

**UCLA**

**UCLA Electronic Theses and Dissertations**

**Title**

Information Dynamics in Social Interactions: Hidden Structure Discovery and Empirical Case Studies

**Permalink**

<https://escholarship.org/uc/item/75p3651m>

**Author**

Zhou, Zicong

**Publication Date**

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

**Information Dynamics in Social Interactions: Hidden  
Structure Discovery and Empirical Case Studies**

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

by

**Zicong Zhou**

2013

© Copyright by  
Zicong Zhou  
2013

ABSTRACT OF THE DISSERTATION

**Information Dynamics in Social Interactions: Hidden  
Structure Discovery and Empirical Case Studies**

by

**Zicong Zhou**

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2013

Professor Vwani Roychowdhury, Chair

As collective human activity and knowledge continues to be digitized and stored, it provides an unprecedented opportunity to understand information dynamics, how they evolve, and how individuals and organizations interact to form groups and make decisions. The petabytes of data collected everyday, however, underscores the need for new computational tools to help organize and understand these vast amounts of information. The focus of this dissertation has been to develop such tools, and present empirical case studies that both establish the efficacy of the developed tools, and provide new insights into the data sets themselves. For example, (i) We analyze a publicly accessible movie database and find global patterns in the underlying collaboration dynamics, and then show how such emergent patterns can be generated from stochastic decisions made at the level of the actors; (ii) We analyze the so called Twitter revolution that was precipitated by the 2009 elections in Iran, and determine a model for the spread of news on Twitter; (iii) We analyze the various aspects of online conversations and demonstrate how they are effective in revealing information dynamics; and finally, (iv) We develop a novel methodology for Topic Models, where given a large corpus of documents, it automatically infers the underlying topics and computes a distribution of words over the computed topics.

The dissertation of Zicong Zhou is approved.

Alan Laub

Yingnian Wu

Kung Yao

Vwani Roychowdhury, Committee Chair

University of California, Los Angeles

2013

*To my parents*

## TABLE OF CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview and Motivation	1
1.2	Outline	2
<b>2</b>	<b>Network Dynamics in Loosely Connected Social Industry</b>	<b>5</b>
2.1	Introduction	5
2.2	Overview and Related Work	6
2.2.1	Basic Concepts and Definitions	7
2.2.2	Statistical Properties of Network	9
2.2.3	Models of Network Structures and Evolution	11
2.3	Structure of Collaboration Network	14
2.4	Link Dynamics	16
2.4.1	Measuring Preferential Attachment	17
2.4.2	Measuring Double Preferential Attachment	18
2.5	Nodes Dynamics	19
2.5.1	Measuring Node Dynamics	20
2.5.2	Evidence of Preferential Survival	22
2.6	Dynamic Model of Network Structure	23
2.7	Conclusion and Discussion	25
<b>3</b>	<b>Information Dynamics on Social Media</b>	<b>27</b>
3.1	Introduction	28
3.2	Related Work	30

3.3	Measurement Methodology . . . . .	32
3.3.1	Data Collection . . . . .	32
3.3.2	Coverage Estimation . . . . .	33
3.3.3	Data Cleaning . . . . .	33
3.3.4	Link Inference for Tweet Networks . . . . .	34
3.4	Friends Followers Network . . . . .	34
3.4.1	Network Structure . . . . .	35
3.4.2	Degree of Separation . . . . .	37
3.4.3	Community Structure . . . . .	39
3.5	Information Propagation . . . . .	44
3.5.1	Tweet Rate as Information Resonance . . . . .	44
3.5.2	Tweet Network as Information Propagation . . . . .	46
3.5.3	Influential Users in Information Propagation . . . . .	49
3.6	Medium of Information Propagation . . . . .	50
3.6.1	Information Propagation via F-F Network . . . . .	50
3.6.2	Information Propagation via Public Timeline . . . . .	53
3.7	Content Taxonomy of Cascades . . . . .	55
3.8	Models Based on Damped Percolation . . . . .	57
3.9	Validation on Other Topics . . . . .	61
3.10	Conclusion and Discussion . . . . .	63
<b>4</b>	<b>Model Content on Social Conversation . . . . .</b>	<b>64</b>
4.1	Introduction . . . . .	64
4.2	Related Work . . . . .	67
4.2.1	Topic Models . . . . .	67



4.2.2	Sentiment Analysis . . . . .	68
4.3	Dataset Description . . . . .	70
4.4	Topic Discovery for Social Conversation . . . . .	72
4.4.1	Choosing Number of Topics . . . . .	72
4.4.2	Labeling Topic Models . . . . .	76
4.4.3	Studying Topics Dynamics . . . . .	79
4.4.4	Modeling User Interest . . . . .	83
4.5	Sentiment Extraction for Social Conversation . . . . .	85
4.5.1	Our Approach and Result . . . . .	85
4.5.2	Topical Sentiment Analysis . . . . .	85
4.6	User Interaction Network and its Implications . . . . .	86
4.7	Sentiment-based Interaction Network and Implications . . . . .	89
4.7.1	Identify Key Users . . . . .	90
4.7.2	User Clustering . . . . .	91
4.7.3	Hierarchical User Clustering . . . . .	92
4.8	Conclusion and Discussion . . . . .	93
<b>5</b>	<b>Discover Hidden Structure on Large-scale Corpus . . . . .</b>	<b>95</b>
5.1	Introduction . . . . .	95
5.2	Related Work . . . . .	97
5.2.1	Topic Models . . . . .	97
5.2.2	Limitation of Topic Models . . . . .	99
5.2.3	Semantic Network . . . . .	104
5.3	Our Approach: Associative Network . . . . .	105
5.3.1	Our Topic Model . . . . .	105

5.3.2	Estimation of Associative Network . . . . .	106
5.3.3	Discovery of Hidden Clusters . . . . .	109
5.3.4	Topic Representation . . . . .	111
5.4	Experiments and Results . . . . .	113
5.5	Model Evaluation . . . . .	116
5.5.1	Automatic Evaluation for Topic Coherence . . . . .	117
5.5.2	External Evaluation for Document Classification . . . . .	118
5.6	Computational Cost . . . . .	119
5.7	Conclusion and Discussion . . . . .	120
<b>6</b>	<b>Conclusion . . . . .</b>	<b>121</b>
	<b>References . . . . .</b>	<b>124</b>

## LIST OF FIGURES

2.1	Cumulative degree distribution of actor/actress collaboration network . . . . .	16
2.2	External links measurement . . . . .	18
2.3	Internal links measurement . . . . .	19
2.4	Power law exponent . . . . .	20
2.5	Dynamics of nodes in actor collaboration network . . . . .	21
2.6	Dynamics of deletion rate over the years . . . . .	22
2.7	Degree distribution of removed nodes . . . . .	23
2.8	PL component in the removed nodes and all nodes . . . . .	24
3.1	Cumulative distribution of number of tweets and retweets per user. Power law fit to the data with exponents -1.92 and -1.94. . . . .	35
3.2	Cumulative in-degree and out-degree distribution of Twitter's friends followers network . . . . .	36
3.3	In-degree and out-degree comparison on Twitter friends followers net- work . . . . .	37
3.4	On Twitter, on average a user could reach 91.3% of others within 4 steps or shorter. For 99.8% user pairs, the shortest distance is 5 or shorter.	39
3.5	Modularity as the number of steps to merge communities . . . . .	42
3.6	Dendrogram shows hierarchical community structure . . . . .	43
3.7	Plot of social connections between a group of users from Eastern Asian	44
3.8	Plot of three communities extracted from social connections between a group of users from Eastern Asian. Further tweets studies show three groups of users are Chinese, Japanese and Korean speaking respectively	44

3.9	Number of tweets by day from June 1 2009 to Aug 1 2009. The rate gradually increased as the events unfolded in Iran and the use of Twitter provoked attention, spiking dramatically in relation to political events inside Iran as well as in relation to new events and incidents particular to the web. . . . .	45
3.10	Cumulative distribution of out-degree in tweet network. Power law fit to the data with exponents -2.33 . . . . .	47
3.11	Top ten common nontrivial cascade shapes ordered by the frequency. For each graph we show the number of nodes, the number of edges and frequency. . . . .	48
3.12	Real cascades observed (a) 'StopAhmadi' wrote: Please @Twitter and @ev don't take down Twitter, for the iranian ppl #iranelection (b) 'Re-alTalibKweli' wrote: Pray for the protesters in Iran. Regardless of your politics (c) 'Stephenfry' wrote: Functioning Iran proxies 218.128.112.18:8080 218.206.94.132:808 218.253.65.99:808 219.50.16.70:8080 #iranelection - feel free to RT . . . . .	48
3.13	Cumulative distribution of cascade size and audience size. More than 10% of the cascades have 10k recipients or more although more than 99% of the cascades have size less than 20. . . . .	49
3.14	Percentage of followers' retweets. As the whole issue provoked attention, the percentage dropped and approached to 63.7% in the end. . . .	52
3.15	Cumulative distribution of retweet rate decays with a stretched-exponential law. . . . .	52
3.16	Number of tweets versus percentage of nonfollowers' retweets per day. Once the number of tweets posted exceeded 10k per day, the percentage of nonfollowers' retweets increased by 10%. . . . .	53
3.17	Larger cascades included more retweets of nonfollowers' retweets. . . .	54

3.18	Cumulative distribution of in-degree and out-degree in the event-specific F-F network. . . . .	58
3.19	Comparison of the real data and the our model based on damped percolation. We plotted the distribution of the real cascades with circles and the simulation of our model with plus signs . . . . .	61
3.20	Cascade size and indegree distribution of event-specific F-F network in death of Michael Jackson and Swine Flu breakout . . . . .	62
4.1	The number of posts per thread and the number of posts per user follow power law distribution . . . . .	70
4.2	The number of posts per week shows the temporal information dynamics on the site . . . . .	71
4.3	The number of joined users per week shows the temporal user dynamics on the site . . . . .	71
4.4	Perplexity per word under different number of topics . . . . .	73
4.5	Overall topic distribution . . . . .	76
4.6	Temporal dynamics of Topic 22 (in green) and Topic 20 (in blue) . . . . .	81
4.7	Topic users . . . . .	84
4.8	Four user clusters represented by topic centers according on K-means . . . . .	84
4.9	Temporal dynamics of sentiment by week . . . . .	86
4.10	Sentiment score for 25 topics . . . . .	87
4.11	Out-Degree distribution in user interaction network . . . . .	88
4.12	Distribution of user activity time . . . . .	88
4.13	The temporal dynamics of posts from each community in user interaction network by probability and by counts . . . . .	89

4.14	Distribution of positive edge weights which fits a power law with exponent 2.61 . . . . .	90
4.15	The temporal dynamics of posts for each community in user sentiment-based network by probability and by counts . . . . .	92
4.16	The temporal dynamics of posts for each subcommunity of community 4 by probability and by counts . . . . .	92
5.1	Word topic distribution as the number of topic is 2 . . . . .	100
5.2	Word topic distribution as the number of topic is 3 . . . . .	100
5.3	Document classification error versus the average document length . . .	103
5.4	Word topic distribution for top 10 words in scientific documents . . . .	103
5.5	The binomial z score cutoff versus the proportion of largest community	111
5.6	Automatic evaluation of topic coherence in four corpus . . . . .	118
5.7	External evaluation for document classification . . . . .	119
5.8	Comparsion of computational cost for four corpus . . . . .	120

## LIST OF TABLES

2.1	Table of definitions . . . . .	26
3.1	Twitter friends followers network statistics . . . . .	36
3.2	Hierarchical community structure . . . . .	42
3.3	Influential users in information propagation . . . . .	51
4.1	25 topics represented by words topic probability . . . . .	75
4.2	25 topics represented by normalized probability . . . . .	78
4.3	25 topics represented by binomial z score . . . . .	80
4.4	Top top sites referred by HPV posts . . . . .	83
5.1	A simple corpus consists of four documents . . . . .	99
5.2	Statistics of four experiment dataset . . . . .	113
5.3	Top 10 topics discovered on Cafemom associative network . . . . .	114
5.4	Top 10 topics discovered on Cafemom using LDA . . . . .	114
5.5	Top 10 topics discovered on NSF associative network . . . . .	115
5.6	Top 10 topics discovered on Reuters associative network . . . . .	115
5.7	Top 10 topics discovered on IranElection associative network . . . . .	116

## ACKNOWLEDGMENTS

First, it is my great fortune to work with Professor Vwani Roychowdhury throughout my doctoral work. I would like to thank him for his advice and support. This thesis would not have been possible without him and without freedom and encouragement he has given me over the last six years at UCLA.

I would like to thank my thesis committee members, Professor Kung Yao, Professor Alan Laub and Professor Yingnian Wu for their advice and comments.

I have had an amazing group of collaborators and colleagues. I also learned a lot from our discussions, particularly Michael Wu, Hai Qian, Joseph Kong, Roja Bandari and Lichao Chen. Working with them was a great experience and I feel extremely lucky to be surrounded by these outstanding individuals.

Finally, I would like to thank my parents for endless love, encouragement, advice and support. I wouldn't be able to finish this work without their love and support.



## VITA

- 2007            B.S. (Information Engineering), Zhejiang University.
- 2007            University Fellowship, UCLA
- 2009            M.S. (Electrical Engineering), UCLA.
- 2011            Henry Samueli Excellence in Teaching Award, UCLA
- 2009–2013     Teaching Assistant/Associate/Fellow, Electrical Engineering Department, UCLA.
- 2007–2013     Graduate Student Researcher, Electrical Engineering Department, UCLA.

## PUBLICATIONS

*Information Resonance on Twitter: Watching Iran.* Z. Zhou, R. Bandari, J.S. Kong, H. Qian, and V.P. Roychowhury. In KDD Workshop on Social Media Analytics (SOMA '10), ACM, 2010.

# CHAPTER 1

## Introduction

### 1.1 Overview and Motivation

As Internet is fast becoming a natural part of everyday life, we are facing with an ever-growing amount of available data that can no longer be handled without new computational tools. A 2010 study by the International Data Corporation estimated that the world generated 800,000 petabytes of digital information in 2009, and that we were on track to generate 1.2 zettabytes in 2010.

These numbers, however, do not necessarily mean that the amount of available information has increased at the same rate. We know data is simply a record of events that took place. It is the raw data that described what happen, when, where, how, whos involved, etc. However, the fallacy of big data is that more data doesn't mean you will get proportionately more information. In fact, the more data you have, the less information you gain as a proportion of the data. That means the information you can extract from any big data is asymptotically a diminishing return as your data volume increases.

In this area of big data, the first challenges we have to cope with is the heterogeneity of data. Traditional database technology requires an a priori knowledge of what data can be expected. All too often, data is available already, but it is spread over different sources, is in many different formats, and is often incomplete. For the Internet, such a structured index is not an option. The information sources are simply too diverse to capture in one index, the data is not stable, the content can constantly be changed, and the amount of data is unprecedented. How to deal with incomplete data or semi-

structure data is becoming a major question for researchers to solve.

At the same time, the area of big data also provides an unprecedented opportunity to develop new computational tools to find useful and novel patterns and structure in large amounts of data. We consider the Internet as the realization of an old artificial intelligence dream: a database storing the collective knowledge base of humankind. Researchers have already shown that publicly available and collaboratively generated information repositories can be used to semantically enrich information retrieval queries.

The focus of this dissertation has been to develop such tools, and present empirical case studies that both establish the efficacy of the developed tools, and provide new insights into the data sets themselves.

## **1.2 Outline**

In Chapter 2, we study the network dynamics in loosely connected social industry. Although the online social network have made us more densely networked than ever, researcher shows human have never been lonelier. Therefore understanding the network dynamics in a loosely connected social network becomes a very important topic. Towards this goal, we focus on collaboration network between actors based by analyzing a publicly accessible movie database and find global patterns in the underlying collaboration dynamics. We study the emergent patterns that exists in this collaboration network and developed models to explain these phenomenon. In particular, we present a microscopic analysis of the edge-by-edge evolution as well as node evolution for this large scale collaboration network. From empirical data, we show how such emergent patterns can be generated from stochastic decisions made at the level of the actors. These findings are vital to a range of important applications, from the development of better collaboration recommendation algorithms, to designing better systems for social forums that address different aspects for the online society.

In Chapter 3, we study information cascades dealing with specific events, such as

the Iranian election, death of Michael Jackson and the Swine Flu outbreak. Toward this, we determine the set of all active users for a topic, and analyze their status messages to build the tweet networks. The event-specific cascade size distribution is a power-law with exponent equal to  $-2.51$  and more than 98.7% of the cascades have depth less than three, and hence shallow. We found that at most 63.7% of all retweets in Iranian election (78% for the other two topics) were reposts of someone the user was following directly, thus the friendship network plays a major role. Surprisingly, more than 34% of retweets for Iranian election (around 20% for the other topics) are from the public timeline or the broadcast channel. We also study the underlying event-specific Friends-Followers network comprising only the active users, and investigate its role in determining the cascade size distribution. Our results show how real-time popular news propagates over Twitter and it can help us with link prediction as well as viral marketing.

In Chapter 4, we present our approaches to model the various aspects of online conversations to study information dynamics. In recent years, online conversation happens in various aspects of the forms, from the discussion forums, social media, to social customer relationship management. Stimulated by these changes, we study the patterns of information dynamics on online conversation in a vaccination forum, which has grown in prominence as an important resource for parents concerned with health care decisions related to their children. In particular, we applied statistical natural language processing to model the online conversations on social forums, and demonstrate how they are effective in revealing information dynamics. We also study information dynamics between users based on the user interaction networks inferred from their online activities, and find the forum is a fairly homogeneous and consensus driven community as mostly populated by anti-vaccination oriented mothers.

In Chapter 5, we develop a novel methodology for Topic Models, where given a large corpus of documents, it automatically infers the underlying topics and computes a distribution of words over the computed topics. Our approach is very different from the highly popular and widely used existing topic models: Instead of using a bag of words

model, it is inspired by how knowledge is organized in our brains as an associative network, and it exploits the idea of source coding from information theory to infer the latent networks directly from text data. We apply our algorithms on large-scale corpuses, and using automatic evaluation techniques, show that our topic organization is not only more coherent semantically, compared to the state-of-the-art Latent Dirichlet Allocation (LDA) results, but is also computationally more efficient.

Finally, Chapter 6 summarizes the contributions of this dissertation.

## **CHAPTER 2**

### **Network Dynamics in Loosely Connected Social**

### **Industry**

In this chapter, we study the network dynamics in loosely connected social industry. Although the online social network have made us more densely networked than ever, researcher shows human have never been lonelier. Therefore understanding the network dynamics in a loosely connected social network becomes a very important topic. Towards this goal, we focus on collaboration network between actors based by analyzing a publicly accessible movie database and find global patterns in the underlying collaboration dynamics. We study the emergent patterns that exists in this collaboration network and developed models to explain these phenomenon. In particular, we present a microscopic analysis of the edge-by-edge evolution as well as node evolution for this large scale collaboration network. From empirical data, we show how such emergent patterns can be generated from stochastic decisions made at the level of the actors. These findings are vital to a range of important applications, from the development of better collaboration recommendation algorithms, to designing better systems for social forums that address different aspects for the online society.

#### **2.1 Introduction**

In the recent years, a wide variety of models have been proposed for the growth of social networks to reproduce statistical properties observed in real-world data. These models are not only important to discover new mechanisms that play an important role in or-

ganic real-world networks, but also are useful in designing engineered networks and protocols. For instances, several works have shown that the network dynamic models have applications beyond merely modeling real-world systems. It has been shown that randomized protocols can be used to design and engineer systems, with peer-to-peer networks being the primary example [ALP01, SR04, SBR04].

Well-known examples of such data-inspired dynamic models, include preferential attachment and its variants [BA99, BE01, PFL02], copying [KKR99] and double preferential attachment of links [DM00]. However, these dynamics models did not consider the effect of node deletion in modeling a growing network. In the real-world, many networks exhibits significant rates of node deletions rate. For example, in actors collaboration network, actors join collaboration network when producing their first movie while depart from it when they end their careers, effectively removing themselves from collaboration networks. Therefore developing a dynamic model for networks with a significant node dynamics is very important to understand real life social dynamics.

Also from sociology perspective, studying publicly accessible movie database provides a partial but informative window into the entire social system to understand the social dynamics on loosely connected social network. In the recent years, social media and social media, have made us more densely networked than ever. Yet for all this connectivity, new research suggests that we have never been lonelier (or more narcissistic) and that this loneliness is making us mentally and physically ill. Therefore, understanding network dynamics in a loosely connected social network provides us a unique window into our social life.

## **2.2 Overview and Related Work**

In this section, we review the basic concepts and terminologies used in this dissertation and introduce all the notations. Then we survey the works on properties of networks and models to explain their structures.

## 2.2.1 Basic Concepts and Definitions

From mathematical point of view, a network is often modeled or represented as a graph. A graph  $G = (V, E)$  is defined with a vertex set  $V$  connected via an edge set  $E$ . The number of nodes and the number of edges in the network are defined by  $|V|$  and  $|E|$ . In this dissertation, we use terms vertex or node to refer to elements of the vertex set  $V$ , and similarly edge, link or connection to refer to elements of the edge set  $E$ . Now we define the terminology and several basic graph-theoretic concepts:

**Directed and undirected graph:** A graph is undirected if  $(i, j) \in E, (j, i) \in E$ , i.e., edges are unordered pairs of nodes. If pairs of nodes are ordered, i.e., edges have direction, then the graph is directed.

**Bipartite graph:** A graph  $G$  is bipartite if its vertex set can be partitioned into two disjoint sets  $V_1, V_2$ , so that there are only edges connecting nodes across the sets  $V_1$  and  $V_2$ . Or equivalently, there exist no edges between the nodes of the same partition.

**Adjacency matrix:** It indicates which of vertices in the network are connected and it is a convenient way to represent a graph  $G$ . It is a square  $N \times N$  matrix where  $N$  is the total number of vertices in the network. For directed network, its element  $A_{i,j} = 1$  if  $(i, j) \in E$  and 0 otherwise. The adjacency matrix of an undirected network is symmetrical  $A_{i,j} = A_{j,i}$ .

**Connectedness:** We say that two nodes in a network are connected if there exists an undirected path between them.

**Weakly and strongly connected graph:** A graph is connected if there is a path between all pairs of nodes in a graph. If the graph is directed, then it is weakly connected if there exists an undirected path connecting any pair of nodes. Similarly graph is strongly connected if there exists a directed path connecting any pair of nodes in a graph.

**Connected component:** A connected component or just a component is a maximal set of nodes where for every pair of the nodes in the set there exist a path connect-



ing them. Analogously, for directed graphs we have weakly and strongly connected components.

**Complete graph:** A graph is complete if all pairs of nodes are connected.

**Node degree:** We say that a node has degree  $d$  if it has  $d$  incident nodes. For directed graphs we talk about out-degree  $d_{out}$ , which is the number of edges pointing from the node. Similarly, in-degree  $d_{in}$  denotes the number of edges pointing towards the node.

**Degree distribution:** Probability distribution of these degrees over the whole network.

**Diameter:** Graph  $G$  has the diameter  $D$  if the maximum length of undirected shortest path over all connected pairs of nodes is  $D$ . The length of the path is the number of links it contains.

**Clustering coefficient:** Clustering coefficient of a node is the ratio between the total number of the edges connecting its nearest neighbours and the total number of all possible edges. Or equivalently, clustering coefficient is the fraction of triangles centered at node among the  $d(d - 1)/2$  triangles that could possibly exist.

**Betweenness:** Betweenness of a vertex is the total number of shortest paths between all possible pairs of vertices that pass through this vertex.

**Assortativity coefficient:** Assortativity refers to a preference for a network's nodes to attach to others that are similar or different in some way. Correlations between nodes of similar degree are often found in the mixing patterns of many observable networks. For instance, in social networks, highly connected nodes tend to be connected with other high degree nodes. This tendency is referred to as assortative mixing, or assortativity. On the other hand, technological and biological networks typically show disassortative mixing, or disassortativity, as high degree nodes tend to attach to low degree nodes

## 2.2.2 Statistical Properties of Network

Networks are composed of nodes and edges connecting them. Depending on the domain network data comes from they can be represented by directed or undirected networks. Examples of networks include the Internet, World Wide Web, social networks of acquaintance, collaboration or other connections between individuals, organizational networks, metabolic networks, language networks, food webs, distribution networks such as water distribution networks, blood vessels or postal delivery routes, networks of citations between papers, software networks where edges represent dependencies or function calls.

Research over the past few years has identified classes of properties that can be found in many real-world networks from various domains. While many patterns have been discovered, two of the principal ones are heavy-tailed degree distributions and small diameters.

**Degree distributions:** The degree-distribution of a graph is a power law if the number of nodes  $N_d$  of degree  $d$  is given by  $N_d \propto d^{-\gamma} (\gamma > 1)$ , where  $\gamma$  is called the power law degree exponent.

Typically for most datasets the degree exponent takes values  $2 < \gamma < 3$ . For example, in-degree distribution of web graph has  $\gamma_{in} = 2.1$  and out-degree  $\gamma_{out} = 2.4$  [BA99], while autonomous systems have  $\gamma = 2.4$  [FFF99]. However, deviations from the power law pattern have been noticed [PFL02], which can be explained by the "DGX" distribution

Most of large real-world networks have heavy-tailed or power law degree distributions, and are thus often called scale-free networks. This discovery [BA99] is important as it shows that real networks are not random (as we will more precisely define below). Moreover, in scale-free networks there are many vertices with a degree that greatly exceeds the average (a direct result of power law degree distributions). These highest-degree nodes are often called hubs, and are thought to serve specific purposes in their

networks, although this depends greatly on the domain.

**Small diameter:** Most real-world graphs exhibit relatively small diameter, which is also known as the small-world phenomenon: A graph has diameter  $d$  if every pair of nodes can be connected by a path of length at most  $d$ . The diameter  $d$  is susceptible to outliers. Thus, a more robust measure of the pairwise distances between nodes of a graph is the effective diameter. The effective diameter has been found to be small for large real-world graphs like Internet, Web, and social networks [WS98].

**Clustering coefficient:** Clustering coefficient is a measure of transitivity in networks and especially in social networks [WS98], i.e., friend of a friend is more likely to be also my friend. In many networks it is found that if node  $u$  is connected to  $v$  and  $v$  is further connected to  $w$  then there is a higher probability that node  $u$  is connected to  $w$ . In terms of network topology, transitivity means the presence of a heightened number of triangles in the network, i.e., sets of fully connected triples of nodes.

It has been found that clustering coefficient in real networks is significantly higher than for random networks (conditioned on same degree distribution). Moreover, it has been also observed [DM02] that in real networks clustering coefficient  $C_d$  decreases as the node degree  $d$  increases. Moreover,  $C_d$  scales as a power law,  $C_d \propto d^{-1}$ .

**Community structure:** A large body of work has been devoted to defining and identifying communities in social and information networks. Communities, modules or clusters are most often thought as sets of nodes that has more and/or better-connected edges between its members than between members of that set and the remainder of the network [GN02].

The problem of community identification is often formulated as unsupervised learning, some form of clustering or graph partitioning where the idea is to partition the network into disjoint but sometime also overlapping sets of nodes, where there few edges need to be cut to separate internally densely linked set of nodes, i.e., a community. For example, see the reviews on community identification [NG04]. It has been observed

that community-like sets of nodes tend to correspond to organizational units in social networks [New06], functional modules in biological networks [RSM02], and scientific disciplines in collaboration networks between scientists [GN02].

**Small World:** A small world network is a type of mathematical graph in which most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of hops or steps. Specifically, a small-world network is defined to be a network where the typical distance  $L$  between two randomly chosen nodes (the number of steps required) grows proportionally to the logarithm of the number of nodes  $N$  in the network.  $L \propto \log N$ . Last family of network models we describe here strives for small diameters and local structures, like triangles, in networks that arise from geographical proximity or homophily. Such models include the small-world model [WS98]. In a small world model one starts with a regular lattice (e.g., a grid). The lattice models local short-range links. Then for each edge with probability  $p$  we move its endpoint to a uniformly at random chosen node. The model offers a nice way of interpolating between regular ( $p = 0$ ) and random graphs ( $p = 1$ ). For low  $p$  graphs will have lots of local structure with many short range links, clustering will be high but the diameter will be also large. As one increases  $p$  long range edges will start to appear which will have the effect to destroy the local structure (clustering will decrease) but at the same time the diameter of the network will also decrease.

### 2.2.3 Models of Network Structures and Evolution

In parallel with empirical studies of large networks, there has been considerable work on models for graph generation. Both deterministic and stochastic models have been explored. Most often the models do not "force" the network to have a certain property but rather give general principles or mechanisms of edge creation that consequently lead to the global statistical property or distribution to arise in the network.

**Erdos-Renyi random graph model** The earliest probabilistic generative model for

graphs was a random graph model introduced by Erdos and Renyi [ER60]. The model states that given a number of nodes each pair of nodes has an identical, independent probability of being joined by an edge. There are two variants of the model:  $G_{n,p}$  is defined to have  $n$  nodes, and each edge appears independently with probability  $p$ . Similarly, the  $G_{n,m}$  is defined to have  $n$  nodes and  $m$  uniformly at random placed edges. There exists a close correspondence between the models, as in practice most theorems hold for both variants.

One can show that degree distribution of Erdos-Renyi random graph follows a binomial distribution with mean  $d = 2m/n$  [AB02]. Moreover, the diameter (longest shortest path) of a random graph increases with the number of nodes  $n$  as  $O(\log n)$ , and the average shortest path length grows as  $O(\log \log n)$ .

There is a rich mathematical theory about this model; however, the model is not realistic as it produces graphs that fail to match real-world networks in a number of respects (e.g., it does not produce power law degree distributions)

**Preferential attachment** The discovery of degree power laws led to the development of random graph models that exhibited such degree distributions, including the family of models based on preferential attachment [BA99]. The model operates in the following way. Nodes are arriving one at a time. And when a new node  $u$  arrives to the network it creates  $m$  edges ( $m$  is a parameter and is constant for all nodes). The edges are not placed uniformly at random but preferentially.

There are also many extensions to the preferential attachment model. We mention three of them: the fitness model, winners don't take all, and the geometric preferential attachment.

