INFORMATION DIFFUSION ON ONLINE SOCIAL NETWORKS

Lilian Weng

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements for the degree
Doctor of Philosophy
in the Center for Complex Networks and Systems Research,
School of Informatics and Computing,
Indiana University

April 2014

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Filippo Menczer, Ph.D.

Alessandro Flammini, Ph.D

Yong-Yeol Ahn, Ph.D

Steven Myers, Ph.D

April 4, 2014

Copyright © 2014
Lilian Weng

Lilian Weng

INFORMATION DIFFUSION ON ONLINE SOCIAL NETWORKS

Thanks to the advent of the Internet, we can track, model, and predict communication and information propagation. The thesis aims to provide insights into information diffusion on online social networks from three aspects: people who share information, features of spreading information, and the interplays between network structure and diffusion process.

The first part delves into the consequences of limited human attention. Using a model that simulates meme diffusion while each agent is equipped with a finite attention span, we demonstrate that limited attention can intensify the competition, leading to heterogeneous dynamics of meme popularity. Besides, we find that people allocate their attention differently among strong and weak ties, driven by two competing tendencies for information gathering and social force.

The second part investigates properties of transmissible content, particularly into the topic space. We propose a measure of topical diversity and scrutinize its role in content and user popularity. High topical diversity of early adopters or early co-occurring tags is shown to imply high future virality of content, but low topical diversity helps an individual accumulate social impact. We also study how people communicate with strong and weak ties differently in terms of conversation topics.

Finally we present studies of how network structure, particularly community structure, influences the propagation of Internet memes and how the information flow in turn affects social link formation. We find that network communities trap information flows in general, but not viral memes. By characterizing the early spreading patterns of memes in terms of community concentration, we show that future meme virality can be predicted. We then examine traffic-driven shortcuts in social link formation and compare it with other strategies using Maximum-Likelihood Estimation. Triadic closure has a strong effect, but traffic-based shortcuts are another indispensable factor in interpreting network evolution.

We believe that the presented work can contribute to a better and more comprehensive understanding of information diffusion among online social-technical systems and yield advantages to viral marketing, advertisement, and social media analytics. Many complex dynamics of human society could be better unraveled by continued investigation of network structure.

Acknowledgements

First I would like to thank my advisor Professor Filippo Menczer. He has given me so much advise, encouragement, and endless support during the five-year journey of the PhD study at Indiana University. I appreciate all the time he has spent on editing my papers, discussing my research ideas, listening to my problems, and talking away all my doubts and worries. Without his patience, prudence, and wisdom, I would not be able to finish and eventually get my PhD degree.

Second, I would like to give my sincere thanks to my research committee, Professors Alessandro Flammini, Yong-Yeol Ahn, and Steven Myers. I have acquired a rich set of skills and experiences from their classrooms, knowledge, and meticulous attitude to research during my PhD expedition. Every piece of helpful suggestion makes me better and makes me enjoy research more.

Without my collaborators, I could not have completed my research projects. I give my thanks to Professors Alessandro Flammini and Alessandro Vespignani for working with me when I was still new to the field on my very first project related to the topic of information diffusion. The experience greatly inspired and motivated me, encouraging me to advance through this research direction until now. The work on Yahoo! Meme could not have achieved such a nice ending without the long-lasting efforts of Nicola Perra, Jacob Ratkiewicz, and Bruno Gonçalves. For a recent work on attention on weak ties, I would like to thank Nicola Perra and Márton Karsai for patient and thoughtful discussion during every Skype meeting. I'm particularly grateful for working with Professor Yong-Yeol Ahn on predicting meme virality. His working style and sparkling ideas are always inspiring and informative. Finally I want to thank Rossano Schifanella, with whom I've worked on a series of papers on human computation and social game design at the early time of my PhD; though the work is out of the scope of the thesis, the collaboration experience was very much appreciated.

Thanks to everyone working on the Truthy project (truthy.indiana.edu), I was able to easily access a large collection of Twitter data to facilitate my research. Besides, I want to thank Carlos Castillo and Francesco Bonchi from Yahoo! Research Barcelona for their support on providing a nice dataset from Yahoo! Meme. I also value my experience at Facebook in the Data Science team. In addition to their awesome data sources and computational frameworks, it was enjoyable to work with my mentor and collaborator there, Thomas Lento. I also thank Moira Burke for her contribution to and helpful suggestions on the project at Facebook.

The tech support team at the School of Informatics and Computing, and especially Bruce Shei and Rob Henderson, have provided tremendous help and made all our data-driven projects easier than ever.

All my dear friends and colleagues have offered me a great amount of joy, happiness, and courage along the way. I enjoyed lunch talks, tea time chats, weekly NaN meetings, so many shots of espresso, and foosball tournaments (though the games made me very nervous every time) with Jasleen Kaur, Emilio Ferrara, Onur Varol, Azadeh Nematzadeh, Giovanni Luca Ciampaglia, and other members of the NaN (Networks & agents Network) group. I should not forget to mention my perfect roommate Chaoqun Ni. I would not have been able to survive without her professional cooking skills. Chatting with her about research projects of our own is always interesting and encouraging for both of us as PhD students.

Finally I'm thankful for my beloved parents. I feel sorry that I cannot go back home often to stay around and take care of them. Their endless love and care from distant China reminds me all the time that someone at home is missing me and trusting me. I wish them to remain forever young and healthy.

*This thesis was supported in part by NSF grants IIS-0811994 and CCF-1101743, the project "Information Dynamics in Complex Data Structures" (PRIN), the Lilly Endowment (Data to Insight Center grant), Facebook, the James S. McDonnell Foundation, DARPA Grant W911NF-12-1-0037, and the School of Informatics and Computing at Indiana University, Bloomington.

The more comfortable we become with being stupid, the deeper we will wade into the unknown and the more likely we are to make big discoveries.

— Martin A. Schwartz, The importance of stupidity in scientific research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

With the advent of the Internet, mobile platforms, online commerce, and social media services, the footprints of human behavior are easily recorded in the digital world, generating data on an extremely large scale [Watts, 2007, Vespignani, 2009, Lazer et al., 2009]. *Big data* documents online discussion in which we participate, the people with whom we interact, how we buy and download a virtual product, and many other aspects of daily routines. The era of big data has brought in many opportunities as well as challenges for researchers by providing details of our lives in many perspectives: from characterization of human genome sequences to message exchanges between online users, and from mobilization of large population recovered through traffic data to easy tracking of information broadcasting. We can observe not only individual behavioral patterns, but also interaction and communication among mass by investigating these digital records. Questions about intricate and complex collective behavior of humans and resulting phenomena, which were hardly captured in self-reported surveys and interviews, can be better answered through the exploration of big data.

Our ability to store, track, analyze, and understand massive amounts of data is changing computer science, business, biology, technology, and many other fields. It demands cloud computing frameworks and applications [Dean and Ghemawat, 2008, Ghemawat et al., 2003, Chang et al., 2008] for handling large datasets, ranging from a few dozen terabytes to many petabytes, as these datasets cannot be processed on a single machine within a tolerable elapsed time. On the other hand, a deeper understanding of data on human behavior and interaction requires an interdisciplinary viewpoint, combining theorems and techniques from computer science, social science, physics, biology, cognitive science, and other related disciplines. In a way, the availability of large-scale data prompts new research questions, new computational frameworks, interdisciplinary approaches, and great opportunities to quantitatively explore many fundamental questions. Watts [2007] have predicted that "if handled appropriately, data about Internet-based communication and interactivity could revolutionize our understanding of collective human behavior." Also, as presented in

The Petabyte Age[1], one issue of the Wired magazine in 2008, "*In the era of big data, more isn't just more. More is different*."

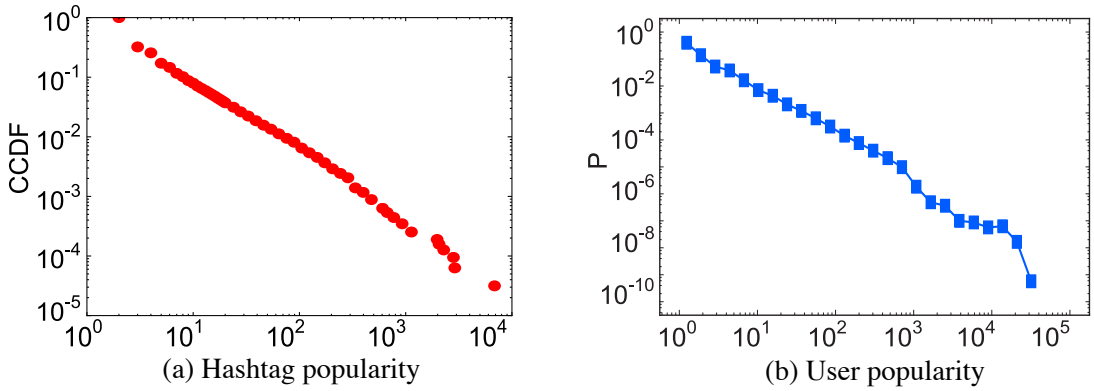
The concept, *social computing*, can be broadly defined as computational facilitation of social studies and human social dynamics, as well as the design and use of information and communication technologies that consider social context [Surowiecki, 2005]. With the deluge of large-scale digitalized data and the establishment of cloud computation infrastructure, we shift attention from the underlying computational facilities to deep mining of the data, nurturing the emergence of a new interdisciplinary field of research—*computational social science*. It leverages the capacity to collect and analyze data on a large scale in order to reveal patterns of individual and group behaviors [Lazer et al., 2009]. Although there still exist several substantial barriers, such as privacy issues and incomplete relevant policies to regulate personal privacy in the digital world, the field of computational social science is booming in leading Internet companies like Google and Facebook, research institutes in academia, and government agencies [Lazer et al., 2009, Mason et al., 2013].

Complex systems are, in particular, a powerful approach to explore and understand complex collective human behavior. We can summarize the strategy of individual behavior to interpret the observation about a single participant. However, in a complex and evolving system with numerous actors, it is hard to inspect the behavior of every single actor. Instead, we measure and model the system by setting up a common rules for all the participants and deduce the collective behavior [Bonabeau, 2002, Weng and Menczer, 2013].

As the cost of communication and information propagation for ordinary users has been considerably lowered by the information technologies, an increasing portion of human activities moves to the online world. How people are sharing information and what kind of information can widely spread become trackable. The major theme of the thesis is to ***quantitatively study how people spread information on the online social networks***. Can we model the way that information is propagated? Can we tell which messages will be more popular than others? Which individual features affect the diffusion process? What do people talk about? Can we summarize the topics of online conversations? We would like to harness the power of big data to answer these questions.

However, the ease of online communication creates side effects as well. According to the *Dunbar's number* [Dunbar, 1998, Huberman et al., 2009, Gonçalves et al., 2011], there is a cognitive limit on the number of stable social relationships that one can sustain and the amount of information an individual can produce and process is finite. Hence, the abundance of information to which we are exposed through various socio-technical systems is exceeding our capacity to consume it. As predicted in the seminal paper by Simon [1971], attention scarcity is a major problem in an information-rich world, and attention has been described as the "new currency of business" [Davenport and Beck, 2001]. Information requires enough amount of attention to survive and furthermore spread widely to reach numerous viewers. Hence, the scarcity of attention results in fierce competition among a large quantity of online content. For example, in Twitter, when the popularity of a hashtag is quantified by the number of daily retweets containing that tag, the distribution of hashtag

[1]http://www.wired.com/science/discoveries/magazine/16-07/pb_intro



(a) Hashtag popularity (b) User popularity

Figure 1.1: The distributions of (a) Twitter hashtag popularity measured as the number of daily retweets, and (b) user popularity in Yahoo! Meme measured as the number of subscribers in the system.

popularity follows a power-law (see Fig. 1.1a). It implies that a large majority of online content cannot grow popular, and only a few pieces may go extremely viral. Meanwhile, an individual can pay attention to other people, triggering formation of new social links. We consider the number of subscribers or followers as a measure of how much attention a user can attract from others and observe a power-law distribution of user popularity. A similar message can be drawn that most people are not popular, but a very small proportion of them may obtain extremely high volume of attention, such as celebrities (see Fig. 1.1b). These observations raise several outstanding questions: Can we predict which pieces of information will go viral in the future? What factors decide to whom an individual pay attention? The path towards the solutions involve a variety of factors: individual attributes, the content of transmissible messages, social network topology, and the mechanism of diffusion process. The thesis aims at providing a systematic description of *information diffusion on online social networks* from all these perspectives. The presented studies are expected to have great implications in many fields, including viral marketing, grassroot movement, online advertisement, and social media analytics.

1.2 Research Questions and Thesis Overview

The studies of information diffusion on online social networks incorporate four critical components: *actors*, *content*, *underlying network structure*, and *diffusion process*, as illustrated in Fig. 1.2; studies on network topology and diffusion mechanisms are combined in the third part, because we focus on the interplays between them. Social media users are connected through predefined online interactions such as Facebook friendship or Twitter following, forming a network. People in such a network can share messages according to their interests with connected neighbors, propagating information through social links and creating cascades.

In Part I, we investigate the role of limited user attention in the diffusion of online memes,

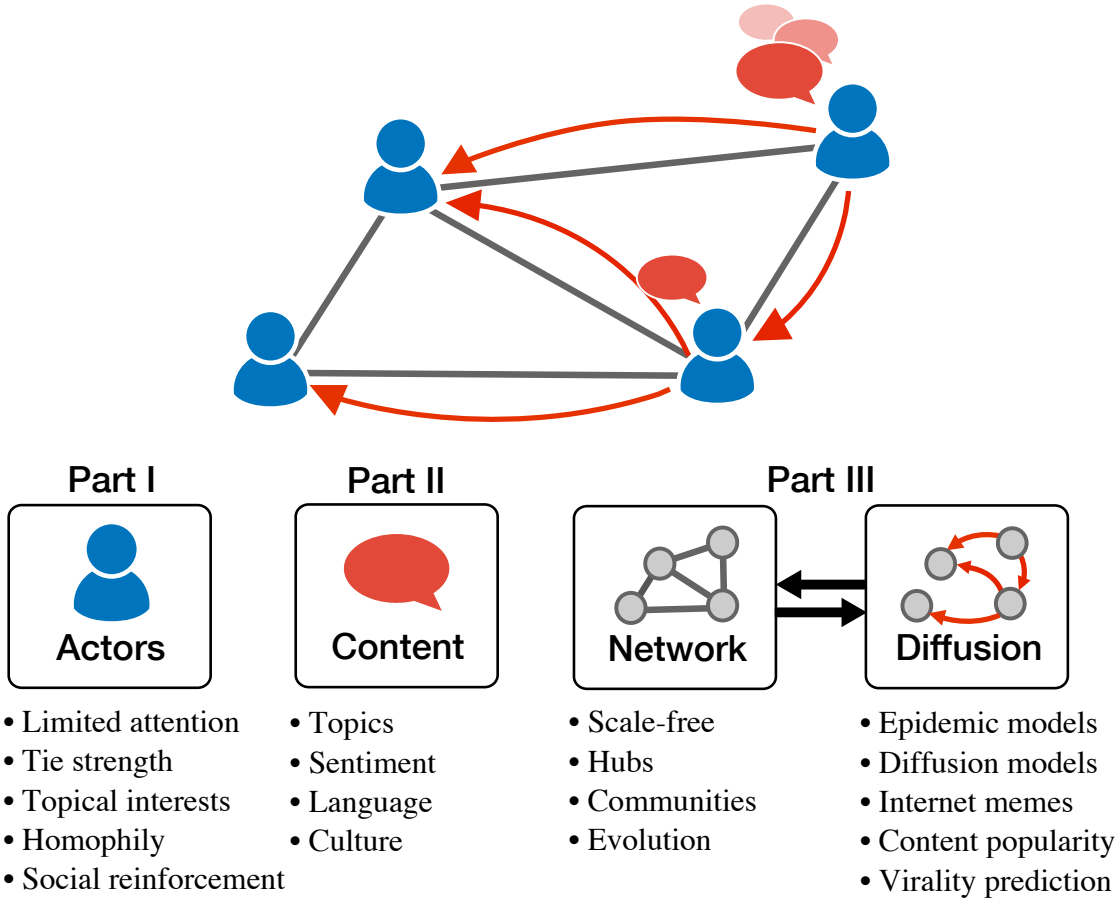


Figure 1.2: The study of information diffusion on social networks incorporates four important components: actors, content, underlying network structure, and the diffusion process.

particularly how it drives the fierce competition among memes and leads to heterogeneous distributions of meme popularities and life spans. Then we investigate the difference between attention allocated on strong and weak ties, and find that the observed difference can be interpreted by distinguishing social-oriented links from information-oriented ones. Part II focuses on the topic space extracted from online conversation. The diversity of topics assigned to an actor or a message is shown to be a nice predictor of the future popularity. While examining conversation topics in the context of local neighborhood, we study how people communicate differently through weak and strong ties in terms of topic diversity and popularity. Finally, the studies on the interplays between network structure and diffusion process are presented in Part III. It explores how the network topology affects the spread of information and how the traffic flow, in turn, shapes the network evolution.

Each part contains two chapters and each chapter follows the logical flow from observation to model and from model to application. As shown in Table 1.1, all the chapters follow the same logic, but do not necessarily finish the whole cycle: Chapter 4 and 5 present the empirical findings and the models; Chapter 7 only reports empirical observations. Next

Table 1.1: Summary of three parts and six main chapters in the thesis.

Part	Name	Observation	Model	Application
I	**Actors**	S4.1-S4.2	S4.3	-
		S5.1-S5.3	S5.4	-
II	**Content**	S6.1	S6.2	S6.3
		S7.1-S7.4	-	-
III	**Diffusion on Network**	S8.2	S8.2	S8.3
		S9.1	S9.2	S9.3

section introduces the research questions examined in each part with more details.

1.2.1 Part I Actors and Limited Attention

Starting from the prediction on the poverty of attention by Simon [1971]:

- **How do memes compete for limited human attention?**
- **How does limited attention affect information diffusion?**

The work on the role of limited attention in the dynamics of meme popularity is motivated by the observation that the wide adoption of online social media has increased the competition among ideas. Using an agent-based model, we demonstrate that finite individual attention is a key ingredient to explain the heterogeneity of global patterns such as meme popularity, meme lifetime, and user activities. We find that the combination of such competition and social network topology alone can explain the observed global patterns without even assuming the different intrinsic values among ideas [Weng et al., 2011, 2012].

Then we have questions on the allocation of limited attention:

- **Can we measure attention?**
- **Do people assign attention differently on weak and strong ties?**

Our interest in how people allocate attention among weak and strong ties is driven by studies on limited attention and the weak tie hypothesis. Granovetter [1973] hypothesized that weak ties do not carry as much communication as strong ties, but they act as bridges between distant groups and propagate innovative information. We propose a measure of attention with the assumption that each individual has a fixed amount of attention. We find that only very strong and very weak ties attract much attention; the former tend to be close connections built up for maintaining social relationships, while the latter are constructed mainly for receiving information [Weng et al., 2014a].

1.2.2 Part II Content and Topic Space

We look into questions on the content of spreading content:

- **Can we detect topics in social media?**
- **How does topical diversity affect user and content popularity?**

Information flows can be mapped into the topic space where each meme is represented as a node and clusters of memes form topics. We can thus learn topics by detecting communities in the meme co-occurrence network. According to the association between the social network and the topic space, we assign each user or message with a variety of topic and quantify topical diversity. We find that high topic diversity is a good predictor of popularity of messages, while low diversity benefits people for accumulating social influence [Weng and Menczer, 2014].

From the individual viewpoint, the question on the conversation topics is:

- **Do people talk through weak and strong ties differently in terms of topics?**

Many types of friends read and respond to our posts on social media. Close friends may be more likely to respond to posts than acquaintances, but how much does the post topic matter? With an analysis of a sample set of Facebook users, we show that strong ties reply to a wider variety of topics than weak ties, even after controlling for the number of posts a tie responds to. Popular topics, such as memes and holiday greetings, are equally popular among strong and weak ties [Weng et al., 2014b]. The results validate and expand previous theoretical and survey-based work, demonstrating that strong ties are characterized by a diversity of interests.

1.2.3 Part III Diffusion on Social Networks

The first section studies the role of social network structure on the spread of information with questions:

- **How does network structure affect the spread of information?**
- **Can we predict the future meme virality based on its early spreading pattern?**

In this line of research, we hypothesize that a community-centric viewpoint can provide unique insights to information diffusion, specifically to the questions of how memes spread and which memes will go viral. Communities, clusters of densely connected nodes, are expected to trap information flow due to a combined effect of structural trapping, social reinforcement [Centola, 2010], and homophily [McPherson et al., 2001]. We find that a viral meme is less likely to be trapped from the early stage, spreading like infectious diseases. By quantifying how early adopters of a meme concentrate among communities, we demonstrate that it is possible to predict the future virality of the meme [Weng et al., 2013a, 2014c].

The second section asks the following questions to explore the role of information flows in shaping the network topology:

- **How does information diffusion affect network evolution?**

- **What are individual link creation strategies in accord with information flows?**

A social network in reality is evolving dynamically and continuously. Such evolution is coupled with the spread of information on top of the network: the network topology affects the channels of information diffusion; the birth and death of connections in the network is, in turn, triggered by the traffic. We consider traffic-based shortcuts—the formation of social links according to reposted or seen traffic—as a key link formation strategy, and then characterize the network growth with several other strategies using Maximum-Likelihood Estimation. Our findings indicate that triadic closure strongly affects link formation, while shortcut based on information flows is another indispensable factor in interpreting network evolution. Yet, we also show that individual strategies for following other users are highly heterogeneous. Their link creation behaviors can be summarized by classifying users into different categories with distinct structural and behavioral characteristics [Weng et al., 2013b].

Chapter 2

Concepts and Methods

2.1 Datasets

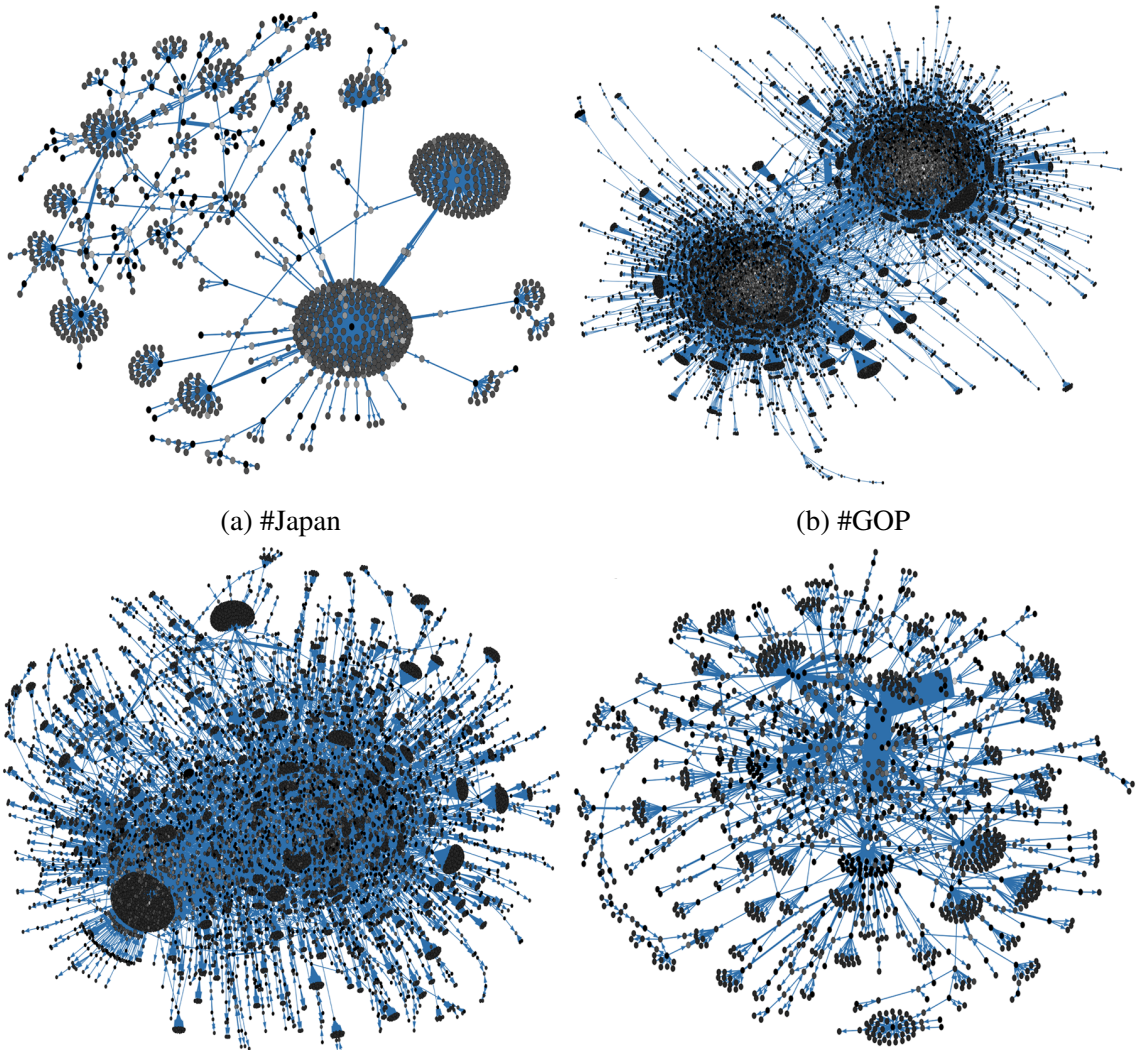
The thesis presents studies using datasets from Twitter, Yahoo! Meme, mobile phone call network, Enron email collection, and Facebook. This section introduces these platforms and how the data is collected from each platform.

2.1.1 Twitter

Twitter[1] is a micro-blogging system that allows many millions of people to broadcast short messages through social connections. Users can subscribe to (or "follow") interesting people, by which a directed social network is formed. Short messages (or "Tweets") that are limited to contain no more than 140 characters appear on the screen of followers. People can forward (or "retweet") selected posts on the screen to their followers. They are also allowed to "mention" the screen name of another user in tweets by using the "@" symbol (e.g. @obama). Furthermore, users often mark their posts with topic labels (or "hashtags")—explicit topical tags, words, or phrases following a hash symbol (e.g. #oscar2014). We consider hashtags as operational proxies to identify memes. A retweet carries a meme from user to user. As a meme spreads in this way, it forms a cascade or diffusion network such as those illustrated in Fig. 2.1. The data of tweets is gathered through Twitter streaming API[2] which provides about 10% random sample of public tweets in real time. The information about following connections is collected using Twitter GET followers API[3].

Chapters 4, 5, 6 and 8 analyze data from Twitter. Social networks used in these chapters are built on either directed following or reciprocal following relationships (see Fig. 2.2a). Edges are directed to account for asymmetric relations between users; a node can follow another without being followed back. Reciprocal following relationships are considered,

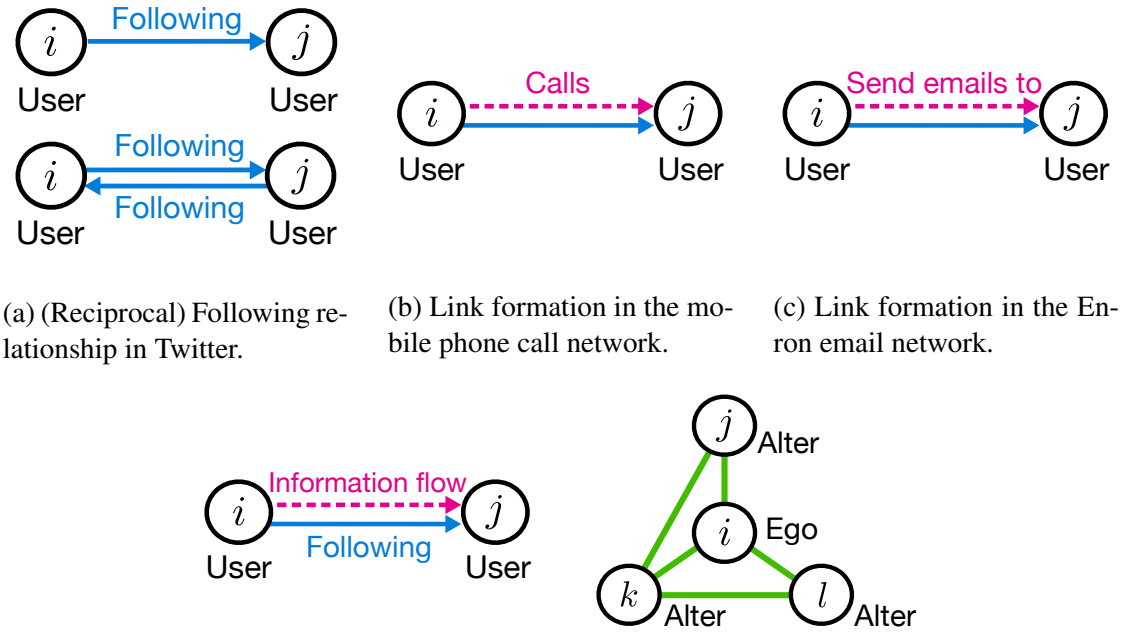
[1] http://www.twitter.com
[2] http://dev.twitter.com/docs/streaming-apis
[3] https://dev.twitter.com/docs/api/1/get/followers/ids



(a) #Japan (b) #GOP

(c) #Egypt (d) #Syria

Figure 2.1: Visualizations of meme diffusion networks for different topics. Nodes represent Twitter users, and directed edges represent retweeted posts that carry the meme. The brightness of a node indicates the activity (number of retweets) of a user, and the weight of an edge reflects the number of retweets between two users. (a) The #Japan meme shows how news about the March 2011 earthquake propagated. (b) The #GOP tag stands for the US Republican Party and as many political memes, displays a strong polarization between people with opposing views. Memes related to the "Arab Spring" and in particular the 2011 uprisings in (c) #Egypt and (d) #Syria display characteristic hub users and strong connections, respectively.



(a) (Reciprocal) Following relationship in Twitter.

(b) Link formation in the mobile phone call network.

(c) Link formation in the Enron email network.

(d) Information flow and following link in Yahoo! Meme.

(e) Ego network centered at an individual Facebook user.

Figure 2.2: Illustrations of various social connections on (a) Twitter, (b) Yahoo! Meme, and (c) Facebook.

as bi-directional links reflect more stable and reliable social connections. In the tasks of predicting future popularity of hashtags, we only include memes that emerge during our data collection time window to ensure that we examine only 'new' memes. *Emergent memes* are defined as hashtags that are used during the first week of the data window and have fewer than a threshold number of tweets during the previous month.

2.1.2 Mobile Phone Call Network

The mobile phone call dataset used in Chapter 5 was recorded by a single operator with 20% market share in an undisclosed European country (ethic statement was issued by the Northeastern University Institutional Review Board). This dataset records about 487 millions directed call events during 120 days with one second resolution. The social network is recovered in a way that a directed edge $(i, j) \in E$ is added for each call from the caller i to the callee j, as $i, j \in V$ (see Fig. 2.2b).

2.1.3 Enron Email Collection

The Enron email corpus[4] used in Chapter 5 was made publicly available during the legal investigation concerning the Enron corporation [Klimt and Yang, 2004]. This email corpus records 246,391 emails exchanged inside the Enron corporation. While recovering the email network, an edge $(i, j) \in E$ is established if there is at least one email observed from the user i to the user j, $i, j \in V$ (see Fig. 2.2c).

2.1.4 Yahoo! Meme

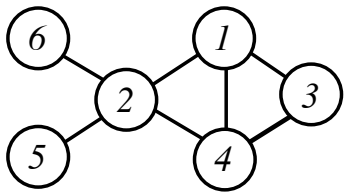
Yahoo! Meme is a social micro-blogging system with functionality similar to Twitter and Tumblr, which was active between 2009 and 2012.[5] We have access to the entire history of the system from April 2009 until March 2010, including full records of every message propagation and link creation event. This dataset is used for studying how information diffusion affects network evolution in Chapter 9.

In the Yahoo! Meme dataset, a user j following a user i is represented in the follower network by a directed edge $\ell = (i, j)$, indicating j can receive messages posted by i. We adopt this notation, as illustrated in Fig. 2.2d, in which the link creator is the target, to emphasize the direction of information flow. In our notation, the in-degree of a node i is the number of people followed by i, and the out-degree is the number of i's followers. When user j reposts content from i, we infer a flow of information from i to j. Each link is weighted by the numbers of messages from i that are reposted or seen by j.

2.1.5 Facebook

Facebook[6] is an online social networking sites built on mostly mutual social relationships. There are several types of connections on Facebook.[7] In the micro-blogging applications like Twitter and Yahoo! Meme, people can follow essentially any body, resulting in directed social links. Different from these platforms, to build up a "friendship" in Facebook, an individual has to get approved by the invite d other to confirm this mutual and undirected relationship. An unverified friendship invitation is invalid and cannot initiate any information sharing. Directed "follow" relationships also exist but not prevalent, mostly between users and public pages or celebrities. Facebook users can update statuses, upload photos, and share both internal and external information. Then the posted content appear in the newsfeeds of their friends. As a response, friends might "like" or "comment" on the content in their newsfeeds to reveal interests. Besides, users are allowed to "tag" others in statuses, photos, and link share posts. Finally, Facebook provides "lists" as a way to

[4] http://www.cs.cmu.edu/ ./enron/
[5] http://en.wikipedia.org/wiki/Yahoo!_Meme
[6] http://www.facebook.com
[7] http://www.facebook.com/help/366702950069221/



(a) Labelled graph

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(b) Adjacency matrix

Figure 2.3: Graph representations of an undirected and unweighted network in (a) a labelled graph and (b) an adjacency matrix.

organize one's social relationships. Users may create a close friend list, containing "best friends you want to see more of on Facebook".[8]

To test the relationship between tie strength and topic diversity in Chapter 7, we gather a dataset of sampled Facebook users and their ego networks. An ego network is centered at one individual containing all connected neighbors and social links among these neighbors, as illustrated in Fig. 2.2e.

2.2 Graph Theory

2.2.1 Graph Representation

The structure of a social network is commonly framed as a *graph*. A graph G consists of a set of vertices (or nodes) V and a set of edges (or links) $E = \{e_{ij} \mid i, j \in V\}$, and $G = (V, E)$. Each edge in E is represented as a pair of vertices, so we have $E \subseteq V \times V$. These edges may or may not have directions; given a undirected graphs, $e_{ij} \in E \iff e_{ji} \in E$, but it is not necessarily true for a directed graph. A graph can also be represented in an adjacency matrix A, in which each entry A_{ij} (row i, column j) labels the number of edges between nodes i and j. Figure 2.3 exemplifies two different representations of a simple, unweighted, and undirected graph.

2.2.2 Basic Concepts

Degree Given a node i, it is the number of edges connected to it, noted as k_i. In a directed graph, we can define in-degree k_i^{in} as the number of edges pointing to it and out-degree k_i^{out} as ones from it. We have $k_i = k_i^{\text{in}} + k_i^{\text{out}}$. In many real-world social networks, the degree distribution follows a power law (see more in Sec. 3.3.1).

[8] http://www.facebook.com/help/200538509990389

Weight A graph can have each edge e_{ij} associated with a number w_{ij}. The empirical meaning is determined by the properties of the network. For instance, the weight might refer to the cost traveling from one city to the other when we map a state as a graph; in the email network, the weight can label the number of emails that one user has sent out to the other.

Strength It is the sum of the weights of all edges adjacent to a given node i, labelled as s_i. Similar to how we define in-degree and out-degree, we can have in-strength s_i^{in} and out-strength s_i^{out} in a directed graph, and $s_i = s_i^{\text{in}} + s_i^{\text{out}}$.

Distance The distance $d(i, j)$ between two given nodes, i and j, is the shortest path between them. It is the minimum number of edges one needs to go through, if he travels from one node to the other.

Clustering coefficient Both the global and local clustering coefficients are defined in an undirected graph. A triplet consists of three nodes with either two (open triplet) or three (closed triplet) undirected edges in-between.The global clustering coefficient is defined for a graph as the proportion of existing closed triplets among all triplets [Wasserman, 1994].

$$C_G = \frac{|\{\langle i, j, k\rangle \mid e_{ij}, e_{jk}, e_{ik} \in E\}|}{|\{\langle i, j, k\rangle \mid e_{ij}, e_{jk} \in E\}|} \tag{2.1}$$

A triangle refers to a set of three nodes with three undirected edges among them. The local clustering coefficient is defined for a node i as the fraction of triangles among all the triples of nodes in i's neighborhood, while both the selected triangles and node triples should contain i.

$$C(i) = \frac{|\{e_{jk} \mid j, k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)} \tag{2.2}$$

where $N_i = \{j \mid e_{ij} \in E \vee e_{ji} \in E\}$ is the set of i's neighbors. The local clustering coefficient measures how close one's neighbors are to being a clique [Watts and Strogatz, 1998].

2.2.3 Community Detection

A *community* is a group of densely connected nodes in a graph. The community structure is claimed to be one key property of social networks, suggesting that a social network can be partitioned into several (potentially overlapped) clusters so that nodes in one cluster are densely connected internally but not externally; such clustering might derive from common interests, family memberships, or geographical divisions [Girvan and Newman, 2002, Newman and Park, 2003]. How to detect communities has been widely studied [Fortunato, 2010], popular methods including modularity optimization [Newman, 2006], Louvain method [Blondel et al., 2008], infomap [Rosvall and Bergstrom, 2008], clique percolation [Palla et al., 2005], and link clustering [Ahn et al., 2010]. The methods applied in the thesis are introduced as follows.

Lounvain method The method attempts to optimize the modularity of a partition of the network using a greedy optimization approach. Modularity is used to qualify how well a partition is by measuring the density of links inside communities as compared to links between communities [Newman, 2006].

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \text{ where } m = \frac{1}{2} \sum_i k_i \tag{2.3}$$

where c_i and c_j are communities of nodes i and j, respectively; $\delta(c_i, c_j)$ equals to 1 if $c_i = c_j$ otherwise 0. $\frac{k_i k_j}{2m}$ computes the expected number of edges between i and j, irrespective of the community structure. The values of Q ranges between -1 and 1. The optimization has two phases: first, it forms small communities to achieve local maxima of the modularity; second, a new network is constructed whose nodes are communities found in the first phase. These two phases are performed repeatedly until a maximum of modularity is attained, producing a hierarchy of communities [Blondel et al., 2008].

Infomap The infomap community detection method is built on the assumption that a random walker is more likely to be trapped in communities than to travel between communities. The path of a random walker can be encoded, and then compressed given a hierarchical network partition so that the encoded description is minimum. The duality between finding community structure in a network and the coding problem is: to find an efficient code, it looks for a module partition $\mathbf{M}$ of n nodes into m modules so as to minimize the expected description length of a random walk. By using the module partition $\mathbf{M}$, the average description length of a single step is given by

$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{L}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i) \tag{2.4}$$

where $H(\mathcal{L})$ is the entropy of module names in $\mathbf{M}$; $H(\mathcal{P}^i)$ is the entropy of intra-module movements; $q_{\curvearrowright}$ gives the probability that the random walk switches modules on a given step; $p^i_{\circlearrowright}$ is the sum of the probability of intra-module movements inside the module i and the probability of exiting i. The first part of the formula describes the entropy of the movement between communities, and the second part sums up the entropy within each community. Eventually infomap applies computational search algorithm to find the best partition as the outcome [Rosvall and Bergstrom, 2008].

Link clustering Different from previous two, the link clustering method aims at discovering overlapped communities in which a node is allowed to belong to multiple groups. The link clustering algorithm reinvents communities as groups of links rather than nodes. The set of neighbors of a node i is denoted as N_i. Given a pair of links with one shared node, e_{ij} and e_{jk}, the similarity between these two links is the Jaccard similarity between neighbor sets of distinct nodes:

$$S(e_{ij}, e_{jk}) = \frac{|N_i \cap N_k|}{|N_i \cup N_k|} \tag{2.5}$$

Then a dendrogram is built up according to these similarities using single-linkage hierarchical clustering and cutting the dendrogram at some level produces the overlapped community structure. Given a partition $P = \{P_1, P_2, \ldots, P_C\}$, a partition density D can be computed by the average partition density weighted by the fraction of present links in each partition:

$$D = \sum_c \frac{m_c}{M} D_c = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \tag{2.6}$$

where m_c and n_c are the numbers of edges and nodes in the partition P_c, respectively. The cutting threshold in the dendrogram can be determined by achieving a maximum partition density.

2.3 Agent-Based Modeling and Simulation

Agent-based modeling (ABM) is a class of computational analysis tools that is widely used for simulating system dynamics when the system consists of multiple autonomous and interacting individual components—named "agents". Each agent follows its own decision-making processes according to a set of rules and contextual information from the history, other agents, and possibly other environmental settings. The sets of behavioral rules can be identical for all agents (homogeneous) or different from agent to agent (heterogeneous). For example, in prisoner dilemma games, every agent follows the same strategy to negotiate; in an ecosystem, some agents play the roles of producers while others are consumers [Bonabeau, 2002, Weng and Menczer, 2013].

The key point of ABM is to describe a system by setting up the behavioral strategies of its constituent agents. ABMs are often applied to validate individual-level configurations by reproducing the patterns that result at the system level with empirical data. Model predictions are typically obtained by computational simulations, in which the outcomes of interactions between agents are repetitively calculated. This approach makes it possible to make predictions that reach beyond those derived by pure mathematical methods, when the model cannot be solved analytically.

Agent-based modeling has now obtained a central role in the study of natural systems. A large body of literature has been developing in the past few years about the internal characteristics of agents, their activities, connectivity, and multi-agent features [Castellano et al., 2009]. Biological, ecological, human collaborative systems, and society can be naturally translated into an agent-based framework. ABM techniques are therefore employed in domains that include biology, ecology, cognitive science, epidemiology, and the social sciences. Let us consider a few domains to demonstrate the usage of ABMs in practice.

Social dynamics The Axelrod model [Axelrod, 1997] investigated cultural dynamics by modeling individuals as nodes (agents) in the network in which whether a person interacts with another depends on the similarity between their statuses. Global convergence and the persistence of diversity are two important ingredients explored in

this setting. Holme and Newman [2006] proposed a model in which each agent is associated with an opinion. At each time step, agents either change their opinions to match neighbors, or re-wire links toward agents with similar opinions. The model can capture the dual process of social influence via opinion changes and selection via re-wiring of connections. Agent-based models have also been applied to the study of the birth and decline of scientific disciplines. The evolution of disciplines is guided by social interactions among agents representing scientists. Disciplines emerge from splitting and merging of social communities in a collaboration network. This model is capable of reproducing various empirical observations about the relationships between disciplines, scholars, and publications [Sun et al., 2013]. Many more models of social dynamics are reviewed in the literatures [Castellano et al., 2009].

Network evolution In the above examples, agents are connected in a network structure. This is often the case in ABMs, especially in the context of social systems. The edges in the network represent social relationships between pairs of individuals. Some models focus specifically on the local rules that regulate the growth and evolution of the network and lead to its observed global topology. Models have explored many different strategies of how an agent creates connections [Erdös and Rényi, 1960, Watts and Strogatz, 1998, Barabási and Albert, 1999, Weng et al., 2013b]. For example, the phenomenon of linking to well-connected nodes (e.g., people or Web pages) is described by preferential attachment mechanisms [Barabási and Albert, 1999].

Diffusion Information and innovation spread on networks, and we can observe the cascades that ensue as agents are infected. The diffusion process is affected by both the actions of agents and the underlying network structure. Watts [2002] studies the cascade sizes and vulnerability of the system to global cascades using a simple spreading process on random networks. Pastor-Satorras and Vespignani [2001] simulated classical epidemic models on scale-free networks, revealing that infections always survive no matter how small the spreading rate is. Goetz et al. [2009] proposed an agent-based model of blog dynamics, where each agent is associated with mechanisms capturing both the topology and temporal features.

2.4 Maximum-Likelihood Estimation

Maximum-likelihood estimation (MLE) [Cowan, 1998] is a method of estimating the parameters θ of a statistical model M, given the independent observed data $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$. Let us assume the probability of observing x_i in the model M given parameters θ is $f(x_i|\theta)$. Then the likelihood of having parameters θ equals to the probability of observing $\mathbf{X}$ given

θ:

$$\mathcal{L}(\theta|\mathbf{X}) = f(\mathbf{X}|\theta) = f(x_1|\theta)f(x_2|\theta)\dots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2.7}$$

$$\log \mathcal{L}(\theta|\mathbf{X}) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{2.8}$$

$$\hat{\theta} = \arg\max_{\theta} \log \mathcal{L}(\theta|\mathbf{X}) = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i|\theta) \tag{2.9}$$

To find the optimal parameter $\hat{\theta}$ which best describes the observed data given the model and thus provides the largest log-likelihood value, we can solve the equation 2.9 or computationally search for the best solution in the parameter space.

2.5 Classification Models and Evaluation

In machine learning, *classification* is to categorize new items given a training set of items with known categories. Classification is supervised learning and requires a set of existing categories and pre-labelled observations to train the model (or classifier). For example, given a group of hashtags each belonging to two classes, either popular or unpopular, to determine whether a new hashtag is popular or unpopular frames a classification task. This example only contains two classes, to which is often referred as *binary classification*.

Popular classifiers include linear regression, naive Bayes, support vector machine (SVM), and random forest [Bishop, 2006]. To avoid noises and outliers in the training set, in most cases, the classifier runs with k-fold cross validation. The training data is randomly partitioned into k groups. At each run the model is trained on $k-1$ groups and tested in the left one; the process is repeated k times until each subgroup has been used for testing once. The evaluation of outcomes is the average results among k runs.

Given a set of items with observed labels $D = \{d_1, d_2, \dots, d_k\}$, and corresponding predicted labels $P = \{p_1, p_2, \dots, p_k\}$. Accuracy is the fraction of correctly labelled items among all present ones.

$$\text{accuracy} = \frac{1}{k}|\{d_i \mid d_i = p_i, 1 \le i \le k\}| \tag{2.10}$$

For binary classification (with two classes 0 and 1), we have several additional measures for evaluation.

$$TP = |\{d_i \mid d_i = p_i = 1, 1 \le i \le k\}| \tag{2.11}$$

$$TN = |\{d_i \mid d_i = p_i = 0, 1 \le i \le k\}| \tag{2.12}$$

$$FP = |\{d_i \mid d_i = 0 \wedge p_i = 1, 1 \le i \le k\}| \tag{2.13}$$

$$FN = |\{d_i \mid d_i = 1 \wedge p_i = 0, 1 \le i \le k\}| \tag{2.14}$$

$$\text{precision} = \frac{TP}{TP+FP}, \quad \text{recall} = \frac{TP}{TP+FN} \tag{2.15}$$

Precision and recall are two most common evaluation quantities for binary classification. F_1 score is the harmonic mean of precision and recall:

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.16}$$

When the threshold of the classifier is not determined, we can get a set of outcomes with a variety of threshold settings. The receiver-operating-characteristic (ROC) plot displays these results together by plotting how the true positive rate (TPR, $\frac{TP}{TP+FN}$) changes with the false positive rate (FPR, $\frac{FP}{FP+TN}$). Other than presenting the sensitivity of the model outcomes to various thresholds, ROC analysis can also compare multiple models to help select the optimal. The more the curve approaches to the top left, the better the result. AUC is the area under the ROC curve, summarizing the ROC plot with one number between 0 (worst) and 1 (best).

