this approach. First is that this approach will focus on the large structure at the expense of complexities and sophistication that exist in the real world. The simple answer is that we are not seeking to reduce the complex and multifaceted actions and interactions in society to a single structure. This is why a layered structure that allows further exploration of nuances is important. Neither are we presenting that structure as the only correct or true version of reality. On the contrary, this structure is one of the possible structures, one way to extract meaning, and an entryway to further investigation and inquiry into more complex questions.

Another important consideration is whether we should be combining individual actions which may have different meanings and purposes. Detecting meaning and intention of actions is very difficult if not impossible in its purest form. This concern is an important practical one to take it into account. For example, the action that we have considered here is a vote; thus an underlying assumption in this work is that a vote implies a preference for the article that is posted. Yet if some users vote at random, others vote based on friendship, or vote with intent to manipulate, the votes will not be combinable. Consequently, the results will become muddled and we may need additional data on users to separate these actions[3]. Nevertheless, votes in this website appear to be clearly defined and if not all users, a large enough number of them are voting based on their support of the content. It is therefore important to choose actions that are simple and well-defined so that once they are aggregated they will demonstrate the trends as closely as possible.

[3]Deriving such structures from language can be even more precarious, and even harder is the discovery of user perceptions based on language with tools like sentiment analysis.

In the end, the central question is whether there is value in studying this vast footprint of human actions or not. To the author of this manuscript, it seems foolish to ignore the possibilities that may come out of investigating this data. Thus the goal of this work was to advocate a practical path to enter the study of large data on user behavior in an attempt to bridge the gap between those who study big data and those who have been studying human behavior and society in the real world.

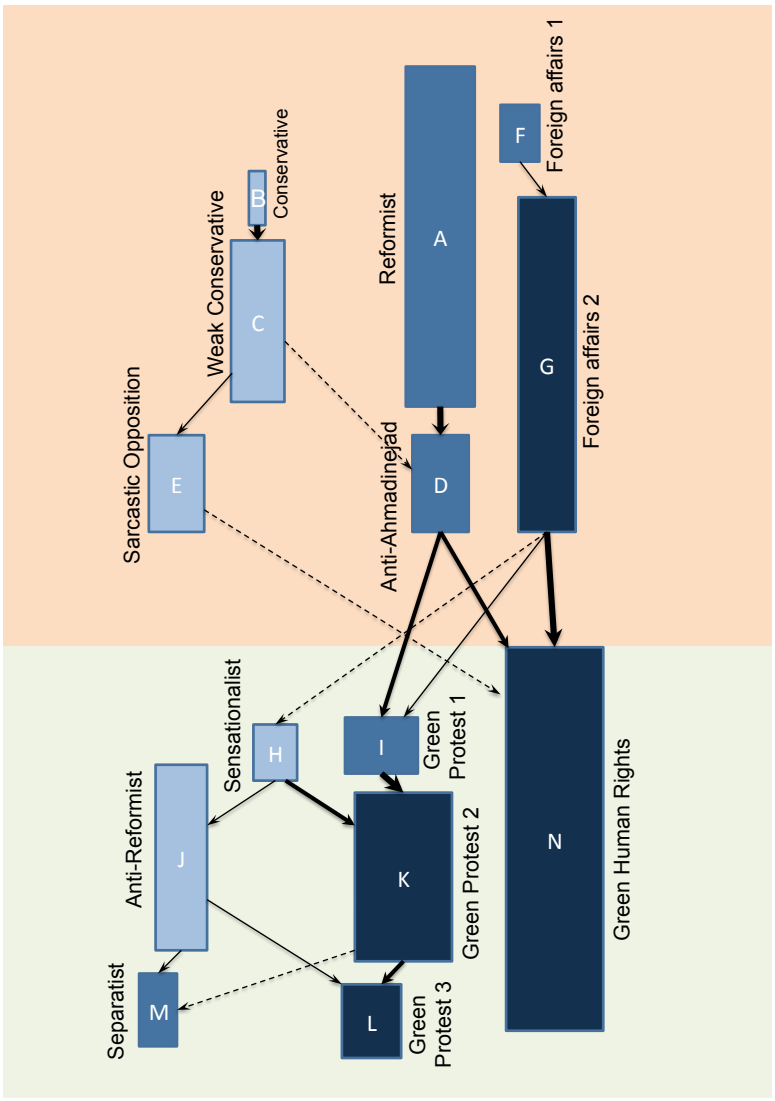


Figure 6.6: This figure summarizes several measurements relating to community evolution paths. Time begins on top and progresses downward with the change in the background color marking the June 2009 presidential election in Iran. Path width corresponds with the number of unique users in the path and arrows mark inter-path migrations. The darkness of each path marks its user retention. A highly simplified description of each path is noted next to it in rotated text.

