공학박사 학위논문

Understanding and Predicting User Behavior and Content Propagation Patterns in Internet: A Data-Scientific Approach

데이터 과학 분석 방법에 기반한 온라인 상의 사용자 행동 및 콘텐트 전파 패턴 이해 및 예측

2019년 2월

서울대학교 대학원

전기.컴퓨터공학부

최 대 진

Understanding and Predicting User Behavior and Content Propagation Patterns in Internet: A Data-Scientific Approach

데이터 과학 분석 방법에 기반한 온라인 상의 사용자 행동 및 콘텐트 전파 패턴 이해 및 예측

지도교수 권 태 경

이 논문을 공학박사 학위논문으로 제출함

2018 년 11 월

서울대학교 대학원

전기.컴퓨터공학부

최 대 진

최 대 진의 박사 학위논문을 인준함

2018 년 12 월

위 원 장	최 양 희	(인)
부위원장	권 태 경	(인)
위 원	김 종 권	(인)
위 원	나 종 연	(인)
위 원	차 미 영	(인)

Abstract

Understanding and Predicting User Behavior and Content Propagation Patterns in Internet: A Data-Scientific Approach

Daejin Choi
Department of Electrical Engineering & Computer Science
The Graduate School
Seoul National University

It becomes a norm for people to communicate with one another through various online social channels, such as message boards, online social networks, and social media. As these online digital channels of communications are producing a deluge of social data, computational data-driven studies have in turn spurred to understand human behaviors and communication patterns. As part of such studies, this thesis studies online communications from the following topics: (i) characterizing threaded conversations in terms of content, user, and community perspectives, (ii) characterizing popular and viral image propagation, and (iii) understanding content publishing and sharing patterns. To this end, three large-scale datasets that contain (i) 0.7 million threaded conversations from 1.5 million users from Reddit, (ii) 0.3 million images shared by 1 million users from Pinterest, and (iii) 4.2 billion requests for 80 million URLs created through Bitly are collected. The data-driven analysis on the datasets reveals that content, user behavioral, and topical community factors (e.g., difficulties of texts, portion of reciprocal communications, or discussion-encouraged communities) are highly associated with the large, responsive, or viral conversations. Through in-depth analysis on Pinterest dataset, this thesis shows that structural virality of image cascade differentiates large cascades in terms of its shape (i.e., broadcast or diffusion) and factors such as propagating time are differently related to the volume and virality. By modeling the relations among web sites (e.g., twitter.com, amazon.com) for content sources and publishing spaces from Bitly dataset, this thesis finds that they play different roles in publishing short URLs. For example, search engines, online social networks, and computer & electronics sites like newsfeed services are popular spaces for content publishing while news and streaming services are widely used as content sources. The analysis of content publishing and sharing patterns through URL shortening reveals that users are likely to access different types of content via different websites. For example, adult or malicious content tend to be requested from search engines, shopping content is primarily accessed through online social networks, and news content is usually clicked through computer & electronics websites. This thesis also reports that news or shopping content, published through online social networks, tend to be requested quickly and virally. Lastly, based on the lessons learned, a learning-based model to predict whether a conversation or an image cascade would be large or viral is proposed, which achieves a high performance. By giving valuable insights on understanding (i) how different users interact with others across different content, topics, and communities, (ii) what and how content is propagated in a viral manner, and (iii) how different content is published and accessed through different online spaces, this thesis is believed to contribute to better online services such as marketing or novel platform design.

Keywords: Computational Social Science, Machine Learning, Reddit, Pinterest, Bitly, Content Propagation, User Behavior Analysis, Data Science

Student Number: 2012-23243

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The recent advances in information technology have been revolutionizing how people communicate with one another. *Online* communication channels, such as messengers, online social networks (OSNs), and social media, have become indispensable in everyday life. These online digital channels of communication are not only facilitating interactions among people and the dissemination of new content, but also producing a deluge of social data. Such data in turn enables computational data-driven studies on human behaviors and communication patterns, often dubbed as "Computational Social Science" [50].

From traditional message boards such as USENET and BBS to more recent OSNs such as Facebook, Twitter, and Pinterest, there have been many computational data-driven studies that provide valuable insights into human behaviors and content dissemination patterns, which examine creating, publishing, and sharing content on various online spaces [55, 54, 48, 34, 64, 70, 3, 4, 72, 11, 62, 17]. Some studies have investigated online communication behaviors across different platforms including online chatting [55], online forums [54, 48, 34], and OSNs [70]. Otheres have paid huge efforts in studying information adoption and propagation in various OSNs [3, 4, 72, 11, 62, 17].

As part of the computational social science on understanding user behavior and content propagation patterns, this thesis studies online communications from the following topics: (i) characterizing threaded conversations in terms of content, user, and community perspectives, (ii) characterizing popular and viral image propagation, and (iii) understanding content publishing and sharing patterns. Using the result from the analysis, the machine learning-based applications to predict whether a given threaded conversation or an image will be popular or viral in very early time are designed and evaluated. By giving valuable insights on understanding (i) how different users interact with others across different content, topics, and communities, (ii) what and how content is propagated in viral manner, and (iii) how different content is published and accessed through different online spaces, this thesis is believed to contribute to better online services such as marketing or novel platform design.

For the starting point of such data-driven analysis, three large-scale datasets that contain (i) 0.7 million posts and 18 million associated comments generated by 1.5 million users) collected from Reddit, (ii) more than 0.3 million images shared by 1 million users collected from Pinterest, and (iii) 4.2 billion requests for 80 million URLs created through `Bit.ly`, an URL shortening service, are collected. To this end, we keep track of all the newly-uploaded posts and their follow-up comments in Reddit and a propagation path of each image in Pinterest, then model each of threaded conversation and a propagation (or cascade) of an image as a *comment tree* and a *pin tree*, respectively. Using the three datasets and models, this thesis explore (i) how the content properties and user participation behaviors are associated with threaded conversations by characterizing comment trees in terms of volume, responsiveness, and virality, then comparing properties of content and user behavior with three characteristics, (ii) how macro-level virality (i.e., volume) is different from micro-level virality of image cascades and what features are related to the both types of virality, and (iii) what types of web pages (e.g., news, adult, or sports) are shortened and shared through different publishing spaces. To leverage the implications of the findings, this thesis investigates how content and user behavioral features from initial observation can predict large and viral conversations or image cascades, by evaluating the machine learning-based model.

The main contributions of this thesis are as follows:

- **Content and User Behavior Analysis on Threaded Conversations:** This thesis explores whether content properties (e.g., sentiment or text difficulty), users' participation behaviors, and characteristics of topical communities (i.e., subreddit) are associated with the volume, responsiveness, and virality of conversations. Interestingly, the difficulty of content texts is an important indicator that can differentiate large/viral and responsive conversations; a large/viral conversation is likely to have difficult texts, whereas a responsive conversation tends to have plain texts. The further analysis in this thesis also finds that a large and viral comment tree is often generated by a small portion of users who reciprocally communicate with each other in the tree. This thesis further reveals that subreddits that deal with similar conversation topics tend to show similar communication patterns in terms of volume, responsiveness, and virality; the news-related and image-based subreddits are more likely to have large and responsive conversations, respectively, and the conversations in discussion-driven subreddits tend to be viral, implying that discussions are likely to recursively elicit many other users to join the conversations. Interestingly, users participating in multimedia-related subreddits tend to react mainly to the original content (i.e., post) rather than derived conversations (i.e., threaded comments), and likely to use emotional words in their posts.

- **Propagation Pattern and Factor Analysis on Image Cascades:** This thesis sheds light on understanding how macro-level virality (i.e., volume) of image cascade in Pinterest is different from micro-level virality (i.e., Weiner Index). The analysis in this thesis reveals that although there is overall a positive correlation between the volume and structural virality of an image cascade, popular images are not necessarily viral, by indicating two images of similar popularity propagate through very different scenarios, i.e., one by broadcast and the other by a person-to-person contagion process.

- **Content Publishing and Sharing Behavior Analysis through URL Shortening Service:** This thesis explores who shortens and publishes content URLs reveals that web pages are primarily shortened through the third party companies, e.g., Twitter (twitterfeed, tweetdeckapi, and twipple), Bit.ly (bitly and zatbitly), Facebook (rssgraffiti), substantially more than by individual user accounts. The further investigation on the request patterns of short URLs find that short URLs are proliferated mostly across OSNs, news/media sites, and computer/electronics sites (e.g., newsfeed service), and OSN pages tend more to be requested. This thesis also reveals that the URL shortening practice and request patterns show disparate patterns. For example, while there are not so many short URLs for the shopping websites or adult content, they are likely to be requested notably. To shed light on the practice of content publishing through short URLs, this thesis models the relations among content and referrer domains[1], in the form of *content-referrer graph.* The analysis of the content-referrer graph reveals that different domains play different roles in publishing short URLs. For example, search engines, OSNs, and computer/electronics sites are popular spaces for content publishing while news and streaming services are widely used as content sources in general. This thesis also shows that users are likely to access different types of content pages through different referrer domains; e.g., adult or malicious content pages tend to be requested from search engines; shopping content is primarily accessed through OSNs; news is usually clicked through computer/electronics domains. Also, news and shopping pages, published through OSNs, tend to be requested quickly and virally.

- **Predicting Large and Viral Conversations and Image Cascade:** We leverage the implications of our findings on threaded conversations

[1] A referrer domain indicates a domain where a short URL is published, while a content domain represents a domain whose content is created.

(and image cascades) for predicting the large and viral comment trees (and pin trees) in practical and engineering standpoint. In particular, this thesis assumes three different scenarios in terms of available information : (i) If only a post (an image) is given and available, (ii) At the moment when a post (an image) is published and its meta and/or poster information is available, and (iii) If an initial image propagation pattern is observable. Our proposed model shows that features of a post (an image) itself is not as effective in predicting popular or viral cascades, which is in line with previous work in other OSNs [15, 68]. However, this thesis shows that the prediction model can be improved by combining the information of post (image) itself, meta features (such as category and source domain), poster (pinner) features, and initial propagation patterns can accurately identify large or viral conversations (image cascades), Throughout the extensive feature study, this thesis reveals that user participation behaviors are important in predicting the large conversations, while the content features are good predictors in predicting viral conversations. In case of predicting large or viral image cascades in Pinterest, this thesis indicates that image meta and pinner features are the strongest predictors in predicting large image cascades, implying that we can accurately forecast the image popularity using the image meta information (e.g., category, source, or title) and poster's information (e.g., his/her connectivity or activity), at the moment when the image is posted. On the other hand, the initial propagation pattern of an image and its meta information are the best predictors in predicting viral image cascades, implying that if we observe the *initial propagation* *pattern* of an image, we can accurately predict whether the image will go viral in the future.

The rest of this thesis is organized as follows. We present the background of Reddit, Pinterest, and Bitly and review the related work in Chapter 2. We explain our measurement and analysis methodologies in Chapter 3. We start our analysis on threaded conversations in Chapter 4, followed by investigation of image cascade in Chapter 5. We then report the analysis result on content publishing and sharing patterns through Bitly in Chapter 6. We finally propose and evaluate models to predict large or viral conversations (and image cascades), followed by concluding remarks in Chapter 8.

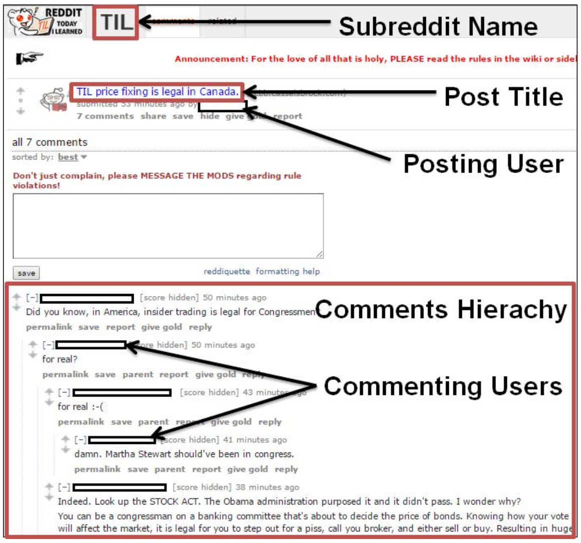
Chapter 2

Background

2.1 Reddit, Pinterest, and Bitly

2.1.1 Reddit

Reddit allows users to share news, articles, and opinions with each other in the areas of interests. The areas of topical interests in Reddit are called “subreddits”, each of which serves as an independent community. A subreddit can be created by any user who is interested in any particular topic, e.g., game, politics, or sports. Each subreddit is managed by several “moderators” under its own roles and policies. In each subreddit, users can (i) submit content (i.e., write a post), (ii) write a comment to a post, or (iii) write a comment to another comment. Figure 2.1 shows an illustration of a post and its associated comments in “Today I Learned (TIL)” subreddit. Note that we collectively refer to both a post and a comment as a “message”.



Announcement: For the love of all that is holy, PLEASE read the rules in the wiki or side

TIL price fixing is legal in Canada.
submitted 53 minutes ago by [redacted]
7 comments share save hide give gold report

all 7 comments
sorted by: best

Don't just complain, please MESSAGE THE MODS regarding rule violations!

save reddiquette formatting help

[–] [redacted] [score hidden] 50 minutes ago
Did you know, in America, insider trading is legal for Congressmen
permalink save report give gold reply

[–] [redacted] [score hidden] 50 minutes ago
for real?
permalink save parent report give gold reply

[–] [redacted] [score hidden] 43 minutes ago
for real :-(
permalink save parent report give gold reply

[–] [redacted] [score hidden] 41 minutes ago
damn. Martha Stewart should've been in congress.
permalink save parent report give gold reply

[–] [redacted] [score hidden] 38 minutes ago
Indeed. Look up the STOCK ACT. The Obama administration purposed it and it didn't pass. I wonder why?
You can be a congressman on a banking committee that's about to decide the price of bonds. Knowing how your vote will affect the market, it is legal for you to step out for a piss, call you broker, and either sell or buy. Resulting in huge

Figure 2.1 A post with its associated comments in the subreddit "Today I Learned (TIL)" is illustrated.

2.1.2 Pinterest

Pinterest [74, 10, 30, 59] is a pinboard-style content sharing platform that allows users to exhibit collections of images or videos. The main idea of Pinterest is to collect, organize, and share content (mostly images since image content is dominant in Pinterest) that users find interesting; Pinterest focuses on collecting and sharing content (i.e., pins in Pinterest). That is, Pinterest's basic function is to let users collect, organize, and share pins by their tastes or interests. Direct communications (e.g., private messages in Facebook or Twitter) between users are not available in Pinterest. Instead, user interactions mostly occur at the time they write feedbacks on or share pins (e.g., a user can *like* or *comment* on someone's pin). We describe key terminologies in Pinterest below.

- Pin/Repin: Each image/video is called as a pin, and the act of posting a pin is referred to as pinning. If the posted pin is shared by someone, the shared pin is called as a repin, which is similar to *retweet* in Twitter, and the act of sharing other user's pin is called repinning. Users who post and repins a pin are the (original) pinner and repinner, respectively.

- Source: A user can directly upload a pin or fetch a pin from other websites like *Tumblr.com.* In the latter case, the URL of the pin is referred to as a source.

- Like/Comment: Similar to Facebook, a user can push a like button for a pin that she likes and leave a comment on a pin.

- Pinboard/Category: A pinboard is a collection of pins organized by a user. Each pinboard belongs to one of the categories in Pinterest. At the moment there are 32 categories in Pinterest, varying from "animals" to "history" to "women's fashion".

- Following/Follower: Like Twitter, the relation between two users in Pinterest is not symmetric. The fact that user A follows user B does not necessarily mean B follows A. If A follows B, A can see the updated news (e.g., the act of posting a new pin) of B.

2.1.3 Bit.ly: A URL Shortening Service

URL shortening services assist in publishing and sharing content by providing a short equivalent URL that is redirected to an original URL [2]. A user (who wishes to share content) can submit an original content URL to a URL shortening service, and he/she can obtain a short URL as a concatenation

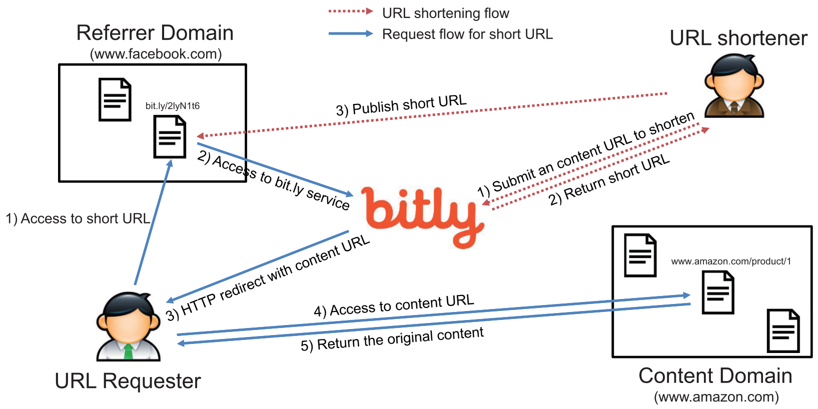


Figure 2.2 We illustrate how a URL shortener shortens an original content URL, and how a URL requester accesses the short URL.

form of the name of the service domain and the hash value, e.g., `bit.ly/2gXUgJI`. The user can then publish the obtained short URL to any place in which he/she wishes to publish, such as instant messages, e-mails, OSNs, and newsfeed services. Then, a person who wishes to access the content makes an HTTP request by clicking the corresponding short URL, the URL shortening service redirects the request to the original content URL. Figure 2.2 illustrates how a URL shortener shortens an original content URL, and how a URL requester accesses the short URL.