2.6 Clustering Algorithms

Clustering is unsupervised learning, as in this case we do not have pre-labelled items and categories are usually unknown as well. The goal is to split the items into several groups so that each item is similar to those in the same group but dissimilar to others in different groups. Popular methods include k-means algorithm, hierarchical clustering, and expectation-maximization algorithm [Bishop, 2006]. Different from the community detection problem in Sec. 2.2.3, the clustering algorithms are traditionally developed to target at grouping text-based or feature-based documents. The community detection algorithms work for clustering nodes in the setting of a network structure, though often it can be framed in a way that classical clustering algorithms can be applied; for instance, in the link clustering community detection method, a hierarchical clustering algorithm is used after similarities are computed between pairs of edges in the network [Ahn et al., 2010].

Expectation-maximization (EM) algorithm [Dempster et al., 1977] is employed for classifying multi-dimensional user data in Chapter 9. The EM algorithm is commonly used to compute maximum-likelihood estimates of unknown parameters in probabilistic models involving latent variables. When applied in the context of clustering, the EM algorithm helps find solutions for mixture models. Let us assume that we are given a set of data points $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ for clustering and k potential clusters $\mathbf{Z} = \{z_1, z_2, \ldots, z_k\}$. The probability of picking the class z_j, $1 \leq j \leq k$, is determined by $P(z_j)$. Given a selected cluster z_j, to generate a data point x_i, $1 \leq i \leq n$, is sampled from $p(x_i|z_j)$, where $p(x_i|z_j)$ can be any distribution such as gaussian, poisson, and exponential distribution. For instance, the most common mixture model is a Gaussian mixture model, $p(x_i|z_j) \propto \mathcal{N}(\mu_j, \sigma_j)$. Our goal is to learn $P(z_j)$ and $p(x_i|z_j)$ in the underlying model and clusters by estimating the latent parameters:

$$\theta = (P(z_1), P(z_2), \ldots, P(z_k), \mu_1, \mu_2, \ldots, \mu_k, \sigma_1, \sigma_2, \ldots, \sigma_k) \tag{2.17}$$

To do so, we use the EM algorithm which involves two steps, expectation (E) and maximization (M). During E step, we calculate the expected value of the log-likelihood of

observed data given the model parameter θ:

$$\mathcal{Q}(\theta|\theta^{(t)}) = E_{\theta^{(t)}}[\log \mathcal{L}(\theta)] \tag{2.18}$$

where $\theta^{(t)}$ is the θ achieved in the previous step, and

$$\begin{aligned} \log \mathcal{L}(\theta) &= \log p(\mathbf{X}|\theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \\ &= \log \sum_{j=1}^{k} P(z_j) p(\mathbf{X}|z_j; \mu_j, \sigma_j) = \log \sum_{j=1}^{k} \sum_{i=1}^{n} P(z_j) p(x_i|z_j; \mu_j, \sigma_j) \end{aligned} \tag{2.19}$$

During the M step, we find the new parameter that maximizes this quantity $\mathcal{Q}$:

$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{Q}(\theta|\theta^{(t)}) \tag{2.20}$$

The EM algorithm iterates between E step and M step until convergence, and the label t marks the number of iterations. The initial parameters $\theta^{(0)}$ can be heuristically configured. Finally we can compute $P(z_j|x_i)$ so as to assign each data point x_i with a cluster label:

$$P(z_j|x_i) \propto p(x_i|z_j) P(z_j) \tag{2.21}$$

The cluster z_j that gives the maximum probability $P(z_j|x_i)$ should be the most likely cluster for a given data point x_i.

Chapter 3

Related Work

The remarkable advances in computer technology has lowered the cost of communication, fostering new types of social connections and intensive interactions online. The availability of data from the Internet, especially online social media, yields unprecedented opportunities to explore human and social phenomena on a global scale [Lazer et al., 2009, Vespignani, 2009]. The studies of information diffusion on social network involve four essential elements, people who produce and spread online content, characteristics of transmissible information, social network structure, and the mechanisms of diffusion processes (see Fig. 1.2). This chapter summarizes and reviews existing literatures on each part.

3.1 Human Dynamics

People are living in the society and unavoidably interacting with others all the time, motivating research on dynamics of collective human behaviors. *Human dynamics* is an emergent research field to study human behaviors using statistical tools and models in statistical physics. Research in this area has boosted after Barabási and Albert [2005]'s seminal paper on modeling activities to capture the bursty nature and heavy tails of human behaviors. In this model, bursts and the heavy tailed distribution of inter-event times were captured by people making decisions with a queuing process: an individual makes choices among a list of tasks according to some priorities. The following sections introduce several critical aspects in the studies of human dynamics, most about how people interact with or adopt social behaviors from one another in a collective environment.

3.1.1 Threshold Model

The *threshold model* describes how people decide to adopt ideas or behaviors: people tend to follow the same trend as most of their friends do so [Granovetter, 1978, Morris, 2000]. It defines a threshold as the number or fraction of others who must make one decision before a given actor does the same. The threshold marks the point where net benefits begin

to exceed net costs for the actor, triggering adoption which can yield more advantages. Furthermore, considering that influences of friends on a given actor vary according to the relationship between a friend and the actor, these friends may have different contributions to the threshold.

The threshold model can be widely applied to various systems and situations, such as information diffusion, rumors, diseases, strikes, voting, educational attainment, migration, and experimental social psychologies [Granovetter, 1978]. In the online social media sites, an individual is more likely to believe and adopt an idea, if she has been exposed to this idea multiple times from her social connections. Many empirical studies succeeded to demonstrate the existence of such a "threshold" in social and behavioral contagions online [Backstrom et al., 2006, Bakshy et al., 2009, Cosley et al., 2010, Romero et al., 2011b].

3.1.2 Homophily

The principle of *homophily* suggests that similar people are more likely to have contact than dissimilar ones, also known as "birds of a feather flock together" or "similarity breeds connection" [Kandel, 1978, McPherson and Smith-Lovin, 1987, McPherson et al., 2001, Kossinets and Watts, 2009]. The existence of homophily in social groups has been supported by various empirical observations and experiments in the online settings [Fiore and Donath, 2005, Şimşek and Jensen, 2008, Aral et al., 2009, De Choudhury, 2011, Schifanella et al., 2010, Aiello et al., 2012, Gallos et al., 2012, Weng and Lento, 2014]. Crandall et al. [2008] proposed a homophily-based model to predict a user's future activity and interactions with others according to user similarities. A feedback loop (or the Echo-Chamber Effect) was claimed to result in similarity among users: people grow to resemble their friends because of the social influence (peer influence), and meanwhile they are more likely to form links with similar people (homophily) [Crandall et al., 2008, Jamieson and Cappella, 2009, Weng and Lento, 2014]. Though it is hard to fully distinguish these two processes [Shalizi and Thomas, 2011], the homophily effect was suggested to greatly promote behavioral contagion other than the peer influence [Aral et al., 2009]. It is also worthy mentioning that dissimilarity, disagreement, and heterogeneity also exist among people close to each other at the same time [Munson and Resnick, 2010, Goel et al., 2010, Aral and Van Alstyne, 2011, Brzozowski et al., 2008], which might cause segregation in social circles.

3.1.3 Weak Tie Hypothesis

Friendships vary in their intensity and intimacy, triggering the concept of tie strength; "strong ties" are our closest confidants and supporters, while "weak ties," with whom we feel less close, comprise the majority of our personal networks. In the seminal paper "The strength of weak ties", Granovetter [1973] defined the strength of social ties proportionally to the size of shared social circles and proposed the well-known *weak tie hypothesis* [Granovetter, 1973, 1995]. He hypothesized that weak ties do not carry as much communication as strong ties, but they act as bridges between communities and thus as important

channels for novel information that people rarely get from close social circles. As follow-ups of Granovetter's studies, there were many empirical testings on the weak-tie hypothesis, mostly using surveys and interviews [Friedkin, 1980, Lin et al., 1981, Granovetter, 1983, Brown and Reingen, 1987, Nelson, 1989, Levin and Cross, 2004, Gilbert and Karahalios, 2009]. Brown and Reingen [1987] found an important bridging function of weak ties in word-of-month referral behavior, allowing information to travel from one distinct subgroup of referral actors to another. Levin and Cross [2004] investigated dyadic social ties in transferring useful knowledge. They found that strong ties lead to the receipt of useful knowledge more than weak ties, but weak ties benefit knowledge transmission when the trustworthiness is controlled. Gilbert and Karahalios [2009] tested several dimensions of tie strength on social media, revealing that both intensity of communication and intimate language are strong indicators of relationship closeness. In summary, strong ties were believed to provide greater emotional support [Wellman and Wortley, 1990] and are more influential [Brown and Reingen, 1987, Bakshy et al., 2012, Bond et al., 2012], while weak ties provide novel information and connect us to opportunities outside our immediate circles [Granovetter, 1973, Putnam, 2001, Burt, 2009]. We communicate more frequently with our strong ties [Granovetter, 1973] and we talk to them over a variety of media [Haythornthwaite and Wellman, 1998, Wellman and Wortley, 1990].

However, only a handful of studies were performed recently taking advance of large-scale datasets of human behavior [Onnela et al., 2007b, Burke and Kraut, 2013, Bakshy et al., 2012]. Onnela et al. [2007b] analyzed a mobile call network and showed that individuals in clusters tend to communicate more, while ties between clusters have less traffic. The stimulation results of their model demonstrated similar results that weak ties, acting as bridges, slow down the diffusion process. Bakshy et al. [2012] experimented on Facebook comparing individual adoption rate when an external URL shared by friends is or is not included in the newsfeed. They found that although stronger ties are individually more influential in persuading others to adopt and spread information, more abundant weak ties are responsible for the propagation of novel information. Furthermore, by considering the temporal nature of communication, it was shown that the Granovetter type weight-topology correlations are mostly responsible for the slow spreading of information in communication networks besides bursty temporal behavior [Karsai et al., 2011, Miritello et al., 2011].

3.1.4 Limited Attention

Our limited attention is constrained by a cognitive limit on the number of stable social relationships that one can sustain, as postulated by Dunbar [1998] and later supported by analysis of Twitter data [Huberman et al., 2009, Gonçalves et al., 2011]. Huberman et al. [2009] defined friends of a given Twitter user as others who have been mentioned by the user at least twice. Then they found out that most users have a very small number of friends compared to a large number of followers, and the friend network is more influential than the follower network in driving Twitter usage. Wu and Huberman [2007] analyzed the dynamics of collective attention on Digg.com and modeled the delay of collective attention with a single novelty factor. Their measurements indicated that novelty within groups

decays with a stretched-exponential law, suggesting the existence of a natural time scale over which attention fades.

The dynamics of information is driven more than ever before by the *economy of attention* [Simon, 1971], due to the availability of abundant information through online social media and other socio-technical systems. Simon [1971] first introduced the notion of "attention economy" to describe the mutating balance between the increasing supply of information and the essentially rigid demand that is limited by our finite attention. Attention has been described as the *new currency of business* [Davenport and Beck, 2001], raising the question of what are the laws that govern it [Goldhaber, 1997] and motivating efforts to model these laws quantitatively [Falkinger, 2007]. The most obvious consequence of living in a world with abundant information and scarce attention is that pieces of information need to compete for our individual and collective attention to survive. In this context, one of the most challenging problems is the study of the competition dynamics of ideas, information, knowledge, and rumors [Crane and Sornette, 2008, Lerman and Ghosh, 2010, Wu and Huberman, 2007, Moussaid et al., 2009, Weng et al., 2011, 2012].

In the context of social media, several authors explored the temporal evolution of popularity. Wu and Huberman [2007] studied the decay in news popularity, and showed that temporal patterns of collective attention are well described by a multiplicative process with a single novelty factor. While the decay in popularity was attributed to competition for attention, the underlying mechanism was not modeled explicitly. Crane and Sornette [2008] introduced a model to describe the exogenous and endogenous bursts of attention toward a video, by combining an epidemic spreading process with a forgetting mechanism. Hogg and Lerman [2009] proposed a stochastic model to predict the popularity of a news story via the intrinsic interest of the story and the rates at which users find it directly and through friends. These models described the popularity of a single piece of information, and were therefore unsuitable to capture the competition for our collective attention among multiple simultaneous information epidemics. Although recent epidemiological models have started considering the simultaneous spread of competing strains [Sneppen et al., 2010, Karrer and Newman, 2011], they did not provide a framework to deal with an arbitrary number of new topics continuously injected into the system.

3.1.5 Social Influence

The concept of *social influence* has been discussed extensively in social media research. There is no concrete and unified definition of social influence yet. Most of the studies in the literature consider people who are more popular or more likely to persuade others to adopt certain social behaviors as those with higher social influence. Many methods for quantifying social impact and identifying influential users have been proposed, for instances, in terms of high in-degree in the follower network [Cha et al., 2010, Suh et al., 2010], information forwarding activity [Romero et al., 2011a, Kwak et al., 2010], seeding larger cascades [Kitsak et al., 2010, Bakshy et al., 2011], PageRank scores [Kwak et al., 2010], or topical interests [Tang et al., 2009, Weng et al., 2010, Yang et al., 2012, Weng and Menczer, 2014].

Cha et al. [2010] compared three influence measures based on the number of followers (in-degree), retweets, and mentions. These three measures did not show much overlap among top influentials, as they represented different aspects of social impact. In-degree alone revealed little about the user influence, and comparatively the numbers of mentions and retweets had higher correlations between each other. Their findings also supported that influence is not gained spontaneously or accidentally, but through concerted long-term efforts. In addition, user ranks by retweets were found to differ from ranks by PageRank scores [Kwak et al., 2010]. Bakshy et al. [2011] defined social influence as users' abilities to seed contents containing URLs that generate large cascades of reposts, precisely measured as the logarithm of the average size of all cascades for which that user was a seed. Then future influences were predicted according to individual features and their past influences. However, the intuition that influential people are more likely to be influential in the future was found to be only correct on the average level, and content features were not helpful on making better predictions. Weng et al. [2010] claimed that social influence is topic-sensitive and proposed a PageRank-based algorithm, TwitterRank, to incorporate both the topical similarity and the link structure between users.

Crandall et al. [2008] pointed out that the concept of homophily should be two-folded: people grow to resemble their friends because of the social influence, and they are more likely to form links with similar people. Bond et al. [2012] ran a randomized controlled experiment of political mobilization messages in Facebook. According to the experiment, the messages directly influenced political self-expression, information seeking, and real-world voting behavior of people, and furthermore the messages had transmissible influences on the users who received them, users friends, and friends of friends. A recent experiment by Muchnik et al. [2013] showed that positive and negative votes can have different effects in terms of influencing others' votes: false positive votes lead to inflated subsequent scores, whereas false negative ones have small long-term effects.

3.2 Content Features

3.2.1 Content Innate Appeal

The *innate appeal* of a message is commonly believed to contribute to how likely people are willing to share it with others and essentially influence its future virality. Berger and Milkman [2009] studied how emotion hidden in the content of news articles affects whether content is highly shared. They found that positive content is more viral than negative content, as is content that inspires awe; however, while sad content is less viral, anger or anxiety inducing articles are both more likely to be popular. Guerini et al. [2011] characterized various aspects that indicate virality of text-based contents, based on the assumption that virality is an intrinsic trait of text. Tsur and Rappoport [2012] analyzed a rich set of content-based features extracted from hashtags, such as the number of words contained, spelling, lexical items, location in tweets, emotional and cognitive dimensions, in order to predict future popularity of the hashtags. However, a controlled experiment on music

choices suggested that innate quality may play only a minor role in determining future success due to strong effect of social influence [Salganik et al., 2006].

3.2.2 Topic Detection and Topic Locality

The ability to identify *topics* of online content can endow us with many applications and studies concerning a finer-grained categorization of information. Several studies examined the recognition of topics in the online scenario and social media [Leskovec et al., 2009, Xie et al., 2011, Simmons et al., 2011, Agarwal et al., 2012, Ferrara et al., 2013]. Leskovec et al. [2009] grouped short, distinctive phrases by single-rooted directed acyclic graphs used as signatures for different topics. Features extracted from content, metadata, network, and their combinations were leveraged to detect events in social streams [Aggarwal and Subbian, 2012, Ferrara et al., 2013]. Another approach is based on the discovery of dense clusters in the inferred graph of correlated keywords, extracted from messages in a given time frame [Agarwal et al., 2012, Tang et al., 2009]. Our approach in Chapter 6 inherits a similar strategy to identify clusters of similar hashtags by detecting communities in the network topology [Blondel et al., 2008, Newman, 2012] on account of topic locality.

Topic locality (or *topicality*) in the Web describes such a phenomenon that most Web pages tend to link with related content [Davison, 2000, Menczer, 2004]. The effect of topic locality is utilized in focused Web crawlers [Menczer and Belew, 1998], collaborative filtering [Goldberg et al., 1992, 2001], interest discovery in social tagging [Schifanella et al., 2010, Aiello et al., 2012], and many other applications [Haveliwala, 2003, Michlmayr and Cayzer, 2007, Tang et al., 2009, Weng et al., 2010]. In the context of social media sites like Twitter, topic locality refers to the assumption that semantically similar hashtags are more likely to be mentioned in the same posts and therefore to be close to each other in the hashtag co-occurrence network.

We see a growing literature on discovering *user interests* and topics [Java et al., 2007, Phelan et al., 2009, Michelson and Macskassy, 2010, Chen et al., 2010, Weng et al., 2010, Xu et al., 2011]. A common approach is using a vector representation generated from all the posts by a user to represent his interests. Then whether a user would be interested in a newly incoming message is estimated by the similarity between the feature vectors of user interests and the message [Chen et al., 2010, Weng et al., 2012]. LDA has also been applied to extract user interests from user generated content [Weng et al., 2010]. Java et al. [2007] looked into communities of users in the reciprocal Twitter follower network and summarized user intent into several categories (daily chatter, conversations, information sharing, and news updates); a user could talk about various topics with friends in different communities. Michelson and Macskassy [2010] discovered entities mentioned in tweets according to predefined folksonomy-based categories to allocate topics so as to build an entity-based topic profile. The diversity of user interests has not yet been thoroughly investigated. An exception is the work of An et al. [2011], who explored which news sources Twitter users are following and correlated the observation with the diversity of their political opinions.

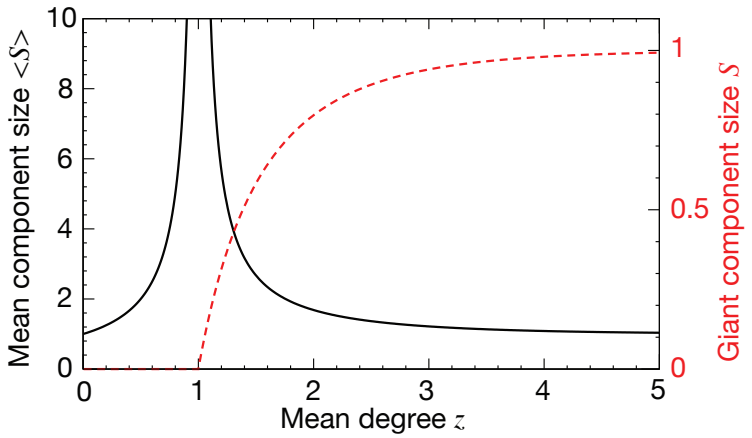


Figure 3.1: The plot of the mean component size excluding the giant component if there is one (black solid line), and the giant component size (red dashed line), for the ER random network [Newman, 2003]. The mean degree $z = p(n-1)$.

3.3 Network Dynamics

Network structure specifies the underlying topology of linkages among individuals, which is critical for the dynamics of the diffusion process on top [Barrat et al., 2008].

3.3.1 Network Evolution Models

Various models devoted to reproducing the growth and evolution of network topology have been presented to capture different characteristics of complex networks. Most such models focused on defining basic mechanisms that drive link creation [Albert and Barabási, 2002, Newman, 2003, Wasserman and Faust, 1994, Newman, 2010, Barrat et al., 2008].

Random Networks

The first network evolution model proposed in 1959 by Erdös and Rényi [1960] described the process of growing a *random network*: n nodes connected by m edges randomly selected from all $n(n-1)/2$ possible edges with equal probability p. The degree of *Erdös-Rényi (ER) network* has a Poisson degree distribution. The other key feature is a sudden change of the network connectivity with the increase of p: when p is small, many clusters are small and isolated, but once p increases to be larger than a critical value, the network suddenly becomes very dense where almost all the nodes are linked to each other in a giant connected component (see Fig. 3.1).

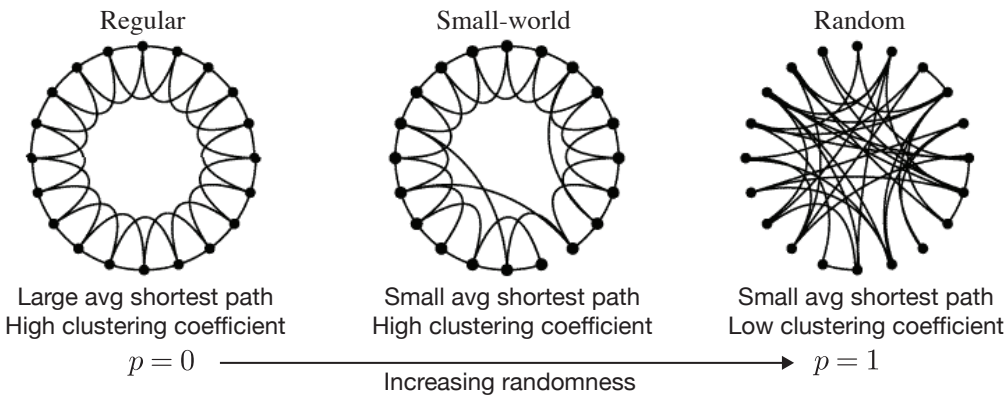


Figure 3.2: The Watts-Strogatz model reproduces the small-world phenomenon by rewiring edges in a regular network according to the randomness parameter p [Watts and Strogatz, 1998].

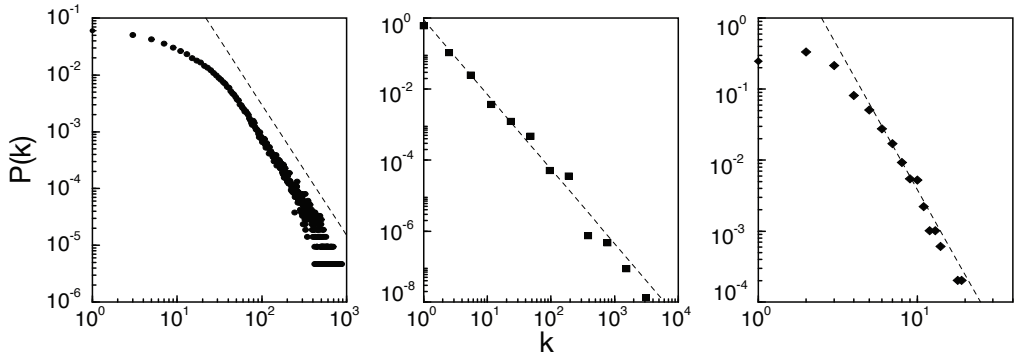
Small-World Networks

The *small-world network* originated from the experiment of Milgram [1967], in which selected persons were asked to deliver a letter to a target receiver by only passing the letter to their acquaintances. Among all the successful instances, the average length of these communication chains was short, around six steps. The phenomenon is well known as "small-world effect" or "six degrees of separation". A small-world network has acquaintanceship-based edges and the distance between a random pair of people is smaller than expected. In the real world setting, the small-world effect implies that most friends of an individual are people living around, but he may also have a few friends far away. People are moving around, but the geographic distance limits the strength of social relationships. The *Watts-Strogatz model* was designed to reproduce the small-world phenomenon by rewiring each link in a regular network with a probability p [Watts and Strogatz, 1998]. As shown in Fig. 3.2, when $p = 0$, the network is fully ordered; when $p = 1$, every edge is rewired so as to create a random network; when $0 < p < 1$, we obtain a small-world network with small average shortest path and high clustering coefficient [Watts and Strogatz, 1998].

Scale-Free Networks

A *scale-free network* has a power-law degree distribution, commonly seen in many real-world networks, such as the Internet, the film actor network, the scientific collaboration network, the citation network, and many others (see Fig. 3.3) [Barabási and Albert, 1999, Albert and Barabási, 2002, Newman, 2003]. Highly unbalanced degree distribution in a social network indicates that, in a large group of people, only a few are extremely popular and most others do not have too many contacts. It has been suggested to be the most critical feature of social networks [Newman et al., 2002].

Among many models that can capture the heterogeneous distribution in connectivity [Kleinberg et al., 1999, Kumar et al., 2000, Newman et al., 2002, Krapivsky and Redner, 2001,



(a) Actor (b) WWW (c) Powergrid

Figure 3.3: The connectivities of various large real-world networks have scale-free distributions, (a) actor collaboration graph, (b) the World Wide Web, and (c) the power grid network [Barabási and Albert, 1999].

Dorogovtsev et al., 2000, Fortunato et al., 2006], *Barabási-Albert model* was the first to generate a scale-free network with two simple mechanisms: continuously adding new nodes into the system ("growth") and connecting with other nodes with preference to the high-degree ones ("preferential attachment") [Barabási and Albert, 1999]. Motivated by the structure of the Web graph, the *copying model* added a new node into the network and linked it to a random existing node or its neighbors [Kleinberg et al., 1999, Kumar et al., 2000]. Another model proposed by Newman et al. [2002] aimed to build up a random graph with the arbitrary degree setting. The *ranking model* grew the network according to a rank of the nodes by any given prestige measure; the probability of linking a target node could be any power law function of its rank, resulting in a power-law degree distribution [Fortunato et al., 2006].

Network Models with Other Features

In the social context, the rationale behind the preferential attachment mechanism is that people prefer to form edges with well connected individuals such as celebrities [Barabási and Albert, 1999]. However, this prescription alone is not sufficient to reproduce several other important social features of real networks. Other models have been put forth to fill this gap, including ingredients such as homophily [Holme and Newman, 2006, McPherson et al., 2001, Papadopoulos et al., 2012, Aiello et al., 2012, Gallos et al., 2012], triadic closure [Simmel and Wolff, 1950, Granovetter, 1973, Leskovec et al., 2008, Krackhardt and Handcock, 2007, Romero and Kleinberg, 2010], hierarchical structure [Clauset et al., 2008], and information diffusion [Barbieri et al., 2013, Weng et al., 2013b].

The impact of homophily on link creation in large-scale online networks is interpreted as people with a propensity for linking with similar others [Holme and Newman, 2006, Papadopoulos et al., 2012, Aiello et al., 2012, Gallos et al., 2012]. The triadic closure mechanism is based on the intuition that two individuals with mutual friends have a higher probability to establish a new contact [Simmel and Wolff, 1950, Granovetter, 1973]. This

tendency was observed in both undirected and directed online social networks and incorporated into several network growth models [Leskovec et al., 2008, Krackhardt and Handcock, 2007, Romero and Kleinberg, 2010]. In particular, Leskovec et al. [2008] tested triadic closure against many other mechanisms in four different large-scale social networks. By using Maximum Likelihood Estimation (MLE) [Cowan, 1998], they identified triadic closure as the best rule, among those considered, to explain link creation and to reproduce the clustering coefficient and the degree distribution.

3.3.2 Community Structure

Social networks naturally consist of *communities* corresponding to certain social circles or interest groups, differentiating social networks from other complex networks [Newman and Park, 2003]. Although there is no clear consensus, a community is often defined as a densely connected subgraph. Community structure is a common approach to understand social network as a mesoscopic description of the topology [Newman, 2006, Fortunato, 2010, Newman, 2012].

The problem of community detection originated from the graph partitioning problem in graph theory, such as the Kernighan-Lin algorithm [Kernighan and Lin, 1970], spectral partitioning [Fielder, 1973, Hagen and Kahng, 1992], and hierarchical clustering [Scott, 2000, Newman, 2012]. Various detection methods have been proposed for either disjoint communities [Girvan and Newman, 2002, Newman, 2006, Blondel et al., 2008, Rosvall and Bergstrom, 2008] or overlapping ones [Palla et al., 2005, Ahn et al., 2010], where each node can belong to only one or multiple communities, respectively. Modularity is widely adopted as an efficient measure to detect communities in graph [Newman, 2006, Blondel et al., 2008]. It is defined as the number of edges in one group minus the expected edges placed at random, to evaluate how well a set of vertices are grouped together [Newman, 2006]. The maximization of the modularity score is leveraged to do the community detection. The infomap method decomposed a network into modules by optimally compressing a description of information flows on the network [Rosvall and Bergstrom, 2008]. Clique percolation built up overlapping communities from k-cliques, each corresponding to a complete sub-graphs of k nodes [Palla et al., 2005]. Link clustering method reinvented communities as groups of links and constructed a hierarchy based on similarities between links [Ahn et al., 2010].

3.4 Information Diffusion

3.4.1 Internet Memes

A *meme* usually refers to a concept, a topic, an idea, or a piece of information amenable to be shared and transferred from one to another. The concept was first proposed by Richard Dawkins in his book *The Selfish Gene* [Dawkins, 1989], referring to "a unit of cultural

transmission, or a unit of imitation." Operationally a meme could be identified by a word, a phrase, a tag, or an URL on the Internet, all of which are transmissible among online users and replicated on the social media platforms. A widely adopted example is using Twitter hashtag as a meme indicator [Romero et al., 2011b, Lehmann et al., 2012, Yang et al., 2012, Weng et al., 2012, 2013a, Ma et al., 2013] (see more explanations in Sec. A.2).

3.4.2 Epidemic Models

Early models concerning communication dynamics were inspired by studies of epidemic spreading [Rapoport, 1953, Goffman and Newill, 1964, Daley and Kendall, 1964, Bailey, 1975, Anderson et al., 1992]. Similar to how an infectious disease is transmitted among the population, a piece of information can pass from one individual to another through social connections and "infected" individuals can, in turn, propagate the information to others, possibly generating a full-scale contagion. The SIS [Bailey, 1975] and SIR [Anderson et al., 1992]models are two classical models in epidemiology, in which the infected population grows exponentially until the rate of infection is balanced by the rate of recovery, or the contagion finally dies off when the recovery rate prevails.

Neither the SIS or SIR model considers the underlying network structure of how individuals are connected. Pastor-Satorras and Vespignani [2001] studied the epidemic spreading in the scale-free networks, particularly with the data of computer viruses propagating on the Internet. Different from what the SIS model states, according to the empirical data, there was no such a critical threshold of the spreading rate that separated the persistence and the vanishing of viruses. No matter how small the spreading rate was, viruses could always survive and infect a small proportion of nodes.

Many studies were accomplished on modeling the information spread in the online milieu. Gruhl et al. [2004] proposed a model of information diffusion, in which users were connected by a directed network and each edge was associated with a topic adoption rate and a visiting frequency. Galuba et al. [2010] designed a similar propagation model with a target probability associated with each pair of users. These models are akin to the traditional approach for disease propagation, but very complex due to a large number of free parameters. Leskovec et al. [2007b] extended the SIS model to apply to the online social environment. It contained one parameter β, similar to the infection probability in SIS, representing how infectious a post was to others in the blogosphere. This simple model could predict top cascade shapes as well as several corresponding distributions of cascade sizes. Later a more complex model based on a random walk mechanism attempted to capture both the topology and temporal dynamics of information diffusion through blog links [Goetz et al., 2009].

3.4.3 Information Diffusion Models

Early diffusion models were designed with flavors similar to epidemic models. They were later extended to include cascade phenomena [Goldenberg et al., 2001], factors that influence the speed of spreading such as information recency [Moreno et al., 2004], the hetero-

geneity in connectivity patterns [Pastor-Satorras and Vespignani, 2001], clusterings [Onnela et al., 2007b], user-created content [Bakshy et al., 2009], and temporal connectivity patterns [Morris and Kretzschmar, 1995, Butts, 2008, 2009, Perra et al., 2012].

Most recent work on the analysis and modeling of online information diffusion aimed to reproduce statistical features of the cascades as in the empirical data or learn the mechanism of how a message is propagated. Leskovec et al. [2009] modeled the dynamics of the news cycles by a process in which different sources copy one another and meanwhile topics are governed by strong recency effects. The simulation results seemed to capture the highly non-uniform topic volume distribution without any exogenous influence. The roles of user influence and resource virulence were also studied for the spread of URLs; URLs retweeted by powerful users in Twitter were found to be more likely to generate big cascades [Galuba et al., 2010].

Another different set of models were derived from the threshold model [Granovetter, 1978, Centola, 2010, Romero et al., 2011b]. The threshold model believes that people are more likely to adopt an idea when many friends share it; In other words, the adoption rate increases with more adopted neighbors in the social network. Bakshy et al. [2009] studied the user-to-user content sharing in Second Life and found that the adoption rate increases as more friends adopt, but users with many friends are less likely to be influenced by any particular one. Cosley et al. [2010] measured user influence in Wikipedia by expressing the probability of adoption as a function of the number of neighbors who have adopted. Similar measures of the adoption probability were also explored in Twitter [Romero et al., 2011b, Hodas and Lerman, 2012]. Furthermore, the diversity in the local neighborhood was shown to help increase the individual adoption and engagement rate in Facebook, suggesting that a user prefers to adopt a social behavior if he has received reinforcement from various social groups [Ugander et al., 2012].

3.4.4 Temporal Patterns

A large proportion of studies on temporal patterns were about bursty dynamics of online activities or content popularities [Ratkiewicz et al., 2010, Leskovec et al., 2009, Crane and Sornette, 2008, Lehmann et al., 2012, Yang and Leskovec, 2011, Aral et al., 2009]. Ratkiewicz et al. [2010] found that bursty dynamics of Wikipedia document popularities, including the distributions of burst sizes and inter-burst times, cannot be reproduced by neither the preferential attachment mechanism [Barabási and Albert, 1999] or the ranking model [Fortunato et al., 2006], because neither is able to capture the shift of user attention by exogenous factors. Instead, they proposed a rank-shift model to simulate attention shifting events, using a re-ranking probability to boost some topics at random. Leskovec et al. [2009] studied the dynamics and persistent temporal patterns of online news. The highly non-uniform volume distribution of news was proposed to be captured by two factors, imitate (different sources copy one another) and recency (topics are governed by strong recency effects) without any exogenous influence. Crane and Sornette [2008] modeled bursts of human activities by both endogenous and exogenous influences and stipulated four dynamic classes according to the type of disturbance (endogenous or exogenous) and the

ability of people influencing others (critical or subcritical). A study on bursty patterns of collective attention in Twitter classified hashtags into four categories according to the fractions of tweets before, during, and after the spikes [Lehmann et al., 2012]. They proposed that the hashtag popularity was mostly driven by exogenous factors, and the epidemic spreading played a minor role.

3.4.5 Content Virality

The questions of which memes will go viral in the future have attracted much attention across disciplines. Existing research attempted to characterize viral memes in terms of message content [Berger and Milkman, 2009, Guerini et al., 2011, Salganik et al., 2006], temporal variation [Leskovec et al., 2009, Szabo and Huberman, 2010], influential users [Kitsak et al., 2010, Aral and Walker, 2011], finite user attention [Weng et al., 2012, Lehmann et al., 2012], and local neighborhood structure [Ugander et al., 2012].

User behaviors and characteristics are important aspects in the context of meme virality. Limited individual attention causes the competition among memes, inducing strong heterogeneity in meme popularity and life span longevity [Weng et al., 2012]. Each user has different interests and it alters adoption preference and meme popularity [Yang et al., 2012]. Influential users were expected to promote memes more so as to make them popular in the future [Weng et al., 2010, Bakshy et al., 2011, Kitsak et al., 2010].

The structure of the underlying network has a crucial impact on the spreading process in general [Daley and Kendall, 1964, Goffman and Newill, 1964, Christakis and Fowler, 2007, Barrat et al., 2008, Pastor-Satorras and Vespignani, 2001]. The existence of hubs, nodes with extremely large degree, was known to affect the persistence of infections, the distribution of cascade sizes, and the vulnerability of the system [Pastor-Satorras and Vespignani, 2001, Watts, 2002]. In addition to the hubs, another important network structure in most real social networks is community [Newman, 2006, Girvan and Newman, 2002, Rosvall and Bergstrom, 2008, Ahn et al., 2010, Fortunato, 2010]. Communities is commonly believed to constrain the information flow or slow down the spread of diseases [Granovetter, 1973, Onnela et al., 2007b, Rosvall and Bergstrom, 2008, Weng et al., 2013a, 2014c].

The spread of memes is often considered as social contagion, defined as the spread of information or behavior on a social network where an individual serves as the stimulus for the imitative action of another [Lindzey and Aronson, 1985]. Meanwhile, memes bear lots of similarities to infectious diseases, as both travel through social ties from one person to another [Daley and Kendall, 1964, Goffman and Newill, 1964]. However, studies have shown that information contagion may spread differently from diseases, as multiple exposures can significantly increase the chances of adoption [Granovetter, 1978, Centola, 2010, Romero et al., 2011b]. The speed and ease of meme transmission is also affected by characteristics of social ties. Strong and homophilous ties were considered to be more effective than weak ties for spreading messages [Brown and Reingen, 1987], while weak ties could transmit novel information [Granovetter, 1973]. In viral marketing and consumer studies, researchers actively apply network approaches to analyze and model local and global

patterns of social network structure [Leskovec et al., 2007a, Mason et al., 2008, Aral and Walker, 2011].

One of the mainstream approaches to detect viral memes is *time series analysis*, which examines temporal patterns such as growth, bursts, and decay of meme traffic [Wu and Huberman, 2007, Romero et al., 2011b, Asur et al., 2011].Temporal patterns of memes could be summarized into a few categories with predictive power to spot trendy or bursty memes [Yang and Leskovec, 2011, Lehmann et al., 2012]. Classification of temporal patterns was also seen as an extended application of trajectory clustering [Gaffney and Smyth, 1999, Lee et al., 2007]. Many existing virality prediction algorithms tried to forecast the future time series based on the sequences of all the past values [McNames, 1998, Lenser and Veloso, 2005, Kaltenbrunner et al., 2007]. Some event detection methods grouped memes together to form topics and use temporal activity to detect trending topics [Becker et al., 2011, Cataldi et al., 2010].

In another approach, the prediction problem is treated as a *classification* task. Several studies claimed that the early popularity of online content is strongly correlated with its future popularity [Jamali and Rangwala, 2009, Szabo and Huberman, 2010, Lerman and Hogg, 2010, Tatar et al., 2011]. Szabo and Huberman [2010] proposed a model predicting future popularity of a Youtube video based on its early popularity. Jamali and Rangwala [2009] made use of daily user activities, user interest peaks, and comment information attached to each Digg story to estimate the future usage. Design elements of a Website were shown to be informative as well; Lerman and Hogg [2010] found that incorporating design features of the website can improve the outcomes of their stochastic prediction model. The numbers of URLs and hashtags in a tweet were suggested to be strongly correlated with the tweet's retweetability, while the number of followers, followees, and the account age weakly affected its retweetability [Suh et al., 2010]. Yang et al. [2012] quantified how a user selects content tags using individual interests, relevance, and behavior of neighbors. Romero et al. [2013] predicted popularity of a tag based on the social connections of its early adopters. Some other notable features are content properties (terms, language, semantics, category, and so on) [Berger and Milkman, 2009], user influence [Bakshy et al., 2011, Salganik et al., 2006], source authority [Bandari et al., 2012], and the graph topology of early adopters [Romero et al., 2013, Ma et al., 2013]. In a recent paper, Cheng et al. [2014] formulated social virality prediction as a sequence of binary classification problems, while a cascade is tracked over time. In spite of the different problem formulation, our results seem to be consistent with their finding that initially, breadth is a strong indicator of larger cascades.

Part I

Actor: Limited Attention

How does limited human attention affect meme competition?
Is attention assigned differently on strong and weak ties?

Chapter 4

Limited Attention

Ideas have formidable potential to impact public opinion, culture, policy, and profit. However, the abundance of information yields the scarcity of individual and collective attention, driving fierce competition among ideas and memes for attention to survive [Simon, 1971, Davenport and Beck, 2001]. Understanding the competition dynamics of ideas, information, knowledge, and rumors is crucial in a broad range of settings, from viral marketing to scientific discovery acceleration. Aspects of competition for limited attention have been studied through news, movies, and topics posted on blogs and social media [Crane and Sornette, 2008, Leskovec et al., 2009, Lerman and Ghosh, 2010]. The popularity of news decreases with the number of competing items that are simultaneously available [Wu and Huberman, 2007, Moussaid et al., 2009, Asur et al., 2011]. This chapter inquires into limited attention for producing and processing information and examines how finite attention intensifies competition among memes, using empirical data from Twitter and the technique of agent-based modeling.

We first outline a number of empirical findings on finite individual attention and user interests, motivating both our question and the main assumptions behind the model. We then propose an agent-based toy model of meme diffusion and compare its predictions with the empirical data. Finally we show that the social network structure and our finite attention are both key ingredients of the diffusion model, as their removal leads to results inconsistent with the meme dynamics observed in the empirical data.

4.1 Limited Attention

We first explore the competition among memes. In particular, we test the hypothesis that the attention of a user is somewhat independent from the overall diversity of information discussed in a given period. Let us quantify the *breadth of attention* of a user through Shannon entropy $S = -\sum_i f(i) \log f(i)$ where $f(i)$ is the proportion of tweets generated by the user about meme i. Given a user who has posted n messages, her entropy can be as small as 0, if all of her posts are about the same meme; or as large as $\log n$ if she has posted a

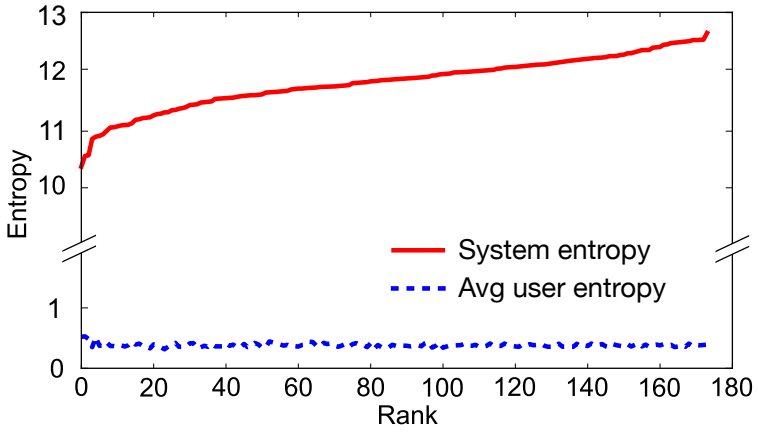


Figure 4.1: The plot of daily system entropy (solid red line) and average user breadth of attention (dashed blue line). Days in our observation period are ranked from low to high system entropy, therefore the former is monotonously increasing.

message about each of n different memes. We can measure the diversity of the information available in the system analogously, defining $f(i)$ as the proportion of tweets about meme i across all users. Note that these entropy-based measures are subject to the limits of our operational definition of a meme; finer or coarser definitions would yield different values.

In Fig. 4.1 we compare the daily values of the system entropy to the corresponding average user entropy. The key observation here is that a user's breadth of attention remains essentially constant irrespective of system diversity. It is a clear indication that the diversity of memes to which a user can pay attention is bound. With the continuous injection of new memes, this indirectly suggests that memes survive at the expense of others. We explicitly assume this in the information diffusion model presented later.

4.2 User Interests

It has been suggested that topical interests affect user behavior in social media [Ienco et al., 2010, Michelson and Macskassy, 2010, Yang et al., 2012, Romero et al., 2013]. This is a potentially important ingredient in a model of meme diffusion, as an interesting meme may have a competitive advantage. Therefore we wish to explore whether user interests, as inferred from past behavior, are predictive of future behavior.

Let us consider every user and all the retweets they produce in our dataset. When a user u emits a new retweet, we define her *interests* I_u as the set of all memes about which she has tweeted up to that moment. We also collect the set M_0 of memes associated with the new retweet. The n most recent posts across all users prior to the new retweet are considered as a set of potential candidates that might have been retweeted, but were not. The corresponding sets of memes $M_1, M_2, \ldots, M_n$ are recorded ($n = 10$). We compute the similarity $sim(M_0, I_u), sim(M_1, I_u), \ldots, sim(M_n, I_u)$ between the user interests and the actual and candidate posts, and recover the conditional probability $P(retweet(u, M)|sim(M, I_u))$

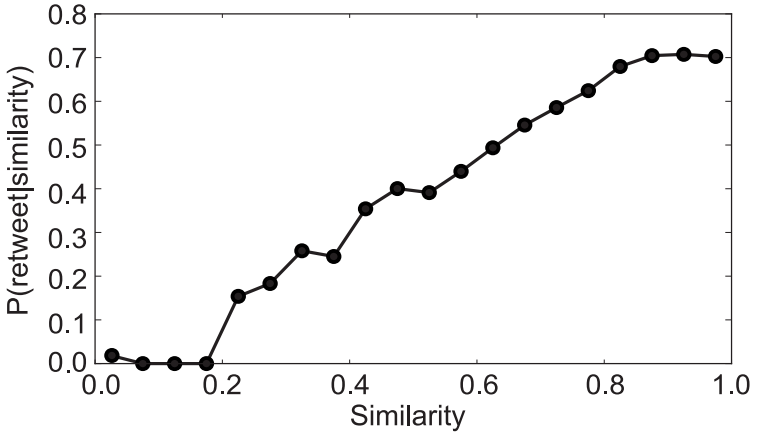


Figure 4.2: The relationship between the probability of retweeting a message and its similarity to user interests, inferred from prior posting behavior.

that u retweets a post with memes M given the similarity between the memes and her user interests. We turn to the Maximum Information Path (MIP) similarity measure [Markines et al., 2009, Markines and Menczer, 2009] that considers shared memes but discounts the more common ones:

$$sim_{MIP}(M, I) = \frac{2\log[\min_{x \in M \cap I} f(x)]}{\log[\min_{x \in M} f(x)] + \log[\min_{x \in I} f(x)]} \tag{4.1}$$

where x is a meme and $f(x)$ is the proportion of messages about x.

Fig. 4.2 shows that users are more likely to retweet memes about which they posted in the past (Pearson correlation coefficient $\rho = 0.98$). This suggests that memory is an important ingredient for guiding users' reposting activities in a model of meme competition, and we explicitly take this aspect into account in the model presented below.

4.3 Competition Model

We propose an agent-based model in this section to study the role of the limited attention of individual users in the diffusion process, and in particular whether competition for our finite attention may affect meme popularity, diversity, and lifetime. Although competition among ideas has been implicitly assumed as a factor behind, e.g., the decay in interest toward news and movies [Leskovec et al., 2007b, Wu and Huberman, 2007, Crane and Sornette, 2008], to the best of our knowledge nobody has attempted to explicitly model the mechanisms of competition and how they shape the spread of information. Particularly we show that a simple model of competition on a social network, without any further assumptions about meme merit, user interests, or explicit exogenous factors, can account for the massive heterogeneity in meme popularity and persistence.