In preferential attachment model nodes that arrive early will end up having highest degrees. However, one could envision that each node has an inherent competitive factor that nodes may have, capable of affecting the networks evolution. This is called node fitness [BB01]. The idea is that intrinsic ability of a node to attract links in the network

varies from node to node. The most efficient nodes are able to gather more edges at the expense of others. In that sense, not all nodes are identical, and they claim their degree increase in the number of edges accordingly to the fitness they possess every time. Fitness parameter is usually considered as not varying over time and is multiplicative to the edge probability. [KSR08] use sequential large-scale crawl data to empirically investigate and validate the dynamics that underlie the evolution of the structure of the web. The web is conservative in judging talent and the overall fitness distribution is exponential, showing low variability. The small variance in talent, however, is enough to lead to experience distributions with high variance: The preferential attachment mechanism amplifies these small biases and leads to heavy-tailed power-law inbound degree distributions over all pages, as well as over pages that are of the same age. The balancing act between experience and talent on the web allows newly introduced pages with novel and interesting content to grow quickly and surpass older pages. In this regard, it is much like what we observe in high-mobility and meritocratic societies: People with entitlement continue to have access to the best resources, but there is just enough screening for fitness that allows for talented winners to emerge and join the ranks of the leaders. Finally, the authors show that the fitness estimates have potential practical applications in ranking query results.

In spirit similar is the winners don't take all [PFL02] model where the intuition is taken from the web. It has been observed that for web communities of interest the distribution of links no longer follows a power law but rather resembles a normal distribution [PFL02]. Based on this observation, the authors then propose a generative model that mixes preferential attachment with a baseline probability of gaining a link.

A last variant of preferential attachment that we also describe is the geometric preferential attachment [FFV04], where the idea is to incorporate geography into the preferential attachment model. Intuition is that probability of linking to a node of degree  $d$  should be higher if the node is closer rather than farther. In this model nodes belong to some underlying geometry and then each node connects preferentially to other nodes

inside some local ball of radius  $r$ . For example, one can scatter nodes uniformly on a sphere, and each node uses preferential attachment mechanism to attach to other nodes in some local neighborhood as defined by the sphere.

Last family of network models we describe here strives for small diameters and local structures, like triangles, in networks that arise from geographical proximity or homophily. Such models include the small-world model [WS98]. In a small world model one starts with a regular lattice (e.g., a grid). The lattice models local short-range links. Then for each edge with probability  $p$  we move its endpoint to a uniformly at random chosen node. The model offers a nice way of interpolating between regular ( $p = 0$ ) and random graphs ( $p = 1$ ). For low  $p$  graphs will have lots of local structure with many short range links, clustering will be high but the diameter will be also large. As one increases  $p$  long range edges will start to appear which will have the effect to destroy the local structure (clustering will decrease) but at the same time the diameter of the network will also decrease.

### **2.3 Structure of Collaboration Network**

Towards our research goal, we focus on collaboration network between actors by analyzing a publicly accessible movie database from The Internet Movie Database (IMDB). IMDB is an online database of information related to films, television programs, direct-to-video products, and video games. This includes actors, production crew personnel, and fictional characters featured in these four visual entertainment media. The website consists of one of the largest accumulations of data about these categories, reaching back to each medium's respective beginning. In many cases, the information goes beyond simple title and crew credit, but also includes data on uncredited personnel, production and distribution companies, plot summaries, memorable quotes, awards, reviews, box office performance, filming locations, technical specs, promotional content, trivia, and links to official and other websites. Furthermore, the IMDb tracks titles in

production, including major announced projects still in development.

The complete database provided by IMDB offers a great opportunity for us to study the dynamic of networks with a significant deletion component as actor joins and leave the movie industry in a swift fashion. We studied movie actor collaboration network obtained from the IMDB, which consists of more than 800,000 actors, 500,000 actresses and 700,000 movies by the research was done.

By assuming the actor/actress who cast in the same movie together know each other, we build actor/actress collaboration network <sup>1</sup> based on the movie they cast in. In actor collaboration network, each actor is represented by a vertex and actor being connected if they were cast together in the same movie. In IMDB datasets by the research was done, there are 821,649 actors, 3,731,135 pair of actors and movie. There are 532,813 actresses, 2,203,763 pair of actress and movie. We first studied the overall actor collaboration network regardless of the year of produce.

We plot the degree distribution of the actor collaboration network in Figure 2.1 and it exhibits the power law degree distribution with an exponential decay, which is similarly reported in [ASB00]. In log-log plot of the cumulative distribution, it suggests that for values of number of collaborations between 10 and 500, the data are consistent with power law decay. The apparent exponent of this cumulative distribution is equal to 1.51. For larger numbers of collaborations, the power law decay is truncated and from 2000 to 30,000 we plot the cumulative distribution of actor collaboration network on linear-log scale. The distribution falls on a straight line, indicating an exponential decay of the distribution of connection.

The exponential decay of degree distribution shows the **aging of the vertices** in actor collaboration network, which can be explained that every actor will stop acting eventually. The fact implies that even a very highly connected vertex will stop receiving new links, even through it is still part of the network and contributes to network statistics. The aging of the vertices thus limits the preferential attachment preventing a

---

<sup>1</sup>We refer actor/actress collaboration network as actor collaboration network



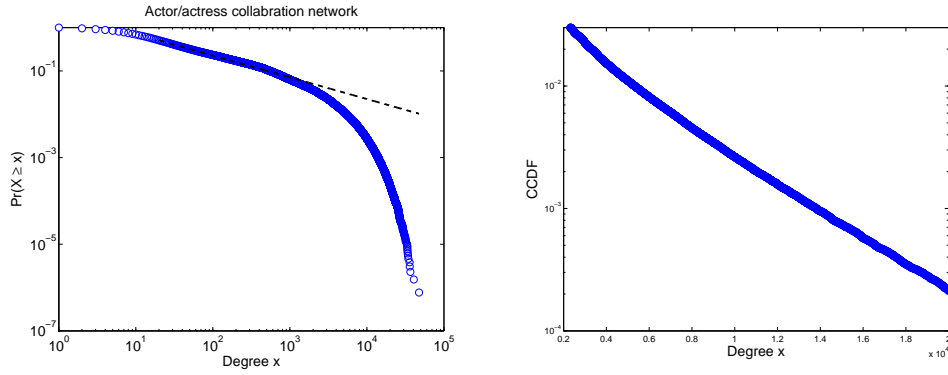


Figure 2.1: Cumulative degree distribution of actor/actress collaboration network

scale-free distribution of connection [ASB00]. The aging of vertices may occur differently for different kind of vertices, we studied this effect in actor collaboration network in order to develop dynamic model for node deletions in large scale network. In the next two sections, we perform the empirical measurement of link dynamics and node dynamics in actor collaboration network.

## 2.4 Link Dynamics

Actor collaboration networks continuously expand through the addition of new nodes and links between the nodes, while the preferential attachment [BA99, BE01, PFL02] hypothesis states that the rate  $\Pi(k)$  with which a node with  $k$  links acquires new links is a monotonically increasing function of  $k$ . We are going to measure the time evolution of degree  $k_i$  of node  $i$ , which can be obtained from

$$\frac{dk_i}{dt} = m\Pi(k) \quad (2.1)$$

where  $m$  is constant and  $\Pi(k) = \frac{k_i^\alpha}{\sum_j k_j^\alpha}$  with  $\alpha > 0$  is an unknown scaling exponent we want to measure.

For  $\alpha = 1$  these models reduce to the scale-free model [BA99], for which the degree distribution  $P(k)$ , giving the probability that a node has  $k$  links, follows  $P(k) \propto k^{-\gamma}$

with  $\gamma = 3$ .

### 2.4.1 Measuring Preferential Attachment

To measure  $\Pi(k)$  we consider a network for which we know the order in which each node and link joins the system. According to Eq. 2.1 the function  $\Pi(k)$  gives the rate at which an existing node with  $k$  links acquires new links as the network grows. To measure  $\Pi(k)$  we need to monitor to which old node new nodes link to, as function of the degree of the old node. This measurement of the preferential attachment can be broke down into following steps:

1. Define existing nodes in the system at time  $T_0$ , called “ $T_0$  nodes”.
2. Select a group of “ $T_1$  nodes”, added between  $[T_1, T_1 + \Delta T]$ , where  $\Delta T \ll T_1$  and  $T_1 > T_0$ .
3. When a  $T_1$  node joins the system, we record the degree  $k$  of the  $T_0$  node to which the new node links to. Then the probability of degree  $k$  node  $T_0$  get links can be used to measure  $\Pi(k)$ . Since  $m$  is constant, the probability of degree  $k$  node to get links is proportional to  $\Pi(k)$ . Therefore we can find:

$$\Pi(k) \propto \frac{\text{number of links aquired by } T_0 \text{ nodes with exactly } k \text{ degree}}{\text{number of nodes with degree } k}$$

4. To avoid the sparsity, we estimate  $\Pi(k)$  from its cumulative function  $K(k)$ :

$$K(k) = \int_0^k \Pi(k) dk$$

we would like to expect  $K(k) \propto k^{\alpha+1}$

We choose “ $T_0$  nodes” as the actor that debut between 1920 and 1940. And we measure preferential attachment for “ $T_1$  nodes” which is added between  $[T_1, T_1 + 1]$

where  $T_1$  is taken from 1940 to 1970. We follow the above steps for measurement and obtained  $K(k)$  as function of  $k$  in Figure 2.2. In Figure 2.2, we can see  $K(k)$  fits a straight line in log-log scale and the slope of is measured to be 2.1. Therefore the exponent  $\alpha$  in  $\Pi(k)$  is found to be 1.1. We measure the exponent  $\alpha$  from 1941 to 1970 and the average is found to be 1.0467. The above empirical result suggests linear preferential attachment offers a good approximation for actor collaboration network.

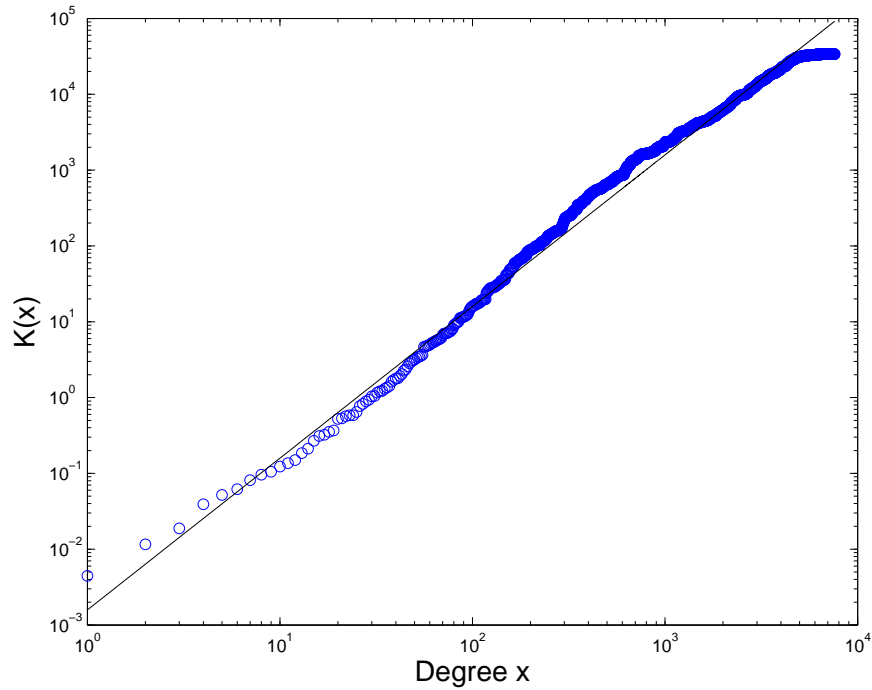


Figure 2.2: External links measurement

## 2.4.2 Measuring Double Preferential Attachment

For the actor collaboration network, the new links not only comes from new nodes added to the network, but also comes from connecting previously existing nodes as well. In this section, we focus only on new internal links, which are new links that connect two previously present but disconnected nodes.

Researchers proposed double preferential attachment [BJN02] to model the new

links between previously disconnected nodes. Double preferential attachment implies that the probability that a new internal link appears between two nodes with  $k_1$  and  $k_2$  degree scales with the product of  $k_1$  and  $k_2$ . We focus on the internal links between “ $T_0$ ” nodes and compute the probability of node with degree of  $k_1$  connecting to node with degree of  $k_2$ . We plot  $K(k_1k_2)$  as a function of  $k_1k_2$  in Figure 2.3, and it fits a straight line in log-log scale as well. The average exponent is found to be 0.9125 for actor collaboration network. This observation show the connecting preference in collaboration network and validate the internal links in the actor collaboration network follow double preferential attachment.

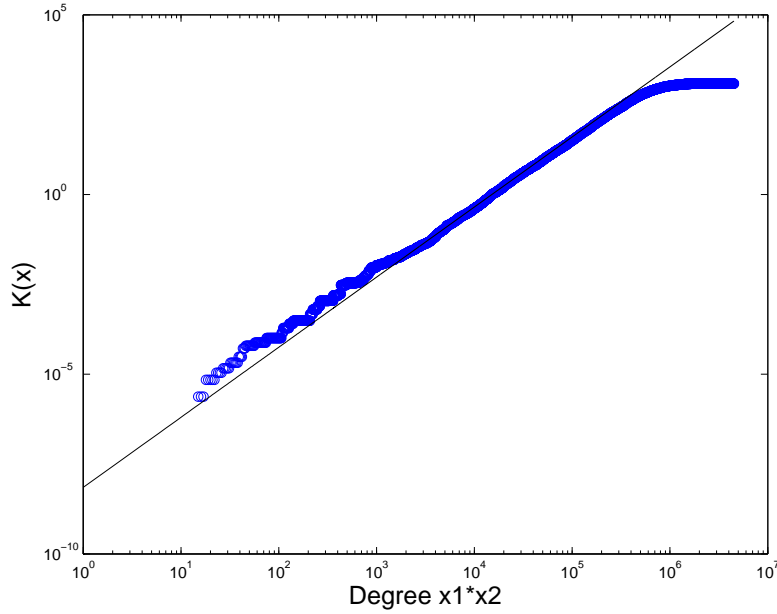


Figure 2.3: Internal links measurement

## 2.5 Nodes Dynamics

To further investigate the dynamics process of collaboration network, we consider the ad hoc characteristics of actors in collaboration network. The frequent joining and leaving of actors in collaboration network leads us to first study the evolution of collab-

oration network over time. To start with, we measured the overall power law exponent in the collaboration network and plot the dynamics of power law component in Figure 2.4. The figure shows the statistics in overall network structure is very stable despite the dynamics of links as well as nodes.

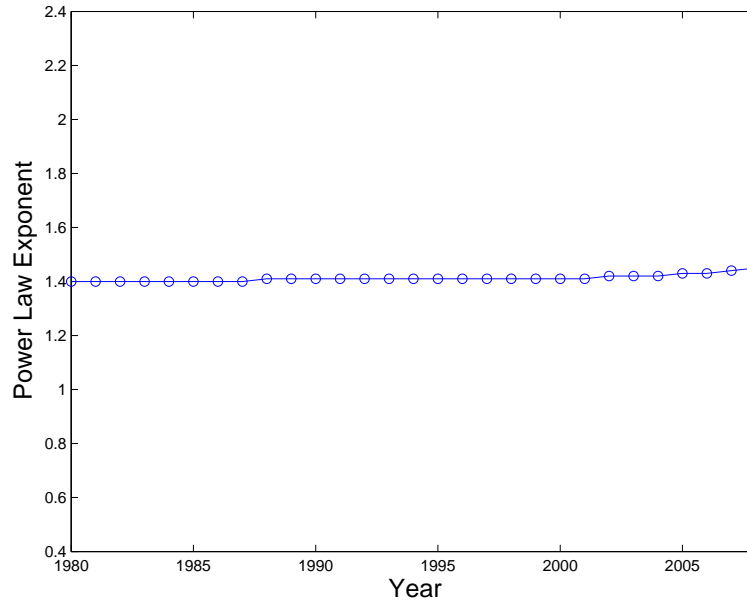


Figure 2.4: Power law exponent

### 2.5.1 Measuring Node Dynamics

To measure the node dynamics, we define the joining time of the node as the production year of actors debut movie and the removal year of the node as the production year of actors last movie. By assuming the actors who have not acted in any movies from 1990 are no longer active any more. To further classify the active from inactive nodes, we take the following steps:

1. We find the total number of nodes in historical actor collaboration network from year 1930 to year 1990. The historical actor collaboration network in year  $t$  is referred to collaboration network constructed from the actor movie information

up to year  $t$ .

2. We find the inactive nodes in historical actor collaboration network in year  $t$  by picking up actors whose last appearance is no later than year  $t$ . We consider these nodes to be inactive and will model them to be removed from in historical network.
3. We find the active nodes by subtracting the inactive nodes from all nodes.

The Figure 2.5 shows the dynamic of all nodes, active nodes and inactive nodes in historical actor collaboration network from year 1930 to year 1990.

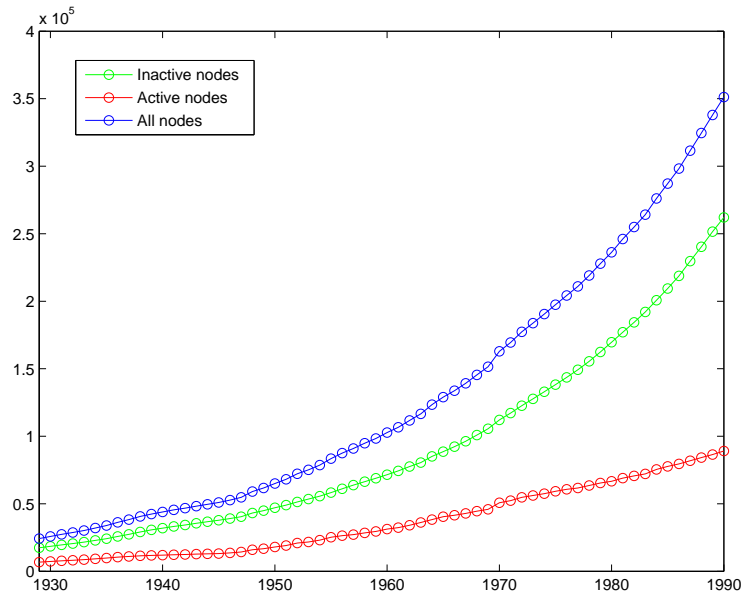


Figure 2.5: Dynamics of nodes in actor collaboration network

Deletion rate is defined as the average number of nodes removed per node added.

In our dataset, we measured the deletion rate as follows:

1. We find the number of nodes that joins the collaboration network when actor took part in their first movie.
2. We consider the inactive nodes to be removed from the collaboration network.

3. We find ratio between removed nodes and new joined nodes.

In our dataset, the deletion rate is measured to be  $c=0.74$ . Figure 2.6 shows the dynamics of deletion rate in actor collaboration network from year 1930 to year 1990.

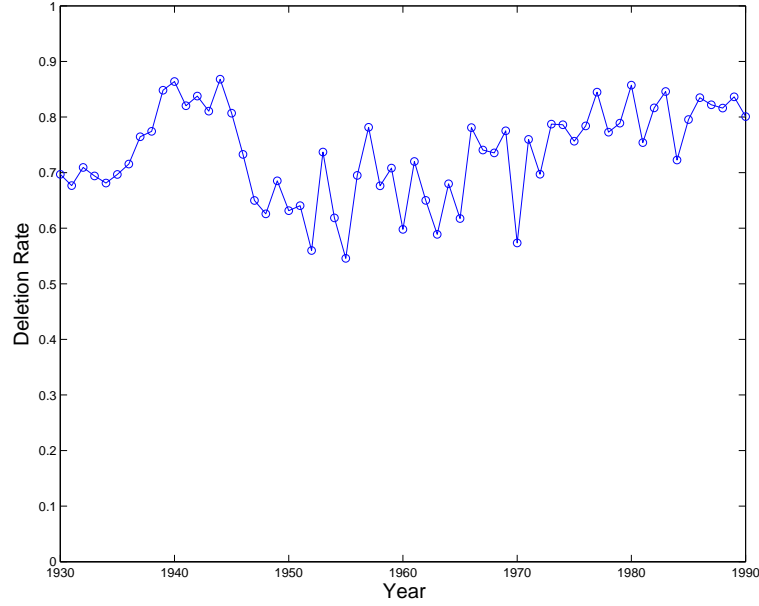


Figure 2.6: Dynamics of deletion rate over the years

## 2.5.2 Evidence of Preferential Survival

We find our dataset for direct empirical evidence of the preferential survival mechanism by studying the degree distribution of the deleted nodes of a given year in Figure 2.7.

If nodes were to be deleted uniformly randomly, the degree distribution of the set of deleted nodes would be identical to the network's degree distribution. For our dataset, we found that the power law exponent of the degree distribution of the set of deleted nodes to be  $\gamma_{del} \approx 1.62$  (Figure 2.7), which is different from the power law exponent for the entire network  $\gamma \approx 1.4$  (Figure 2.1). Our finding suggests that a node is removed according to the deletion probability kernel:  $D(k) \propto k^{-\alpha}$ , where  $\alpha = \gamma_{del} - \gamma \approx 0.2$  in our case. We will show in our model that a deletion kernel with  $\alpha = 0.2$  leads to the

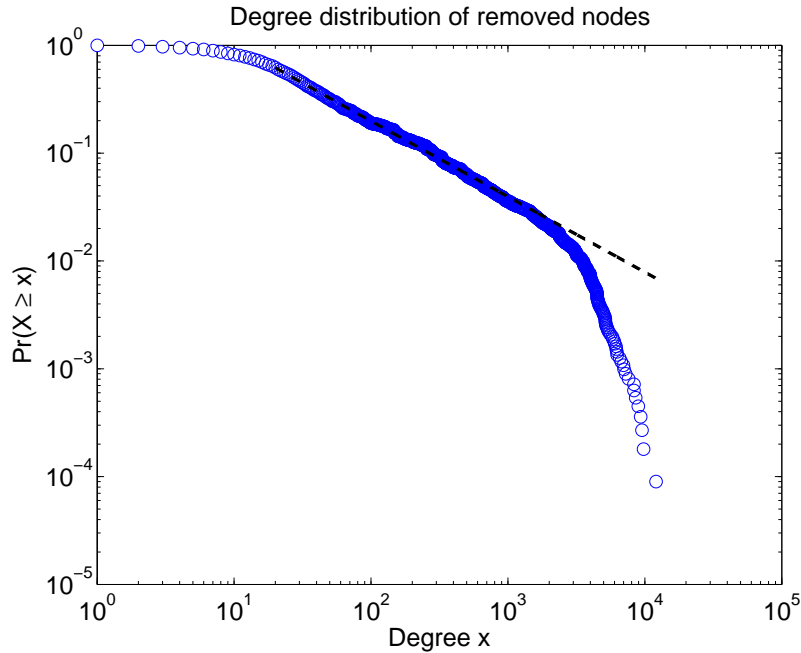


Figure 2.7: Degree distribution of removed nodes

overall degree distribution of network. Figure 2.8 shows the dynamics of power law component for deleted nodes and entire network.

## 2.6 Dynamic Model of Network Structure

Inspired by the empirical results from node dynamics and link dynamics sections, we propose a dynamic model to explain the degree distribution of actor collaboration network. The generative model is found as follows: at each time step, a node joins the network and makes  $m$  links to  $m$  nodes preferentially; with probability  $c$ , a node is chosen to be removed, according to the deletion kernel  $D(k) \propto k^{-\alpha}$ , along with all of its associated links;  $bm$  new internal edges link in a double preferential attachment principle to existing nodes. The parameter  $c$  denotes the turnover rate or the deletion rate, which is defined as the rate of node removal divided by the rate of node addition.

Each node in the network is labeled by its insertion time. Let  $D(i, t)$  be the probability that the  $i$ th node is still in the network at time  $t$ , where  $t > i$ . Note that  $D(i, t)$



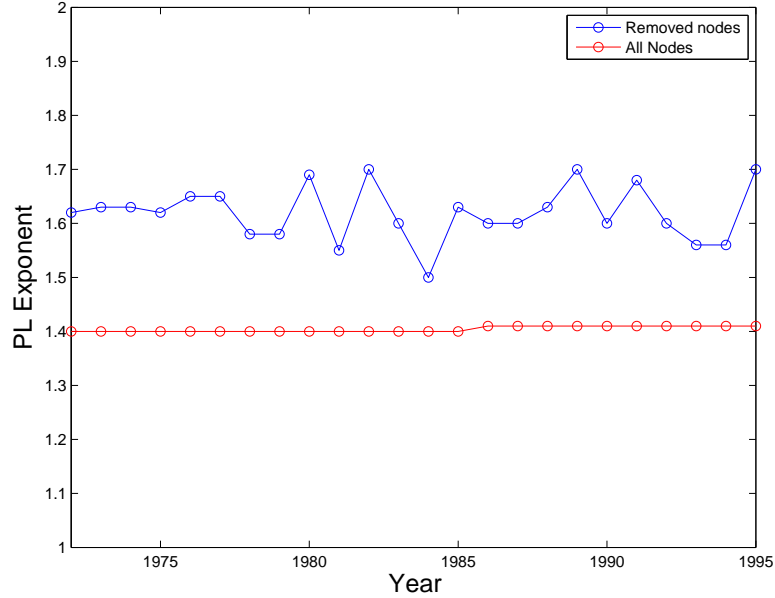


Figure 2.8: PL component in the removed nodes and all nodes

yields the lifetime distribution of node  $i$ . Then we have:

$$D(i, t + 1) = D(i, t) \left[ 1 - c \frac{k(i, t)^{-\alpha}}{N(t) \langle k^{-\alpha}(t) \rangle} \right]. \quad (2.2)$$

The initial condition is  $D(i, i) = 1$  and  $\langle k^{-\alpha}(t) \rangle = \sum_k k^{-\alpha} P(k, t)$ , which can be considered as the “ $-\alpha$ ” moment of the degree distribution at time  $t$  (see Table 2.1 for the definition of symbols).

Assuming the  $i$ th node is still in the network at time  $t$ , the evolution of its expected degree is described by the following equation:

$$\begin{aligned} \frac{\partial k(i, t)}{\partial t} &= m \frac{k(i, t)}{S(t)} - ck(i, t)P(\text{a neighbor is removed}) \\ &+ 2bm \frac{k(i, t)}{S(t)}, \end{aligned} \quad (2.3)$$

where the sum of node degrees at time  $t$  is described by  $S(t) = \langle k(t) \rangle N(t)$ , with  $\langle k(t) \rangle$  denoting the average node degree at time  $t$  and  $N(t) = (1 - c)t$  is the number of nodes

at time  $t$ .

The initial condition is:  $k(i, i) = m$ . Eq. (2.3) gives the rate at which the  $i$ th node gains connections at time  $t$ . The first term in Eq. (2.3) describes the attachments of the  $m$  preferential links as a result of the joining node; the second term denotes the deletion of node  $i$ 's neighbors according to the deletion kernel; the third term describes the appearance of  $bm$  new internal edges attaching in a double preferential manner to  $2bm$  target nodes. Furthermore, the evolution of  $S(t)$  is described by:

$$\frac{\partial S(t)}{\partial t} = 2(1+b)m - 2c\langle k_{del}(t) \rangle \quad (2.4)$$

where  $\langle k_{del}(t) \rangle$  is the average degree of a deleted node at time  $t$ .

Eq. (2.4) gives the rate of increase for the sum of node degrees at time  $t$ ; the first term on the right hand side describes the addition of  $(1+b)m$  edges, hence  $2(1+b)m$  degrees are added to the sum of degrees; the second term describes the loss of edges as a result of the removed node.

Now to calculate the power-law exponent, we note that

$$\begin{aligned} P(k, t) &= \frac{\text{No. of nodes with degree} = k}{\text{Total number of nodes}} \\ &= \frac{1}{N(t)} \sum_{i:k(i,t)=k} D(i, t) \\ &= \frac{1}{N(t)} D(i, t) \left| \frac{\partial k(i, t)}{\partial i} \right|_{i:k(i,t)=k}^{-1} \end{aligned} \quad (2.5)$$

The general model stated above appears to be very difficult to solve analytically.

## 2.7 Conclusion and Discussion

In this chapter, we study the network dynamics in loosely connected social industry. Although the online social network have made us more densely networked than ever,

<b>Var.</b>	<b>Definition</b>
$k(i, t)$	expected degree of the $i$ th node at time $t$
$S(t)$	sum of node degrees at time $t$
$N(t)$	size of the network at time $t$
$\langle k(t) \rangle$	average node degree at time $t$
$\langle k_{del}(t) \rangle$	average degree of a deleted node at time $t$
$\langle k^{-\alpha}(t) \rangle$	$\sum_k k^{-\alpha} P(k, t)$
$m$	number of connections of the joining node
$c$	turnover rate or number of nodes deleted in each time step
$b$	ratio of number of internal edges added per time step and number of connections per joining node
$\alpha$	exponent in the deletion kernel $D(k) \propto k^{-\alpha}$
$a_0$	the "-0.2" moment of the degree distribution: $\sum_k k^{0.2} P(k)$

Table 2.1: Table of definitions

researcher shows human have never been lonelier. Therefore understanding the network dynamics in a loosely connected social network becomes a very important topic. Towards this goal, we focus on collaboration network between actors based by analyzing a publicly accessible movie database and find global patterns in the underlying collaboration dynamics. We study the emergent patterns that exists in this collaboration network and developed models to explain these phenomenon. In particular, we present a microscopic analysis of the edge-by-edge evolution as well as node evolution for this large scale collaboration network. From empirical data, we show how such emergent patterns can be generated from stochastic decisions made at the level of the actors. These findings are vital to a range of important applications, from the development of better collaboration recommendation algorithms, to designing better systems for social forums that address different aspects for the online society.

## CHAPTER 3

### Information Dynamics on Social Media

One of the distinguishing features of social networks and social media is their potential for information propagation. In this chapter, we study information propagation dealing with specific events, such as the Iranian election. Specifically we ask questions, how do large numbers of users collaborate to spread some messages widely? Does information dissemination on Twitter follow patterns similar to other known cases? Are different avenues through which users can access information, unique characteristics that create interesting and different dynamics of information dissemination on this network.

Toward this, we determine the set of all active users for a topic, and analyze their status messages to build the information networks. The event-specific cascade size distribution is a power-law with exponent equal to  $-2.51$  and more than 98.7% of the cascades have depth less than three, and hence shallow. We found that at most 63.7% of all retweets in Iranian election were reposts of someone the user was following directly, thus the friendship network plays a major role. Surprisingly, more than 34% of retweets for Iranian election are from the public timeline or the broadcast channel. We also study the underlying event-specific Friends-Followers network comprising only the active users, and investigate its role in determining the cascade size distribution. We also compare the dynamics of information propagation through the study of tweets about other specific events, such as death of Michael Jackson and the Swine Flu outbreak and found cascade size distribution is very similar between different topics. Our results show how real-time popular news propagates over network and it can help us with link prediction as well as viral marketing.