The main benefit of using a URL shortening service is that a user can publish a short, manageable, and human-unreadable URL to share content [2, 57, 37]. Hence, for example, microbloggers often use short URLs to share content in their microblogs which have length limits, e.g., 140-characters limit in Twitter [2]. Also, users who want to remove semantics from original URLs usually use short URLs for content sharing purposes. As a side effect, the short URLs are also used by spammers, attackers, or users who would like to hide original URLs [37].

`Bit.ly` is one of the most popular URL shortening services. It has received attention since 2009 when Twitter has used it as a default URL shortening service [37]. According to *New York Times*, people have created about 600 M URLs through `Bit.ly`, which have been requested over 8 B times [56]. `Bit.ly` also offers supporting functions for companies such as custom domain supports (e.g., nyti.ms for New York Times, pep.si for Pepsi) and analytics tools, which increases the popularity of the services for companies as well as for individuals.

2.2 Related Works

Online communication: Online communications through diverse channels (e.g., messengers, social media) have begun to dominate everyday social interactions, which has spurred studies on online communication behaviors across different platforms including online chatting [55], online forums [54, 48, 34], and OSNs [70]. Mayfield *et al.* investigated a way to disentangle the conversation threads from multi-part chatting [55]. With Yahoo!, USENET, and Twitter datasets, Kumar *et al.* investigated (i) the volume, depth, and degree of posts, and (ii) the number of users in each conversation thread, and proposed a conversation growth model based on the properties [48]. Marcoccia investigated conversation threads in USENET newsgroups [54], similar to subreddits in Reddit, and found that their sizes tend to be small and sometimes messages are misplaced. Gomez *et al.* explored discussion patterns on Slashdot, a technology-related news website where users can post and comment, and found that the degree distribution of conversations follows a log-normal distribution, and conversation threads show the strong heterogeneity and self-similarity [34]. Wang *et al.* proposed a model to predict the volume of conversations in `Digg.com`, and applied the model to other platforms such as Twitter and Reddit [70]. In this context, we focus on analyzing which factors (e.g., participant or content properties) are associated with the volume, responsiveness, and virality of a threaded conversation in Reddit, which can provide important implications on modeling, understanding, and predicting online conversation patterns.

Information cascades in OSNs: As OSNs have become one of the popular platforms to spread information such as news, photo, URL, or product, there has been a huge effort in studying information adoption and propagation in various OSNs [3, 4, 72, 11, 62, 17]. Bakshy *et al.* studied the role of social networks in information diffusion in Facebook, and showed that exposed users in the network are more likely to spread information [4]. Aral *et al.* identified influential individual and susceptible users in adopting the product in Facebook [3]. Rahman *et al.* [62] analyzed the adoptions and propagations of Facebook gifting applications, and showed that the evolutionary perspectives of cascades such as their initial growth rates are important factors for predicting the final population size of the application cascades. Choi *et al.* characterized online conversations in Reddit, and revealed how content properties and user participation behaviors are associated with successful conversation [17].

A few recent studies have shifted focus to micro-level dynamics of viral cascades [15, 31, 35, 24, 45]. The structural virality of cascades was measured based on the user dynamics information in Twitter [31]. For predicting image virality, some work used image features [24, 35, 45, 15], revealing a possibility to use content information in predicting viral diffusion. Deza and Parikh studied the viral image prediction from a computer vision perspective [24]. They evaluated several image features for predicting image virality. The most relevant work to this paper is that by Cheng *et al.* [15], which studied models for predicting whether a given cascade (with size k) *grows* beyond the median size of all the cascades with at least k reshares, which is a *growth prediction problem*. They showed that temporal and structural features are key predictors of the photo reshare cascade growth in Facebook [15]. While the work by Cheng *et al.* [15] provided an important theoretical insight into cascade prediction, this paper goes one step further from a practical and engineering standpoint; we focus on a *popularity or virality prediction problem* in Pinterest, based on the following feature sets which can be observed in different scenarios: (i) image features that can be obtained before posting, (ii) image meta and poster's information that can be obtained at the moment of posting, and (iii) initial propagation pattern. We explore which factors are strong predictors in predicting popular or viral image cascades, respectively.

Online Content Publishing and Sharing: People share various online content such as images, videos, or news through different Internet systems, e.g., online communities, OSNs, e-commerces, or social curating services. Online communities or news services are one of the popular places where users share news or new information by uploading external content or URLs [18]. Other users can write comments and exchange their thoughts to the uploaded content. Many researchers have investigated content sharing patterns in such online communities or news services [48, 54, 34]. Choi *et al.* analyzed posts and comments in Reddit, a popular online community, and characterized commenting patterns in terms of volume, responsiveness, and virality [18]. They explored how characteristics of content and user participation behavior are associated with commenting patterns in Reddit. Using Yahoo!, USENET, and Twitter datasets, Kumar *et al.* observed content propagation in terms of volume and depth, and proposed a propagation growth model based on the observations [48]. Gomez *et al.* explored content propagation patterns in Slashdot [34], a technology-related news website, and found that the degrees of comments follow a log-normal distribution. Wang *et al.* proposed a model to predict the volume of comments in Digg.com, a popular social news service, and applied the proposed model to different platforms such as Twitter and Reddit [70].

As OSNs have become one of the most popular places where various content types are shared, there have been many efforts in understanding and predicting content sharing patterns. Rodrigues *et al.* analyzed the word-of-mouth exchange of URLs among Twitter users and showed that URLs are likely to be shared among users who are geographically close together [63]. Bakshy *et al.* examined the patterns of information sharing in Facebook, and found that weak ties play a more important role in dissemination of content in Facebook [4]. Cheng *et al.* showed that temporal and structural features are key factors to predict the size of a photo cascade generated by resharing in Facebook [15]. Cha *et al.* analyzed propagation patterns of photo content in Flickr and showed that photos do not spread widely and quickly [11]. Goel *et al.* investigated the propagation patterns of URLs in Yahoo! and Twitter, and found that the majority of the diffusions occur within one hop from a seed node [33].

Recently, social curation services such as `Pinterest` have been reported as the vibrant places that encourage users to collect, organize, and share content by their tastes or interests [40, 27], which reveal distinct consumption patterns compared to other online services. Han *et al.* investigated content propagation patterns in `Pinterest` using the collected large-scale data, and showed that sharing pins in Pinterest is mostly driven by pin's properties like its topic, not by users' characteristics such as the number of followers [40]. This was confirmed by Gelley and John [27], who showed that 'following' is not significantly utilized in content sharing in `Pinterest`. Chang *et al.* investigated which categories are popular to male and female users in `Pinterest`, and showed that male and female users differ in collecting content across different topics. Han *et al.* showed that content creation and diffusion patterns are associated with users' different motivations and genders in `Pinterest` [39].

While these studies provide valuable insights into understanding content publishing and sharing patterns in online systems, we focus on how content is published and shared in a form of a *short URL*. In particular, we explore how the content pages with different categories (e.g., news, shop, adult content) are shortened, published, and shared across different online systems.

Reddit: Reddit has recently received a great attention as it becomes one of the most popular website that hosts a large number of online communities about almost all kinds of topics [66, 29]. Recently, many researchers have investigated user behaviors [43, 22, 6, 67], commenting patterns [73, 19, 21], and content popularity [66, 51, 29, 49] on Reddit. Singer *et al.* investigated the user preferences for topics shared among Reddit users and showed that the topics that users are interested in have diverged over 5 years [66]. The two case studies on "Hurricane Sandy" [51] and "duplicated image submissions" [49] showed some distinct factors that affect content popularity in Reddit. Gilbert showed popular images attract more attention and newly-uploaded images are under-provisioned in Reddit [29]. Weninger *et al.* analyzed top-scoring posts and their comments in Reddit, showing that comments that are closer in a comment tree are topically more similar than farther ones [73]. Choudhury *et al.* investigated posts and comments that contain self-discourse about mental health and found that posts with higher emotional intensity tend to receive more comments [19]. Danish *et al.* studied Q&A patterns in the `IAmA` subreddit, and showed that the posting user – who answers questions about themselves – is likely to answer the earlier and/or non-redundant questions [21]. In this paper, we characterize conversations in terms of volume, responsiveness, and virality, and explore how content, user, and topical communities are associated with the characteristics. In addition, we develop a learning-based model to predict large and viral conversations in Reddit.

Pinterest – an interest-driven OSN: Unlike other popular friendship-based OSNs such as Facebook, interests drive user activities or connectivities in Pinterest [40, 28]. Han *et al.* revealed that pin propagation in Pinterest is mostly driven by content properties like its topic, not by users' characteristics [40]. Gelley and John also showed that 'following' is not significantly associated with content sharing in Pinterest [28]. Zhong *et al.* proposed models to predict whether a user will be interested in repinning the given pin [76].

Han *et al.* [38] proposed a method to predict which topics an individual user will be interested in. Totti *et al.* evaluated the predictive power of different features on image popularity [68], and showed that visual properties have a lower predictive power than social cues. This paper proposes models for predicting popular or viral images based on two factors – *human social context* and *properties of content*, which can give an important insight into resource allocation for content providers and marketers in Pinterest-like interest-driven OSNs.

URL Shortening Services: The characteristics of short URLs motivate people to use URL shortening services and its variants with different goals, which in turn has spurred active research into usage patterns of URL shortening services [2, 57]. Demetris *et al.* investigated how short URLs are shared based on the information pages of short URLs, providing daily statistics of short URLs, and tweet/retweets including the URLs [2]. They provided *a macro-level view* of the short URL usage shared in Twitter, such as the daily click counts of tweets. On the other hand, we perform *a micro-level analysis* of the short URL usage including how content pages are created and published through short URLs, and how short URLs are shared through various types of domains (e.g., search engine, computer & electronics, not to mention OSNs), based on the detailed request logs for `Bit.ly` short URLs.

As short URLs themselves can hide their original content URLs, they are often used for sharing malicious content such as spams or phishing. Hence, many studies have focused on the potential risks of sharing short URLs. Using the dataset of `qc.rx`, a well-known URL shortening service, Klien and Strohmaier studied how short URLs are used for spamming from a geographical perspective [46]. Chhabra *et al.* investigated how phishing URLs are shared and propagated in online services based on the `Bit.ly` and `PhishTank` datasets, and showed that short URLs in Twitter tend to be more requested from more countries for a longer time than other services [16]. Wang *el al.* observed the spam short URLs published in Twitter, and developed a model for detecting spams [71]. Using the two-years large-scale dataset from several URL shorting services, Maggi *et al.* analyzed how many users are exposed to malicious short URLs, and found that the threats of using short URLs are not as serious as those of using long URLs [53]. Nikiforakis [57] reported a high portion of short URLs created from ad-based URL shortening services are likely to be used for infecting users with malware and exfiltrating private data. On the contrary, we comprehensively analyze (i) what content types (e.g., news, adult, or sports) are shortened and shared, (ii) how and where short URLs are published, and (iii) how content pages are shared through different publishing spaces.

Chapter 3

Methodology

3.1 Data Collection

3.1.1 Reddit

We first analyze the patterns of posting/commenting activities in Reddit and then derive user interactions from the activities. To retrieve posts and associated comments, we developed our measurement system for data collection

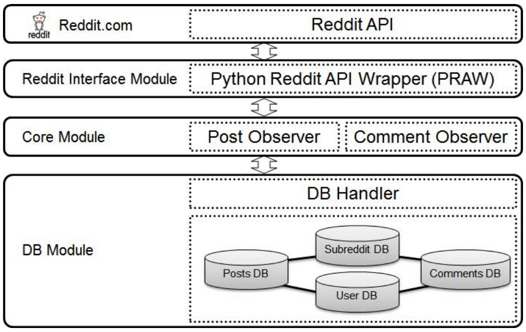


Figure 3.1 The architecture of the Reddit measurement system is depicted.

and analysis as shown in Figure 3.1. The measurement system consists of three parts: (i) Reddit interface module, (ii) core module, and (iii) DB module. The Reddit interface module communicates with `Reddit.com` through the APIs[1] provided by Reddit. We utilize the 'Python Reddit API Wrapper (PRAW)[2]' package.

To monitor posts and their follow-up comments, we developed two key submodules in the core module: the post observer and comment observer. Every minute, the post observer monitors and fetches all new posts in each subreddit. At the time of our data collection, Reddit APIs provided up to $1,000$ recent posts in each subreddit in the chronological order; hence our crawler fetches up to $1,000$ posts every minute, which was enough to monitor all new posts. Whenever the post observer identifies a new post, the comment observer begins to monitor the comments relevant to the post. Similarly, the comment observer monitors and collect every comment associated with the posts that we have fetched. We collected every single post and comment during our measurement period. The observed maximum number of messages per minute was 722, which did not exceed the rate limit of the Reddit API. The collected dataset is stored in the DB module.

Our measurement focuses on the top 100 subreddits in terms of the number of subscribers, which are responsible for a large portion of Reddic conversations. Note that the top 100 subreddits account for more than 60% of all subscribers (out of 378,293 subreddits, as of Oct. 22, 2014) in Reddit. We collected the dataset for 35 days from March 13 to April 18, 2014, which con-

[1] Reddit provides public APIs, through which third-party applications such as crawlers and readers are supported.

[2] http://praw.readthedocs.org/en/v2.1.16/

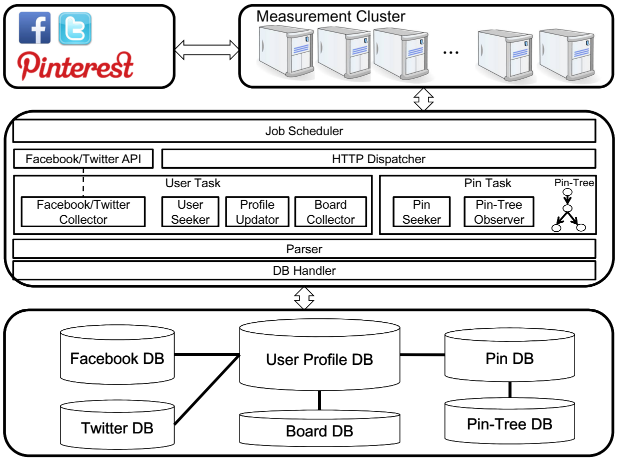


Figure 3.2 The architecture of our Pinterest crawling and analysis system is described.

tains 1,016,342 posts and 18,626,530 comments, shared by 1,531,247 users. We then extracted 695,857 (68.5%) posts that each have at least one comment, and their 18,093,422 comments; posts and comments are written by 1,455,293 users. Each post contains the `author id`, `title`, `subreddit id`, and `timestamp`, while each comment contains the `original post id`, `user id`, `comment text`, and `parent id` from which the comment is generated.

3.1.2 Pinterest

Since Pinterest does not provide an official API for data collection, we developed our measurement system by crawling Pinterest pages as shown in Figure 3.2. We fetch web pages in Pinterest, from which the relevant information is extracted; the data about each pin or board[3] can be extracted from a web page. This is challenging since we need to crawl a large number of web pages from Pinterest. For example, if a user has 1,000 boards, we need to

[3] In this paper, a pinboard and a board are used interchangeably.

make 1,000 HTTP requests to collect the data about her board. To address this problem, we designed a distributed crawling system. Our measurement cluster consists of 25 PCs, which continuously send HTTP requests assigned by the job scheduler. The HTTP dispatcher processes the HTTP requests and responses according to the tasks explained below.

There are two main tasks in our system: *pin task* and *user task*. Unlike prior measurement studies (e.g., [30, 59]), we focus on pin propagation patterns. To this end, we periodically (every five minutes) monitor all the newly-posted pins in the menu of each category (e.g., animal, kids, women's fashion). Since Pinterest shows all the recent activities including posting a pin, repinning, and leaving a comment in the menu of each category in the chronological order, our pin seeker fetches 10 recent web pages not to miss newly-posted pins. The pin-tree observer keeps track of each pin and its associated repins to build a pin propagation tree, which is called a *pin-tree*. If a user repins the original pin, Pinterest provides a link to the board that includes the repinned one; we can find and fetch the associated web page of the repinned one among other pin pages in the board, so that we can keep track of the chain of the pin-tree. The collected information of each pin-tree is stored in the pin-tree database. The pin (and repin) dataset consists of the number of likes, number of comments, its category, its source, and its description, which is stored in the pin database.

In the user task, we collect the information (e.g., number of pins, number of followers, number of boards, gender, country, etc.) of each user. In addition to the 1 M users found in pin-trees, the user seeker additionally finds 2 M users using a breadth first search (BFS) in Pinterest. For the discovered

3 M users, we collect the information of each user, including her name, her description, gender, number of followers, number of followings, number of boards, number of pins, number of likes, her external website, location, and Facebook/Twitter links, which are stored in the user profile database. Along with the user profiles, the board collector collects the information of each board including its category, and number of pins, which are stored in the board database. To identify the gender and country of users, we use external links to Facebook and Twitter, which can be found in the profile pages of users. The Facebook/Twitter collector sends queries to Facebook and Twitter through their APIs and fetches the gender and country information of each user if available. We finally decide the gender and country of each user by collectively combining information from Pinterest, Facebook, and Twitter.

3.1.3 Bitly

To investigate the practice of using URLs shortened by `Bit.ly`, we perform a measurement study using a large-scale dataset from `Bit.ly`. Our dataset consists of two parts – (i) short URL data and (ii) request data for the short URL. The short URL data includes the content (or original) URL a user shortens, the global hash of the target URL, a user id, and its creation time. Each request log consists of a global hash of short URL, its original URL, its referrer URL where the short URL is published, and the temporal, geographical request information such as request time, country, city, and timezone. Note that only anonymized user data is used for this research, and no personally identifiable information is used.