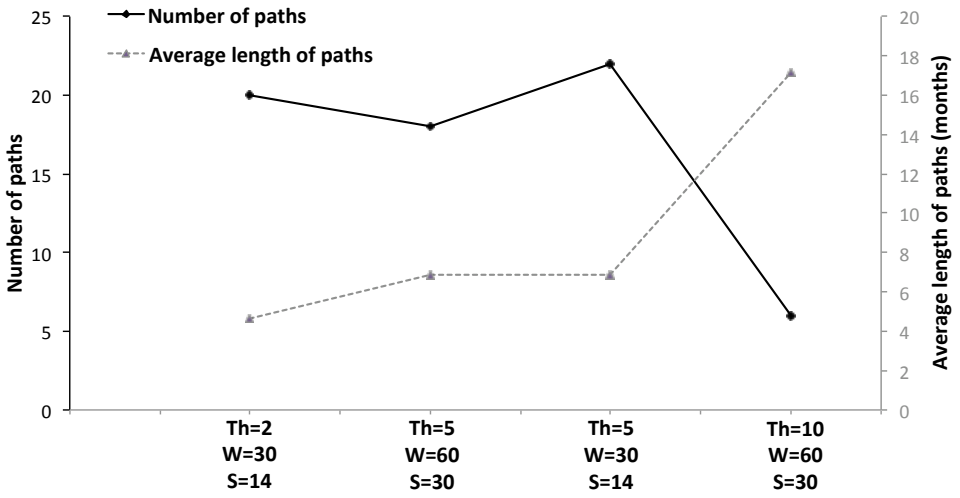
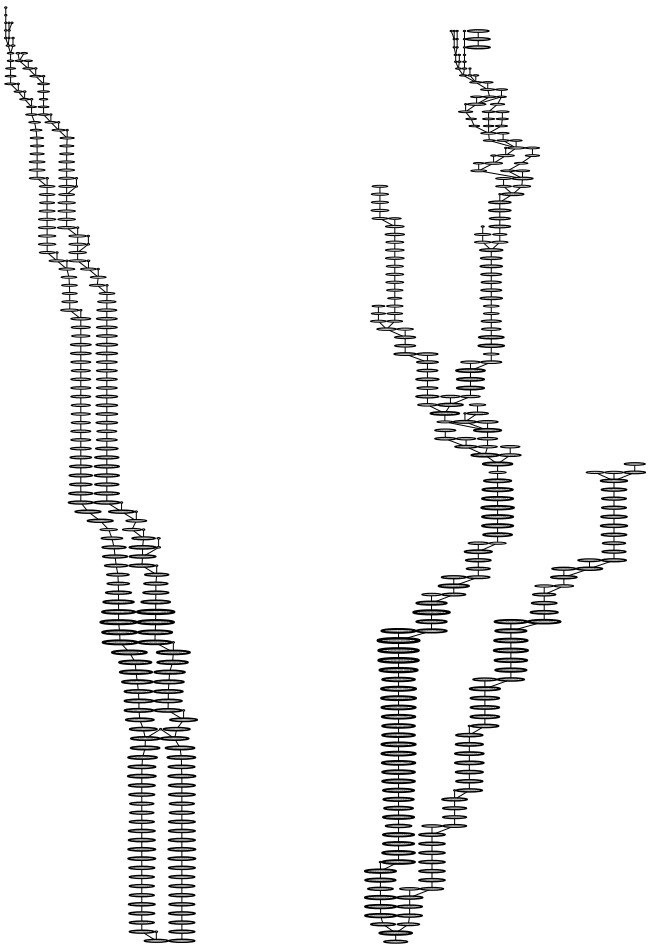


Figure 6.7: Sensitivity of results to parameter choices: window length (W), shift length (S), and vote threshold (Th). Community detection is performed for overlapping windows of W days shifted S days at each time epoch. Inactive users are defined as those with less than Th votes in each window and are eliminated. Parameter sets that produce more paths offer more granular representations of the dynamics, yet the resulting paths may be shorter and less significant. We therefore choose the parameter set that produces more and at the same time longer paths (i.e. Th=5, W=30, S=14).