### 3.1 Introduction

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 across Iran. With international news reporters purged from the country shortly after the election, Iranian citizen journalism 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.

Twitter is a microblogging service that allows each user to post tweets of a maximum 140 characters on their profile page. Each user can then follow a collection of other users of her or his choice in order to view their tweets aggregated in a home page. We will call those who follow a user, his or her followers, and we will call those whom the user follows, his or her friends. Since following someone's tweets does not automatically mean that they will follow you back, Twitter's friends followers network is a directed graph.

Some conventions, without being required by Twitter, have been widely adopted by users. Using the “#” sign to tag a post according to its content (called a hashtag) is one such convention used in many tweets. The hashtag can be used as a search keyword to access a public listing (called the public timeline) of all the tweets that use that specific hashtag. When a keyword becomes very popular at any point in time, it appears as a trending topic on all users' home pages and on the twitter front page, giving all users direct access to all the tweets on that topic. Another convention that became a widely used standard (and recently implemented in the service as a proper function) was retweeting, where user2 would repeat user 1's tweet almost exactly, and adding “RT @user1” at the beginning of the tweet to give credit. For a more detailed guide to Twitter, please see [OM09].

Twitter has undoubtedly caught the attention of both the general public, and academia

[JSF07, KGA08, HRW09, KLP10, CHB10] as a microblogging [PHB08] service worthy of study and attention. Twitter has several features that sets it apart from other social media/networking sites, including its 140 character limit on each user's message (tweet), and the unique combination of avenues via which information is shared: directed social network of friends and followers and public timeline. A directed social network of friends and followers, where a user's message is sent to all its followers, provides a viral or point-to-point channel for information dissemination via *retweet* (RT) [BGL10]. The *public timeline* [HP09], which provides real time access to tweets on specific topics for everyone, provides a broadcast channel for information transfer, which can then feed into the point-to-point channel. While the character limit plays a role in shaping the type of messages that are posted and shared, the dual mode of sharing information (public timeline vs posts to one's followers) provides multiple pathways in which a posting can propagate through the user landscape via retweet, leading us to ask the following questions: How does a message resonate and spread widely among the users on Twitter? That is, how do the information cascades form and propagate? Is there a pattern to the types of messages that find large resonance and spread through the network? Are the resulting cascade dynamics different from those reported in the literature- [WHA04, GGL04, AA05, LMF07, LK08, BKA09], due to the unique features of Twitter? Given the existence of both broadcast and point-to-point channels of communication in Twitter, what relative roles do they play in the cascades?

Since different content can create different dynamics of information propagation, we focus our study on this very specific, yet large set of data. We know that during the 2009 post-election protests in Iran, Twitter and its large community of users played an important role in disseminating news, images, and videos worldwide and in documenting the events [Gro09]. We analyzed over three million tweets related to the Iranian election posted by around 500K users during June and July of 2009. Our results provide several key insights into the dynamics of information propagation that are special to Twitter. For example, the tweet cascade size distribution is a power-law with expo-

ment of  $-2.51$  and more than  $98.7\%$  of the cascades have depth less than 3. The exponent is different from what one expects from a branching process (usually used to model information cascades) and so is the shallow depth, implying that the dynamics underlying the cascades are potentially different on Twitter. Next we rank users by the number of tweets, the number of retweets, PageRank in F-F network as well as information network, and present comparison among them. The ranking by PageRank in information network gives us better understanding about the most influential users in this topic, who are Iranian tweeters and news media. Furthermore, we found that at most  $63.7\%$  of all retweets in this case were reposts of someone the user was following directly, therefore are able to show that Twitter's F-F network structure plays an important role in information propagation through retweets. Similarly, we also found that at least  $34\%$  of retweets are from the public timeline, thus the public timeline on Twitter's front page offer other significant avenues for the spread of information outside the explicit F-F network. By introducing event-specific F-F network comprising only the active users, we develop and evaluate the damped percolation model to explain the cascade size distribution. We also present a brief taxonomy of cascade content that gained the attention of users and leading to large cascades. Finally, we showed similar cascading behavior on death of Michael Jackson and Swine Flu outbreak, therefore our conclusion about news spreading on Twitter are universal for the topics of different genres.

## **3.2 Related Work**

Online social network systems have emerged recently as the most popular forums for user participation, social intercourse, and content generation. Research work conducted on modeling and analyzing various aspects of social networks have identified many recurring patterns, such as power law degree distributions, small world, local clustering and communities structures [BA99, GN02, KNT06, MMG07, NP03], in the underlying friendship or contact networks. Moreover, microscopic network evolution models have

been proposed [BHK06, LBK08].

Twitter has attracted much attention from researchers since it became an important social network as well as social media. Java et al. [JSF07] study the topological and geographical properties of Twitter's social network, and show how users with similar intentions connect with each other. Huberman et al. [HRW09] point out that the use of @user is a form of conversation, which indicates the hidden network of connections underlying the "declared" set of friends and followers. Boyd et al. [BGL10] present various conventions and styles of retweeting prevalent today and examine the emergence of retweeting as a conversational practice. Kwak et al. [KLP10] crawl the entire Twittersphere to study its topological characteristics and retweet trees between different users. However, we find that the cascades and information mechanisms for tweets are highly topic and content dependent, and hence, we chose to study a particular event that comprises a medium size network, and provides a window into various subtler aspects of information propagation on Twitter. For example it allows us to study the role played by the public timeline vs the F-F network in propagating information.

One of the distinguishing features of online social networks and social media is their potential for information propagation. It 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 recommendation network and study efficacy of viral product recommendation strategies [LAH07]. Leskovec et al. examine information propagation structure [LMF07] on blogosphere and propose algorithms for identifying influential nodes [LKG07]. Bakshy et al. [BKA09] trace the spread of influence in a multi-player online games and found patterns similar to our findings with social news dynamics on Twitter. However, in these previous studies, the underlying network is defined by message passing among the users and agents, i.e., an



edge connects two nodes,  $A$  and  $B$  if the message or link posted by  $A$  is copied or published by node  $B$ , and thus the cascades studied are analogous to the Tweet Networks studied in this paper. On Twitter, however, we have visibility of the Friends-Followers (F-F) network as well, and it provides us with a unique opportunity to study the role played by the F-F network vs the role played by the public timeline (or the posted messages) and the links to trending topics accessible to all viewers. For example, it provides us with an opportunity of finding what types of content led to cascades of significant size, and that the tweet infection rate is a function of the content type and hence a notion of fitness has to be introduced.

### **3.3 Measurement Methodology**

We used the Twitter API <sup>1</sup> to crawl the social network and download a large number of public user pages on Twitter. Since our goal here is to study the topological characteristics of information propagation regarding the Iranian election, our data sampling process is highly biased toward users who have tweeted about this topic.

As a summary, we collected the tweets as well as the F-F network of about 20 million public users on Twitter. Using most widely used keywords related to the Iranian election [BB10] to filter the tweets first, we focused on a total of more than 3 million tweets posted by 500K users between June 1 2009 and August 1 2009.

#### **3.3.1 Data Collection**

We began with a list of 100 most active users on the topic of Iranian election as reported by the Web Ecology Project [BB10]. Using these users as seeds, we traversed their directed F-F network (friends and followers) and reached about 126K valid users who were one step away from the seed users, which we will call depth-1 users. We continued to traverse the F-F network of these depth-1 users, and this gave us 23 million distinct

---

<sup>1</sup><http://apiwiki.Twitter.com/>

depth-2 users. We then crawled the F-F network of these 23 million users and finally collected about 20 million users' F-F network (Some of the target users were invalid or had protected profile, so we were not able to download their F-F network).

Since Twitter API only allows access to a maximum of 3,200 tweets per user, we collected as many tweets for these users as the API could provide. In total we collected the tweets as well as the F-F network for about 20 million users.

### 3.3.2 Coverage Estimation

We did not cover the entire connected component of Twitter but we had a qualitative coverage examination of our crawl. Since the IDs of user on Twitter are assigned sequentially, we uniformly selected 200K random IDs between the first ID and the last one. Among the IDs we tried to collect, there are 130K (65.0%) users with public profiles, 13K (6.5%) users with protected profiles and the remaining 57K (28.5%) IDs were invalid for different reasons. Based on these statistics, there should be around 55M valid users on Twitter by the end of September 2009 as maximum ID was 77M by then. Among the 130K users we downloaded, there were 1738 users who tweeted about the Iranian election 1558 of which were included in our crawled dataset; there were 11108 tweets related to the Iranian election and our dataset covers 10760 of them. Therefore, it appears that our dataset covers 89.6% of users and 96.9% of tweets relevant to the subject of Iranian election on Twitter.

### 3.3.3 Data Cleaning

Before the analysis, we applied the following procedures to clean the data in order to better represent the structures of information propagation.

**Only consider the tweets that have related keywords.** We used most widely used keywords related to the Iranian election [BB10] to filter the tweets first. As a result, we focused on a total of more than 3 million tweets posted by 500K users between June 1

2009 and August 1 2009.

**Only consider the RT tag.** In this paper we study information propagation as retweeting and only restrict the tweets that have form of 'RT @user'. On Twitter users may get similar news or messages from different sources, and it is possible for them to come up with the similar tweets without reference each other, which is not regarded as information propagation in our case.

**Remove self retweet.** Users sometimes retweet themselves in order to emphasize their message or increase the number of people who view their tweet, but self-retweets do not represent any information propagation.

### 3.3.4 Link Inference for Tweet Networks

Although a retweet explicitly mentions the user who posted the original tweet (RT @user), there is no mention of or link to the specific tweet that is being retweeted. In order to build tweet network we need to find links between a tweet and its retweet. So when a retweet mentions a certain user (RT @user) we must search that user's messages and find the tweet that has similar textual content with the retweet. On blogosphere, text analysis technique is proposed to infer relationship among posts [AA05]. In this paper, we adopt the digests technique [DDP04] to determine if two messages contain the same textual content. The activity by each user and observe that the distributions of user's activity follows power law with exponent about -1.92 (The K-S metric D is equal to 0.0078) Figure 3.1(a). The cutoff in power law degree distribution in Figure 3.1(a) is due to the limit of downloading 3200 status messages per user in Twitter API.

## 3.4 Friends Followers Network

The major difference between social network and social media like Twitter is the direction of relationship. We want to answer questions how does the directed relationship

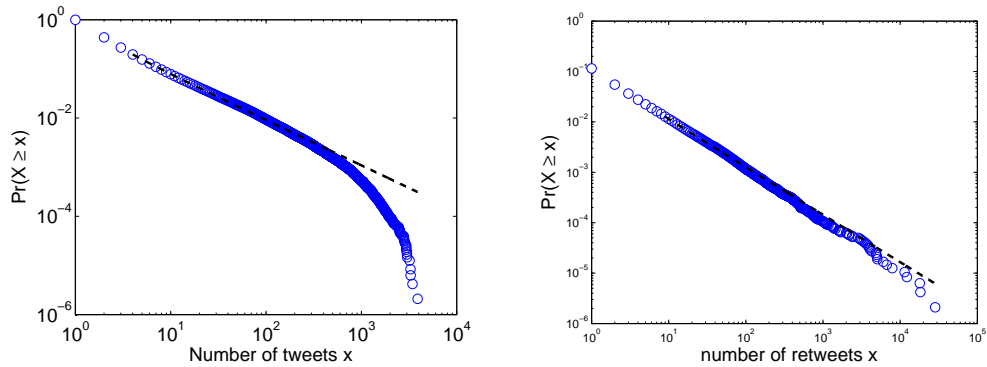


Figure 3.1: Cumulative distribution of number of tweets and retweets per user. Power law fit to the data with exponents  $-1.92$  and  $-1.94$ .

between users impacts the topological characteristics on Twitter space? In this selection, we focus our studies on studying the degree distribution, degree of separation and community structure on Twitter and discuss the unique characteristic of Twitter structure that made it a valuable place for information diffusion.

### 3.4.1 Network Structure

Networks with power law degree distribution have been the focus of attention in the literature and they are sometimes referred to as a class of "scale-free network" [BA99]. Studies [BJN02] show that the degree distribution of social collaboration networks follows a power-law as well. Does the direct relationship change the power-law distribution between users? We construct a directed network based on the following graph and analyze its in-degree and out-degree distributions. In Figure 3.2, the x-axis shows the in-degree and out-degree respectively and the y-axis represents the complementary cumulative distribution function (CCDF). The straight line in log-log axes shows both in-degree and out-degree distributions of this friends/followers network, following a power-law distribution.

To test how well the degree distributions are modeled by a power-law, we calculated the best power-law fit using maximum likelihood [CSN09]. Table 3.1 shows the estimated power-law coefficients, the corresponding Kolmogorov-Smirnov goodness-

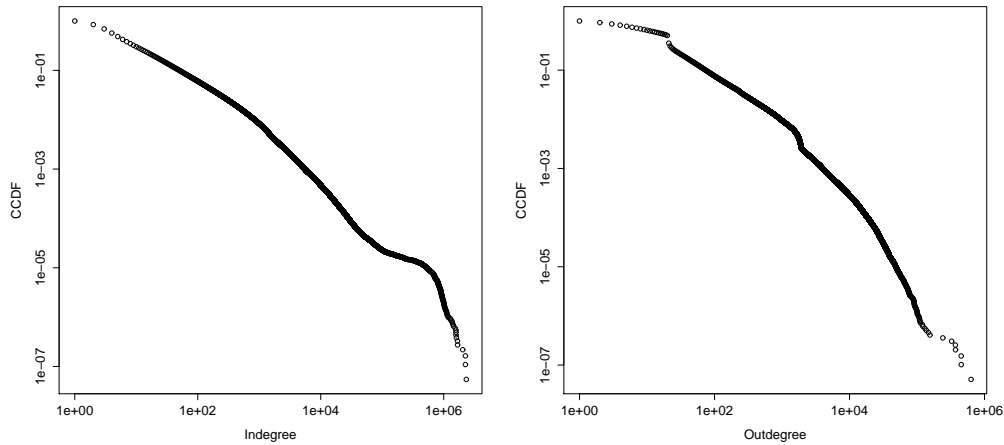


Figure 3.2: Cumulative in-degree and out-degree distribution of Twitter’s friends followers network

Property	Statistics
Number of Nodes	20045911
Average In-degree/Out-degree	58.42
In-degree Distribution $\alpha$	-2.25
Out-degree Distribution $\alpha$	-1.90
In-degree Distribution $D$	0.0084
Out-degree Distribution $D$	0.0098
Correlation of in-degree and out-degree	0.2696

Table 3.1: Twitter friends followers network statistics

of-fit metrics  $D$  (K-S metrics  $D$ ) as well as other properties of network we studied. It is interesting to note that the slope  $\gamma_{in}$  and  $\gamma_{out}$  are approximately -2.25 and -1.90. This value for the power law exponent is similar to that found for the Web.

The Pearson’s correlation coefficient between in-degree and out-degree is 0.2696, which motivates us to study the effect of reciprocity on Twitter friends followers network. We analyzed the average number of friends each user have given the number of their followers. The result in Figure 3.3 shows generally in-degree and out-degree correlates well, meaning users who get a lot of followers also have a lot of friends. However, we can see users who have more than 10,000 followers do not have many followers. In fact, these accounts are usually ofcial pages of new media, politicians and

celebrities, who get a lot of attentions. At the same time, they are not necessarily to follow back others. We also studies the reciprocal properties on Twitter friends followers network, about 79% of users with any link between them are connected one-way, and only 21% have reciprocal connection between them, which is similarly as reported in [KLP10].

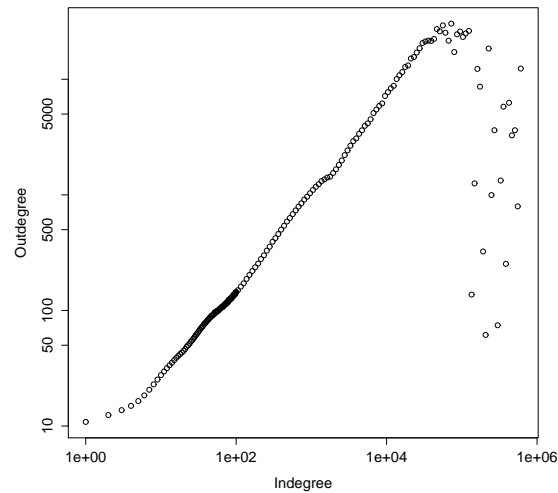


Figure 3.3: In-degree and out-degree comparison on Twitter friends followers network

### 3.4.2 Degree of Separation

The concept of degrees of separation has become a key to understanding the social network, ever since Stanley Milgrams famous six degrees of separation experiment [Mil67]. In his work he reports that any two people could be connected on average within six hops from each other. Watts and Strogatz have found that many social and technological networks have small path lengths too and they call them a small-world. Recently Leskovec et al. [LH08] study on the MSN messenger network of 180 million users and find the median and the 90% degrees of separation are 6 and 7.8 respectively.

Our goal here is to study degree of separation on Twitter social graph. As we point out, the directed nature of Twitter connection made us wonder if small world

exists on friends followers network too. In MSN a link represents a mutual agreement of a relationship, while on Twitter a user is not obligated to reciprocate followers by following them. Thus a path from a user to another may follow different hops or not exist in the reverse direction. To study the degree of separation, we consider a subset of complete Twitter F-F network that is active during certain period. Despite of Twitter's explosive growth in recent years, more than 60 percent of new users fail to return to Twitter the following month after they sign up. Studies also show the retention rate, which is the number of Twitter's returning users, still does not outnumber those giving up on Twitter after trying out the site.

We sampled our dataset and find 470,040 active users who posted at least one tweet about the certain topic between June 1 2009 and August 1 2009. Among these active users, there are 40,938,802 edges between them. and the clustering coefficient is 0.1052 which is an indication of how densely neighbors are connected. Assortativity, a measure of the likelihood for nodes to connect to others with similar degrees, has been shown to be positive in social networks [NP03]. However, the friends follower network of our dataset has a negative assortativity, which means nodes are likely to connect to nodes with different degree than their own.

As we point out in the previous section, about 21% of connections are reciprocal, so we expect the average path length between two users in Twitter to be longer than other social networks. To completely find statistics for the degree of separation, we need to find shortest distance between all possible pairs of users, which is huge amount of work. To estimate the path-length distribution, we randomly sample pairs of users and find their shortest distance. Figure 3.4 exhibits the distributions of the shortest paths on Twitter with sampling rate of 5%. Surprisingly for 91.3% of user pairs, the path length is 4 or shorter, and for 99.8% of user pairs the shortest distance is 5 or shorter.

The average path length of 3.82 is quite short considering the directed nature of Twitter. This interesting phenomenon speaks for Twitters role as information diffusion. Users follow others not only for social networking, but for latest information. The low

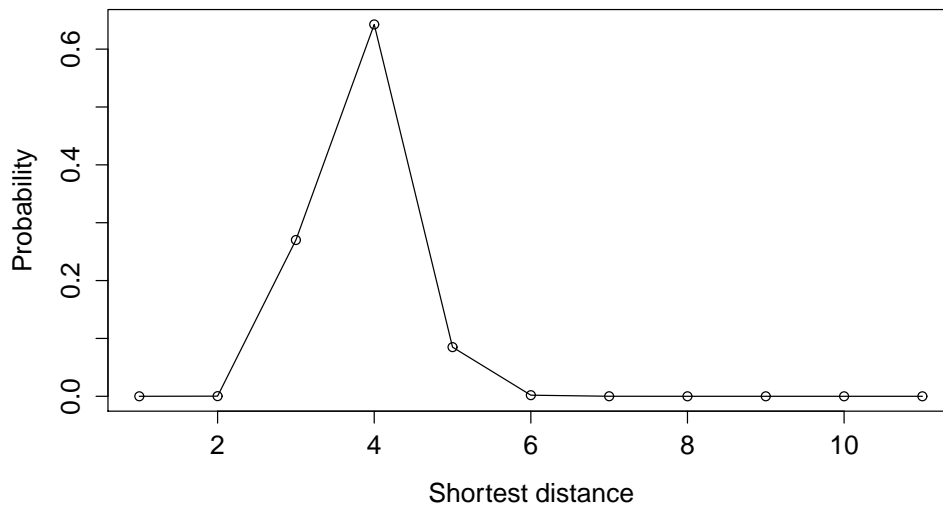


Figure 3.4: On Twitter, on average a user could reach 91.3% of others within 4 steps or shorter. For 99.8% user pairs, the shortest distance is 5 or shorter.

shortest distance easily makes information flow on Twitter graph very fast.

### 3.4.3 Community Structure

Many networks in nature, society and technology are characterized by a mesoscopic level of organization, with groups of nodes forming tightly connected units, called communities or modules, that are only weakly linked to each other. Uncovering this community structure is one of the most important problems in the field of complex networks. For instance, in biochemical or neural networks, communities may be functional groups, and separating the network into such groups could simplify functional analysis considerably. In the case of social networks, networks of friendships or other acquaintances between individuals, communities might represent real social groupings, perhaps by interest or background. Communities in a citation network might represent related papers on a single topic. Communities on the web might represent pages on related topics. Being able to identify these communities could help us to understand and exploit these networks more effectively.



Generally community structure exists if the nodes of the network can be easily grouped into sets of nodes such that each set of nodes is densely connected internally. Basically, community detection is nodes clustering algorithm that divides network naturally into groups of nodes with dense connections internally and sparser connections between groups. The ability to detect such groups is significantly important for social network as groups in social networks might correspond to social units.

Newman [GN02] tried to construct a measure that tells us which edges are most central to communities. They focus the edges that are most between communities using vertex betweenness. Vertex betweenness has been studied in the past as a measure of the centrality and influence of nodes in networks. The betweenness centrality of a vertex  $i$  is defined as the number of shortest paths between pairs of other vertices that run through  $i$ . It is a measure of the influence of a node over the flow of information between other nodes, especially in cases where information flow over a network primarily follows the shortest available path.

To find which edges in a network are most between other pairs of vertices, Newman et al. in [GN02] generalize betweenness centrality to edges and define the edge betweenness of an edge as the number of shortest paths between pairs of vertices that run along it. If there is more than one shortest path between a pair of vertices, each path is given equal weight such that the total weight of all of the paths is unity. If a network contains communities or groups that are only loosely connected by a few intergroup edges, then all shortest paths between different communities must go along one of these few edges. Thus, the edges connecting communities will have high edge betweenness. By removing these edges, authors separate groups from one another and so reveal the underlying community structure of the graph.

Newman [New06] proposed using modularity as measurement to quantify the goodness of community structure. Modularity is one measure of the structure of networks or graphs. The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed

at random. Networks with high modularity have dense connections between the nodes within communities but sparse connections between nodes in different communities.

In [New06], author uses modularity as optimization methods for detecting community structure in networks even through exact modularity optimization is known to be NP-hard. It uses greedy optimization method that attempts to optimize the modularity of a partition of the network. The optimization is performed in two steps. First, the method looks for small communities by optimizing modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced. Although the exact computational complexity of the method is not known, the running time of method is in time  $O(n \log n)$  with most of the computational effort spent on the optimization at the first level.

We perform this algorithm on the social graph of active users and plot modularity under the number of merges in Figure 3.5. We can see the maximum modularity the algorithm achieved is 0.3269, which is a good indicator of significant community structure in a network. Under the maximum modularity, there are 9 major communities and each has the size of 181225, 20133, 215908, 29873, 4827, 7115, 2363, 271 and 403 respectively.

### **3.4.3.1 Hierarchical Community Structure**

It also has been shown that there is a hierarchical structure of complex networks with communities embedded within other communities. Essentially, small communities group together to form larger ones, which in turn group together to form even larger ones [LFK09]. We investigate this problem in our social graph and break up the individual communities into smaller ones by maximizing modularity. Table 3.2 shows the maximum modularity found for the top 9 communities, which shows there exists small

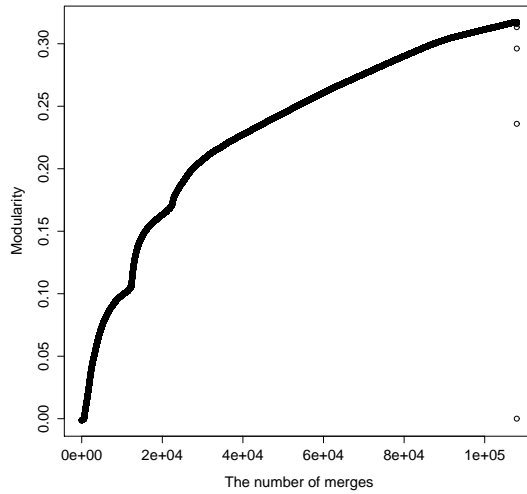


Figure 3.5: Modularity as the number of steps to merge communities

communities within in most of communities.

Size of community	maximum modularity
181225	0.343130
20133	0.1251577
215908	0.3727419
29873	0.5871593
4827	0.4131485
7115	0.4015921
2363	0.5763832
271	0.1114815
403	0.5327359

Table 3.2: Hierarchical community structure

Not only observing the small communities inside big ones, we also study the hierarchical community structure. A natural way to represent the hierarchical structure of a graph is to draw a dendrogram. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. At the bottom, each vertex is its own community. By moving upwards, groups of vertices are successively aggregated. Mergers of communities are represented by horizontal lines and the uppermost level represents the whole graph as a single community. Cutting the

diagram horizontally at some height, displays one possible partition of the graph.

To better visualize the dendrogram, we take one community with high modularity and cluster nodes hierarchically use dendrogram. As Figure 3.6 suggests, there are three communities densely connected inside community.

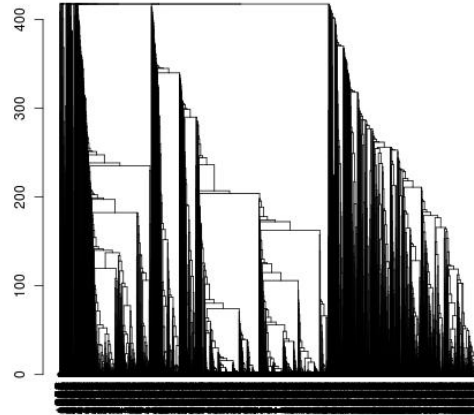


Figure 3.6: Dendrogram shows hierarchical community structure

### 3.4.3.2 Community Formulation

The nature for users to form groups lies in users certain geolocation, affiliation, language and so on. To confirm this hypothesis, we extract a group of users from Eastern Asian and visualize the social connection in Figure 3.7. We can see clearly from the network structures, there are three small communities inside. Three small communities are extracted completely by community detection algorithm and we show three individually in Figure 3.8.

To further study the subgroups among these users, we retrieve all of users' tweets and study the reason for community formulation. Further tweets studies show three groups of users are Chinese, Japanese and Korean speaking respectively. We confirm our hypothesis that users in the same community are connected together because most

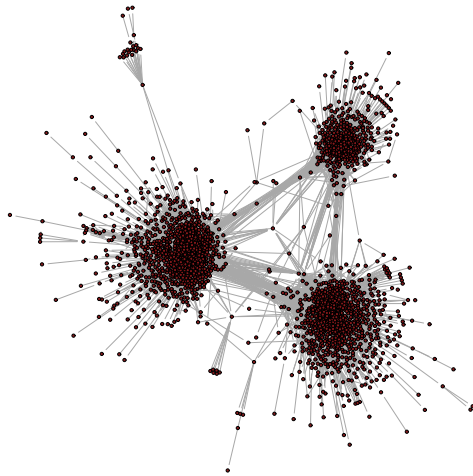


Figure 3.7: Plot of social connections between a group of users from Eastern Asian of the members might come from the same country and share similar interest.

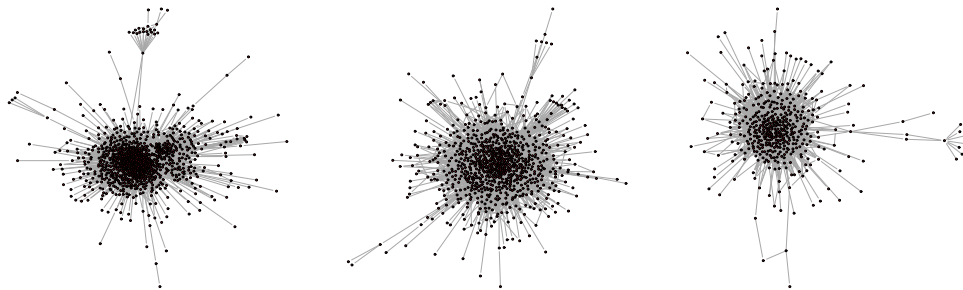


Figure 3.8: Plot of three communities extracted from social connections between a group of users from Eastern Asian. Further tweets studies show three groups of users are Chinese, Japanese and Korean speaking respectively

## 3.5 Information Propagation

### 3.5.1 Tweet Rate as Information Resonance

Social media feeds can be effective indicators of real-world performance, for example, the rate at which movie tweets are generated can be used to predict movie box-office revenue [AH10]. In Figure 3.9 we plot the number of tweets per day on the Iranian

election from June 1 to August 1, 2009. We observe that the rate at which users post relevant tweets gradually increased as the events unfolded in Iran and the use of Twitter provoked attention, spiking dramatically in relation to political events inside Iran as well as in relation to new events and incidents particular to the web. For example, on June 20 mass protests took place in Tehran and security forces responded with violence; a young Iranian woman named Neda Agha-Soltan was shot and killed by the Basij - government militia- in Tehran. Videos of the killing taken with mobile camera were posted on youtube and rapidly spread across the Internet. On that day, tweet rate around the topic of Iranian election reached its peak of about 300K tweets per day.

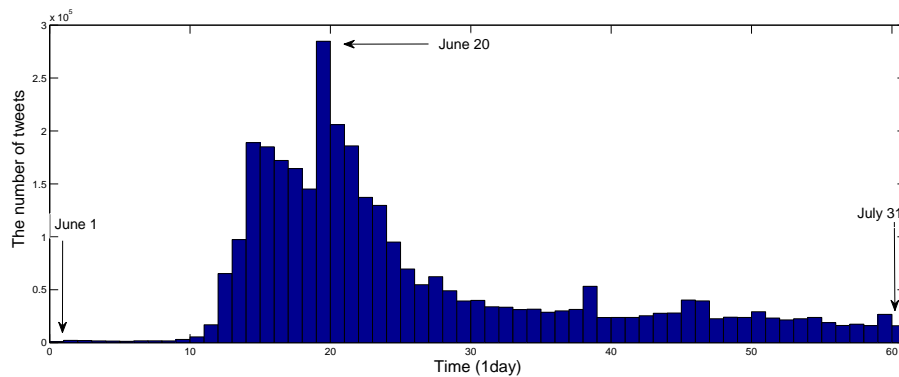


Figure 3.9: Number of tweets by day from June 1 2009 to Aug 1 2009. The rate gradually increased as the events unfolded in Iran and the use of Twitter provoked attention, spiking dramatically in relation to political events inside Iran as well as in relation to new events and incidents particular to the web.