To characterize the URL properties, we additionally investigated the category of each of content and referrer domains. To this end, we first extracted the domain name of a content (or referrer) URL by removing all characters after the first delimiter '/'. For example, the domain name of a content URL `'www.facebook.com/video/abc.mp4'` is `'www.facebook.com'`. We then submitted the domain name of the content to a commercial URL scanner, VirusTotal[4], which scans a submitted URL over a corpus of five website scanning engines and returns the category for the given URL. Note that the returned category name is usually different across the five engines, and the categorization is often not consistent even within a single engine. Also, some engines even require users to manually label categories. To address this problem, we perform a semi-manual categorization. That is, we made the standardized set of categories, each of which is provided by SimilarWeb[5]. Note that we mostly used the second-level categories in SimilarWeb. If there are only 1st level categories, we used them as they are. In addition, we added 'Violence & Illegal', 'Blogs', and 'Streaming' categories to the standardized category set, which are provided by VirusTotal engines but not by SimilarWeb. Finally, we have total 64 categories in our standardized category set.

Our dataset contains more than 80 M short URLs and their 4.2 B requests generated from more than 2.1 B devices and more than 220 countries during one month, June 2012. The top 3 countries by the number of requests are USA, China, and Japan. Considering the report that the portion of Internet users in these countries are 10.2%, 22.4%, and 4.2%, respectively [25], the result indicates that the short URLs are more heavily requested from the USA. Note that the numbers of content and referrer domains are 3.1 M and

[4] https://www.virustotal.com/

[5] https://www.similarweb.com/category

2.2 M, respectively.

3.2 Models

3.2.1 Comment Tree: Threaded Conversation Model

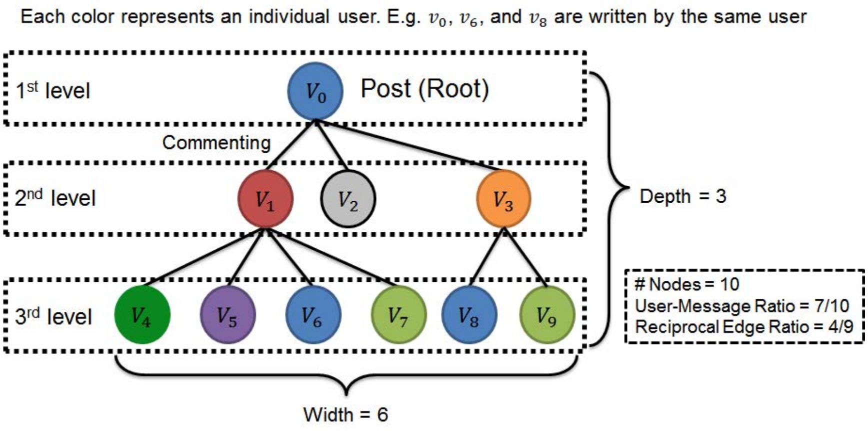

Each color represents an individual user. E.g. v_0, v_6, and v_8 are written by the same user
1st level
V_0
Post (Root)
Commenting
2nd level
V_1
V_2
V_3
Depth = 3
3rd level
V_4
V_5
V_6
V_7
V_8
V_9
# Nodes = 10
User-Message Ratio = 7/10
Reciprocal Edge Ratio = 4/9
Width = 6

Figure 3.3 A comment tree is illustrated for a post that has 9 comments.

To model a conversation thread from a given post and its follow-up comments, we define a *comment tree* as an undirected rooted tree, $T = (V, E)$, where V is the set of all messages, which includes the original post (root) and all the follow-up comments in the thread, and E is the set of edges, each of which connects two messages that are linked by commenting. Figure 3.3 illustrates a comment tree that consists of a post and nine comments.

We characterize a comment tree T based on the following three metrics:

- **Volume** (N_T): The volume of tree T is the number of nodes, $|V|$, in the tree. For instance, N_T of the tree in Figure 3.3 is 10.

- **Responsiveness** (R_T): To capture how quickly users participate in (or respond to) a conversation, we first calculate the time differences between a comment and its parent node (the post or comment). We only consider the time differences within the range of $[\mu - 2\sigma, \mu + 2\sigma]$ to exclude outliers, where the μ is the average time difference of parent-child edges of the given tree. We then calculate the average of the inverse of (comment) inter-arrival time differences (in minutes), *responsiveness* R_T, which quantifies how fast comments are written to a given post (or its tree). Hence, the higher responsiveness a tree has, the faster users add comments to the tree.

- **Virality (V_T)**: The *(structural) virality* of a cascade, also known as Wiener Index (WI) [32, 15, 14, 41], seeks to quantify the degree of multi-generativity of the conversations. That is, given the same number of nodes, the WI becomes the minimum when all comments are directly added to the root, and the maximum when the tree becomes a chain (the depth of a tree is the number of nodes in the tree). The former indicates that no subsequent spreading has occurred except at the first generation and the latter indicates that every comment (except the last one) is followed by another comment, as shown in Figure 3.4 (the leftmost and rightmost ones, respectively). Formally, the WI of a tree is defined by the average hop count over all node pairs in the tree. WIs are calculated for the four 10-node comment trees in Figure 3.4 for illustration purposes.

Figure 3.5 shows the distributions of the volume, responsiveness, and virality values of the comment trees. The volume distribution exhibits a heavy tail that spans several orders of magnitudes. For instance, as shown in Figure 3.5(a), while 72.8% of trees consist of less than 10 nodes, top 0.1% of the trees attract more than $2,211$ messages, indicating a large deviation among

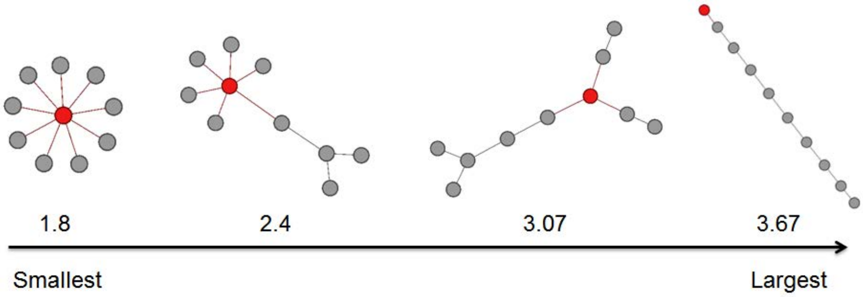
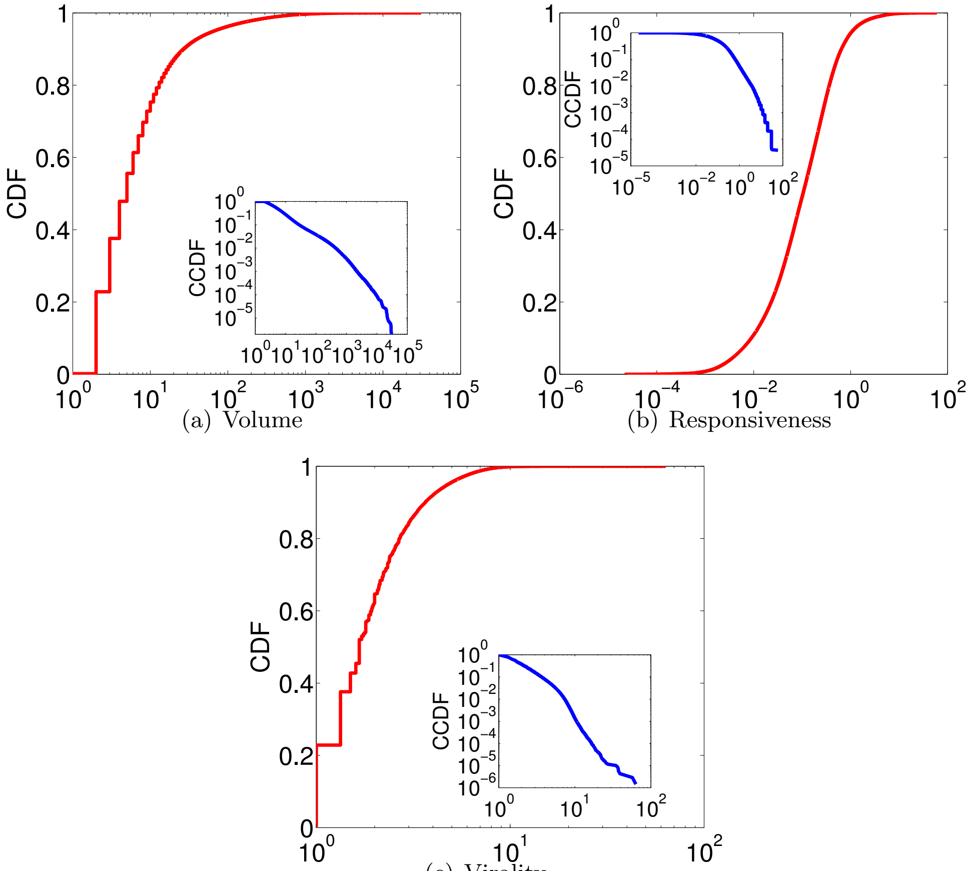


Figure 3.4 Virality values (i.e., WIs) are calculated for 10-node comment trees ($N_T = 10$).

threads. The virality distribution also exhibits a heavy-tailed distribution although the range of virality values only spans two orders of magnitudes. As shown in Figure 3.5(c), around 99.8% of the virality values are smaller than 10, and top 0.1% of the virality values greater than 50 (the maximum is 63.44). The average of the virality values of the comment trees is 2.09, which implies a comment in a tree is likely to span around 2 levels on average. On the other hand, the responsiveness distribution follows a Gaussian-like distribution. The average of responsiveness values is 0.32, whose unit is the inverse of a minute. In addition, the responsiveness values of the top 5% of trees are larger than one (i.e., more than one comment every minute), meaning that those trees are highly responsive, while the comments of the bottom 15% of trees are generated once per hour on average.

3.2.2 Pin Tree:Image Cascade Model

We first model an *image cascade* as an undirected tree, $T = (U, R)$, where U is the set of users including a pinner and follow-up repinners for *a given image* (or a pin) posted by the pinner, and R is the set of repinnig activities, i.e., pin propagation. We characterize the image cascade T based on the following



(a) Volume

(b) Responsiveness

(c) Virality

Figure 3.5 Distributions of volume, responsiveness, and virality of comment trees are plotted for all the comment trees.

two metrics:

- **Volume** ($|U|$) of cascade T is the number of nodes in the tree. For example, $|U|$ of the cascade in Figure 3.6 is 11.

- **Structural virality (or WI)** of cascade T represents the average range of a node's effect in an image cascade. To quantify the structural virality, we adopt a well-known metric 'Wiener Index (WI)' [75, 31, 15, 17], which is defined as the average hop count between all pairs of

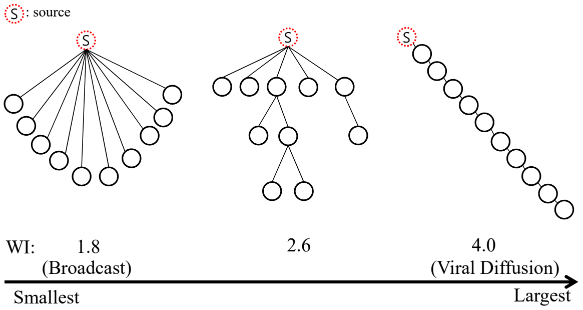


Figure 3.6 We illustrate different types of 11-node diffusions from a simple broadcast (on the far left) to a viral diffusion through a chain (on the far right). The bottom axis shows the Wiener Index (WI) values calculated for the 11-node cascades, $V_T = 11$.

nodes in a tree T. More specifically, WI of cascade T can be calculated as follows:

$$WI = \frac{1}{|U|(|U|-1)} \sum_{i,j \in U, i \neq j} d_{ij} \tag{3.1}$$

where U is the set of users in T, and d_{ij} is the distance (or hop count) of the shortest path between users i and j in T. Figure 3.6 shows the three 11-node trees with their WIs. As shown in Figure 3.6, given the same number of nodes, i.e., an image reaches to 10 audiences through different propagation scenarios, the WI becomes the minimum if all the repinners directly get the image from the pinner (i.e., the leftmost scenario in Figure 3.6), and the maximum if T becomes a chain (i.e., the rightmost scenario in Figure 3.6).

3.2.3 Content-Referrer Graph Model

To explore how short URLs are published through domains, we model the relations among content and referrer domains as a *Content-Referrer graph*, a directed graph $G = (V, E, W)$, where V is the set of all domains, including content and referrer domains, and E is the set of edges. Each edge connects from a content domain to a referrer domain, where a short URL of a content URL is published. Note that any domain can be a content domain, a referrer domain, or both. The weight of an edge is the number of content URLs published in the referrer domain. Here, we consider only the content URLs requested at least once. Figure 3.7 illustrates a relation between a referrer domain (Twitter) and a content domain (Facebook), which is modeled as the directed edge between the two nodes in the content-referrer graph. Note that the content-referrer graph is a forest that consists of multiple distinct components across which there is no reachable path.

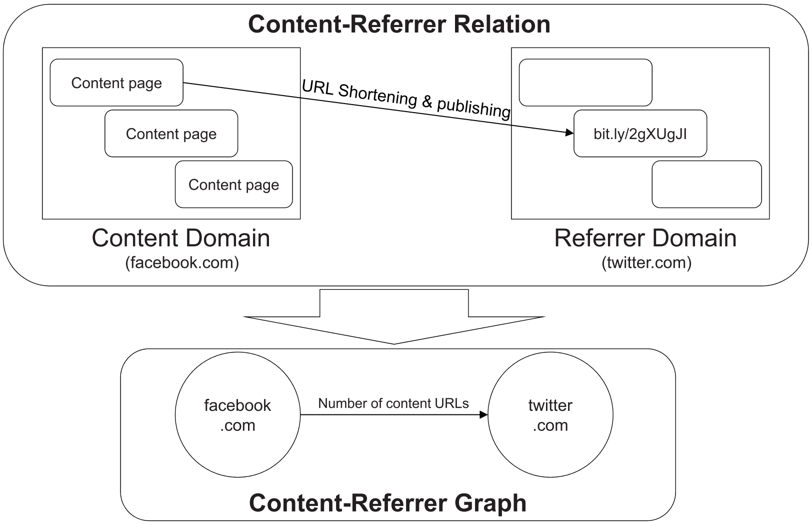


Figure 3.7 We model the relations among content and referrer domains in the form of a *content-referrer* graph. If a tweet has a short URL for a content page in Facebook, there is a directed edge from Facebook (content domain) to Twitter (referrer domain). The weight (of an edge) is the number of content URLs published in a referrer domain.

We finally build a content-referrer graph based on more than 3 M content domains and 2 M referrer domains. The number of nodes, edges, and components are about 4.3 M, 12 M, and 48 K, respectively. Note that requests from non-websites (e.g., Instant Message and Apps) are labeled as ‘direct’ in Bit.ly, and are removed in constructing the content-referrer graph.

Chapter 4

Analysis on Online Conversations in Reddit

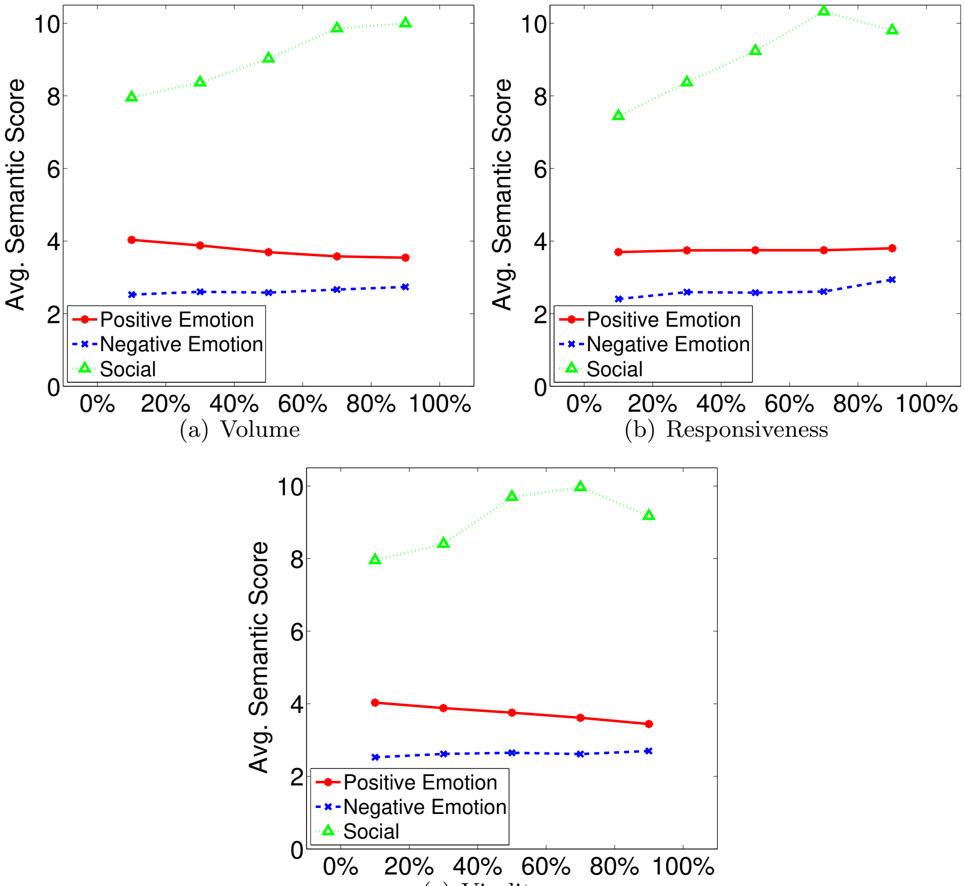
4.1 Comment Tree Analysis

In this section, we analyze the conversations (i.e., comment trees) in terms of content and user participation properties. To this end, we first divide comment trees into five intervals in terms of volume, responsiveness, and virality, respectively, and then explore the characteristics of the comment trees in each interval. Note that we perform one-way ANOVA tests for our analyses, and verify that all the p-values are smaller than 0.05.