(a) Threshold = 10. (b) Threshold = 2.

Figure 6.8: Renditions of evolving communities for Window size = 30 Shift size= 14 with different thresholds.

(a) Threshold = 5. (b) Threshold = 10.

Figure 6.9: Renditions of evolving communities for Window size = 60 Shift size= 30 with different thresholds.

(a) Sports. (b) Society.

Figure 6.10: Renditions of evolving communities for Sports and Society sections of the website.

References

[ABD08] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. "Online LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking." In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pp. 3–12, Washington, DC, USA, 2008. IEEE Computer Society.

[ABT12] James Abello, Peter Broadwell, and Timothy R. Tangherlini. "Computational folkloristics." *Commun. ACM*, **55**(7):60–70, July 2012.

[AG05] Lada A. Adamic and Natalie Glance. "The political blogosphere and the 2004 U.S. election: divided they blog." In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pp. 36–43, New York, NY, USA, 2005. ACM.

[AHE11] Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J. Smola, and Choon Hui Teo. "Unified analysis of streaming news." In *Proceedings of the 20th international conference on World wide web*, WWW '11, pp. 267–276, New York, NY, USA, 2011. ACM.

[AHK12] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. "Effects of user similarity in social media." In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pp. 703–712, New York, NY, USA, 2012. ACM.

[Ali08] Alias-i. "LingPipe 4.1.0." http://alias-i.com/lingpipe, 2008.

[ALT08] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. "Identifying the influential bloggers in a community." In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pp. 207–218, New York, NY, USA, 2008. ACM.

[Bar07] Michael J. Barber. "Modularity and community detection in bipartite networks." *Phys. Rev. E*, **76**:066102, Dec 2007.

[Bee09] David Beer. "Power through the algorithm? Participatory web cultures and the technological unconscious." *New Media & Society*, **11**(6):985–1002, 2009.

[BGL08] Vincent D Blondel, Jean-loup Guillaume, and Etienne Lefebvre. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment*, pp. 1–12, 2008.

[BGL10] D. Boyd, S. Golder, and G. Lotan. "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter." In *hicss*, pp. 1–10, 2010.

[BL06] David M. Blei and John D. Lafferty. "Dynamic topic models." In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pp. 113–120, New York, NY, USA, 2006. ACM.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *J. Mach. Learn. Res.*, **3**:993–1022, March 2003.

[BRB05] R. Boscolo, B. A. Rezaei, P. O. Boykin, and V. P. Roychowdhury. "Functionality Encoded In Topology? Discovering Macroscopic Regulatory Modules from Large-Scale Protein-DNA Interaction Networks." *eprint arXiv:q-bio/0501039*, January 2005.

[BRB10] C. Betsch, F. Renkewitz, T. Betsch, and C. Ulshofer. "The influence of vaccine-critical websites on perceiving vaccination risks." *J Health Psychol*, **15**(3):446–455, Apr 2010.

[CHK10] Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, Xiangyang Lan, and Siddharth Suri. "Sequential Influence Models in Social Networks." In *4th International Conference on Weblogs and Social Media*, 2010.

[CKT06] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. "Evolutionary clustering." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pp. 554–560, New York, NY, USA, 2006. ACM.

[CL11] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines." *ACM Transactions on Intelligent Systems and Technology*, **2**:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[CNM04a] A. Clauset, M. E. J. Newman, and C. Moore. "Finding community structure in very large networks." *Physical Review E*, **70**(6):066111, December 2004.

[CNM04b] Aaron Clauset, M. Newman, and Cristopher Moore. "Finding community structure in very large networks." *Physical Review E*, **70**(6):1–6, December 2004.

[Dah07] Lincoln Dahlberg. "Rethinking the fragmentation of the cyberpublic: from consensus to contestation." *New Media & Society*, **9**(5):827–847, 2007.

[DLP03] Kushal Dave, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: opinion extraction and semantic classification of product reviews." In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pp. 519–528, New York, NY, USA, 2003. ACM.

[FCB10] Gary L. Freed, Sarah J. Clark, Amy T. Butchart, Dianne C. Singer, and Matthew M. Davis. "Parental Vaccine Safety Concerns in 2009." *Pediatrics*, **125**(4):654–659, 2010.

[FKB11] C. Fountain, M. D. King, and P. S. Bearman. "Age of diagnosis for autism: individual and community factors across 10 birth cohorts." *J Epidemiol Community Health*, **65**(6):503–510, Jun 2011.