In addition, we observe that the number of tweets per user follows a power law with exponent of -1.92 and the K-S metric D is equal to 0.0078 [CSN09]. The number of retweets received by each user is also heavy-tailed and we can fit a power-law distribution with exponent of -1.94 with a K-S metric D equal to 0.0110.

One might expect that users who posted a lot of tweets regarding Iranian election would have a lot of followers who also post on this topic. Intuitively, we expect the attention (number of followers) a user gets to be correlated with the user's activity (number of tweets). However this does not seem to be the case in our Iran related users.

The Pearson correlation coefficient between the number of tweets and the number of followers is only 0.040. Furthermore, a user's activity is not correlated with how many friends he or she has, as the correlation coefficient is only 0.041. We analyze users who are authoritative or prominent within the community in this case. The distribution of retweets is heavy-tailed and we can fit a power-law distribution with exponent of -1.94 with a K-S metric D equal to 0.0110. Correlation between the number of retweets and number of followers is 0.1824 while the correlation between number of retweets and the number of tweets is 0.2327. Therefore if a user has more tweets and more followers, she or he will get more retweets.

### **3.5.2 Tweet Network as Information Propagation**

A tweet network is a collection of cascades where every node represents a tweet and there is a directed edge from tweet  $u$  to  $v$  if tweet  $v$  retweeted tweet  $u$ . There are a total of 3,219,038 nodes (tweets) in our tweet network and 2,600,295 nodes are isolated, meaning that they did not get retweeted by others. These nodes represent the most common cascade in our dataset and we call it trivial cascades. After ignoring 2374 self edges, we got 433,088 edges in our tweet network. We found out-degree distribution follows a power law with exponent of -2.33 and the K-S metric D is equal to 0.0045 3.10. Note the in-degree of node in our tweet network is 1 or 0, meaning that tweet is a retweet or not.

We can decompose the tweet network into weakly connected components and every component represents cascades of different content. We want to see what are the common cascade shapes and how do the real cascades look like. We consider the number of nodes, the number of edges, the sorted in-and out-degree sequence as well as the singular value of the adjacency matrix obtained from singular value decomposition as a good signature, since the isomorphic graphs would have the same signature [LSK06]. Then we hash on these signatures to obtain the frequency and examples of the common cascade shapes. The top ten common nontrivial cascade shape is presented in Figure 3.11

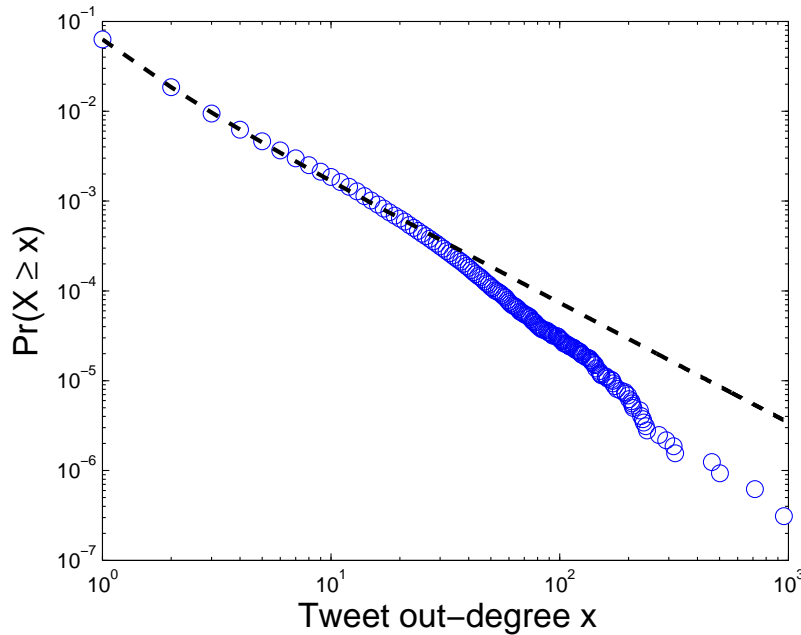


Figure 3.10: Cumulative distribution of out-degree in tweet network. Power law fit to the data with exponents -2.33

ordered by frequency, and the script of the label gives frequency rank. For example,  $G_9$  is 9th most frequency cascade with 1424 occurrences. We find that there are 173,282 non-trivial cascades with total of 1817 different shapes. The distribution of cascade shape frequency also follows the power law distribution as exponent equal to -1.6 and the K-S metric  $D$  is equal to 0.0281.

We notice that real cascades tend to propagate as certain shape and there are some interesting observations. *Cascades tend to be wide and shallow.* For example,  $G_3$  is more common than  $G_5$ , and  $G_4$  is more common than  $G_7$ . Statistically, 88.5% of the cascades have depth one and more than 98.7% of the cascades have depth less than 3. *Most of the cascade have a central hub.* For example, a central hub cascade (Figure 3.12(a)) is more likely to occur than a two hubs cascade (Figure 3.12(b)). As the central hub is usually the cascade source, users are always retweeting the influential users of this topic from the cascade source. *The largest cascade in the Iranian election (Figure 3.12(c)) is initiated by Stephen Fry about spreading proxies that help Iranians*













ID	Graph	# of Nodes	# of Edges	Frequency
G <sub>2</sub>		2	1	112895
G <sub>3</sub>		3	2	21814
G <sub>4</sub>		4	3	7269
G <sub>5</sub>		3	2	5591
G <sub>6</sub>		5	4	3482
G <sub>7</sub>		4	3	3194
G <sub>8</sub>		6	5	1977
G <sub>9</sub>		5	4	1424
G <sub>10</sub>		7	6	1315
G <sub>11</sub>		8	7	932

Figure 3.11: Top ten common nontrivial cascade shapes ordered by the frequency. For each graph we show the number of nodes, the number of edges and frequency.

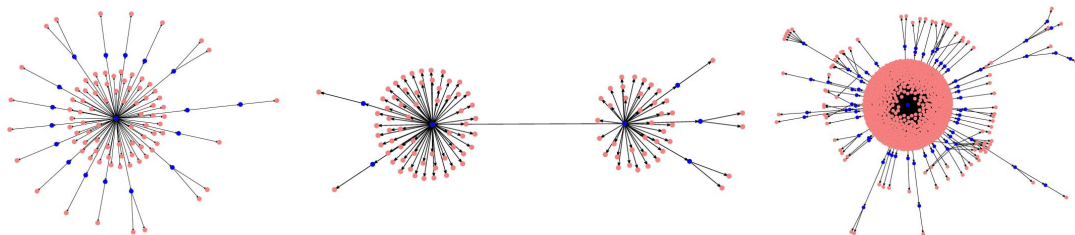


Figure 3.12: Real cascades observed (a) 'StopAhmadi' wrote: Please @Twitter and @ev don't take down Twitter, for the iranian ppl #iranelection (b) 'RealTalibKweli' wrote: Pray for the protesters in Iran. Regardless of your politics (c) 'Stephen-fry' wrote: Functioning Iran proxies 218.128.112.18:8080 218.206.94.132:808 218.253.65.99:808 219.50.16.70:8080 #iranelection - feel free to RT

*bypass Internet filters*. More than 1000 retweets are following this message.

Furthermore, we observe that overall cascade size (how many tweets are in each cascade) follows a power-law with exponent equal to -2.51 while audience size (how many users have been reached in each cascade) distribution is also a heavy tail distribution 3.13. In addition, more than 10% of the cascades have 10k recipients or more and the maximum audience size is 2M although more than 99% of the cascades have size less than 20. Because of this amazing power of retweeting, individual users have the power to spread important information by the form of retweet, which collectively determines the importance of the original tweet.

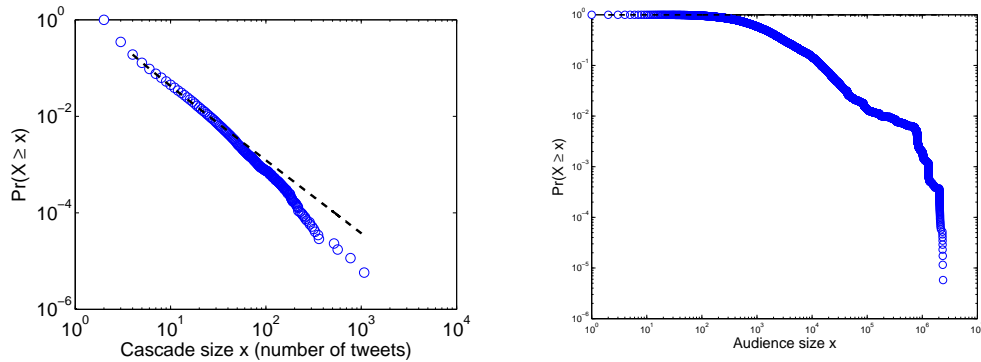


Figure 3.13: Cumulative distribution of cascade size and audience size. More than 10% of the cascades have 10k recipients or more although more than 99% of the cascades have size less than 20.

### 3.5.3 Influential Users in Information Propagation

Identified a given relevant topic on Twitter, we would like to study the influence in this space, an important concept in fields of sociology, communication and marketing. Despite the large number of theories of influence in sociology [Rog95, KLR55], there is no concrete definition of how to measure the influence, in the spread of news. To analyze Twitter as a social news media and study its role in information propagation, we focused on an individual’s potential to lead others to engage in this topic. We consider user’s activity, attention as well as PageRank [BP98] in the associated networks

as metrics to measure the influence and rank users based on the following criterion: *the number of tweets* (indicates the ability of that user to engage others in information propagation), *the number of retweets* (indicates the ability of that user to generate cascades with pass-along value), *PageRank in F-F network* (measures the user's relative importance within F-F network) and *PageRank in information network* (measures the users' relative importance within information network, where the edge between users means the information propagation).

The top 10 users in each category are listed in 3.3. As a summary, the most influential authors in this topic are *Iranian tweeters* and *news media*. Iranian tweeters, some of them tweeting inside Iran (@persiankiwi, @mousavi1388) and some others tweeting from other countries (@oxfordgirl, @iranriggedelect, @stopahmadi), are actively tweeting for this topic. They provided a lot of real-time and accurate information regarding this topic and are well-known among the active users in this space, making their influence through reputation they built in this domain. As a result, these domain experts would consistently appear in many medium-size cascades (between 30 and 150 retweets) even if they do not have a huge number of followers. Official news media (@breakingnews, @cnnbrk) and social news blog (@Tehranbureau, @Mashable) are usually tweeting much of breaking news. Although they do not tweet a lot about this topic, they have a substantial number of followers, making their influence through F-F network. As a result, most of the large cascades in this topic are generated by these news media and celebrities.

## **3.6 Medium of Information Propagation**

### **3.6.1 Information Propagation via F-F Network**

A directed F-F network provides a point-to-point channel for users to read tweets directly from their friends and retweet it. To study this kind of information propagation, for all the retweets, we check whether the retweeters are the author's followers and

Ranking	Rank by # of tweets	Rank by # of retweets	Rank by PageRank in F-F network	Rank by PageRank in Information Network
1	ahuramazda	oxfordgirl	barackobama	persiankiwi
2	oxfordgirl	stopahmadi	cnnbrk	stopahmadi
3	zozizz	persiankiwi	theonion	nytimeskristof
4	greentips1388	iranriggedelect	aplusk	mousavi1388
5	realgreen1388	iranbaan	johncmayer	tehranbureau
6	greenscreen1388	breakingnews	sarahksilverman	oxfordgirl
7	nedaagain	tehranbureau	stephenfry	iranriggedelect
8	a_iran_election	lotfan	rainnwilson	cnnbrk
9	razzmichi	cnnbrk	maddow	iran09
10	a_ie_pics	laraabcnews	mrtweet	mashable

Table 3.3: Influential users in information propagation

call these retweets *followers' retweets*, otherwise we call them *nonfollowers' retweets*. The percentage of followers' retweets over the span of our dataset suggests that before the election when there were little traffic, most of the retweets were coming from the friends' posts. As Iranian election received more attention, the percentage of followers' retweets dropped and finally approached to 63.7% in the end. Therefore F-F network plays an important role for spreading information and Twitter serves a role of social networking by forwarding users' tweets to their followers.

To study the information propagation via F-F network qualitatively, we analyze the retweet characteristic of tweets by estimating *retweet rate*  $T(x)$  of tweet  $x$  as follows:

$$T(x) = \frac{\text{number of followers' retweets of } x}{\text{number of followers that } x\text{'s author has}} \quad (3.1)$$

We measured the retweet rate for each cascade source and show their cumulative distribution in 3.15. As we expected, different tweet has different retweet rate, which suggests that different content may have different popularity among the audience.

Our observation that more than 98.7% of the cascades have depth less than 3 suggests the retweet rate may decay as the cascades spreads away from the source. We define the *retweet rate decay factor* at hop  $N$  as the ratio between retweet rate at hop

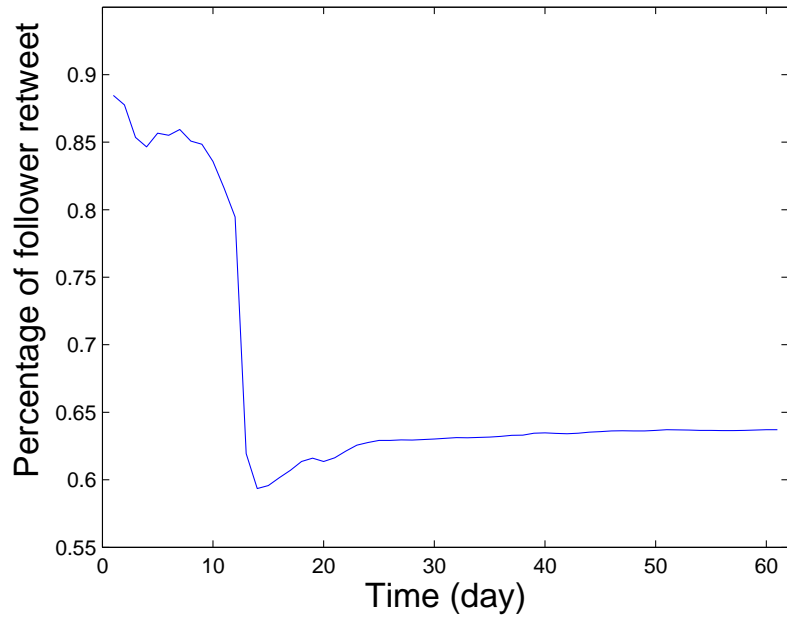


Figure 3.14: Percentage of followers' retweets. As the whole issue provoked attention, the percentage dropped and approached to 63.7% in the end.

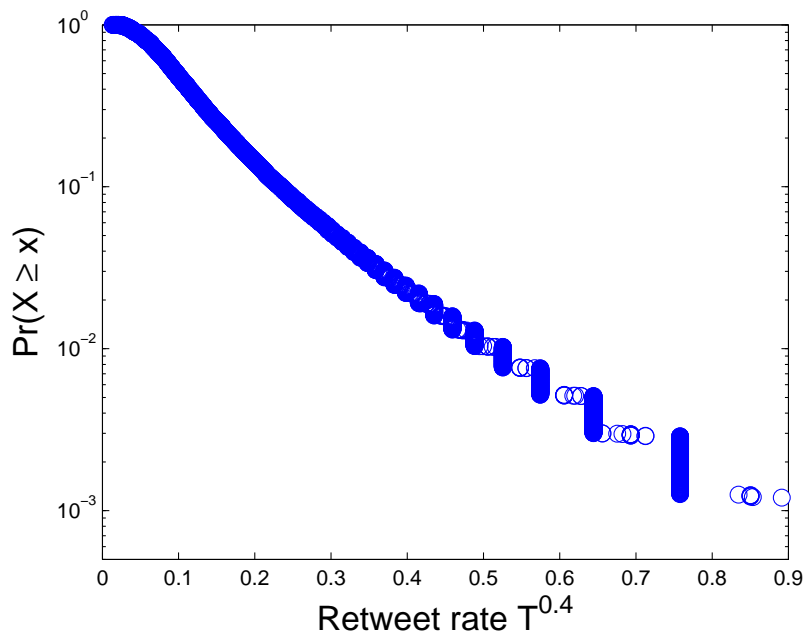


Figure 3.15: Cumulative distribution of retweet rate decays with a stretched-exponential law.

$N$  and retweet rate at hop  $N - 1$ . For example, a factor of 0.5 means the retweet rate at hop  $N$  is half of retweet rate at hop  $N - 1$ . We find that the mean of decay factors are all about 0.2 while the standard deviations are high, which suggests the variance of the tweet content affects the decay factor. *Therefore the retweet rate decays exponentially as the cascades spreads away from the source.* One possible explanation would be that the freshness of tweet tends to fade with time and the attention that people pay to it would drop as the time goes on [WH07].

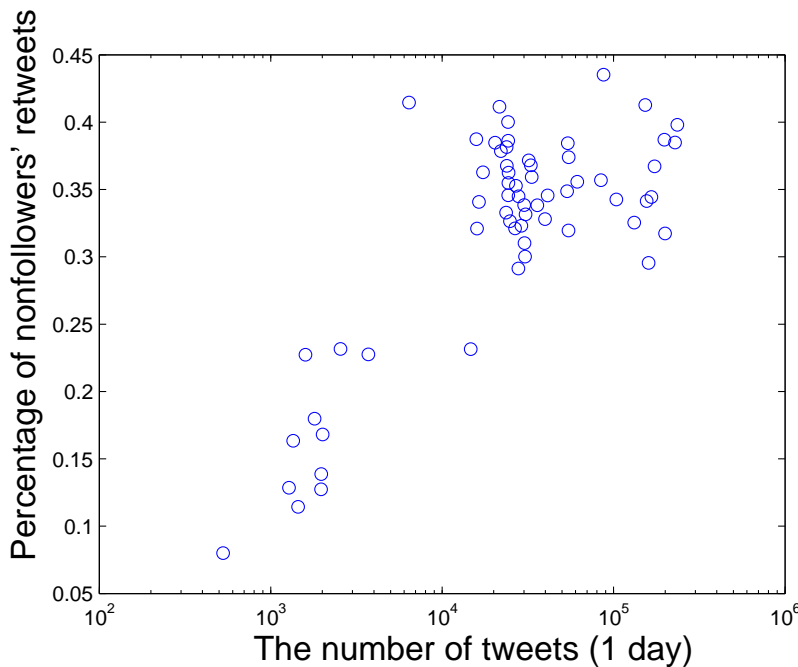


Figure 3.16: Number of tweets versus percentage of nonfollowers' retweets per day. Once the number of tweets posted exceeded 10k per day, the percentage of nonfollowers' retweets increased by 10%.

### 3.6.2 Information Propagation via Public Timeline

Twitter tracks keywords and hashtags that are most often mentioned and posted under the title of trending topic on public timeline, which provides a broadcast channel for information propagation. Iranian election is a popular trending topic most of the time during our dataset span, so potentially more users were retweeting it from the public

timeline. Figure 3.16 shows the relation between number of tweets and percentage of nonfollowers' retweets per day. Once the number of tweets posted exceeded 10k per day, the percentage of nonfollowers' retweets increased by 10%. This is consistent with our hypothesis that after topics are promoted to be trending topics, a lot of users read tweet from the public timeline and retweet it. We also analyze how nonfollowers' retweets contribute to the overall cascade. We show the number of nonfollowers' retweets for each cascade and the slope in the log-log plot is about 1.317. The linear relationships between the number of nonfollowers' retweets and the size of cascade suggests tweets are equally likely to get retweeted via public timeline.

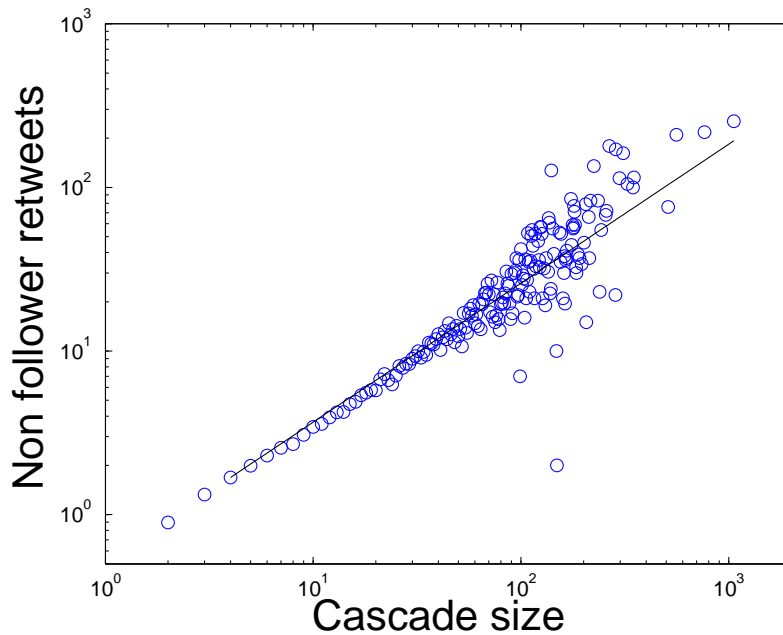


Figure 3.17: Larger cascades included more retweets of nonfollowers' retweets.

Excluding the possibility that some users would like to give credits to the original users when they see their friends retweeting, we find that at least 34% of retweets are from the public timeline. Therefore Twitter also serves a role of social media where users can get fresh and related tweets from the public timeline. To further study these nonfollowers' retweets, we calculated the shortest paths it took the retweeters to reach the author through F-F network. There are 57.47% of nonfollowers' retweets can be

reached by two hops so most of the time, these users are just retweeting their friends' friends, however there are still 6.79% of the nonfollowers' retweets cannot be reached through F-F network.

### 3.7 Content Taxonomy 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 tweets) in our data set and observed several noteworthy characteristics. The contents of tweets in medium and large cascades 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.

**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 [BA99] (with K-S metric  $D$  of 0.0047). The most popular URL found in our dataset is <http://helpiranelection.com/> (appearing about 200K times). The website 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 easily 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 hashtags came with our tweets, for example #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 <sup>2</sup>.

**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 Barack 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

---

<sup>2</sup><http://news.bbc.co.uk/2/hi/uk/8116869.stm>

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.

### 3.8 Models Based on Damped Percolation

The fact that the cascade size via followers' retweets exhibits a power law distribution with exponent equal to -3, which is different from what one expects from a branching process -1.5 [Wat02, SR07], indicates that the dynamics of information dissemination are different in this case. What is the underlying process that generates real cascades?

We first define the *event-specific F-F network* to be the Friends-Followers network between active users who posted at least one tweet regarding the specific topic. The event-specific F-F network is among the enthusiasts who are interested in posting anything on the topic and the likelihood for someone outside to participate in this topic is almost zero. Figure 3.18(a) and Figure 3.18(b) show the in-degree and out-degree distributions of this network, both following a power law distribution. We also define *influence basin* to a tree obtained by performing damped percolation of breadth-first search (BFS) on event-specific F-F network of a particular user, which determines the potential shape of cascades generated by that user.

Then we introduce an epidemic model with transmission probability of exponential decay on event-specific F-F network. Following the analysis for the susceptible, infected, removed (SIR) model [New02], we consider the problem of information transmission through a random network with given degree distribution  $p_k$  with transmissibility  $T$ . The mean degree is given by  $z$  and the higher moments are given by:

$$\langle k^n \rangle = \sum_k k^n p_k.$$

We define the generating function for the degree of a randomly chosen node in the

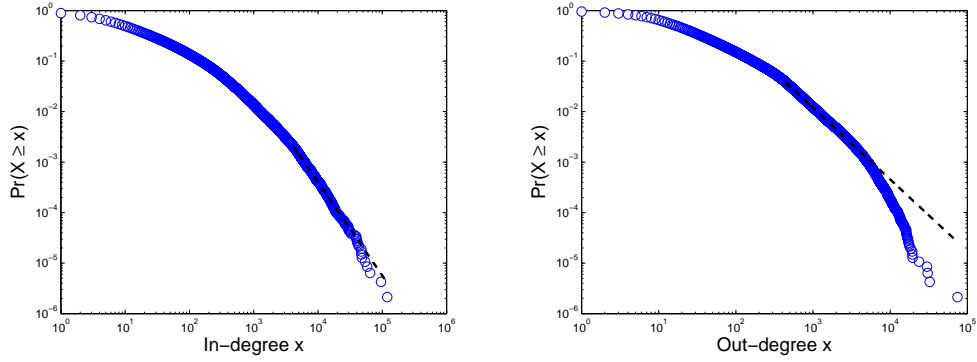


Figure 3.18: Cumulative distribution of in-degree and out-degree in the event-specific F-F network.

network:

$$G_0(x) = \sum_k p_k x^k \quad (3.2)$$

Following a randomly chosen edge to reach a node, the probability that the node has degree  $k$  is proportional to  $k p_k$ , since high degree nodes have more edges attached to them. Excluding the edge that we arrived to the node along, the *remaining degree* is given by  $(k - 1)$ . Thus, the generating function for the distribution of the remaining degree of a node reached by following a randomly selected edge is:

$$G_1(x) = \frac{1}{z} \sum_k k p_k x^{k-1} \quad (3.3)$$

We estimate the probability  $p_l^{(1)}$  that a randomly selected node has  $l$  *infected* edges attached to the node. Using the binomial distribution, we find:

$$p_l^{(1)} = \sum_{k=l}^{\infty} p_k \binom{k}{l} T^l (1 - T)^{k-l} \quad (3.4)$$

then the generating function  $G_0(x; T)$  for the distribution of the number of infected

edges attached to a randomly chosen node is given by:

$$\begin{aligned}
G_0(x; T) &= \sum_{l=0}^{\infty} \sum_{k=l}^{\infty} p_k \binom{k}{l} T^l (1-T)^{k-l} x^l \\
&= \sum_{k=0}^{\infty} p_k \sum_{l=0}^k \binom{k}{l} (xT)^l (1-T)^{k-l} \\
&= \sum_{k=0}^{\infty} p_k (1 + (x-1)T)^k \\
&= G_0(1 + (x-1)T)
\end{aligned} \tag{3.5}$$

Following similar steps of derivation, the generating function  $G_1(x; T)$  for the distribution of the number of infected edges attached to a node arrived by following a randomly selected edge is:

$$G_1(x; T) = G_1(1 + (x-1)T) \tag{3.6}$$

Suppose the transmissibility decays exponentially as the cascades spreads away from the source. Then the probability that an  $m$ th neighbor will transmit the information to a person with whom he has contact is given by

$$T^{(m)} = \alpha^m T_0 \tag{3.7}$$

where  $\alpha$  is the decay factor.  $T^{(m)} = T_0$  at the initiator ( $m = 0$ ) and decays to zero as  $m \rightarrow \infty$ . We define  $G^{(m)}(x)$  to be the generating function for the distribution of the number of  $m$ th neighbors affected by following a randomly chosen node, then for the generating function for transmission probability to the first neighbors is  $G^{(1)}(x) = G_0(x; T_0) = G_0(1 + (x-1)T_0)$ . The generating function for the transmission probability to 2nd neighbors can be written as

$$G^{(2)}(x) = \sum_k p_k^{(1)} [G_1^{(1)}(x)]^k = G^{(1)}(G_1^{(1)}(x)) \tag{3.8}$$

where  $G_1^{(m)}(x) = G_1(x; \alpha^m T_0) = G_1(1 + (x - 1)\alpha^m T_0)$ . Similar, we can have

$$G^{(m+1)}(x) = \sum_k p_{(k)}^{(m)} [G_1^{(m)}(x)]^k = G^{(m)}(G_1^{(m)}(x)) \quad (3.9)$$

It is found [PV01] that the absence of an epidemic threshold and its associated critical behavior in a dynamical model for the spreading of infections on scale-free networks. However, by assuming the transmissibility decays exponentially imply that the spread of information is limited. Assume  $z_{m+1}$  to be the average number of  $(m + 1)$ th neighbors, then

$$z_{m+1} = G^{(m+1)'}(1) = G_1^{(m)'}(1)G^{(m)'}(1) = G_1^{(m)'}(1)z_m \quad (3.10)$$

so the condition that the size of the outbreak(the number of affected individuals) remains finite is given by

$$\frac{z_{m+1}}{z_m} = G_1^{(m)'}(1) = \alpha^m T_0 G_1'(1) < 1 \quad (3.11)$$

For any given  $T_0$  and  $G_1'(1)$ , the left-hand side of the inequality above goes to zero when  $m \rightarrow \infty$ , so the condition is eventually satisfied for large  $m$ . Therefore, the average total size is always finite if the transmissibility decays with distance.

We validate our model by numerical simulation. We propagate cascades over event-specific F-F network with exponentially decay transmissibility of  $T_0 = 0.006$  and  $\alpha = 0.2$  from the empirical data. We show the results on cascade size distribution in Figure 3.19 and there is a good agreement between cascade size distribution of the real data and model. Therefore, by introducing the event-specific F-F network, we validate the model of damped percolation on event-specific F-F network to understand information propagation via F-F network.