4.1.1 Content Perspectives

We first perform the text analysis for every comment tree by measuring its sentiment and other properties to characterize the content of the trees. We then investigate how these characteristics are relevant to the volume, responsiveness, and virality of the comment trees.

Revealed Sentiment



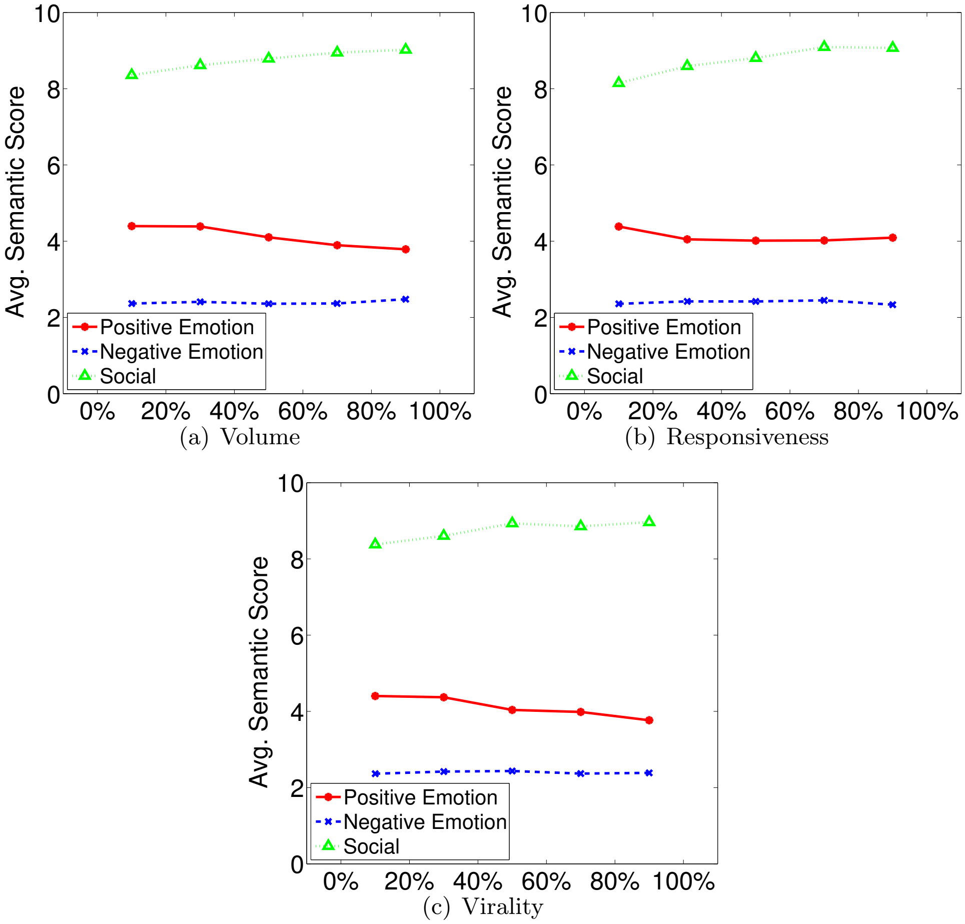
(a) Volume

(b) Responsiveness

(c) Virality

Figure 4.1 The distributions of emotional scores of posts are plotted.

We first perform a sentiment analysis by using LIWC (Linguistic Inquiry and Word Count), which is text analysis software that counts words that belong to psychologically meaningful categories. For a given text, the LIWC tool provides various sentimental scores, each of which is calculated as the relative frequency of the words in the given sentiment category on a percentile scale, out of all the words in the text. We use three categories: social, positive and negative emotions. For example, the words “family” and “friends” belong to the social category, and “love” and “sweet” are in the positive emotion category, while “hurt” and “nasty” belong to the negative emotion category. Note that we compute the LIWC scores for (i) titles of posts (since there are some posts containing only multimedia content without any text), and (ii) all the texts written in comments.



(a) Volume

(b) Responsiveness

(c) Virality

Figure 4.2 Sentiment scores of texts in conversation trees are plotted.

Figures 4.1 and 4.2 indicate that social words are more frequently used than words of positive emotions, which in turn are more frequently used than words of negative emotions. We notice that this trend is also in line with the sentiment analysis on blogs, emotional writing, and talking [60]. There are no significant differences in the two emotional scores as the volume, responsiveness, and virality increase. On the other hand, the higher social scores the titles of posts have, the larger, more responsive, and more viral trees they tend to become. This implies that a post whose title contains more social words is likely to generate a large, responsive, and viral tree to a certain degree.

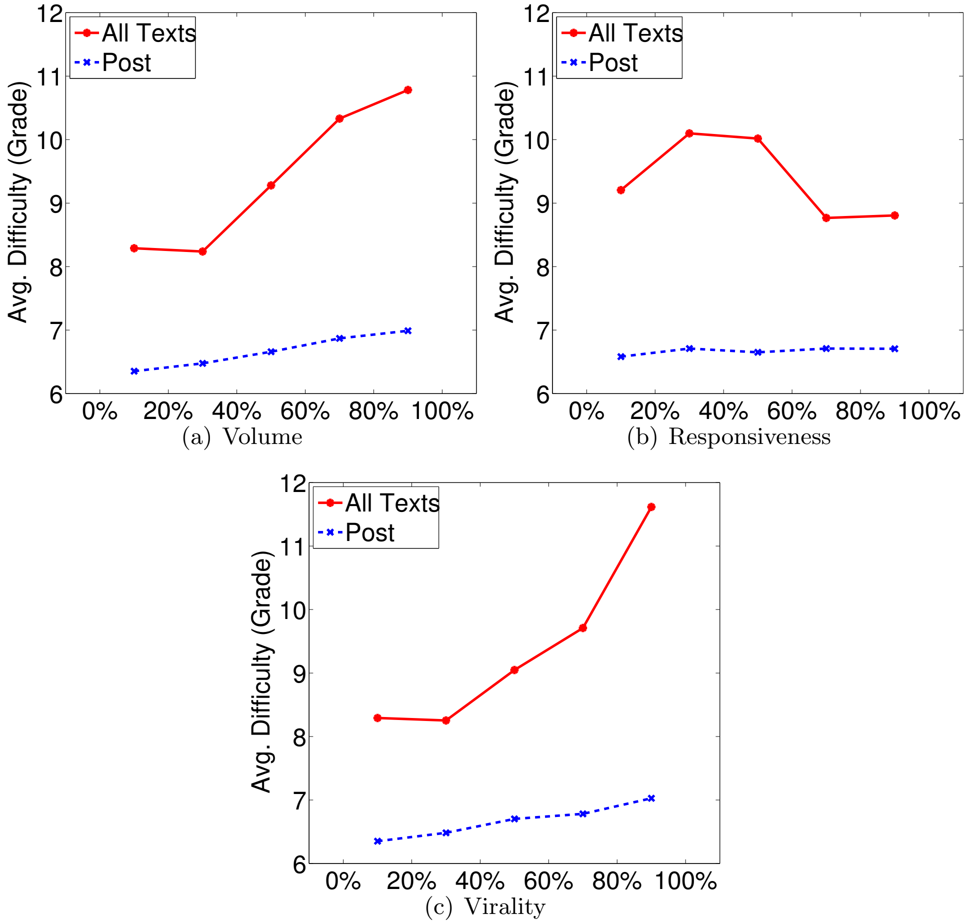
Document Difficulty

We next examine whether the (readability) difficulties of titles and texts of trees are relevant to their volume, responsiveness, and virality. To this end, we compute *Gunning-Fog Index*, a popular readability score to estimate what grade of students is suitable to read the text [36]. That is, if the index of a text is 12, the text requires the 12th-grade ability (around 18 years old). The Gunning-Fog index of a comment tree T is defined by:

$$G_T = 0.4[(\frac{N^T_{words}}{N^T_{sentences}}) + 100(\frac{N^T_{complex}}{N^T_{words}})] \tag{4.1}$$

where N^T_{words}, $N^T_{sentences}$, and $N^T_{complex}$, are the numbers of words, sentences and complex words in texts, respectively. A complex word is defined as the word that contains three or more syllables excluding proper nouns, familiar jargon, compound words, and words with common suffixes such as -es, -ed. Similarly to the sentiment analysis, we calculate the difficulties of comment trees for (i) the title of a post and (ii) all texts of a comment tree (including its title).

Figure 4.3 shows that the average difficulty of the texts of a tree ranges mostly from 8 to 12, and is generally larger than that of its title (around 6 to



(a) Volume

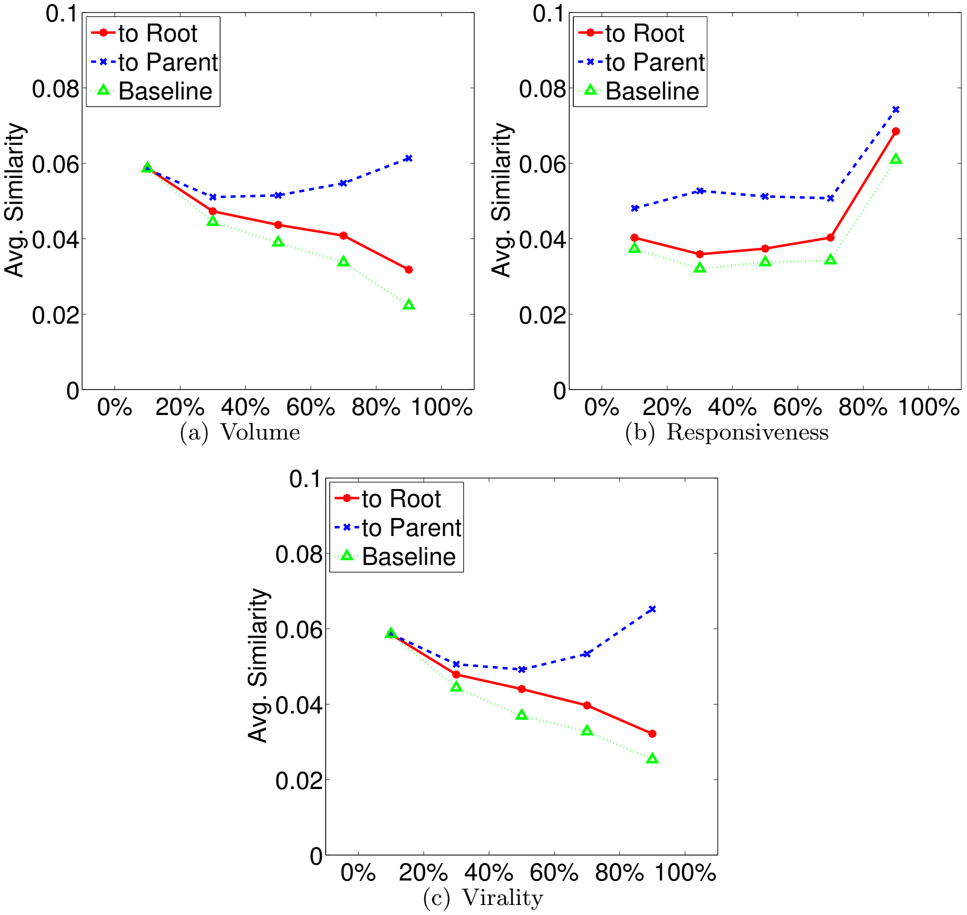
(b) Responsiveness

(c) Virality

Figure 4.3 The average difficulties of trees and posts are plotted.

7), probably because titles are usually short and consist of a few keywords. Interestingly, as the difficulties of both the titles of posts and the texts of comment trees increase, their volumes increase significantly, and more rapidly as the virality increases. This implies that larger and more viral trees tend to contain comments with more difficult words on average. On the other hand, the difficulties of the texts of the top 40% responsive trees are lower than those of the less responsive trees, which implies using easier words is positively related to the quick responsiveness.

Document Similarity



(a) Volume

(b) Responsiveness

(c) Virality

Figure 4.4 The similarity among messages of the same comment tree is plotted.

We finally investigate whether the similarity between two messages is relevant to volume, responsiveness, and virality. To this end, we compute the message similarity in two cases: (1) between a post (or root) and its child comments, and (2) between a parent and its child comments, by using the Term Frequency-Inverse Document Frequency (TF-IDF) similarity, one of the popular metrics to measure the similarity between two documents in information retrieval [1]. For each word, its TF-IDF is defined as the product of TF and IDF, each of which quantifies how frequently the word is used in a document, and whether the word is common or rare between two documents, respectively. Thus, a TF-IDF similarity score (of a given word) is high (i) if the word is used in the document frequently and/or (ii) if the word is rarely used in the two documents, and vice versa. Before calculating the TF-IDF similarity, we remove stop-words (e.g., at, which), and perform Porter stemming by using *Natural Language Toolkit*[1]. After measuring the TF-IDF score for each word, we then compute the cosine similarity of two score vectors between two documents (or messages). (The vector dimension is the number of distinct words in the two documents.) The cosine similarity being 1 means the two documents are almost identical, while 0 indicates no words are shared.

Figure 4.4 shows the averages of document similarity. As a reference, we measure the cosine similarity between any pair of messages in a tree (even if there is no parent-child link), labeled as baseline. As shown in Figure 4.4, the average document similarity in the first case decreases as the volume and virality increase, while the one in the second case increases. This result reveals that topics may somewhat digress in large and viral conversations although the parent-child comments become increasingly similar as the volume and virality grow from their medians. Furthermore, highly responsive trees exhibit high similarity in both cases, which implies quickly-generated comments are more similar to their parent messages.

[1] We use a python package as its implementation.

4.1.2 User Participation in Comment Trees

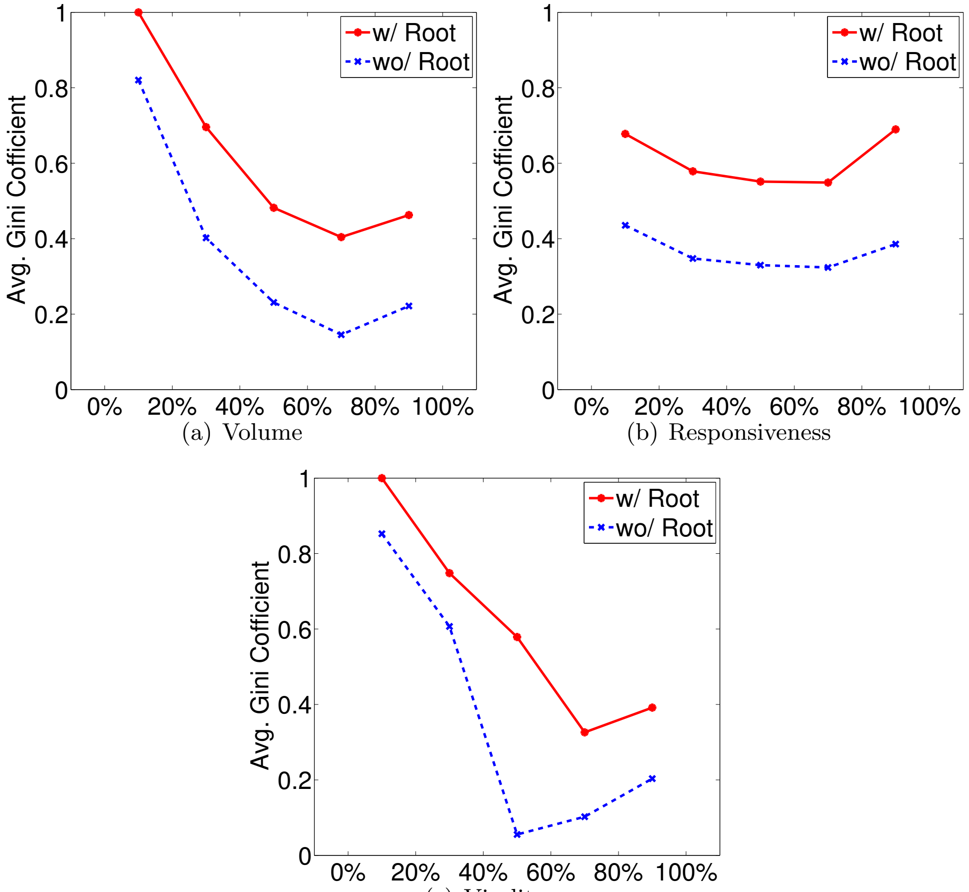
We seek to understand how (user) participation behaviors are associated with volume, responsiveness, and virality of comment trees. To quantify participation behaviors, we compute Gini coefficients, user-message ratio, and reciprocal edge ratio across the five intervals. We then investigate how different roles of users are related to generating large, responsive, and viral comment trees.

Participation Behaviors of Users

We first quantify the heterogeneity of the nodes in a tree by computing the *Gini Coefficient*, a metric that is most commonly used to capture inequality of income distribution in Economics [20]. The Gini coefficient, represented in the range of [0, 1], increases as the distribution of incomes is increasingly skewed. In our case, the coefficient becomes 0 if every node in a tree has the same number of child nodes, and the coefficient is 1 if only the post has all the child nodes. Note that we calculate two kinds of Gini coefficients for a tree: with or without a root (i.e., a post).

Figure 4.5 shows the average Gini coefficients for each interval in terms of volume, responsiveness, and virality. Overall, the Gini coefficient with roots is higher than the one without roots, which implies users are more likely to reply to posts in general. For both cases, the coefficient sharply decreases as the volume and virality increase, except for the rightmost interval. This indicates that comments in large and viral trees uniformly attract other comments to a certain degree, but extremely large and viral trees have comments that elicit much more follow-up comments than others.

On the other hand, as the responsiveness increases, the Gini coefficients of trees with and without roots do not decrease as much as in the case of volume and virality, showing more symmetric convex patterns. Note that moderately viral trees show low Gini coefficients, which means that messages with the relatively uniform distribution of follow-up comments take somewhat longer inter-message time.