[For10a] Santo Fortunato. "Community detection in graphs." *Physics Reports*, **486**(3–5):75–174, 2 2010.

[For10b] Santo Fortunato. "Community detection in graphs." *Physics Reports*, **486**(35):75 – 174, 2010.

[Fre10] Deen G. Freelon. "Analyzing online political discussion using three models of democratic communication." *New Media & Society*, **12**(7):1172–1190, 2010.

[FSS12] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. "Mining contrastive opinions on political texts using cross-perspective topic model." In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pp. 63–72, New York, NY, USA, 2012. ACM.

[GHS09] André Gohr, Alexander Hinneburg, Rene Schult, and Myra Spiliopoulou. "Topic Evolution in a Stream of Documents." In *SDM*, pp. 859–872. SIAM, 2009.

[GLG04] Daniel Gruhl, David Liben-Nowell, Ramanathan V. Guha, and Andrew Tomkins. "Information diffusion through blogspace." *SIGKDD Explorations*, **6**(2):43–52, 2004.

[GN02] M Girvan and M E J Newman. "Community structure in social and biological networks." *Proceedings of the National Academy of Sciences of the United States of America*, **99**(12):7821–6, June 2002.

[GPL06] Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. "Text mining for product attribute extraction." *SIGKDD Explor. Newsl.*, **8**(1):41–48, June 2006.

[GS04] Thomas L. Griffiths and Mark Steyvers. "Finding scientific topics." *Proceedings of the National Academy of Sciences of the United States of America*, **101**(Suppl 1):5228–5235, 2004.

[GSA07] Roger Guimerá, Marta Sales-Pardo, and Luís A. Nunes Amaral. "Module identification in bipartite and directed networks." *Phys. Rev. E*, **76**:036102, Sep 2007.

[GV12] M. Giatsoglou and A. Vakali. "Capturing Social Data Evolution via Graph Clustering." *Internet Computing, IEEE*, **PP**(99):1, 2012.

[GZV12] Kavita Ganesan, ChengXiang Zhai, and Evelyne Viegas. "Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions." In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pp. 869–878, New York, NY, USA, 2012. ACM.

[HFH09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." *SIGKDD Explor. Newsl.*, **11**:10–18, November 2009.

[Hil73] M. O. Hill. "Diversity and Evenness: A Unifying Notation and Its Consequences." *Ecology*, **54**(2):427–432, 2012/08/06 1973.

[Him10] Itai Himelboim. "Civil Society and Online Political Discourse: The Network Structure of Unrestricted Discussions." *Communication Research*, 2010.

[Hin09] M.S. Hindman. *The Myth of Digital Democracy*. Princeton University Press. Princeton University Press, 2009.

[HRW09] B.A. Huberman, D.M. Romero, and F. Wu. "Social networks that matter: Twitter under the microscope." *First Monday*, **14**(1):8, 2009.

[HTF08] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, 2008.

[JA08] Maojin Jiang and Shlomo Argamon. "Exploiting subjectivity analysis in blogs to improve political leaning categorization." In *SIGIR*, pp. 725–726, 2008.

[JHL11] Yookyung Jo, John E. Hopcroft, and Carl Lagoze. "The web of topics: discovering the topology of topic evolution in a corpus." In *Proceedings of the 20th international conference on World wide web*, WWW '11, pp. 257–266, New York, NY, USA, 2011. ACM.

[JR09] S. Jamali and H. Rangwala. "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis." In *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, pp. 32 –38, nov. 2009.

[JSF07] A. Java, X. Song, T. Finin, and B. Tseng. "Why we twitter: understanding microblogging usage and communities." In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65. ACM, 2007.

[KKC11] Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho. "Predicting the Virtual Temperature of Web-Blog Articles as a Measurement Tool for Online Popularity." In *IEEE 11th International Conference on Computer and Information Technology (CIT)*, pp. 449 –454, 31 2011-sept. 2 2011.

[KKT03a] D. Kempe, J. Kleinberg, and É. Tardos. "Maximizing the spread of influence through a social network." In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. ACM New York, NY, USA, 2003.

[KKT03b] David Kempe, Jon M. Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." In *KDD*, pp. 137–146. ACM, 2003.

[KLN11] Allison Kennedy, Katherine LaVail, Glen Nowak, Michelle Basket, and Sarah Landry. "Confidence About Vaccines In The United States: Understanding Parents Perceptions." *Health Affairs*, **30**(6):1151–1159, 2011.