4.3.1 Empirical Regularities

In Fig. 4.3 we observe several regularities in the empirical data. We first consider meme lifetime, defined as the maximum number of consecutive time units in which posts about the meme are observed; meme popularity, defined as the number of users per day who tweet about a meme, measured over a given time period; and user activity, defined as the number of messages per day posted by a user, measured over a time period. These three quantities all display long-tailed distributions (Fig. 4.3(a,b,c)). The excellent collapse of the curves demonstrates that the distributions are robust even if measured over different time units or observed over different periods of time. We further measure the breadth of user attention, defined earlier through the meme entropy. Although the entropy distribution is peaked, some users have broad attention while others are very focused (Fig. 4.3d). This distribution is also robust with respect to different periods of time.

All of these empirical findings point to extremely heterogenous behaviors; some memes are extremely successful (popular and persistent), while the great majority die quickly. A small fraction of memes therefore account for the great majority of all posts. Likewise, a small proportion of users generate most of the traffic.

These heterogeneities can in principle be attributed to a variety of causes. The broad distributions of meme popularity could result from a diversity in some intrinsic meme values, with "important" memes attracting more attention. Long-lived memes might be sustained exogenously by traditional media and real-world events. User activity and breadth of attention distributions could be a reflection of innate behavioral differences. What is, then, a minimal set of assumptions necessary to interpret this empirical data? One way to tackle this question is to start from a minimalist model of information spreading that assumes none of the above externalities. In particular we will explore to what extent the statistical features of memes and users can be accounted by the limited attention capacity of the users coupled with the heterogeneity of their social connections.

4.3.2 Model Description

Our basic model assumes a frozen network of agents. An agent maintains a time-ordered list of *posts,* each about a specific *meme*. Multiple posts may be about the same meme. Users pay attention to these memes only. Asynchronously and with uniform probability, each agent can generate a post about a new meme or forward some of the posts from the list, transmitting the corresponding memes to neighboring agents. Neighbors in turn pay attention to a newly received meme by placing it at the top of their lists. To account for the empirical observation that past behavior affects what memes the user will spread in the future, we include a memory mechanism that allows agents to develop endogenous interests and focus. Finally, we model *limited attention* by allowing posts to survive in an agent's list or memory only for a finite amount of time. When a post is forgotten, its associated meme become less represented. A meme is forgotten when the last post carrying that meme disappears from the user's list or memory. Note that list and memory work like first-in-first-

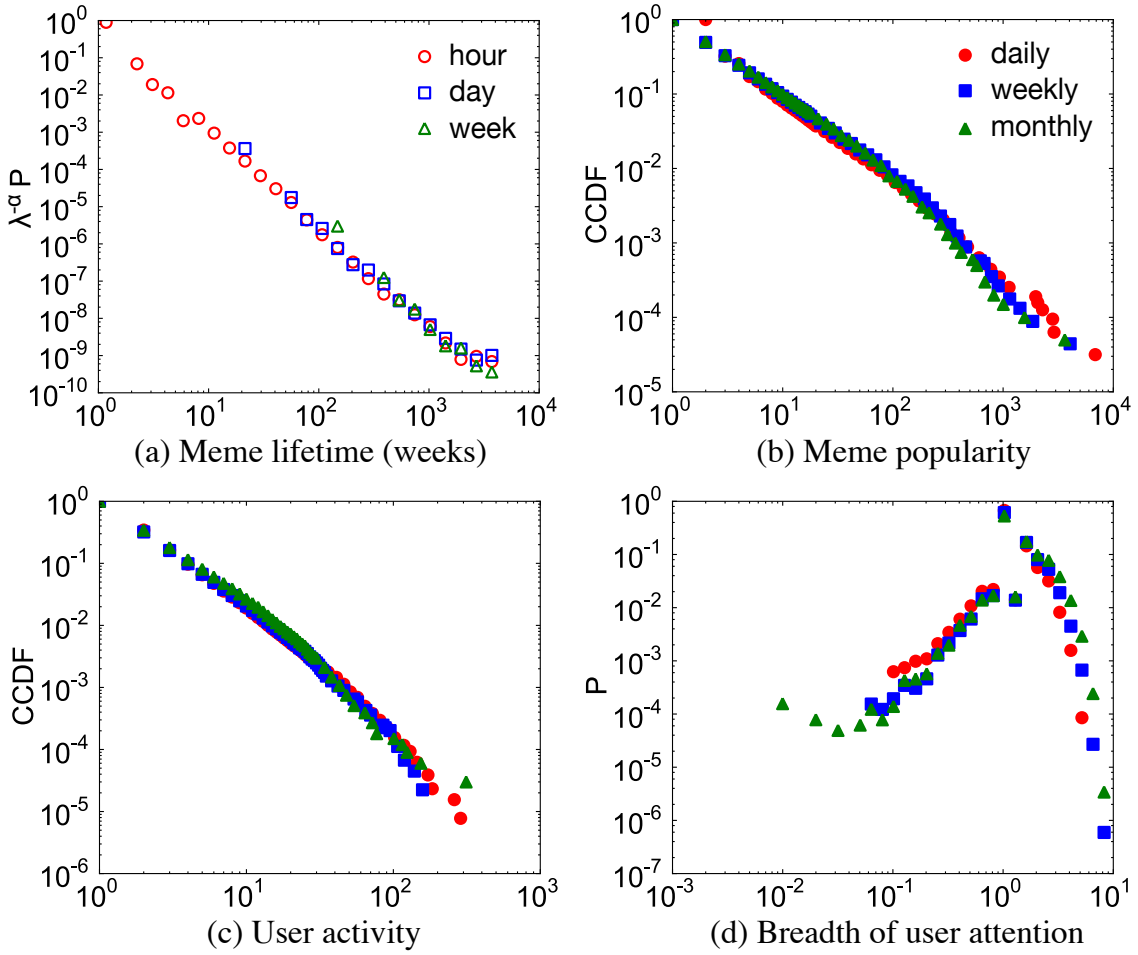


Figure 4.3: Empirical regularities in Twitter data. (a) Probability distribution of the lifetime of a meme using hours (red circles), days (blue squares), and weeks (green triangles) as time units. In the plot, units are converted into hours. Since the distributions are well approximated by a power law, we can align the curves by rescaling the y-axis by $\lambda^{-\alpha}$, where λ is the ratio of the time units (e.g., $\lambda = 24$ for rescaling days into hours) and $\alpha \approx 2.5$ is the exponent of the power law (via maximum likelihood estimation [Clauset et al., 2009]). This demonstrates that the shape of the lifetime distribution is not an artifact of the time unit chosen to define the lifetime. (b) Complementary cumulative probability distribution of the popularity of a meme, measured by the total number of users per day who have used that meme. This and the following measures were performed daily (filled red circles), weekly (filled blue squares), and monthly (filled green triangles). (c) Complementary cumulative probability distribution of user activity, measured by the number of messages per time unit posted by a user. (d) Probability distribution of breadth of user attention (entropy), based on the memes tweeted by a user. Note that the larger the number of posts produced, the smaller the non-zero entropy values recorded for users who focus on a small set of memes. This explains why the distributions for longer periods of time extend further to the left.

out rather than priority queues, as proposed in models of bursty human activity [Barabási and Albert, 2005].

The model is illustrated in Fig. 4.4. Agents interact on a directed social network of friends or followers. Each user node is equipped with a *screen* where received memes are recorded, and a *memory* with records of posted memes. An edge from a friend to a follower indicates that the friend's memes can be read on the follower's screen (#x and #y in Fig. 4.4a appear on the screen in Fig. 4.4b). At each step, an agent is selected randomly to post memes to neighbors. The agent may post about a new meme with probability p_n (#z in Fig. 4.4b). The posted meme immediately appears at the top of the memory. Otherwise, the agent reads posts about existing memes from the screen. Each post may attract the user's attention with probability p_r (the user pays attention to #x, #y in Fig. 4.4c). Then the agent either retweets the post (#x in Fig. 4.4c) with probability $1 - p_m$, or tweets about a meme chosen from memory (#v triggered by #y in Fig. 4.4c) with probability p_m. Any post in memory has equal opportunity to be selected, therefore memes that appear more frequently in memory are more likely to be propagated (the memory has two posts about #v in Fig. 4.4d). To model limited user attention, both screen and memory have a finite capacity, which is the time in which a post remains in an agent's screen or memory. For all agents, posts are removed after one *time unit*, which simulates a unit of real time, corresponding to N_u steps where N_u is the number of agents. If people use the system once weekly on average, the time unit corresponds to a week.

4.3.3 Parameter Tuning

The model has three parameters: p_n regulates the amount of novelty that enters the system (number of cascades), p_r determines the overall retweet activities (size of cascades), and p_m accounts for individual focus (diversity of user interests). We estimate all three parameters directly from the data (see Table 4.1). The parameter p_n characterizes the probability of tweeting about a new meme; to estimate p_n, we count new hashtags by examining whether hashtags have been observed in previous time units (weeks). The proportion of posts with new hashtags is approximately 0.45 ± 0.05, and thus we set $p_n = 0.45$ for all the simulations. For each simulation, the parameter p_r is tuned to capture the average number of posted memes per user per time unit. Finally, the parameter p_m represents the proportion of all memes tweeted by an individual that match the content of the memory; we compare each hashtag with those produced by a user in the previous time unit (week). We use the average value across all users (0.4 ± 0.01) and set $p_m = 0.4$.

4.3.4 Simulation Results

The underlying social network is a critical component of the model of meme diffusion. To obtain a network of manageable size while preserving the structure of the actual social network, we sample a directed graph with 10^5 nodes and about 3×10^6 edges from the Twitter follower network. The sampling procedure is a random walk with occasional

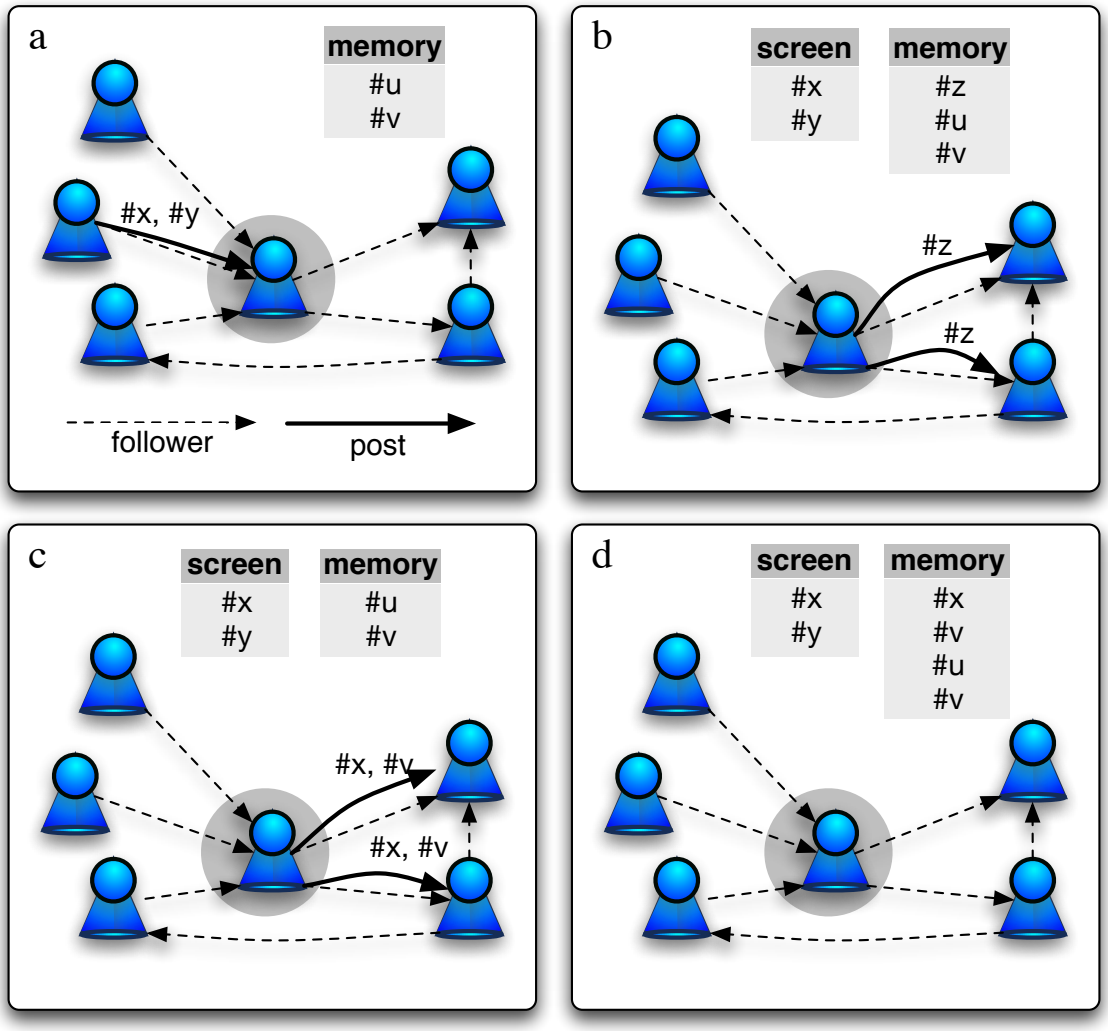


Figure 4.4: Illustration of the meme diffusion model. Each user has a memory and a screen, both with sizes limited by time. (a) Memes are propagated along follower links. (b) The memes received by a user appear on his screen. With probability p_n, the user posts a new meme, which is stored in memory. (c) Otherwise, with probability $1 - p_n$, the user scans the screen. Each meme x in the screen catches the user's attention with probability p_r. Then with probability p_m a random meme from memory is triggered, or x is retweeted with probability $1 - p_m$. (d) All memes posted by the user are also stored in memory.

Table 4.1: Parameter settings for different simulations. "Avg user activity" is the average number of posts per user per time unit.

Simulation	Network	t_w	p_r	Avg user activity
Standard	Sampled	1.0	0.016	2.5
ER Network	Synthetic	1.0	0.029	2.1
Weak Competition	Sampled	5.0	0.205	2.3
Strong Competition	Sampled	0.1	0.001	2.7

restarts from random locations (teleportation factor 0.15). Though no sampling method is perfect, the modified random walk is efficient in terms of API queries and reproduces the salient topological features of the sampled network [Leskovec and Faloutsos, 2006]. The empirical retweets generated by the users in the sample display trends similar to those from the entire dataset, therefore we expect the model predictions to be consistent not only with the sample but also with the full dataset.

To evaluate the predictions of our model, we compare them with empirical data that includes only the retweets of the same subset of users. To study the role played by the network structure in the meme diffusion process, we also simulate the model on a random Erdös-Rényi (ER) network with the same number of nodes and edges. As shown in Fig. 4.5, the model captures the main features of the empirical distributions of meme lifetime and popularity, user activity, and breadth of user attention. The comparison with the corresponding distributions generated using the ER network shows that in general, the heterogeneity of the observed quantities is greatly reduced when memes spread on a random network. This is not unexpected. Consider for example meme popularity (Fig. 4.5b); the real social network has a broad (scale free) distribution of degree, with a consistent number of hub users who have a large number of followers. Memes spread by these users are likely to achieve greater popularity. This does not happen in the ER network where the degree distribution is narrow (Poissonian). The difference observed in the distribution of breadth of user attention, for both low and high entropy values (Fig. 4.5d), may be explained by the heterogeneity in the number of friends. Users with few friends may have low breadth of attention while those with many friends are exposed to many memes and thus may exhibit greater entropy.

The second key ingredient of our model is the competition among memes for limited user attention. To evaluate the role of such a competition on the meme diffusion process, we simulate variations of the model with stronger or weaker competition. This is accomplished by tuning the length t_w of the time window in which posts are retained in an agent's screen or memory. A shorter time window ($t_w < 1$) leads to less attention and thus increased competition, while a longer time window ($t_w > 1$) allows for attention to more memes and thus less competition. As we can observe in Fig. 4.6, stronger competition ($t_w = 0.1$) fails to reproduce the large observed number of long-lived memes (Fig. 4.6a). Weaker competition ($t_w = 5$), on the other hand, cannot generate extremely popular memes (Fig. 4.6b) nor extremely active users (Fig. 4.6c).

We also simulate our model without user interests, by setting $p_m = 0$. The most noticeable difference in this case is the lack of highly focused individuals. Users have no memory of their past behavior, and can only pay attention to memes from their friends. As a result, the model fails to account for low entropy individuals (not shown but similar to the random network case in Fig. 4.6d).

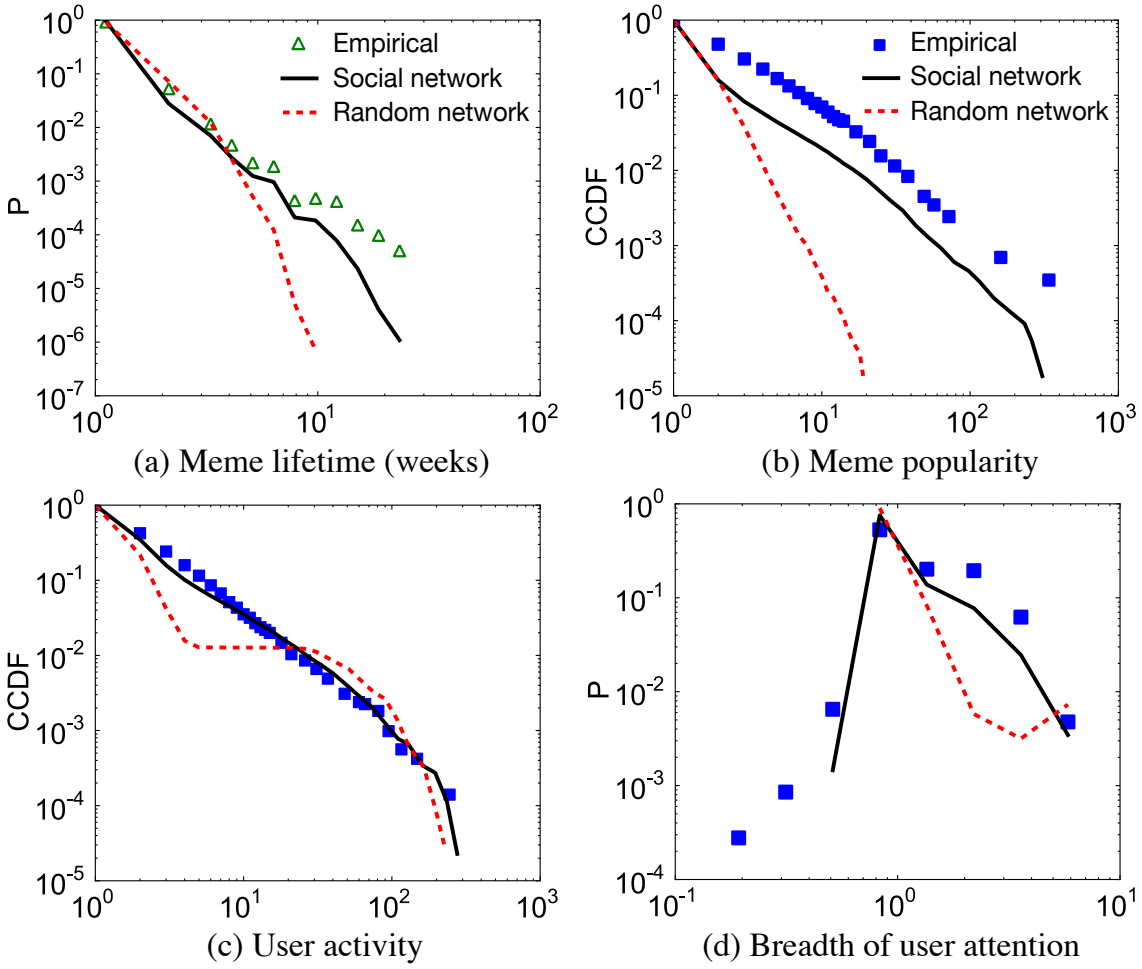


Figure 4.5: Evaluation of model by comparison of simulations with empirical data (same panels and symbols as in Fig. 4.3). To study the role played by the network structure in the meme diffusion process, we simulate the model on the sampled follower network (solid black line) and a random network (dashed red line). Both networks have 10^5 nodes and about 3×10^6 edges. (a) The definition of lifetime uses the week as time unit. (b,c,d) Meme popularity, user activity, and user entropy data are based on weekly measures.

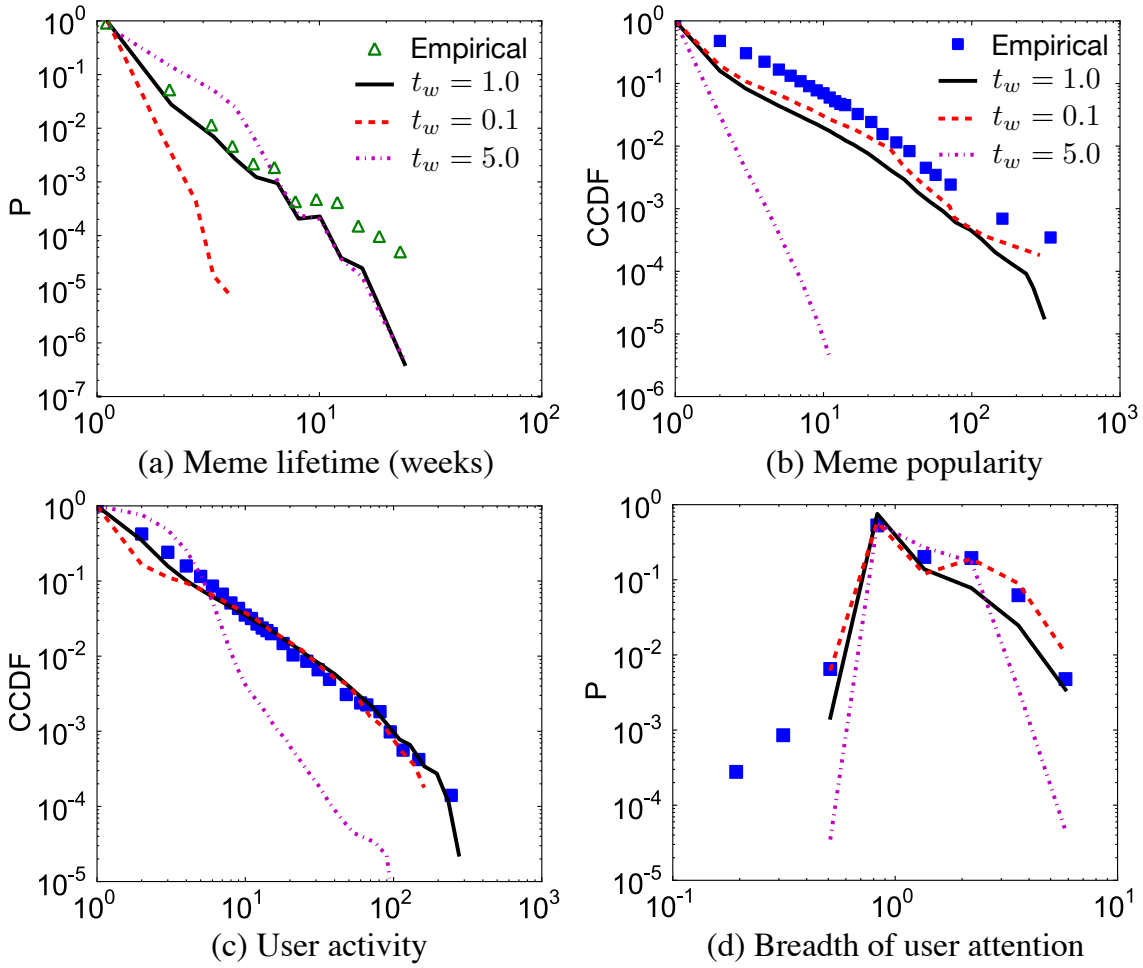


Figure 4.6: Evaluation of model by comparison of simulations with empirical data (same panels and symbols as in Fig. 4.3). To study the role of meme competition, we simulate the model on the sampled follower network with different levels of competition; posts are removed from screen and memory after t_w time units. We compare the standard model ($t_w = 1$, solid black line) against versions with less competition ($t_w = 5$, dot-dashed magenta line) and more competition ($t_w = 0.1$, dashed red line). (a) The definition of lifetime uses the week as time unit. (b,c,d) Meme popularity, user activity, and user entropy data are based on weekly measures.

4.4 Discussion

The present findings demonstrate that the combination of social network structure and competition for finite user attention is a *sufficient condition* for the emergence of broad diversity in meme popularity, lifetime, and user activity. This is a remarkable result: one can account for the often-reported long-tailed distributions of topic popularity and lifetime [Watts, 2002, Goetz et al., 2009, Ratkiewicz et al., 2010, Ienco et al., 2010] without having to assume exogenous factors such as intrinsic meme appeal, user influence, or external events. The only source of heterogeneity in our model is the social network; users differ in their audience size but not in the quality of their messages.

Although we have seen many models on reproducing the temporal evolution of popularity in the context of social media [Wu and Huberman, 2007, Crane and Sornette, 2008, Hogg and Lerman, 2009] and recent epidemiological models have started considering the simultaneous spread of competing strains [Sneppen et al., 2010, Karrer and Newman, 2011], our framework is the first attempt to deal with a virtually unbounded number of new "epidemics" that are continuously injected into the system. A closer analogy to our approach is perhaps provided by neutral models of ecosystems, where individuals (posts) belonging to different species (memes) produce offspring in an environment (our collective attention) that can sustain only a limited number of individuals. At every generation, individuals belonging to new species enter the ecosystem while as many individuals die as needed to maintain the sustainability threshold [Pigolotti et al., 2004].

Since Simon's seminal paper [Simon, 1971], the economy of attention has been an enormously popular notion, yet it has always been assumed implicitly and never put to the test. The model proposed in this chapter provides a first attempt to focus explicitly on mechanisms of competition, and to evaluate the quantitative effects of making attention more scarce or abundant.

Our results do not constitute a proof that exogenous features, like intrinsic values of memes, play no role in determining their popularity. However, we have shown that at the statistical level it is not necessary to invoke external explanations for the observed global dynamics of memes. This appears as an arresting conclusion that makes information epidemics quite different from the basic modeling and conceptual framework of biological epidemics. While the intrinsic features of viruses and their adaptation to hosts are extremely relevant in determining the winning strains, in the information world the limited time and attention of human behavior are sufficient to generate a complex information landscape and define a wide range of different meme spreading patterns. This calls for a major revision of many concepts commonly used in the modeling and characterization of meme diffusion and opens the path to different frameworks for the analysis of competition among ideas and strategies for the optimization and suppression of their spread.

Chapter 5

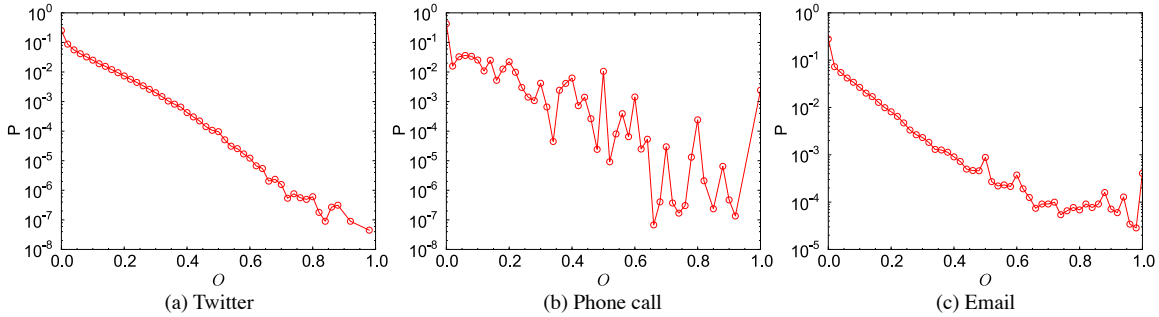
Attention on Weak Ties

In this chapter we continue the investigation into finite human attention, but from an individual viewpoint. The seminal paper "The strength of weak ties" by Granovetter [1973] defined the strength of social ties proportionally to the size of shared social circles. The more common friends two individuals have, the stronger the tie is between them. Based on this definition, the prominent weak tie hypothesis was proposed, according to which social ties of different strengths have distinct roles in the dynamics of social structure and information sharing [Granovetter, 1973, 1995].

In general, we are surrounded by a wide variety of socio-technical systems [Vespignani, 2009], yielding an abundant amount of information that exceeds our capacity to consume it. Thus attention has become an important and valuable resource that we share parsimoniously [Dunbar, 1998, Gonçalves et al., 2011, Backstrom et al., 2011, Hodas and Lerman, 2012, Weng et al., 2012]. Our interactions are steered more than ever before by the "economy of attention" [Simon, 1971, Davenport and Beck, 2001].

As Simon predicted in 1971, "Information consumes the attention of its recipients. Hence a wealth of information creates *a poverty of attention* and *a need to allocate that attention* efficiently among the overabundance of information sources" [Simon, 1971]. As the individual attention is limited, either strong or weak ties cannot acquire infinite attention. We believe that weak ties act as bridges between communities and thus as important channels for novel information that people rarely get from close social circles; hence, if there exists a need for allocating attention among information sources as Simon predicted, weak ties would attract much attention even though they do not carry much traffic.

Starting from these considerations, it is then natural to ask: *How is attention allocated among ties? Do people favor strong over weak ties? What are the factors behind one or the other tendency?* In this chapter, we investigate these questions with respect to the role of attention and attention allocation within Granovetter's hypothesis, using three large-scale datasets describing different types of human interactions, information sharing in online social media, cell phone calls, and email exchanges. We first quantify the *strength* of each social tie using the number of shared neighbors. The results verify the weak tie hypothesis by showing that the largest fraction of interactions happen on strong ties while weak ties



(a) Twitter (b) Phone call (c) Email

Figure 5.1: Probability distribution of link overlaps.

carry much less traffic [Granovetter, 1973, Onnela et al., 2007b]. A further step is made in scrutinizing the role of *attention* and its relationship with tie strength. We invent a measure of attention based on the assumption that each user has an finite amount of attention, proportional to how active the user is [Weng et al., 2012], to distribute to information sources and preserve social contacts. Interestingly, we find that only very weak or very strong ties attract a good amount of attention, implying two potentially competing trends. On one hand, people frequently interact with close friends due to their social needs; on the other hand, people look for novel and useful information through weak ties driven by the demand for information, in accordance with Granovetter and Simon. The former assigns more attention to strong ties, while the latter prefers weak ones. The relative strength of these two tendencies vary according to the mechanisms of the systems.

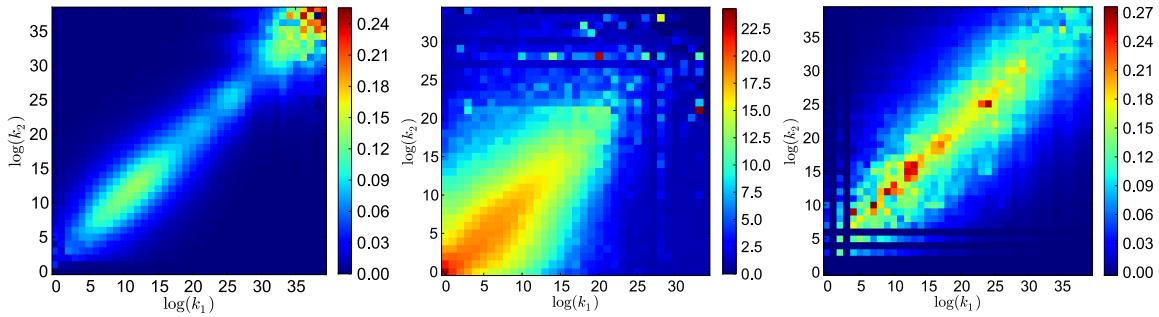
5.1 Tie Strength, Weight, and Attention

In consonance with Granovetter's theory, we measure the *tie strength*—the closeness between two users i and j—as the Jaccard overlap between their friend sets [Granovetter, 1973, Onnela et al., 2007b]:

$$O_{ij} = \frac{|N_i \cap N_j \setminus \{i,j\}|}{|N_i \cup N_j \setminus \{i,j\}|} \quad \text{and} \quad N_i = \{u \mid (i,u) \in E \wedge (u,i) \in E\} \tag{5.1}$$

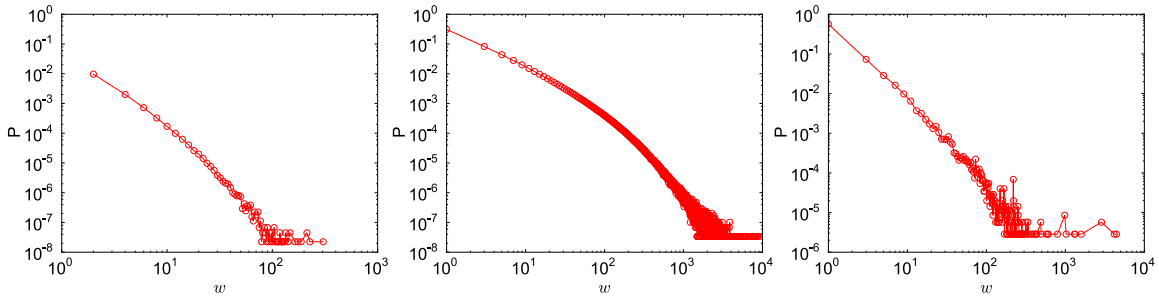
where N_i and N_j are the set of neighbors of nodes i and j, respectively, which are evaluated independently of the link directions. We also refer to tie strength as link *overlap* in the sequent discussion. In Fig. 5.1 we plot the probability distribution of link overlaps in three datasets and all of them present broad distributions: most ties are weak with little overlap, while only a very small fraction of ties are strong.

Figure 5.2 illustrates heat maps of the link overlap as a function of degrees of two nodes connected by the link, in which all three networks reveal strong assortative mixing patterns [Newman, 2002a]. In Twitter, high link overlap is more likely to appear between two high-degree nodes. In the cell phone call network, ties between inactive users tend to have higher strength. In the Enron email network, people send emails more frequently to others with whom they are of similar activities.



(a) Twitter (b) Phone call (c) Enron email

Figure 5.2: Heat maps of tie strength (link overlap) as a function of the source node with degree k_1 and the target node with degree k_2 in (a) Twitter, (b) cell phone network and (c) email network. Note that the degree is the sum of in-degree and out-degree, i.e. the number of neighbors of a given node irrespective of direction.



(a) Twitter (b) Phone call (c) Email

Figure 5.3: Probability distribution of link weights.

The intensity of communication on a tie (i, j) is quantified by the total frequency of i reposting, calling, or emailing j, denoted as link *weight* w_{ij}. Figure 5.3 shows power-law distributions of link weights, suggesting that in all three networks, Twitter follower, phone call, and email networks, a large majority of links carry little traffic but a few can attract extremely high volume of interactions.

Attention is the core concept in this chapter and we measure this quantity by considering the fraction of communication directed to a link with respect to all the activities started by the individual. According to how active a user is in the network, the amount of individual attention is varying among people but finite [Gonçalves et al., 2011, Hodas and Lerman, 2012]. As demonstrated in Fig. 5.4(a-c), higher out-degree indicates more activities (i.e., follow a large number of Twitter users, call many people, or send emails to many others), when the out-degrees are relatively small; however, the amount of activities seem to approach saturation quickly given a range of large out-degrees—active users have more attention but the total amount is bounded. Such an effect in the email network is less noticeable due to a large variation. The total activities of an individual can be approximated by a linear dependence of her out-degree in logarithm scale (see Fig. 5.4(d-f)), and therefore we have:

$$\text{Amount of attention of user } i \propto \log k_{\text{out}}(i) \tag{5.2}$$

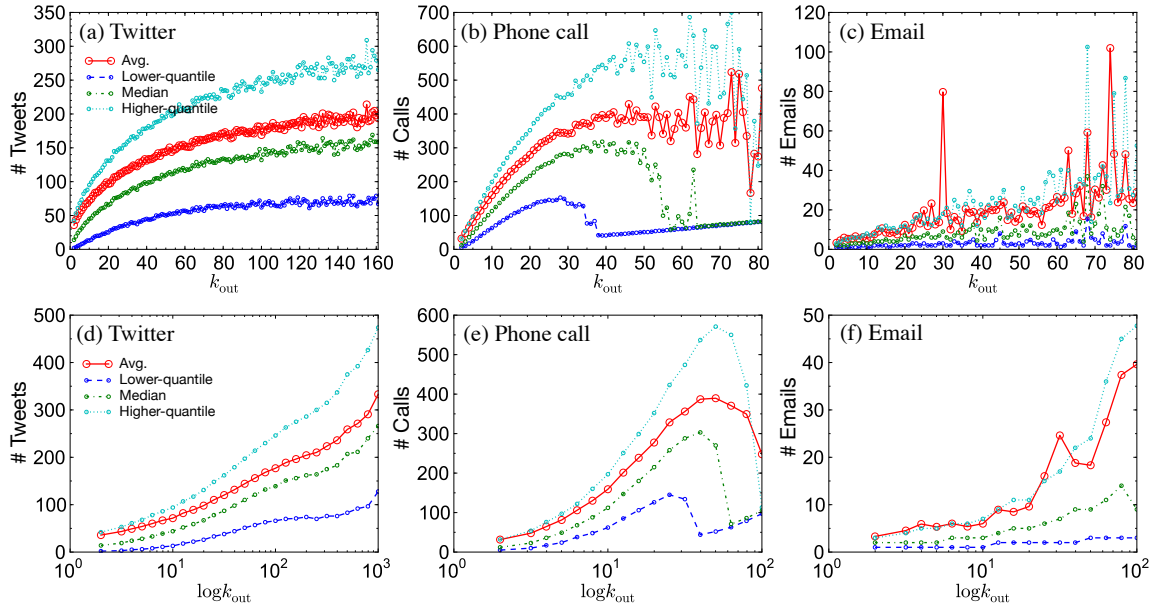


Figure 5.4: The total activities of an individual as a function of the user's out-degree in logarithm.

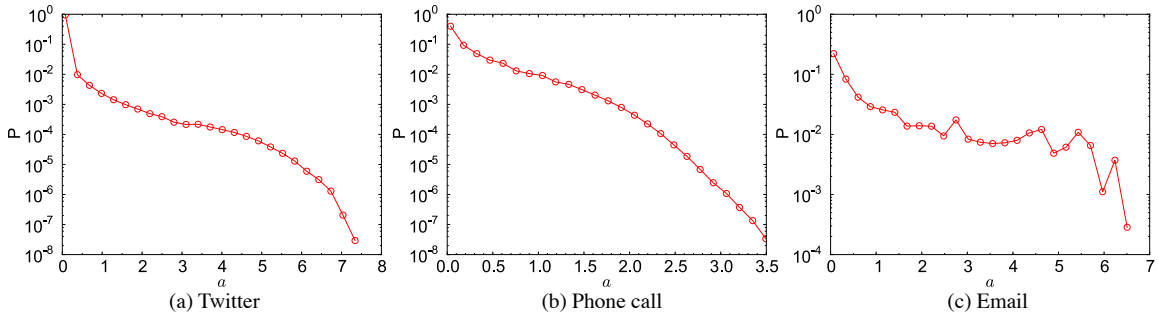


Figure 5.5: Probability distribution of link attention.

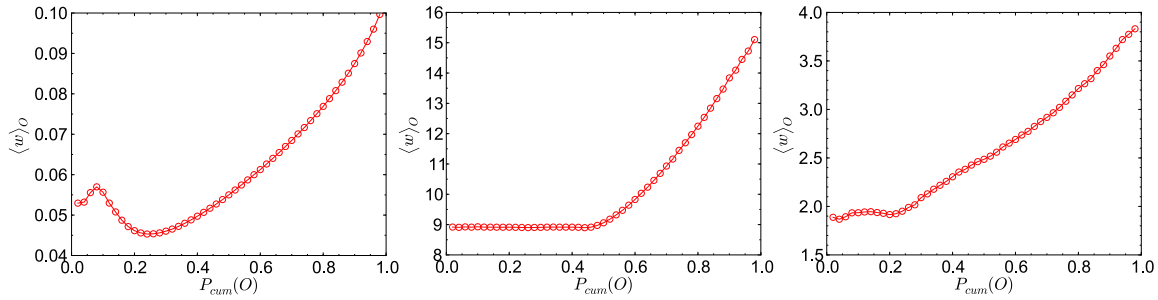
Finally the *attention* assigned to a link (i, j) as a portion of the total amount of attention of a user i is computed as:

$$a_{ij} = \log k_{\text{out}}(i) \cdot \frac{w_{ij}}{\sum_{u \in N_i} w_{iu}} \tag{5.3}$$

Attention on social ties also has a broad distribution as shown in Fig. 5.5.

5.2 Weak Ties Hypothesis and the Role of Attention

The weak tie hypothesis claims that strong ties carry a majority of interactions, while weak ties act as bridges and indispensable channels for transferring novel informations. We aim to enrich such a characterization of weak ties in the process of information diffusion with respect to attention.



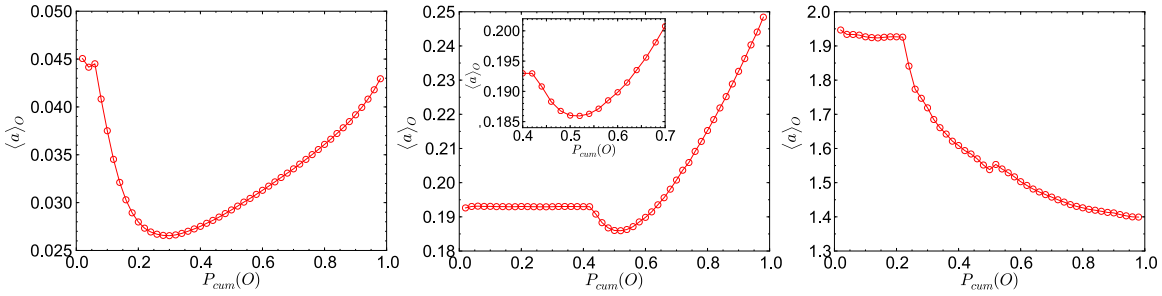
(a) Twitter (b) Phone call (c) Email

Figure 5.6: The average weight, $\langle w \rangle_O$, of first $P_{cum}(O)$ percent links which are sorted by link overlap. The quantity $P_{cum}(O)$ represents a fraction of links with smallest tie strength and the corresponding link weight $\langle w \rangle_O$ is computed averaging weights of those links.

5.2.1 Traffic on Strong Ties

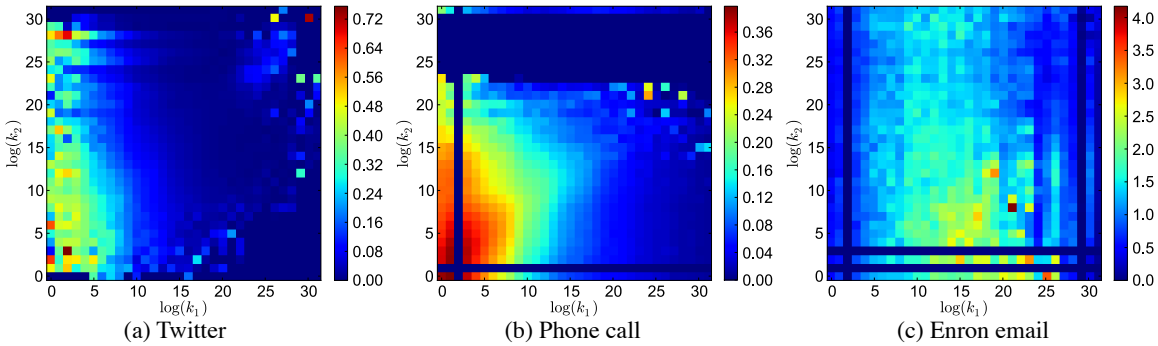
As a first step, we start with the verification of Granovetter's theory by plotting the average link weight, $\langle w \rangle_O$, as a function of the cumulative link overlap $P_{cum}(O)$. The quantity $P_{cum}(O)$ represents a fraction of links with smallest tie strength and the corresponding link weight $\langle w \rangle_O$ is computed averaging weights of those links [Onnela et al., 2007b]. As shown in Fig. 5.6, the average link weights in three datasets increase as a function of cumulative tie strengths in general. The observed pattern validates the weak tie hypothesis and several previous empirical studies [Friedkin, 1980, Onnela et al., 2007b,a, Cheng et al., 2010, Grabowicz et al., 2012, Pajevic and Plenz, 2012]. Strong ties carry more traffic than weak ties, confirming that people tend to communicate more with close friends, or others with whom they share very similar social circles. The emerging plateaus of the average curves at the beginning are due to links with zero overlap which are in random order (5.5% of links with zero overlap in Twitter; 40% in the cell phone network; and 23.6% in the Enron email network).

It is important to stress the diversity of the datasets considered. Indeed, they describe very different human activities and in general they could be subject to different dynamics, but the tie creation mechanisms and evolution processes are similar. In the case of Twitter, the result implies that users are more likely to adopt and repost messages from neighbors with similar circles of connectivities. While in the phone call network, people tend to call individuals with very similar contact lists more frequently. In the email network, people working in the same or close divisions of the corporation and thus sharing many common coworkers have many more email exchanges. Overall the emerging picture suggests that despite their diverse forms of interactions, the dynamics of human communication in these systems is driven by the same forces, aligned with the weak tie hypothesis.



(a) Twitter (b) Phone call (c) Email

Figure 5.7: The average link attention $\langle a \rangle_O$ as a function of the cumulative tie strength $P_{cum}(O)$.



(a) Twitter (b) Phone call (c) Enron email

Figure 5.8: The heat maps of the amount of attention allocated on a link as a function of degrees of two nodes connected by this link.

5.2.2 Attention on Weak Ties

The ever increasing amount of stimuli, generated by the wide variety of socio-technical systems we are part of, are pushing our limits. Indeed, we are all characterized by finite cognitive capacities due to biological and environmental constrains that force us to carefully allocate our limited attention among our connections and activities [Simon, 1971, Dunbar, 1998, Weng et al., 2012]. To investigate the role of attention in communication with different ties, we compute the average link attention $\langle a \rangle_O$ as a function of the cumulative overlap $P_{cum}(O)$, defined similarly as in Fig. 5.6. As reported in Fig. 5.7, interestingly, while the three datasets are almost indistinguishable in the analysis of the behavior of link weights, they show crucial differences on the distribution of attention among weak ties. The attention distribution reveals a 'U'-shape curve in the Twitter network—a positive correlation between attention and overlap for strong ties but a negative correlation for weak ties, suggesting that people are equally likely to allocate much attention on very weak ties as on very strong ones. In fact, we do observe that low-degree users pay attention to both people with low degree and celebrities with high degree (see Fig. 5.8a). The 'U'-shape is less evident in the phone call network. Except for the starting plateaus resulted from links with zero overlaps, weak ties acquire attention slightly more than intermediate ties while the majority of attention is assigned to strong connections. However, the trend is reversed

in the Enron email network where weak ties are dominant in attracting attention and there is a negative correlation between the amount of attention per link and the tie strength.