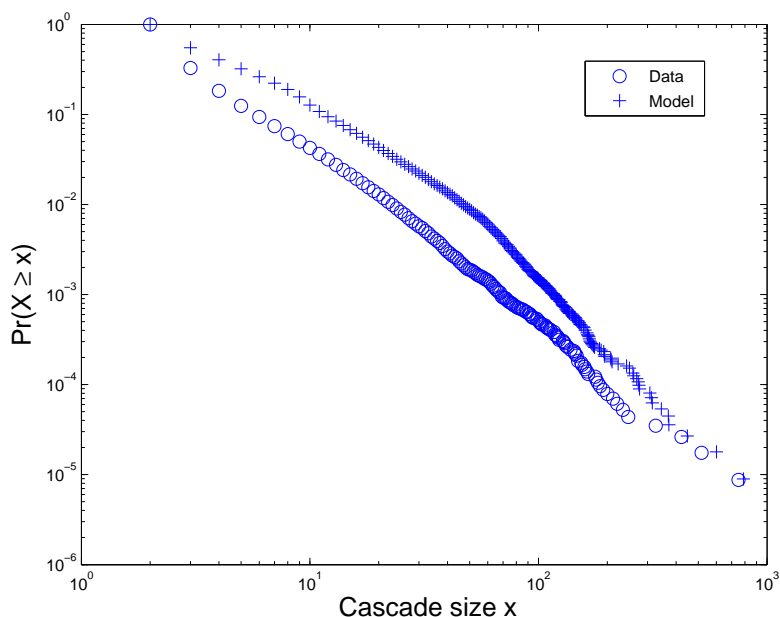


Figure 3.19: Comparison of the real data and the our model based on damped percolation. We plotted the distribution of the real cascades with circles and the simulation of our model with plus signs

### 3.9 Validation on Other Topics

In order to contrast the results of the Iranian election (political genre) stream against the general news events, we analyzed the cascading behavior for other two topics, that is death of Michael Jackson (social genre) and breakout of Swine Flu (health genre). It is very interesting that the cascade size distribution in these topics (Figure 3.20 (a) and (b)) are very similar to what we observed in Iranian election. Furthermore, we find that the event-specific F-F network for these two events (Figure 3.20 (c) and (d)) exhibit similar structure to the one in Iranian election as well. It not only expands our empirical results for other topics, but also validates our damped percolation model on event-specific F-F network for cascade size distribution.

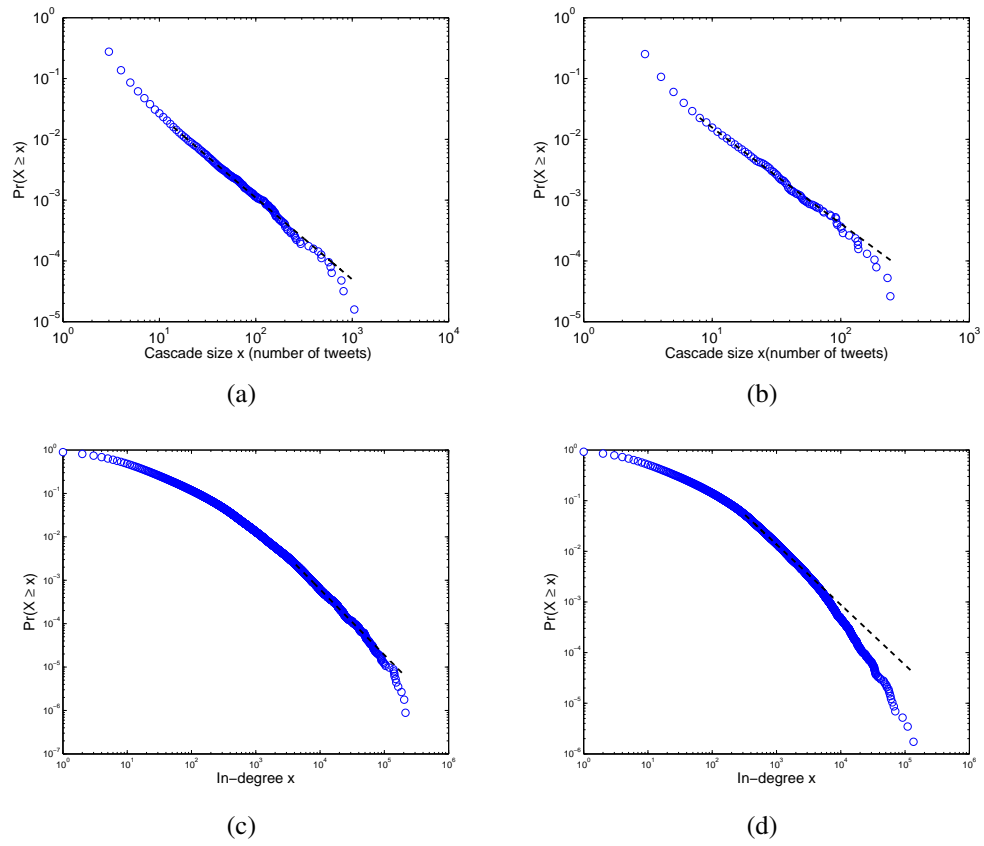


Figure 3.20: Cascade size and indegree distribution of event-specific F-F network in death of Michael Jackson and Swine Flu breakout

### 3.10 Conclusion and Discussion

In this Chapter, we use specific events as a window to study information cascades. First of all, we find the *structure of cascades* tends to be wide, and shallow, with a central hub being more common. The overall cascade size distribution follows a power-law distribution with exponent equal -2.51 and more than 98.7% of the cascades have depth less than three. Due to broadcasting of tweets, cascades reach a lot of audience on the network we studied, even although user participation rate is not high. Then we study the *medium of cascades* on Twitter. We found that at most 63.7% of all retweets in Iranian election (78% for the other two topics) are from F-F network, thus the friendship network plays a major role. More than 34% of retweets for Iranian election (around 20% for the other topics) are from the public timeline, therefore public timeline offers other avenues for the spread of information outside the explicit friendship network. Last not the least, we study the *mechanism of cascade* by the underlying event-specific F-F network with a power law structure, and investigate its role in determining the cascade size distribution. We formulate the damped percolation model on event-specific F-F network to study information propagation and validate it through extensive simulation.

Understanding the principles of information propagation via F-F network as well as public timeline will be help design better application systems that address different aspects of the social media. Our findings about structure of information propagation has a significant impact on determining and managing Internet traffic, and hence Internet infrastructure backbone. Also we can take advantage of real-time trending topics on public timeline for viral marketing.



## CHAPTER 4

### Model Content on Social Conversation

In this chapter, we present our approaches to model the various aspects of online conversations to study information dynamics. In the recent years, Internet has become one of the most important channels for information exchange and idea sharing. People spend as much time online as offline to communication with different groups of friends. Online conversation happens in various aspects of our forms, from the discussion forums, social media, to social customer relationship management. Stimulated by these changes, many researchers start to study the patterns of information dynamics from online conversation, and here we focus our research on a vaccination forum, called *mothering.com*, which has grown in prominence as an important resource for parents concerned with health care decisions related to their children (mostly young and pre-teen). In particular, we applied statistical natural language processing to model the online conversations on social forums, and demonstrate how they are effective in revealing information dynamics.

#### 4.1 Introduction

In this Information Age, discussion forums, social network messages, video comments, and social media services provide rich information for users to discuss ideas and exchange information. These huge volumes of information discussion and exchange circulating on multiple information channels that may affect the user's understanding, decisions as well as behaviors of individuals. The study of characteristics and dynam-

ics of online conversations has fueled an increasing amount of research in recent years, however research to model and analyze this dynamic, complex system of social interaction in a scientific manner is still developing. We apply novel methods in computation to achieve a major advance towards harnessing and understanding this information and its impact.

We started with addressing the question that most people ask “What is this online conversation about?”. To address this summarization problem, we used topic models to discover the latent topics that were discussed among users. The advantage of topic model is that it discovers insights from a collection of text without any information pre-specified by the user. To extract the topics structure with the best semantic coherence, we first focused on picking the best number of topics for a corpus using nonparametric bayesian method. We also proposed different approaches to label the multinomial topic distribution with keywords based on the topic words distribution and overall words distribution. Our automatic labels give a better interpretation of topics than the approach purely based on topic words distribution. By projecting online conversation into topics, we studied the temporal dynamics of different topics and identified the hot and cold topics at each time. Extraction of the temporal dynamics of a given topic allow users to trace information dynamics between online and offline social events. Furthermore, we studied topic distribution of the online conversations each user involved, which we show users have different preference towards different topics. We use this piece of information to model users, which would allow us to build a better recommendation system for knowledge sharing and social assistance.

Understanding user’s sentiment in online conversation is another important aspect to understand the information dynamics. Sentiment analysis refers to the application of natural language processing to identify and extract subjective information in source materials. We extract opinions from user generated content in an automatic fashion using unsupervised learning approach and project sentiment on different topics for opinion mining at aspect level.

To study information dynamics between users, we study user interaction networks inferred from online conversation. We find the degree distribution of network follow power laws, which implies a few highly active users and a majority of less active ones. We further extract community structure from user interaction network and it reveals interesting properties of this site. For example, each of the large four communities comprises users who were active together temporally (i.e., at the same time), and did not overlap with most of the users in other communities. Thus, while there are a few users who have been active for a long period, the users fall into four temporally non-overlapping groups, and within each such group there is no significant clustering of users based on their post-response patterns. This lack of clustering and patterns other than on a temporal basis, makes it imperative that one goes beyond the simplistic measure of who-responded-to-whom relationships and look at the content of the posts as well.

Furthermore, we use the content of posts to infer user interaction network. The a weighted and directed network is constructed as before, except that the edge is now assigned a weight of +1 if the sentiments of the two posts agree and -1 if they disagree. Thus, if sum of the weights of all the out-going edges from a node B is highly positive, then it implies that user B is agreed to by most of her responders. We find there are users that belong to all these four categories: (1) Mostly agreed to, (2) Mostly disagreed with, (3) Most of her responses disagreed with the person she was responding to, and (4) Most of her responses agreed with the person she was responding to. Section 4.7.1 summarizes some of our findings, including the fact that the most active posters are also the most agreed-with posters. This points to the fact that this forum is a fairly homogeneous and consensus driven community (mostly populated by anti-vaccination oriented mothers), which is also borne out by our manual reading of many of the posts. Overall, the users that were most-disagreed-with turned out to be either pro-vaccine enthusiasts or posted a link that was mostly pro-vaccine in orientation. Clearly, content analysis helps one to identify several key features of the dynamics of the group.

## 4.2 Related Work

### 4.2.1 Topic Models

With more and more knowledges being digitized online, news, blogs, web pages, scientific articles, books, images, sound, video, and social networks have reached to scale people never imagined before. Information overload has become a major concern for everyone in this Information Age. According to the report, there are 168,000,000 emails are sent every 60 seconds, 98,000 tweets go onto Twitter every 60 seconds, 695,000 Facebook status updates happen every 60 seconds. However, 91% of U.S. workers have deleted information without reading it, the average U.S. citizen consumes over 100,000 printed words a day (the size of a novel) and the average U.S. citizen receives over 63,000 words of new information per day. It becomes more difficult for us to discover useful information and still able to looking the information we want for given the size and growth of online collections. To develop the necessary tools for exploring and browsing online digital collections, we require automated methods of organizing, managing, delivering and understanding these vast amounts of information.

The form of information discovery techniques called topic modeling [SG07, Ble12] have been developed to address this issue efficiently. Topic models [BNJ03, GS04, BJ04, Hof99] are a suite of algorithms that uncover the hidden thematic structure in document collections . Latent Dirichlet allocation (LDA) [BNJ03] is a successful example of applying probabilistic graphical models on analyzing text collections. LDA is a generative model each document is viewed as a mixture of various topics. The generative process has documents represented as mixtures over latent topics, where each topic is characterized by a distribution over words.

The computational challenge in LDA is to compute the conditional distribution of the topic structure given the observed documents(called the posterior probability). The posterior probability can be express as the ratio between joint and marginal probability. While the joint distribution can be easily computed for any setting of the hidden vari-

ables, the marginal probability of the observations, which is the probability of seeing the observed corpus under any topic model is the one that is difficult to compute. In theory, marginal probability can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure. However that number of possible topic structures is exponentially large, which makes the posterior inference intractable to compute in this case. The general approach of posterior inference in topic models is to form an approximation by adapting an alternative distribution over the latent topic structure to be close to the true posterior. These algorithms generally fall into two categories: sampling-based algorithms [GS04] and variational algorithms [BNJ03].

With the research progress on generative models and posterior inference, probabilistic topic models have been applied to many kinds of content, including email documents, scientific abstracts [BNJ03, GS04], and newspaper archives [WC06]. By discovering patterns of word use and connecting documents that exhibit similar patterns, topic models have emerged as a powerful new technique for finding latent structure in a unstructured collection.

#### **4.2.2 Sentiment Analysis**

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks.

Document sentiment classification is the most widely studied problem, which is basically a text classification problem. [Tur02] presents a simple unsupervised learning

algorithm for classifying reviews, and the classification of a review is predicted by the average semantic orientation of the phrases that contain adjectives or adverbs. The semantic orientation of a phrase is calculated as the mutual information between the given phrase and the word excellent minus the mutual information between the given phrase and the word poor. [PLV02] directly applies three machine learning methods (Naive Bayes, maximum entropy classification, and support vector machines) to classify reviews into positive and negative. The feature they used is unigram, which has been studied by numerous researchers subsequently. [CMD06, DN09, PL04, NDA06] have tried a large set of features, like terms frequency and different IR weighting schemes, part of speech tags, opinion words and phrases, negations, syntactic dependency. At the same time, sentiment classification is sensitive to the domain of the training data, and existing research has used labeled data from one domain and unlabeled data from the target domain and general opinion words for learning [BDP07].

Document-level level sentiment classification is too coarse for most applications, so researchers study at sentence level as well. [WBO99] first identifies subjective sentences, which uses supervised learning and then sentiment classification of subjective sentence. [RW03] presents a bootstrapping approach that uses a high-precision classifier to label unannotated data to automatically create a large training set, which is then given to an extraction pattern learning algorithm.

Sentiment classifications at both the document and sentence levels are useful, but they do not find what people liked and disliked. Therefore we need to find both entities/aspect and opinions. To extract aspect, a frequency-based approach is proposed in [HL04], nouns that are frequently talked about are likely to be true aspects. To improve recall due to loss of infrequent aspects, it uses opinions words to extract them. [QLB11] extracts aspects using double propagation, exploiting the relations between sentiment words and product features that the sentiment words modify, and also sentiment words and product features themselves to extract new sentiment words.

### 4.3 Dataset Description

We focus our attention on the vaccinations forum of mothering.com, which has been active for more than ten years. We crawl its complete social forum and extract the online conversation in automatic way. We find there have been a total of 12,367 users (i.e., each such users has posted on the forum at least once) contributing 299,778 posts that comprise 26,942 threads. In addition to the active users, the forum draws readers who read but do not post, and the posts have a received a total of 16,329,543 views.

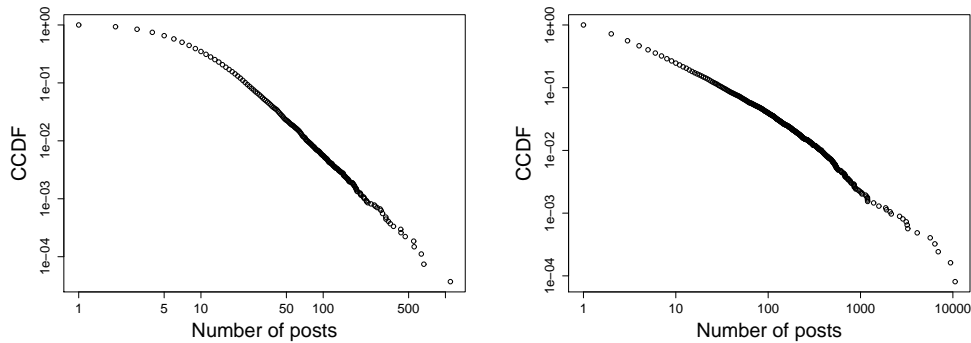


Figure 4.1: The number of posts per thread and the number of posts per user follow power law distribution

Figure 4.1 shows the distribution of the size of threads (measured by number posts in a thread) and the number of posts made per user. It is interesting to note that both fit power laws, with exponents of 2.87 and 1.77, respectively. Such power law distributions are typical of social network and social media and often points to its organic growth and development.

Figure 4.2 shows temporal dynamic of user activities based on the number posts user posted per week. Typical of such sites, it exhibits high volatility with a major spikes. It is interesting to observe the extent that outside events and conversations drive conversations on these forums. For this purpose we studied links included in posts. Results show that 24% of users post links, 10% of all posts include a link and 18% of threads are initiated with a post that includes a link.

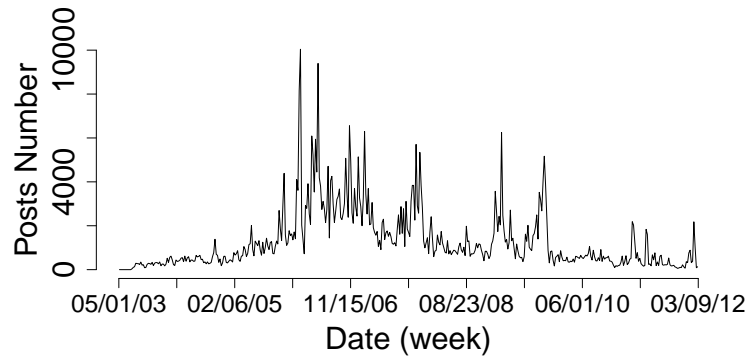


Figure 4.2: The number of posts per week shows the temporal information dynamics on the site

The number of posts per week not only depends on outside events, but also the number of active users on this social form. We plot the temporal dynamics of user joining based on the number of user registrations per week in Figure 4.3. We can see the social forum reached a high popularity between 2005 and 2007, attracting thousands of users to the forum. This dynamic has a major affect on the user interactions in Figure 4.2 as we can see the number of posts per week keeps at a high level during that period. These dynamics also show the social forum reached a pretty stable state after 2009 as both the number of posts and joined users no longer change dramatically.

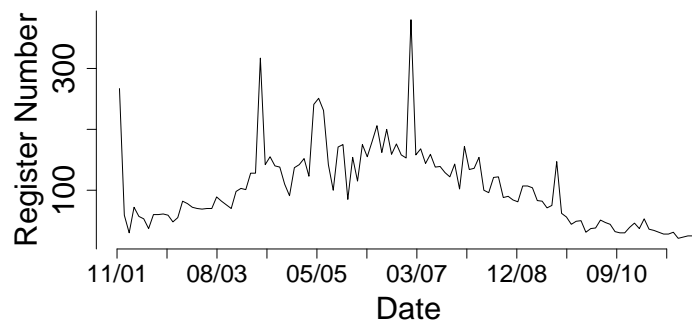


Figure 4.3: The number of joined users per week shows the temporal user dynamics on the site

These overall temporal dynamics present us interesting patterns for online activity between users on social forum, however it is difficult to discover information dynamics



without identifying information itself. To identify the online conversations on the social forums, we need computation tools such as topic models to discovery information and cluster conversations into different categories so that we can understand the information dynamics in a topical way. In the next section, we will present our topic modeling approach for online conversation.

## **4.4 Topic Discovery for Social Conversation**

### **4.4.1 Choosing Number of Topics**

The number of topics are assumed to be given in LDA model and it is shown in [CXL09] the number of topics would affect the interpretation of topics. As we expect, in a good topic structure of LDA, every topic is an meaningful and compact semantic cluster. Also conceptually the topic structures are hierarchical and corpus-specific. On the higher layer, we need fewer topics, but the topics are abstract and overlap with each other, which results in a lot of correlations to retain the discriminability. On the other hand, on the lower layer the topics are more concrete, then the information implicated in one topic is too little (every topic is a sparse vector in the large word space) to retain the discriminability. The number of topics determines the layer of the topic structure, therefore find the optimal number of topics is very important for applying topic models.

To find the number of topics is equivalent to a problem of model selection, which we use a standard method from bayesian statistics. Faced with a choice between a set of statistical models, the natural response is to compute the posterior probability of that set of models given the observed data. The key constituent of this posterior probability will be the likelihood of the data given the model, integrating over all parameters in the model.

In our case, the data are the words in the corpus  $w$ , and the model is specified by the number of topics  $K$ . We wish to compute the likelihood  $P(w|K)$  under different

possible value of  $K$ . In our case, we used  $\alpha = 0.1$  and  $\beta = \frac{50}{K}$  [GS04], keeping constant the sum of the Dirichlet hyperparameters, which can be interpreted as the number of virtual samples. We computed  $P(w|K)$  of testing documents for  $K$  values starting from 1 topics to 50 topics based on the topic models learnt from training data, and plot the perplexity under different the number of topic  $K$  in Figure 4.4. The result shows the perplexity per word goes down as the number of topic increases and then it goes up as the number of topic continue to increase. We determine the number of topics from this dynamics and pick the number of topics when the perplexity reached minimum at topic of 48.

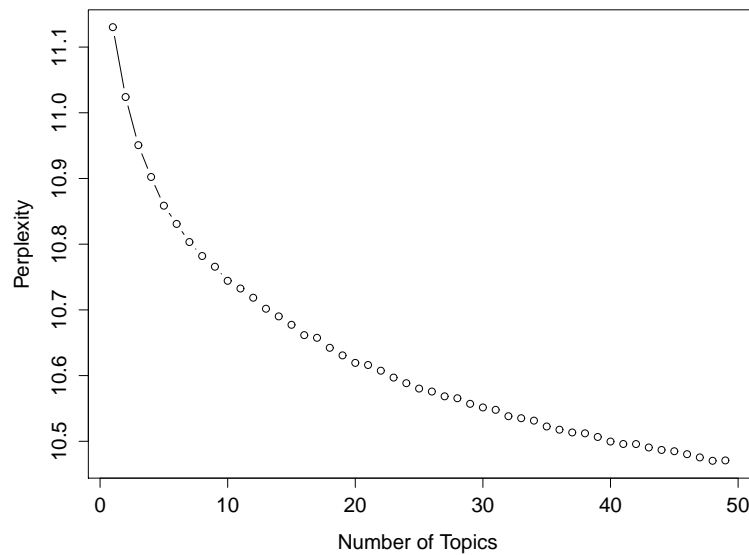


Figure 4.4: Perplexity per word under different number of topics

This model section approach is effective for picking up the best number of topic, however it is pretty slow in practice and it also depends on the parameters of prior. Teh et al. [TJB06] proposed hierarchical Dirichlet process(HDP) to determine the best number of topics in LDA. HDP is intended to model groups of data that have a pre-defined hierarchical structure, where each pre-defined group is associated with a dirichlet process whose base measure is sampled from a higher-level dirichlet process. Based on the similarity between HDP and LDA in structure, Teh et al. [TJB06] used the non-

parametric nature to resolve the problem of selecting appropriate number of topics for LDA. HDP replaces the finite topic mixture in LDA with a dirichlet process, and gives the different mixing proportions to each document-specific dirichlet process.

HDP [TJB06] presents a nonparametric approach to the problem of model section, however it is shown to be slow in practice as well. [BGJ10] presents the nested Chinese restaurant process (nCRP), a stochastic process that assigns probability distributions to ensembles of infinitely deep, infinitely branching trees. It is shown that this stochastic process can be used as a prior distribution in a bayesian nonparametric model of document collections. The posterior inference algorithm of nCRP finds an approximation to a posterior distribution over trees, topics and allocations of words to levels of the tree.

We applied nCRP for our online conversations and finds the best number of topics to be 25. We will use it as the number of topic of LDA model for online conversations going forwards. Table 4.1 shows the 25 topics based on word topic distribution and as expected most of topics discussed in this social forum is about child, health related issue. These topics present 25 word clusters with good semantic coherence and meanings.

Figure 4.5 shows the overall topic distribution in online conversation. We can see some topics are hot while the others do not get the similar amount of attention. Topics on scientific discussion, argument about vaccine are trending, which are expected as many users on this social forum are concerned about vaccine use for their children. More and more parents begin to realize the side effect of vaccine as scientific research has pointed it out, and oppose the idea of vaccination. Topic 18 is also about discussion of vaccine, in terms of friends talk. Topic 2 is also popular, which is about the disease caused by vaccine and parents' decision whether their children should have vaccine.

The above results show topic models only help us for information discovery, but also shows the natural distribution of the topics that been discussed. Social forum administer could use this information to quickly find the topics that have been discussed a lot. Users who are new to the social forums can quickly find the topics they are

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
school	vaccine	vax	vaccine	test
exemption	vax	kids	companies	dog
state	disease	sick	health	vax
vax	child	months	drug	rabies
form	research	baby	money	vaccine
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
vax	book	measles	doctor	vaccine
vaccine	read	pox	vax	cases
months	vaccine	immune	ped	children
shot	link	chicken	doc	reported
dose	info	vax	dr	death
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
food	baby	immune	vaccine	cough
vitamin	birth	vaccine	polio	pertussis
eat	hospital	disease	disease	fever
good	shot	hib	country	infection
body	nurse	system	travel	whooping
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
child	vaccine	vax	reaction	cancer
vaccine	mercury	people	vaccine	vaccine
parents	aluminum	kids	shot	hpv
sign	cells	thing	vax	hep
medical	injected	make	allergies	cervical
Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
people	flu	tetanus	autism	religious
vaccine	shot	wound	study	exemption
make	vaccine	shot	children	beliefs
thing	years	worry	vaccine	vaccine
study	people	years	link	state

Table 4.1: 25 topics represented by words topic probability

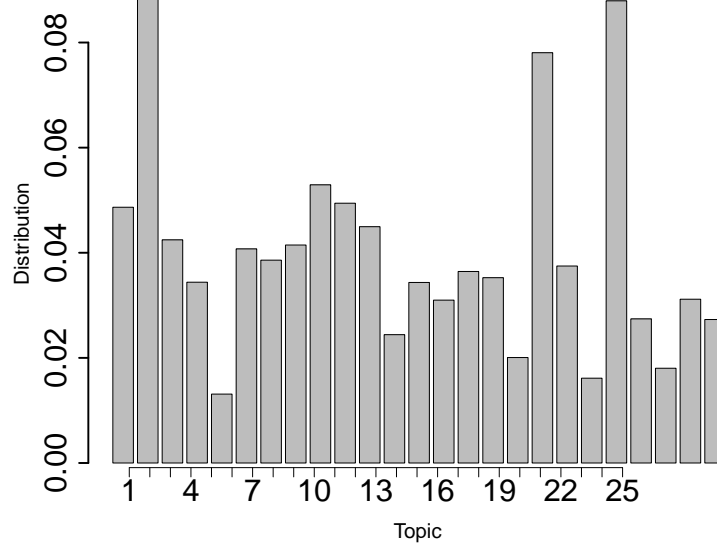


Figure 4.5: Overall topic distribution

interested and join the online conversation. Also topic models are truly unsupervised learning algorithms, no human interaction is required in the process, which offers another advantage considering nowadays' information overload.

#### 4.4.2 Labeling Topic Models

In previous section, we represent 25 topics according to their words topic distribution. The discovered word topic distributions are often intuitively meaningful, however a major challenge shared is to accurately interpret the meaning of each topic. As shown in Table 4.1, LDA models gives high probabilities to common words, such as vaccines, autism and people. These general words in this corpus does not give user more information. Overall, the topics are still hard to interpret for users who has little knowledge about what has been discussed on this social forum. Representing a topic merely based on the multinomial distribution might not be a ideal situation, especially users are expected to get more semantic meanings from each topic.

There is little work on how to automatically label the topic models. Mei et al. [MSZ07] propose probabilistic approaches to automatically labeling multinomial topic

models in an objective way. They cast this labeling problem as an optimization problem involving minimizing Kullback-Leibler divergence between word distributions and maximizing mutual information between a label and a topic model. Experiments with user study have been done on two text data sets with different genres. In this section, we propose two simple but effective algorithms to generate labels that are meaningful and useful for interpreting the discovered topic models.

At first approach, we take the global words distribution into consideration and normalize the words topic probability based on their word distribution in the other topics. We define *normalized word topic probability* as the word topic probability multiplied by the difference between logarithmic of topic probability minus the corpus probability. The normalized probability  $\beta'_{w,k}$  is given as follows where  $\beta_{w,k}$  represents the probability of word  $w$  in topic  $k$ .

$$\beta'_{w,k} = \beta_{w,k} \left( \log \beta_{w,k} - \frac{1}{K} \sum_{k'} \log \beta_{w,k'} \right) \quad (4.1)$$

We compute normalized word topic probability  $\beta'_{w,k}$  for each topic and further rank the words based on the  $\beta'_{w,k}$ . Figure 4.2 shows the top 5 words for the same 25 topics according to normalized word topic probability. We can see this label shows more semantic coherence than the one purely based on word topic probability. Also we can tell these topic labels are natural for people to understand and there are less overlap between labelled keywords among different topics.

Another approach we proposed is based on Pearson's chi-squared test. Pearson's chi-squared test  $\chi^2$  is the best-known of many chi-squared tests, a statistical procedures whose results are evaluated by reference to the chi-squared distribution. It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The Pearson's chi-squared test  $\chi^2$  is given as follows:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
school	vax	vax	companies	test
exemption	vaccine	kids	money	dog
state	disease	sick	government	tb
form	child	unvaxed	bill	rabies
religious	decision	months	vaccine	cat
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
months	book	measles	ped	cases
vax	read	pox	doctor	reported
hib	link	chicken	doc	death
dtap	info	cp	vax	study
schedule	information	rubella	visit	vaccine
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
vitamin	baby	immune	polio	cough
food	birth	hib	country	pertussis
eat	hospital	meningitis	travel	fever
diet	pregnancy	disease	smallpox	whooping
water	nurse	infection	disease	infection
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
child	mercury	people	reaction	hpv
sign	cells	kids	allergies	cancer
parents	aluminum	vax	shot	hep
medical	thimerosal	told	seizures	cervical
form	injected	friend	months	hepatitis
Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
people	flu	tetanus	autism	religious
study	shot	wound	study	exemption
science	h1n1	clean	children	beliefs
point	influenza	shot	autistic	religion
argument	season	cut	disorders	letter

Table 4.2: 25 topics represented by normalized probability

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (4.2)$$

In our case, we consider the significance of a word in a topic given the corpus. The null hypothesis in our application is the probability of words belongs to topic follows binomial distribution with probability of  $p(w)$  where  $p(w)$  is the probability of  $w$  in the whole collection. Empirically we can estimate  $p(w)$  by maximum likelihood as follows:

$$p(w) = \frac{\text{Total occurrence of word } w}{\text{Total number of words in corpus}} \quad (4.3)$$

Under binomial case, we use the notation of *binomial z score* and then  $\chi^2$  scores becomes as follows. Note  $D(k)$  represents the number of words in document  $k$ .

$$\chi^2(w, k) = \frac{(n(w, k) - p(w)D(k))^2}{(1 - p(w))p(w)D(k)} \quad (4.4)$$

The binomial z score measures the degree of independence of the word from the topic and Table 4.3 presents 25 topics based on binomial z score. We compare with the topics presented in Table 4.2 and find it gives very similar results and both representations give more details about the topic itself. In these two examples, we show both normalized word topic probability and binomial z score are good metrics to label the topic.