[KLP10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a Social Network or a News Media?" In *WWW'10: Proceedings of the 19th International World Wide Web Conference*, April 2010.

[KOW08] Jan Kostka, YvonneAnne Oswald, and Roger Wattenhofer. "Word of Mouth: Rumor Dissemination in Social Networks." In AlexanderA. Shvartsman and Pascal Felber, editors, *Structural Information and Communication Complexity*, volume 5058 of *Lecture Notes in Computer Science*, pp. 185–196. Springer Berlin Heidelberg, 2008.

[KSB11] Michael Kaschesky, Pawel Sobkowicz, and Guillaume Bouchard. "Opinion mining in social media: modeling, simulating, and visualizing political opinion formation in the web." In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, dg.o '11, pp. 317–326, New York, NY, USA, 2011. ACM.

[LAH07a] J. Leskovec, L.A. Adamic, and B.A. Huberman. "The dynamics of viral marketing." *ACM Transactions on the Web (TWEB)*, **1**(1):5, 2007.

[LAH07b] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. "The dynamics of viral marketing." *TWEB*, **1**(1), 2007.

[LBK09] Jure Leskovec, Lars Backstrom, and Jon M. Kleinberg. "Meme-tracking and the dynamics of the news cycle." In *KDD*, pp. 497–506. ACM, 2009.

[LG10a] K. Lerman and R. Ghosh. "Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks." *4th International Conference on Weblogs and Social Media*, 2010.

[LG10b] Kristina Lerman and Rumi Ghosh. "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks." In *ICWSM*. The AAAI Press, 2010.

[LH10] Kristina Lerman and Tad Hogg. "Using a model of social dynamics to predict popularity of news." In *WWW*, pp. 621–630. ACM, 2010.

[LHC05] Bing Liu, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the Web." In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pp. 342–351, New York, NY, USA, 2005. ACM.

[LHM02] Peter Levine, R. Hayduk, and K. Mattson. *Can the Internet Rescue Democracy? Toward an On-Line Commons*, pp. 121–137. Rowman & Littlefield, Lanham, MD, 2002.

[LKG07] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. "Cost-effective outbreak detection in networks." In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 429. ACM, 2007.

[LMF07] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. "Cascading behavior in large blog graphs." In *In SDM*, 2007.

[LMS10] Jong Gun Lee, Sue Moon, and Kavé Salamatian. "An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors." In *Web Intelligence*, pp. 623–630. IEEE, 2010.

[LSK06] J. Leskovec, A. Singh, and J. Kleinberg. "Patterns of influence in a recommendation network." *Advances in Knowledge Discovery and Data Mining*, pp. 380–389, 2006.

[LSK09] Y.R. Lin, H. Sundaram, and A. Kelliher. "Jam: Joint action matrix factorization for summarizing a temporal heterogeneous social network." In *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[LZM10] Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, and Jiawei Han. "PET: a statistical model for popular events tracking in social communities." In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pp. 929–938, New York, NY, USA, 2010. ACM.

[McC02] Andrew Kachites McCallum. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu, 2002.

[MM12] Panagiotis T. Metaxas and Eni Mustafaraj. "Social Media and the Elections." *Science*, **338**(6106):472–473, 10 2012.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, 2008.

[MZ05] Qiaozhu Mei and ChengXiang Zhai. "Discovering evolutionary theme patterns from text: an exploration of temporal text mining." In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pp. 198–207, New York, NY, USA, 2005. ACM.

[New04] M. E. J. Newman. "Analysis of weighted networks." *Phys. Rev. E*, **70**(5):056131, Nov 2004.

[Pap04] Zizi Papacharissi. "Democracy online: civility, politeness, and the democratic potential of online political discussion groups." *New Media & Society*, **6**(2):259–283, 2004.

[PBV07] Gergely Palla, Albert-Laszlo Barabasi, and Tamas Vicsek. "Quantifying social group evolution." *Nature*, **446**(7136):664–667, 04 2007.

[PL04] Bo Pang and Lillian Lee. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts." In *Proceedings of the ACL*, pp. 271–278, 2004.

[RMK11] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter." In *Proceedings of the 20th international conference on World wide web*, WWW '11, pp. 695–704, New York, NY, USA, 2011. ACM.

[Rog95] E.M. Rogers. *Diffusion of innovations*. Free Pr, 1995.