In fact, the 'U'-shape curve can be potentially interpreted by the emergence of two competing trends: on one hand, people are actively maintaining their social relationships by frequent interactions with close friends, so that strong ties capture much attention; on the other hand, people are approaching novel and useful information by paying attention to weak ties and thus weak ties are able to obtain a fair amount of attention. In Twitter, people communicate with close friends and, in the meantime, they also follow other important information sources although the tie is weak. Hence we can observe a combined effect of tendency for preserving strong ties and leaning towards weak ones. Different from Twitter, in the phone call network people often make calls to their closest social contacts, which accounts for assigning much attention to strong ties. People do occasionally call weak connections such as consumer service hotlines but it cannot be prevalent. In contrast, the email exchanges happen within a corporation and therefore the system is expected to be much more information-driven. The tendency for maintaining social relationships would be hardly noticed, leading to little attention on strong ties. The distinction between information-driven and social-driven communication will be further explored in the next section.

This finding has several implications. First, attention plays an important role in the dynamics of social networks. Second, the concentration of attention on weak ties is well aligned with Granovetter's hypothesis where weak ties are key ingredients in information transmission. Though the underlying mechanism of a network determines the extent to which weak ties may succeed to attract attention; weak ties in a system designed for more information-driven purposes may be more capable for channeling information and collecting attention. Finally, attention potentially distinguishes two competing trends in ties formation, one driven by social forces and the other instead by information gathering. The last point is examined in the following section.

5.3 Social and Information Links

Attention is concentrated among very weak or very strong ties, as seen in Fig. 5.7. A possible explanation of this observed pattern is the existence of two different, potentially competing, tendencies in communication, one for keeping social bonds and the other for novel information. In order to test this hypothesis, let us first look into the mechanisms of link formation in the three networks, respectively.

Micro-blogging system like Twitter, Weibo, and Google+ has several fundamental differences from offline social networks. First, people build up directed connections freely instead of mutual-trust-based reciprocal links. Second, the system is designed for efficient information sharing, not only for maintaining mutual friendships. Other than communicating with real-world friends, many users in the micro-blogging platforms do follow unknown but interesting others, such as celebrities, musicians, politicians, experts in tech fields, and

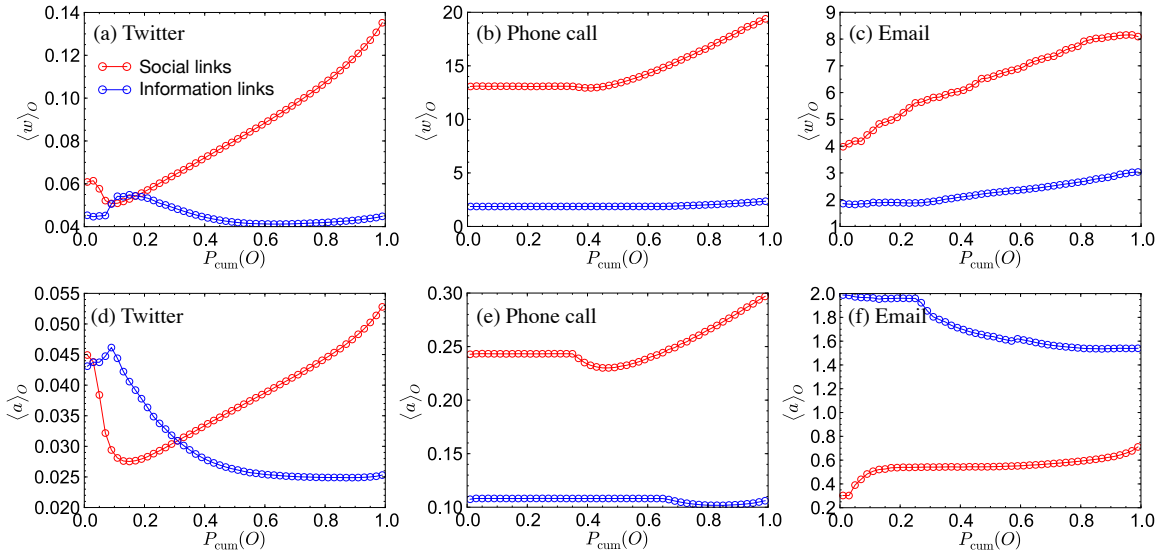


Figure 5.9: Social links versus information links in terms of the distributions of link weight and attention in three networks.

business accounts, to get updates. Owing to this special mechanism in the micro-blogging systems, [Huberman et al., 2009] distinguished friends from followers based on the number of "@"-interactions and pointed out that friends are the crucial underlying social networks that convey most traffic.

Similar phenomenon can be found in the phone call network. Real-world friends frequently talk to each other on phone and the interactions are usually intensive, mutual, and long-lasting. Meanwhile, business hotlines and customer services get calls from individual callers once or a few times, and the tie between them would be expected to be weak and non-mutual.

In the Enron email network, all the emails are supposed to be business- or information-driven, and therefore the social force is much weaker than in Twitter or phone call network. The intensity of collaboration on a tie would be dependent on how much overlap two individuals have in work, and these routine email exchanges are more likely to go through both directions. However, cross-division communication on a weak tie is expected to be more crucial and of higher priority (i.e., announcement from the board), thus obtaining more attention though maybe not mutual.

The social relationship between real-world friends is expected to be different from one between unknown people or coworkers (i.e., a Twitter user following a celebrity, a consumer calling business a hotline, or two coworkers with no contact in personal lives). The former aims at maintaining existing social ties, while the latter targets at information gathering. We therefore name the former as the *social link* and the latter as the *information link*. A simple way to distinguish these two types of ties is based on whether the link is mutual: for a given link $(i, j) \in E$, if i and j are linked (i.e., by following in Twitter, initiating at least one phone call, or sending at least one email) to one the other, (i, j) is deemed to be *so-*

cial; otherwise, it is an *information* link [Huberman et al., 2009, Granovetter, 1973, Onnela et al., 2007b, Karsai et al., 2011]. Using this denotation we then compute the average link weight and attention as a function of the cumulative link overlap, for social and information links, separately. Interestingly, as shown in in Fig. 5.9, we observe clear distinctions between these two types of links in terms of the allocation of both traffic and attention. More importantly, the distinctions can well interpret the differences between the behaviors of three networks in the distribution of attention (see Fig. 5.7).

Let us start with a discussion of link weights in Fig. 5.9(a-c). In all three networks, *social* links have larger weights than information ones, irrespective of tie strengths; meanwhile, their average weights increase with the cumulative tie strengths. The average weights of *information* links instead show different trends in three networks, suggesting that the mechanism behind their formation is different. Fig. 5.9(d-f) display the attention distributions on social ties of different natures. Among *social* links, strong ties attract more attention than weak ones. Among *information* links, weak ties are more appealing with regards to attention. Furthermore, considering that links with zero overlap serve a special topological role—a "perfect bridge"[1] connecting distant groups, we expect to see more zero-overlap ties among information links than among social ones. In Twitter, there are 7.5% of information links with zero overlap, compared with 4.4% of social links; in the phone call network, about 65% of information links have overlap zero but only about 40% among social links; this effect is the strongest in the email network, 27.5% of information links have overlap zero but only 4.1% for social links. The final noteworthy point is that the distinctions between information and social links in terms of attention allocation can potentially decipher the differences between the different patterns observed in Fig. 5.7. The Twitter network allows users to maintain social contacts and information sources at the same time, and therefore the volumes of attention on information and social links are comparable. The phone call network is more commonly used for social purposes, so information links only win little attention overall. The email exchanges in the Enron corporate are designed for gaining information and processing business issues, thus making information links much more dominant.

Information links are formed for efficient information spread, not for social relationship maintenance. As a consequence, an information link with high overlap is not necessarily to be "strong" in the social sense, due to the unbalanced communication. Furthermore, whether information links or social links can dominate the system is determined by the underlying mechanisms of the network.

5.4 Conclusion

In this chapter, we verify the weak tie hypothesis [Granovetter, 1973] on various large empirical networks. We use attention—the fraction of an individual's total attention directed to a given link—to quantify the importance of a social tie in information diffusion. While

[1]Note that in our calculation, all the leave nodes (nodes with only one out-link) are removed.

strong ties do carry more traffic, weak ties succeed to attract quite amount of attention similar to or even more than strong ties, suggesting that people do listen to weak ties often, for instance, for gathering innovative information from distant groups. The extent to which weak ties acquire attention is determined by the underlying mechanism of the network. By distinguishing social and information links based on link mutuality, we observe two potentially competing trends: on one hand, people interact with close friends to maintain close *social* relationships; on the other hand, people look for novel and useful information through weak ties driven by their needs for valuable *information*. In a system that is designed for information-driven communication (i.e., corporation email network), information links are dominant and assigned with more attention; however, in a system aimed for more social-driven communication (i.e., cell phone call network), social links yield more attention.

Part II

Content: Topic Space

Can we detect topics in social media?
Do topical diversity effect user and content popularity?
Do people discuss different topics with strong and weak ties?

Chapter 6

Topical Diversity

The characteristics of transmissible content is another essential factor in shaping the information diffusion process. The theme and generality of a message would affect the willingness of people to view and broadcast it, as well as to decide with whom to share it. In this chapter, we investigate the semantic space of online information and correlate conversation topics with user and content popularity. We particularly look into questions like whether users who comment on a variety of matters are more likely to achieve high influence than those who delve into one focused field and whether general Twitter hashtags, such as #lol, tend to be more popular than novel ones, such as #instantlyinlove. These questions demand a way to detect topics hidden behind messages associated with an individual or a hashtag, and a gauge of similarity among these topics.

In a social media site like Twitter, the social network topology is determined by how people are following each other. Each individual is represented as a node and each following relationship as an edge linking a pair of users (see the bottom layer in Fig. 6.1). Messages in social media involve a variety of *topics*. Content, messages, or ideas are deemed semantically similar if they discuss, comment, or debate about the same topic; conversely, we can detect a topic by clustering a group of similar messages observed. Hashtags spread among people through connections in the social network layer and can be mapped into a semantic space, in which each node is a tag and similar ones are coupled forming topic clusters (see the top layer in Fig. 6.1). By examining which topics are attached to a user's messages, we can infer her interests; by examining the topics of tags that co-occur with a given hashtag, we can learn what that hashtag is about. In reality we are able to observe the social network structure and information diffusion flows, but not topic formation in the semantic space. To the best of our knowledge, the connection between these two layers of information diffusion is not yet well explored [Serrano et al., 2008, Romero et al., 2013].

Here we propose an approach to identify clusters of similar hashtags by detecting communities in the hashtag co-occurrence network. Then the topical diversity of one's interests is quantified by the entropy of her hashtags across different topic clusters. A similar measure is applied to hashtags, based on co-occurring tags. We find that high topical diversity of early adopters or co-occurring tags implies high future popularity of hashtags. In contrast,

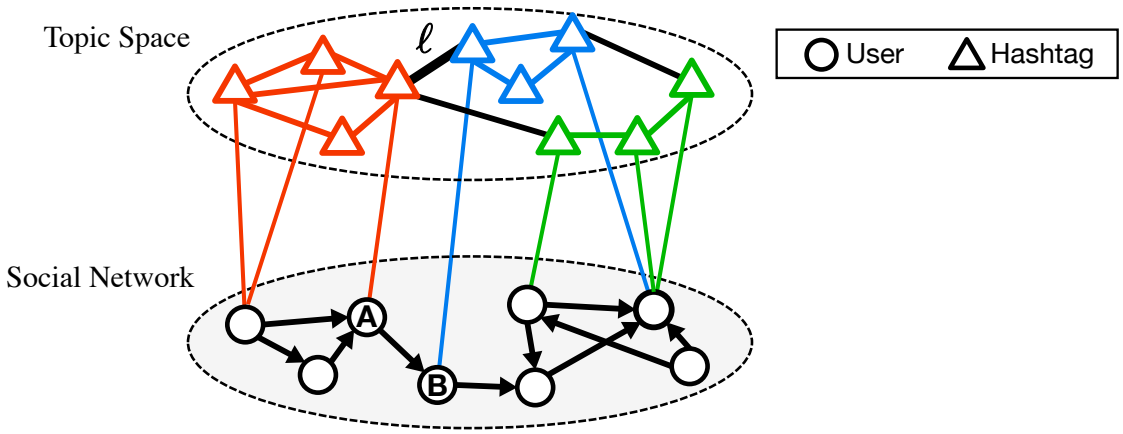


Figure 6.1: We can represent the topics of online conversations in social media by a multi-layer network. The social network connects people. In the topic network, nodes represent hashtags that are linked when they co-occur; clusters represent topics (shown in colors). A person and a hashtag are connected when the person uses the hashtag.

low diversity helps an individual accumulate social influence. In short, diverse messages and focused messengers are more likely to gain impact.

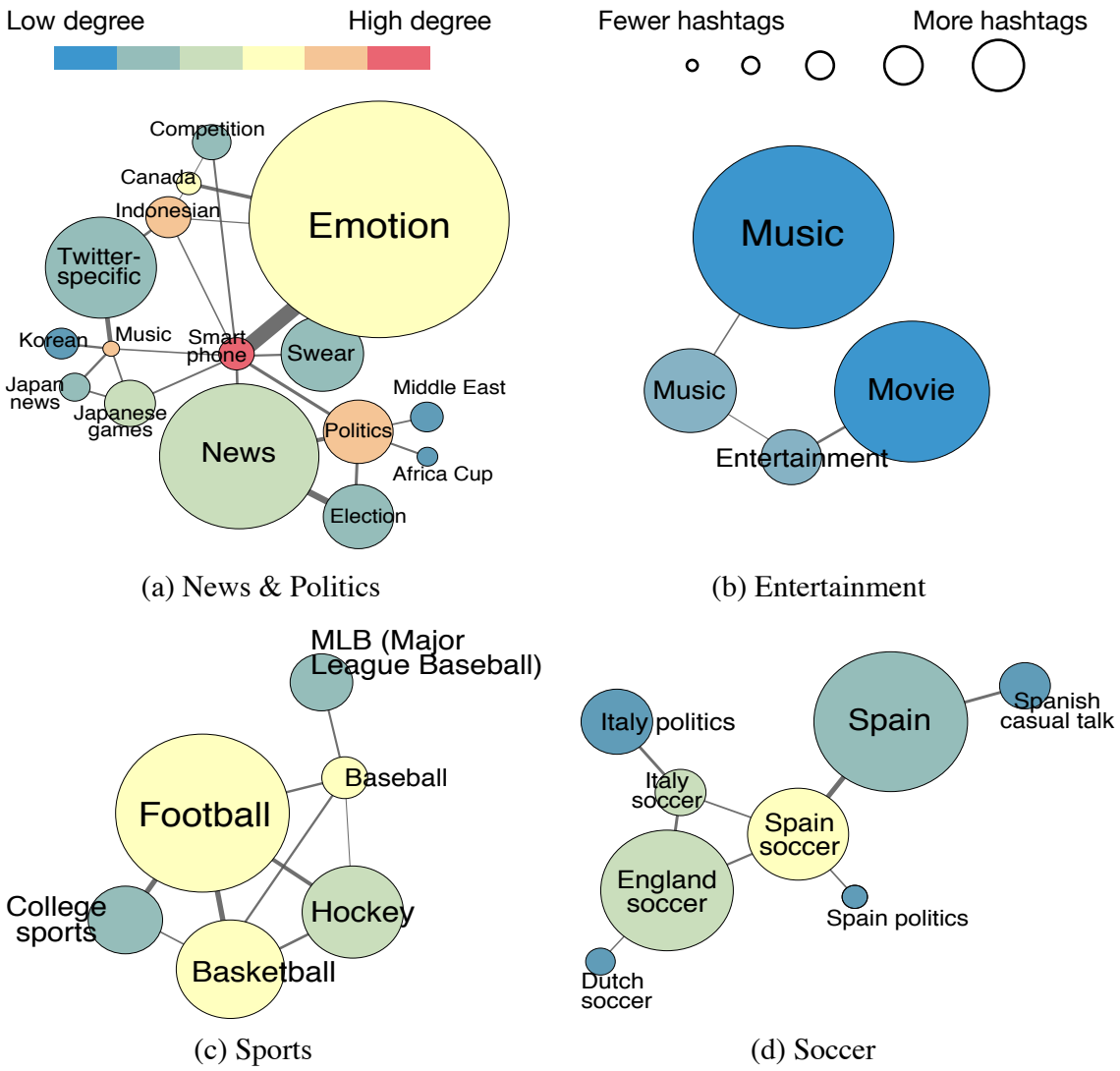
6.1 Definitions

In this section we define several key concepts to facilitate the subsequent discussion. We analyzed a dataset of public tweets from January to March 2013. We set the first two months of our dataset as the *observation period* and the last month as the *test period*; the former is used to build up the topic network and quantify user topical interests, and the latter works for evaluating the results of prediction tasks.

6.1.1 Topic Clusters

Hashtags are explicit topic identifiers on Twitter that are invented autonomously by millions of content generators. Since there is no predefined consensus on how to name a topic, multiple duplicate hashtags may be developed to represent the same event, theme, or object. For instance, #followback, #followfriday, #ff, #teamfollowback, and #tfb are all about asking others to follow someone back or suggesting people to follow; #tcot, #ttxcot, #twcot, and #ccot label politically conservative groups on Twitter. To reduce the duplication, we shift attention from single hashtags to more general categories—clusters of semantically similar hashtags—that we call *topic clusters*.

With the topic locality assumption that semantically similar hashtags are more likely to appear in the same tweets together, such topic clusters are expected to be densely connected.



(a) News & Politics

(b) Entertainment

(c) Sports

(d) Soccer

Figure 6.2: Examples of connected topic clusters of related themes: (a) news and politics, (b) sports, (c) soccer, and (d) music and entertainment. Each node represents a cluster of hashtags on the topic as labelled; the area is proportional to the number of hashtags that the topic cluster contains; the color is assigned according to the degree so that high degree is more red and low degree is more blue. All these examples support the existence of topic locality, even at the macroscopic level.

Table 6.1: Examples of topic clusters in the hashtag co-occurrence network.

Semantics	Example hashtags
Technology	#google, #microsoft, #supercomputers, #ibm, #wikipedia, #pinterest, #startuptip, #topworkplace
Politics	#tcot, #p2, #top, #usgovernment, #dems, #owe, #politics, #teaparty
Lifestyle	#pizza, #pepsi, #cheese, #health, #vacation, #caribbean, #ford, #honda, #volkswagen, #hm, #timberland
Twitter-specific	#followme, #followback, #teamfollowback, #followfriday, #friday, #justretweet, #instantfollower, #rt
Mobile devices	#apple, #galaxy3, #note2, #iosapp, #mp3player, #releases

We detect these clusters by finding communities in the hashtag co-occurrence network. First we recover the network by only considering hashtags used by at least three distinct users and join occurrences observed in at least three messages. We do this to filter out noise from accidental co-occurrence and spam. The recovered network contains 974,529 nodes and 7,325,492 edges. Then communities are detected using the Louvain community detection method [Blondel et al., 2008], which was selected because of its efficiency. We obtain 37,067 communities (the level 2 in the hierarchical structure found by the Louvain method). As exemplified in Table 6.1, communities in the hashtag co-occurrence network capture coherent topics. At the macroscopic level we can still observe strong topic locality (see Fig. 6.2).

6.1.2 Diversity of User Interests

Given a messenger u, we can track the sequence of hashtags (with repetition) that he used in the past, $h_1, h_2, \ldots, h_{n_u}$. Each hashtag h_i is attached to a topic $T(h_i)$, given by:

$$T(h_i) = \begin{cases} C(h_i) & \text{if } h_i \text{ exists in the hashtag co-occurrence network} \\ h_i & \text{otherwise} \end{cases} \tag{6.1}$$

where $C(h)$ is a community containing h in the hashtag co-occurrence network. The set of distinct topics associated with all of u's hashtags is denoted as $\mathbb{T}_u$, $T(h_i) \in \mathbb{T}_u$. The topical diversity of a user's interests can be estimated by computing the entropy of hashtags across topics:

$$H_1(u) = -\sum_{T_j \in \mathbb{T}_u} P(T_j) \log P(T_j) \tag{6.2}$$

$$P(T_j) = \frac{1}{n_u} |\{h_i | T(h_i) = T_j, 1 \leq i \leq n_u\}|. \tag{6.3}$$

Table 6.2 compares two people, both having used ten distinct hashtags for twenty times. User A was interested in trendy Twitter-specific tags almost exclusively (low H_1), while user B paid attention to a set of very diverse conversations about countries, movies, books,

Table 6.2: Comparison of two users with different diversity of topical interest.

User	C	Hashtag (usage count)
A	20	#nowplaying(1)
	96	#rt(4), #follow(3), #tfb(2), #ff(2), #500aday(2), #teamfollow(2), #teamfollowback(2), #f4f(1), #rt2gain(1)
		$n_A = 20,\ \lvert\mathbb{T}_A\rvert = 10,\ H_1(A) = 0.2864$
B	9	#australia(1)
	20	#cosmicconsciousness(1), #thenotebook(1)
	33	#thedescendants(1)
	57	#friendswithbenefits(1)
	79	#thepowerofnow(2)
	139	#gemini(8), #geminis(2)
	806	#tdl(2)
	–	#tipfortheday (1)
		$n_B = 20,\ \lvert\mathbb{T}_B\rvert = 10,\ H_1(B) = 2.3610$

and horoscope (high H_1). Note that the opposite (and wrong!) conclusion, $H_1 > H_2$, would be drawn had we measured entropy based on hashtags rather than topic clusters.

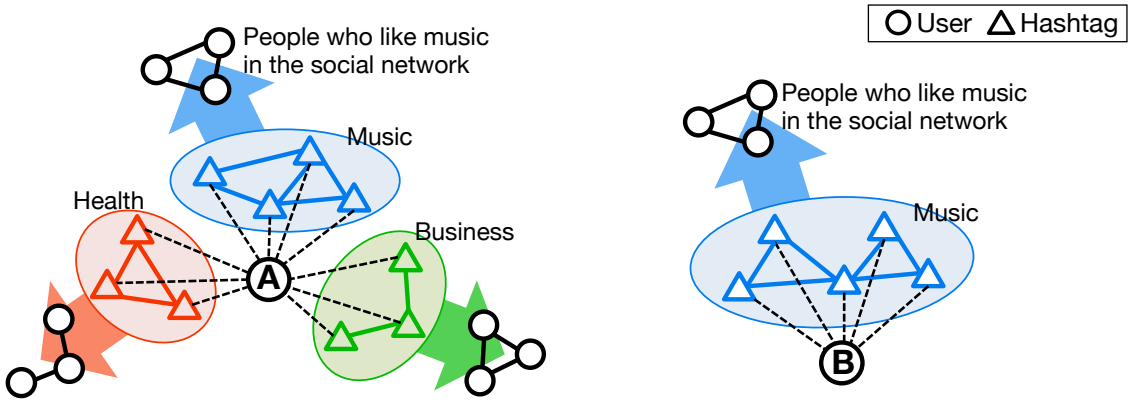
6.1.3 Diversity of Content

Similarly, given a hashtag h, we recover the sequence of other hashtags (with repetition) that co-occurred with it, $h_1, h_2, \ldots, h_{m_h}$. Each co-occurring hashtag (*co-tag*) h_i is assigned to topic $T(h_i)$ based on the topic cluster to which it belongs (see Equation 6.1). Then the co-tag diversity of h, $H_2(h)$, is measured in the same way as the user diversity H_1 (see Equation 6.2).

6.2 Predicting Hashtag Popularity

Do diversity measures help us detect hashtags that will go viral in the future? In this section we explore whether the topical diversity of a hashtag's adopters or co-tags predicts its future popularity.

Hashtags during the test period are used for prediction tasks. We are interested in newly emergent tags, so that we are able to identify the start of their lifetime and track their growth for at least three weeks. We select hashtags that do not appear during the observation period, but are used by at least three distinct users during the test period. In addition, only tags with the first tweet observed during the first week of March are considered, so that we can track their usage during the whole month. Eventually, about 3.03% of all hashtags in the test period are chosen as *emergent* hashtags.



(a) User A with diverse interests

(b) User B with focused interests

Figure 6.3: (a) User A has diverse topical interests, each connected group corresponding to a social circle with common interests. (b) User B displays more focused interests.

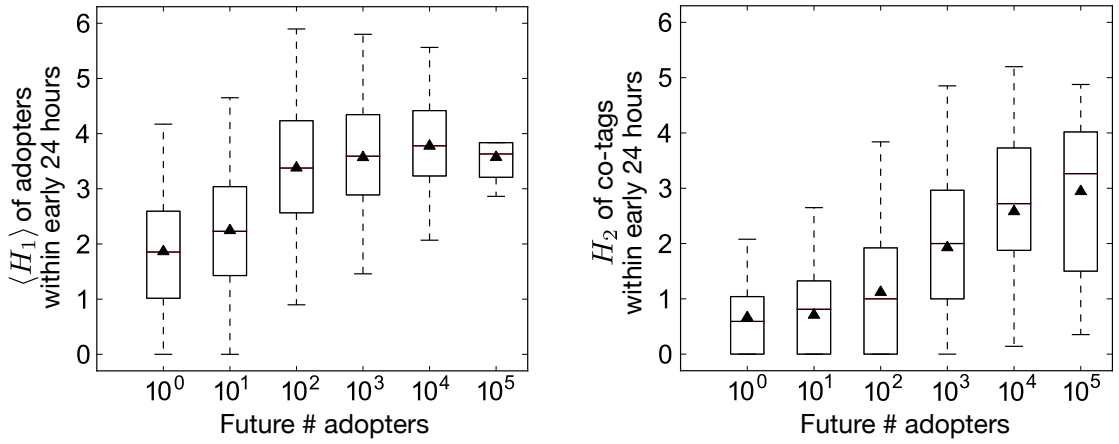
6.2.1 Prediction via User Diversity

Hashtags in Twitter can be treated as channels connecting people with shared interests, because hashtags label and index messages enabling people to easily retrieve information and broadcast to certain groups. As illustrated in Fig. 6.3, users with focused interests are linked with few groups, while people who care about diverse issues are exposed to a larger number of interest groups through hashtag channels. We expect the latter category of users to play a critical bridging role, connecting many groups in the network. This would allow them to spread innovative information to multiple groups, as suggested by the weak tie hypothesis [Granovetter, 1973], thus boosting the diffusion of hashtags [Onnela et al., 2007b, Weng et al., 2013a, 2014c]. In other words, we hypothesize that if a hashtag has early adopters with diverse topical interests, it is more likely to go viral.

Given a hashtag h, we track the users who adopt it within t hours after h is created and compute the average interest diversity among these early adopters as a simple predictor. Irrespective of how long we track, we observe a positive correlation between the average user diversity and the future popularity of the hashtags, measured as the total number of adopters after one month (see Fig. 6.4a).

To better evaluate the predictive power of adopter diversity, let us run a simple prediction task based on information at the early stage to forecast which hashtags from the test period will be popular in the future. A hashtag is deemed popular if the number of distinct adopters at the end of the test period is above a given threshold. Our evaluation algorithm has three steps:

i) For each feature, we compute its value for each newly emergent hashtag h in the test period based on the set of early adopters of h within t hours after the birth of h. A hashtag is born when the first tweet containing it appears. The feature is either a measure of user characteristics averaged among early adopters, or a linear combination of several such measures. We track adoption events for $t = 1, 6$, and 24 hours since



(a) User diversity (b) Content diversity

Figure 6.4: (a) Correlation between the average topic-based entropy H_1 of adopters in the first 24 hours and the total number of future hashtag adopters. (b) Entropy H_2 of hashtags co-occurring in the first 24 hours across topic clusters as a function of future popularity of emergent hashtags.

birth.

ii) Hashtags are ranked by the feature values in descending order.

iii) We set a percentile threshold for labeling popular hashtags. The most popular hashtags are deemed "viral." Based on this ground truth, we can measure false positive and true positive rates and draw a receiver-operating-characteristic (ROC) plot. The area under the ROC curve (AUC) is our evaluation metric. The higher the AUC value, the better the feature as a predictor of future hashtag popularity.

We consider several user attributes of early adopters that have been shown in the literature to be strong predictors of virality [Cha et al., 2010, Suh et al., 2010, Szabo and Huberman, 2010, Romero et al., 2011a]. These include the number of early adopters n, number of followers *fol* (potential audience), and number of tweets *twt* that a user has produced during the observation period (activity). We additionally consider the diversity of topical interests, H_1. The goal of our experiment is not to achieve the highest accuracy (a task for which different learning algorithms could be explored). We aim to compare the predictive powers of different features. Therefore we focus on the relative differences between AUC values generated by single or combined features rather than on the absolute AUC values. AUC values measured using different features are listed in Table 6.3. Among individual features, n is the most effective. When we combine it with other features, *fol* yields high AUC consistently, but the differences are very small. The performance of the diversity metric is competitive, matching the top results in several experimental configurations. These results are not particularly sensitive to the popularity threshold or the duration of the early observation window.

Table 6.3: AUC of prediction results using different adopter features within t early hours. Prediction features include the number of followers (fol), the number of tweets (twt), the diversity of topical interests of adopters (H_1), and the number of early adopters (n). The threshold is expressed as a top percentile of most popular hashtags that are deemed viral for evaluation purposes. Best results for each column are bolded.

Threshold	50%			10%			1%			0.1%		
t (hour)	1	6	24	1	6	24	1	6	24	1	6	24
twt	0.57	0.58	0.59	0.62	0.63	0.64	0.67	0.68	0.68	0.70	0.71	0.72
fol	0.57	0.58	0.59	0.62	0.64	0.65	0.67	0.69	0.70	0.74	0.75	0.76
H_1	0.55	0.55	0.55	0.58	0.58	0.58	0.60	0.60	0.61	0.63	0.63	0.64
n	0.64	**0.68**	0.71	0.75	0.79	0.84	0.82	0.86	**0.90**	0.86	**0.89**	0.91
$n + twt^\dagger$	0.64	**0.68**	0.71	0.75	**0.80**	0.84	**0.83**	**0.87**	**0.90**	**0.87**	**0.89**	**0.92**
$n + fol^\dagger$	**0.65**	**0.68**	**0.72**	**0.76**	**0.80**	**0.85**	**0.83**	**0.87**	**0.90**	**0.87**	**0.89**	**0.92**
$n + H_1^\dagger$	0.64	**0.68**	0.71	0.75	**0.80**	**0.85**	**0.83**	0.86	**0.90**	0.86	**0.89**	**0.92**

† A linear combination with coefficients determined by regression fitting using least squared error.

6.2.2 Prediction via Content Diversity

In this section we examine whether the future popularity of a hashtag is affected by the topical diversity of its early co-occurring tags. How people apply hashtags to label their messages depicts their topical interests and determines the topology of the tag co-occurrence network. In Fig. 6.1, a link between the topic layer and the social layer of the network marks an association between a user and a hashtag. This tag may attract an audience in the social network. The co-occurrence of two tags extends the audience groups of both. For example, link ℓ in Fig. 6.1 exposes user A to the blue topic and user B to the red cluster. Therefore we expect a hashtag to be exposed to more potential adopters, making it more likely to go viral, if it often co-occurs with many other hashtags. To test this hypothesis, we measure the number m of co-tags. Furthermore, if co-tags are very popular, we would expect a stronger effect because they would provide a larger audience. We therefore measure the popularity of co-tags in terms of numbers of tweets T and adopters A during the observation period. And if co-tags are about diverse topics, this may further boost the effect by extending the audience to many groups with small overlap. In conclusion, we hypothesize that many popular co-tags about diverse topics should be a sign that a hashtag will grow popular.

Given an emergent hashtag h, we track other tags that co-occur with h within t hours after h is born and measure the topical diversity H_2 of these co-tags. We observe a positive correlation between the diversity of early co-tags and the future popularity of the tag (see Fig. 6.4b). Then we apply the same method as in Sec. 6.2.1 to test the predictive power of different traits associated with early co-tags. In this case, the prediction features for each target hashtag are computed based on early co-tags instead of adopters. Again, the goal of our experiment is to compare the predictive powers of different features, thus we examine the relative differences in AUC values generated by the various traits. The results are reported in Table 6.4. The number m of co-tags observed in the early stage is the best

Table 6.4: AUC of prediction results using different features among co-tags within t early hours. Prediction features include the number of tweets containing the co-tags (T), the number of co-tag adopters (A), the diversity of co-tags (H_2), and the number of observed co-tags (m). The threshold is expressed as a top percentile of most popular hashtags that are deemed viral for evaluation purposes. Best results for each column are bolded.

Threshold	50%			10%			1%			0.1%		
t (hour)	1	6	24	1	6	24	1	6	24	1	6	24
T	0.50	0.51	0.52	0.51	0.53	0.55	0.58	0.62	0.66	0.66	0.72	0.75
A	0.50	0.50	0.52	0.50	0.52	0.54	0.58	0.62	0.65	0.65	0.71	0.74
H_2	0.50	0.51	0.53	0.52	0.54	0.57	0.61	0.66	0.70	0.70	0.77	0.82
m	0.52	0.53	0.55	0.55	0.58	0.61	0.64	**0.70**	**0.75**	0.72	**0.81**	**0.86**
$m + T^\dagger$	0.52	0.53	0.55	0.54	0.57	0.61	0.64	**0.70**	**0.75**	0.72	**0.81**	**0.86**
$m + A^\dagger$	0.52	0.53	0.55	0.55	0.57	0.61	0.64	**0.70**	**0.75**	0.72	**0.81**	**0.86**
$m + H_2^\dagger$	**0.55**	**0.55**	**0.57**	**0.58**	**0.60**	**0.63**	**0.66**	**0.70**	**0.75**	**0.74**	**0.81**	**0.86**

† A linear combination with coefficients determined by regression fitting using least squared error.

single predictor of virality. When we combine m with a second feature, co-tag diversity provides the best results irrespective of the threshold or the duration of the early observation window. Interestingly, m and H_2 are both about diversity and perform better than the popularity-based features T and A.

6.2.3 Summary

In the discussion above, we evaluate the predictive powers of two categories of features for identifying future popular hashtags. These two sets of features, based on early adopters and co-occurring tags, have different effectiveness. By comparing the AUC values in Tables 6.3 and 6.4, we find that adopter features yield better results. However, they also require additional prerequisite knowledge: in addition to tracking hashtag co-occurrences for building the topic network, we also need to record user-generated content. The features built upon early co-tags are less expensive, but the performance is slightly worse; a possible interpretation for this is that few tweets in the observation window may contain co-occurring tags, while they all have associated users. Therefore co-tag features are more sparse. Depending on what type of information is available, one might choose either approach or a combination of both.

6.3 Social Influence

High topical diversity of adopters and co-occurring tags is a positive sign that a hashtag is growing popular, as shown in the previous section. However, does high topical diversity also signal a growth in individual influence? On one hand, when an individual talks about various topics, she may have contact with many others through shared interests or

hashtags, thus attracting more attention (see Fig. 6.3). On the other hand, focused interest may enhance expertise in specific fields, thus increasing the content interestingness and retweetability. In this light, low diversity triggered by expertise might help people become popular. In this section we evaluate these two contradictory hypotheses.

Some people are more influential than others in persuading friends to adopt an idea, an action, or a piece of information. The concept of *social influence* has been discussed extensively in social media research. Most of the studies in the literature have considered users who are active [Cha et al., 2010], have many followers [Cha et al., 2010, Romero et al., 2011a], are able to trigger large cascades [Kitsak et al., 2010, Bakshy et al., 2011], or get retweeted or mentioned a lot [Cha et al., 2010, Suh et al., 2010, Romero et al., 2011a] as signals of high social influence. Which user characteristics make people popular and influential? Does the diversity of individual topical interests play a role in the social influence processes? Let us consider several individual properties:

Number of retweets (RT) How many times an individual is retweeted during the observation time period. We consider RT as a direct indicator of social influence, since it quantifies how many times the user succeeds in making others adopt and spread information.[1] The number of retweets is dependent on the length of the observation window, because we believe that social influence is accumulated in time and requires long-term endeavor [Cha et al., 2010].

Number of followers (fol) The number of followers suggests how many people can potentially view a message once the user posts it.

Number of tweets (twt) The number of tweets generated by the user; the higher the number, the more active the user.

Content interestingness (β) How interesting is the content posted by the user. Lerman studied the interestingness of online content on Digg and defined it as "the probability it will get retweeted when viewed" [Lerman, 2007]. To measure β in the Twitter context, we assume that the value of RT for an individual is proportional to the number of tweets twt he produced, the number of followers fol, the chance α that a message is seen by a follower, and the appeal of the content. Treating α as a constant for simplicity, we obtain

$$\beta = \frac{RT}{twt \cdot fol \cdot \alpha} \propto \frac{RT}{twt \cdot fol}. \tag{6.4}$$

Diversity of interests (H_1) See Sec. 6.1.2.

Table 6.5 lists the results of a linear regression estimating how many times a user is retweeted according to several user features. Intuitively, users with many followers are more likely to spread their messages and thus get retweeted more frequently, because they

[1] Due to the settings of the Twitter API, the number of retweets per user that we collect includes all the retweeters in every cascade. That is, suppose user B retweets user A and then C retweets B; both tweets are counted in RT for A, even though C did not directly retweet A. However, since the majority of information cascades are very shallow [Bakshy et al., 2011], RT is a good approximation of the direct retweet count.

Table 6.5: Linear regression estimating how many times a user is retweeted. For efficiency, the regression is based on a random sample of 10% of the users ($N = 2,171,624$).

	How many times a user is retweeted	
	Coefficient	SE
(Intercept)	20.9	0.5
Num. followers $(fol)^{\dagger}$	193.0 ***	0.5
Num. tweets $(twt)^{\dagger}$	51.1 ***	0.5
Content interestingness $(\beta)^{\dagger}$	3.9 ***	0.5
Diversity of interests $(H_1)^{\dagger}$	-9.1 ***	0.5

† Variables are normalized by Z-score. *** $p < 0.001$

have many more potential viewers. The number of followers is the most important factor, as supported by the largest positive coefficient in the regression. The number of generated tweets also has a positive coefficient in the regression, implying that being active helps users get retweeted more. The result confirms several existing studies suggesting that high social influence requires long-term, consistent effort [Cha et al., 2010, Suh et al., 2010]. The interestingness of the story is positively correlated with social influence as well, although not as strongly as the other factors. Finally, the negative coefficient of diversity in Table 6.5 suggests that users with diverse interests tend to have low influence. This supports the hypothesis that social influence is topic-sensitive, requiring expertise in a specific field [Weng et al., 2010]; posting about the same topic is more effective for gaining social influence, compared to commenting on many different subjects. In summary, people can acquire social influence by having a big audience group, being productive, creating interesting content, and staying focused on a field. Unfortunately, it seems that there is no simple recipe of success.

We illustrate how several user properties are related to the number of followers and the topical diversity of user interests in Fig. 6.5. Most users have a small number of followers and low entropy (Fig. 6.5a). Active users tend to have high diversity, as expected by the nature of entropy (Fig. 6.5b). The number of followers is shown in Fig. 6.5c to be a powerful factor to get retweeted more often, consistently with the regression results in Table 6.5. Finally, the content interestingness appears to be correlated with the number of followers but strongly with user diversity (Fig. 6.5d).

6.3.1 Active vs. Inactive Users

Let us explore how the number of followers a user can attract is affected by the diversity of topical interests. The entropy measure for diversity is biased by user activity: generating more tweets with more hashtags tends to yield higher entropy. Thus we group users by productivity, so that individuals in the same group have comparable values of topical diversity. For users in the same group, we compute the Spearman rank correlation between the number of followers and diversity. We use Spearman because, unlike Pearson, it does not require that both variables be normally distributed. According to Fig. 6.6, low-engagement

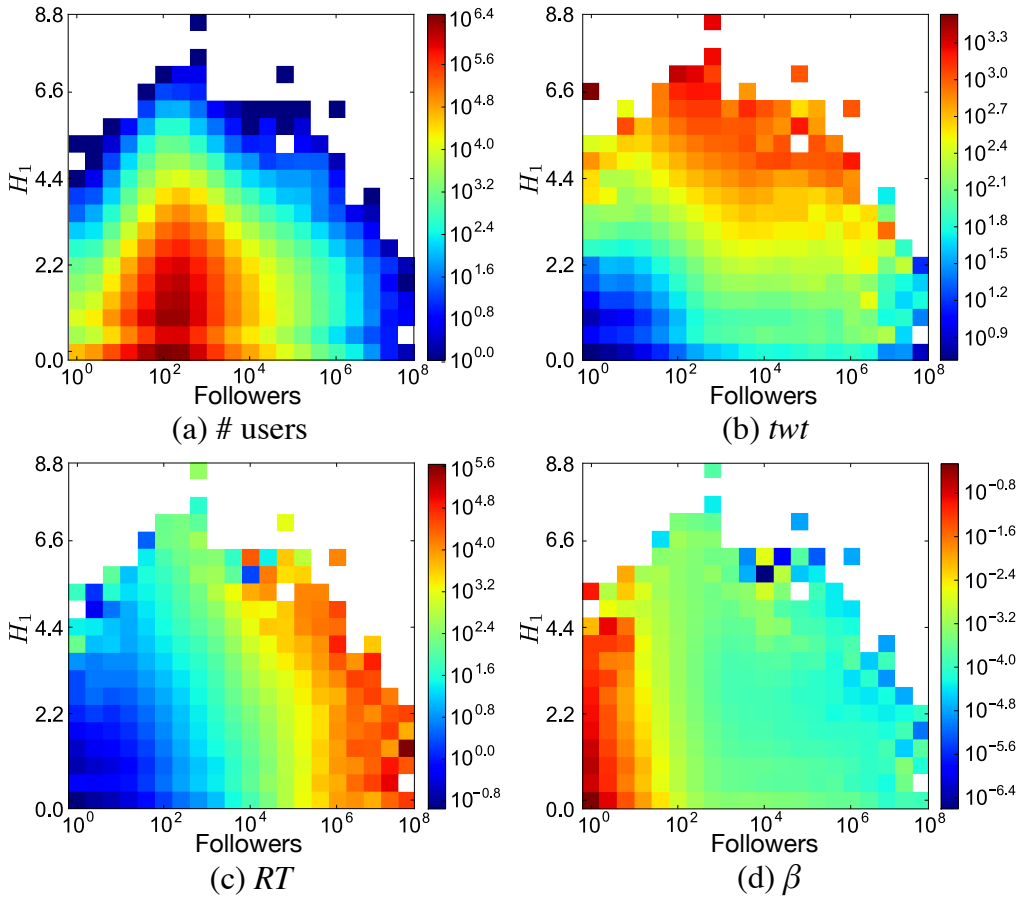


Figure 6.5: Heatmaps of (a) the number of users, (b) the number of tweets generated, (c) how many times a user is retweeted, and (d) the content interestingness of a user, as a function of the diversity of topical interests, H_1, and the number of followers in the observation window.

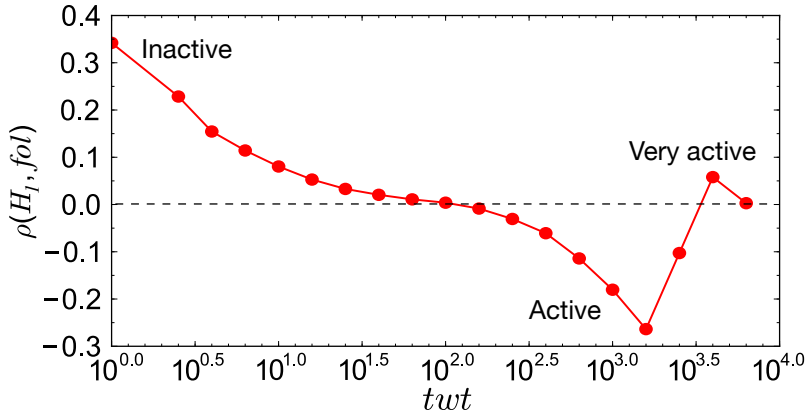


Figure 6.6: Spearman rank correlation between the number of followers and the topical diversity of user interests as a function user activity. All shown correlation values are significant ($p < 0.05$).

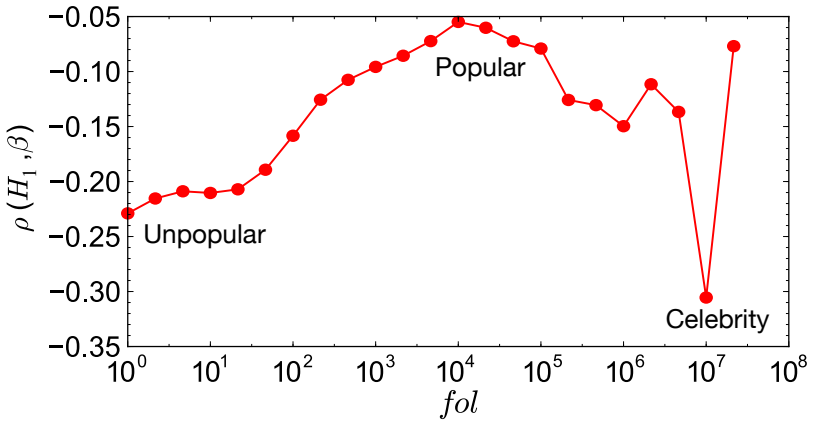


Figure 6.7: Spearman rank correlation between content interestingness and the topical diversity of user interests as a function of how many followers users have. All shown correlation values are significant ($p < 0.05$).

users attract followers by talking about various topics, while active users tend to obtain many followers by maintaining focused topical interests. For the most active users, topical diversity is not relevant; many of these accounts are spammers and bots.

6.3.2 Celebrities and Ordinary Users

When looking into the effect of interest diversity on content appeal, we need to control for the number of followers, since our interestingness measure is strongly correlated with the number of followers (see Fig. 6.5d). The negative correlations shown in Fig. 6.7 suggest that in general, focused posts promote content appeal. One possible interpretation is that people follow someone for a reason. Content has to be consistent in order to match such expectations; i.e., one is less likely to share a tip on cosmetics from a politician. This effect is stronger for users with few followers and celebrities; people with moderate popularity generate retweets with focused and diverse content.

6.4 Discussion

In this chapter, we propose methods to identify topics using Twitter data by detecting communities in the hashtag co-occurrence network, and to quantify the topical diversity of user interests and content, defined by how tags are distributed across different topic clusters.

We find that popular hashtags tend to have adopters who care about various issues and to co-occur with other tags of diverse themes at the early stage. One practical application evaluated here is to predict viral hashtags using features built upon the topical diversity of early adopters or co-tags. In the prediction using information on early adopters, the performance of topical diversity is competitive with other user traits while combined with the number of

early adopters. In the prediction with early co-occurring hashtags, features about diversity, including the number of early co-tags and their topical diversity, excel the popularity-based features. However, high topical diversity is not a positive factor for individual popularity. High social influence is more easily obtained by having a big audience group, producing lots of interesting content, and staying focused. In short, diverse messages and focused messengers are more likely to generate impact.

The interesting observation that high diversity helps a hashtag grow popular but does not help develop personal authority originates from the different mechanisms by which a hashtag and a user attract attention. In the diffusion process of a hashtag, adopters with diverse interests play a role as bridges connecting different groups and thus positively improve the visibility of the tag. These results are consistent with Granovetter's theory [Granovetter, 1973], as well as our recent findings on the strong link between community diversity and virality [Weng et al., 2013a]. On the other hand, a user gains social influence through expertise or authority within a cohesive group with common interests.