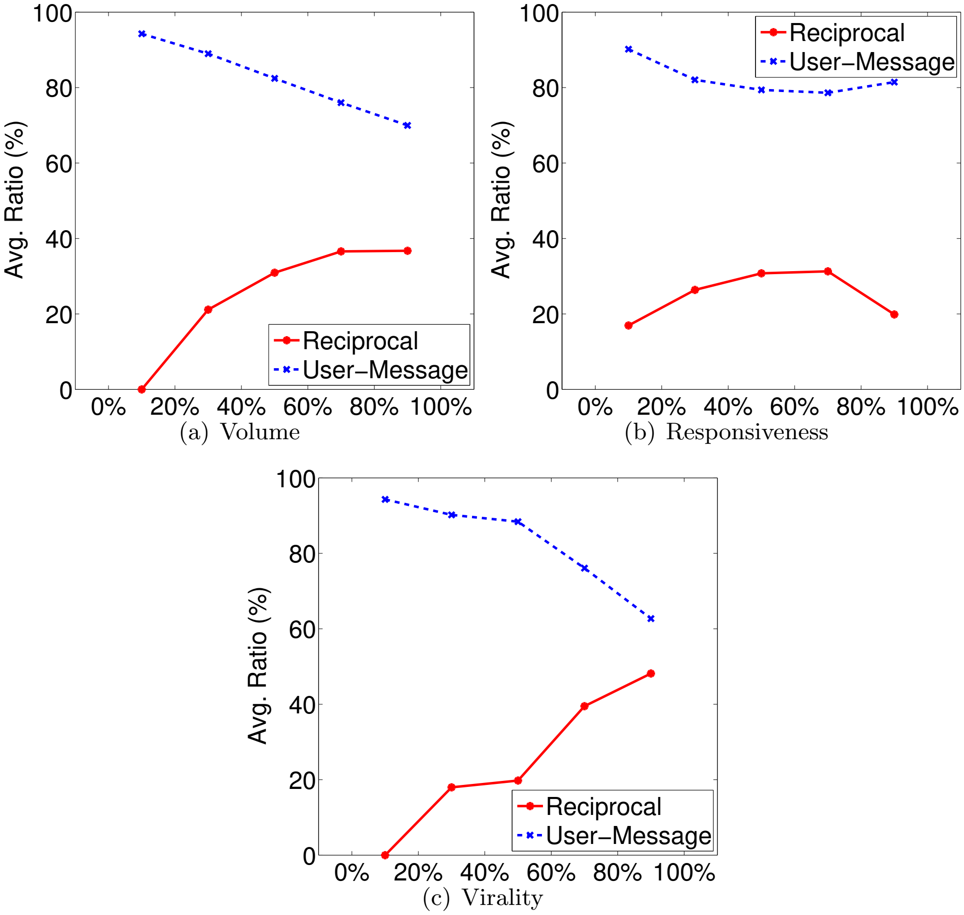
(a) Volume

(b) Responsiveness

(c) Virality

Figure 4.5 Average of Gini coefficients of comment trees are plotted.

We next investigate how many users are likely to make comments and how reciprocally users communicate in a tree by computing the user-message ratio and reciprocal edge ratio, respectively. The user-message ratio for a tree T is defined as the ratio of the number of users participating in T to the volume of T. If every user in a tree submits only one message, its user-message ratio is 1, meaning that every participating user generates exactly one message for the tree. The reciprocal edge ratio is the ratio of the number of edges generated by reciprocal user pairs (i.e., they exchange comments) to the number of all the edges in the given tree.



(a) Volume

(b) Responsiveness

(c) Virality

Figure 4.6 Reciprocal edge ratio and User-Message Ratio are plotted.

Figure 4.6 shows the reciprocal edge ratio and user-message ratio as the volume, responsiveness, and virality increase. Interestingly, the user-message ratio drops in larger and more viral trees, whereas the reciprocal edge ratio increases. This result implies that comments of a large and viral tree are usually generated by a small portion of users who reciprocally communicate to one another. Note that the tendency is more noticeable as the virality increases, which means that extremely viral trees tend to result from intensively reciprocal communications.

Figure 4.6(b) reveals that the top 20% and bottom 20% responsive trees have the smaller reciprocal edge ratio and the higher user-message ratio than the trees in other intervals, respectively. This result is in line with Figure 4.5 in the sense that the portion of reciprocal communications in a comment tree is low since users are more likely to respond to a post in moderately responsive trees.

Roles of Users

To investigate users' special roles in large, responsive, and viral comment trees, we first identify users based on behavioral types as follows:

- U_{post} are the top 1% of users measured by the number of uploaded posts. They can be considered as active initiators since they initiate conversations by writing many posts.

- U_{cmt} are the top 1% users in terms of the number of comments. They participate in conversations by actively commenting to other messages.

- U_{rcvcmt} are the top 1% users identified by the number of received comments from others. These users attract many comments from others, and may play a major role in developing large, responsive, or viral conversations.

- U_{uni} are the users who participate in a number of subreddits. These users can be considered as *translators* [8] or *generalists*, who are translating or cross-pollinating content/ideas across multiple communities. To identify such translators, we count the number of messages (i.e., posts and comments) a user has submitted for each subreddit, and then calculate the *subreddit entropy* for each user u, as follows:

$$H_u = -\sum_{m=1}^{N_{sub}^u} p_m^u \log p_m^u \tag{4.2}$$

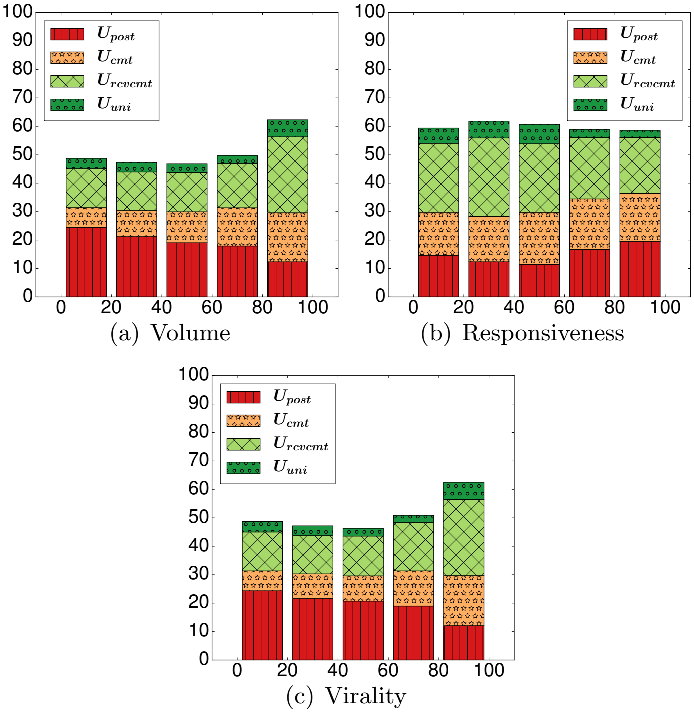
where N_{sub}^u is the number of subreddits where the user u uploaded messages and p_m^u is the fraction of u's messages in the m^{th} subreddit. We then choose the top 1% of users based on the subreddit entropy, called U_{uni}. Since we do not normalize the subreddit entropy by the number of subreddits, the identified translators tend to be those who participate in many subreddits.

Note that the identified users can have multiple role types, and user types can be correlated in principle.

	U_{post}	U_{cmt}	U_{rcvcmt}	U_{uni}
U_{post}	1.0	0.14	0.29	0.07
U_{cmt}	0.14	1.0	0.53	0.18
U_{rcvcmt}	0.29	**0.53**	1.0	0.12
U_{uni}	0.07	0.18	0.12	1.0

Table 4.1 Conditional probabilities among role types are described.

We first measure how identified role types are overlapped by calculating the conditional probabilities of each pair of role types in Table 4.1. As shown in Table 4.1, 14% of users in U_{post} and U_{cmt} are overlapped, indicating that a small portion of users plays important roles both in posting and commenting on Reddit. Note that the probability $p(U_{cmt}|U_{rcvcmt})$ is larger than 0.5, meaning that the users who comment more also tend to receive more comments, probably as a result of their active commenting behaviors. Interestingly, the probability of U_{uni} and other role types are mostly low, which implies that users who are interested in multiple topics are distinct from the users in other activity-related roles.



(a) Volume

(b) Responsiveness

(c) Virality

Figure 4.7 Contribution ratios of four user role types are plotted.

We investigate how each role type contributes to large, responsive, and viral conversations, respectively. Figure 4.7 shows the portions of comments received by the users in each role type. As shown in Figure 4.7, around 50%

of comments are elicited by the four role types, and this portion increases up to about 60% in the top 20% of large and viral conversations. The portion of comments elicited by U_{post} decreases as the conversations become larger and more viral, whereas the ones elicited by others increase, which indicates the users in U_{post} play diminishing roles in large and viral conversations. Interestingly, U_{uni} attract more comments in the top 20% intervals in terms of both volume and virality, implying that translators who have broad interests are likely to attract more comments in a large or viral conversation. The responsive conversations show distinct patterns; the portion of comments elicited by U_{post} increases as the conversations become more responsive, meaning that heavy-posting users play more roles in attracting others' comments in responsive conversations where many of comments are just quick responses to the post content.

4.2 Conversation Patterns across Communities

In this section, we compare subreddits based on conversation patterns captured in volume, responsiveness, and virality of comment trees. We also extract the top 10 subreddits in terms of each of the three criteria and further analyze the content properties and users' participation behaviors. We then investigate the characteristics of subreddit groups that show similar volume, responsiveness, and virality.

4.2.1 Conversations in Subreddits

We investigate how conversations across communities (or subreddits) show different patterns in terms of the volume, responsiveness, and virality. To this end, we first calculate the averages of the three quantities in each subreddit,

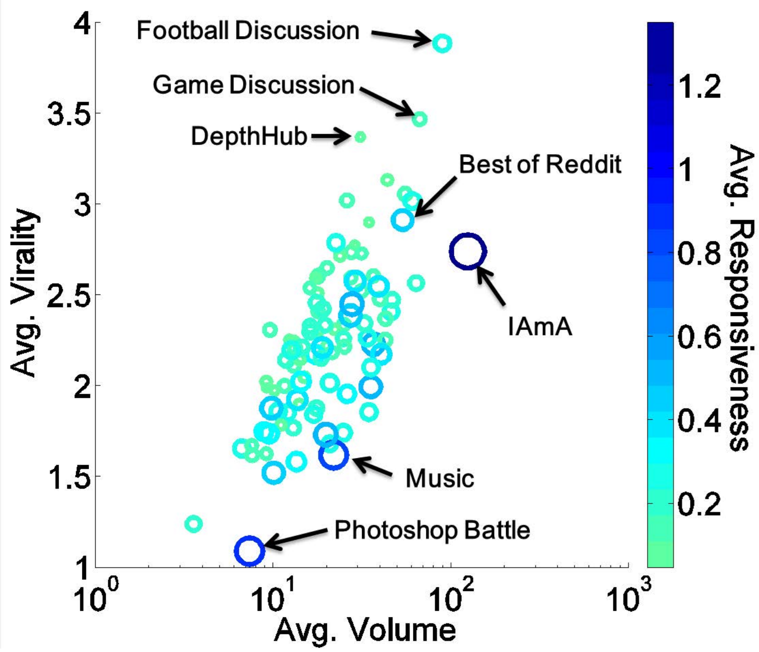


Figure 4.8 We map subreddits by calculating the average values of their trees in terms of the volume, responsiveness, and virality.

and then plot *a subreddit map* in Figure 4.8. The position of a subreddit corresponds to the average volume and virality, while the color and diameter of its circle represent mean responsiveness. For instance, the conversations in subreddit `IAmA` tend to have the highest volume (i.e., 100) and highest responsiveness (i.e., 1.4), but their virality lies in the middle (i.e., 2.7) among subreddits.

Figure 4.8 shows that the average volume and virality values exhibit a strong correlation in general, while there are some outliers. For instance, the conversations in `Music` or `IAmA` show large volumes but their viralities tend to be low, while the conversations in `DepthHub` tend to be viral but their volumes are relatively small. Some subreddits (e.g., `Photoshop Battle` or

Music) show interesting patterns; while their conversations show small volume and low virality, their responsiveness is relatively high, meaning that participants of the conversations in those subreddits are likely to be responsive.

Rank	Volume (S_{vol})	Responsiveness (S_{rsp})	Virality (S_{vrl})
1	IAmA	IAmA	Football Discussion
2	Football Discussion	Photoshop Battle	Game Discussion
3	Game Discussion	Music	DepthHub
4	Technology	Reddit Gold Mine	Android
5	Soccer	Mystery of the soda	You Should Know
6	You Should Know	AskReddit	The Dismal Science
7	Best of Reddit	Science	Soccer
8	World News	Game of Thrones	Best of Reddit
9	TIL	FoodPorn	Frugal Living
10	Android	EarthPorn	Game Deals

Table 4.2 Top 10 subreddits in terms of volume, responsiveness, and virality are listed.

To further analyze conversation patterns across subreddits in detail, we select the top 10 subreddits ranked by the volume, responsiveness, and virality, respectively (See Table 4.2). We refer to the three lists for the volume, responsiveness, and virality as S_{vol}, S_{rsp}, and S_{vrl}, respectively. As shown in Table 4.2, the three lists, S_{vol}, S_{rsp}, and S_{vrl}, are substantially different. In particular, the 9 subreddits in the S_{rsp} exist in neither S_{vol} nor S_{vrl}, which again indicates that the responsiveness is not correlated to volume and virality of conversations.

The two lists, S_{vol} and S_{vrl}, are relatively similar; they share six subreddits. The common subreddits between S_{vol} and S_{vrl} are mostly discussion-driven subreddits such as Football Discussion, Game Discussion, or Soccer. Yet, subreddits such as Technology, World News, and Today I Learned

(TIL) that are focused on sharing news and useful information tend to appear in S_{vol}, whereas discussion-oriented subreddits such as DepthHub[2] and The Dismal Science are found in S_{vrl}.

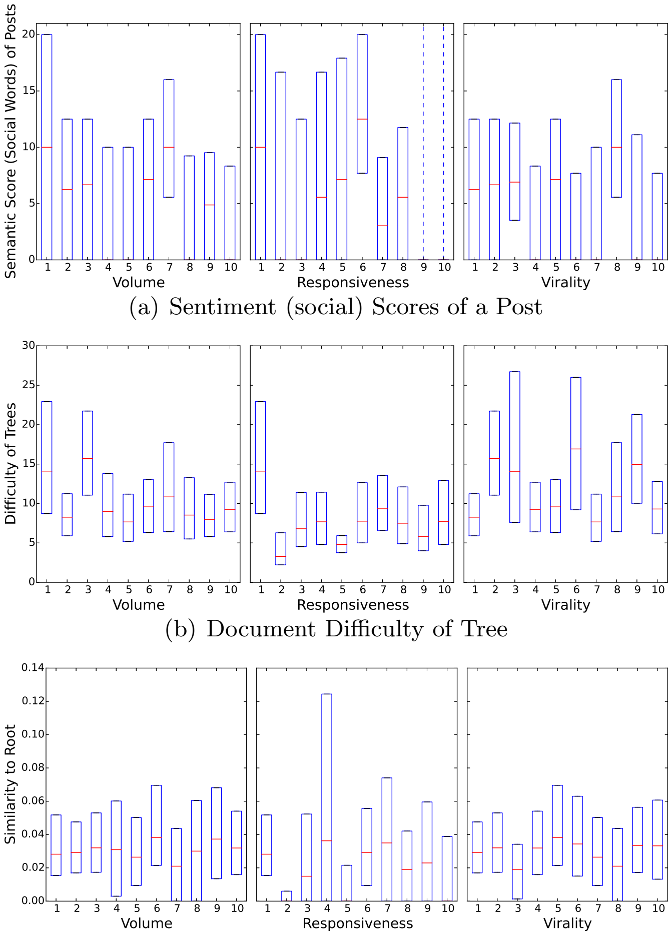
On the other hand, S_{rsp} contains many subreddits associated with multimedia content; users are allowed to only upload photos in Photoshop Battle, Reddit Gold Mine, Mystery of the soda, FoodPorn, and EarthPorn, and to link music streaming in Music. This implies that multimedia content usually leads to users' quick responses, which may not lead to large and viral conversations.

Interestingly, IAmA, where people introduce themselves or find some other people to ask something, shows a unique pattern; it ranks the first in terms of both volume and responsiveness, both of which are two disparate lists. Since the conversations in IAmA are often driven by celebrities and imply real-time interactions where an initiator answers questions from commenters, it often draws huge attention (large volume) and is highly responsive (real-time Q&A).

4.2.2 Content and User Characteristics

We now analyze how content properties and users' participation behaviors are different across the top 10 topical communities in Table 4.2. For the content properties, we report the three representative metrics, which turn out to be relevant to large, responsive, and viral conversations in Section 4.1.1: (i) the sentiment (social) score of a post, (ii) the document difficulty of a conversation by Gunning-Fog indexes, and (iii) the document similarity to a post in a conversation. Note that we exclude outliers and plot the values of

[2] DepthHub gathers the best in-depth submissions and discussions in Reddit.