[SDW06] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science*, **311**(5762):854–856, 02 2006.

[SGH12] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. "Trains of thought: generating information maps." In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pp. 899–908, New York, NY, USA, 2012. ACM.

[SH10] Gábor Szabó and Bernardo A. Huberman. "Predicting the popularity of online content." *Commun. ACM*, **53**(8):80–88, 2010.

[SR08] M. V. Simkin and V. P. Roychowdhury. "A theory of web traffic." *EPL (Europhysics Letters)*, **82**(2):28006, April 2008.

[SRM09] Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas M. Lento. "Gesundheit! Modeling Contagion through Facebook News Feed." In *ICWSM*. The AAAI Press, 2009.

[SWB49] C.E. Shannon, W. Weaver, R.E. Blahut, and B. Hajek. *The mathematical theory of communication*, volume 117. University of Illinois press Urbana, 1949.

[Tan13] Timothy R. Tangherlini. "The Folklore Macroscope: Challenges for a Computational Folkloristics." *Western Folklore*, **72**(1), 2013.

[TBK07] Chayant Tantipathananandh, Tanya Berger-Wolf, and David Kempe. "A framework for community identification in dynamic social networks." In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pp. 717–726, New York, NY, USA, 2007. ACM.

[TLA11] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. "Predicting the popularity of online articles based on user comments." In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pp. 67:1–67:8, New York, NY, USA, 2011. ACM.

[VB05] Marshall Van Alstyne and Erik Brynjolfsson. "Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities." *Management Science*, **51**(6):851–868, 06 2005.

[VOD06] Alexei Vázquez, João Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. "Modeling bursts and heavy tails in human dynamics." *Phys. Rev. E*, **73**:036127, Mar 2006.

[Wat02] D.J. Watts. "A simple model of global cascades on random networks." *Proceedings of the National Academy of Sciences of the United States of America*, **99**(9):5766, 2002.

[Wes98] Tracy Westen. "Can Technology Save Democracy?" *National Civic Review*, **87**(1):47–56, 1998.

[WGG12] Robert E. Wilson, Samuel D. Gosling, and Lindsay T. Graham. "A Review of Facebook Research in the Social Sciences." *Perspectives on Psychological Science*, **7**(3):203–220, 2012.

[WH07] Fang Wu and Bernardo A. Huberman. "Novelty and collective attention." *Proceedings of the National Academy of Sciences*, **104**(45):17599–17601, 2007.

[WHA04] F. Wu, B.A. Huberman, L.A. Adamic, and J.R. Tyler. "Information flow in social groups." *Physica A: Statistical and Theoretical Physics*, **337**(1-2):327–335, 2004.

[WSL02] R. M. Wolfe, L. K. Sharp, and M. S. Lipsky. "Content and design attributes of antivaccination web sites." *JAMA*, **287**(24):3245–3248, Jun 2002.

[WZY09] Bin Wu, Fengying Zhao, Shengqi Yang, Lijun Suo, and Hongqiao Tian. "Characterizing the evolution of collaboration network." In *Proceedings of the 2nd ACM workshop on Social web search and mining*, SWSM '09, pp. 33–40, New York, NY, USA, 2009. ACM.

[YCK11] Bei Yu, Miao Chen, and Linchi Kwok. "Toward Predicting Popularity of Social Marketing Messages." In *SBP*, volume 6589 of *Lecture Notes in Computer Science*, pp. 317–324. Springer, 2011.

[YL11] Jaewon Yang and Jure Leskovec. "Patterns of temporal variation in online media." In *WSDM*, pp. 177–186. ACM, 2011.

[ZBK10] Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, and Vwani Roychowdhury. "Information resonance on Twitter: watching Iran." In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pp. 123–131, New York, NY, USA, 2010. ACM.

[ZJZ06] Ding Zhou, Xiang Ji, Hongyuan Zha, and C. Lee Giles. "Topic evolution and social interactions: how authors effect research." In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pp. 248–257, New York, NY, USA, 2006. ACM.

[ZRM11] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. "Classifying the Political Leaning of News Articles and Users from User Votes." In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2011.

[ZWF05] R. K. Zimmerman, R. M. Wolfe, D. E. Fox, J. R. Fox, M. P. Nowalk, J. A. Troy, and L. K. Sharp. "Vaccine criticism on the World Wide Web." *J. Med. Internet Res.*, **7**(2):e17, 2005.