The proposed measurement of topical diversity provides a simple yet powerful way to connect the social network structure with the semantic space extracted from online conversation. We believe that it holds great potential in applications such as predicting viral hashtags and helping users strengthen their online presence. Several previous studies have supported our intuition. For example, network diversity was shown to be positively correlated with regional economic development [Quigley, 1998, Eagle et al., 2010]; community diversity at the early stage tend to boost the chances of a meme going viral [Weng et al., 2013a, 2014c]. We expect the diversity measurement would prompt new approaches to many research questions in viral marketing and social media analytics.

Chapter 7

Topic Selectivity and Tie Strength

After studying the topic space of online conversation and the relationship between topical diversity and popularity from a global viewpoint, we shift our attention to the conversation topics discussed in *ego networks* centered at individual actors. Do people talk about a wider variety of subjects with close friends than acquaintances? On social network sites like Facebook, people communicate with all kinds of ties—best friends, coworkers, and family members. Much of this communication is comprised of updates we broadcast to our social circles, and anyone who sees a post can respond with a "Like" or a comment. However, some friends may respond to every post, while others only respond to posts about certain topics.

In this chapter, we demonstrate that close friends respond to a wider variety of topics than do acquaintances, even after accounting for the number of posts a tie responds to. Meanwhile, weak ties do not have a stronger preference for more popular (potentially "safer") topics.

7.1 Hypotheses

Friendships vary in their intensity and intimacy, a concept known as tie strength; *strong ties* are our closest confidants and supporters, while *weak ties*, with whom we feel less close, comprise the majority of our personal networks [Granovetter, 1973]. Furthermore, we talk about more intimate topics with our close friends [Granovetter, 1973]. We disclose private information to the people we like, and that disclosure further increases tie strength [Collins and Miller, 1994]. This is true online, as well: intimate language in Facebook messages and wall posts is a strong indicator of relationship closeness [Gilbert and Karahalios, 2009]; teens save discussion of personal problems and romantic relationships for their closest friends, but talk to many ties about school and TV [Mesch and Talmud, 2006].

However, we do not just use more intimate language with close friends; our relationships also demonstrate greater multiplexity, that is, we share multiple facets of our lives with them [Verbrugge, 1979]. Strong ties exchange more kinds of information [Haythornthwaite

and Wellman, 1998, Wellman and Wortley, 1990], and many of the topics that good friends report talking about are lightweight, silly matters [Bearman and Parigi, 2004]. A strong tie is the coworker who is also a neighbor and a hiking buddy, and so conversations might range from serious work issues to hiking boots to jokes told at the neighborhood barbecue.

Given this increased intimacy and multiplexity, we would expect that strong ties would talk about a broader variety of topics than would weak ties, both online and off. We seek to validate these previous survey-based findings using the text of millions of Facebook posts and information about the kinds of ties who respond to them. Strong ties respond to more posts on social media (e.g., [Gilbert and Karahalios, 2009]), but even after accounting for response count, we expect strong ties to span a wider range of topics. Because strong tie relationships are multiplex and have longer histories, strong ties may be more comfortable replying regardless of the topic, while weak ties may only feel comfortable responding to topics that are aligned with the nature of the tie (e.g., work). However, weak ties might be willing to cross certain boundaries if the topic is one that everyone is talking about, and therefore more socially acceptable (e.g., a holiday or a life event where congratulations are expected). We therefore test the following two hypotheses:

H1 *Strong ties will respond to a wider variety of topics than will weak ties, even after accounting for the number of posts they respond to.*

H2 *Weak ties will be especially likely to respond to popular topics.*

7.2 Definitions

We first introduce several key concepts to facilitate our subsequent discussion.

7.2.1 Topic Classification

To assign a topic to each status update, we adapted the topic models constructed by Schwartz et al. [2013]. Their corpus[1] maps terms to topics with a given probability based on applying Latent Dirichlet Allocation (LDA) to the Facebook status updates of 75,000 volunteers. Their models produce 2,000 topics, ranging from emotional support to the weather. In our corpus, each post was automatically assigned the topic with the highest probability score. The most popular topics appear in Table 7.1 along with representative terms.

7.2.2 Tie Strength

Tie strength is operationalized two ways, one binary and one continuous. First, Facebook allows users to create a close friend list. In the binary measure of tie strength, an alter is considered as a strong tie, if he or she appears in ego's *close friend list*, and a weak tie

[1] `http://www.wwbp.org/data.html`

Table 7.1: Most popular topics and key terms from Facebook status updates. Topic labels added by researchers for clarity.

ID	Topic label	Posts%	Egos%	Common terms
1083	Comfort	1.51	1.07	don ll ve won haven worry
213	Emotion/Emoticon	0.83	0.40	:-d ;- :-p :-/ xxx hehehe toooo esp wooohooo
214	Informal	0.82	0.42	hai ki se ko ka ho ke bhi
1156	Father's day	0.48	0.52	happy fathers father's dad dads
107	US Independence day	0.48	0.48	july happy fireworks fourth safe display BBQ holiday
68	Disappointment	0.46	0.39	missing gutted depressed nooooo nooo poorly bummer
300	Smiling	0.44	0.39	: smiles
290	Weather	0.38	0.30	heat hot degrees weather cold AC summer temperature
1851	Slang	0.38	0.12	text haha funn mall mee hmu txt todayy bestfriend
1291	Swimming	0.37	0.32	pool swimming swim water suit dive adult drowning
366	Abbreviations	0.35	0.21	wid gud ma tym frnds fr dis luv frm
1915	Laughter	0.35	0.27	hah :p hahaha ;p sexy ;d hehe btw hehehe aha cheers
1055	Riding	0.32	0.29	ride bike riding horse rode horses motorcycle bicycle
1860	Boredom	0.31	0.24	bored bore text entertain extremely boredom boring
1430	Swearing	0.28	0.14	shit ass bitches fuck bitch niggas nigga hoes

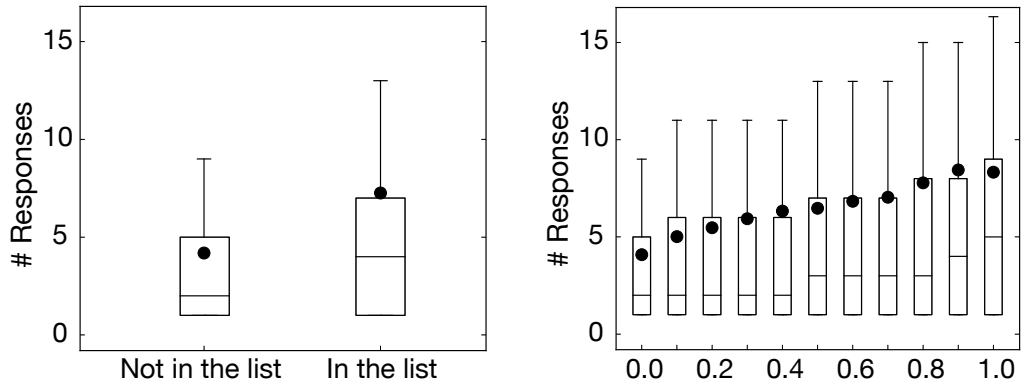
otherwise. The continuous measure of tie strength is the frequency with which ego and alter appear together, based on being *co-tagged* in photos, events, posts, and check-ins [Jones et al., 2013]. Since some egos appear in many photos and others very few, co-tagging is normalized by rank within each ego network and scaled between 0 and 1, with the high end of the scale representing alters who are co-tagged most frequently with ego. Because we use co-tagging as a measure of tie strength, any posts in which alter is tagged are excluded from analysis.

We adopt tie strength measurements independent of interaction among people, so that we can test whether behavioral and mental recognition of weak and strong ties are aligned—self-claimed strong ties attract more communication. Sensitivity testing with other measures of tie strength, e.g., whether ego and alter are in a *romantic relationship* or have listed each other as *family members* produce similar results, with strong ties replying to a wider variety of topics.

7.3 Tie Strength and Topic Diversity

We begin by confirming that strong ties respond to (like or comment on) more posts. In Fig. 7.1, ties that appear on ego's close friend list are, on average, nearly twice as responsive as other ties ($p < 0.001$), and co-tagging rank and number of responses are also highly correlated (Pearson's correlation coefficient $r = 0.96$).

Next we examine Hypothesis 1, evaluating how tie strength relates to the diversity of topics covered by a given ego-alter pair. Figure 7.2 shows that strong ties respond to a wider variety of topics, independent of the number of posts they respond to. Each plot has three



(a) Close friend lists

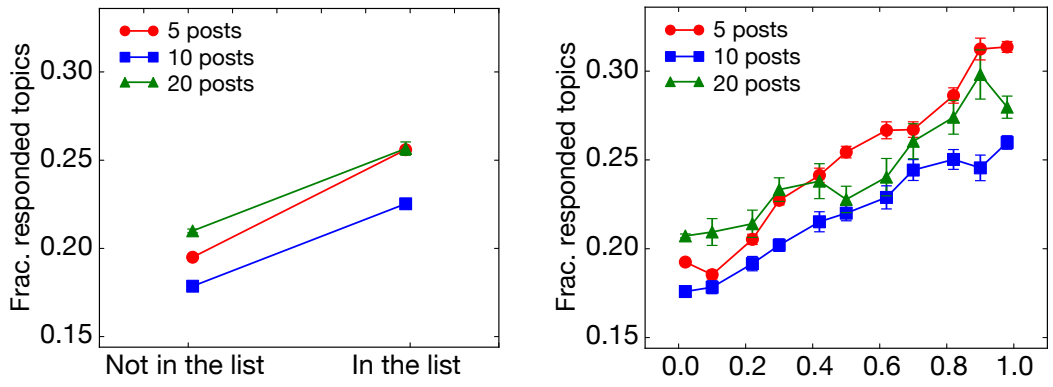
(b) Normalized co-tag rank

Figure 7.1: Strong ties are more responsive than weak ties. Tie strength is measured as (a) being on a close friend list or by (b) the percentage of photos and posts ego and alter are tagged in together.

lines, representing alters who responded to different numbers of posts in total (5, 10, or 20 posts). Consider two friends who both responded to 5 posts over three months (red line): On average, a close friend responded to 25% of ego's topics, while a weak tie (one not on ego's close friend list) responded to only 20%. Similarly, alters who are frequently co-tagged with ego also respond to a greater fraction of ego's topics. A multilevel linear model estimating the fraction of topics to which a given tie will respond confirms these results (see Table 7.2). The model is grouped at the ego level to account for multiple alters per ego, and the outcome is normalized as a fraction of topics which accounts for both ego's posted topic diversity and the number of posts alter responds to. The intercept represents a typical alter who does not appear on ego's close friends list. On average, this weak-tie alter will reply to about 40% of the topics ego posts. An alter who appears on ego's close friends list responds to 5% more topics ($p < 0.001$). A similar regression using the continuous measure of tie strength yields similar results. Each decile in normalized co-tag rank is associated with responding to 0.7% more topics.

7.4 Tie Strength and Topic Popularity

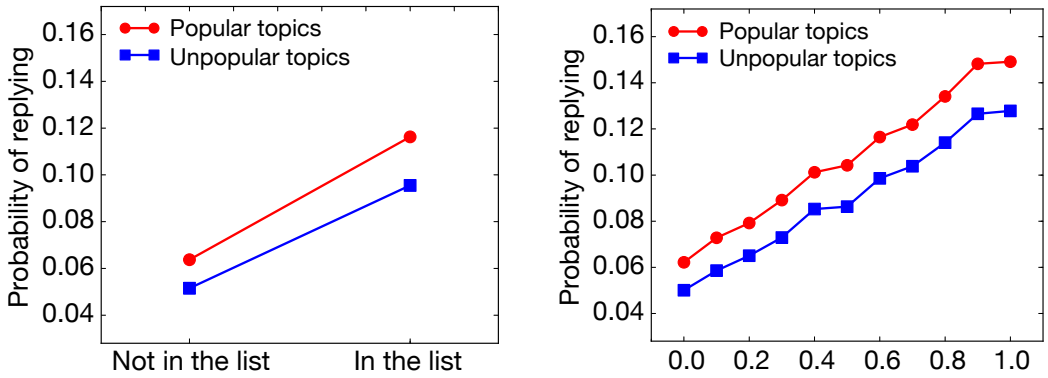
We now turn our attention to Hypothesis 2, and examine whether or not weak ties are more likely to respond to popular topics. Each post is assigned a binary popularity score based on a median split of topics ranked by the number of responses they received during the three-month window from all sampled users. Figure 7.3 shows that topic popularity does not moderate the relationship between tie strength and topic diversity. The y-axis shows the conditional probability that a tie will reply to a topic, and the two lines represent popular and unpopular topics. While strong ties (by either measurement) are more likely to reply to a post, the slopes of the two lines are the same, such that both kinds of ties are more likely to reply to popular topics. We then perform a logistic regression using one randomly



(a) Close friend lists

(b) Normalized co-tag rank

Figure 7.2: The fraction of ego's posted topics that an alter has liked or commented on, as a function of tie strength. Alters are grouped by total number of the ego's posts to which they responded.



(a) Close friend lists

(b) Normalized co-tag rank

Figure 7.3: The conditional probability of replying to popular and unpopular topics as a function of tie strength. Strong and weak ties both respond more to popular topics.

sampled post-alter pair per ego. The model estimates the likelihood that a given alter would reply to a particular post, and includes controls for tie strength, topic popularity, and an interaction between the two (see Table 7.3). While topic popularity and tie strength both have statistically significant, if modest, correlations with likelihood of response, the effect of the interaction between the two is not statistically significant. Both strong and weak ties prefer popular topics, and topic popularity is not especially influential on weak ties.

7.5 Discussion

These results are consistent with the hypothesis that strong ties respond to a wider variety of topics than weak ties. The difference is modest but statistically significant: a strong tie

Table 7.2: Multilevel linear regression estimating the fraction of topics to which an alter replies. Close friends respond to approximately 4.5% more topics.

	Fraction of topics to which alter responds	
	Coefficient	SE
(Intercept)	0.403 ***	0.001
Alter is close friend (0/1)	0.045 ***	0.000
N = 117,091 egos; 3,583,850 dyads.		*** $p < 0.001$

Table 7.3: Logistic regression estimating the likelihood that an alter replied to a post, given tie strength and post popularity. Tie strength is centered at its mean.

	Likelihood of response		
	Coefficient	SE	p-value
(Intercept)	-1.031 ***	0.033	< 0.001
Tie strength (co-tag rank)	0.621 ***	0.142	< 0.001
Topic is popular (0/1)	0.252 ***	0.046	< 0.001
Tie strength $\times$ Is popular	-0.252	0.201	0.209
N = 10,000 posts.			*** $p < 0.001$

responds to about 5% more topics than a weak tie. This finding confirms the results of earlier survey-based studies [Mesch and Talmud, 2006, Bearman and Parigi, 2004], and lends support to the notion that strong ties have a broader range of common interests than weak ties. The relatively small effect size might be the result of a number of factors. The topic model might be too granular, resulting in the appearance of topic breadth for weak ties even if they are responding to posts on separate but related topics. Although this is unlikely to have a larger impact on weak ties than strong ties, additional measurement with less granular topic models might be warranted.

It is more likely that features of Facebook itself are lessening differences between strong and weak ties' response patterns. "Liking" a piece of content is a fairly lightweight form of feedback, and one that might be more popular among weak ties who wish to express affinity or enjoyment of a topic on which they would not otherwise converse with ego. Normalizing for this effect, if it exists, would be difficult, but an additional analysis could tease apart likes and comments and we would expect to see a larger difference in comments.

Contrary to expectations, we do not find evidence that weak ties are more heavily influenced by topic popularity. Intuitively, we expected weak ties to feel more comfortable responding to a topic outside of their normal range of interaction with the poster if the content in question had received responses from many others. Although this intuition proved to be correct, the same effect was observed for strong ties at about the same magnitude. In our case, operationalizing topic popularity using a median split may be masking some distributional differences in the responses posted by strong and weak ties. For example, strong ties might be far more likely to respond to extremely unpopular content, but the median split could mask this effect by emphasizing distinctions between moderately popular

and moderately unpopular topics. A more thorough investigation of responses across the content popularity distribution might produce different results. However, given the general response curves produced with this approach and the marginal effect size of the interaction term we believe it unlikely that an alternative operationalization will yield different results.

On Facebook, popularity and intimacy may be conflated. Popular topics on the site are a mixture of those traditionally considered self-disclosure (e.g., requests for support) and general-interest topics (e.g. holiday greetings). This mixture of content might make topic popularity a less important predictor of strong versus weak tie response rates. Instead, a measure of intimacy might be a better choice. We would expect weak ties to be more likely to respond to less intimate posts, such as holidays or the weather, and strong ties to be drawn to posts about health issues or family matters.

Facebook's feed ranking algorithm might also affect what posts strong and weak ties see. Posts that receive a lot of feedback may be ranked more highly for future viewers, potentially accounting for the finding that both strong and weak ties respond more to popular topics. These effects cannot be eliminated in an observational study, suggesting the need for controlled experimentation. Showing randomly selected content to alters might clarify the extent to which ranking algorithms are affecting these results.

This chapter shows that strong ties respond to a broader variety of topics than weak ties, even after controlling for the overall number of responses. Although the effect size is modest, it is in line with earlier survey-based findings, and it provides additional support for theories that strong ties are characterized by a multiplexity of interests. Weak ties do not have a stronger preference for popular topics, suggesting that the impact of content popularity operates independently from tie strength in an individual's decision to respond to a piece of content. Note that weak ties in the experiment are shown to be able to respond to as many as 20 posts. One interesting question we can look into as future work is what motivates interactions between egos and weak-tie alters; for example, they happened to travel to the same city, or just finished a reunion with old friends. Incorporating these concerns might help better predict responses to posts by weak ties. Furthermore, it would be valuable to investigate the inconsistency between listed strong ties and those who intensely communicate with egos.

Part III

Diffusion on Online Social Networks

How does network structure affect information diffusion?
How does information diffusion affect network evolution?

Chapter 8

Meme Virality

Although numerous memes are created everyday, only a few of them go viral, prompting a question: *can we predict the future popularity of a meme at its early stage?* This question has attracted much attention across disciplines, including marketing [Rogers, 2003, Aral and Walker, 2011], network science [Leskovec et al., 2007a], communication [Berger and Milkman, 2009], and social media analytics [Jamali and Rangwala, 2009, Szabo and Huberman, 2010, Suh et al., 2010, Tsur and Rappoport, 2012]. The structure of the underlying network has been shown to have a significant impact on the spreading process in general [Goffman and Newill, 1964, Daley and Kendall, 1964, Christakis and Fowler, 2007, Barrat et al., 2008, Pastor-Satorras and Vespignani, 2001].

This chapter focuses on the role of network topology, particularly community structure, in shaping the diffusion of online information, showing that a community-centric viewpoint can provide a unique vantage point to the challenge of predicting viral memes [Colbaugh and Glass, 2012, Weng et al., 2013a, 2014c]. We demonstrate that communities allow us to estimate how much the spreading pattern of a meme deviates from that of infectious diseases; meanwhile, viral memes tend to spread like epidemics, less trapped by communities. Our findings naturally lead to a practical application—predicting which memes will go viral in the future—in which we can predict the virality of memes based on their early spreading patterns in terms of community structure. Furthermore, we combine knowledge extracted from network community structure with feature sets built on influence of early adopters and characteristics of adoption time series into a more sophisticated prediction model [Weng et al., 2014c]. Our model outperforms random guessing, majority guessing, and three regression models that use early popularity or expected influence of early adopters. Features based on community structure are found to be the most powerful predictors of meme future success.

8.1 Definitions

Let us first define several key concepts and mathematical notations to facilitate the subsequent discussion.

8.1.1 Meme Popularity

We consider each Twitter hashtag h as a meme. $\mathbb{T}(h)$ is a set of all tweets that contain h and $\mathbb{T}_n(h)$ is a set of the earliest n tweets that contain h. Thus there are $T(h) = |\mathbb{T}(h)|$ tweets in total observed; a large $T(h)$ indicates high popularity. According to the definitions, we have $\mathbb{T}_n(h) \subseteq \mathbb{T}(h)$ and $n = |\mathbb{T}_n(h)| \leq T(h) = |\mathbb{T}(h)|$. Similar definitions can be made for adopters. $\mathbb{A}(h)$ is a set of adopters who tweeted about h. $\mathbb{A}_n(h) \subseteq \mathbb{A}(h)$ is a set of early adopters who tweeted at least one of the first n tweets and $A_n(h) = |\mathbb{A}_n(h)| \leq A(h) = |\mathbb{A}(h)|$. We employ this popularity measure, either based on tweets $T(h)$ or adopters $A(h)$, of a meme as an indicator of its virality; viral memes appear in a large number of messages and are adopted by many people.

We only include *emergent memes* in the measurement of meme concentration among communities, the strength of social reinforcement, and the prediction tasks. Emergent memes are defined as those that are used during the first week of our observation window and have fewer than X tweets during the previous month. Here we set $X = 20$.

8.1.2 Network and Community

We recover the social network based on reciprocal following relationships, and thus the recovered network is simply undirected and unweighted. Such a conservative choice to exclude information about link weights and directions makes the approach more generally applicable to cases where static data about the social network is more readily available than dynamic data about information flows. To demonstrate the robustness of the results across different types of communities, we apply both disjoint and overlapping community detection methods on the reciprocal follower network—*InfoMap* [Rosvall and Bergstrom, 2008] and *Link Clustering* [Ahn et al., 2010]. The results presented in this chapter are consistent irrespective of which community detection method is applied. Communities with fewer than three nodes are removed.

A community $c \in \mathcal{C}$ is a subset of nodes (users) in the network. The set of communities to which user u is assigned is:

$$C_u = \{c \mid u \in c, c \in \mathcal{C}\} \subseteq \mathcal{C}. \tag{8.1}$$

For an edge $(u, v) \in E$, $u, v \in V$ are two users connected by this edge. The sets of intra-community and inter-community edges are defined as, respectively:

$$E_{\circlearrowleft} = \{(u, v) \mid C_u \cap C_v \neq \varnothing\} \tag{8.2}$$

$$E_{\curvearrowright} = \{(u, v) \mid C_u \cap C_v = \varnothing\}. \tag{8.3}$$

Similarly, the sets of intra-community and inter-community edges can be defined for a single community c, respectively:

$$E^c_{\circlearrowleft} = \{(u,v) \mid c \in C_u \wedge c \in C_v,\ (u,v) \in E_{\circlearrowleft}\} \tag{8.4}$$
$$E^c_{\curvearrowright} = \{(u,v) \mid c \in C_u \vee c \in C_v,\ (u,v) \in E_{\curvearrowright}\}. \tag{8.5}$$

For each meme h, a set $\mathbb{T}(h|c)$ contains tweets generated by a group of adopters $\mathbb{A}(h|c)$ in community $c \in \mathcal{C}$; we have $T(h|c) = |\mathbb{T}(h|c)|$ and $A(h|c) = |\mathbb{A}(h|c)|$. We define $\mathbb{T}_n(h|c)$, $T_n(h|c)$, $\mathbb{A}_n(h|c)$, and $A_n(h|c)$, that consider only early tweets, in a similar fashion as in Sec. 8.1.1.

$C(h)$ denotes the number of *infected communities* of h, which are communities with at least one tweet containing h:

$$C(h) = |\{c \mid T(h|c) \geq 1,\ c \in \mathcal{C}\}|. \tag{8.6}$$

Similarly, the infected communities can be counted with only early tweets:

$$C_n(h) = |\{c \mid T_n(h|c) \geq 1,\ c \in \mathcal{C}\}|. \tag{8.7}$$

8.1.3 Network Surface

The neighbors of a given set of users U (not counting U) are deemed to be U's *surface*:

$$S(U) = \{v \mid u \in U \wedge v \notin U \wedge (u,v) \in E\}. \tag{8.8}$$

The definition of the surface can be extended recursively to the k-th surface, which contains users within k steps from any user in the target set U,

$$S^k(U) = S(S^{k-1}(U)) \cup S^{k-1}(U) \text{ and } S^1(U) = S(U). \tag{8.9}$$

8.1.4 Adopter Sequences and Time Series

For a given meme h, we consider the sequence of meme adopters, $\langle a^h_1, a^h_2, \ldots, a^h_{T(h)} \rangle$, where $a^h_i \in \mathbb{A}(h)$ is the creator of i-th tweet of h. A user may appear multiple times in the sequence if the user tweeted about h more than once. We also measure the shortest network path length between each pair of consecutive adopters and call it *step distance*, $d(a^h_i, a^h_{i+1})$.

Similarly we build the tweet time series, $\langle t^h_1, t^h_2, \ldots, t^h_{T(h)} \rangle$ where t^h_i marks the timestamp (in second) of the i-th tweet containing h. The number of tweets within the time τ is labelled as $T^\tau(h)$ where τ is a time duration measured starting from the first tweet. We can also measure the time difference between consecutive tweets, $\Delta t_i(h) = t^h_{i+1} - t^h_i$, known as *step time duration*.

8.1.5 Interactions

Let us define $I(h)$ as the total number of pair-wise user interactions regarding a meme h. Two types of user interactions are considered: retweets (RT), by which a user retweets a message containing h from another user; and mentions ($@$), by which a user mentions another in an tweet of h. Accordingly, the interactions are represented as $I^{\mathrm{RT}}(h)$ and $I^{@}(h)$. We consider interactions within communities $I^{\circlearrowleft}(h)$ and between communities $I^{\curvearrowright}(h)$, respectively:

$$I^{\circlearrowleft \mathrm{RT}}(h) = |\{(u,v) \mid u \text{ retweets } v \text{ about } h,\ (u,v) \in E_{\circlearrowleft}\}| \tag{8.10}$$

$$I^{\circlearrowleft @}(h) = |\{(u,v) \mid u \text{ mentions } v \text{ about } h,\ (u,v) \in E_{\circlearrowleft}\}| \tag{8.11}$$

$$I^{\curvearrowright \mathrm{RT}}(h) = |\{(u,v) \mid u \text{ retweets } v \text{ about } h,\ (u,v) \in E_{\curvearrowright}\}| \tag{8.12}$$

$$I^{\curvearrowright @}(h) = |\{(u,v) \mid u \text{ mentions } v \text{ about } h,\ (u,v) \in E_{\curvearrowright}\}| \tag{8.13}$$

In addition, we have $I^{\mathrm{RT}}(h) = I^{\circlearrowleft \mathrm{RT}}(h) + I^{\curvearrowright \mathrm{RT}}(h)$ and $I^{@}(h) = I^{\circlearrowleft @}(h) + I^{\curvearrowright @}(h)$.

8.2 Trapping Effect of Communities

Community structure is a common approach to study network topology and has been shown to affect information diffusion, including global cascades [Galstyan and Cohen, 2007, Gleeson, 2008], the speed of propagation [Onnela et al., 2007b], and the activity of individuals [Granovetter, 1973, Grabowicz et al., 2012]. One straight-forward effect is that communities are thought to be able to cripple the global spread because they act as traps for random flows [Granovetter, 1973, Onnela et al., 2007b] (see Fig. 8.1a).

Yet, the causes and consequences of the *trapping* effect have not been fully understood, particularly when *structural* trapping is combined with two important phenomena: social reinforcement and homophily. Complex contagions are sensitive to *social reinforcement*: each additional exposure significantly increases the chance of adoption. Although the notion is not new [Granovetter, 1978], it was only recently confirmed in a controlled experiment [Centola, 2010]. A few concentrated adoptions inside highly clustered communities can induce many multiple exposures (see Fig. 8.1b). The adoption of memes within communities may also be affected by *homophily*, according to which social relationships are more likely to form between similar people [McPherson et al., 2001, Centola, 2011]. Communities capture homophily as people sharing similar characteristics naturally establish more edges among them. Thus we expect similar tastes among community members, making people more susceptible to memes from peers in the same community (see Fig. 8.1c). Straightforward examples of homophilous communities are those formed around language or culture (see Fig. 8.1(d-e)); people are much more likely to propagate messages written in their mother tongue. Separating social contagion and homophily is difficult [Aral et al., 2009, Shalizi and Thomas, 2011], and we interpret complex contagion broadly to include homophily; we focus on how both social reinforcement and homophily effects collectively boost the trapping of memes within dense communities, not on the distinctions between them.

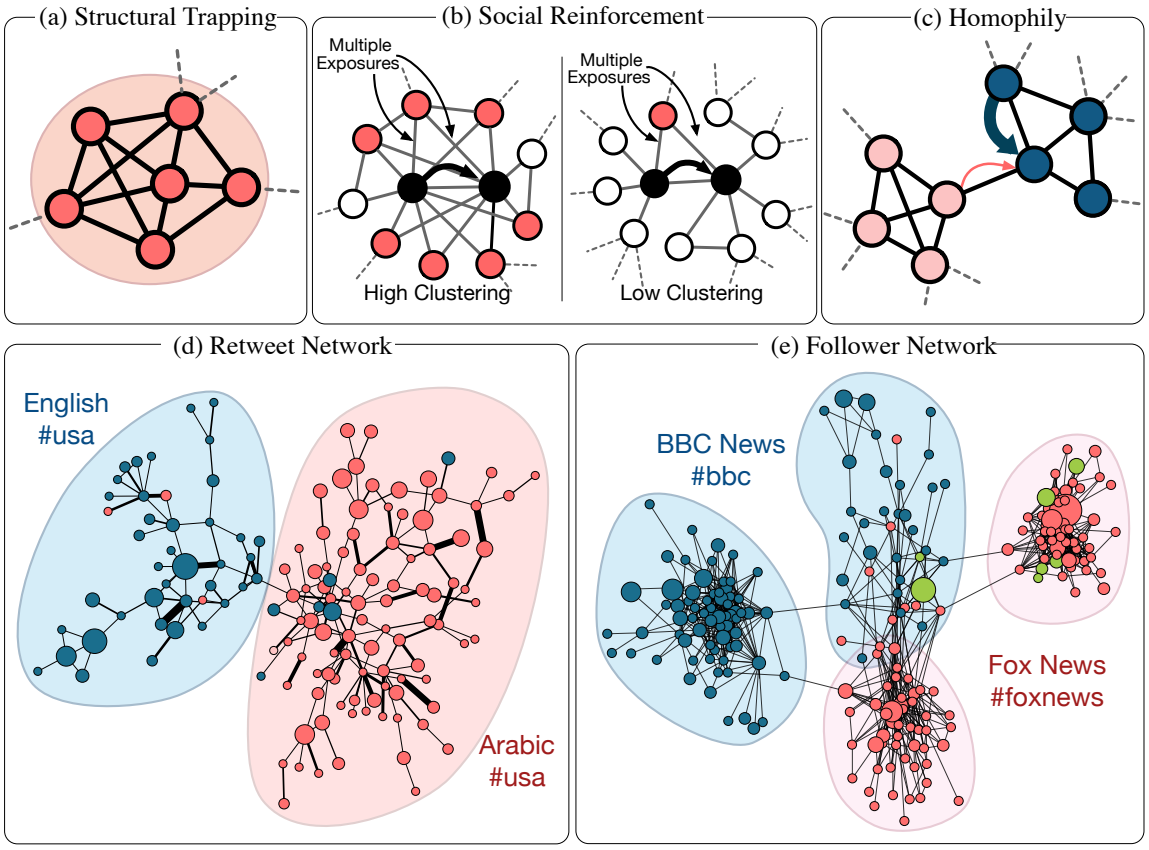
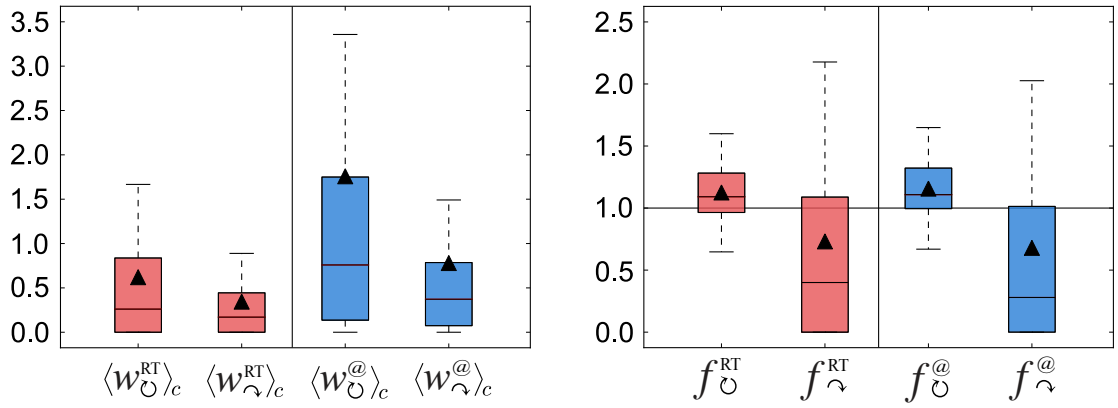


Figure 8.1: The importance of community structure in the spreading of social contagions. (a) *Structural trapping*: dense communities with few outgoing links naturally trap information flow. (b) *Social reinforcement*: people who have adopted a meme (black nodes) trigger multiple exposures to others (red nodes). In the presence of high clustering, any additional adoption is likely to produce more multiple exposures than in the case of low clustering, inducing cascades of additional adoptions. (c) *Homophily*: people in the same community (same color nodes) are more likely to be similar and to adopt the same ideas. (d) Diffusion structure based on retweets among Twitter users sharing the hashtag #usa. Blue nodes represent English users and red nodes are Arabic users. Node size and link weight are proportional to retweet activity. (e) Community structure among Twitter users sharing the hashtags #bbc and #foxnews. Blue nodes represent #bbc users, red nodes are #foxnews users, and users who have used both hashtags are green. Node size is proportional to usage (tweet) activity, and links represent mutual following relations.



(a) Community edge weight

(b) User community focus

Figure 8.2: Meme concentration in communities. We measure weights and focus in terms of retweets (RT) or mentions (@). We show (a) *community edge weight* and (b) *user community focus* using box plots. Boxes cover 50% of data and whisker cover 95%. The line and triangle in a box represent the median and mean, respectively.

8.2.1 Communication Volume

Do memes spread like complex contagions in general? If social reinforcement and homophily significantly influence the spread of memes, we expect more communication within than across communities.

Edge Weight

Let us define the weight w of an edge by the frequency of communication between two users connected by the edge. Nodes are partitioned into dense communities based on the structure of the network, but without knowledge of the weights. For each community c, the average edge weights of intra-community and inter-community links, $\langle w_{\circlearrowleft}\rangle_c$ and $\langle w_{\curvearrowright}\rangle_c$, quantify how much information flows within and across communities, respectively:

$$\langle w_{\circlearrowleft}\rangle_c = \frac{1}{|E^c_{\circlearrowleft}|}\sum_{(u,v)\in E^c_{\circlearrowleft}} w(u,v), \quad \langle w_{\curvearrowright}\rangle_c = \frac{1}{|E^c_{\curvearrowright}|}\sum_{(u,v)\in E^c_{\curvearrowright}} w(u,v). \tag{8.14}$$

where $w(u,v)$ is the weight of an edge (u,v), defined by the frequency of u retweeting ("RT") or mentioning ("@") v, noted as $w^{\mathrm{RT}}(u,v)$ or $w^{@}(u,v)$.

A common intuition that random walks on a graph tend to get trapped inside densely connected components has been employed in many community detection method [Pons and Latapy, 2005, Rosvall and Bergstrom, 2008]. Even if a community does not consist of homophilous people or stronger links, the spreading of a meme can circulate more within communities, driven by dense internal connections. To estimate the structural trapping effect of communities, we consider a random walker traversing the graph. The basic assumption is that if information spreads randomly through links and there is an infinite

number of spreading events (treating every node and link equally), the probability that a given link is used in the transmission of information will approach the probability that a random walker traverses the link. The probability of the random walker moving from a node u to another connected node v given the walker is at u is $p_{u\to v} = 1/k(u)$, where $k(u)$ is the degree of u. We can construct a transition matrix P where $p_{u\to v} = 1/k(u)$ if u and v are connected, and $p_{u\to v} = 0$ otherwise. The stationary probability of the walker stopping at a node u is the element π_u of a vector π such that $P^T\pi = \pi$. It can be shown that $\pi_u = k(u)/\sum_m k(m)$ [Lovász, 1993]. The expected amount of communication carried by edge (u, v), considering structural trapping but without any homophily or social reinforcement effects, can be computed by the probability $w^{\mathrm{rw}}(u, v)$ of the random walker traveling through the edge:

$$\begin{aligned} w^{\mathrm{rw}}(u,v) &= \pi_u p_{u\to v} + \pi_v p_{v\to u} \\ &= \frac{k(u)}{\sum_m k(m)}\frac{1}{k(u)} + \frac{k(v)}{\sum_m k(m)}\frac{1}{k(v)} = \frac{2}{\sum_m k(m)} \propto \mathrm{const.} \end{aligned} \tag{8.15}$$

In other words, a random walker, or a random spreading event, will traverse each edge with the same probability.

We measure actual edge weights by aggregating all the meme spreading events in our data. If memes spread obliviously to community structure, like simple contagions, we would expect no difference between intra-community and inter-community links. By contrast, we observe that the intra-community links carry more messages (see Fig. 8.2a). Similar results have been reported from other datasets [Onnela et al., 2007b, Grabowicz et al., 2012].

User Community Focus

In addition, we define the focus of an individual as the fraction of activity that is directed to each neighbor in the same community, $f_{\circlearrowleft}$, or in different communities, $f_{\curvearrowright}$:

$$f_{\circlearrowleft}(u) = \frac{\frac{1}{k_{\circlearrowleft}(u)}\sum_{(u,v)\in E_{\circlearrowleft}} w(u,v)}{\frac{1}{k(u)}\sum_{(u,v)\in E} w(u,v)}, \quad f_{\curvearrowright}(u) = \frac{\frac{1}{k_{\curvearrowright}(u)}\sum_{(u,v)\in E_{\curvearrowright}} w(u,v)}{\frac{1}{k(u)}\sum_{(u,v)\in E} w(u,v)} \tag{8.16}$$

where $k_{\circlearrowleft}(u)$ and $k_{\curvearrowright}(u)$ are the numbers of u's intra-community and inter-community links:

$$k_{\circlearrowleft}(u) = |\{v \mid (u,v) \in E_{\circlearrowleft}\}| \tag{8.17}$$

$$k_{\curvearrowright}(u) = |\{v \mid (u,v) \in E_{\curvearrowright}\}| \tag{8.18}$$

$$k(u) = k_{\circlearrowleft}(u) + k_{\curvearrowright}(u) \tag{8.19}$$

The ratios $f_{\circlearrowleft}$ and $f_{\curvearrowright}$ characterize how attention is directed toward a person within the same community versus a person in another community. Using the random walk analogy, the user community focus represents the probability of a random walker from a node traveling through each of its links. By definition, the random walker does not distinguish links, and thus we always have $f^{\mathrm{rw}}_{\circlearrowleft} = f^{\mathrm{rw}}_{\curvearrowright} = 1$.

Figure 8.2b shows that $f_{\circlearrowleft} > 1 > f_{\curvearrowright}$, indicating that people interact more with members of the same community. The results are robust across different activity measures and communities, and the differences are statistically significant (Mann-Whitney U test [Mann and Whitney, 1947], $p \ll 0.001$).

8.2.2 Meme Concentration

These results suggest that communities strongly trap communication. To quantify this effect for individual memes, let us define the concentration of a meme in communities. We expect more concentrated communication and meme adoption within communities if the meme spreads like a complex contagion. To gauge this effect, we introduce four baseline models. The *random sampling model* (M_1) assumes equal adoption probability for everyone, ignoring network topology and all activity. The *simple cascade model* (M_2) simulates the spreading of simple contagions [Karsai et al., 2011]. The *social reinforcement model* (M_3) employs a simple social reinforcement mechanism in addition to considering the network structure. In the *homophily model* (M_4), users prefer to adopt the same ideas that are adopted by others in the same community. The simulation mechanisms of the four baseline models are summarized in Table 8.1.

To replicate the Twitter API sampling effect in the baseline models, each simulation runs until 10 times more tweets are generated than the empirical numbers. Then, we select 10% of the tweets at random. Every simulation is repeated 100 times and the 10%-sampling is repeated 10 times on each simulation outcome. Thus, the average values of the measures from our toy models are computed across 100×10 samples.

All measures introduced below are computed only based on tweets containing each meme in its early stage (first $n = 50$ tweets) to avoid any bias from heterogenous meme popularities.

Dominance

Let us first define the concentration of a meme h based on the proportions of *tweets* in each community. The fraction of tweets with hashtag h in community c is

$$r_c(h) = \frac{T_n(h|c)}{n}. \tag{8.20}$$

The *dominant community* that produces most messages with h is:

$$c^T(h) = \arg\max_{c \in \mathcal{C}} r_c(h). \tag{8.21}$$

The *usage dominance* is the proportion of tweets produced in the dominant community out of the total number of tweets of h, quantifying the contribution to the hashtag usage from a single dominant community:

$$r(h) = r_{c^T(h)}(h) = \frac{\max_{c \in \mathcal{C}} T_n(h|c)}{n} \tag{8.22}$$

Table 8.1: Baseline models for information diffusion.

	Community effects			**Simulation implementation**
	Network	Reinforcement	Homophily	
M_1				For a given hashtag h, M_1 randomly samples the same number of tweets or users as in the real data.
M_2	✓			M_2 takes the network structure into account while neglecting social reinforcement and homophily. M_2 starts with a random seed user. At each step, with probability p, an infected node is randomly selected and one of its neighbors adopts the meme, or with probability $1-p$, the process restarts from a new seed user ($p = 0.85$). This model simulates *simple contagions*.
M_3	✓	✓		The cascade in M_3 is generated similarly to M_2 but at each step the user with the maximum number of infected neighbors adopts the meme. If there are multiple candidates with the same number of exposures, we randomly select one.
M_4	✓		✓	In M_4, the simple cascading process is simulated in the same way as in M_2 but subject to the constraint that at each step, only neighbors in the same community have a chance to adopt the meme. If there are multiple neighbors in the same community, we randomly choose one.

Analogous concentration measures can be defined based on *users*. Let $g(h)$ be the *adoption dominance* of h, i.e., the proportion of the $\mathbb{A}(h)$ adopters in the community with most adopters:

$$g_c(h) = \frac{A_n(h|c)}{A_n(h)} \tag{8.23}$$

$$c^A(h) = \arg\max_{c \in \mathcal{C}} g_c(h) \tag{8.24}$$

$$g(h) = g_{c^A(h)}(h) = \frac{\max_{c \in \mathcal{C}} A_n(h|c)}{A_n(h)}. \tag{8.25}$$

The simulations of the baseline models for each meme stop when the equal number of tweets or adopters are produced. A user can adopt the hashtag h multiple times, as a user is able to generate multiple tweets with the same hashtag on Twitter. The usage dominance $r(h)$ produced by the baseline models are labelled as $r_{M_1}(h)$, $r_{M_2}(h)$, $r_{M_3}(h)$, and $r_{M_4}(h)$, respectively. Similarly, for adoption dominance, there are $g_{M_1}(h)$, $g_{M_2}(h)$, $g_{M_3}(h)$ and $g_{M_4}(h)$. The *relative* dominances, $r(h)/r_{M_1}(h)$ and $g(h)/g_{M_1}(h)$, reflect the strength of meme concentration beyond random sampling. The higher the dominance, the stronger the concentration of the meme.

According to Fig. 8.3(a-b), we observe that non-viral memes exhibit concentration similar to or stronger than baselines M_3 or M_4, suggesting that these memes tend to spread like complex contagions. In contrast, viral memes show similar patterns as the model of simple contagion (M_2) with much weaker concentration than models of complex contagions (M_3 and M_4). Note that models M_2, M_3, and M_4 produce stronger concentration than random sampling (M_1), because M_2 incorporates the structural trapping effect in simple cascades, M_3 considers both structural trapping and social reinforcement, and M_4 captures both structural trapping and homophily.

Entropy

We also compute the *usage entropy* based on how tweets containing h are distributed across different communities:.

$$H^T(h) = -\sum_{c \in \mathcal{C}} r_c(h) \log r_c(h) = -\sum_{c \in \mathcal{C}} \frac{T_n(h|c)}{n}) \log \frac{T_n(h|c)}{n} \tag{8.26}$$

Similarly, the diversity of the distribution of meme adopters is quantified by the *adoption entropy*

$$H^A(h) = -\sum_{c \in \mathcal{C}} g_c(h) \log g_c(h) = -\sum_{c \in \mathcal{C}} \frac{A_n(h|c)}{A_n(h)} \log \frac{A_n(h|c)}{A_n(h)} \tag{8.27}$$

Again, we compare the concentration of usage and user engagement with random sampling (M_1) by computing the *relative* usage and adoption entropies, $H^T(h)/H^T_{M_1}(h)$ and

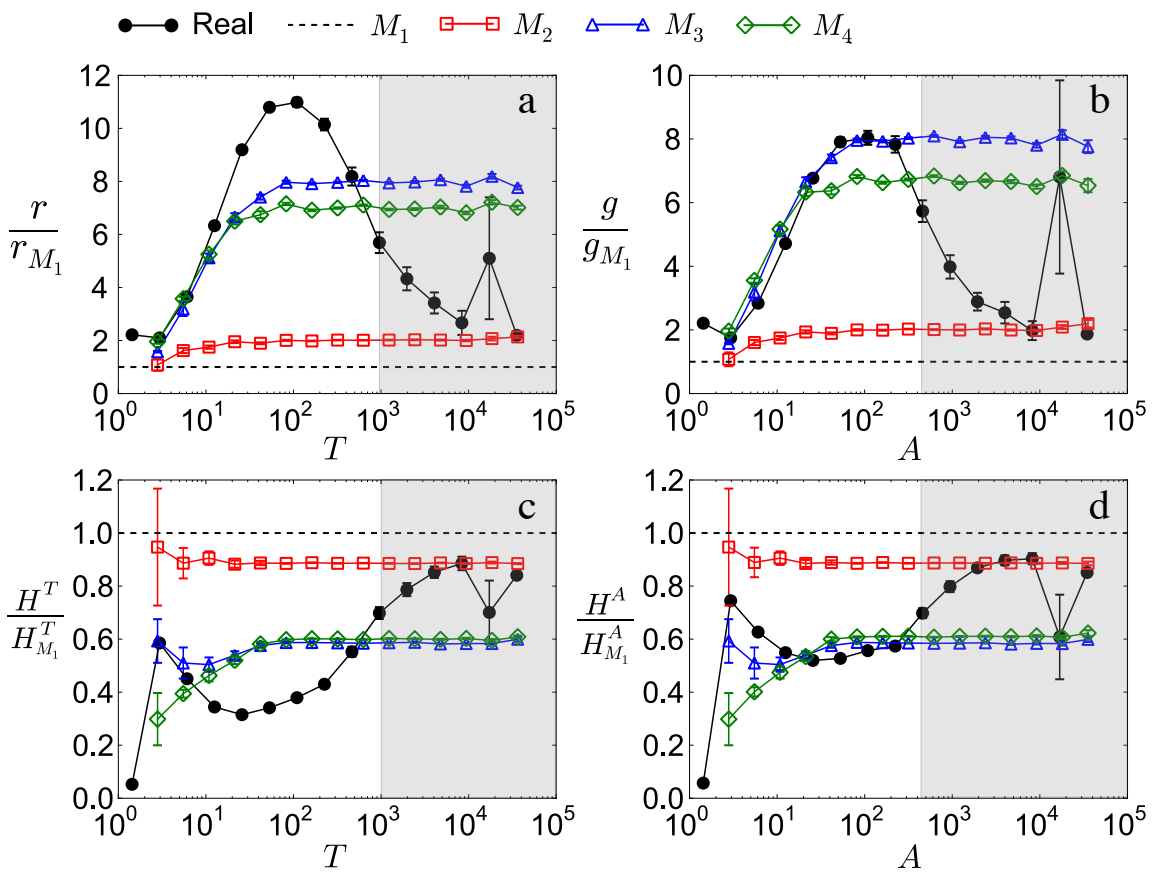


Figure 8.3: Meme concentration in communities. Changes in meme concentration as a function of meme popularity are illustrated by plotting relative (a) *usage dominance*, (b) *adoption dominance*, (c) *usage entropy*, and (d) *adoption entropy*. The relative dominance and entropy ratios are averaged across hashtags in each popularity bin, with popularity defined as number of tweets T or adopters A; error bars indicate standard errors. Gray areas represent the ranges of popularity in which actual data exhibit weaker concentration than both baseline models M_3 and M_4.