(a) Sentiment (social) Scores of a Post

(b) Document Difficulty of Tree

(c) Document Similarity to a Post

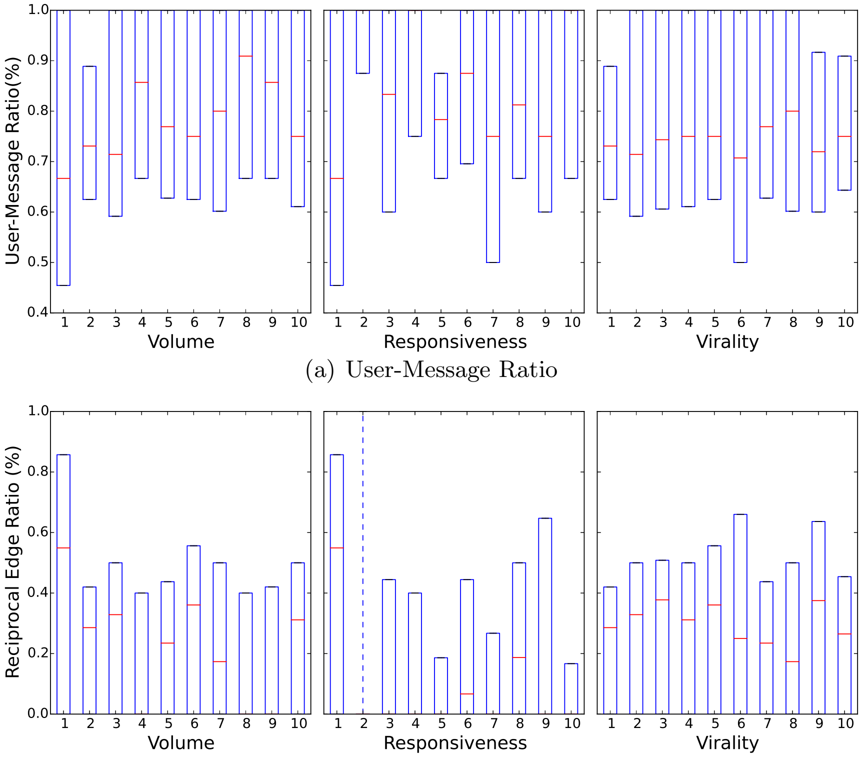
Figure 4.9 Three content properties across subreddits are plotted in terms of sentiment scores of a post, document difficulty of a tree, and document similarity to a post.

ones ranging from 25% to 75% of the distribution (as a box plot) to focus on the normal cases.

Figure 4.9(a) first shows the distributions of the social scores of posts across different topical communities. For brevity, we refer to n^{th} subreddits in S_{vol}, S_{rsp}, and S_{vrl} as $S_{vol}(n)$, $S_{rsp}(n)$, and $S_{vrl}(n)$, respectively. Overall the distributions of the social scores are different across different topical communities. As shown in Figure 4.9(a), the medians of the social scores of both IAmA ($S_{vol}(1)$ or $S_{rsp}(1)$) and AskReddit($S_{vol}(6)$) are higher than 10.0, which means that posts in those subreddits tend to use more social words than other subreddits. On the other hand, the social scores of FoodPorn ($S_{rsp}(9)$) and EarthPorn ($S_{rsp}(10)$) are mostly zero, meaning that the posts in those subreddits tend to have few social words.

When we look at the distributions of the document difficulties of comment trees in Figure 4.9(b), we find that some subreddits in S_{vrl} have higher difficulties than others. For example, the difficulty values of comment trees of Game Discussion ($S_{vrl}(2)$), DepthHub ($S_{vrl}(3)$), The Dismal Science ($S_{vrl}(6)$), and Frugal Living ($S_{vrl}(9)$) are higher than those of other subreddits, most of which are associated with discussion-driven subreddits. Note that the conversations in Photoshop Battle ($S_{rsp}(2)$) and Mystery of the soda ($S_{rsp}(5)$) are likely to be easy, which is associated with responsive conversations. The average document difficulty of IAmA ($S_{vol}(1)$ or $S_{rsp}(1)$) is also high, even though it does not belong to the list S_{vrl}, which suggests that the post of a conversation in IAmA tends to contain social words but its generated comments (including itself) are likely to be sophisticated.

Figure 4.9(c) shows the document similarity to the original post across different subreddits. We find that the document similarity values of subreddits in S_{vol} and S_{vrl} are mostly high. However, we observe that subreddits in S_{rsp} show different: the comments in Photoshop Battle and Mystery of the soda are rarely relevant to their posts, whereas the comments in Reddit Gold Mine ($S_{rsp}(4)$) are closely relevant to their posts, which implies that posts in Reddit Gold Mine ($S_{rsp}(4)$) tend to drive users to make their comments (not on comments, but on posts).



(a) User-Message Ratio

(b) Reciprocal Edge Ratio

Figure 4.10 User-message and reciprocal edge ratios are plotted across subreddits.

We then investigate users' participation behaviors across the top 10 topical communities in Table 4.2 with two user metrics plotted in Figure 4.10: (i) user-message ratio and (ii) reciprocal edge ratio. We find that the user-message ratios of the most subreddits in S_{vol} and S_{vrl} are relatively lower than the ones in S_{rsp} while the reciprocal edge ratios of the most subreddits in S_{vol} and S_{vrl} are substantially higher than the ones in S_{rsp}. This result is in line with Section 4.1.2 that revealed large and viral conversations are

Group No.	Type	Subreddit Groups	Avg. Vol.	Avg. Resp.	Avg. Vir.
1	Discussion	Game Discussion, Football Discussion	81.76	0.23	3.74
2	Multimedia	Photoshop Battle, Music, Reddit Gold Mine, The mystery of the soda	19.81	0.57	1.72
3	IAmA	IAmA	125	1.35	2.74
4	Information	Technology, You Should Know, Android, Soccer, DepthHub	58.59	0.22	2.90

Table 4.3 Groups of subreddits. We report average volume, responsiveness, and virality for each group.

likely to have low user-message ratio and high reciprocal edge ratio. However, `Technology` ($S_{vol}(4)$), `World News` ($S_{vol}(8)$), and `TIL` ($S_{vol}(9)$) show an opposite tendency; their user-message ratios are high but their reciprocal edge ratios are low, which implies that participants tend to submit a small number of comments and not to reciprocally communicate with others, probably because they are focused on sharing new information (news or knowledge) rather than discussion. Note that `IAmA` ($S_{vol}(1)$) shows a noticeable pattern; its user-message ratio is much lower and reciprocal edge ratio is much higher than the other subreddits.

The responsive subreddits (in S_{rsp}) tend to have high user-message ratio and low reciprocal edge ratio in general. However, the user-message and reciprocal edge ratio of `AskReddit` ($S_{rsp}(6)$) and `Game of Thrones` ($S_{vol}(8)$) show the somewhat inconsistent tendency, meaning that participants in those subreddits tend to be not only responsive but also reciprocal with other people. Note that both user-message and reciprocal edge ratio values of conversations in `Science` ($S_{rsp}(7)$) are relatively lower than those of other subreddits, which implies that participants in the science-related subreddit are likely to submit more comments, but they do not actively interact with others.

4.2.3 Groups of Subreddits

We have shown that conversation patterns are different across different communities (subreddits). In this section, we explore how multiple communities can be grouped and what are the characteristics of groups of communities. To this end, we first calculate a vector $v_i = (x_{vol}, x_{resp}, x_{vrl})$ for the subreddit i, where each element represents the average volume, responsiveness, and virality of the subreddit i, respectively. To cluster/group the subreddits, we apply the K-means clustering algorithm [42] that calculate the Euclidean distances among the vectors (of subreddits). Note that we determine the number of groups (i.e., K) as 4.

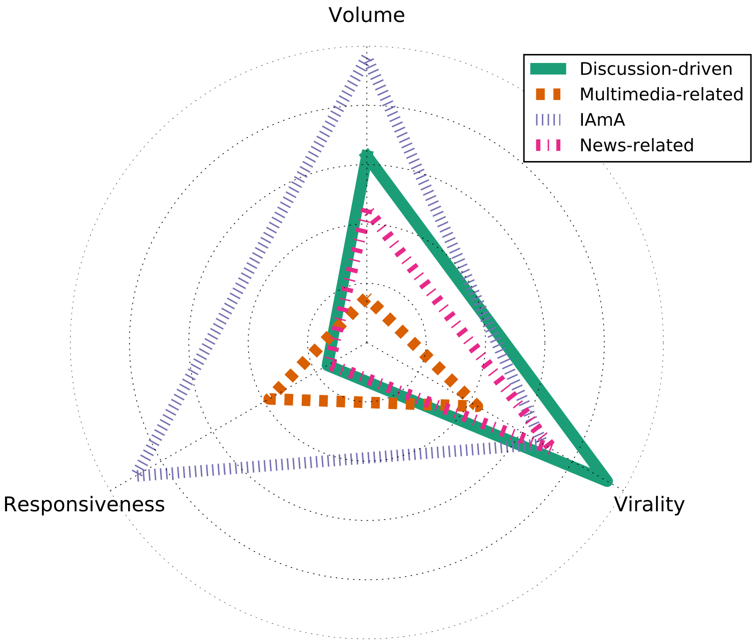


Figure 4.11 The degree of volume, resposiveness, and virality for subreddits with similar topics is described.

Table 4.3 shows the groups of subreddits, with their collective characteristics including average volume, responsiveness, and virality. To illustrate three characteristics, we represent the degree of volume, responsiveness, and virality in Figure 4.11. As shown in Table 4.3 and Figure 4.11, subreddits with similar topics tend to belong to the same group. For example, groups 1 is relevant to discussion-driven subreddits (e.g., Game Discussion) where users in this group tend to interact with others in a conversation and make conversations more viral by generating sub-conversations. On the other hand, groups 2 includes multimedia-related subreddits (e.g., Music), where users in these subreddits are likely to react (i.e., comment) to the original multimedia content (e.g., image), resulting in responsive and relatively less viral conversations. The group 4 subreddits tend to deal with informative topics like new technology or life experiences, generating moderately-large conversations. Note that IAmA shows a distinct pattern compared to other subreddits; thus it is the only member in group 3. These results imply that communities, where conversation patterns are similar, are likely to deal with similar topics.

Group No.	Difficulty		Social		Pos. Emo		Neg. Emo	
	Post	Tree	Post	Tree	Post	Tree	Post	Tree
1	6.77	12.4	7.93	8.77	3.3	4.2	1.95	2.29
2	5.43	8.71	9.62	9.74	4.01	3.9	3.56	2.7
3	9.22	17.06	11.92	12.87	1.49	3.48	1.13	1.36
4	6.62	10	5.48	7.48	3.36	4.12	1.49	2.04

Table 4.4 Content properties of each subreddit group are described.

We next investigate how content properties are different across subreddit groups. To this end, we compute the difficulties and sentiment scores of conversations (for their posts and trees, respectively) in each group. As shown in Table 4.4, conversations posted in different subreddit groups indicates distinct content properties. For example, the conversations in the self-introduction group (i.e., IAmA(3)) tend to be more difficult and social, meaning that users in these subreddits are likely to use difficult and social words. When we look at the multimedia-related group (i.e., 2), the emotion scores (i.e., both positive and negative scores) of posts tend to be high while their difficulties are relatively low, which implies that users in this group tend to use emotional titles and easy words. On the other hand, discussion-related groups (1) show higher difficulty and fewer emotion scores. A possible assumption is that users in the group are likely to use domain-specific or professional words to discuss with others and to abstain using emotional words. Interestingly, information-related group (4) shows the lower social score, but higher positive emotion score, implying that social words are unlikely to be used to describe new information or technology and users are likely to approve the news positively.

No.	Gini Coeff. (with Root)	Gini Coeff. (without Root)	Reciprocal Edge Ratio	User-Message Ratio
1	0.47	0.24	0.3	0.74
2	0.70	0.18	0.20	0.86
3	0.54	0.16	0.47	0.69
4	0.53	0.22	0.26	0.78

Table 4.5 User participation behaviors in each group are described.

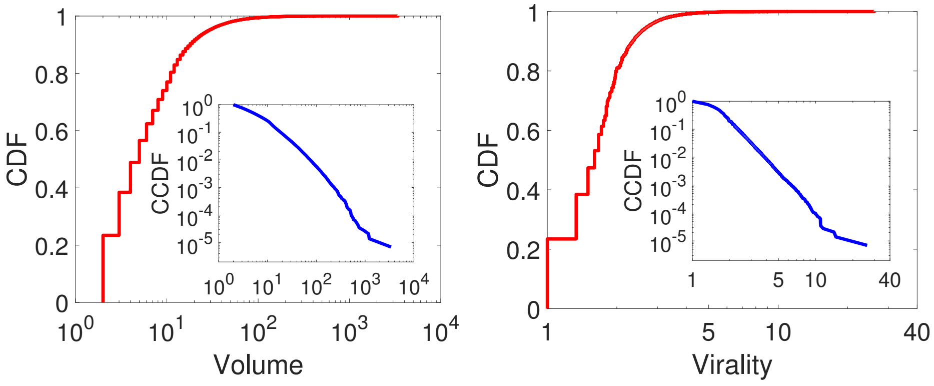
We finally investigate the users' participation behaviors in different subreddit groups, by calculating the Gini coefficient with/without root nodes, reciprocal edge ratio, and user-message ratio. As shown in Table 4.5, users in different subreddit groups show different participation behaviors. We observe that the conversations with roots included in the multimedia-related group tend to show high Gini coefficient values. Also, their reciprocal edge ratios and user-message ratios are likely to be low and high, respectively. This implies that users participating in those multimedia-related groups tend to react to the original content (i.e., posts) rather than communicate with other users. The conversations of the discussion-driven group tend to have high reciprocal edge ratio and low user-message ratio, while the conversations of the information-related group are likely to have low reciprocal edge ratio and high user-message ratio. This signifies the difference between discussion-driven and information communities; discussion-driven conversations are likely to be generated by reciprocal communications among relatively a small number of users, while information-related conversations are likely to be generated by a larger number of unrelated users.

Chapter 5

Analysis on Image Cascade in Pinterest

5.1 Characteristics of Image Cascades

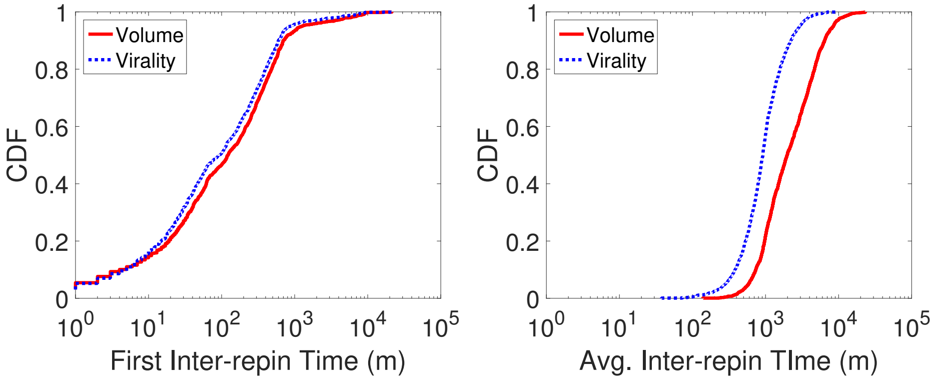
We start characterizing image cascades with investigating the distributions of volume and structural virality of image cascades in Figure 5.1. As shown in Figure 5.1(a), the volume of image cascades shows a heavy tail distribution; the range of volume spans four orders of magnitudes. While 77% of cascades have less than 10 users (including a pinner and 9 repinners), top 1% and 0.1% of cascades consist of more than 74 and 230 users, respectively, signifying a large deviation among cascades. Note that the average, median, and maximum volume is 9.26, 5, and 3,401, respectively. When we look at Figure 5.1(b), the structural virality also shows a heavy tail distribution, which only spans two orders of magnitudes. While around 81% of WI values are smaller than 2 and 99.7% of them are smaller than 5, top 0.1% of WIs



(a) Volume

(b) Structural virality

Figure 5.1 Distributions of volume and structural virality of image cascades are described.



(a) First repin time

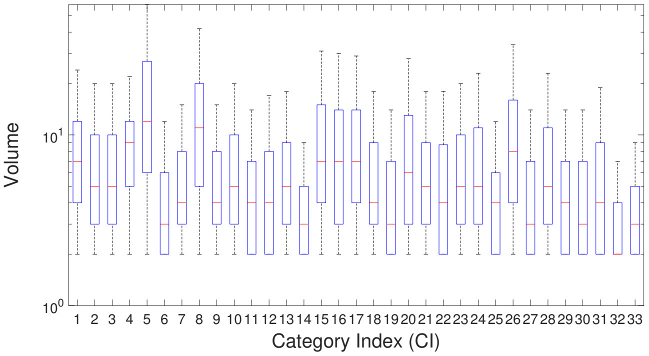
(b) Average inter-repin time

Figure 5.2 Distributions of repin times of image cascades are depicted.

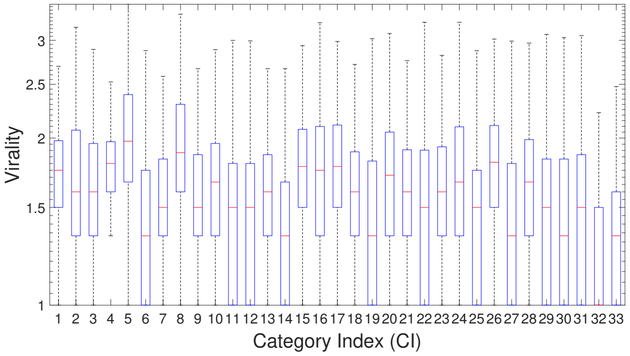
are greater than 6.32. This implies that most of image cascades in Pinterest are not likely to span deep. Note that the average, median, and maximum structural virality is 1.66, 1.6, and 26.12, respectively.