#### 4.4.3 Studying Topics Dynamics

We know at the different time, different topics will have different popularity, most likely to be triggered by the offline news and activities. By projecting the conversation into topic, we can find the temporal dynamics of different topics therefore finding the hot and cold topics at given time. In Figure 4.6, we show the temporal dynamics on



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
school	vaccine	kids	companies	dog
exemption	decision	unvaxed	bill	rabies
form	research	sick	money	test
state	disease	shedding	government	cat
required	vax	daycare	pharma	tb
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
dtap	book	pox	ped	death
dose	read	measles	doctor	reported
schedule	link	chicken	doc	cases
hib	info	rubella	visit	rate
months	site	cp	dr	number
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
food	birth	immune	polio	cough
vitamin	baby	meningitis	travel	pertussis
eat	hospital	bacteria	country	fever
diet	pregnancy	antibodies	smallpox	whooping
supplements	newborn	hib	opv	ear
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
sign	mercury	friend	reaction	hpv
consent	aluminum	talk	allergies	cancer
parents	cells	mom	seizures	cervical
neglect	thimerosal	dh	eczema	hep
refuse	formaldehyde	lol	screaming	gardasil
Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
people	flu	tetanus	autism	religious
science	h1n1	wound	autistic	exemption
argument	shot	tig	disorders	beliefs
scientific	swine	puncture	study	religion
evidence	season	bleeding	genetic	church

Table 4.3: 25 topics represented by binomial z score

Topic 22 (about Flu) in green and Topic 20 (about human papillomavirus) in blue. We extracted the posts that are classified to these two topics and find there are 13523 posts related to Flu and 4736 posts related to human papillomavirus(HPV).

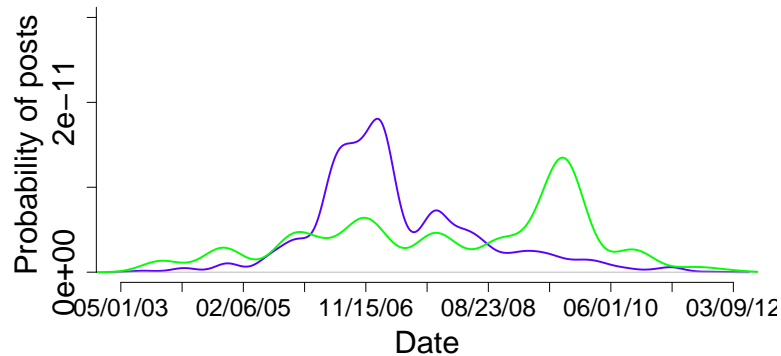


Figure 4.6: Temporal dynamics of Topic 22 (in green) and Topic 20 (in blue)

We know US faces flu vaccine shortage in October 2004 and the outbreak of swine flu in 2009. The observed two peaks in the Figure 4.6 corresponding to these events respectively, one toward the end of 2004 and another at the end of 2009. We can also verify this by studying some example threads, such as the thread titled “Who here is getting a flu shot? copied below:

I cannot believe the hype this year about the flu shot:

<http://www.cnn.com/2004/HEALTH/10/16....ap/index.html>

People are crazed to get a shot! I personally don't want one (even less than usual) this year because I still don't believe that the entire batch isn't tainted! In the mad rush to get the vaccines to market who knows what corners drug companies are willing to cut. As a healthcare provider I am offered a shot at my clinic and expected to get one. I'm not gonna do it.

What are the feelings on this matter?

-Laura

Figure 4.6 shows the peak of HPV discussion happened at the end of January 2007, and the related threads discussed Gardasil from CDC. (Gardasil is a vaccine for use in

the prevention of certain types of HPV). The other peak of discussion happened at the end of September 2007, and the related threads discussed Michigan law requiring 6th grade girls to get HPV vaccine. Discussion on HPV topic on the social forum confirmed our observation of the following offline HPV vaccine events:

- June 2006: FDA approves use of Gardasil in the U.S. for girls 11 to 12 with catch up vaccination for females up to 26 years (and vaccination of girls as young as 9 years old) for prevention of cervical cancer and genital warts. We can see there are some discussions on the forum about this event, for example the following thread discussed the news from cnn health:
- June 2006: CDCs Advisory Committee on Immunization Practices (ACIP) provisionally recommends the quadrivalent HPV vaccine. We can find several examples of discussions on the forum about this event, with a lot of negative sentiment/feeling towards this news item.
- February 2007: Texas governor Rick Perry signs an executive order stating that all girls entering 6th grade would need to receive HPV vaccine; the executive order was quickly overturned by the Texas legislature There is a short thread discussing this issue on the forum, with the associated link.
- March 2007: CDCs Advisory Committee on Immunization Practices (ACIP) officially recommends the quadrivalent vaccine for girls
- September 2009: Death of a school girl in England, Natalie Morton, after receiving Cervarix heightens fears about the vaccine and is reported widely in the media worldwide, autopsy revealed serious underlying health condition (malignant tumor of the chest), which was likely cause of death

Table 4.4 shows the top sites these posts referred to with total counts, which is similar to our overall site statistics.

Site domain	Link count
www.cdc.gov	59
www.mothing.com	52
www.fda.gov	49
www.merck.com	44
us.gsk.com	18
www.medalerts.org	13
www.newstarget.com	11
www.909shot.com	10

Table 4.4: Top top sites referred by HPV posts

#### 4.4.4 Modeling User Interest

We expect different users have different speciality or interests in online discussion. By modeling online conversation, we can model users' interest by studying their posts. For each user, we compute the number of posts he/she posted in 25 topics, and get the topic distribution for each user. In Figure 4.7, we show the histogram of Kullback-Leibler divergence(KL Divergence) between overall topic distribution and individuals topic distribution for all the 12367 users. If people are uniformly posting over 25 topics, the KL Divergence will be around 0.2. Based on statistics shown in Figure 4.7, we can see clearly different users are showing different preference over the different topics. For example, user Sherlock shows a higher interest in discussions about Topic 24(about autism). User Jenelle has preference over discussions of Topic 22(about flu).User Mamaterra has participated more in conversations Topic 20 (about HPV).

Furthermore, we use k-means clustering algorithm [HW79] on user topic vectors and cluster users into four different categories. Figure 4.8 shows the cluster centers for four clusters and it is interesting to see three of them are mainly concentrated on single topic. These three clusters are made up of users who only focus on certain topic while users in cluster 3 are a group of users who actively participated in various kind of topics.

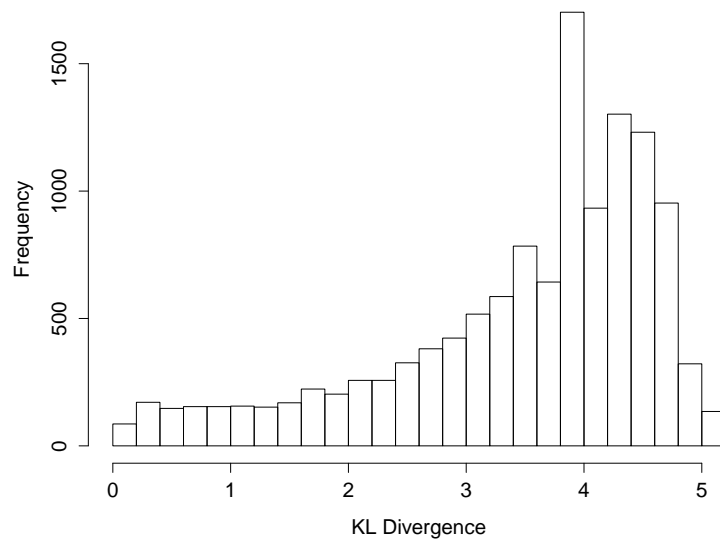


Figure 4.7: Topic users

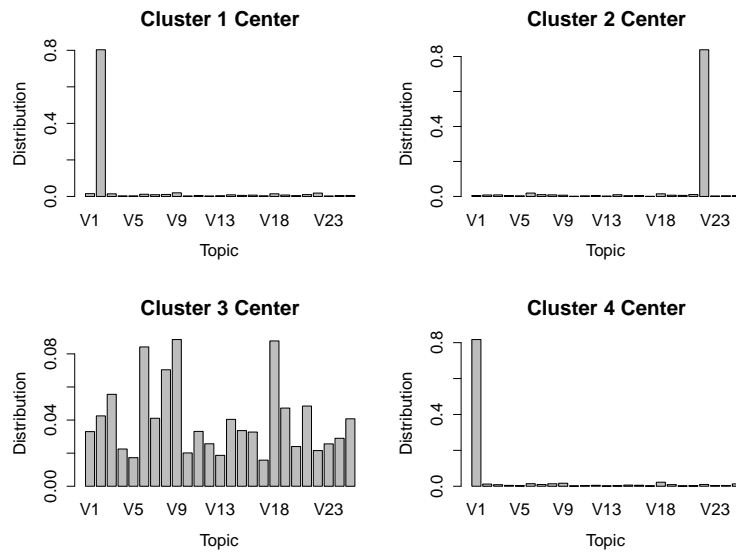


Figure 4.8: Four user clusters represented by topic centers according on K-means

## **4.5 Sentiment Extraction for Social Conversation**

### **4.5.1 Our Approach and Result**

We want to utilize the topic models to uncover latent aspects on the social forum and extract sentiment meanings regarding these aspects. Using LDA to uncover latent topics in a document collection, we want to study user's sentiment towards these topics or aspects.

We present an unsupervised system for extracting aspects and determining sentiment in online conversation. The method is simple and flexible with regard to domain and language, and takes into account the influence of aspect on sentiment polarity. We detect sentiment based on lexical resources such as a dictionary of opinionated terms. SentiWordNet [ES06] is one such resource, containing opinion information on terms extracted from the WordNet database and made publicly available for research purposes.

We apply a Lexicon based algorithm to detect sentiment using SentiWordNet and present our result of sentiment per posts over the time in Figure 4.9. We can see the sentiment per posts increase and decrease in a swift fashion. We believe the ups and downs are triggered by the information dynamics happened outside the social conversation regarding specific topic. To extract the sentiment regarding to specific topic, we combined our results from topic models and extract user sentiment towards topic.

### **4.5.2 Topical Sentiment Analysis**

First, we studied the sentiment by thread. A thread is a collection of post under the same title, which is the discussion of the unique topic in online conversation. We would like to see the overall sentiment contained in each thread and find out the most positive and most negative one. The most positive (happy) thread is about nutrition and Immunology suggestions for mothers. The most negative (sad) thread is about a refusal form for

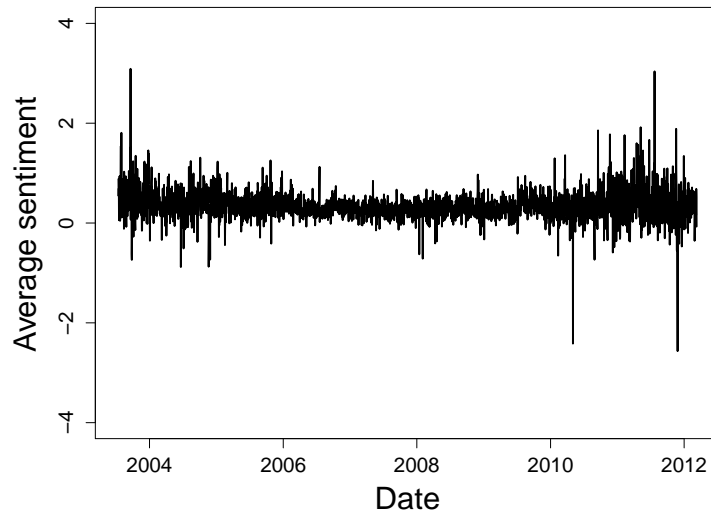


Figure 4.9: Temporal dynamics of sentiment by week

vaccine. This evidence shows there are many users on this social forum are oppose to vaccine.

Then we use the latent topics discovered from topic models to study the sentiment per topic. For each topic, we use the average sentiment score by posts as the measurement for 25 topics and show our result in Figure 4.10. We can see most positive topic is Topic 25 (about religion) while are not that happy about different topics on diseases, which are expected.

## 4.6 User Interaction Network and its Implications

Modeling the information based on topic and sentiment gives us a useful tool to study the information dynamics between users. User friendship connections are necessary to study user dynamics on the network, however on this social forums, the user interaction network is not available through explicit social network connections. We need to infer the user interaction network based on user activities. In our user interaction network, we represent each node as a user, and a directed edge from user A to user B means

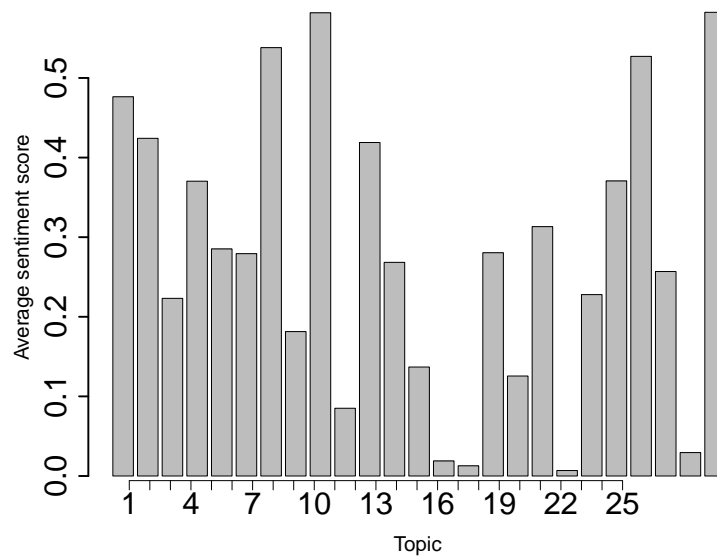


Figure 4.10: Sentiment score for 25 topics

information flows from A to B. In any thread, B might reply directly to another user A by quoting A’s post, which is represented by an edge from A to B. If there is no direct quote of other user’s post, we assume that B is replying to the thread initiator’s post.

We infer user interaction network and find there are 12367 nodes and 158711 edges. The out-degree distribution is a power law with the exponent 2.47 as shown in Figure 4.11. Higher out-degree of a user means that more people responded to this users posts. High out-degree users are a group of users who actively post information on the forum. Interestingly, we find these most active users form a small core in the network and with the high degree nodes removed, the network decomposes into many clusters, where most of the clusters are small with only one node. This confirms that at any particular time, there are a few dominant users responsible for most of the activity of the network.

The distribution of the lengths of user activity time (the interval between the last and first posts) by user is shown in the Figure 4.12. The distribution has an exponential tail, which shows user’s activity decay fast on social forums. This finding confirms that the high turnover rate on found on the other social networks too.

To further study the user connections on this forum, we represent user interaction



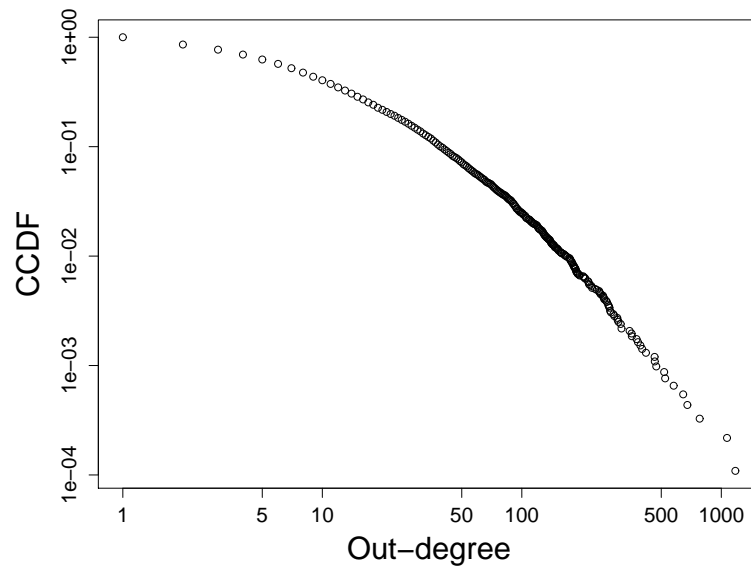


Figure 4.11: Out-Degree distribution in user interaction network

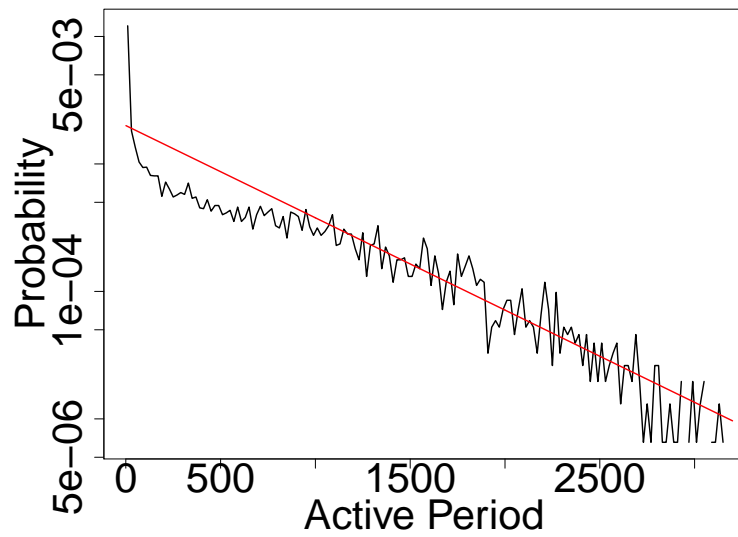


Figure 4.12: Distribution of user activity time

network as weighted network, with weight of an edge from A to B being the number of times that user B responded to user A. We applied a community finding algorithm [GN02] to determine its community structure. The modularity of the community structure is 0.31, which shows strong community structure. There are 4 major communities with nodes numbering larger than 150, inside this network weighted by the number of responses. They have 2087 nodes, 3034 nodes, 1633 nodes and 5343 nodes respectively. The probability distribution for the times of each communitys posts as well as their post activity profiles are shown in Figure 4.13.

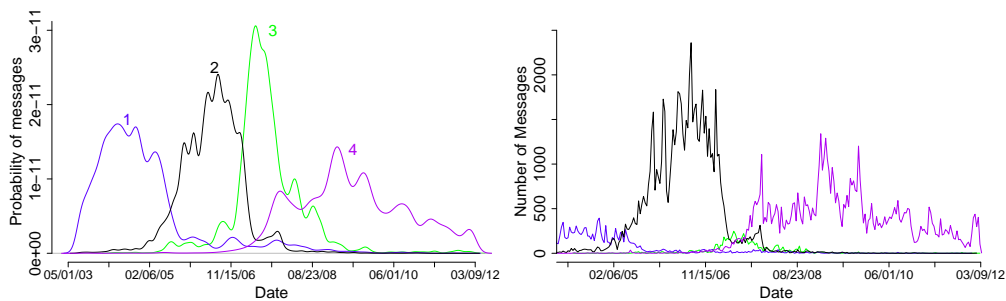


Figure 4.13: The temporal dynamics of posts from each community in user interaction network by probability and by counts

Clearly, the users belonging to communities 1-4 occupy different regions of time periods, thus we reach the following conclusion. At any particular time, a few users dominate the conversations on this site, and there are four such major groups who have operated at different times. Within each such time period there is no distinct clustering of users based on post-response interaction patterns. Clearly, in order to get more understanding of the dynamics of the network we have to turn to analyzing the content of the posts, which will be addressed in the next section.

## 4.7 Sentiment-based Interaction Network and Implications

To take content into consideration, we use our sentiment analysis results and associate an sentiment score to each user interaction. We define the weight of an edge from A

to B as the sentiment score. Suppose user B responded to one of user A's posts, and the sentiment of A's original post is  $S_A$  and the sentiment of B's response is  $S_B$ . We define the score of this particular exchange as  $sign(S_A)sign(S_B)$ . Since smaller values of sentiments are not very accurate, we set a threshold of 0.2, and if  $|S_A| < 0.2$  we set the score of this exchange to zero. We sum up the scores of all the responses from B to A, and this sum is now used as the weight of the edge from A to B. The edge weight distribution is shown in the Figure 4.14. The number of edges with positive weights is 79118; the number of edges with negative weights is 55796. The cumulative distribution for the positive weight has a fat tail with the power law -2.61.

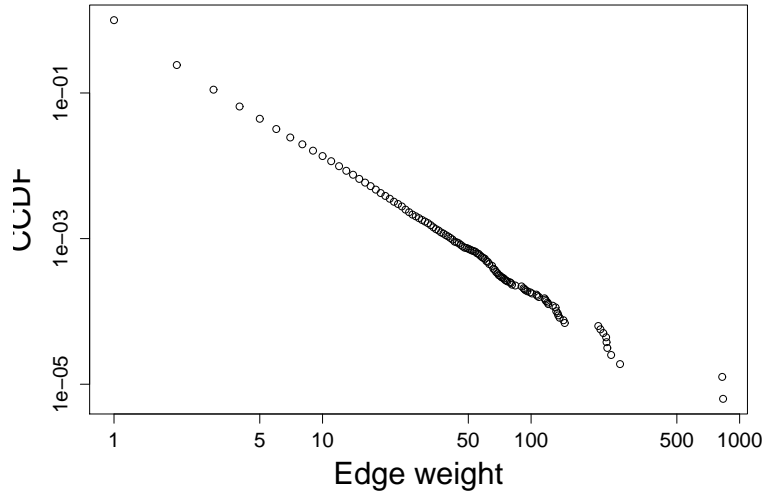


Figure 4.14: Distribution of positive edge weights which fits a power law with exponent 2.61

#### 4.7.1 Identify Key Users

We want to identify key users in user interaction network based on network statistics and study their roles in online conversation. If the edge from A to B has a positive weight, B usually agreed with A. Otherwise, B usually disagreed with A. Thus, the sum of user A's all out-edge weights describes whether A was agreed/disagreed by other users. And the sum of user A's all in-edge weights describes whether A usually

agreed/disagreed with others.

We find there are 964 users whose 95% of posts are agreed by other and 1968 users who agree with others more than 95% of his/her replies. These are a significant number of users that have high total scores and deserve further investigation. We summarize our observations about the top users identified as follows.

**Highest out-degree** High out-degree users represent users are most agreed with with. We find many of them also appear in agreeing with others. These users are also the most popular and active users in forums, which implies a sentiment-wise and opinion-wise homogeneous forum: the most active users are also the most agreed with.

**Lowest out-degree** Users with lowest out-degree (others disagree with them, i.e. replies are mostly opposite of their sentiment) are also equally informative. One poster is a pro-vaccination parent. In fact she ended up changing her mind and deciding not to vaccinate, based on the overwhelming negative reaction to her initial intent to vaccinate. The other user is a pediatrician (also pro-vaccination): But the opposing sentiments were also observed to the user with this following thread (suggesting that vaccines are part of a conspiracy):

**Highest in-degree** High in-degree users are most agreeing with others. Those most agreeing are mostly similar to those who are most agreed with. Those who agree most with others and those whom others agree with most, we see a large overlap, however there are a few users who tend to gain agreement but do not always agree with others. We find one user is quite popular, she does not always try to be pleasant.

Thus, this simple macroscopic analysis can highlight different classes of users that deserve especial attention from a dynamic perspective.

#### 4.7.2 User Clustering

We also performed community structure determination for this sentiment-based interaction network. The modularity of the community structure is 0.34 and there are 6

major communities with nodes numbering larger than 150. They have 1232 nodes, 2561 nodes, 1413 nodes, 3409 nodes, 783 nodes and 602 nodes respectively. The posts posted by the users inside each community have time stamps and the probability distribution for the times of each communitys posts is shown in Figure 4.15. Communities 1 to 4, which are the most dominant in terms of total number of posts, occupy different regions of time periods. Communities 5 and 6 occupy much longer time periods.

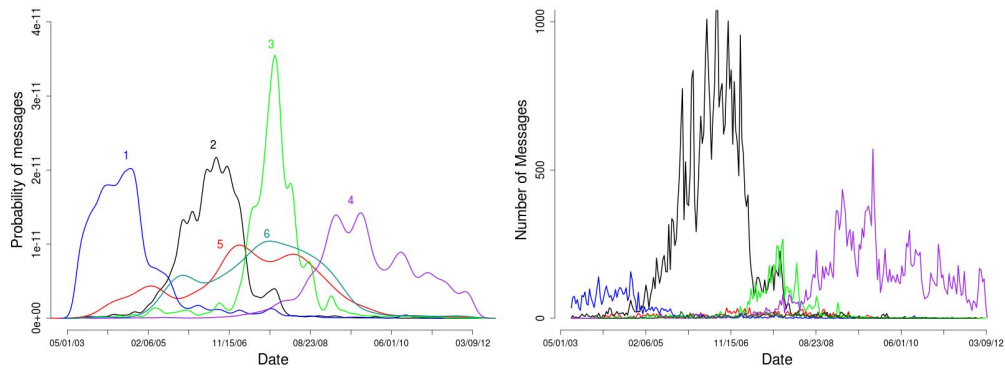


Figure 4.15: The temporal dynamics of posts for each community in user sentiment-based network by probability and by counts

### 4.7.3 Hierarchical User Clustering

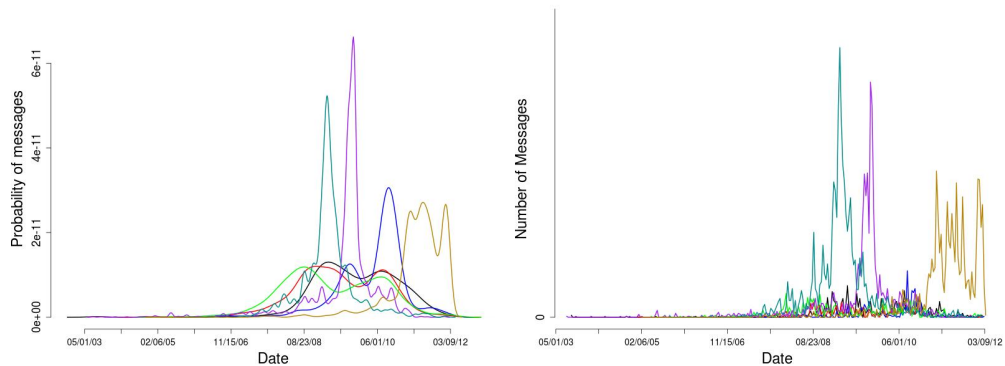


Figure 4.16: The temporal dynamics of posts for each subcommunity of community 4 by probability and by counts

While communities 1 to 4 are similar to the communities found in Section 4.6, the activities of users communities 5 and 6 spread over a much larger time period, and are

basically centered around two dominant users of this forum: Community 5 is formed around the one user. Community 6 is formed around another user. Most of the edges in the two communities start or end at these two users, and they are by far the most long-lived leaders of this community.

In Figure 4.16, we plot the similar measurements for the sub-community structure of community 4. The figures have similar implications as the previous figures: community structure is represents temporally exclusive activity patterns by different groups of users.

## 4.8 Conclusion and Discussion

In this Chapter, we address different aspects of topic models to extract the latent topics with the semantic meanings. We proposed two effective approaches to label the topics with keywords and our automatic labels give a better interpretation of baseline approach. Extraction of the temporal dynamics of a given topic allow users to trace information dynamics between online and offline social events. Furthermore, we show users have different preference towards different topics based on the online conversations they involved in. We use this piece of information to model users, which would allow us to build a better recommendation system for knowledge sharing and social assistance. We also extract opinions from user generated content in an automatic fashion and project sentiment on different topics for opinion mining at aspect level.

To study information dynamics between users, we first inferred user interaction network from online conversation. We find the degree distribution of network follow power laws, which implies a few highly active users and a majority of less active ones. We further extract community structure from user interaction network and it reveals interesting properties of this site. We find there are a few users who have been active for a long period, the users fall into four temporally non-overlapping groups, and within each such group there is no significant clustering of users based on their post-response

patterns. This lack of clustering and patterns other than on a temporal basis, makes it imperative that one goes beyond the simplistic measure of who-responded-to-whom relationships and look at the content of the posts as well.

To address this issue, we use the sentiment of posts to infer user interaction network. The a weighted and directed network is constructed as before, except that the edge is now assigned a weight of +1 if the sentiments of the two posts agree and -1 if they disagree. Thus, if sum of the weights of all the out-going edges from a node B is highly positive, then it implies that user B is agreed to by most of her responders. We find there are users that belong to all these four categories. Our result points to the fact that this forum is a fairly homogenous and consensus driven community (mostly populated by anti-vaccination oriented mothers). Content analysis helps one to identify several key features of the dynamics of the group.

## CHAPTER 5

### Discover Hidden Structure on Large-scale Corpus

As collective human activity and knowledge continues to be digitized and stored, it provides an unprecedented opportunity to understand what topics are important, how they evolve, and how individuals and organizations interact to form groups and make decisions. The petabytes of data collected everyday, however, underscores the need for new computational tools to help organize and understand these vast amounts of information. In this chapter, we develop a novel methodology for Topic Models, where given a large corpus of documents, it automatically infers the underlying topics and computes a distribution of words over the computed topics. Our approach is very different from the highly popular and widely used existing topic models: instead of using a bag of words model, it is inspired by how knowledge is organized in our brains as an associative network, and it exploits the idea of source coding from information theory to infer the latent networks directly from text data. We apply our algorithms on large-scale corpus, and using automatic evaluation techniques, show that our topic organization is not only more coherent semantically, compared to the state-of-the-art Latent Dirichlet Allocation (LDA) results, but is also computationally more efficient.

#### 5.1 Introduction

With more and more data being digitized online, news, blogs, web pages, scientific articles, books, images, sound, video, and social networks have reached to scale people never imagined before. According to the report, there are 168,000,000 emails are sent



every 60 seconds, 98,000 tweets go onto Twitter every 60 seconds, 695,000 Facebook status updates happen every 60 seconds. However, 91% of U.S. workers have deleted information without reading it, the average U.S. citizen consumes over 100,000 printed words a day (the size of a novel) and the average U.S. citizen receives over 63,000 words of new information per day. Information overload has become a hot topic not only for researchers but also for everyone as it becomes more difficult to discover useful information and still able to looking the information we want for given the size and growth of online collections.

However, it also provides an unprecedented opportunity for computer scientists to use powerful machine learning tools to study online collections and to finally understand both: (i) what is important information wise, and how such informational clusters, i.e., topics, interact and evolve, and (ii) how individuals and organizations interact. While the opportunity is very real, the cliché of petabytes of data being generated everyday is also very real, underscoring the need for new computational tools to help organize, search, manage, deliver and understand these vast amounts of information.