$H^A(h)/H^A_{M_1}(h)$. The lower the entropy, the stronger the concentration of the meme. Figure 8.3(c,d) demonstrate similar patterns as dominance measures: the empirical data displays the strongest concentration for memes of intermediate popularities, and, interestingly, viral memes are weakly trapped in communities as simple contagion (M_2).

Do *all* memes spread like complex contagions? While the majority of memes are not viral, viral memes are adopted differently. Their concentration in the empirical data is the same as that of the simple cascade model M_2 (see the gray areas in Fig. 8.3); community structure does not seem to trap successful memes as much as others. These memes spread like simple contagions, permeating through many communities.

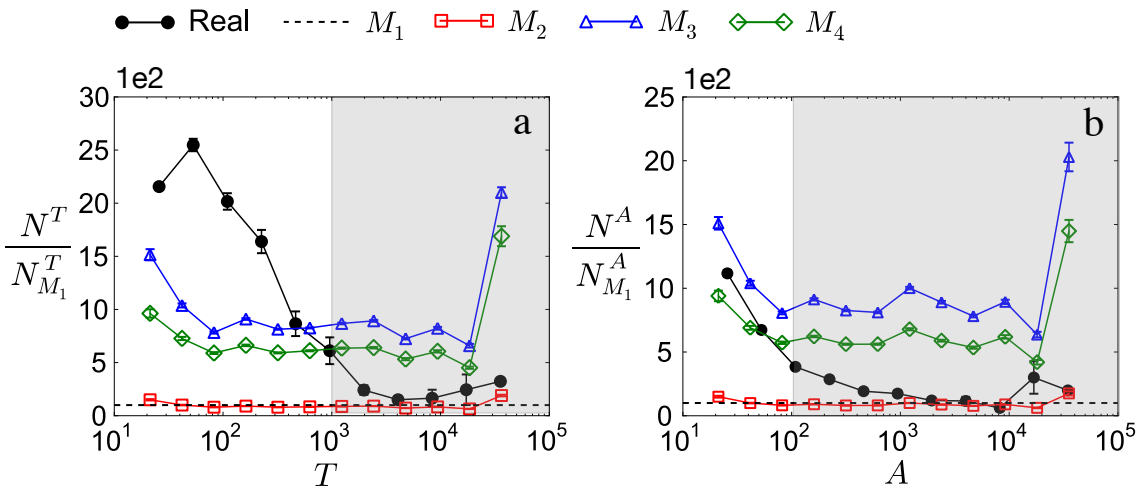


Figure 8.4: The effect of multiple social reinforcement is estimated by *average exposures* for every meme. The exposures can be measured in terms of (a) tweets or (b) users.

8.2.3 Strength of Social Reinforcement

To further distinguish viral memes from others in terms of types of contagion, let us explicitly estimate the strength of social reinforcement. For a given meme h, we count the number of exposures that each adopter has experienced before the adoption and compute the *average exposures* across all adopters, representing the strength of social reinforcement on h, labelled as $N(h)$. The exposures can be measured in terms of tweets $N^T(h)$ or users $N^A(h)$. We compute relative average exposures, $N^T(h)/N^T_{M_1}(h)$ and $N^A(h)/N^A_{M_1}(h)$, using only tweets at the early stages (first $n = 50$ tweets). If this quantity is large, adoptions are more likely to happen with multiple social reinforcement and thus the meme spreads more like a complex contagion. As shown in Fig. 8.4, viral memes require as little reinforcement as the simple cascade model M_2, while non-viral memes need as many exposures as M_3 or M_4. We arrive at the same conclusion: viral memes spread like simple contagions rather than like complex ones.

8.3 Prediction Model

The above findings imply an intriguing possibility: high concentration of a meme would hint that the meme is only interesting to certain communities, while weak concentration would imply a universal appeal and therefore might be used to detect viral memes. Meanwhile, we should notice that many factors contribute to the virality of memes. First, a meme may become viral simply because the meme appeals to many [Berger and Milkman, 2009, Cataldi et al., 2010]. However, given the competition between memes and social influence, innate appeal alone may not be able to paint the whole picture [Salganik et al., 2006, Kitsak et al., 2010, Weng et al., 2012]. The success of a meme also depends on timing, network structure, randomness, and many other factors [Centola, 2010, Weng et al., 2012, Pinto

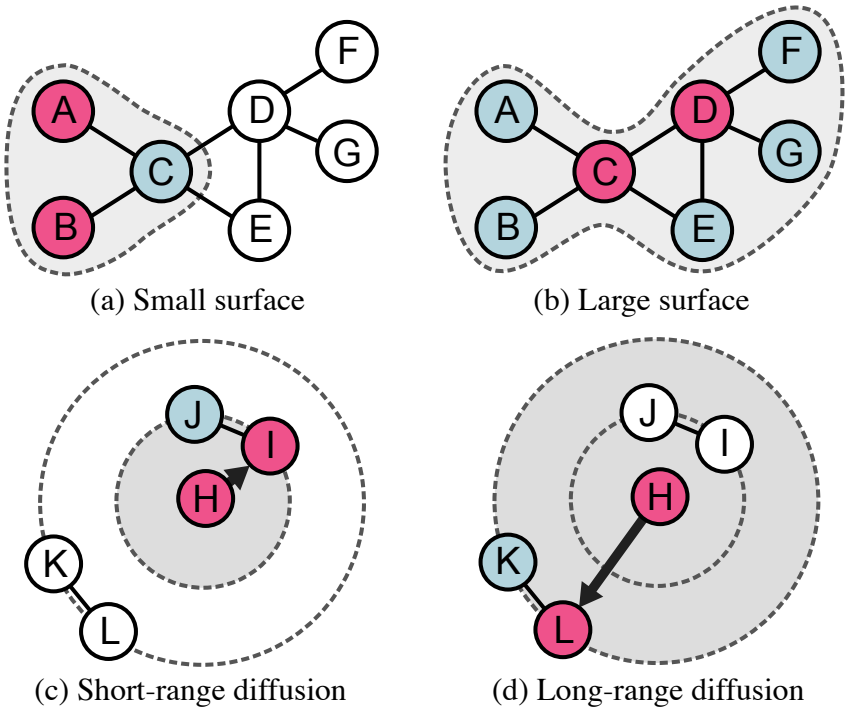
et al., 2013]. Other than the early spreading patterns of viral memes in terms of community structure (see Sec. 8.2), we also present several other categories of features in this section for predicting future meme success and combine them into one prediction model [Weng et al., 2014c].

We identify two major approaches to meme virality prediction: time series analysis and feature-based classification. Time series analyses focus on the patterns of early popularity fluctuation of a meme, assuming that the patterns of a meme's growth and decay tell us whether it will go viral in the future [Jamali and Rangwala, 2009, Asur et al., 2011, Yang and Leskovec, 2011]. Classification approaches commonly aim to discover distinguishing features of successful memes by applying supervised machine learning techniques with labeled datasets. A variety of features have been proposed and tested to differentiate viral memes from others; examples include comments, votes, and user-defined groups [Lerman and Hogg, 2010, Jamali and Rangwala, 2009, Suh et al., 2010, Hong et al., 2011, Yang et al., 2012]. However, most studies have paid little attention to the role of the underlying network structure [Colbaugh and Glass, 2012, Romero et al., 2013, Ma et al., 2013] even though it is natural to expect network topology to affect information diffusion, as memes spread through social ties.

In this section, we examine the predictive power of *community structure*, as it was shown in Sec. 8.2.2 that the spreading pattern of a meme across communities reveals the general appeal of the meme. Besides, we investigate the other two categories of features. First, features based on network topology are designed to capture the *audience size*. As many studies on social influence have assumed, the neighbors of an individual in the network can be considered as his potential audience [Kitsak et al., 2010, Cha et al., 2010, Suh et al., 2010, Bakshy et al., 2011]. For example, one of the common beliefs is that star users with lots of followers are more influential than others with fewer followers. Second, we take into account the *speed of growth* in early meme adoption. By comparing with multiple representative prediction models, we show that our model can accurately predict the popularity of memes (to an order of magnitude) after 2 months with knowledge of only a small number of early tweets. Our model outperforms random guessing, majority guessing, and three regression models that use early popularity or expected influence of early adopters for detecting the most viral memes.

8.3.1 Characterizing Viral Memes

In the following discussion, we identify signatures of viral memes at their early stages in terms of three characteristics: *network topology*, *growth rate*, and *community diversity*. We demonstrate that the information on early adopters, particularly in the context of social network structure, is powerful enough to identify young viral memes. Let us present the rationales for the prediction features used in the model before introducing the detailed definition of each feature in the next section.



(a) Small surface (b) Large surface

(c) Short-range diffusion (d) Long-range diffusion

Figure 8.5: Network surfaces. Red nodes represent the early adopters of a meme; blue ones are neighbors of early adopters; the grey color marks the surface area. (a When A and B adopt a meme, the corresponding network surface is small. (b) The adoption by C and D creates a large network surface. (c) Adopter H spreads the meme to node I (a nearby node), and the potential adopters do not change much. (d) H spread the meme to node L (a node farther away), and consequently the set of potential meme adopter grows a lot.

Network Topology

The position of an adopter in the network determines the size of the potential audience [Kitsak et al., 2010]. The network surface of a given set of adopters S captures the number of neighbors who are directly exposed. As illustrated in Fig. 8.5(a-b), the network surface varies greatly depending on the degrees and positions of the adopters. We also estimate the growth of potential audience in time by examining the distance between consecutive adopters in the network. Note that new adopters are not necessarily connected to existing adopters because a meme can be injected into multiple nodes of the network, and because our collection is based on a sample of the entire public stream. The longer the jump between two consecutive adopters, the more potential spreaders the meme may have. Figures 8.5(c-d) compare the potential adopters of two spreading events.

Meme Growth Rate

Viral memes are expected to spread more quickly than others [Szabo and Huberman, 2010]. To incorporate this intuition, we define the time difference between the first and the n-th

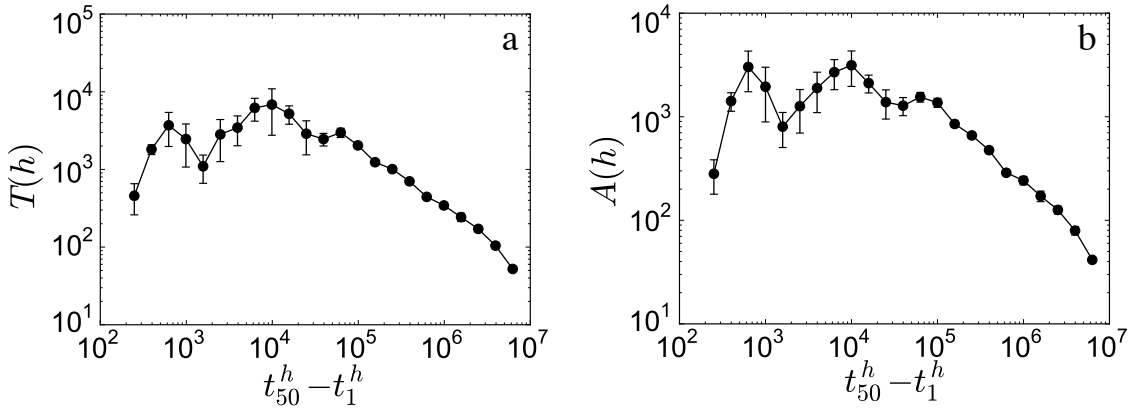


Figure 8.6: The relationship between the meme popularity measured in the number of tweets $T(h)$ and adopters $A(h)$ and the early spreading time with $n = 50$, $t_{50}^h - t_1^h$ seconds.

tweet in the time series of a meme h as the *early spreading time*, $t_n^h - t_1^h$. It gauges the initial *growth rate* of h. Figure 8.6 displays a correlation between the growth rate and meme popularity. We assume that there are only a very small number of memes that go viral after two months (our observation window). Although we observe some fluctuations when the early spreading time $t_{50}^h - t_1^h$ is small, meme popularity significantly decreases when the early spreading is slow.

Community Diversity

The previous examinations propose the predictive power of the community structure, as weak concentration of early meme adopters among communities implies a universal appeal of the meme and high future virality (see Sec. 8.2.2). To illustrate the intuition, we show in Fig. 8.7 how the diffusion pattern of a viral meme differs from that of a non-viral one, when analyzed through the lens of community concentration.

8.3.2 Prediction Features

Based on our preliminary analyses above, we design several features for our prediction model. Network features describe the size of potential audience based on the positions of early adopters in the network. Growth-rate features quantify the initial momentum. Community features measure the community diversity at the early stage. We have 13 features in total, marked as $f.1$–13. All the features are computed based on the first n tweets for each hashtag, where the parameter n is a relatively small number compared to the final number of tweets generated by viral memes.

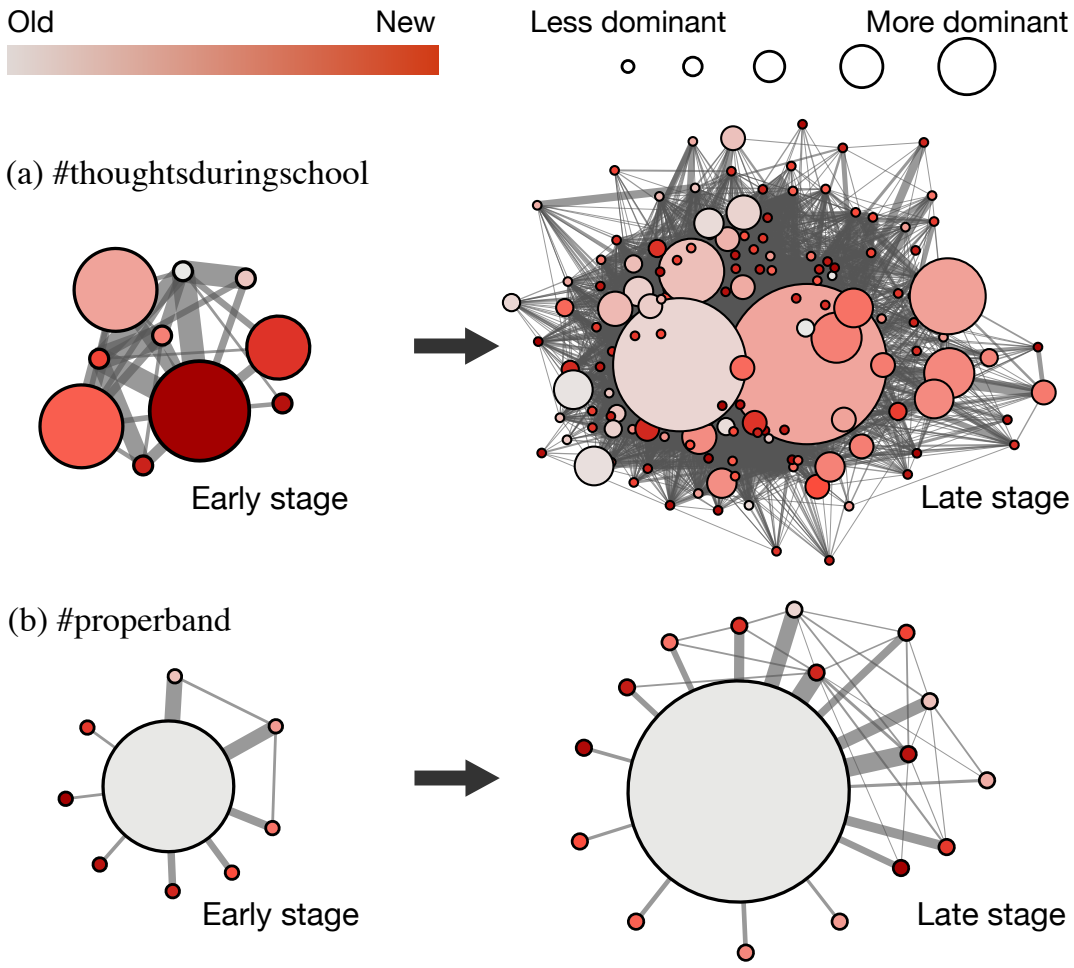


Figure 8.7: Evolution of two contrasting memes (viral vs. non-viral) in terms of community structure. We represent each community as a node, whose size is proportional to the number of tweets produced by the community. The color of a community represents the time when the hashtag is first used in the community. (a) The evolution of a viral meme (#thoughtsduringschool) from the early stage (30 tweets) to the late stage (200 tweets) of diffusion. (b) The evolution of a non-viral meme (#properband) from the early stage to the final stage (65 tweets).

Basic Network Features

Here we use the connectivity of users in the network.

$f.1$. **Number of early adopters**, $A_n(h)$. Among the earliest n tweets of a meme h, it refers to the set of distinct adopters. The number of early adopters is one of the most basic and simple features. A small $A_n(h)$ would indicate that a small number of users generated most tweets and the hashtag is failing to spread.

$f.2$. **Size of first surface**, $|S(\mathbb{A}_n(h))|$. The first surface contains all the uninfected neighbors of early adopters of h. It is a set of the most immediate adopter candidates [Ma et al., 2013].

$f.3$. **Size of second surface**, $|S^2(\mathbb{A}_n(h))|$. The second surface includes uninfected users in the second surface of early adopters, characterizing the number of potential adopters within two steps.

Distance Features

Here we use the position of adopters in the network.

$f.4$. **Average step distance**, $\overline{d_n(h)}$. We examine the average distance between consecutive adopters of h in time:

$$\overline{d_n(h)} = \frac{1}{n-1}\sum_{i=1}^{n-1} d(a_i^h, a_{i+1}^h). \tag{8.28}$$

$f.5$. **CV of step distances**, $C_v(d_n(h))$. The coefficient of variation (C_v) of a variable is the ratio of its standard deviation to the mean. We use it to measure the relative variability in step distance:

$$C_v(d_n(h)) = \frac{1}{\overline{d_n(h)}}\sqrt{\frac{\sum_{i=1}^{n-1}(d(a_i^h, a_{i+1}^h) - \overline{d_n(h)})^2}{n-2}}. \tag{8.29}$$

$f.6$. **Diameter**, $D_n(h)$. The diameter is the maximum distance between any two adopters of h within the first n tweets. It is a measure of audience coverage in the network:

$$D_n(h) = \max_{1\le i\neq j\le n-1} d(a_i^h, a_j^h). \tag{8.30}$$

Growth Rate Features

The mean and fluctuations of the sequence of step time durations, $\Delta t_i(h)$, where $1 \le i \le n-1$, are implemented as two prediction features.

$f.7$. **Average step time duration**, $\overline{\Delta t_n(h)}$:

$$\overline{\Delta t_n(h)} = \frac{\sum_{i=1}^{n-1} t_{i+1}^h - t_i^h}{n-1} = \frac{t_n^h - t_1^h}{n-1}. \tag{8.31}$$

$f.8$. **CV of step time durations**, $C_v(\Delta t_n(h))$:

$$C_v(\Delta t_n(h)) = \frac{1}{\overline{\Delta t_n(h)}} \sqrt{\frac{\sum_{i=1}^{n-1}(t_{i+1}^h - t_i^h - \overline{\Delta t_n(h)})^2}{n-2}}. \tag{8.32}$$

Community Features

Community-based features are computed at the prediction time, based on the predefined community structure; the community detection algorithm is executed once on the network built upon the historical data, as the network structure does not evolve much within a short time period.

$f.9$. **Number of infected communities**, $C_n(h)$. It is the number of communities with at least one adopter of h among first n tweets.

$f.10$–11. **Usage and adopter entropy**, $H_n^T(h)$ and $H_n^A(h)$. See definition in Sec. 8.2.2. Large entropy indicates high diversity and low concentration.

$f.12$–13. **Fraction of intra-community user interactions**, $I_n^{\circlearrowleft}(h)/I_n(h)$. The likelihood of a user adopting information from members of the same community increases with the strength of the community trapping effect. We expect to observe weaker community trapping and higher community diversity among early adopters of viral memes. Here we quantify this by measuring the fraction of intra-community user interactions. The interactions can be retweets or mentions:

$$\frac{I_n^{\circlearrowleft \mathrm{RT}}(h)}{I_n^{\mathrm{RT}}(h)}, \quad \frac{I_n^{\circlearrowleft @}(h)}{I_n^{@}(h)}. \tag{8.33}$$

A high fraction of intra-community user interactions implies a strong effect of homophily and social reinforcement on the spread of the meme. Figure 8.8 suggests that a meme with high $\frac{I_n^{\circlearrowleft}}{I_n}$ may have less potential in spreading out to multiple communities to become popular. Interestingly, highly viral memes ($T(h) \geq 10^4$ or $A(h) \geq 10^4$) also tend to have many intra-community transmissions—viral memes circulate well inside communities while having small dominance and high entropy.

8.3.3 Experiments

In this section we predict the magnitude of a meme's future popularity using the features introduced above, calculated on the basis of early observation, and compare the results with five baselines.

Task Definition

We define the popularity (virality) of a meme h as the number of tweets $T(h)$ or adopters $A(h)$. We use both definitions, as they highlight different perspectives of a meme: the

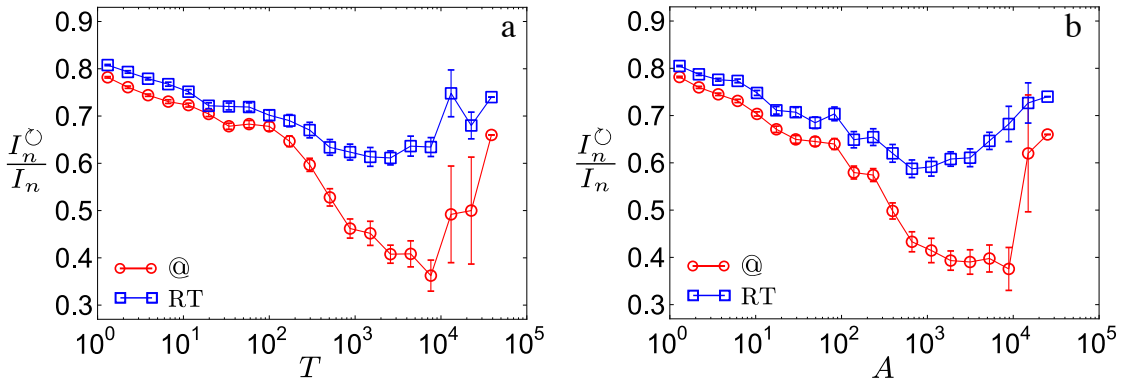


Figure 8.8: The relationship between the fraction of intra-community user interactions and popularity measured in the number of (a) tweets $T(h)$ and (b) adopters $A(h)$. The measures of $\frac{I_n^{\circlearrowleft@}}{I_n^{@}}$ and $\frac{I_n^{\circlearrowleft\mathrm{RT}}}{I_n^{\mathrm{RT}}}$ are computed using the first $n = 50$ tweets of each hashtag.

former characterizes the amount of discussion a meme triggers; the latter tells us about the size of the crowd participating in the discussion. Large $T(h)$ does not necessarily implies large $A(h)$, because a single user may generate many tweets.

Meme popularity exhibits a broad and skewed distribution, as observed in many previous studies [Lerman and Ghosh, 2010, Weng et al., 2012]. We partition all the memes into classes based on the order of magnitude of the total popularity ($\lceil \log_{10} T(h) + 0.5 \rceil$ or $\lceil \log_{10} A(h) + 0.5 \rceil$). The prediction task is therefore a *multi-label classification*. Given the information about the early stage of a hashtag, the task is to predict which class it belongs to after about two months, at the end of the observation period of our dataset.

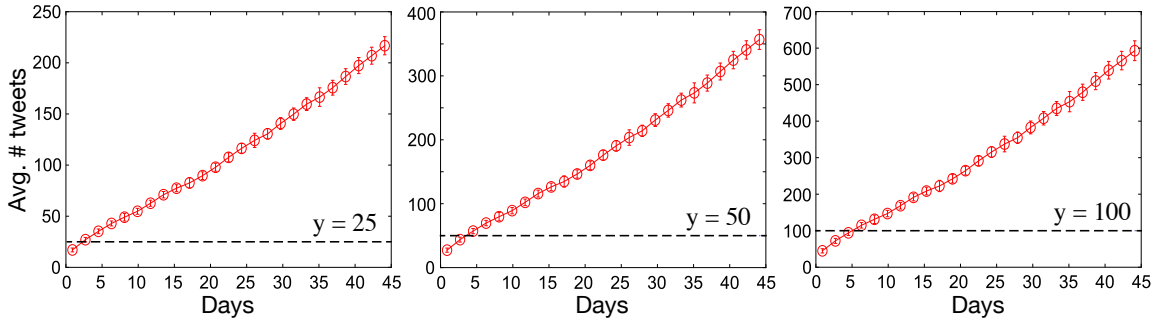
Baselines

We evaluate our prediction results by comparing them with five baseline prediction models: B_1 and B_2 are trivial baselines; B_3, B_4, and B_5 are regression models that use features such as social influence of adopters, cumulative popularity, and the growth sequence of memes. Note that content-based prediction models, such as the model proposed by Tsur and Rappoport [2012], are not considered as we focus on the prediction problem using only network spreading patterns, without looking into the content.

1. *Random guess* (B_1): Assuming that we know the exact number of memes in each class, B_1 randomly assign the class label to each meme with the prior probability.

2. *Majority guess* (B_2): Due to the imbalanced distribution of meme popularity, simply assigning the dominant class label to every meme yields high accuracy. Note that, however, B_2 fails to capture the most important but not dominant class—the most viral memes. This simple but 'powerful' baseline has been ignored in most existing studies.

3. *Social influence model* (B_3): This is built on the common notion that influential users play a key role in the wide adoption of a meme [Kitsak et al., 2010, Cha et al., 2010,



(a) $n = 25$ (b) $n = 50$ (c) $n = 100$

Figure 8.9: The average number of tweets for memes with a given minimum n as a function of time since creation. The dashed lines mark where memes get n tweets. We consider (a) $n = 25$, (b) $n = 50$, and (c) $n = 100$.

Suh et al., 2010, Bakshy et al., 2011]. We calculate each user's PageRank score [Brin and Page, 1998] and number of followers, which approximately captures the importance of the user in the network and the size of potential viewers of his content, respectively.[1] According to the social-influence perspective, if a meme is reposted by more influential people at the early stage, it is more likely to go viral. For each given meme, we therefore compute the maximum, mean, median, and coefficient of variation of PageRank scores of its n early adopters; the other feature set is built similarly with the follower counts, but on a logarithmic scale. We then apply multivariate linear regression using these eight features as one baseline model.

4. *LN model* (B_4): Szabo and Huberman [2010] proposed a linear regression (LN) model that uses the logarithm of the early popularity of a meme at time τ, $\log T^{\tau}(h)$, to predict its popularity in logarithm in the future, $\log T(h)$. Given that we use parameter values $n = 25, 50, 100$, we set $\tau = 7$ days, as it takes at most 7 days on average to obtain the numbers of tweets required (see Fig. 8.9).

5. *ML model* (B_5): The multivariate linear (ML) model, built upon Szabo and Huberman's linear regression model, was proposed by Pinto et al. [2013]. Instead of using the cumulative popularity reached by a meme on a given day, the model takes the popularity measured on each day up to time τ to form a vector as the predictor for the future popularity. We set $\tau = 7$ same as in B_4.

Network-based Prediction Model (P_n)

Since we focus on identifying the predictive features, we choose one of the most widely adopted methods—the random forest algorithm—that has been shown to be robust and reliable [Breiman, 2001]. We construct 300 decision trees, each with 5 random features from those introduced earlier. Our prediction model P_n uses the features computed with the

[1]The network is undirected and unweighted and therefore PageRank is proportional to the degree but not generally identical [Grolmusz, 2012].

Table 8.2: The number of emergent memes in each class with different n values. Note that only 48 memes in the dataset reach the order of 10^4 tweets and only 33 memes reach the order of 10^4 adopters.

n	$\lceil \log T(h) + 0.5 \rceil$				$\lceil \log A(h) + 0.5 \rceil$				
	1	2	3	$\geq$4	0	1	2	3	$\geq$4
25	2,853	6,227	224	48	157	5,202	3,810	149	33
50	-	2,761	224	48	21	723	2,106	149	33
100	-	676	224	48	4	118	643	149	33

first n tweets of each emergent meme. Note that hashtags with fewer than n tweets are not considered in the calculation. We experiment with multiple values of n; the corresponding number of emergent memes in each class is listed in Table 8.2.

Evaluation with F_1 Score

Simply computing the accuracy, the percentage of correctly predicted items among all the items, is not good enough for evaluation in our prediction task, because the classes in our task are imbalanced (see Table 8.2). When class sizes are skewed, a high accuracy does not necessarily indicate good performance. Overlooking small classes, as done by the majority guess baseline B_2, can yield good accuracy, when one or a few dominant classes are over-represented in the dataset.

Instead we measure F_1 scores for predicting the future usage or adopter popularity of memes and the results are displayed in Fig. 8.10 and Fig. 8.11. For both P_n and all baselines, we employ 10-fold cross validation. To quantify and compare how each set of features in P_n performs, we also run the models with only basic features (f.1-3), distance features (f.4-6), timing features (f.7-8), and community-based features (f.9-13).

Results

All models, including the two trivial baselines (B_1 and B_2), achieve good results for dominant classes ($\lceil \log T(h) + 0.5 \rceil = 2$ or $\lceil \log A(h) + 0.5 \rceil = 2$), due to the imbalanced class sizes. Note that B_2 can only achieve non-zero F_1 score in the dominant class. Regression models (B_3, B_4, and B_5) in general have similar performance. We find that the LN baseline model does not work well for the most viral hashtags, because the popularity at the early stage does not guarantee future popularity, in contrast to the common premise of many studies. The correlation between the early popularity $T^\tau(h)$ and the final popularity $T(h)$, as illustrated in Fig. 8.12a, is weak. This suggests that many initially unpopular hashtags eventually become popular later (cf. upper left quadrant in Fig. 8.12a). It should be noted that the LN model was originally designed for predicting the popularity of a single piece of online content, such as a Digg story, a YouTube video, or a single tweet, which tends to have swift growth and decay within a shorter lifetime. By contrast, hashtag usage

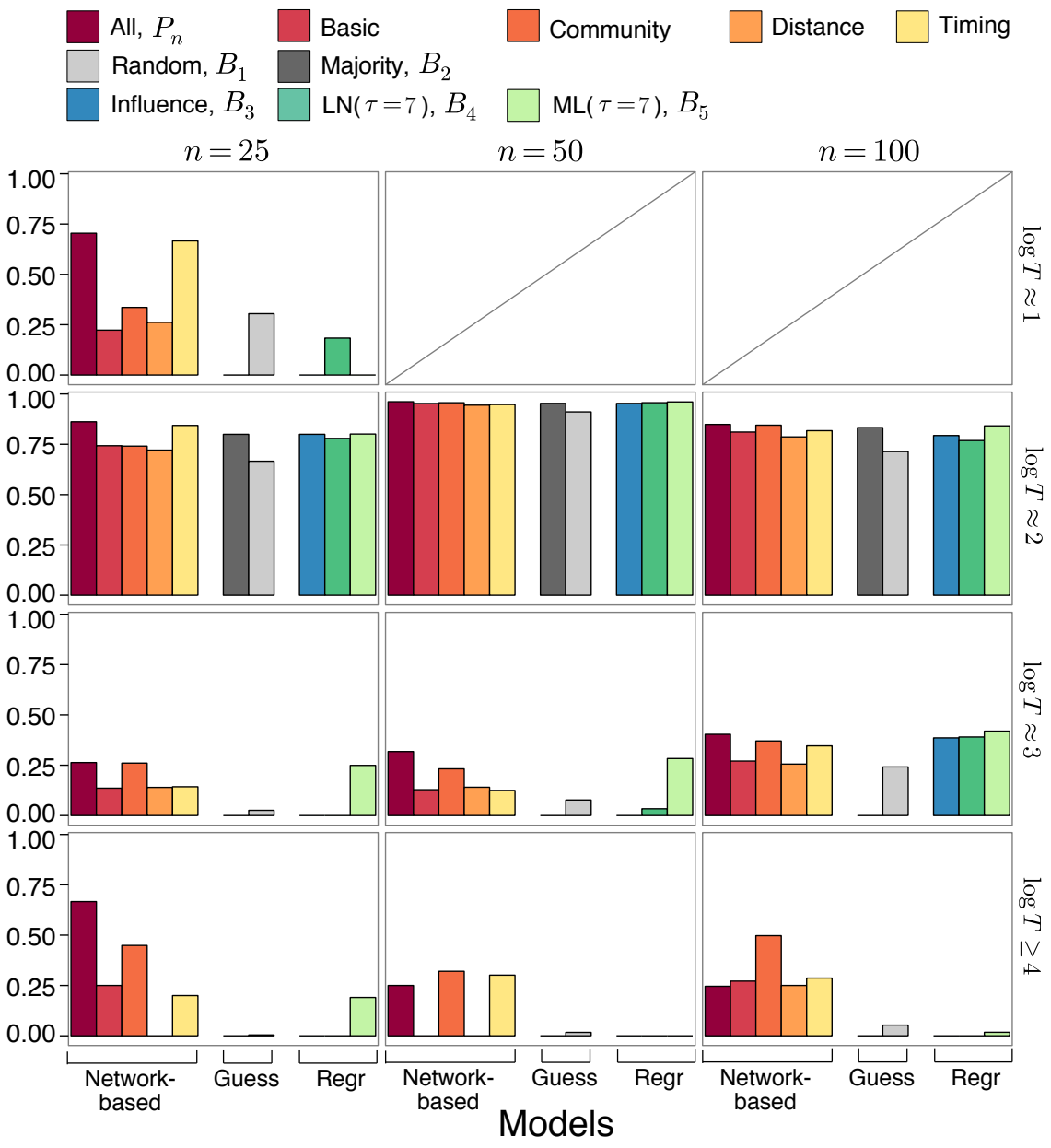


Figure 8.10: F_1 scores of various models predicting future meme popularity classes measured in the number of tweets, $\lceil \log T + 0.5 \rceil = 1, 2, 3, 4$. The observation window is set to $n = 25, 50, 100$ tweets, respectively. Here we only demonstrate the results using the Infomap community detection method; link clustering yields similar results.

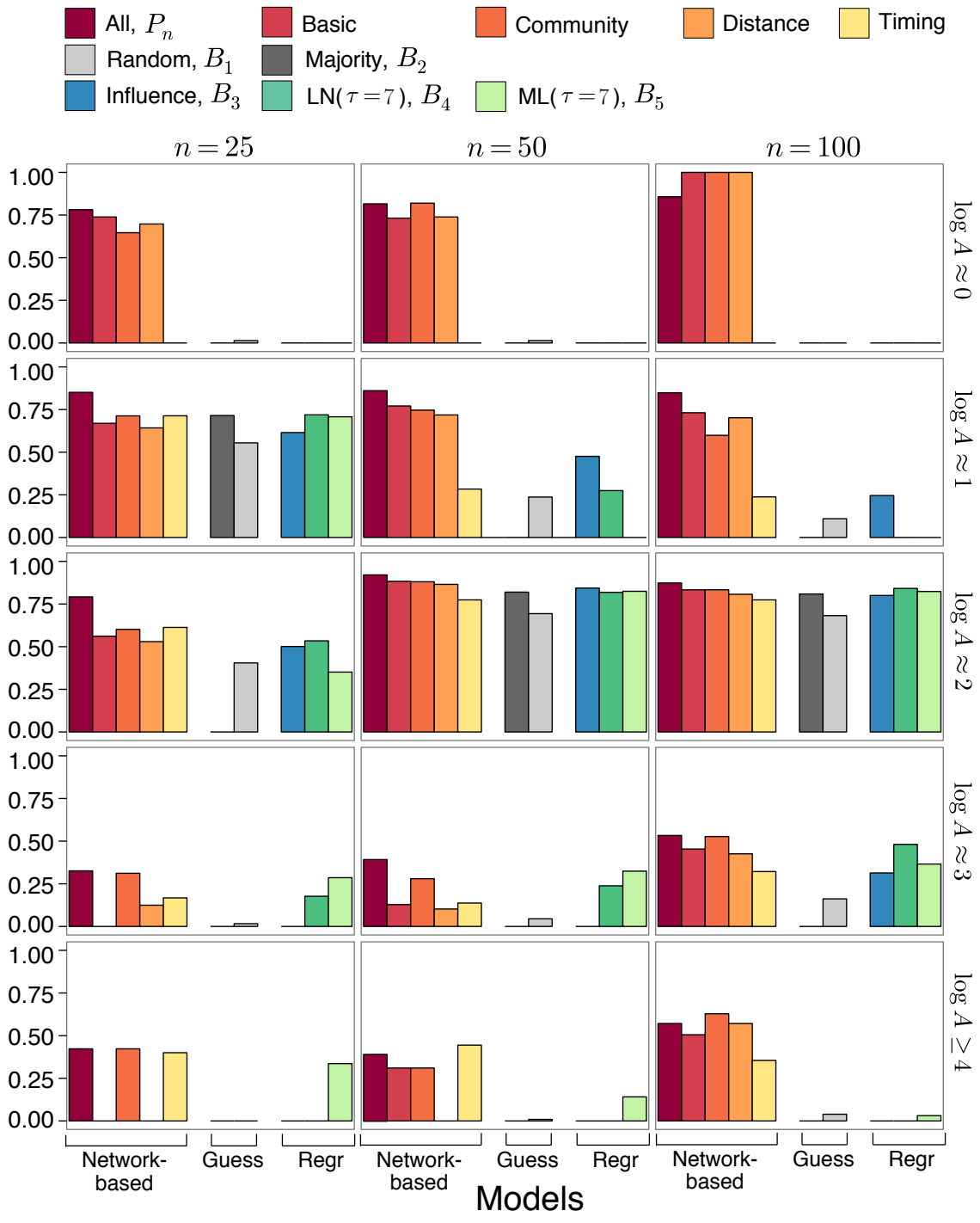


Figure 8.11: F_1 scores of various models predicting future meme popularity classes measured in the number of adopters, $\lceil \log A + 0.5 \rceil = 0, 1, 2, 3, 4$. The observation window is set to $n = 25, 50, 100$ tweets, respectively. Here we only demonstrate the results using the Infomap community detection method; link clustering yields similar results.

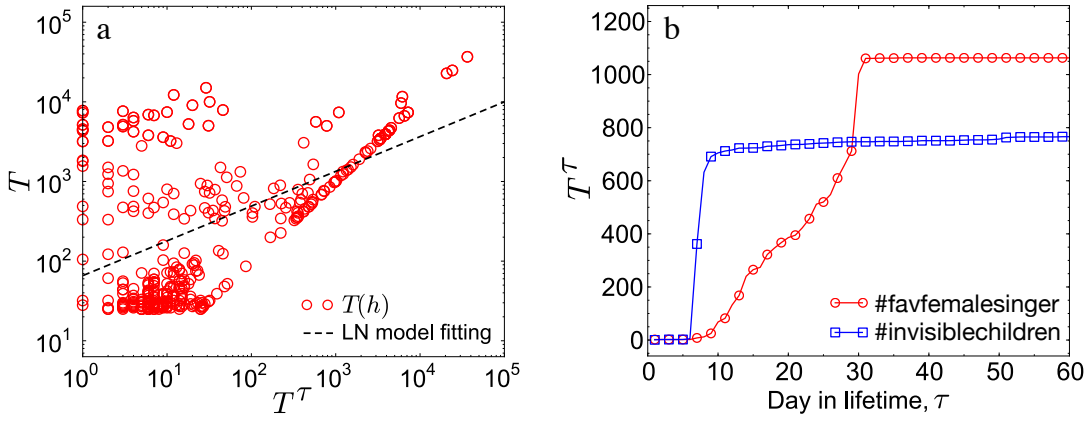


Figure 8.12: (a) Scatter plot of early popularity T^τ versus T for each meme; the black dashed line is the regression line by the LN model. (b) Cumulative popularity for two hashtags, #favfemalesinger and #invisiblechildren.

seems to be affected more by long-term endogenous diffusion processes on the network. For instance, in Fig. 8.12b, the hashtag #favfemalesinger had fewer tweets than #invisiblechildren during the first 2 weeks, but it continued to grow and eventually became more popular than #invisiblechildren, while #invisiblechildren obtained new tweets slowly after the early burst. The LN model may work better for foretelling the future popularity (number of retweets) of a single tweet, but not for hashtags. The ML baseline (B_5) captures more viral memes compared to the other two regression models. The richer description of early growth patterns contained in a meme's daily usage vector yields improved prediction quality.

Our network-based approach P_n outperforms the five baselines in most cases, especially for the most viral hashtags ($\lceil \log T(h) + 0.5 \rceil \geq 4$ or $\lceil \log A(h) + 0.5 \rceil \geq 4$) or hashtags with a small number of adopters ($\lceil \log A(h) + 0.5 \rceil \leq 1$) when all other baseline models fail to correctly classify any instances. Basic network features are weak for viral memes, but good enough for dominant classes. Timing-based features work better for estimating future usage, while distance-based features are more helpful for predicting the number of adopters. Community-based features yield the best results in general, particularly when detecting the classes of very popular memes. By combining all the features together, P_n provides the best overall results. The network-based approach outperforms all baselines in detecting rare events—extremely popular and extremely unpopular memes.

The influence model and features of basic network topology, distance, and community structure require knowledge about the network and the positions of early adopters, while the LN model, ML model, and timing features need the timestamps of early messages containing the meme. It is noteworthy that community-based features extract extended values out of the network topology by detecting community structure so as to provide better prediction outcomes. Depending on what type of information is available, one might choose different approaches.

8.4 Discussion

Despite the vast and growing literature on network communities, the importance of community structure has not yet been fully explored and understood. Our findings expose an important role of community structure in the diffusion of memes. While the role of weak ties between different social groups in information diffusion has been recognized for decades [Granovetter, 1973, Onnela et al., 2007b], we provide a direct approach for translating data about community structure into predictive knowledge about what information will spread virally. We then test and compare features built upon community diversity, network topology, and time series of early growth with machine learning techniques. The evaluation is executed against two simple baselines, as well as three more sophisticated regression models. Our proposed model excels all baselines in most cases, especially for detecting memes in minor but crucial classes. Our method does not exploit message content, and can be easily applied to any socio-technical systems using a small sample of data.

The work presented in this chapter offers a novel view on the relationship between information diffusion and community structure: viral memes are less trapped by communities and powerful features for predicting meme success can be built accordingly. We also present the first comprehensive analysis comparing multiple approaches for early prediction of viral memes. The ability to predict whether a meme can grow popular in the future by just observing a few early messages provides us with many potential applications in social media analytics, marketing, and advertisement. Further analyses of network community structure in relation to social processes hold potential for characterizing and forecasting social behavior. We believe that many other complex dynamics of human society, from ethnic tension to global conflicts, and from grassroots social movements to political campaigns [Borge-Holthoefer et al., 2011, Conover et al., 2013b,a], could be better understood by continued investigation of network structure.

Chapter 9

Network Evolution

Network topology indeed affects the spread of information among people, providing powerful indicators for predicting future meme popularity. Meanwhile, traffic flow in turn influences the formation of new links and ultimately alters the shape of the network. Much effort in the past was devoted to modeling the evolution of complex networks [Wasserman and Faust, 1994, Barabási and Albert, 1999, Albert and Barabási, 2002, Barrat et al., 2008, Newman, 2010]. Among proposed mechanisms of how a link is created, *triadic closure* is a simple but powerful principle to model the evolution of social networks based on shared friends: two individuals with mutual friends have a higher than random chance to establish a new contact [Simmel and Wolff, 1950, Granovetter, 1973, Leskovec et al., 2008, Krackhardt and Handcock, 2007, Romero and Kleinberg, 2010]. However, most existing models do not take user activity—how information spreads on the network—into consideration. Social micro-blogging networks, such as Twitter, Google Plus, Sina Weibo, and Yahoo! Meme, are designed for information sharing. As illustrated in Fig. 9.1, the dynamics on the network directly affects the dynamics of the networks and vice versa. In this chapter, we probe into the effect of information diffusion in shaping the evolution of the social network structure, using the dataset of Yahoo! Meme. With the exploration of information sharing and social link formation at the same time, we are able to depict a better and more comprehensive view of the network evolution by understanding how the structural growth of the system is coupled with information diffusion processes [Weng et al., 2013b].