To capture how quickly users propagate images in the top 1% image cascades in terms of volume and structural virality, we calculate the first inter-repin times (i.e., the time difference between the original pinning and the first repinning) and the average inter-repin times of the top 1% cascades in Figure 5.2. In calculating the average inter-repin time, we only consider the time differences (in a cascade) within the range of $[\mu - 2\sigma,\ \mu + 2\sigma]$ for excluding outliers. As shown in Figure 5.2, the inter-repin times of the top popular cascades (with high volumes) are higher than those of the top viral cascades (with higher WIs), meaning that users in viral cascades tend to propagate images more quickly. This implies that the propagation speed can be used to predict popular or viral image cascades. For example, if we observe the initial propagation speed of a cascade, we may forecast whether the cascade goes viral in the future.

We next investigate the volume and structural virality of image cascades across 33 categories and the top 20 sources in Figures 5.3 and 5.4, respectively. As shown in Figure 5.3, the volume and structural virality are different across categories. This implies that category and source information of an image cascade can be one of the important factors in predicting whether the cascades grow much or goes viral. Interestingly, 'humor' (category index (CI) 8), 'quotes' (CI 5), and 'tattoos' (CI 26) show higher volume and structural virality than others while there are relatively less number of pins in those categories. Also, the volume and structural virality are different across sources as shown in Figure 5.4. Interestingly, the images from 'themetapicture.com' (source index (SI) 20) are much more popular and viral than others; the website is not so popular in general (the Alexa rank is 20,328 as of October 2016) and provides funny pictures. Also, food-related sources such as 'food.com' (SI 14) and 'allrecipes.com' (SI 15) are likely to provide popular and viral images to Pinterest.



(a) Volume

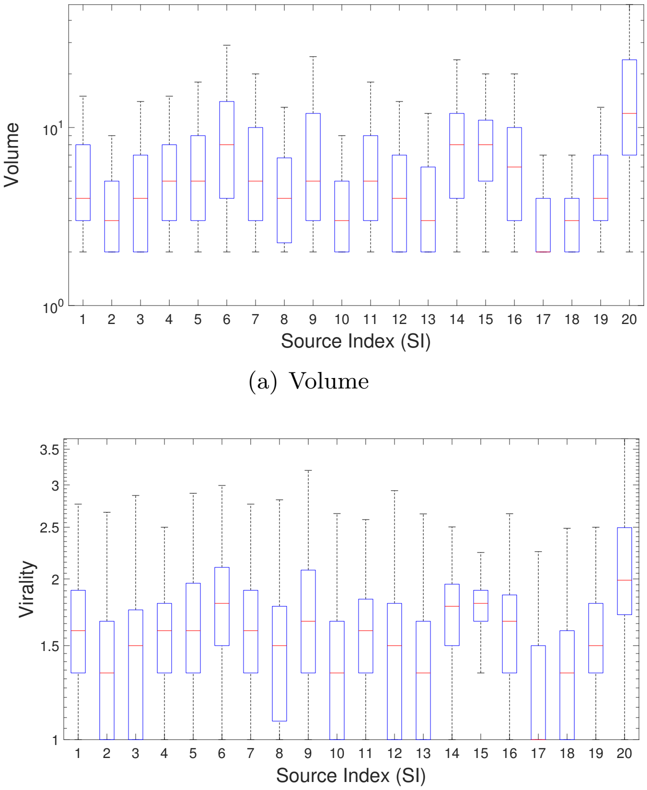


(b) Sturctural virality

Figure 5.3 We illustrate distributions of volume and structural virality of image cascades across 33 categories.

5.2 Are Popular Images Also Viral?

We now investigate whether the popular images (i.e., cascades having high volume) are also viral. To this end, we first plot the volume/virality of each image cascade in Figure 5.5(a). As shown in Figure 5.5(a), there is an overall positive correlation between the volume and virality, which means cascades with higher volumes tend to have higher structural viralities. However, the viralities of cascades with high volumes (e.g., over 100) tend to radiate. The Pearson correlation between the volume and virality of the top 1% cascades



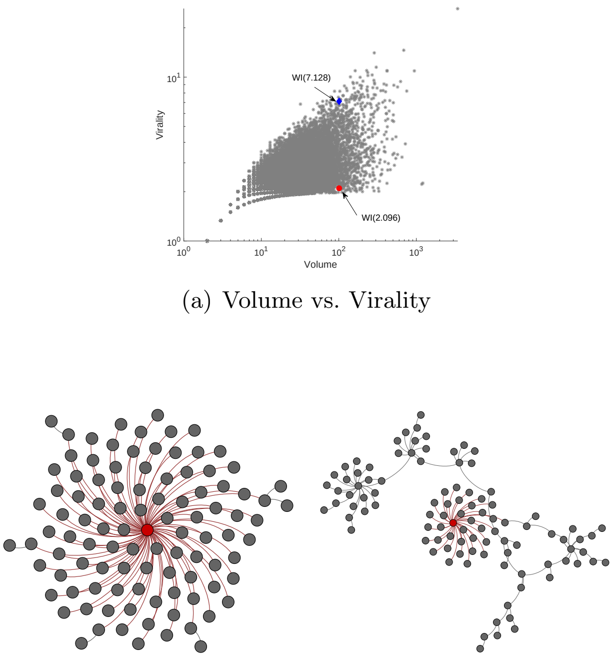
(a) Volume

(b) Sturctural virality

Figure 5.4 Distributions of volume and structural virality of image cascades across the top 20 sources are illustrated.

(whose volumes are higher than 74) is 0.42. This implies that popular images are not necessarily viral. Note that the Pearson correlation between the volume and virality of the bottom 99% cascades (whose volumes are 74 or smaller) is 0.8. When we look at the volume-based top 1% cascades and the virality-based top 1% cascades, only 17.4% of the cascades are overlapped, which signifies that top popular and viral cascades are disparate. This implies that different factors may be useful for predicting popular or viral image cascades, which will be discussed in the next section.

As an example, we illustrate two cascades with same volume ($N = 101$)



(a) Volume vs. Virality

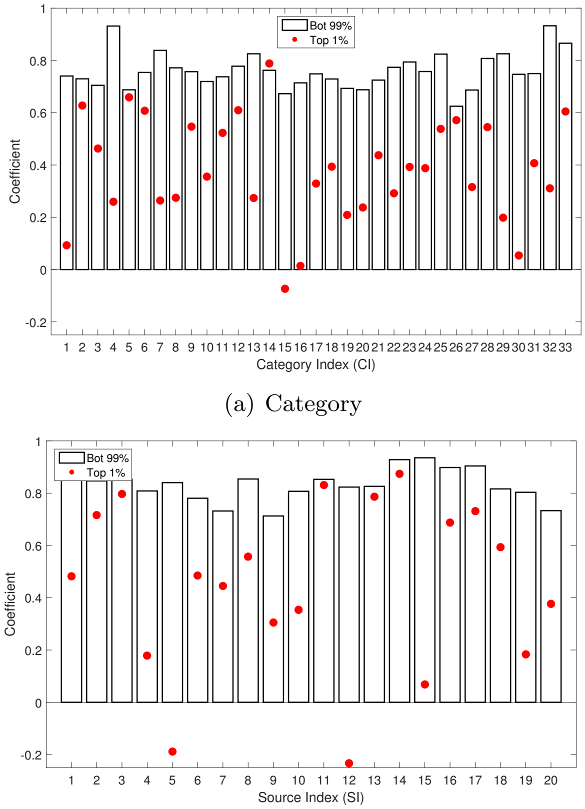
(b) Broadcast

(c) Viral diffusion

Figure 5.5 We describe volume and virality of each image. Overall, there is a positive correlation between the volume and virality. However, popular images are not necessarily viral, e.g., two cascades with same volume ($N = 101$) have different viralities: $WI = 2.096$ for the red circle (b) and $WI = 7.128$ for the blue diamond (c).

but different viralities, $WI = 2.096$ for the red circle and $WI = 7.128$ for the blue diamond in Figures 5.5(b) and 5.5(c), respectively. As shown in Figures 5.5(b) and 5.5(c), two similarly popular images can be propagated through different scenarios: (i) broadcast where a pinner mostly spreads an image to the most of recipients and (ii) viral diffusion where an image propagates via the person-to-person contagion process. This confirms that an image can be popular but not viral.

To investigate whether popular images are viral in different categories



(a) Category

(b) Source

Figure 5.6 We show pearson correlation coefficients between volume and virality of the top 1% (with high volumes, represented as red dots) and bottom 99% (represented as bar) cascades, respectively, in each category and source.

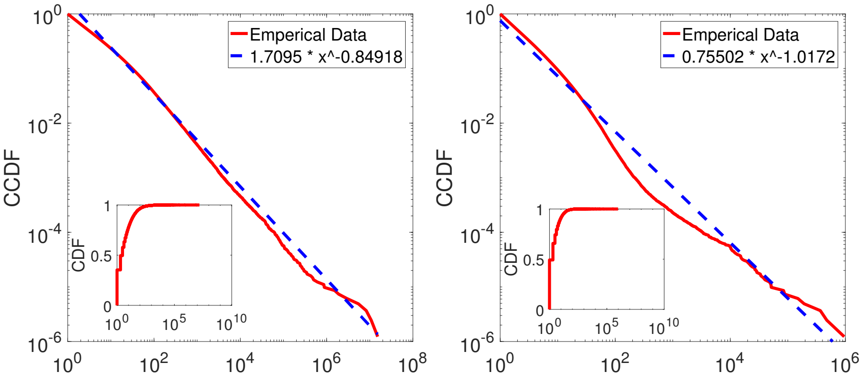
and sources, we calculate the Pearson correlation coefficients between the volume and structural virality in each category and source. Figure 5.6 shows the two coefficient values for the top 1% (with high volumes, represented as redo dots) and bottom 99% (represented as bar) cascades, respectively, in each category and source. Overall, the Pearson correlation coefficients of the bottom 99% cascades are very high, i.e., mostly over 0.7. The coefficients of the bottom 99% cascades in 'food & drink' (CI 4) and 'shop' (CI 32) are even higher than 0.9 as shown in Figure 5.6(a). However, the coefficients of the top 1% cascades are substantially lower than those of the bottom 99%. Especially, the coefficients of the top 1% cascades in 'men's fashion' (CI 15), 'science & nature' (CI 16), 'sports' (CI 30), and 'diy & crafts' (CI 1) are lower than 0.1, meaning that popular images in those categories are not necessarily viral. We observe a similar pattern in Figure 5.6(b) that shows popular images from particular sources (e.g., 'imdb.com' (SI 4), 'greatist.com' (SI 5), 'houzz.com' (SI 12), 'allrecipes.com' (SI 15), 'wikipaintings.org' (SI 19)) are not necessarily viral. Interestingly, top 1% popular images in 'greatist.com' (SI 5) and 'houzz.com' (SI 12) show even weak negative correlations, which is a significantly disparate pattern with the bottom 99% (unpopular) images from the sources.

Chapter 6

Analyzing Content Publishing and Sharing Patterns through Bitly

6.1 Content Sharing Patterns thorough Bit.ly

6.1.1 URL Shortening Patterns



(a) Number of short URLs

(b) Number of Domains

Figure 6.1 Numbers of short URLs and domains for each user are plotted.

We first investigate how people create short URLs through Bit.ly. Figures 6.1(a) and 6.1(b) show the distributions of the number of short URLs created by and that of domains shortened by each user, respectively. As shown in Figure 6.1(a), 35% of users shorten only 1 URL while 3.66% of users create more than 100 short URLs. Figure 6.1(b) shows that around half of users create short URLs of only a single domain, but 0.32% of users shorten content pages in more than 100 domains, meaning that URL shortening users are likely to shorten content for a small number of domains. Interestingly, the CCDF of empirical data is under the fitting function in $[100, 10000]$ ranges, but is over the fitting function when the number of domains is greater than 10000, meaning that URL shorteners tend to create short URLs for either only a small number of domains or a large number of domains. Note that the two distributions both follow the power-law [5] with (1.7095, -0.84918, $x \geq 10$) and (0.75502, -1.0172) as parameters.

Rank	User Name	Number of Created Short URLs
1	twitterfeed	15,276,864
2	dens	10,820,680
3	bitly	8,289,480
4	rssgraffiti	5,448,646
5	tweetdeckapi	2,952,763
6	addthis	1,908,226
7	ameba	1,539,152
8	ifttt	884,396
9	twipple	863,334
10	zatbitly	615,117

Table 6.1 Top 10 Bit.ly users in terms of number of created short URLs are shown.

We next investigate relatively ‘active’ Bit.ly users who create more short URLs than others. Here, a Bit.ly user can be either an individual user or a company account. Table 6.1 shows the top 10 `Bit.ly` users in terms of number of created short URLs. As shown in Table 6.1, the top shorteners are likely to be the third party companies rather than individual users. For example, the third party services of Twitter (twitterfeed, tweetdeckapi, and twipple), Bit.ly (bitly and zatbitly), and Facebook (rssgraffiti) are ranked in the top 10 list. Moreover, 'dens' and 'ameba', made by service providers (`foursquare.com` and `'ameblo.jp'`, respectively) to encourage their users to publish content by short URLs for their services, are also heavily used for URL shortening. Interestingly, management tools for web services such as 'addthis', 'ifttt' are widely used in shortening URLs.

Rank	Category	Portion of URLs (%)
1	Social Network	25.59
2	News & Media	15.39
3	Computer & Electronics	9.51
4	Shopping	8.04
5	Business & Industry	4.49
6	Blogs	4.10
7	Search Engine	3.24
8	Sports	3.11
9	Arts & Entertainment	3.03
10	File Sharing	2.59

Table 6.2 Top 10 categories in terms of number of short URLs are shown.

We then examine the top categories and top domains in terms of number of URLs in Tables 6.2 and 6.3, respectively. Table 6.2 shows that content in 'Social Network', 'News & Media', and 'Computer & Electronics' (e.g., newsfeed service) is likely to be published through short URLs. The top domain (in terms of number of short URLs) in Table 6.3 is `'foursquare.com'`. We find that most of the URLs associated with `'foursquare.com'` are the check-in information, which are published by users who wish to share their current location information with others. Interestingly, `‘ameblo.jp’` (a Japanese microblog service) ranks higher than other global companies such as Google, Facebook, and Twitter, which implies a heavy usage of short URLs for content in `‘ameblo.jp’`. This may be partially because global companies provide their own URL shortening services: goo.gl, fb.me, and t.co for Google, Facebook, and Twitter, respectively. The portion of content URLs for the ‘Shopping’ category is over 8%, and they are mostly originated from `‘www.amazon.com’`, implying that short URLs are widely used in publishing shopping content.

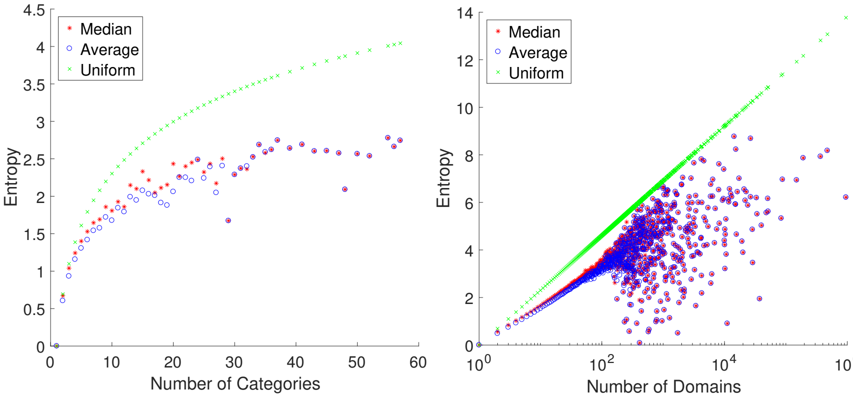
Rank	Domain	Category	Portion of URLs (%)
1	foursquare.com	Social Network	13.39
2	ameblo.jp	Social Network	2.30
3	feedproxy.google.com	Computer & Electronics	2.28
4	www.amazon.com	Shopping	1.76
5	www.google.com	Search Engine	1.32
6	www.facebook.com	Social Network	1.27
7	www.youtube.com	Streaming	1.25
8	twitter.com	Social Network	1.11
9	news.google.com	News & Media	1.06
10	apps.facebook.com	Social Network	1.05

Table 6.3 Top 10 domains (in terms of number of short URLs) and their associated categories are shown.

To investigate how uniformly each user shortens content pages across the categories and domains, we count the number of content URLs a user has shortened, and calculate the *category entropy* and *domain entropy* for each user u as follows:

$$Entropy_u = -\sum_{m=1}^{N_u} p_m^u \log p_m^u \tag{6.1}$$

where N_u is the number of categories/domains associated with the URLs of user u, and p_m^u is the URL portion of the m^{th} category/domain of user u.



(a) Category Entropy

(b) Domain Entropy

Figure 6.2 The median, average, and uniform values of category and domain entropy are plotted as the number of categories/domains increases.

Figure 6.2 shows the median, average, and uniform entropy values as the number of categories/domains (of a single user) increases. Note that the uniform values are calculated when the numbers of content URLs are equal across the categories/domains. The gap of entropy values between uniform and median (and average) increases as the number of categories (or domains) increases. This signifies the skewness of users' interests in shortening URLs – although there are a small number of users who shorten URLs in many categories or domains, most users are likely to focus on a few categories (and domains) in shortening URLs.

6.1.2 URL Request Pattern

We next investigate what types (or categories) of content are requested through short URLs. Tables 6.4 and 6.5 show the top 10 categories and domains in terms of number of short URL requests (i.e., through URL clicks), respectively. Overall, short URLs for 'Social Network' are heavily requested; content pages in 'www.facebook.com' are most requested through short URLs. In addition, 'www.youtube.com' and 'www.amazon.co.jp' are also in top 10 by the number of requests. These results seem to be related to the global popularity of websites reported in [9]. That is, popular websites (e.g., Google, Facebook, Amazon, and Youtube), whose content pages are heavily accessed in general, are also more likely to be requested through short URLs. However, interestingly, the portions of requests of the content pages in relatively less popular domains such as www.pornhub.com, mlks.co, and www.lapatilla.com are also high, meaning that content pages in these domains tend to be accessed through short URLs rather than direct access. The gap of popularity may come from the functionalities of short URLs; For example, since the informative text (e.g., domain name or content title) represented in URL can be hidden through URL shortening, adult content pages are likely to be published and shared through short URLs. Note that the URL publishing practice and access patterns are disparate when we compare Tables 6.2 and 6.4; while there are not many short URLs for content in the 'Shopping' and 'Adult' categories, they are likely to be requested many times.