We propose a novel methodology for Topic Models, where given a large corpus of documents, it automatically infers the underlying topics and computes a distribution of words over the computed topics. Our approach is very different from the highly popular and widely used existing topic models: Instead of using a bag of words model, it is inspired by how knowledge is organized in our brains as an associative network, and it exploits the idea of source coding from information theory to infer the latent networks directly from text data. We propose generative models for organizing topics and generating documents based on human knowledge. Given the observed documents, we develop scalable inference algorithms to estimate latent structure of associative network and discover the hidden topic organization by minimizing the description length. We apply our algorithms on large-scale corpus and compare our results with LDA model using automatic evaluation based on external resources. Our topic organization is shown to be more coherent semantically compared LDA results, and the inference algorithm

is proven to be computationally more efficient as well.

## **5.2 Related Work**

Processing human language requires the retrieval of concepts from memory in response to an ongoing stream of information. This retrieval is facilitated if one can infer the topic of a sentence, conversation, or document and use that topic to predict related concepts and disambiguate words. Characterizing the content of documents is a standard problem addressed in information retrieval, statistical natural language processing, and machine learning. A representation of document content can be used to organize, classify, or search a collection of documents. In this section, we review the related work on solving this problem based on probabilistic topic models and semantic network.

### **5.2.1 Topic Models**

As documents can be presented as term-document matrix which describes the occurrences of terms in documents; it is a sparse matrix whose rows correspond to terms and whose columns correspond to documents. A typical example of the weighting of the elements of the matrix is tf-idf (term frequency inverse document frequency): the element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up weighted to reflect their relative importance.

Researchers often apply a statistical method such as Latent Semantic Analysis (LSA) [DDL90, LD97, LFL98] to analyze relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text. Singular value decomposition (SVD) is applied on term-document matrix to reduce the number of terms while preserving the similarity structure among document. The LSA approach makes three claims: that semantic information can be derived from a word-document co-occurrence matrix; that dimensionality reduction is an essential

part of this derivation; and that words and documents can be represented as points in Euclidean space.

Recently another form of information discovery techniques called topic models [SG07, Ble12] have been developed to address this issue efficiently. Topic models [GS03, Hof99, Hof01] is consistent with the first two of these claims of LSA, but differs in the third, describing a class of statistical models in which the semantic properties of words and documents are expressed in terms of probabilistic topics. Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Statistical inference techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents.

Representing the content of words and documents with probabilistic topics has one distinct advantage over a purely spatial representation. Each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms. Latent Dirichlet allocation (LDA) [BNJ03] is a successful example of applying probabilistic graphical models on analyzing text collections. And it has been applied to various kinds of content, including email documents, scientific abstracts [BNJ03, GS04], and newspaper archives [WC06]. By discovering patterns of word use and connecting documents that exhibit similar patterns, topic models have emerged as a powerful new technique for finding latent structure in a unstructured collection.

## 5.2.2 Limitation of Topic Models

### 5.2.2.1 Choice of Number of Topics

The number of topics are assumed to be given in LDA model and it is shown in [CXL09] the number of topics would affect the interpretation of topics. As we expect, in a good topic structure of LDA, every topic is an meaningful and compact semantic cluster. Also conceptually the topic structures are hierarchical and corpus-specific. On the higher layer, we need fewer topics, but the topics are abstract and overlap with each other, which results in a lot of correlations to retain the discriminability. On the other hand, on the lower layer the topics are more concrete, then the information implicated in one topic is too little (every topic is a sparse vector in the large word space) to retain the discriminability. The number of topics determines the layer of the topic structure, therefore find the optimal number of topics is very important for applying topic models.

Document 1	money stock finance
Document 2	finance money bank
Document 3	water bed stream
Document 4	water stream bank

Table 5.1: A simple corpus consists of four documents

Let us use a simple example with four short documents as shown in Table 5.1. We can see document 1 and document 2 are about topic of finance while document 3 and document 4 are about topic of water stream. The word “bank” is a polysemy here, which has meanings of both topics. Let us see how the choice of number of topic could affect topic representation.

Figure 5.1 shows the word topic distribution for all seven words if the number of topic is chosen as two. LDA model inference the words topic distribution exactly what we expected and word “bank” is involved in both topics with equal probability.

However if user has not prior knowledge about what number of topic they should specify and choose 3 to be the number of topics in this case, the topic clusters might

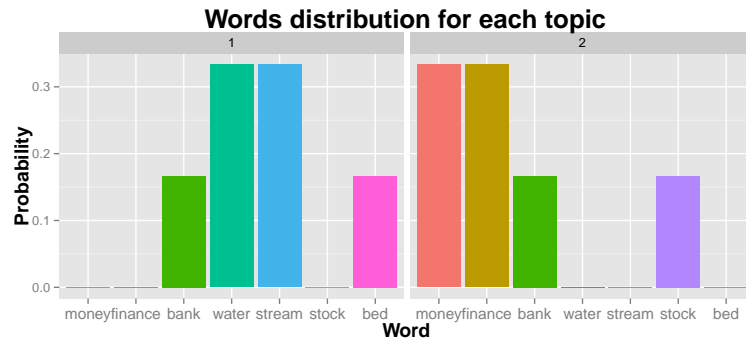


Figure 5.1: Word topic distribution as the number of topic is 2

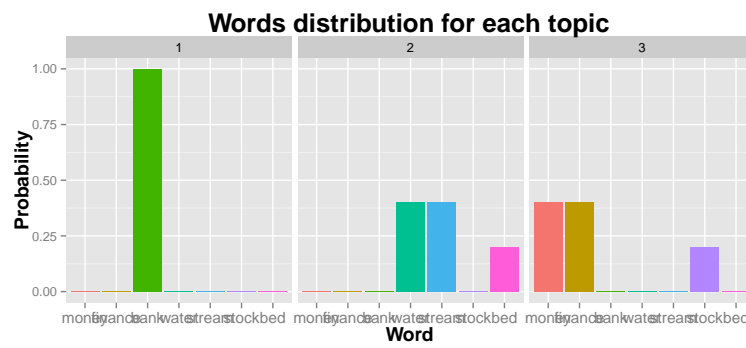


Figure 5.2: Word topic distribution as the number of topic is 3

not have the same interpretation as we expect. We applied LDA model and the word topic distribution is shown in Figure 5.2 for three topics. The word “bank” is clustered into a single topic and as a result the document 2 is treated as the mixture of topic 1 with topic 2. These simple observations show that the choice of number of topics can really affect the topic coherence in the LDA results.

Teh et al. [TJB06] proposed hierarchical Dirichlet process(HDP) to resolve the problem of selecting appropriate number of topics for LDA. HDP replaces the finite topic mixture in LDA with a dirichlet process, and gives the different mixing proportions to each document-specific dirichlet process. However HDP is shown to be slow in practice.

#### **5.2.2.2 Topic Correlation**

The LDA model assumes that the words of each document arise from a mixture of topics, each of which is a distribution over the words. Another limitation of LDA is the inability to model topic correlation. This limitation stems from the use of the Dirichlet distribution to model the variability among the topic proportions. Under a Dirichlet, the components of the proportions vector are nearly independent; this leads to the strong and unrealistic modeling assumption that the presence of one topic is not correlated with the presence of another.

In many indeed most text corpora, it is natural to expect that subsets of the underlying latent topics will be highly correlated. In a corpus of news, for instance, an article about health may be likely to also be about disease, but unlikely to also be about astronomy. To address this limitation, Blei et al. [LB05] develop the correlated topic model (CTM), where the topic proportions exhibit correlation via the logistic normal distribution. Li et al. [LM06] introduce the pachinko allocation model (PAM), which captures arbitrary, nested, and possibly sparse correlations between topics using a directed acyclic graph (DAG). The leaves of the DAG represent individual words in the

vocabulary, while each interior node represents a correlation among its children, which may be words or other interior nodes (topics).

### **5.2.2.3 The Length of Document**

It is reported LDA model does not work well on short texts [HD10] for topic discovery. Hong et al. [HD10] propose several schemes to train a standard topic model and compare their quality and effectiveness through a set of carefully designed experiments from both qualitative and quantitative perspectives. They show that by training a topic model on aggregated messages we can obtain a higher quality of learned model which results in significantly better performance in two real world classification problems.

To qualitatively investigate how the length of document influences LDA results, we generated synthetic documents according to LDA generative models and study the performance of LDA model through document classification. We created a set of synthetic documents with a pre-defined topic structure. We first set up  $n$  groups of words with or without some overlaps, and here each group is a topic. Then we extract words randomly from the groups to make a set of documents according to LDA generative model, which has the desired topic structure.

To evaluate the document classification result, we apply LDA on the set of documents, and compare the LDA result with the pre-defined topic structure. We plot the document classification error versus the average length of document in Figure 5.3 and it turns out that the average document length must be large enough (about 80 words) to obtain the correct result.

### **5.2.2.4 Understanding words topic distribution**

The inference algorithm of topic models finds the words topic distribution. Given a topic, usually words distribution are often intuitively meaningful, however a major challenge shared is to accurately interpret the meaning of each topic.

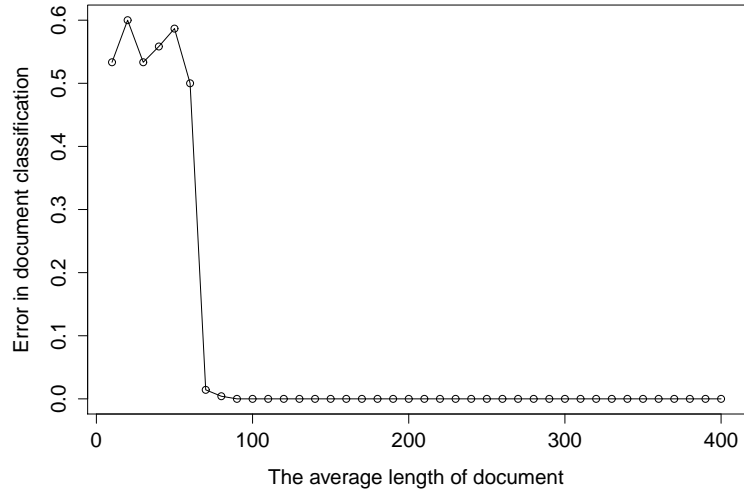


Figure 5.3: Document classification error versus the average document length

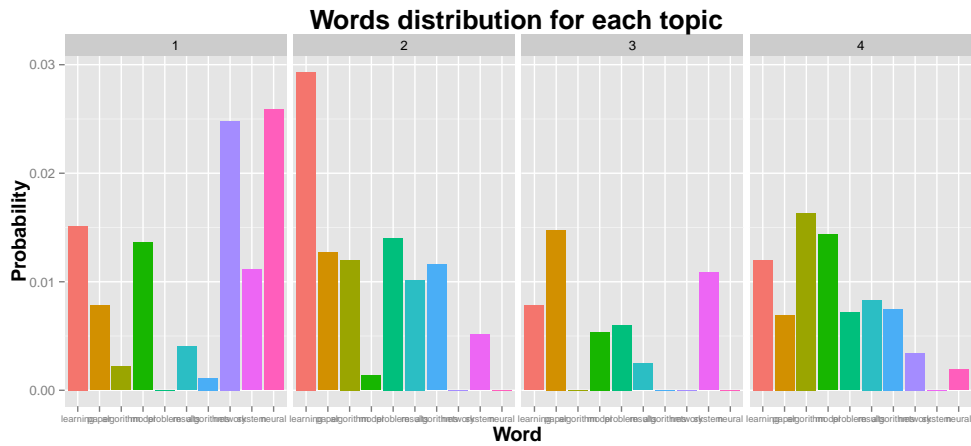


Figure 5.4: Word topic distribution for top 10 words in scientific documents



In order to show this problem, we train a LDA model on a collection of 2410 scientific documents with links and titles from the computer science research paper using four topics assignment. We show the topic distribution for top 10 words in each topic respectively in Figure 5.4. As expected, LDA models gives high probabilities to these top 10 words in all four topics, which are also common words in the respective corpus. From the inference algorithm perspective, posterior inference of LDA model is trained to maximize the posterior probability of latent variables given data observation, so maximizing these common words is a part of optimization goal. However these words are usually generic words in the respective corpus (like learning, Bayesian in our example), which does not give detail information for each topic. As we can see, all four topics are pretty general in terms of top words and topics are pretty hard to interpret.

To address this issue, there are some researches efforts on how to automatically label the topic models given word topic distribution in topic models. Mei et al. [MSZ07] propose probabilistic approaches to automatically labeling multinomial topic models in an objective way.

### **5.2.3 Semantic Network**

A semantic network is used when one has knowledge that is best understood as a set of concepts that are related to one another. It was introduced in the theory of Collins and Quillian [CQ69], whose work centers on how natural language is understood and how the meanings of words can be captured in a machine. In terms of representation, a set of words or concepts is represented as nodes connected by edges that indicate pairwise associations. The edges are directed and labeled; thus, a semantic network is a directed graph. This represents the simplest form of a semantic network, a collection of undifferentiated objects and arrows. The structure of the network defines its meaning, which are merely which node has a pointer to which other node. Most semantic networks are cognitively based and they also consist of arcs and nodes which can be organized into a taxonomic hierarchy.

Semantic networks provide an intuitive framework for expressing the semantic relationships between words. The structure of semantic networks such as networks formed by word association has proven to be useful to predict performance in a variety of experimental tasks such as recall and recognition [Dee66, NMS04]. Steyvers et al. [ST05] investigate the large scale structure of several semantic networks constructed by different means by measuring a few statistical properties. These statistical properties can then be used to distinguish semantic networks from other networks such as random networks where concepts are linked by random connections. Semantic networks also contributed ideas of spreading activation, inheritance, and nodes as proto-objects. They are intractable for large domains.

### **5.3 Our Approach: Associative Network**

Our approach of topic models is to use associative network, as inspired by how knowledge is organized in our brains, to represent semantic relations between words. We estimate the corpus-specific associative networks based on different sources of large scale corpus and develop scalable inference algorithms to discover the hidden topic organization by minimizing the description length.

#### **5.3.1 Our Topic Model**

Human memory has a vast capacity, storing all the semantic knowledge, facts, and experiences that people accrue over a lifetime. Given this huge repository of data, retrieving any one piece of information from memory is a challenging computational problem. In fact, it is the same problem faced by libraries [And90] and internet search engines [GSF07] that need to efficiently organize information to facilitate retrieval of those items most likely to be relevant to a query. It thus becomes interesting to try to understand exactly what kind of algorithms and representations are used when people search their memory.

Recently Hill et al. [HJT12] find that the evidence that human memory search is similar to animals foraging for food in patchy environments, with people making a rational decision to switch away from a cluster of related information as it becomes depleted. Abbott et al. [AAG12] demonstrate that these results that were taken as evidence for this account also emerge from a random walk on a semantic network, much like the random web surfer model used in internet search engines.

This offers a simpler and more unified account of how people search their memory, postulating a single process rather than one process for exploring a cluster and one process for switching between clusters. Psychological studies have revealed clear regularities in how people search their memory, with clusters of semantically related items tending to be retrieved together. These conclusions help us build memory search process for our model and we know our memory search is local, relating similar concepts together. In our model, topics are modeled as a cluster of semantically related concepts tending to be retrieved together when people search their memory. Document is modeled as a random walk on human memory.

The topic generation model and document generation model are jointly dependent, which makes inference intractable. We propose an simple inference algorithm decouple these two processes and solve the following two inference problems. From document generation model, we infer the structure of associative network. From topic generation model, we discover the hidden modules within associative network.

### **5.3.2 Estimation of Associative Network**

We want to infer the structure of associative network based on all the document given in the corpus. For simplicity, we will represent our corpus specific associative networks using the word as node and the edge between two nodes reflect the connection between them, which is unlabeled. We build our inference algorithm based on maximum likelihood estimator (MLE) with regularization to infer the structure of associative network.

We build unigram model by picking single words from corpus as nodes, and represent the weight  $w_{ij}$  as the conditional probability of recalling word  $j$  given the presence of word  $i$ . The weight  $w_{ij}$  can be estimated from the following equation where  $c_{ij}$  is the co-occurrence count between word  $i$  and word  $j$ .

$$w_{ij} = \frac{c_{ij}}{\sum_{j \in I_i} c_{ij}} \quad (5.1)$$

Using MLE to estimate all weights between nodes in a large scale associative network might overfit the training data. There are  $O(N^2)$  parameters to estimate where  $N$  is the number of nodes in associative network. To avoid overfitting problem, we add certain regularization terms added into our model and the simple idea we use here is based on Pearson's chi-squared test.

Pearson's chi-squared test  $\chi^2$  is the best-known of many chi-squared tests, a statistical procedures whose results are evaluated by reference to the chi-squared distribution. It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The Pearson's chi-squared test  $\chi^2$  is given as follows:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (5.2)$$

### 5.3.2.1 Node Filtering

We want to extract the important words to represent each document so that the connections we consider between words are organic. Keyword extraction from a given document is an important research topic for various reason. Matsuo et al. [MI04] present a method on keyword extraction algorithm that applies to a single document without using a corpus. Matsuo et al. first extracted frequent terms, then a set of co-occurrences between each term and the frequent terms. Co-occurrence distribution shows importance of a term in the document as follows. If the probability distribution

of co-occurrence between term  $a$  and the frequent terms is biased to a particular subset of frequent terms, then term  $a$  is likely to be a keyword. The degree of bias of a distribution is measured by the  $\chi^2$  measure.

Similarly, we consider the significance of a word in a document given the corpus. The null hypothesis in our application is the probability of words belongs to topic follows binomial distribution with probability of  $p(w)$  where  $p(w)$  is the probability of  $w$  in the whole collection. Empirically we can estimate  $p(w)$  by maximum likelihood as follows:

$$p(w) = \frac{\text{Total occurrence of word } w}{\text{Total number of words in corpus}} \quad (5.3)$$

Under binomial case, we use the notation of binomial z score and then  $\chi^2$  scores becomes as follows. Note  $D(k)$  represents the number of words in document  $k$ .

$$\chi^2(w, k) = \frac{(n(w, k) - p(w)D(k))^2}{(1 - p(w))p(w)D(k)} \quad (5.4)$$

For a given document, we compute the binomial z score for each word in the document and extract the words whose z score are above certain threshold (as we will explore how this threshold affect structure of associative network later). We use these words to represent the document and compute the co-occurrence count between word  $i$  and word  $j$  in our corpus.

### 5.3.2.2 Edge Filtering

Similarly, we could find the co-occurrence count between word  $i$  and word  $j$  first and see if this co-occurrence is significant enough. According to the Pearson's chi-squared test  $\chi^2$ , the null hypothesis in this case is the probability of word  $i$  and probability of word  $j$  are independent. Assume word  $i$  and word  $j$  follows binomial distribution with probability of  $p(w_i)$  and  $p(w_j)$ , then word  $i$  and word  $j$  would occur with probability

of  $p(w_i, w_j) = p(w_i)p(w_j)$ , where  $p(w_i)$  and  $p(w_j)$  can be estimated from the whole corpus.

In this joint case,  $\chi^2$  scores becomes as follows. Note  $D(k)$  represents the number of words in document  $k$ .

$$\chi^2(w_i, w_j) = \frac{(n(w_i, w_j) - p(w_i, w_j)D(k))^2}{(1 - p(w_i, w_j))p(w_i, w_j)D(k)} \quad (5.5)$$

For a given document, we compute  $\chi^2$  score for each pair of words in the corpus and extract the pair whose z score is above certain threshold. We consider their co-occurrence as significant and use them to represent the co-occurrence count between word  $i$  and word  $j$ .

### 5.3.3 Discovery of Hidden Clusters

Based on the maximum likelihood estimation with certain regularization, we compute the co-occurrence between any pair of words. We estimate the structure of associative network based on Equation 5.1 and discovery the hidden clusters based on the estimated associative network.

As shown in our model, we model human knowledge as information flow on our associative network. From information theory perspective, succinctly describing information flow is a coding or compression problem. The key idea in coding theory is that a data stream can be compressed by a code that exploits regularities in the process that generates the stream. We use a random walk as a proxy for the information flow, because information flow on human memory can be represented as a random walk on semantic network.

Taking this approach, we develop an efficient code to describe a random walk on a network. We thereby show that finding topic clusters in associative networks is equivalent to solving a coding problem. If maximal compression were our only objective,

we could encode the path at or near the entropy 5.6 rate of the corresponding Markov process. Shannon [Sha01] showed that one can achieve this rate by assigning to each node a unique dictionary over the outgoing transitions.

$$H(X) = - \sum_{i=1}^n p_i \log(p_i) \quad (5.6)$$

But compression is not our only objective; here, we want our language to reflect the topic structure, we want the words we use to refer to things in the world. Shannon’s approach does not do this for us because every codeword would have a different meaning depending on where it is used. To solve this problem, Rosvall et al. [RB08] propose a two-level description of the network. They retain unique names for large-scale objects, the clusters or modules to be identified within our network, but we reuse the names associated with fine-grain details, the individual nodes within each module. This two-level description allows us to describe the path in fewer bits than we could do with a one-level description. The optimization goal of our inference is to minimize the description length of human memory, which can be formulated using the following equation:

$$L(M) = qH(Q) + \sum_{i=1}^m p_i H(P_i) \quad (5.7)$$

where  $q$  is per step probability of module switch and  $H(Q)$  is the entropy of movement between modules.  $p_i$  is per step probability of staying within the module  $i$  and  $H(P_i)$  is the entropy of movement within module  $i$ .

Applying inference algorithm in [RB08], we discover the hidden semantic clusters within associative network. To show how the binomial  $z$  score would affect our network structure. We explore the threshold of our binomial  $z$  score and discover the hidden semantic clusters for its network structure. Figure 5.5 shows as binomial  $z$  score cutoff increases, the proportion of largest community drops sharply. We use this piece of useful

information to discover the critical point for network structure change and discover separate clusters on the given threshold.

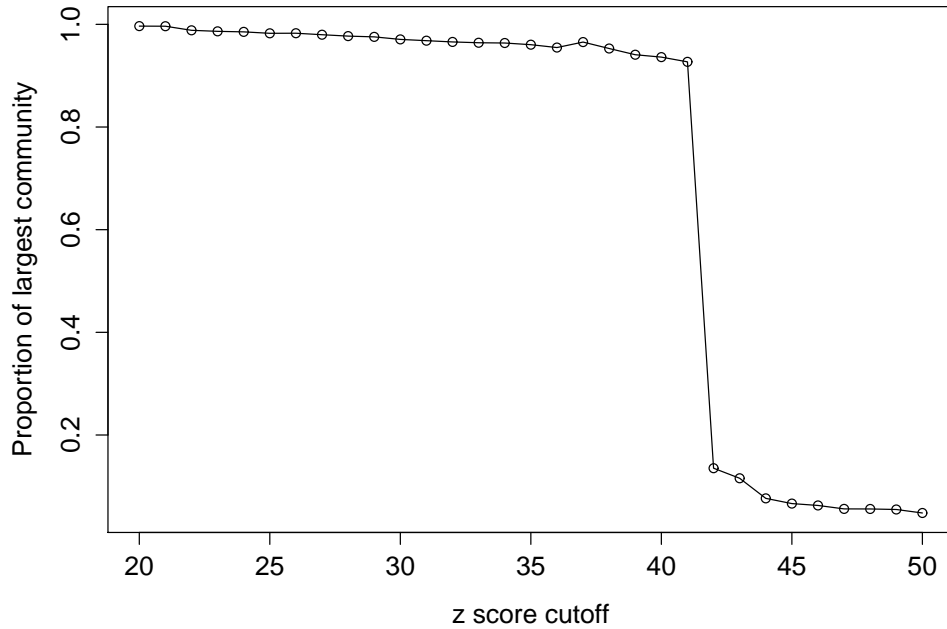


Figure 5.5: The binomial z score cutoff versus the proportion of largest community

We compute the structure of semantic network at the critical point and capitalize on the network’s structure that a random walker is statistically likely to spend long periods of time within certain clusters of words. Thus, we assign unique names to coarse-grain topic but reuse the names associated with fine-grain details.

### 5.3.4 Topic Representation

#### 5.3.4.1 PageRank for Human Memory

PageRank [PBM99] is proposed to rank the importance of webpages by considering the behavior of a random surfer visiting webpages on the network being considered. Suppose at step  $i$  the surfer is at webpage  $d_i$ . At each step, with probability  $\alpha$  the surfer picks (uniformly at random) one of the webpages  $d_i$  links to and goes there,



and with probability  $1 - \alpha$  he teleports to any one of the webpages in the network (again chosen uniformly at random.) The PageRank of a webpage  $d$  is then equal to the probability of the surfer visiting  $d$  at any given step in the long run. It can be understood as a Markov chain in which the states are webpages, and the transitions are the links between webpages. The PageRank vector  $r$  is then the stationary or limiting distribution of this Markov chain, which is the solution to the following equations:

$$r = Ar \tag{5.8}$$

where  $A = \alpha W + \frac{1-\alpha}{N} E$  where  $W$  is the adjacent matrix representation of network and  $E$  is the matrix with all 1s.

Griffiths et al. [GSF07] shows human memory and Internet search engines face a shared computational problem, needing to retrieve stored pieces of information in response to a query. They show that PageRank, computed on a semantic network constructed from word-association data, outperformed word frequency and the number of words for which a word is named as an associate as a predictor of the words that people produced in this task.

We compute the PageRank for each individual modules in associative network and use the PageRank vector as words distribution for the topics associated with this module. As we know, some words might have multiple meanings like the example of “bank” we have earlier, the hard clustering potentially cause big problem in case of words sense disambiguation. In order to determine words with multiple meanings, we applied the personalized PageRank algorithm to resolve this problem.

### 5.3.4.2 Personalized PageRank for Topic Overlaps

Jeh et al. [JW03] present new graph-theoretical results, and a new technique based on these results, that encode personalized views as partial vectors. In the random walk model, random walk return with the preference set  $P$  to a preference vector  $u$ , where

$|u| = 1$  and  $u(p)$  denotes the amount of preference for page  $p$ . They formalize personalized PageRank scoring using matrix-vector equations. Let  $A$  be the matrix corresponding to the web graph  $G$ . For a given  $u$ , the personalized PageRank equation can be written as follows:

$$r = (1 - \alpha)Ar + \alpha u \quad (5.9)$$

For each topic, we use the vector  $u$  according to PageRank of that module and compute the personalized PageRank as words distribution for the topics associated with this module. The issue of words sense disambiguation could be solved successfully using personalized PageRank.

## 5.4 Experiments and Results

To evaluate our topic assignment using associative networks, we applied our inference algorithms on four large scale datasets of different genre. Table 5.2 summarizes the size of document, the size of vocabulary, and number of tokens in each dataset. Cafemom [AHB12] is a online social forum where young mothers discuss different issues about parenting and their children and we consider a whole thread with all the posts as a document. The National Science Foundation datasets [BL13] are a collection of NSF funding abstracts, which represent the corpus of scientific abstract. The Reuters datasets [LYR04] are a collection of various news, from politics, science, finance to sports. Lastly, 2009 Iran post-election tweets [ZBK10] is a collection of short texts from online microblogging system Twitter.

Corpus	Cafemom	NSF	Reuters	Tweets
Document size	139,455	132,371	297,141	2,665,947
Vocabulary size	24,389	11,829	22,141	11,304
Token size	52,188,703	15,267,732	36,363,812	19,693,219

Table 5.2: Statistics of four experiment dataset

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
time	vaccine	food	business	church
people	vaccines	milk	team	jesus
child	health	eat	company	christian
kids	disease	water	income	bible
son	flu	foods	website	religion
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
owner	north	white	fever	breastfeeding
admin	south	black	cough	co-sleeping
vaccinate	california	racist	vomiting	natural
educate	city	color	nose	sahm
informed	texas	race	throat	diapering

Table 5.3: Top 10 topics discovered on Cafemom associative network

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
school	home	money	vaccines	autism
kids	free	people	vaccine	mom
teacher	business	pay	child	son
child	work	work	vax	group
year	team	job	shots	love
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
god	food	baby	time	health
people	eat	birth	son	body
religion	milk	women	day	cancer
life	make	pregnant	back	drugs
church	foods	hospital	things	drug

Table 5.4: Top 10 topics discovered on Cafemom using LDA

Table 5.3 and Table 5.4 compare the top 10 topics on Cafemom using our associative network approach and LDA. We can see LDA discovers these broad topics while our approach extracts fine-grained and specific topics. Overall we can see the topics extracted using network approach (in Table 5.3) are highly interpretable, and reveal the different meanings of a term in the corpus.

Table 5.5 show the top 10 topics discovered on the NSF associative network. As expected, we can see the top topics are about general research, biology, geology, math and so on. Table 5.6 show the top 10 topics discovered on the Reuters associative network. The topics are about finance, politics, sport etc. Table 5.7 show the top 10 topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
research	protein	ice	mantle	geometry
project	proteins	climate	seismic	groups
students	gene	ocean	crust	algebraic
university	genes	sea	rocks	spaces
science	cell	global	crustal	manifolds
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
reactions	laser	species	films	high
metal	optical	genetic	thin	temperature
organic	electron	populations	devices	low
complexes	spectroscopy	evolutionary	growth	energy
compounds	light	evolution	liquid	power

Table 5.5: Top 10 topics discovered on NSF associative network

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
percent	pct	police	world	party
million	uk	people	cup	opposition
market	dec	killed	league	parliament
year	nov	army	beat	house
bank	bln	forces	match	election
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
european	oil	court	coupon	israel
eu	gas	case	maturity	peace
commission	crude	trial	approx	israeli
union	fuel	charges	aaa	talks
states	gasoline	judge	date	palestinian

Table 5.6: Top 10 topics discovered on Reuters associative network

discovered on the IranElection tweets associative network. Interestingly, one top topic is about michael jackson, harry potter etc. We find death of michael jackson and release of harry potter film happened within the same time as IranElection. In our dataset, there are many tweets about michael jackson using meaningful hashtags from IranElection to attract attentions from users. However, our associative network approach does not get confused about the content and successfully detect this topic. Overall considering all four corpus, our approach extract topics that are highly interpretable, and the algorithm performs consistently in different genres of text.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
iran	green	location	embassy	sq
rt	support	gmt	injured	azadi
iranian	twitter	time	british	enghelab
people	democracy	change	accepting	square
tehran	add	zone	staff	st
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
gas	leader	michael	july	votes
water	khamenei	jackson	day	council
tear	supreme	mj	diplomats	guardian
eyes	ayatollah	potter	global	recount
helicopters	ali	harry	action	cities

Table 5.7: Top 10 topics discovered on IranElection associative network

## 5.5 Model Evaluation

Evaluation of topic model is becoming an open research topic as topic models have been shown successful in practice. Most earlier work is based on intrinsically evaluating learned topics on the basis of perplexity results. A model is learned on a collection of training documents, then the log probability of the unseen test documents is computed using that learned model. Usually perplexity is reported, which is the inverse of the geometric mean per-word likelihood. Wallach et al. [WMS09] presented efficient and unbiased methods for computing perplexity and evaluating almost any type of topic model.