Particularly we study the role of information flow in determining network growth, and the individual strategies that bring about this effect by way of creating social links. We characterize link creation processes with a set of parameters associated with different link creation strategies, estimated by a Maximum-Likelihood approach [Cowan, 1998], showing that triadic closure does have a strong effect on link formation, but shortcuts based on traffic are another indispensable factor in interpreting network evolution. This suggests a link creation mechanism whereby Alice is more likely to follow Charlie after seeing many messages by Charlie. However, individual strategies for following other users are highly heterogeneous. Link creation behaviors can be summarized by classifying users in different categories with distinct structural and behavioral characteristics. Users who are popular, active, and influential tend to create traffic-based shortcuts, making the information diffusion process more

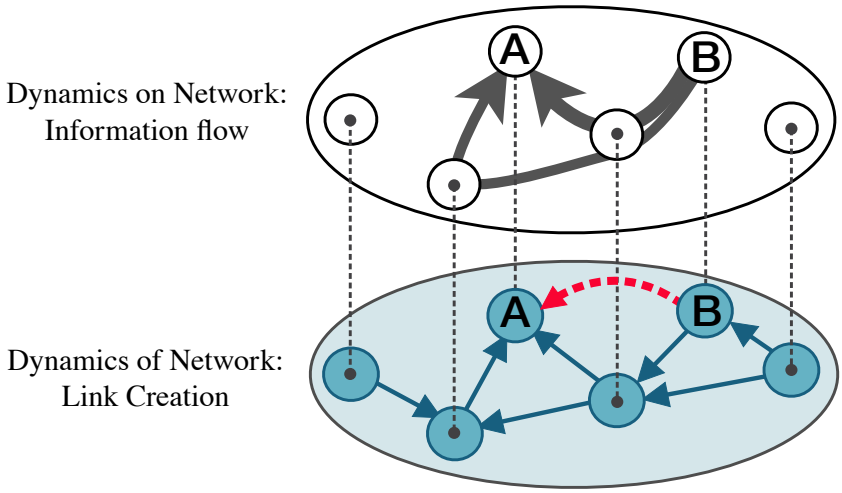


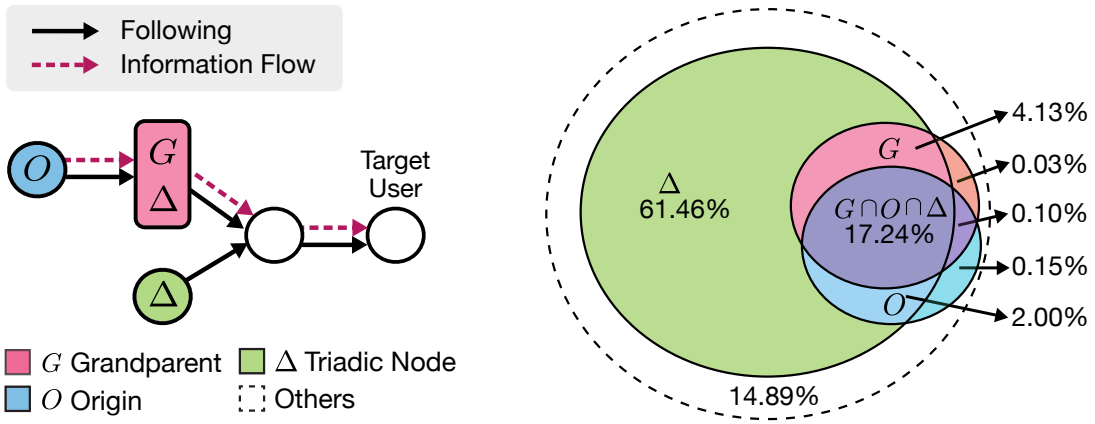
Figure 9.1: The dynamics *of* and *on* the network are strongly coupled. The bottom layer illustrates the social network structure, where the blue arrows represent "follow" relationships with the direction of information flow. The dashed red arrow marks a newly created link. The upper layer depicts the flow of information between people in the same group, leading to the creation of the new link. The social network structure constrains communication patterns, but information propagated through the network also affect how agents behave and ultimately how the network changes and grows.

efficient in the network.

9.1 Link Creation Mechanisms

When users post or repost messages, all their followers can see these posts and might decide to repost them, generating paths that together form cascade networks. When receiving a reposted message, a user in such a path can see both the *grandparent* (G, the user two steps ahead in the path) and the *origin* (O, original source). A user may decide to follow a grandparent or origin, receiving their future messages directly. These new links create *shortcuts* connecting users at any distance in the network. A triadic closure occurs when a user follows a *triadic node* (Δ, the user two steps away in the follower network). The definitions of different types of link creation mechanisms are illustrated in Fig. 9.2a.

The Venn diagram in Fig. 9.2b shows the proportions of links of different types and the logical relationships between these sets of links in our Yahoo! Meme dataset. We observe that 84.8% of new edges consist of triadic closures, 21.5% form shortcuts to grandparent, and 19.5% to origins. Note that not all the grandparents are triadic nodes, because users are allowed to repost messages from people they are not following in Yahoo! Meme. This account for 0.03% of links. There is a large overlap between triadic closure links and traffic-based shortcuts. This can be explained by the phenomenon that most real-world information cascades are shallow [Bakshy et al., 2011] and thus triadic closure links and



(a) Definition of traffic-based shortcuts

(b) Venn diagram

Figure 9.2: (a) Illustration of the link creation mechanisms. (b) Venn digram of the proportions of grandparent, origin, and triadic closure links among all existing edges.

traffic-based shortcuts coincide.

This evidence suggests that traffic-based link creation mechanisms are an important complement to the triadic closure in modeling network evolution. Actions of posting and reposting induce the creation of shortcuts, shaping the structure of the network. Newly created links in turn determine what messages are seen by users, making the network more efficient at spreading information.

9.1.1 Statistical Analyses of Shortcuts

To quantify the statistical tendency of users to create shortcuts, let us consider every single link creation in the data as an independent event. We test the null hypotheses that links to grandparents, origins, and triadic nodes are generated by choosing targets at random among the users not already followed by the creator.

We label each link ℓ by its creation order, $1 \leq \ell \leq L$, where L is the total number of links. For each link, we can compute the likelihood of following a grandparent by chance:

$$p_G(\ell) = \frac{N_G(\ell)}{N(\ell) - k(\ell) - 1} \tag{9.1}$$

where $N_G(\ell)$ is the number of distinct grandparents seen by the creator of ℓ at the moment when ℓ is about to be created; $N(\ell)$ is the number of available users in the system when ℓ is to be created; $k(\ell)$ is the in-degree of ℓ's creator at the same moment; and the denominator is the number of potential candidates to be followed.

The indicator function for each link ℓ denotes whether the link connects with a grandparent

or not in the real data:

$$\mathbf{1}_G(\ell) = \begin{cases} 1 & \text{if } \ell \text{ links to a grandparent in the data} \\ 0 & \text{otherwise.} \end{cases} \tag{9.2}$$

The expected number of links to grandparents according to the null hypothesis can be then computed as:

$$E_G = \sum_{\ell=1}^{L} p_G(\ell) \tag{9.3}$$

and its variance is given by:

$$\sigma_G^2 = \sum_{\ell=1}^{L} p_G(\ell)\,(1 - p_G(\ell)) \tag{9.4}$$

while the corresponding empirical number is:

$$S_G = \sum_{\ell=1}^{L} \mathbf{1}_G(\ell). \tag{9.5}$$

According to the Lyapunov central limit theorem,[1] the variable $z_G = (S_G - E_G)/\sigma_G$ is distributed according to a standard normal $\mathcal{N}(0,1)$. For linking to origins (O) or triadic nodes (Δ), we can define z_O and z_Δ similarly. In all three cases, using a z-test, we can reject the null hypotheses with high confidence ($p < 10^{-10}$). We conclude that links established by following grandparents, origins or triadic nodes happen much more frequently than by random connection. These link creation mechanisms have important roles in the evolution of the social network.

9.1.2 User Preference

The variables z_G, z_O, and z_Δ, as defined above, measure how much more likely a type of links are formed than by chance—in other words, how strong individual preferences are for following grandparents, origins or triadic nodes. To study the dependence of the link formation tendencies on the different stages of an individual's lifetime, let us compute z_G^k, z_O^k and z_Δ^k for links created by users with in-degree k, that is, those who are following k users at the time when the link is created. Figure 9.3 shows that the principle of triadic closure dominates user behavior when one follows a small number of users ($k < 75$). In the early stages, one does not receive much traffic, so it is natural to follow people based on local social circles, consistently with triadic closure. However, users who have been active for a long time and have followed many people ($k > 75$) have more channels through which they monitor traffic. This creates an opportunity to follow others from whom they have seen messages in the past.

[1] Lyapunov's condition, $\frac{1}{\sigma_n^4}\sum_{\ell=1}^{n} E[(X(\ell) - p(\ell))^4] \overset{n\to\infty}{\longrightarrow} 0$ where $X(\ell)$ is a random Bernoulli variable with success probability $p(\ell)$ [Billingsley, 1995], is consistent with numerical tests. Details are omitted for brevity.

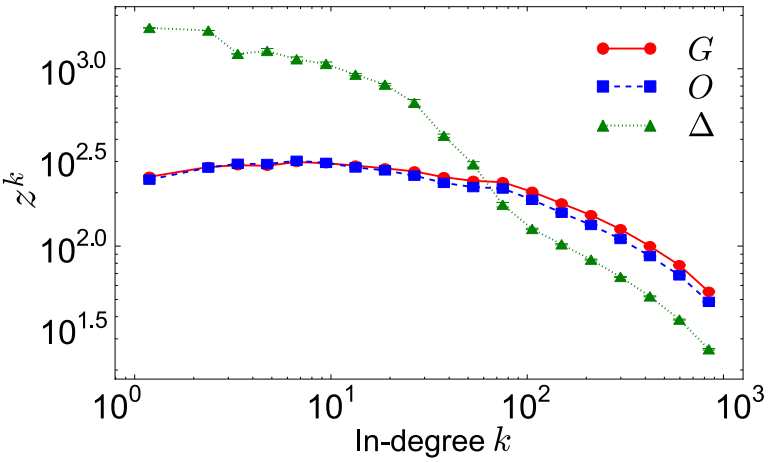


Figure 9.3: Individual preferences for following grandparents (red circles), origins (blue squares) and triadic nodes (green triangles) change with the in-degree of the link creator.

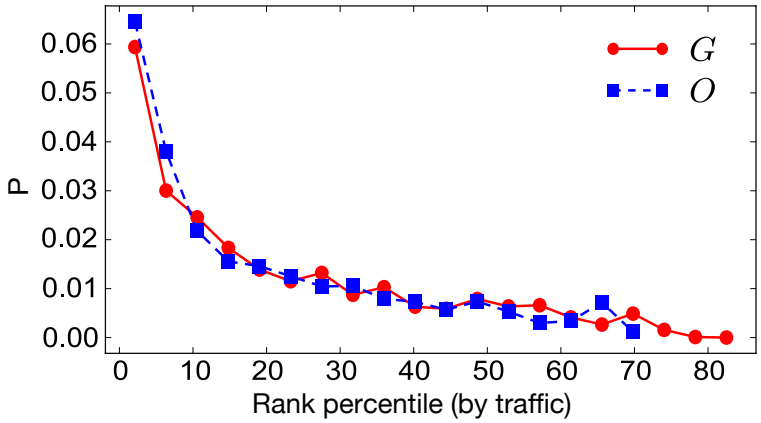
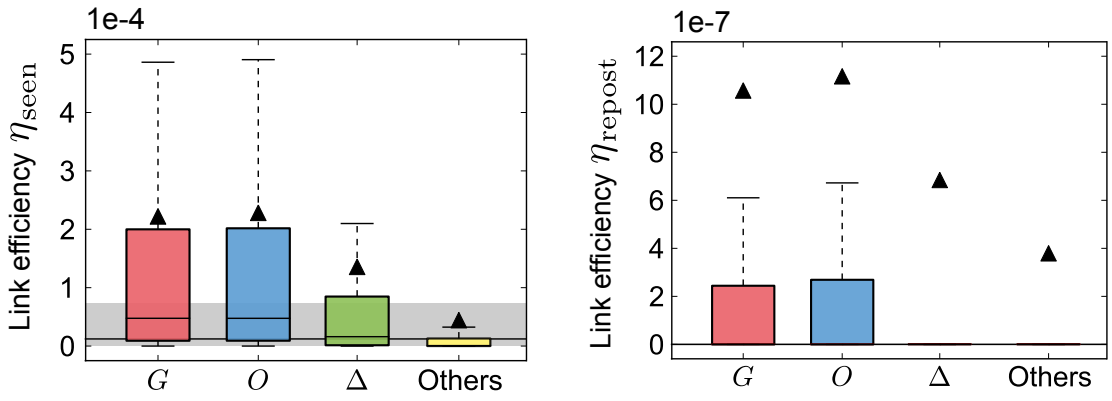


Figure 9.4: Probability density of followed grandparents (red circles) or origins (blue squares) having a certain rank percentile. Link targets are ranked so that the link creator has seen more messages from a user with smaller rank percentile.

9.1.3 Traffic Bias

Further inspection of the empirical data reveals that not all shortcuts are equally likely; users tend to follow those who have often been sources of seen messages. To investigate this, consider all new shortcuts to grandparents or origins. For each shortcut, we rank all the available grandparent or origin candidates according to how many of their messages have been seen by the creator prior to the link formation. We plot the probability of a followed grandparent or origin having a certain rank percentile in Fig. 9.4. The plot clearly demonstrates that repeated exposure to contents posted by a user increases the probability of following that user. This is analogous to the way in which we are more likely to adopt a piece of information or behavior to which we are exposed multiple times [Bakshy et al., 2009, Centola, 2010, Romero et al., 2011b, Hodas and Lerman, 2012]. This observation



(a) Link efficiency by seen traffic

(b) Link efficiency by repost traffic

Figure 9.5: Efficiency of links created according to different mechanisms, or average number of messages (a) seen or (b) reposted per time unit. Each box shows data within lower and upper quartile. Whiskers represent the 99-th percentile. The triangle and line in a box represent the median and mean, respectively. Note that the mean can fall outside the shown quantiles for skewed distributions. The grey area and the black line across the entire figure mark the interquartile range and the median of the measure across all links, respectively.

shows that topology alone is insufficient to explain the evolution of the network; activity patterns—the dynamics *on* the network—are a necessary ingredient in describing the formation of new links.

9.1.4 Link Efficiency

In information diffusion networks like Twitter and Yahoo! Meme, social links may have a key efficiency function of shortening the distance between information creators and consumers. An efficient link should be able to convey more information to the follower than others. Hence we define the *efficiency* of link ℓ as the average number of posts seen or reposted through ℓ during one time unit after its creation:

$$\eta_{\text{seen}} = \frac{w_{\text{seen}}(\ell)}{T - t(\ell)}, \quad \eta_{\text{repost}} = \frac{w_{\text{repost}}(\ell)}{T - t(\ell)} \tag{9.6}$$

where $w(\ell)$ is the number of messages seen or reposted through ℓ; $t(\ell)$ is the time when ℓ was created; and T is the time of the last action recorded in our dataset. Both seen and reposted messages are considered, as they represent different types of traffic; the former are what is visible to a user, and the latter are what a user is willing to share. We compute the link efficiency of every grandparent, origin, and triadic closure link. As shown in Fig. 9.5, both grandparent and origin links exhibit higher efficiency than triadic closure links, irrespective of the type of traffic. By shortening the paths of information flows, more posts from the content generators reach the consumers.

9.2 Rules of Network Evolution

To infer the different link creation strategies from the observed data, we characterize users with a set of probabilities associated with different actions, and approximate these parameters by Maximum-Likelihood Estimation (MLE) [Cowan, 1998]. For each link ℓ, we know the actual creator and the target; we can thus compute the likelihood $f(\ell|\Gamma, \Theta)$ of the target being followed by the creator according to a particular strategy Γ, given the network configuration Θ at the time when ℓ is created. The likelihoods associated with different strategies can be mixed according to the parameters to obtain a model of link creation behavior. Finally, assuming that link creation events are independent, we can derive the likelihood of obtaining the empirical network from the model by the product of likelihoods associated with every link. The higher the value of the likelihood function, the more *accurate* the model.

9.2.1 Single Strategies

Let us consider five link creation mechanisms and their combinations:

Random (Rand): follow a randomly selected user who is not yet followed.

Triadic closure (Δ): follow a randomly selected triadic node.

Grandparent (G): follow a randomly selected grandparent.

Origin (O): follow a randomly selected origin.

Traffic shortcut ($G \cup O$): follow a randomly selected grandparent or origin.

Other mechanisms for link creation could be similarly incorporated, such as social balance [Easley and Kleinberg, 2010] and preferential attachment [Barabási and Albert, 1999]. However, preferential attachment is built on the assumption that everyone knows the global connectivity of everyone else, which is not realistic. The strategies considered here essentially reproduce and extend the copy model [Kumar et al., 2000], approximating preferential attachment with only local knowledge.

To model link creation with a single strategy, we can use a parameter p for the probability of using that strategy, while a random user is followed with probability $1 - p$. The calculation of maximum likelihood, taking the single strategy of grandparents as an example, is as follows:

$$\begin{aligned}
\mathcal{L}_G(p) &= \prod_{\ell=1}^{L} \left(pf(\ell|G,\Theta) + (1-p)f(\ell|\text{Rand},\Theta)\right) \\
&= \prod_{\ell=1}^{L} \left(p\frac{\mathbf{1}_G(\ell)}{N_G(\ell)} + (1-p)\frac{1}{N(\ell)-k(\ell)-1}\right) \\
&= \prod_{\mathbf{1}_G(\ell)=1} \left(\frac{p}{N_G(\ell)} + \frac{1-p}{N(\ell)-k(\ell)-1}\right) \prod_{\mathbf{1}_G(\ell)=0} \frac{1-p}{N(\ell)-k(\ell)-1}.
\end{aligned} \tag{9.7}$$

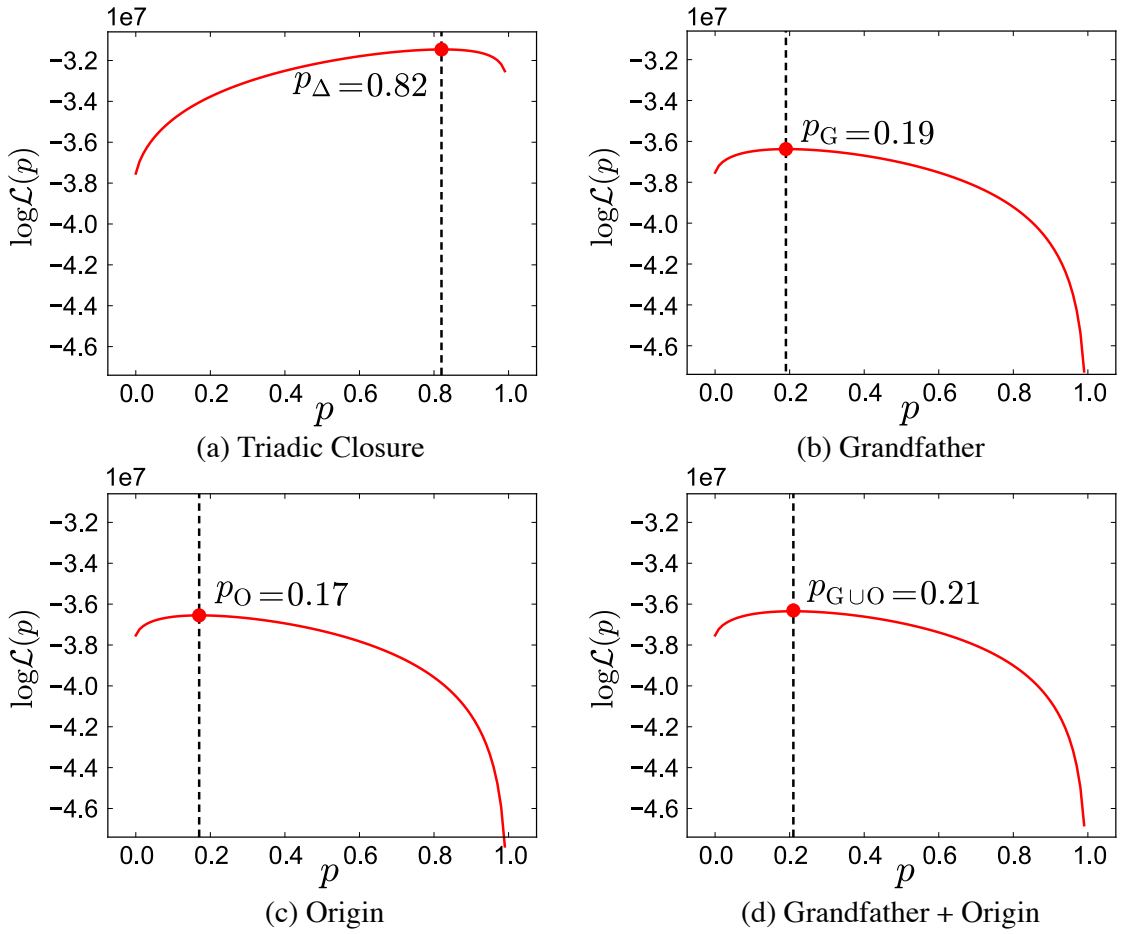


Figure 9.6: The plot of the log-likelihood $\log \mathcal{L}(p)$ as a function of link creation strategy probabilities for models with a single strategy. The red circles mark the maximized $\log \mathcal{L}(p)$.

Note that since a follow action can be ascribed to multiple strategies, it can contribute to multiple terms in the log-likelihood expression. For instance, a link could be counted in both $f(\ell|G, \Theta)$ and $f(\ell|\text{Rand}, \Theta)$. For numerically stable computation, we maximize the log-likelihood:

$$\begin{aligned} &\log \mathcal{L}_G(p) \\ &= \sum_{\mathbf{1}_G(\ell)=1} \ln \left(\frac{p}{N_G(\ell)} + \frac{1-p}{N(\ell) - k(\ell) - 1} \right) + \sum_{\mathbf{1}_G(\ell)=0} \ln \frac{1-p}{N(\ell) - k(\ell) - 1}. \end{aligned} \tag{9.8}$$

Similar expressions of log-likelihood can be obtained for other strategies (Δ, O, and $G \cup O$).

It is not trivial to obtain the best p analytically, so we explore the values of $p \in (0, 1)$ numerically (see Fig. 9.6). Triadic closure dominates as a single strategy, with $p_\Delta = 0.82$, consistently with the large number of triadic closure links observed in the data. Traffic-based strategies alone account for about 20% of the links.

9.2.2 Combined Strategies

For a more realistic model of the empirical data, let us consider combined strategies with both triadic closure and traffic-based shortcuts. For each link ℓ, the follower with probability p_1 creates a shortcut by linking to a grandparent (G), an origin (O), or either of them ($G \cup O$); with probability p_2 follows a triadic node (Δ); and with probability $1 - p_1 - p_2$ connects to a random node.

Taking the combined strategy with grandparent as an example, we compute the log-likelihood as:

$$\begin{aligned}
&\log \mathcal{L}_{G+\Delta}(p_1, p_2) \\
&= \log \prod_{\ell=1}^{L} [p_1 f(\ell|G, \Theta) + p_2 f(\ell|\Delta, \Theta) + (1 - p_1 - p_2) f(\ell|\text{Rand}, \Theta)] \\
&= \sum_{\substack{\mathbf{1}_G(\ell)=1 \\ \mathbf{1}_\Delta(\ell)=1}} \log \left(\frac{p_1}{N_G(\ell)} + \frac{p_2}{N_\Delta(\ell)} + \frac{1 - p_1 - p_2}{N(\ell) - k(\ell) - 1} \right) \\
&\quad + \sum_{\substack{\mathbf{1}_G(\ell)=1 \\ \mathbf{1}_\Delta(\ell)=0}} \log \left(\frac{p_1}{N_G(\ell)} + \frac{1 - p_1 - p_2}{N(\ell) - k(\ell) - 1} \right) \\
&\quad + \sum_{\substack{\mathbf{1}_G(\ell)=0 \\ \mathbf{1}_\Delta(\ell)=1}} \log \left(\frac{p_2}{N_\Delta(\ell)} + \frac{1 - p_1 - p_2}{N(\ell) - k(\ell) - 1} \right) + \sum_{\substack{\mathbf{1}_G(\ell)=0 \\ \mathbf{1}_\Delta(\ell)=0}} \log \frac{1 - p_1 - p_2}{N(\ell) - k(\ell) - 1}.
\end{aligned} \tag{9.9}$$

Once again, many following actions can create both triadic closure links and traffic shortcuts, so they can contribute to multiple terms in the log-likelihood expression.

It is hard to obtain the optimal solution analytically. We numerically explore the values of p_1 and p_2 in the unit square to maximize the log-likelihood. The best combined strategy is the one considering both grandparents and origins as well as triadic closure (Fig. 9.7) The parameter settings and the maximum likelihood values for all tested models are listed in Table 9.1. We can compare the quality of these models by comparing their maximized $\log \mathcal{L}$'s. The combined models with both traffic shortcuts and triadic closure yield the best accuracy. In these models, triadic closure accounts for 71% of the links, grandparents and origins for 12%, and the rest are created at random.

Thus far we have assumed that each user has the same behavior; in the next section we model each user separately.

9.3 User Behavior

The MLE models for describing the system behavior can be similarly employed to characterize the strategy of an individual user. Let us focus on the model $G \cup O + \Delta$ that

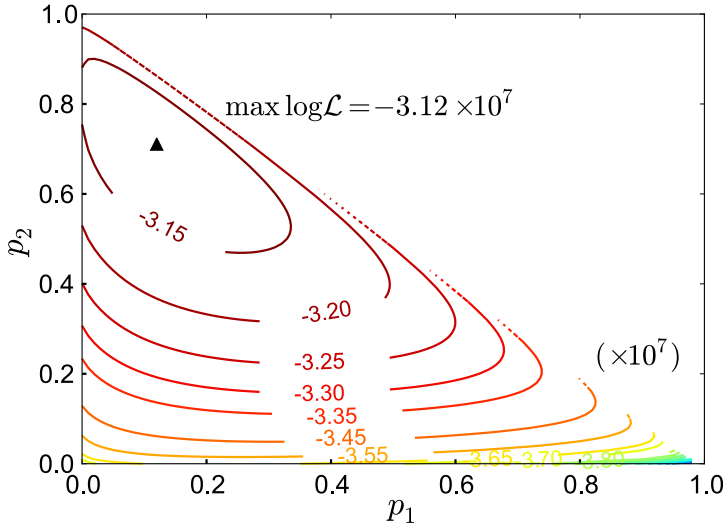


Figure 9.7: The contour plot of log-likelihood $\log \mathcal{L}(p_1, p_2)$ for the combined strategy of creating traffic shortcuts ($G \cup O$) with probability p_1 and triadic closure links (Δ) with probability p_2. The black triangle marks the optimum.

Table 9.1: The best parameters in different models and corresponding values of maximized log-likelihood function.

Strategy	Model	Parameters		$\max \log \mathcal{L}$
Baseline	Rand	–		-3.75×10^7
Single	Δ	$p = 0.82$		-3.15×10^7
	G	$p = 0.19$		-3.64×10^7
	O	$p = 0.17$		-3.65×10^7
	$G \cup$ O	$p = 0.21$		-3.63×10^7
Combined	$G + \Delta$	$p_1 = 0.12$	$p_2 = 0.71$	-3.12×10^7
	O $+ \Delta$	$p_1 = 0.10$	$p_2 = 0.73$	-3.13×10^7
	$G \cup$ O $+ \Delta$	$p_1 = 0.12$	$p_2 = 0.71$	-3.12×10^7

best reproduces the empirical data at the global level. We run MLE to explain the links created by each user independently. We consider users with at least 20 in-links, such that MLE is meaningful. For easier interpretation, let us call $p_{\text{traffic}} = p_1$, $p_{\text{structure}} = p_2$ and $p_{\text{random}} = 1 - p_1 - p_2$. Each user has her own set of parameters.

9.3.1 User Strategy Classification

Using the Expectation-Maximization (EM) algorithm [Celeux and Govaert, 1992, Dempster et al., 1977], users are clustered into several classes based on p_{traffic}, $p_{\text{structure}}$ and p_{random}. EM iteratively preforms an expectation step to compute the probability that each instance belongs to each class, and a maximization step in which latent variables of classes are altered to maximize the expected likelihood of the observed data. EM decides how

Table 9.2: Classes of user link creation strategy

Class	#Users	$\langle p_{\text{traffic}} \rangle$	$\langle p_{\text{structure}} \rangle$	$\langle p_{\text{random}} \rangle$
All Users	45,708	0.07	0.77	0.17
Info	4,750	0.52	0.36	0.13
Friend	12,797	0.00	0.96	0.04
CFrd	23,469	0.01	0.80	0.19
Mix	2,524	0.07	0.63	0.30
Rand	2,168	0.09	0.32	0.59

many clusters to create by cross validation. This procedure yields five classes:

Information-Oriented (Info): People prefer to follow someone from whom or through whom they have received messages.

Friend of a Friend (Friend): People follow users two steps away to form triadic closure, almost exclusively.

Casual Friendship (CFrd): People tend to follow a set of users their friends are following; they also link to random users occasionally.

Mixture (Mix): Miscellaneous behavior of creating traffic shortcuts, connecting others by triadic closure, and following random people.

Random Browsing (Rand): People have a much higher preference for following a random user who is not close in either the follower or the message flow network. "Random" does not necessarily imply the absence of any rule; there can be other strategies not explored in our model, i.e., following a celebrity on purpose (similar to preferential attachment).

Table 9.2 displays the parameter averages for users in each class, representing the overall behavior pattern in that class. Users in the mixture category behave similarly to the average across all users. Figure 9.8 illustrates how users in different classes are mapped into the parameter space with the probability of each link creation strategy as one dimension.

9.3.2 Characterization of User Classes

To further differentiate users with different link creation strategies, let us look at several structural and behavioral characteristics of each class. Figures 9.9(a-c) show how users in different classes create social links by comparing p_{traffic}, $p_{\text{structure}}$ and p_{random}.

As shown in Fig. 9.9d, information-oriented users have been active longer than users in friendship classes. Similarly, information-oriented users tend to follow more people (see Fig. 9.9e). Information-oriented users have even more followers compared to friendship-driven users (see Fig. 9.9f). This suggests that they tend to be more influential, as confirmed by considering the number of times that their messages are reposted (see Fig. 9.9g). Friendship-driven users follow a few people while essentially nobody is following them.

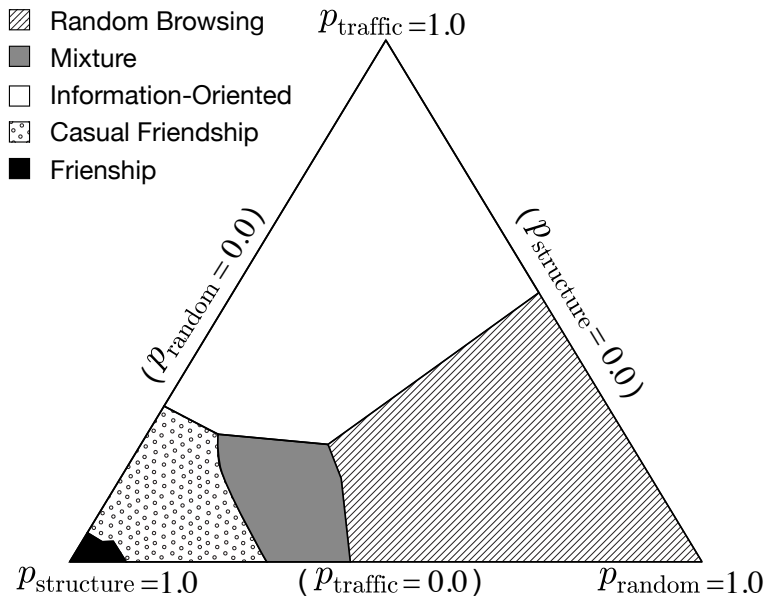


Figure 9.8: Ternary plot of users according to p_{traffic}, $p_{\text{structure}}$ and p_{traffic}.

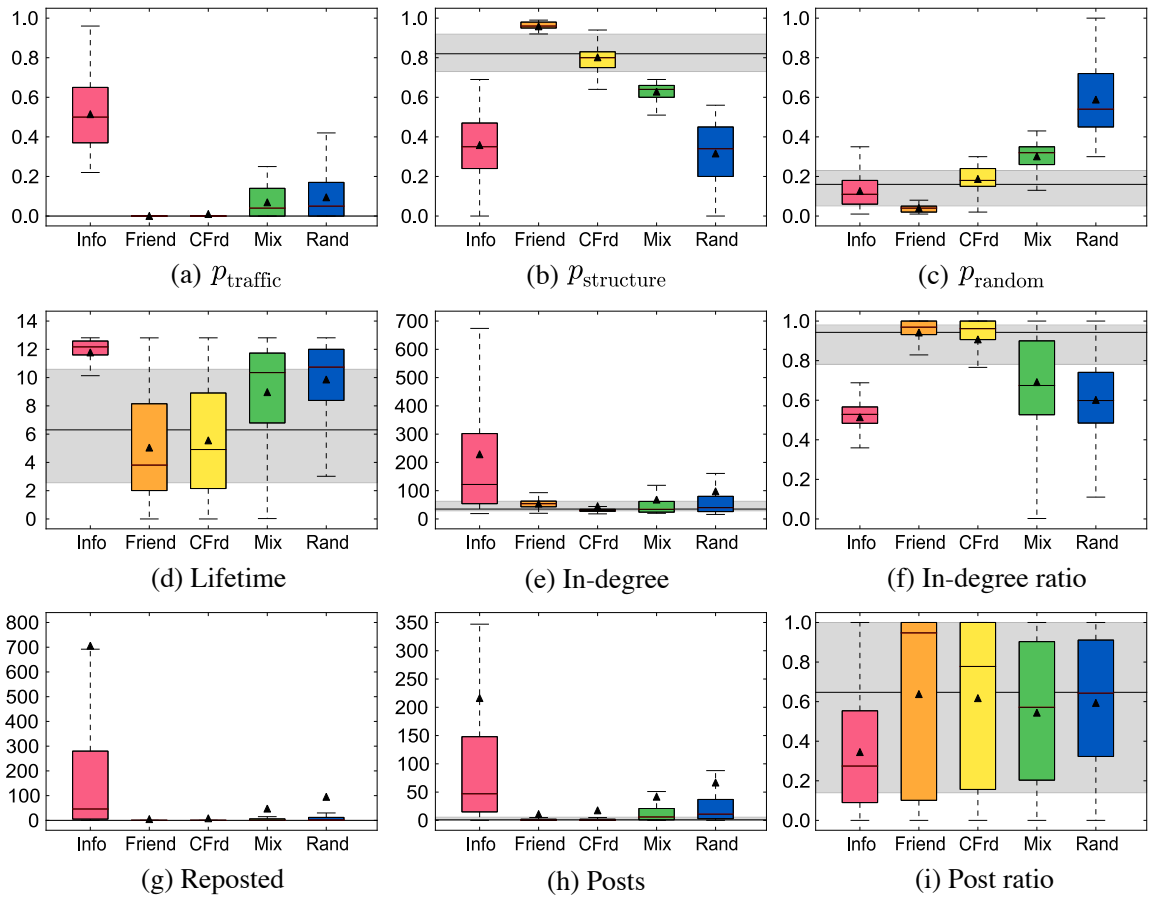
Such a passive role can be explained by their short lifetime. All of these results are consistent with Fig. 9.3. Finally, Figs. 9.9(h-i) suggest that, while information-oriented users tend to produce more messages, their role is more that of spreaders than producers of information compared to other classes.

In this analysis, the parameters are fit according to the entire lifetime of each user. Focusing on the first 20 or 50 links does not yield qualitatively different results.

9.4 Discussion

The study of the feedback loop between the dynamics *of* and *on* the network—how the network grows and how the information spreads—offers a promising framework for understanding social influence, user behavior, and network efficiency in the context of microblogging systems.

The results presented in this chapter show that while triadic closure is the dominant mechanism for social network evolution, it is mainly relevant in the early stages of a user's lifetime. As time progresses, the traffic generated by the dynamics of information flow on the network becomes an indispensable component for user linking behavior. As users become more active and influential, their links create shortcuts that make the spread of information more efficient in the network. Users whose following behavior is driven by the information they see are a minority of the population, but play a key role in the information diffusion process. They produce more information, but, even more importantly, they act as spreaders of the information they collect widely across the network.



(a) p_{traffic} (b) $p_{\text{structure}}$ (c) p_{random}

(d) Lifetime (e) In-degree (f) In-degree ratio

(g) Reposted (h) Posts (i) Post ratio

Figure 9.9: Various features of users in different classes. The lifetime of a user is measured by how many others join the system after him. The in-degree is the number of people a user is following, k, and the in-degree ratio is $k/(k + k_{out})$ where k_{out} is the number of followers. "Reposted" refers to the number of times that a user's messages are reposted by others. "Posts" denotes the number of messages generated by a user excluding reposts. The post ratio is the fraction of all posts by a user (including reposts) that are originated by that user. Each box shows data within lower and upper quartile. Whiskers represent the 99th percentile. The triangle and line in a box represent the median and mean, respectively. The grey area and the black line across the entire figure mark the interquartile range and the median of the measure across all links, respectively.

While existing link prediction algorithms [Liben-Nowell and Kleinberg, 2007, Clauset et al., 2008, Backstrom and Leskovec, 2011, Lou et al., 2010, Schifanella et al., 2010] are not designed to explain the network evolution in a dynamic setting, the MLE framework could in principle be used to assess which link prediction methods are more consistent with the longitudinal structural changes observed in the network, by treating the prediction at each step as a link creation strategy. These approaches will be explained more in details in the future work section (see Sec. 10.2).

We believe our findings apply generally to techno-social networks, and in particular information diffusion networks and (micro)blogs. Analyses of other micro-blogging systems, such as Twitter and Weibo, would be needed to confirm this, but will be challenging due to the difficulty of obtaining full longitudinal data about user actions on the social network.

Chapter 10

Conclusion

Information is a valuable treasure. A fast and efficient information spread benefits knowledge dissemination, optimization of decision making processes, innovation activity, and many other important aspects of human society. In the era of big data, we are endowed with the capability to track, observe, analyze, model, and predict the spread of information using a large amount of digital data. With the techniques from network science, computer science, and social science, we can obtain a more comprehensive understanding of how information is propagated among people in the online community from an interdisciplinary viewpoint.

This thesis presents several studies on the topic of *information diffusion on online social networks*. The information diffusion process can be naturally decomposed into four components: actors who share the information, transmissible content that travels among people, the structure of online social networks, and the diffusion mechanism. Accordingly the thesis is structured to contain three parts on (I) actors, (II) content, and (III) the mutual effects between network structure and diffusion. Each part has two chapters and each chapter follows the logical flow from observation to model and then from model to application. The first part delves into the limited attention of people with empirical analyses and proposes an agent-based model simulating the effect of finite attention on the dynamics of meme diffusion. The second part studies topical diversity measure as well as its role in user and content popularity, and the selectivity of different conversation topics when people communicate with strong-tie and weak-tie friends. The last part presents how the network structure, particularly community structure, influences the propagation of Internet memes and how the information flow, in turn, affects social link formation.

We believe that, other than information diffusion, many complex dynamics of human society, from ethnic tension to global conflicts, and from grassroots social movements to political campaigns, could be better understood by continued investigation of network structure. A better and more comprehensive knowledge set of information diffusion among online communities would help us build a better information-based socio-technical system and bring in advantages to online viral marketing, advertisement, and social media analytics.

10.1 Summary of Contributions

The following sections summarize the contributions of each part in the thesis.

10.1.1 Actor Attention

- We observe that the breadth of individual attention is bounded irrespective of system diversity using the entropy measure. (Sec. 4.1)
- Whether new content can attract a user is shown to be determined by her interests, represented as a memory of things she has shared before. (Sec. 4.2)
- We propose an agent-based model on meme diffusion, in which limited user memory and social network structure are two key components. The model is able to reproduce the heterogeneity of meme popularity, meme life span, user activity, and the breath of user attention, suggesting that limited attention and social network structure are sufficient to interpret a complex information landscape of Internet memes without assuming different intrinsic values of memes. (Sec. 4.3)
- We propose a measure of attention as the fraction of an individual's total amount of attention directed to a given link to quantify the importance of a social tie in information gathering. (Sec. 5.1)
- We show that strong ties carry a large amount of traffic, as presented in the weak tie hypothesis [Granovetter, 1973]. Meanwhile, weak ties succeed to attract quite amount of attention similar to or even more than strong ties, suggesting the important role of weak ties in information propagation as a channel for innovation. (Sec. 5.2)
- The distribution of attention—heavily resting on very strong or weak ties—can be explained by distinguishing social links from information ones. The former are built for maintaining social relationships and therefore prefer strong ties, while the latter seek for information and favor weak ties. The extent to which weak ties acquire attention is determined by the underlying mechanism of the network. (Sec. 5.3)

10.1.2 Content Topics

- We develop a way to extract topics from online conversation in Twitter. Communities in the hashtag co-occurrence network are found to well represent topics as clusters of semantically similar or relevant tags. (Sec. 6.1.1)
- Given a user, we gauge the diversity of his topical interests by examining to which topic clusters his tags belong, so that we can distinguish users with diverse interests from those with focused attention. Given a hashtag, its topical diversity is measured similarly with co-occurring tags. (Sec. 6.1.2 & 6.1.3)

- When a hashtag is adopted by people with diverse interests or co-occurs with other tags on assorted themes, this hashtag is more likely to be popular. We evaluate the average topical diversity of early adopters or co-occurring tags as predictors of hashtag future popularity. (Sec. 6.2)
- However, high topical diversity does not help the growth of individual social impact, as staying focused on one or a few topics could be a sign of expertise. Inactive users attract followers by mentioning a variety of topics, while active users tend to obtain many followers by maintaining focused interests. Focused topical preference promotes content interestingness irrespective of ordinary users or celebrities. (Sec. 7.3)
- Strong ties are found to respond to a broader variety of topics than weak ties in the context of an ego network in the online social network site, even after controlling for the overall number of responses. (Sec. 7.3)
- Strong and weak ties have similar preference to popular topics, implying that the impact of content popularity operates independently from tie strength. (Sec. 7.4)

10.1.3 Network Topology and Diffusion Mechanism

- The spread of information within highly clustered communities is enhanced, while diffusion across communities is hampered, because complex contagions, unlike infectious diseases (simple contagions), are affected by homophily and social reinforcement. (Sec. 8.2)
- We observe significantly more communication within than across communities and people interact more with members of the same community than with others in different groups. (Sec. 8.2.1)
- The trapping effect of communities is triggered by a combination of three key factors in social contagion: structural trapping, social reinforcement, and homophily. We propose and compare four diffusion models, each incorporating none or a subset of these factors. The simulation results suggest that viral memes spread like simple contagions, permeating through many communities. (Sec. 8.2.2 & 8.2.3)
- We propose and analyze a comprehensive set of features based on the early spreading patterns of memes, categorized into three groups: positions of early adopters in the network, community concentration, and characteristics of adoption time series. (Sec. 8.3.1 & 8.3.2)
- We then develop an accurate model to predict future popularity of memes. Our model outperforms other approaches, particularly in the tasks of detecting very popular or unpopular memes. Besides, we find that community-based features are the most powerful predictors of future success and early popularity of a meme is not a good predictor of its future popularity, contrary to the common beliefs. (Sec. 8.3.3)
- A considerable portion of new links in social media site are shortcuts based on information flow—people tend to link to others from whom they have seen messages.

These links happen much more frequently than by random connection with significant statistical evidence. (Sec. 9.1.1)

- Furthermore, not all individuals apply the same strategy to grow their social connections; users who follow many others tend to pay more attention to traffic. (Sec. 9.1.2)
- However, shortcuts are not equally probable. We find that people tend to follow the most active sources of content. (Sec. 9.1.3)
- Traffic-based social links make the social network more efficient in terms of shortening the path of information diffusion. (Sec. 9.1.4)
- We perform a Maximum Likelihood Estimation analysis to quantify the system-wide prevalence of different link creation strategies including triadic closure, traffic-based shorts and random choices. (Sec. 9.2)
- The categorization of individual link formation behavior suggests the existence of several distinct link formation strategies: information-oriented, friend of a friend, casual friendship, mixture, and random browsing. (Sec. 9.3)

10.2 Future Work

Several potential projects are listed in this section as extensions of presented work in the thesis. All the proposed future work is in the domain of information diffusion on social network. The following projects are ordered by the operational difficulties, from the easiest to the hardest.

10.2.1 Evaluation Tool of Missing Link Prediction Algorithms with Longitudinal Data

Existing link prediction algorithms [Liben-Nowell and Kleinberg, 2007, Clauset et al., 2008, Backstrom and Leskovec, 2011, Lou et al., 2010, Schifanella et al., 2010] are not designed to explain the network evolution in a dynamic setting. Most evaluations are performed in such a way that a proportion of links are removed and an accuracy is reported by applying the algorithm to recover the removed relationships. Our Maximum Likelihood Estimation framework proposed in Sec. 9.2 could in principle be used to assess which link prediction methods are more consistent with the longitudinal structural changes observed in the real-world network by treating the prediction at each step as a link creation strategy.

Let us demonstrate how to evaluate a given link prediction algorithm A using our framework. We have a dataset describing the longitudinal topological changes of a social network, such as the Yahoo! Meme dataset in Ch. 9. We label each link ℓ by its creation order, $1 \leq \ell \leq L$, where L is the total number of links in the data. For each link ℓ in time, we consider the network with links built before ℓ as the network configuration Θ_ℓ and compute the probability of creating this link using the algorithm A as $f(\ell|\Theta_\ell, A)$.

The computation of likelihood entails an non-zero probability at each step, but it might happen that the algorithm produces a zero probability for ℓ. Here are two possible solutions.

Adding small constant The likelihood of a single link ℓ is computed in such a way:

$$f(\ell|\Theta_\ell, A) = \frac{A(\ell|\Theta_\ell) + \epsilon}{\sum_{\ell' \in M(\Theta_\ell)} A(\ell'|\Theta_\ell) + \epsilon} \tag{10.1}$$

where $A(\ell|\Theta_\ell)$ is the likelihood associated with the missing link ℓ given by the algorithm A given the early network setting Θ_ℓ; $M(\Theta_\ell)$ contains all the potential future links in Θ_ℓ; and ϵ is a very small constant. In this way, $f(\ell|\Theta_\ell, A)$ can stay non-zero and the log-likelihood for evaluating A is:

$$\log \mathcal{L}_A = \sum_{\ell=1}^{L} \log(A(\ell|\Theta_\ell) + \epsilon) - \sum_{\ell=1}^{L} \log(\sum_{\ell' \in M(\Theta_\ell)} A(\ell'|\Theta_\ell) + \epsilon) \tag{10.2}$$

In the comparison of prediction performances of multiple algorithms, we would compare the values of log-likelihood associated with each algorithm.

Including random choices The other solution is to add random choices as the secondary strategy at each step. Given a missing link ℓ, we may predict it according to A with probability p, or make a random guess with probability $1 - p$, then:

$$f(\ell|\Theta_\ell, A) = \frac{A(\ell|\Theta_\ell)}{\sum_{\ell' \in M(\Theta_\ell)} A(\ell'|\Theta_\ell)} \tag{10.3}$$

$$\mathcal{L}_A(p) = \prod_{\ell=1}^{L} \left(p f(\ell|\Theta_\ell, A) + \frac{1-p}{n_\ell(n_\ell - 1)/2 - \ell + 1} \right) \tag{10.4}$$

$$\log \mathcal{L}_A(p) = \sum_{\ell=1}^{L} \log \left(\frac{pA(\ell|\Theta_\ell)}{\sum_{\ell' \in M(\Theta_\ell)} A(\ell'|\Theta_\ell)} + \frac{2(1-p)}{n_\ell^2 - n_\ell - 2\ell + 2} \right) \tag{10.5}$$

where n_ℓ is the number of nodes in the existing network Θ_ℓ. Thise approach demands a parameter fitting process to find the optimal $\hat{p}$. The final outcomes for the algorithm A is the maximized likelihood $\log \mathcal{L}_A(\hat{p})$ and the probability $\hat{p}$ for random choices. A large likelihood and a small random probability are favorable signs for a strong algorithm.

We should apply multiple existing missing link algorithms on the real dataset using both approaches and see which produce more reasonable and persuasive results, in order to decide which approach is better at eliminating zero probabilities.

10.2.2 Shrinkage of Human Attention Span

The overwhelming amount of online information starts to drive people impatient more than ever before. As one evidence, content on the Internet is evolving to be shorter and shorter.

During the late 1990s and early 2000s, Web blogs had the most popular online communities in which people wrote their commentaries, reviews, dairies, and articles. Meanwhile they read others' blogs and left comments as a way to communicate. Many articles were pretty long (comparatively), but bloggers at that time were 'patient' enough to write them down or read them through. Later, Facebook was born in 2004, a platform for people sending out quick updates, photos, and liking each others' posts. Twitter became popular in about 2009 and it explicitly forces all the tweets to contain fewer than 140 characters. Short and informal messages are all around.

Compared with blog-based systems, Facebook and Twitter incorporate more components on social networking, emphasizing communication and social relationship maintenance. Due to the limited individual attention, it is possible that people cannot spend much time on content production as before, as they need to chat with friends and sustain relationships. Besides, inter-personal conversations tend to be shorter and more casual than formal articles. Meanwhile, the explosion of information is expected to be a strong factor in shortening human attention span. Given that so many news are available and updates are provided so fast, people have to skip many of them or quickly scan through most, suggesting that intensive reading of long articles may gradually become unusual.

Therefore these questions stand out: can we measure the change of online content length in time? Why do we observe a trend of online content becoming shorter than before? Is it caused by the shrinkage of human attention span or system designs with emphasis on social networking?

I would like to propose two studies on blogging and micro-blogging systems, respectively. First, in a blogging system such as Blogger or Wordpress, we can measure the average length of articles in time to depict the change of average attention span. The measurement can be done at the global level by only considering long-term users or at the individual level by normalizing the article length for each individual. One issue to be taken care of is that the length of an article could be relevant to its theme; for instance, a tutorial is more likely to be longer than a diary on daily routines. The average length of comments is another quantity for estimating the attention span and should be computed similarly. The second study is proposed for a micro-blogging system like Twitter and Weibo, where social networking is a key component. With the assumption that both content production and social relationship maintenance consume attention, we may define the attention dedicated to each action accordingly—a sum of the length of original posts and the frequency of reposts or replies. The former measures the amount of content production and the latter represents how ofter the user interacts with others. How to combine these two metrics of very different natures requires a careful investigation.

Note that we have proposed attention measures in Sec. 4.1 and 5.1. In Sec. 4.1, we focus on the upper limit of topics (hashtags) that one user can post about, in comparison with the system behavior. Sec. 5.1 considers each user equipped with a fixed amount of attention and focuses on how one individual assigns attention to neighbors but not on the change of the amount of attention in time. We interpret attention differently because of different goals and research questions. To understand the change of human attention span in time may provide helpful insights into and suggestions for the design of modern social-technical

systems and social good.

10.2.3 Online Social Network and Real-World Friendship

In the online scenario, we may befriend people whom we know offline, denoted as *real-world friends*, but it is often the case that not all of our real-world friends are available in the system. Meanwhile, we become friends with others by online communication, those labelled as *virtual friends*. Because of these missing links with real-world friends and new links with virtual friends, the structure of social networks on the Internet would look differently from the one in the real world. I'm interested in the exploration of both consistencies and inconsistencies between online and offline social networks.

Can we find a way to gauge the difference? Does the online social network well reflect the real-world friendships? From an individual viewpoint, what user characteristics are implied by a high difference between online and offline ego networks? What if it is low?

Let us consider the timing of the friending behavior. For a pair of connected individuals, either offline or online, we label t_{real} as the timestamp of their meeting and knowing each other in the real world, i.e., in any meeting, party, or appointment. Similarly, t_{virtual} represents the timestamp of their online connection, i.e., becoming friends in Facebook or following one another in Twitter. Either t_{real} or t_{virtual} can be infinite, if the corresponding connection is not observed (To eliminate the infinity, we may use closeness instead, defined as $c_{\text{real}} = 1/t_{\text{real}}$ and $c_{\text{virtual}} = 1/t_{\text{virtual}}$). The timestamp t_{virtual} can be extracted out of the data gathered from online social network sites. However, t_{real} is not easy to measure. Even we ask everyone such a question, "when did you become friends with this person?", it could be hard to recall accurately. One potential estimation is the timestamp of the first phone call or text message between the given pair of people, so as to take advantage of available data. Then social links can be classified into three categories, corresponding to $t_{\text{real}} > t_{\text{virtual}}$, $t_{\text{real}} = t_{\text{virtual}}$ and $t_{\text{real}} < t_{\text{virtual}}$. The distribution of friendship links in different categories helps us understand whether online social network is predictive of, equivalent to, or posterior to the real-world social network.

10.3 Further Challenges

The availability of big data has brought about both opportunities and challenges to research on human dynamics and social phenomena [Watts, 2007, Lazer et al., 2009, Manovich, 2011, Labrinidis and Jagadish, 2012].

Data sampling. The data used in most existing research is generated from a sampling process. The collection of data can hardly exhaust the whole history of the entire system, due to the policy, privacy problems, and operational issues. An inappropriate sampling method can lead to strong biases. We have only seen a scant amount of work on data

sampling algorithms [Lakhina et al., 2003, Leskovec and Faloutsos, 2006, González-Bailón et al., 2012, Morstatter et al., 2013] and the understanding is still limited and insufficient.

Universality. Another problem with the datasets is that most of them are about a single system or a snapshot of the system. We are in need of investigation into whether these results can be applied to other stages in time or other systems without a strong effect of "blind men feeling the parts of an elephant" [Lazer et al., 2009], thus inducing broader impact. Therefore, more future work is expected to study the longitudinal patterns on data with long history and to compare multiple platforms.

Privacy. So many aspects of human society and individual everyday lives have been recorded in big data. It can be dangerous when people are able to look across data from multiple sources to decipher the trace of an individual user. Knowledge about detailed personal schedules and private information may easily facilitate crimes; i.e. breaking into one's house knowing the whole family is on vacation, or illegal access to personal bank account knowing one's address, birth date, and social security number. The prevention of data-based crime deserves enough attention from researchers.

Open access. Data is crucial in quantitative research. However, when a small set of researchers work on private data and produce results, it is important for external people to replicate, verify, or question the findings due to the lack of access to the data. Although many datasets or systems cannot be open to everyone with consideration of privacy, an open environment for scientists and researchers in different institutes, countries, and fields can greatly enhance the vitality and health of academia, encouraging more innovative and meticulous research.

Gap between online and offline systems. User behavior on the Internet is not a transparent reflection of who they are in the real world, because online behavior is usually well curated and systematically managed [Ellison et al., 2006]. This rises up the question of the gap between online and offline environments, while applying classical sociological theorems to the study of online systems, or extending the implication of findings derived from online big data to offline social movements and events. The proposed future work in Sec. 10.2.3 is relevant to this challenge.

Bibliography

M. K. Agarwal, K. Ramamritham, and M. Bhide. Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments. *Proc. of VLDB Endowment*, 5(10):980–991, 2012.

C. C. Aggarwal and K. Subbian. Event detection in social streams. In *SIAM Intl. Conf. on Data Mining (SDM)*, pages 624–635, 2012.

Y.-Y. Ahn, J. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2): 9, 2012.

R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97, 2002.

J. An, M. Cha, P. K. Gummadi, and J. Crowcroft. Media landscape in twitter: A world of new conventions and political diversity. In *Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2011.

R. M. Anderson, R. M. May, and B. Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Oxford University Press, 1992.

S. Aral and M. Van Alstyne. The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1):90–171, 2011.

S. Aral and D. Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011.

S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Nat. Acad. Sci. (PNAS)*, 106 (51):21544–21549, 2009.

S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in social media: Persistence and decay. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2011.

R. Axelrod. *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton University Press, 1997.

L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 635–644. ACM, 2011.

L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 44–54, 2006.

L. Backstrom, E. Bakshy, J. Kleinberg, T. Lento, and I. Rosenn. Center of attention: How facebook users allocate attention across friends. In *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*, pages 1–8, 2011.

N. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 2nd edition, 1975.

E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *Proc. ACM Conf. on Electronic Commerce*, pages 325–334, 2009.

E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 65–74. ACM, 2011.

E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proc. ACM Intl. World Wide Web Conf. (WWW)*, pages 519–528, 2012.

R. Bandari, S. Asur, and B. Huberman. The pulse of news in social media: Forecasting popularity. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, pages 26–33, 2012.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

A.-L. Barabási and R. Albert. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 33–42, 2013.

A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

P. Bearman and P. Parigi. Cloning headless frogs and other important matters: Conversation topics and network structure. *Social Forces*, 83(2):535–557, 2004.

H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2011.

J. Berger and K. L. Milkman. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205, 2009.

P. Billingsley. *Probability and measure*, page 362. John Wiley & Sons, 1995.

C. M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10):P10008, 2008.

E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Nat. Acad. Sci. (PNAS)*, 99(Suppl 3):7280–7287, 2002.

R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

J. Borge-Holthoefer, A. Rivero, I. García, E. Cauhé, A. Ferrer, D. Ferrer, D. Francos, D. Iñiguez, M. P. Pérez, G. Ruiz, F. Sanz, F. Serrano, C. Viñas, A. Tarancón, and Y. Moreno. Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. *PLOS ONE*, 6(8):e23883, 2011.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

J. Brown and P. Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3):350–362, 1987.

M. J. Brzozowski, T. Hogg, and G. Szabo. Friends and foes: ideological social networking. In *Proc. ACM Intl. Conf. on Human factors in computing systems (CHI)*, pages 817–820, 2008.

M. Burke and R. Kraut. Using facebook after losing a job: Differential benefits of strong and weak ties. In *Proc. ACM Conf. on Computer supported cooperative work (CSCW)*, pages 1419–1430, 2013.

R. S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, 2009.

C. Butts. Revisting the foundations of network analysis. *Science*, 325:414–416, 2009.

C. T. Butts. Relational event framework for social action. *Sociological Methodology*, 38 (1):155–200, 2008.

C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009.

M. Cataldi, L. D. Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proc. Intl. Workshop on Multimedia Data Mining (MDMKDD)*, pages 4:1–4:10, 2010.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

D. Centola. The spread of behavior in an online social network experiment. *Science*, 329 (5996):1194–1197, 2010.

D. Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, pages 10–17, 2010.

F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):4, 2008.

J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proc. ACM Intl. Conf. on Human factors in computing systems (CHI)*, pages 1185–1194, 2010.

J. Cheng, L. Adamic, A. Dow, J. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proc. Intl. World Wide Web Conf. (WWW)*, 2014.

X.-Q. Cheng, F.-X. Ren, H.-W. Shen, Z.-K. Zhang, and T. Zhou. Bridgeness: a local index on edge significance in maintaining global connectivity. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(10):P10011, 2010.

N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370 – 379, 2007.

A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(1):98–101, 2008.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

R. Colbaugh and K. Glass. Early warning analysis for social diffusion events. *Security Informatics*, 1(1):1–26, 2012.

N. L. Collins and L. C. Miller. Self-disclosure and liking: a meta-analytic review. *Psychological bulletin*, 116(3):457, 1994.

M. D. Conover, C. Davis, E. Ferrara, K. McKelvey, F. Menczer, and A. Flammini. The geospatial characteristics of a social movement communication network. *PLOS ONE*, 8 (3):e55957, 2013a.

M. D. Conover, E. Ferrara, F. Menczer, and A. Flammini. The digital evolution of occupy wall street. *PLOS ONE*, 8(3):e55957, 2013b.

D. Cosley, D. Huttenlocher, J. Kleinberg, X. Lan, and S. Suri. Sequential influence models in social networks. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2010.

G. Cowan. *Statistical Data Analysis*. Oxford Science Publications, 1998.

D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 160–168, 2008.

R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Nat. Acad. Sci. (PNAS)*, 105(41):15649–15653, 2008.

O. Şimşek and D. Jensen. Navigating networks by using homophily and degree. *Proc. Nat. Acad. Sci. (PNAS)*, 105(35):12758–12762, 2008.

D. J. Daley and D. G. Kendall. Epidemics and rumours. *Nature*, 204(4963):1118–1119, 1964.

T. H. Davenport and J. C. Beck. *The Attention Economy: Understanding the New Currency of Business*. Harvard Business School Press, 2001.

B. D. Davison. Topical locality in the web. In *Proc. ACM SIGIR Intl. Conf. on Information retrieval (SIGIR)*, pages 272–279, 2000.

R. Dawkins. *The Selfish Gene*. Oxford University Press, 1989.

M. De Choudhury. Tie formation on twitter: Homophily and structure of egocentric networks. In *Proc. Privacy, security, risk and trust (PASSAT), IEEE Intl. Conf. on Social computing (SocialCom)*, pages 465–470, 2011.

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

S. Dorogovtsev, J. Mendes, and A. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(21):4633–4636, 2000.

R. I. M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 9(10):178–190, 1998.

N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

N. Ellison, R. Heino, and J. Gibbs. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, 11(2):415–441, 2006.

P. Erdös and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61, 1960.

J. Falkinger. Attention economies. *Journal of Economic Theory*, 133(1):266–294, 2007.

E. Ferrara, M. JafariAsbagh, O. Varol, V. Qazvinian, F. Menczer, and A. Flammini. Clustering memes in social media. In *IEEE/ACM Intl. Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, 2013.

M. Fielder. Algebraic connectivity of graphs. *Czech Math. J.*, 23:298–305, 1973.

A. T. Fiore and J. S. Donath. Homophily in online dating: when do you like someone like yourself? In *Proc. ACM CHI Extended Abstracts on Human Factors in Computing Systems (CHI EA)*, pages 1371–1374, 2005.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

S. Fortunato, A. Flammini, and F. Menczer. Scale-free network growth by ranking. *Physical Review Letters*, 96(21):218701, 2006.

N. Friedkin. A test of structural features of granovetter's strength of weak ties theory. *Social Networks*, 2(4):411–422, 1980.

S. Gaffney and P. Smyth. Trajectory clustering with mixtures of regression models. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 63–72, 1999.

L. Gallos, D. Rybski, F. Liljeros, S. Havlin, and H. Makse. How people interact in evolving online affiliation networks. *Physical Review X*, 2(3):031014, 2012.

A. Galstyan and P. Cohen. Cascading dynamics in modular networks. *Physical Review E*, 75(3):036109, 2007.

W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers – predicting information cascades in microblogs. In *Proc. Workshop on Online Social Networks (WSON)*, 2010.

S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. *ACM SIGOPS Operating Systems Review*, 37(5):29–43, 2003.

E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proc. ACM Intl. Conf. on Human Factors in Computing Systems (CHI)*, pages 211–220, 2009.

M. Girvan and M. Newman. Community structure in social and biological networks. *Proc. Nat. Acad. Sci. (PNAS)*, 99(12):7821–7826, 2002.

J. P. Gleeson. Cascades on correlated and modular random networks. *Physical Review E*, 77(4):046117, 2008.

S. Goel, W. Mason, and D. J. Watts. Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology*, 99(4):611, 2010.

M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2009.

W. Goffman and V. A. Newill. Generalization of epidemic theory: an application to the transmission of ideas. *Nature*, 204(4955):225–228, 1964.

D. Goldberg, D. Nichols, B. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of ACM*, 35:61–70, 1992.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 3(12):211–223, 2001.

M. H. Goldhaber. The attention economy and the net. *First Monday*, 2(4), Apr. 1997.

B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PLOS ONE*, 6(8):e22656, 2011.

S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno. Assessing the bias in communication networks sampled from twitter. 1212.1684, arXiv, 2012.

P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz. Social features of online networks: The strength of intermediary ties in online social media. *PLOS ONE*, 7(1):e29358, 2012.

M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1, 1973.

M. Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, 1(1):201–233, 1983.

M. Granovetter. *Getting a job: A study of contacts and careers*. University of Chicago Press, 1995.

M. S. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1433, 1978.

V. Grolmusz. A note on the pagerank of undirected graphs. Technical report, arXiv:1205.1960, 2012.

D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. Intl. World Wide Web Conf. (WWW)*, pages 491–501, 2004.

M. Guerini, C. Strapparava, and G. Özbal. Exploring text virality in social networks. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, pages 506–509, 2011.

L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11 (9):1074 – 1085, 1992.

T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):784–796, 2003.

C. Haythornthwaite and B. Wellman. Work, friendship, and media use for information exchange in a networked organization. *Journal of American Society for Information Science and Technology (JASIST)*, 49(12):1101–1114, 1998.

N. O. Hodas and K. Lerman. How visibility and divided attention constrain social contagion. In *Proc. ASE/IEEE Intl. Conf. on Social Computing*, pages 249–257, 2012.

T. Hogg and K. Lerman. Stochastic models of user-contributory web sites. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2009.

P. Holme and M. E. J. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.

J. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proc. Intl. World Wide Web Conf. (WWW)*, pages 57–58, 2011.

B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), Jan 2009.

D. Ienco, F. Bonchi, and C. Castillo. The meme ranking problem: Maximizing microblogging virality. In *IEEE Intl. Conf. on Data Mining Workshops (ICDMW)*, pages 328–335, 2010.

S. Jamali and H. Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. In *Proc. Intl. Conf. on Web Information Systems and Mining (WISM)*, pages 32–38, 2009.

K. H. Jamieson and J. N. Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, USA, 2009.

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proc. ACM 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, pages 56–65, 2007.

J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring tie strength from online directed behavior. *PLOS ONE*, 8(1):e52168, 2013.

A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *Proc. Latin American Web Conf.*, pages 55–57, 2007.

D. B. Kandel. Homophily, selection, and socialization in adolescent friendships. *American Journal of Sociology*, pages 427–436, 1978.

B. Karrer and M. Newman. Competing epidemics on complex networks. *Physical Review E*, 84(3):036106, 2011.

M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.

B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal*, pages 291–307, 1970.

M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6 (11):888–893, 2010.

J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: measurements, models and methods. *Lecture Notes in Computer Science (LNCS)*, 1627: 1–18, 1999.

B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: European Conf. on Machine Learning (ECML)*, pages 217–226. Springer, 2004.

G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network1. *American Journal of Sociology*, 115(2):405–450, 2009.

D. Krackhardt and M. S. Handcock. Heider vs. simmel: Emergent features in dynamic structure. In *Statistical Network Analysis: Models, Issues, and New Directions*, pages 14–27. Springer, 2007.

P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. IEEE Annual Symposium on Foundations of Computer Scienc*, pages 57–65, 2000.

H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. ACM Intl. World Wide Web Conf. (WWW)*, pages 591–600, 2010.

A. Labrinidis and H. Jagadish. Challenges and opportunities with big data. *Proc. VLDB Endowment*, 5(12):2032–2033, 2012.

A. Lakhina, J. W. Byers, M. Crovella, and P. Xie. Sampling biases in ip topology measurements. *Annual Joint Conf. IEEE Computer and Communications*, 1:332–341, 2003.

D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *Proc. ACM SIGMOD Intl. Conf. on Management of data (SIGMOD)*, 2007.

J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proc. ACM Intl. World Wide Web Conf. (WWW)*, pages 251–260, 2012.

S. Lenser and M. Veloso. Non-parametric time series classification. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2005.

K. Lerman. Social information processing in news aggregation. *IEEE Internet Computing*, 11(6):16–28, 2007.

K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, pages 90–97, 2010.

K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. Intl. World Wide Web Conf. (WWW)*, 2010.

J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 631–636, 2006.

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

J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Pattern and a model. In *Proc. SIAM Intl. Conf. on Data Mining (SDM)*, pages 551–556, 2007b.

J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 462–470, 2008.

J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 497–506, 2009.

D. Z. Levin and R. Cross. The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management science*, 50(11):1477–1490, 2004.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of American Society for Information Science and Technology (JASIST)*, 58(7):1019–1031, 2007.

N. Lin, W. M. Ensel, and J. C. Vaughn. Social resources and strength of ties: Structural factors in occupational status attainment. *American Sociological Review*, pages 393–405, 1981.

G. Lindzey and E. Aronson, editors. *Handbook of Social Psychology: Group Psychology and the Phenomena of Interaction*. Longman Higher Education, 1985.

T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to predict reciprocity and triadic closure in social networks. *ACM Trans. on Embedded Computing Systems*, 9(4), 2010.

L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos Is Eighty*, 2 (1):1–46, 1993.

Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of American Society for Information Science and Technology (JASIST)*, 64(7):1399–1410, 2013.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947.

L. Manovich. Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, pages 460–75, 2011.

B. Markines and F. Menczer. A scalable, collaborative similarity measure for social annotation systems. In *Proc. ACM Conf. on Hypertext and Hypermedia (HT)*, pages 347–348, 2009.

B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proc. Intl. World Wide Web Conf. (WWW)*, pages 641–650, 2009.

W. Mason, A. Jones, and R. Goldstone. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422, 2008.

W. Mason, J. W. Vaughan, and H. Wallach. Computational social science and social computing. *Machine Learning*, pages 1–4, 2013.

J. McNames. A nearest trajectory strategy for time series prediction. In *Proc. Intl. Workshop on Advanced Black-Box Technique for Nonlinear Modeling*, pages 112–128, 1998.

J. M. McPherson and L. Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review*, 52(3): 370–379, 1987.

M. McPherson, L. Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

F. Menczer. Lexical and semantic clustering by web links. *Journal of American Society for Information Science and Technology (JASIST)*, 55(14):1261–1269, 2004.

F. Menczer and R. K. Belew. Adaptive information agents in distributed textual environments. In *Proc. Intl. Conf. on Autonomous Agents*, pages 157–164, 1998.

G. Mesch and I. Talmud. The quality of online and offline relationships: The role of multiplexity and duration of social relationships. *The Information Society*, 22(3):137–148, 2006.

M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proc. ACM Workshop on Analytics for noisy unstructured text data*, pages 73–80, 2010.

E. Michlmayr and S. Cayzer. Learning user profiles from tagging data and leveraging them for personal (ized) information access. In *Proc. Workshop on Tagging and Metadata for Social Information Organization*, 2007.

S. Milgram. The small world problem. *Psychoology Today*, 2(1):60–67, 1967.

G. Miritello, E. Moro, and R. Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, 2011.

Y. Moreno, M. Nekovee, and A. Vespignani. Efficiency and reliability of epidemic data dissemination in complex networks. *Physical Review E*, 69(5):055101, 2004.

M. Morris and M. Kretzschmar. Concurrent partnerships and transmission dynamics in networks. *Social Networks*, 17(3):299–318, 1995.

S. Morris. Contagion. *Review of Economic Studies*, 67(1):57–78, 2000.

F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*, 2013.

M. Moussaid, D. Helbing, and G. Theraulaz. An individual-based model of collective attention. In *Proc. European Conference on Complex Systems*, 2009.

L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

S. A. Munson and P. Resnick. Presenting diverse political opinions: how and how much. In *Proc. ACM Intl. Conf. on Human factors in computing systems (CHI)*, pages 1457–1466, 2010.

R. E. Nelson. The strength of strong ties: Social networks and intergroup conflict in organizations. *Academy of Management Journal*, 32(2):377–401, 1989.

M. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701, 2002a.

M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2): 167–256, 2003.

M. E. J. Newman. Modularity and community structure in networks. *Proc. Nat. Acad. Sci. (PNAS)*, 103(23):8577–8582, 2006.

M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

M. E. J. Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–32, 2012.

M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proc. Nat. Acad. Sci. (PNAS)*, 99(Suppl 1):2566–2572, 2002.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. De Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179, 2007a.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proc. Nat. Acad. Sci. (PNAS)*, 104(18):7332–7336, 2007b.

S. Pajevic and D. Plenz. The organization of strong links in complex networks. *Nature Physics*, 8(5):429–436, 2012.

G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

F. Papadopoulos, M. Kitsak, M. Ángeles Serrano, M. Boguña, and D. Krioukov. Popularity versus similarity in growing networks. *Nature*, 489(7417):537–540, 2012.

R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001.

N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. Time scales and dynamical processes in activity driven networks. *Nature Scientific Reports*, 2:469, 2012.

O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proc. ACM Intl. Conf. on Recommender Systems*, pages 385–388, 2009.

S. Pigolotti, A. Flammini, and A. Maritan. A stochastic model for the species abundance problem in an ecological community. *Physical Review E*, 70:011916, 2004.

H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 365–374, 2013.

P. Pons and M. Latapy. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS*, pages 284–293. Springer, 2005.

R. D. Putnam. *Bowling alone: The collapse and revival of American community*. Simon and Schuster, 2001.

J. M. Quigley. Urban diversity and economic growth. *Journal of Economic Perspectives*, 12:127–138, 1998.

A. Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *Bulletin of mathematical biophysics*, 15(523–533), 1953.

J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Physical Review Letters*, 105(15): 158701, 2010.

E. Rogers. *Diffusion of Innovations*. Free Press, 5th edition, 2003.

D. M. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2010.

D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *Proc. 20th Intl. Conf. on World Wide Web (WWW Companion Volume)*, pages 113–114, 2011a.

D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proc. Intl. Conf. on World Wide Web (WWW)*, 2011b.

D. M. Romero, C. Tan, and J. Ugander. On the interplay between social and topical structure. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 2013.

M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Nat. Acad. Sci. (PNAS)*, 105(4):1118–1123, 2008.

M. Salganik, P. Dodds, and D. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: social link prediction from shared metadata. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 271–280, 2010.

H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, E. Blanco, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar. Toward personality insights from language exploration in social media. In *AAAI Spring Symposium Series*, 2013.

J. Scott. *Social Network Analysis: A Handbook.* Sage, London, 2000.

M. A. Serrano, D. Krioukov, and M. Boguná. Self-similarity of complex networks and hidden metric spaces. *Physical Review Letters*, 100(7):078701, 2008.

C. Shalizi and A. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

G. Simmel and K. H. Wolff. *The Sociology of Georg Simmel*. The Free Press, 1950.

M. P. Simmons, L. A. Adamic, and E. Adar. Memes online: Extracted, subtracted, injected, and recollected. In *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*, 2011.

H. Simon. Designing organizations for an information-rich world. In M. Greenberger, editor, *Computers, Communication, and the Public Interest*, volume 72, pages 37–52. The Johns Hopkins Press, 1971.

K. Sneppen, A. Trusina, M. H. Jensen, and S. Bornholdt. A minimal model for multiple epidemics and immunity spreading. *PLOS ONE*, 5(10):e13326, 2010.

B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proc. IEEE Intl. Conf. on Social Computing (SocialCom)*, pages 177–184, 2010.

X. Sun, J. Kaur, S. Milojevic, A. Flammini, and F. Menczer. Social dynamics of science. *Nature Scientific Reports*, 3(1069), 2013.

J. Surowiecki. *The wisdom of crowds*. Random House Digital, Inc., 2005.

G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 807–816, 2009.

A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida. Predicting the popularity of online articles based on user comments. In *Proc. Intl. Conf. on Web Intelligence (WI)*, 2011.

O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 643–652, 2012.

J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proc. Nat. Acad. Sci. (PNAS)*, 109(16):5962–5966, 2012.

L. M. Verbrugge. Multiplexity in adult friendships. *Social Forces*, 57(4):1286–1309, 1979.

A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939): 425–428, 2009.

S. Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

S. Wasserman and K. Faust. Social network analysis. *Cambridge University Press.*, 1994.

D. J. Watts. A simple model of global cascades on random networks. *Proc. Nat. Acad. Sci. (PNAS)*, 99(9):5766–5771, 2002.

D. J. Watts. A twenty-first century science. *Nature*, 445(7127):489–489, 2007.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393 (6684):440–442, 1998.

B. Wellman and S. Wortley. Different strokes from different folks: Community ties and social support. *American Journal of Sociology*, 96(3):558–588, 1990.

J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 261–270, 2010.

L. Weng and T. Lento. Topic-based clusters in egocentric networks on facebook. In *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*, 2014.

L. Weng and F. Menczer. Computational analysis of collective behaviors via agent-based modeling. In *Handbook of Human Computation*, pages 761–767. Springer, 2013.

L. Weng and F. Menczer. Topicality and social impact: Diverse messages but focused messengers. *Under review*, arXiv 1402.5443, 2014.

L. Weng, A. Flammini, and F. Menczer. An information propagation model based on user interests. In *Proc. 8th Intl. Conf. on Complex Systems (ICCS)*, 2011.

L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Nature Scientific Reports*, 2(335), 2012.

L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Nature Scientific Reports*, 3(2522), 2013a.

L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini. The role of information diffusion in the evolution of social networks. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 356–364, 2013b.

L. Weng, M. Karsai, N. Perra, F. Menczer, and A. Flammini. Attention on weak ties. In preparation, 2014a.

L. Weng, T. Lento, and M. Burke. Tie strength and topic diversity on facebook. In preparation, 2014b.

L. Weng, F. Menczer, and Y.-Y. Ahn. Predicting meme virality in social networks using network and community structure. In *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*, 2014c.

F. Wu and B. A. Huberman. Novelty and collective attention. *Proc. Nat. Acad. Sci. (PNAS)*, 104(45):17599–17601, 2007.

L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith. Visual memes in social media: tracking real-world news in youtube videos. In *Proc. ACM Intl. Conf. on Multimedia*, pages 53–62, 2011.

Z. Xu, R. Lu, L. Xiang, and Q. Yang. Discovering user interest on twitter with a modified author-topic model. In *Proc. IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 422–429, 2011.

J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proc. ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 177–186, 2011.

L. Yang, T. Sun, and Q. Mei. We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In *Proc. ACM Intl. World Wide Web Conf. (WWW)*, pages 261–270, 2012.

Appendices

Appendix A

Appendix

A.1 Basic Statistics of Datasets

The present studies involve three major datasets from Twitter, Yahoo! Meme, mobile phone call network, Enron email collection, and Facebook. This section introduces the basic statistics of each dataset.

A.1.1 Twitter

The dataset analyzed in Chapter 4 contains more than 120 millions retweets from October 2010 to January 2011, involving 12.5 millions distinct users and 1.3 million hashtags. Each post carries information about who generated and who retweeted it. The following relations were also recovered, yielding an underlying social network with about 4.5 millions nodes and 71.5 millions edges. As expected in a social network, the follower graph has scale-free degree distributions; in-degree and out-degree are correlated (Pearson's correlation coefficient $\rho = 0.5$).

Chapter 5 examines a dataset composed of tweets from the Twitter public stream during April 2012. The social network built upon directed following relationships incorporates 0.6 million nodes and 44.6 millions directed edges.

For studying the role of community structure in meme diffusion in Chapter 8, we harnessed a set of tweets from March 24 to April 25, 2012 with 121.8 millions tweets generated by 14.6 millions unique users and containing 10.4 millions hashtags. Only tweets written in English were extracted. We then constructed an undirected, unweighted network based on reciprocal following relationships among a sample set of Twitter users in this dataset. Basic statistics of the network and communities are reported in Table A.1. This dataset is publicly available at http://carl.cs.indiana.edu/data/#virality2013.

We collected public tweets from January to March 2013 for the study in Chapter 6. We set the first two months as the *observation period* and the last month as the *test period*; the

Table A.1: Basic statistics of the reciprocal follower network used in Chapter 8. Node coverage measures the proportion of nodes belonging to communities that have at least three nodes.

		Reciprocal follower network
	# Nodes	400,020
	# Edges	10,012,989
	Clustering coefficient	0.2093
	Density	12.42×10^{-5}
InfoMap	# Communities	6,569
	Node coverage	99.08%
LinkClust	# Communities	193,805
	Node coverage	43.30%

Table A.2: Basic statistics of the dataset used in Chapter 6, which is split into two periods: observation and testing. About 13% of the tweets contain hashtags.

	Jan-Feb 2013 (Observation)	Mar 2013 (Test)
# Tweets	2,449,711,388	1,339,702,599
# Tweets with hashtags	316,668,998	173,823,786
# Hashtags	27,923,499	16,802,087
# Users	92,356,790	72,963,020

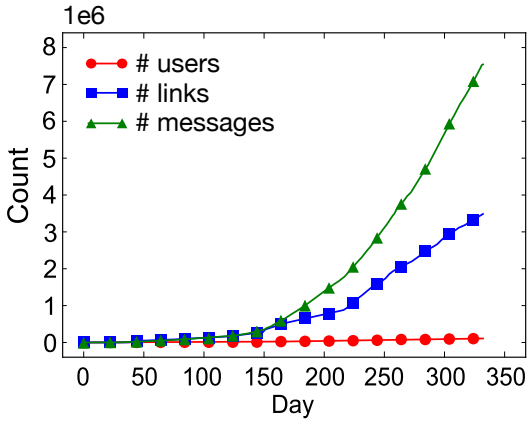
former is used to build up the topic network and quantify user topical interests, and the latter works for evaluating the results of prediction tasks. Table A.2 shows several basic statistics of the dataset, which is publicly available at http://carl.cs.indiana.edu/data/#topic2014.

A.1.2 Mobile Phone Call Network

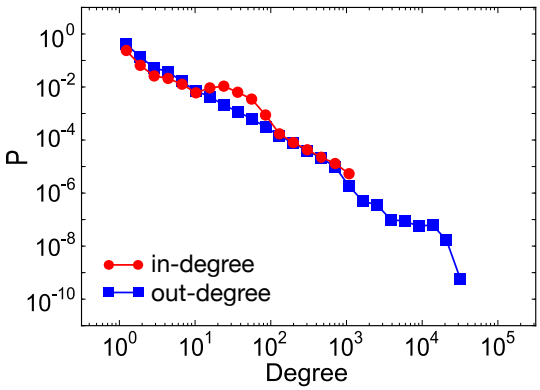
The mobile phone call dataset records about 487 millions directed call events during 120 days with one second resolution. The cumulated social network consists of 6,101,641 nodes and 19,013,221 directed edges which include 11,581,152 mutual links (where at least one call was initiated by each party) and 7,432,069 non-mutual ones (where communication was initiated by only one of the parties).

A.1.3 Enron Email Collection

The Enron email network records 246,391 emails exchanged inside the Enron corporation. The recovered social network based on email exchanges incorporates 86,818 nodes and 359,817 directed edges among which 59,306 links are mutual (where two individuals have sent at one email to one the other) and the rest 299,731 links are non-mutual (where only one party has sent one or more emails to the other).



(a) Network growth



(b) Degree distributions

Figure A.1: General statistics of the Yahoo! Meme dataset. (a) The growth of the system in time, including the number of users (red circles), links (blue squares) and messages (green triangles). (b) Broad distributions of in-degree and out-degree in the follower network of Yahoo! Meme. The maximum in-degree of the nodes is 1,000 due to a limit imposed by the system: users were not allowed to follow more than 1,000 people.

A.1.4 Yahoo! Meme

The Yahoo! Meme follower network in Chapter 9 consists of 128,199 users with at least one edge, connected by a total of 3,485,361 directed edges. Figure A.1 displays general statistics about the growth and structure of the network.

A.1.5 Facebook

To test the relationship between tie strength and topic diversity in Chapter 7, a dataset was constructed of 2.2 millions status updates written by a random sample of 117,091 English-speaking Facebook users (or *egos*) from March 16 to July 15, 2013. People do not actively interact with their entire social networks, so the data includes only Facebook friends (or *alters*) who liked or commented on at least one status update by any sampled egos during the observation time, defined as *active alters*, ending with 30.2 active alters per ego on average.

A.2 Using Twitter Hashtags as Meme Identifiers

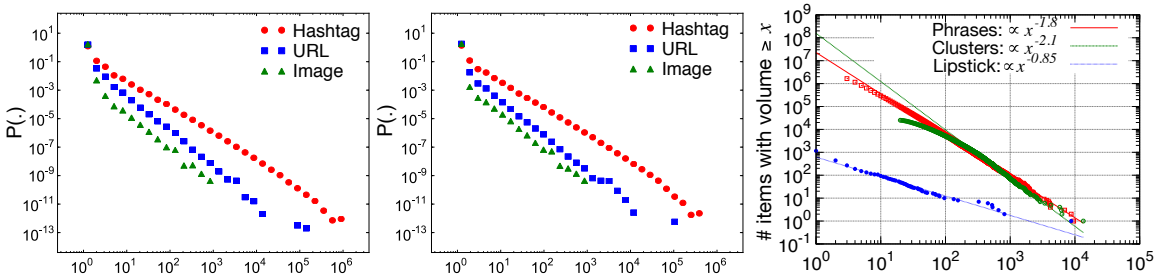
According to the definition in Sec. 3.4.1, hashtags are memes. Most hashtags are unique phrases that clearly spread by imitation, as shown in Table A.3. Moreover, they mutate, compete, and survive: Twitter users quickly reach consensus on hashtags to represent specific or broad topics. For instance, the hashtag #ows was almost instantaneously adopted by

Table A.3: Top 50 common hashtags

#YouGetMajorPointsIf	#YouGetPointsIf
#MyThoughtsDuringSchool	#NationalBestFriendDay
#ThingsISayWhileReadingMytl	#ThingsMostPeopleLikeButIDont
#ThatIrritatesMe	#10ThingsThatILike
#YouKnowWhatsAnnoying	#WeWontWorkOutIf
#ThingsIGottaTeachMySon	#ImHappyWhen
#IFindThatAttractive	#MyThoughtsDuringSex
#WaysToMakeMeHappy	#10PeopleOnTwitterIWantToMeet
#FourWordsYouDontWantToHear	#LOLAtGirlsWho
#4WordsYouDontWantToHear	#ReasonsWeDontTalkAnymore
#SheAintWifeyMaterial	#IWantToPunchPeopleWho
#YouKnowWhatAnnoysMe	#IDidntTextYouBack
#ThingsPeopleDoThatGetOnMyNerves	#ICantDateSomeoneThat
#ReasonsThatImSingle	#HowTopIsSafeMaleOff
#IKnowThisOneGirl	#HotPeopleIFollow
#thingsthatfrustrateme	#MoreFemalesShould
#WordsYouWillNeverHearMeSay	#NoManShouldEver
#SomeTimesIJustWant	#ThingsBlackPeopleTakeSeriously
#ThingsILoveToSee	#AmITheOnlyOneThat
#BackInTheDayWhenIWasAKid	#HardestThingsInLife
#ArentYouTiredOf	#YouDontBelongOnTwitterIf
#IfWeWereTogetherRightNow	#NameYourFavoriterApper
#ButYouathug	#TweetYourHeight
#ThingsThatGuysLlike	#ImMadBecause
#ICantBeTheOnlyPerson	#SomeFactsYouShouldKnow

a community of hundreds of thousands in discussion about the Occupy Wall Street movement, outcompeting other similar hashtags [Conover et al., 2013b].

There are other types of memes (e.g., phrases, URLs, and images). The statistical features of hashtags match those observed in other types of memes (see Fig. A.2). Note that URLs and images do not spread as widely as hashtags in Twitter.



(b) Popularity in # tweets (b) Popularity in # adopters (c) Volume, x

Figure A.2: Popularity distributions of hashtags, URLs, and images, representing different types of memes on Twitter. Popularity can be measured by number of (a) tweets and (b) adopters. (c) The volumes of Internet memes quantified as phrases in blogs and news are similarly distributed [Leskovec et al., 2009].

Appendix B

Resume

Lilian Weng

Emails: lilian.wengweng@gmail.com | weng@indiana.edu
Webpage: http://lilianweng.github.io
Cell: (812) 361-5449
919 E 10th St., Bloomington IN 47408.

EDUCATION

Indiana University Bloomington, IN, USA. (2009.9–2014.4)
Ph.D. in Complex Systems, School of Informatics & Computing, GPA:4/4.
Thesis title: *Information Diffusion on Online Social Networks.*

Hong Kong University, Hong Kong, China. (2006.9–2007.1, short-term exchange)
B.S. in Information Systems, School of Business, GPA:3.80/4.

Peking University, Beijing, China. (2005.9-2009.6)
B.S. in Information Systems and Computer Science, GPA:3.85/4, Rank:1.
Thesis title: *Social Network Analysis of Online Question-Answer Systems*

PUBLICATION

Lilian Weng and Filippo Menczer. Topicality and Social Impact: Diverse Messages but Focused Messengers. *Under review*. 2014.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Predicting Successful Memes using Network and Community Structure. In: *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*. 2014.

Lilian Weng and Thomas Lento. Topic-based Clusters in Egocentric Networks on Facebook. In: *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*. 2014.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality Prediction and Community Structure in Social Networks. *Nature Scientific Report*. (3)2522, 2013. (Media coverage: [1])

Lilian Weng, Jacob Ratkiewicz, Nicola Perra, Bruno Gonçalves, Carlos Castillo, Francesco Bonchi, Rossano Schifanella, Filippo Menczer, and Alessandro Flammini. The Role of Information Diffusion in the Evolution of Social Networks. In: *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*. 2013.

Lilian Weng and Filippo Menczer. Computational Analysis of Collective Behaviors via Agent-Based Modeling. *Handbook of Human Computation*, Springer, 2013.

Lilian Weng and Filippo Menczer. Emergent Semantics from Game-induced Folksonomies. In: *Proc. ACM SIGKDD Crowdsourcing and data mining workshop (CrowdKDD)*. 2012.

Lilian Weng, Alessandro Flammini, Alessandro Vespignani and Filippo Menczer. Competitions among topics in a world with limited attention. *Nature Scientific Report*, (2)335, 2012. (Media coverage: [1][2][3][4])

Lilian Weng and Filippo Menczer. Context Visualization for Social Bookmark Management. Technical Report 1211.6799 [cs.HC], *arXiv*, 2012.

Lilian Weng, Rossano Schifanella and Filippo Menczer. Design of Social Games for Collecting Reliable Semantic Annotations. In: *Proc. IEEE Intl. Conf. on Computer games (CGAMES)*. 2011.

Lilian Weng, Alessandro Flammini and Filippo Menczer. An Information Propagation Model Based on User Interests. In: *Proc. 8th Intl. Conf. on Complex systems (ICCS)*, 2011.

Lilian Weng, Rossano Schifanella and Filippo Menczer. The Chain Model for Social Tagging Game Design. In: *Proc. ACM Intl. Conf. on Foundation of digital games (FDG)*, 2011.

Li(Lilian) Weng and Filippo Menczer. GiveALink Tagging Game: An Incentive for Social Annotation. In: *Proc. ACM SIGKDD Human computation workshop (HComp)*. 2010.

ACADEMIA SERVICE

2011.10, Extended Reviewer for WWW2012 (Social networks track).

2013.4, Reviewer of Journal Technological Forecasting & Social Change.

2013.6, Reviewer of Nature Scientific Report.

2013.11, Reviewer of New Media & Society

2013.11, Reviewer of Journal Technological Forecasting & Social Change.

2014, PC Member of 5th ACM Web Science Conference 2014.

2014, PC Member of the 6th International Conference on Social Informatics 2014.

2014, PC Member of the 1st Data Visualization Workshop of ACM Hypertext 2014.

PATENT

US Patent IURTC No. 13183; filed March 12, 2014.
Title: Systems and Methods to Predict Meme Virality Using Network Structure
Owners: Yong-Yeol Ahn, Lilian Weng and Filippo Menczer

SKILLS

Proficient in Python, C++.

Familiar with Java, C#, Ruby, Object-C; HTML, CSS, Javascript, Ajax; Django, RubyOnRails; Apache, MySQL, WAMP/MAMP. Experience in C, Perl, PHP.

Cloud computing skills including HIVE, MapReduce, Hadoop.

Data analysis in Python, R, and Matlab.

Rich experience with analyzing big data.

Frequent user of Adobe Photoshop, Flex/Flash.

Experience and good sense in user experience design.

Mobile programming for Window phone, iPhone/iPad.

WORK EXPERIENCE

Data Scientist Intern, Data Science, Facebook Inc. (Summer 2013)

- Study the relationship between Facebook post virality and various innate features of the content, characteristics of early reshare users, and properties of the creators, aiming at improving Ads targeting strategies to trigger bigger cascades.
- Compare users with complex social circles and others with simple ego graphs in terms of behavioral patterns, efforts in maintaining friendship, and effects on boosting content popularity through resharing.

Software Engineer Intern, Data Science, Facebook Inc. (Summer 2012)

- Investigate how friends are clustered according to conversation topics in an ego-centric viewpoint and how the topic selecting behavior is restricted by social relationship. The study is intended to provide insights into several Facebook products like the measure of social tie strengths, friends recommendation, and newsfeed ranking.
- Publication: L. Weng and T. Lento. Topic-based Clusters in Egocentric Networks on Facebook. In: *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*. 2014.

User Research Intern, Mozilla Labs, Mozilla Corporation. (Summer 2011)

- Design, implement, and analyze two user studies for Firefox new tab redesign, aiming to better understand how people use new tabs while navigating the Web through quantitative user research. [Related links: 1, 2, 3, 4]
- Help set up several other user tests on *Test Pilot*, an internal platform collecting structured user feedback through Firefox.

Research Intern, eBay Research Labs, eBay Inc. (Summer 2010)

- Work on data tracking, data analysis and personalization algorithm improvement for eBay Discover. [http://discover.ebay.com]
- Design and develop an iPad app prototype which provides user experience similar to reading a real catalog but with functions of easily sharing and saving eBay products.

User Experience Intern, Yahoo! China, Alibaba.com. (2009.3-4)

- Join the team of Linezing Analytics (original Yahoo! Analytics), with services specially designed for sellers on Taobao.com. [http://www.linezing.com]
- Product prototype design and user interaction design.

RESEARCH INTERESTS

Complex networks and systems; Data mining; Web mining; Machine learning; Network community structure; Information diffusion on social networks; Modeling of dynamical processes on networks.; Social media and social networks analysis; Social web application; User experience research; Human-computer interaction.

RESEARCH EXPERIENCE

Truthy.indiana.edu (2010.12 - Present): a research project for better understanding how memes spread online (Python, C++). Indiana University Bloomington.

- Explore the role of limited user attention in determining the virality of memes by proposing a parsimonious agent-based model to investigate whether competition affects the broadly distributed meme popularity, the diversity of information that people are exposed to, and the fading of our collective interests for specific topics.
- Study the connection between network community structure and information diffusion processes. We are able to estimate and predict the future degree of meme virality by characterizing the early spreading patterns of memes in terms of network community structure.

GiveALink.org (2009.9 - 2011.9): a research-oriented online social tagging system (Ruby on Rails). Indiana University Bloomington.

- System maintenance and optimization of GiveALink.org.
- Work on several issues related with the social annotations, i.e. social tagging games, spam detection, API methods design and implementation, etc.
- Design and implement GiveALink Slider (http://slider.givealink.org), a social tagging game as incentive for generating high-quality social annotation data; people can contribute social annotations when they are having fun in the game.

Research Assistant (2007.8-2008.4), working an Web-based social bookmarking application (C#). KVision Research Group, Peking University.

- Aim to strengthen the loose structure of folksonomy using semantic Web techniques such as ontology.

TEACHING EXPERIENCE

Associate Instructor (2010.9-2010.12) for INFO-I527: Search Informatics, School of Informatics and Computing, Indiana University Bloomington

MAIN COURSES

Algorithm & theory of computing; Machine learning; Web Mining; Introduction to complex systems; Seminars in complex system; Bayesian data analysis; Natural language processing; Cloud computing; Mobile computing; Design and analysis of secure protocols & systems.

SCHOLARSHIP

2010.4, Women in Computing (WIC) Grad Cohort Scholarship.

2009.6, Graduate with Honor of Peking University.

2007.3-2008.11, President's Undergraduate Research Fellowship.

2007-2008, National Scholarship.

2008-2009, National Scholarship.

Fall 2006, Li & Fung Scholarship, by the Li & Fung Foundation Limited.

VOLUNTEER WORKS

2008.8, Volunteer of Beijing 2008 Olympic Games.

2008.9-10, JING Forum 2008 between Peking University and University of Tokyo.