While statistical evaluation of topic models is reasonably well understood, there has been much less work on evaluating the intrinsic semantic quality of topics learned by topic models, which could have a far greater impact on the overall value of topic modeling for end-user applications. Chang et al. [CGW09] presented the first human-evaluation of topic models by creating a task where humans were asked to identify which word in a list of  $v$  topic words had been randomly switched with a word from another topic. This work showed some possibly counter-intuitive results, where in some cases humans preferred models with higher perplexity. This type of result shows the need for further exploring measures other than perplexity for evaluating topic models.

Newman et al. [NLG10] introduces the novel task of topic coherence evaluation, whereby a set of words, as generated by a topic model, is rated for coherence or interpretability. Authors apply a range of topic scoring models to the evaluation task, drawing on WordNet, Wikipedia and the Google search engine, and existing research on lexical similarity/relatedness. In comparison with human scores for a set of learned topics over two distinct datasets, authors show a simple co-occurrence measure based on pointwise mutual information over Wikipedia data is able to achieve results for the task at or nearing the level of inter-annotator correlation, and that other Wikipedia-based lexical relatedness methods also achieve strong results.

In this section, we applied automatic evaluation techniques for topic clustering and document classification.

### **5.5.1 Automatic Evaluation for Topic Coherence**

For a given topic assignment, it is important to achieve words coherence within topics while keep how words coherence between topics. As Newman et al. [NLG10] suggests a simple co-occurrence measure based on pointwise mutual information over Wikipedia data is able to achieve results for the task at or nearing the level of inter-annotator correlation. We compute pointwise mutual information for any pairwise words in each topic and use the average mutual information to represent topic coherence for a given topic assignment. We compute the topic coherence score for all the topics in our result and 100 topics in LDA result. We show the average topic coherence score for these four corpus in Figure 5.6 and we can see our approach outperform LDA consistently. The automatic evaluation shows our approach shows topic with better interpretation than LDA.

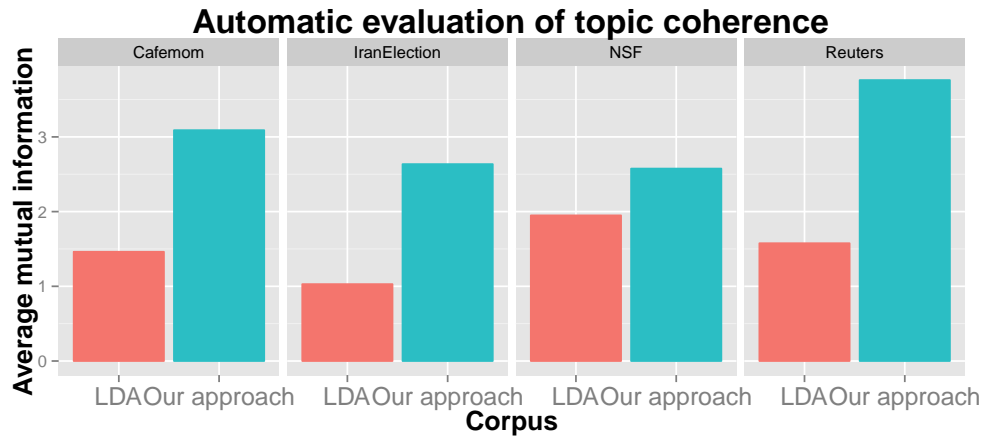


Figure 5.6: Automatic evaluation of topic coherence in four corpus

### 5.5.2 External Evaluation for Document Classification

Another area we want to address is how topic models could help with document classification. We use an external source of data with label topics and evaluate two approach by underlying ground truth. [JCB12] has a sample dataset containing labeled headlines from The New York Times, with total of 3104 documents. We applied both associative network approach and LDA on these documents and extract the most likely topics belong to each document. As the ground truth, we use the topic code associated with the dataset itself and generate the confusion matrix for given 27 topics in on the left side of Figure 5.7.

As we know, any unsupervised learning give a possible cluster assignment and to compare two different cluster assignment we need to consider all possible permutations of the cluster mapping. However it is finding 1-1 matching between two cluster assignment and maximum matching algorithm could be used in our case. We applied [HK73] algorithm and find the maximum matching between two clusters and plot the new confusion matrix on the right side of Figure 5.7.

first, we predict document class using the variation inference of LDA model and find the maximum matching between predict and given label is 541 out of 3104(17%). Then,

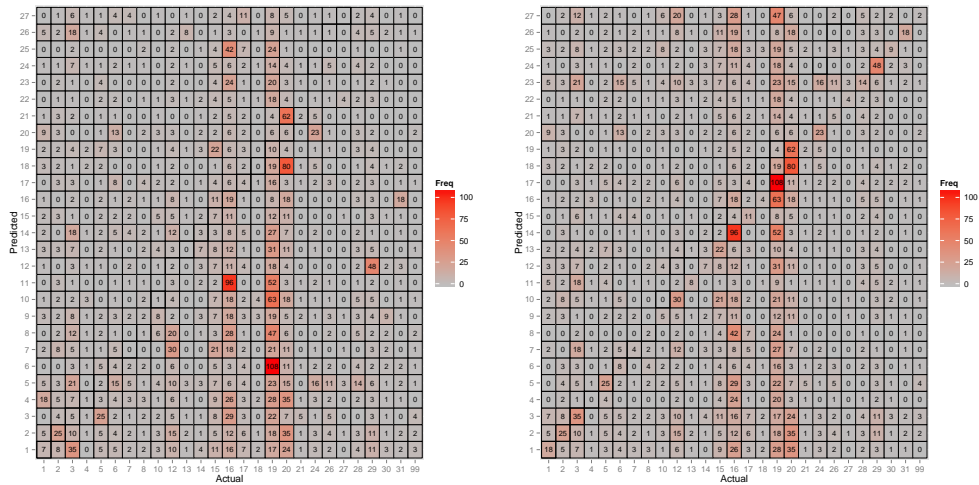


Figure 5.7: External evaluation for document classification

we predict document class using Gibbs sampling of LDA model and find the maximum matching between predict document class and given label is 684 out of 3104(22%). Lastly, we predict document class based on associative network approach, and find the maximum matching between predict document class and given label is 846 out of 3104(27.3%). From these statistics, we can conclude our network approach achieve better accuracy over LDA model on document classification based on the external labels.

## 5.6 Computational Cost

Very large data sets present major opportunities for machine learning, such as the ability to explore much richer and more expressive models, as well as providing new and interesting domains for the application of learning algorithms. However, the scale of these data sets also brings significant challenges for machine learning, particularly in terms of computation time and memory requirements. In this section, we consider the computational cost of two inference algorithms. We compare our network approach to the LDA model on four dataset and infer the hidden topic clusters on the same machine. Figure 5.8 shows the comparison of computational cost for four corpus and we can see



our network approach also outperforms LDA in terms of computational cost.

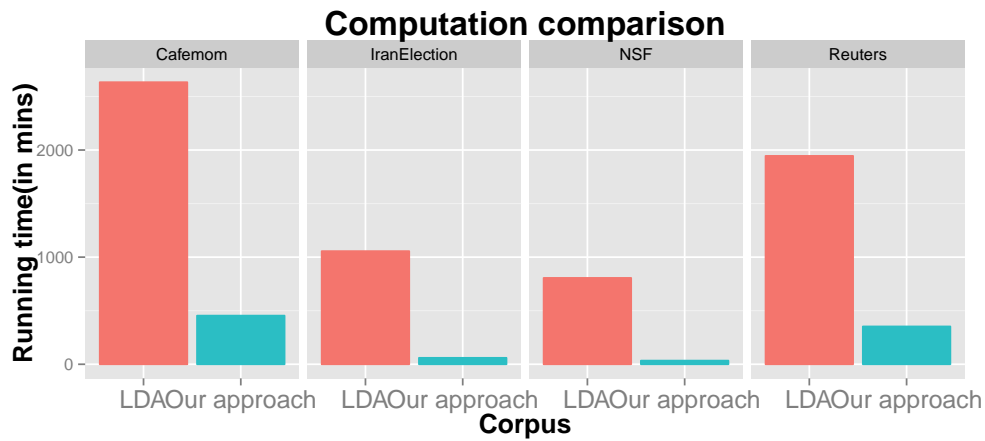


Figure 5.8: Comparison of computational cost for four corpus

## 5.7 Conclusion and Discussion

In this chapter, we develop a novel methodology for topic models, where given a large corpus of documents, it automatically infers the underlying topics and computes a distribution of documents over the computed topics. Our approach is very different from the highly popular and widely used existing topic models: Instead of using a bag of words model, it is inspired by how knowledge is organized in our brains as an associative network, and it exploits the idea of source coding from information theory to infer the latent networks directly from text data. We apply our algorithms on large-scale corpora, and using automatic evaluation techniques, show that our topic organization is not only more coherent semantically, compared to the state-of-the-art Latent Dirichlet Allocation (LDA) results, but is also computationally more efficient.

## CHAPTER 6

### Conclusion

The focus of this dissertation has been to develop such tools, and present empirical case studies that both establish the efficacy of the developed tools, and provide new insights into the data sets themselves.

In Chapter 2, we study the network dynamics in loosely connected social industry. Although the online social network have made us more densely networked than ever, researcher shows human have never been lonelier. Therefore understanding the network dynamics in a loosely connected social network becomes a very important topic. Towards this goal, we focus on collaboration network between actors based by analyzing a publicly accessible movie database and find global patterns in the underlying collaboration dynamics. We study the emergent patterns that exists in this collaboration network and developed models to explain these phenomenon. In particular, we present a microscopic analysis of the edge-by-edge evolution as well as node evolution for this large scale collaboration network. From empirical data, we show how such emergent patterns can be generated from stochastic decisions made at the level of the actors. These findings are vital to a range of important applications, from the development of better collaboration recommendation algorithms, to designing better systems for social forums that address different aspects for the online society.

In Chapter 3, we use specific events as a window to study information cascades. First of all, we find the *structure of cascades* tends to be wide, and shallow, with a central hub being more common. The overall cascade size distribution follows a power-law distribution with exponent equal  $-2.51$  and more than 98.7% of the cascades have

depth less than three. Due to broadcasting of tweets, cascades reach a lot of audience on the network we studied, even although user participation rate is not high. Then we study the *medium of cascades* on Twitter. We found that at most 63.7% of all retweets in Iranian election (78% for the other two topics) are from F-F network, thus the friendship network plays a major role. More than 34% of retweets for Iranian election (around 20% for the other topics) are from the public timeline, therefore public timeline offers other avenues for the spread of information outside the explicit friendship network. Last not the least, we study the *mechanism of cascade* by the underlying event-specific F-F network with a power law structure, and investigate its role in determining the cascade size distribution. We formulate the damped percolation model on event-specific F-F network to study information propagation and validate it through extensive simulation.

Understanding the principles of information propagation via F-F network as well as public timeline will be help design better application systems that address different aspects of the social media. Our findings about structure of information propagation has a significant impact on determining and managing Internet traffic, and hence Internet infrastructure backbone. Also we can take advantage of real-time trending topics on public timeline for viral marketing.

In Chapter 4, we address different aspects of topic models to extract the latent topics with the semantic meanings. We proposed two effective approaches to label the topics with keywords and our automatic labels give a better interpretation of baseline approach. Extraction of the temporal dynamics of a given topic allow users to trace information dynamics between online and offline social events. Furthermore, we show users have different preference towards different topics based on the online conversations they involved in. We use this piece of information to model users, which would allow us to build a better recommendation system for knowledge sharing and social assistance. We also extract opinions from user generated content in an automatic fashion and project sentiment on different topics for opinion mining at aspect level.

To study information dynamics between users, we first inferred user interaction

network from online conversation. We find the degree distribution of network follow power laws, which implies a few highly active users and a majority of less active ones. We further extract community structure from user interaction network and it reveals interesting properties of this site. We find there are a few users who have been active for a long period, the users fall into four temporally non-overlapping groups, and within each such group there is no significant clustering of users based on their post-response patterns. This lack of clustering and patterns other than on a temporal basis, makes it imperative that one goes beyond the simplistic measure of who-responded-to-whom relationships and look at the content of the posts as well.

To address this issue, we use the sentiment of posts to infer user interaction network. The a weighted and directed network is constructed as before, except that the edge is now assigned a weight of +1 if the sentiments of the two posts agree and -1 if they disagree. Thus, if sum of the weights of all the out-going edges from a node B is highly positive, then it implies that user B is agreed to by most of her responders. We find there are users that belong to all these four categories. Our result points to the fact that this forum is a fairly homogenous and consensus driven community (mostly populated by anti-vaccination oriented mothers). Content analysis helps one to identify several key features of the dynamics of the group.

In Chapter 5, we develop a novel methodology for Topic Models, where given a large corpus of documents, it automatically infers the underlying topics and computes a distribution of documents over the computed topics. Our approach is very different from the highly popular and widely used existing topic models: Instead of using a bag of words model, it is inspired by how knowledge is organized in our brains as an associative network, and it exploits the idea of source coding from information theory to infer the latent networks directly from text data. We apply our algorithms on large-scale corpuses, and using automatic evaluation techniques, show that our topic organization is not only more coherent semantically, compared to the state-of-the-art Latent Dirichlet Allocation (LDA) results, but is also computationally more efficient.

## REFERENCES

- [AA05] E. Adar and LA Adamic. “Tracking information epidemics in blogspace.” In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pp. 207–214, 2005.
- [AAG12] Joseph L Austerweil, Joshua T Abbott, and Thomas L Griffiths. “Human memory search as a random walk in a semantic network.” In *Advances in Neural Information Processing Systems*, pp. 3050–3058, 2012.
- [AB02] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks.” *Reviews of modern physics*, **74**(1):47, 2002.
- [AH10] S. Asur and B.A. Huberman. “Predicting the Future With Social Media.” *Arxiv preprint arXiv:1003.5699*, 2010.
- [AHB12] Ansuya Ahluwalia, Allen Huang, Roja Bandari, and Vwani Roychowdhury. “An automated multiscale map of conversations: mothers and matters.” In *Social Informatics*, pp. 15–28. Springer, 2012.
- [ALP01] Lada A Adamic, Rajan M Lukose, Amit R Puniyani, and Bernardo A Huberman. “Search in power-law networks.” *Physical review E*, **64**(4):046135, 2001.
- [And90] John R Anderson. *The adaptive character of thought*. Psychology Press, 1990.
- [ASB00] Luis A Nunes Amaral, Antonio Scala, Marc Barthélémy, and H Eugene Stanley. “Classes of small-world networks.” *Proceedings of the National Academy of Sciences*, **97**(21):11149–11152, 2000.
- [BA99] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks.” *science*, **286**(5439):509–512, 1999.
- [BB01] Ginestra Bianconi and A-L Barabási. “Competition and multiscaling in evolving networks.” *EPL (Europhysics Letters)*, **54**(4):436, 2001.
- [BB10] J. Beilin and M. etc Blake. “The Iranian Election on Twitter: The First Eighteen Days.” *Web Ecology Project*, 2010.
- [BDP07] John Blitzer, Mark Dredze, and Fernando Pereira. “Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification.” In *ACL*, volume 7, pp. 440–447, 2007.
- [BE01] Stefan Bornholdt and Holger Ebel. “World Wide Web scaling exponent from Simons 1955 model.” *Physical Review E*, **64**(3):035104, 2001.

- [BGJ10] David M Blei, Thomas L Griffiths, and Michael I Jordan. “The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies.” *Journal of the ACM (JACM)*, **57**(2):7, 2010.
- [BGL10] D. Boyd, S. Golder, and G. Lotan. “Tweet, tweet, retweet: Conversational aspects of retweeting on twitter.” In *hicss*, pp. 1–10, 2010.
- [BHK06] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. “Group formation in large social networks: membership, growth, and evolution.” In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 54. ACM, 2006.
- [BJ04] Wray Buntine and Aleks Jakulin. “Applying discrete PCA in data analysis.” In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 59–66. AUAI Press, 2004.
- [BJN02] Albert-László Barabási, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. “Evolution of the social network of scientific collaborations.” *Physica A: Statistical Mechanics and its Applications*, **311**(3):590–614, 2002.
- [BKA09] E. Bakshy, B. Karrer, and L.A. Adamic. “Social influence and the diffusion of user-created content.” In *Proceedings of the tenth ACM conference on Electronic commerce*, pp. 325–334. ACM, 2009.
- [BL13] K. Bache and M. Lichman. “UCI Machine Learning Repository.”, 2013.
- [Ble12] David M Blei. “Probabilistic topic models.” *Communications of the ACM*, **55**(4):77–84, 2012.
- [BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation.” *the Journal of machine Learning research*, **3**:993–1022, 2003.
- [BP98] S. Brin and L. Page. “The Anatomy of a Large-Scale Hypertextual Web Search Engine.” In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.
- [CGW09] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. “Reading tea leaves: How humans interpret topic models.” In *Advances in neural information processing systems*, pp. 288–296, 2009.
- [CHB10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. “Measuring User Influence in Twitter: The Million Follower Fallacy.” In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

- [CMD06] Hang Cui, Vibhu Mittal, and Mayur Datar. “Comparative experiments on sentiment classification for online product reviews.” In *AAAI*, volume 6, pp. 1265–1270, 2006.
- [CQ69] Allan M Collins and M Ross Quillian. “Retrieval time from semantic memory.” *Journal of verbal learning and verbal behavior*, **8**(2):240–247, 1969.
- [CSN09] Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. “Power-law distributions in empirical data.” *SIAM Review*, **51**:661–703, 2009.
- [CXL09] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. “A density-based method for adaptive LDA model selection.” *Neurocomputing*, **72**(7):1775–1781, 2009.
- [DDL90] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. “Indexing by latent semantic analysis.” *JASIS*, **41**(6):391–407, 1990.
- [DDP04] E. Damiani, S.D.C. Di Vimercati, S. Paraboschi, and P. Samarati. “An open digest-based technique for spam detection.” In *Proceedings of the 4th IEEE international conference on peer-to-peer computing*. Citeseer, 2004.
- [Dee66] James Deese. *The structure of associations in language and thought*. Johns Hopkins University Press, 1966.
- [DM00] Sergey N Dorogovtsev and José Fernando F Mendes. “Scaling behaviour of developing and decaying networks.” *EPL (Europhysics Letters)*, **52**(1):33, 2000.
- [DM02] Sergey N Dorogovtsev and Jose FF Mendes. “Evolution of networks.” *Advances in physics*, **51**(4):1079–1187, 2002.
- [DN09] Sajib Dasgupta and Vincent Ng. “Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification.” In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 701–709. Association for Computational Linguistics, 2009.
- [ER60] Paul Erdős and A Rényi. “On the evolution of random graphs.” *Publ. Math. Inst. Hungar. Acad. Sci.*, **5**:17–61, 1960.
- [ES06] Andrea Esuli and Fabrizio Sebastiani. “Sentiwordnet: A publicly available lexical resource for opinion mining.” In *Proceedings of LREC*, volume 6, pp. 417–422, 2006.

- [FFF99] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. “On power-law relationships of the internet topology.” In *ACM SIGCOMM Computer Communication Review*, volume 29, pp. 251–262. ACM, 1999.
- [FFV04] Abraham D Flaxman, Alan M Frieze, and Juan Vera. “A geometric preferential attachment model of networks.” In *Algorithms and Models for the Web-Graph*, pp. 44–55. Springer, 2004.
- [GGL04] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. “Information diffusion through blogspace.” In *Proceedings of the 13th international conference on World Wide Web*, pp. 491–501. ACM New York, NY, USA, 2004.
- [GN02] Michelle Girvan and Mark EJ Newman. “Community structure in social and biological networks.” *Proceedings of the National Academy of Sciences*, **99**(12):7821–7826, 2002.
- [Gro09] L. Grossman. “Iran protests: Twitter, the medium of the movement.” *Time Magazine*, 2009.
- [GS03] T Griffiths, Mark Steyvers, et al. “Prediction and semantic association.” *Advances in neural information processing systems*, pp. 11–18, 2003.
- [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.
- [GSF07] Thomas L Griffiths, Mark Steyvers, and Alana Firl. “Google and the mind predicting fluency with pagerank.” *Psychological Science*, **18**(12):1069–1076, 2007.
- [HD10] Liangjie Hong and Brian D Davison. “Empirical study of topic modeling in twitter.” In *Proceedings of the First Workshop on Social Media Analytics*, pp. 80–88. ACM, 2010.
- [HJT12] Thomas T Hills, Michael N Jones, and Peter M Todd. “Optimal foraging in semantic memory.” *Psychological review*, **119**(2):431, 2012.
- [HK73] John E Hopcroft and Richard M Karp. “An  $n^5/2$  algorithm for maximum matchings in bipartite graphs.” *SIAM Journal on computing*, **2**(4):225–231, 1973.
- [HL04] Mingqing Hu and Bing Liu. “Mining and summarizing customer reviews.” In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177. ACM, 2004.
- [Hof99] Thomas Hofmann. “Probabilistic latent semantic indexing.” In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57. ACM, 1999.



- [Hof01] Thomas Hofmann. “Unsupervised learning by probabilistic latent semantic analysis.” *Machine learning*, **42**(1-2):177–196, 2001.
- [HP09] A.L. Hughes and L. Palen. “Twitter adoption and use in mass convergence and emergency events.” *International Journal of Emergency Management*, **6**(3):248–260, 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.
- [HW79] John A Hartigan and Manchek A Wong. “Algorithm AS 136: A k-means clustering algorithm.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1):100–108, 1979.
- [JCB12] Timothy P Jurka, Loren Collingwood, Amber Boydston, Emiliano Grossman, and Wouter van Atteveldt. “RTextTools: Automatic text classification via supervised learning.”, 2012.
- [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.
- [JW03] Glen Jeh and Jennifer Widom. “Scaling personalized web search.” In *Proceedings of the 12th international conference on World Wide Web*, pp. 271–279. ACM, 2003.
- [KGA08] B. Krishnamurthy, P. Gill, and M. Arlitt. “A few chirps about twitter.” In *Proceedings of the first workshop on Online social networks*, pp. 19–24. ACM, 2008.
- [KKR99] Jon M Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S Tomkins. “The web as a graph: Measurements, models, and methods.” In *Computing and combinatorics*, pp. 1–17. Springer, 1999.
- [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.
- [KLR55] E. Katz, P.F. Lazarsfeld, and E. Roper. *Personal influence: The part played by people in the flow of mass communications*. Free Press New York, 1955.
- [KNT06] R. Kumar, J. Novak, and A. Tomkins. “Structure and evolution of online social networks.” In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 617. ACM, 2006.

- [KSR08] Joseph S Kong, Nima Sarshar, and Vwani P Roychowdhury. “Experience versus talent shapes the structure of the Web.” *Proceedings of the National Academy of Sciences*, **105**(37):13724–13729, 2008.
- [LAH07] J. Leskovec, L.A. Adamic, and B.A. Huberman. “The dynamics of viral marketing.” *ACM Transactions on the Web (TWEB)*, **1**(1):5, 2007.
- [LB05] John D Lafferty and David M Blei. “Correlated topic models.” In *Advances in neural information processing systems*, pp. 147–154, 2005.
- [LBK08] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. “Microscopic evolution of social networks.” In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 462–470. ACM, 2008.
- [LD97] Thomas K Landauer and Susan T Dumais. “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.” *Psychological review*, **104**(2):211, 1997.
- [LFK09] Andrea Lancichinetti, Santo Fortunato, and János Kertész. “Detecting the overlapping and hierarchical community structure in complex networks.” *New Journal of Physics*, **11**(3):033015, 2009.
- [LFL98] Thomas K Landauer, Peter W Foltz, and Darrell Laham. “An introduction to latent semantic analysis.” *Discourse processes*, **25**(2-3):259–284, 1998.
- [LH08] Jure Leskovec and Eric Horvitz. “Planetary-scale views on a large instant-messaging network.” In *Proceedings of the 17th international conference on World Wide Web*, pp. 915–924. ACM, 2008.
- [LK08] D. Liben-Nowell and J. Kleinberg. “Tracing information flow on a global scale using Internet chain-letter data.” *Proceedings of the National Academy of Sciences*, **105**(12):4633, 2008.
- [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.
- [LM06] Wei Li and Andrew McCallum. “Pachinko allocation: DAG-structured mixture models of topic correlations.” In *Proceedings of the 23rd international conference on Machine learning*, pp. 577–584. ACM, 2006.
- [LMF07] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. “Cascading behavior in large blog graphs: Patterns and a model.” *SIAM International Conference on Data Mining (SDM)*, 2007.

- [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.
- [LYR04] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. “Rcv1: A new benchmark collection for text categorization research.” *The Journal of Machine Learning Research*, **5**:361–397, 2004.
- [MI04] Yutaka Matsuo and Mitsuru Ishizuka. “Keyword extraction from a single document using word co-occurrence statistical information.” *International Journal on Artificial Intelligence Tools*, **13**(01):157–169, 2004.
- [Mil67] Stanley Milgram. “The small world problem.” *Psychology today*, **2**(1):60–67, 1967.
- [MMG07] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. “Measurement and analysis of online social networks.” In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, p. 42. ACM, 2007.
- [MSZ07] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. “Automatic labeling of multinomial topic models.” In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 490–499. ACM, 2007.
- [NDA06] Vincent Ng, Sajib Dasgupta, and SM Arifin. “Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews.” In *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 611–618. Association for Computational Linguistics, 2006.
- [New02] M.E.J. Newman. “Spread of epidemic disease on networks.” *Physical Review E*, **66**(1):16128, 2002.
- [New06] Mark EJ Newman. “Modularity and community structure in networks.” *Proceedings of the National Academy of Sciences*, **103**(23):8577–8582, 2006.
- [NG04] Mark EJ Newman and Michelle Girvan. “Finding and evaluating community structure in networks.” *Physical review E*, **69**(2):026113, 2004.
- [NLG10] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. “Automatic evaluation of topic coherence.” In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108. Association for Computational Linguistics, 2010.

- [NMS04] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. “The University of South Florida free association, rhyme, and word fragment norms.” *Behavior Research Methods, Instruments, & Computers*, **36**(3):402–407, 2004.
- [NP03] MEJ Newman and J. Park. “Why social networks are different from other types of networks.” *Physical Review E*, **68**(3):36122, 2003.
- [OM09] T. O’Reilly and S. Milstein. *The Twitter Book*. O’Reilly Media, Inc., 2009.
- [PBM99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. “The PageRank citation ranking: bringing order to the web.” 1999.
- [PFL02] David M Pennock, Gary W Flake, Steve Lawrence, Eric J Glover, and C Lee Giles. “Winners don’t take all: Characterizing the competition for links on the web.” *Proceedings of the national academy of sciences*, **99**(8):5207–5211, 2002.
- [PHB08] A. Passant, T. Hastrup, U. Bojars, and J. Breslin. “Microblogging: A semantic and distributed approach.” In *Proceedings of the 4th Workshop on Scripting for the Semantic Web*. Citeseer, 2008.
- [PL04] Bo Pang and Lillian Lee. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.” In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271. Association for Computational Linguistics, 2004.
- [PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. “Thumbs up?: sentiment classification using machine learning techniques.” In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86. Association for Computational Linguistics, 2002.
- [PV01] Romualdo Pastor-Satorras and Alessandro Vespignani. “Epidemic Spreading in Scale-Free Networks.” *Phys. Rev. Lett.*, **86**(14):3200–3203, Apr 2001.
- [QLB11] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. “Opinion word expansion and target extraction through double propagation.” *Computational linguistics*, **37**(1):9–27, 2011.
- [RB08] Martin Rosvall and Carl T Bergstrom. “Maps of random walks on complex networks reveal community structure.” *Proceedings of the National Academy of Sciences*, **105**(4):1118–1123, 2008.
- [Rog95] E.M. Rogers. *Diffusion of innovations*. Free Pr, 1995.

- [RSM02] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. “Hierarchical organization of modularity in metabolic networks.” *science*, **297**(5586):1551–1555, 2002.
- [RW03] Ellen Riloff and Janyce Wiebe. “Learning extraction patterns for subjective expressions.” In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 105–112. Association for Computational Linguistics, 2003.
- [SBR04] Nima Sarshar, P Oscar Boykin, and Vwani P Roychowdhury. “Percolation search in power law networks: Making unstructured peer-to-peer networks scalable.” In *Peer-to-Peer Computing, 2004. Proceedings. Proceedings. Fourth International Conference on*, pp. 2–9. IEEE, 2004.
- [SG07] Mark Steyvers and Tom Griffiths. “Probabilistic topic models.” *Handbook of latent semantic analysis*, **427**(7):424–440, 2007.
- [Sha01] Claude Elwood Shannon. “A mathematical theory of communication.” *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**(1):3–55, 2001.
- [SR04] Nima Sarshar and Vwani Roychowdhury. “Scale-free and stable structures in complex ad hoc networks.” *Physical Review E*, **69**(2):026101, 2004.
- [SR07] M.V. Simkin and V.P. Roychowdhury. “A mathematical theory of citing.” *Journal of the American Society for Information Science and Technology*, **58**(11):1661–1673, 2007.
- [ST05] Mark Steyvers and Joshua B Tenenbaum. “The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth.” *Cognitive science*, **29**(1):41–78, 2005.
- [TJB06] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. “Hierarchical dirichlet processes.” *Journal of the american statistical association*, **101**(476), 2006.
- [Tur02] Peter D Turney. “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.” In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424. Association for Computational Linguistics, 2002.
- [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.
- [WBO99] Janyce M Wiebe, Rebecca F Bruce, and Thomas P O’Hara. “Development and use of a gold-standard data set for subjectivity classifications.” In

*Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 246–253. Association for Computational Linguistics, 1999.

- [WC06] Xing Wei and W Bruce Croft. “LDA-based document models for ad-hoc retrieval.” In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 178–185. ACM, 2006.
- [WH07] F. Wu and B.A. Huberman. “Novelty and collective attention.” *Proceedings of the National Academy of Sciences*, **104**(45):17599, 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.
- [WMS09] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. “Evaluation methods for topic models.” In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105–1112. ACM, 2009.
- [WS98] Duncan J Watts and Steven H Strogatz. “Collective dynamics of small-worldnetworks.” *nature*, **393**(6684):440–442, 1998.
- [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*, pp. 123–131. ACM, 2010.