We next investigate how many domains the short URLs are requested from, which are called *referrers*. For example, if a user clicks a short URL in Facebook, www.facebook.com is a referrer domain. Figure 6.3(a) shows the distribution of the numbers of referrer domains for a given short URL. About 60% of short URLs are requested from only one referrer domain while 0.01% of short URLs are requested from more than 100 referrer domains. Note that

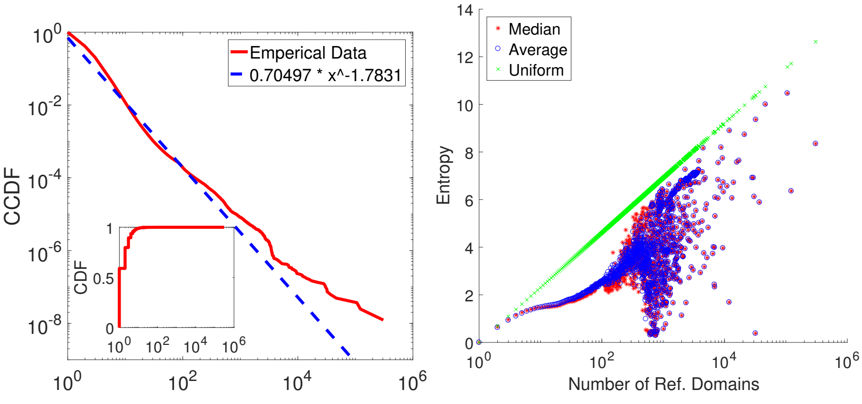
Rank	Category	Portion of URL Requests (%)
1	Social Network	14.95
2	Shopping	14.55
3	News & Media	9.86
4	Computer & Electronics	7.68
5	Adult	7.08
6	File Sharing	5.82
7	Arts & Entertainment	5.20
8	Streaming	5.09
9	Business & Industry	3.67
10	Sports	3.27

Table 6.4 Top 10 categories in terms of number of requests are listed.

Rank	Domain	Category	Portion of URL Requests (%)
1	www.facebook.com	Social Network	8.38
2	www.pornhub.com	Adult	4.96
3	apps.facebook.com	Social Network	1.99
4	rtm.ebaystatic.com	Shopping	1.74
5	www.youtube.com	Streaming	1.56
6	itunes.apple.com	File Sharing	1.42
7	mlks.co	Uncategorized	1.13
8	www.amazon.co.jp	Shopping	1.06
9	www.lapatilla.com	News & Media	1.01
10	feedproxy.google.com	Computer & Electronics	0.69

Table 6.5 Top 10 domains (in terms of number of URL requests) and their associated categories are listed.

the CCDF of empirical data is over the fitting line, meaning that there are a few short URLs accessed from a large number of referrer domains. We also plot the referrer entropy in Figure 6.3(b), which quantifies how uniformly requests are distributed across the referrer domains, whose calculation is similar to Eq. 6.1. As shown in Figure 6.3(b), the median and average values of the request entropies across referrer domains do not increase as much as those of the uniform case, meaning that most URL requests are generated in a few referrer domains.

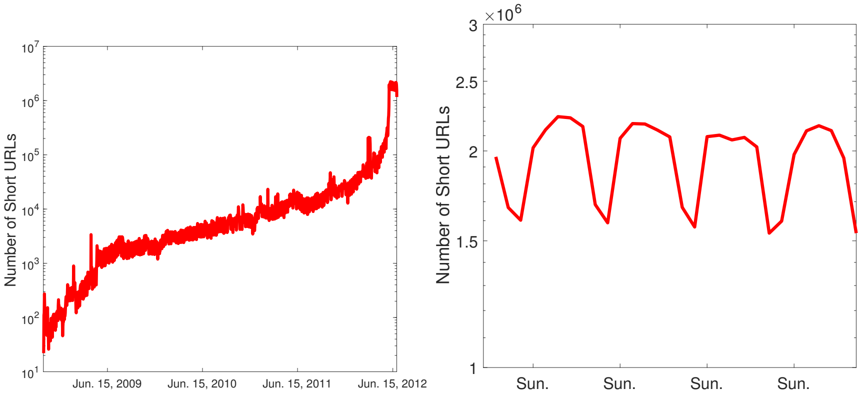


(a) Number of Referrer Domains

(b) Referrer Entropy

Figure 6.3 The distributions of the numbers of referrer domains and referrer entropies are plotted as the number of referrer domains increases.

We next examine the temporal characteristics of the requests to short URLs. To this end, we group the short URLs based on their creation dates and count the numbers of short URLs requested in our measurement period. Note that we describe the number of short URLs created in (i) whole period (Figure 6.4(a)), and (ii) recent 1-month period (Figure 6.4(b)).



(a) Whole Period

(b) Recent 1 Month (June 2012)

Figure 6.4 The numbers of requested short URLs created in (a) whole period and (b) June 2012 are plotted.

As shown in Figure 6.4(a), the number of short URLs increases as the short URLs are continuously created. Note that the numbers of short URLs created in June 2012 are 10 times larger than the ones for short URLs created in the previous month (i.e., May 2012), which implies that short URLs which are relatively recently created are likely to be requested. We also observe that thousands of URLs generated from 2009 to 2011 are still requested in June 2012. When we zoom in the creation time on June 2012, there are distinct temporal patterns between weekdays and weekend; the number of short URLs becomes higher for weekdays and lower for weekends. This is in line with the finding that more URLs are shortened in weekdays than weekends [2].

We also investigate how short URLs are requested from a geographical perspective. We observe the top 5 domains and categories in terms of the number of requests for five representative countries (i.e., USA, Japan, China, Brazil, and GBR) where short URLs are mostly used. As shown in Table 6.6, the URL access patterns are different across the countries. The 'Shopping' URLs are mostly accessed in USA and China; URLs for the 'Social Network' category are actively requested in Japan and Brazil. Note that `'www.lepirata.com'`, a shopping site that sells football jerseys, ranks high in Brazil. When we look at the top 5 domains in USA and China, the CDNs for **Ebay** and **Taobao**, respectively, are the dominant contributors in URL requests. Interestingly, URL requests by Japanese users are likely to go toward localized social services such as `'amablo.jp'` or `blog.livedoor.jp`. Note that people in GBR are likely to request news content, mostly created in `'bbc.co.uk'`, implying that BBC is a major news platform for the short URLs in Britain.

(a) Top 5 domains

Country	Domain	Portion of Requests for the Domain (%)
USA	rtm.ebaystatic.com	10.79
	mobile.ebay.com	3.96
	api.ning.com	2.34
	trib.al	1.75
	www.facebook.com	1.50
Japan	ameblo.jp	11.57
	www.amazon.co.jp	11.54
	cdn1.ustream.tv	10.98
	blog.livedoor.jp	2.51
	p.twipple.jp	2.12
China	www.wmybuy.com	15.38
	img01.taobaocdn.com	6.71
	img03.taobaocdn.com	6.56
	img02.taobaocdn.com	6.52
	img04.taobaocdn.com	5.88
Brazil	www.facebook.com	8.39
	www.lepirata.com	7.32
	www.faston.com.br	6.12
	www.tufos.com.br	2.45
	feedproxy.google.com	2.15
GBR	www.bbc.co.uk	4.05
	www.facebook.com	1.82
	dist1.terasoft.lt	1.78
	viper.w12.org	1.65
	api.ning.com	1.26

(b) Top 5 categories

Country	Category	Portion of Requests for the Category (%)
USA	Shopping	20.47
	News & Media	13.30
	Computer & Electronics	10.82
	Arts & Entertainment	10.67
	Business & Industry	6.17
Japan	Social Network	22.04
	Shopping	15.96
	Streaming	13.89
	Computer & Electronics	7.87
	Search Engine	4.39
China	Shopping	35.95
	Computer & Electronics	21.70
	File Sharing	8.64
	Business Services	4.21
	Games	3.93
Brazil	Social Network	12.83
	Violence & Illegal	10.68
	Shopping	10.27
	Arts & Entertainment	8.27
	File Sharing	7.73
GBR	News & Media	17.74
	Arts & Entertainment	12.57
	Computer & Electronics	11.65
	Sports	7.21
	Business & Industry	5.94

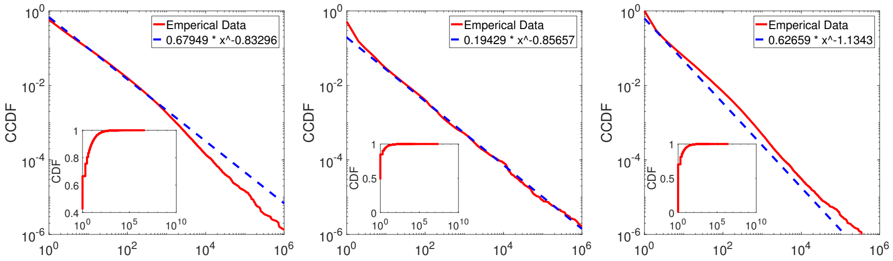
Table 6.6 Top 5 domains and categories in terms of number of requests for five representative countries (i.e., USA, Japan, China, Brazil, and GBR) are summarized.

We note that there exist several 'malicious' or 'black' domains whose content pages are highly requested. For example, `'www.wmybuy.com'`, a gambling portal web pretending to be a shopping portal, is the most highly-requested domain in China. Also, URLs for 'Violence & Illegal' domains are highly requested in Brazil, and `'www.tufos.com.br'`, a site for 'Adult' content, is highly accessed.

6.2 Content-Referrer Graph

6.2.1 Basic Analysis

Based on the proposed graph model described in Section 3.2.3, we investigate (i) how different domains are associated with others, and (ii) which domains play important roles in publishing short URLs. Figure 6.5 shows the distributions of weighted in-degrees, weighted out-degrees, and weights of edges, respectively. The portions of nodes that only have in-degrees (i.e., the 'referrer-only' domains) and out-degrees (i.e., 'content-only' domains) are



(a) In-Degrees (b) Out-Degrees (c) Weights

Figure 6.5 The distributions of weighted in-degrees, out-degrees, and weights of the content-referrer graph are plotted.

48.19% and 42.21%, respectively, indicating that a high portion of domains tends to be used as only either a referrer or a content source. Furthermore, around 15% and 23% of nodes have only 1 weighted out-degree and in-degree, respectively, while 5.8% and 2.7% of domains have more than 100 weighted out-degrees and in-degrees, respectively. This implies that a small number of domains play significant roles in publishing and sharing short URLs.

Figure 6.5 also shows that in-degrees, out-degrees, and weights follow power-law with (0.67949, -0.83296, $x \geq 10$), (0.19429, -0.85657, $x \geq 5$) and (0.62659, -1.1343) as parameters, respectively. Interestingly, the CCDF of in-degrees is below the fitting function and the gap becomes larger as the in-degree increases, implying that short URLs relatively less tend to be published in popular publishing spaces (i.e., referrer domains). Note that weighted in-degrees are larger than out-degrees in general, which indicates that content is generally shortened in fewer content domains and published in more referrer domains.

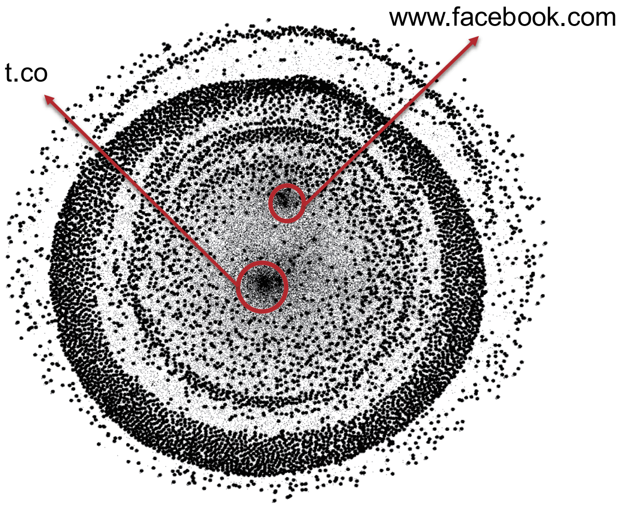


Figure 6.6 The content-referrer graph is plotted. Only the top 0.1% relations of the content-referrer graph in terms of weight are shown for visualization purposes.

6.2.2 Relations among Domains

We next investigate how domains are linked amongst themselves in terms of the content-referrer relation. To this end, we first visualize the content-referrer graph, as shown in Figure 6.6, to reveal relations among domains in a global view. Note that we plot only the top 0.1% edges in terms of the weight for the visualization purposes, and the sum of weights (i.e., total number of content URLs) in this graph is around 40 M, which accounts for 43.8% of the total weights. As shown in Figure 6.6, the content-referrer graph mainly consists of two giant groups – `Facebook.com` and `t.co` (i.e., Twitter), meaning that both representative domains are heavily used to publish short URLs.

To reveal the heavy relations between domains in the content-referrer

Content Domain	Referrer Domain	Portion of URLs (%)
foursquare.com	t.co	3.16%
ameblo.jp	t.co	1.02%
apps.facebook.com	www.facebook.com	0.83%
feedproxy.google.com	t.co	0.80%
feedproxy.google.com	www.facebook.com	0.64%

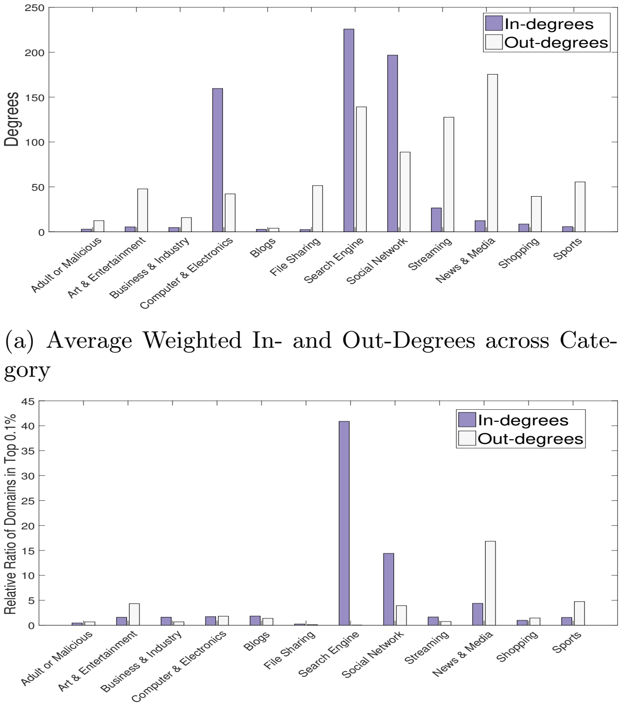
Table 6.7 Top 5 relations in terms of weight are listed.

graph, we analyze the top 5 relations in the content-referrer graph in terms of the number of short URLs (i.e., weight). As shown in Table 6.7, two representative OSNs, Facebook and Twitter, are the dominant referrer domains where short URLs are largely published. However, interestingly, we find that different content domains are likely to use the two OSNs as referral domains. The content URLs in `foursquare.com` and `ameblo.jp` tend to be published through Twitter (t.co), while `app.facebook.com` is likely to be published in Facebook. Note that `feedproxy.google.com`, an online newsfeed service, tends to use both Twitter and Facebook as primary referrers.

6.2.3 Role of Domains

We next investigate how domain categories (or types) play roles in publishing short URLs. To this end, we first classify domains into twelve categories, as described in Section 6.1. Note that 'Adult or Malicious' is a set of the following five categories, which are linked to the malicious or adult content: (i) 'Parked', (ii) 'Spam', (iii) 'Phishing', (iv) 'Violence & Illegal', and (v) 'Adult', whose content URLs for these categories are mostly shortened for hiding the original URLs (e.g., adult content, spamming, etc.) [57, 37].

Figure 6.7(a) shows the average weighted in-degrees and out-degrees for each (domain) category. Obviously, different categories play substantially dif-



(a) Average Weighted In- and Out-Degrees across Category

(b) Normalized Ratio of Domains in Top 0.1% by In- and Out-degrees

Figure 6.7 The average weighted in- and out-degrees across category and the relative ratio of domains in top 0.1% by weighted in- and out-degrees are plotted.

ferent roles in the content-referrer graph. The average in- and out-degrees of both **Search Engine** and **Social Network** domains are higher than others mostly, implying that these domains play roles as both content sources and publishing spaces. Also, the average in-degree of **Computer & Electronics** domains are higher than its average out-degree, while the tendency is reversed in the case of **News & Media** and **Streaming**. This indicates that content URLs in **News & Media** and **Streaming** domains tend to be published in many referrer domains while content URLs from multiple content domains tend to be published in **Computer & Electronics** domains such as newsfeed services.

We next investigate how many domains (in different categories) play crucial roles in the content-referrer graph, by extracting the top 0.1% of domains in terms of weighted in- and out-degrees. We calculate the relative ratio of the top 0.1% domains in each category. That is, if there are a thousand of domains whose category is A and three of them are in the top 0.1% of all the domains in terms of weighted in-degrees, then the relative ratio is $3/(1K * 0.001) = 3$, which indicates that there are three times more domains whose category is A in the top 0.1% list compared to the other categories.

As shown in Figure 6.7(b), the relative ratio of in-degrees of **Computer & Electronics** is almost zero while the relative ratios for **Search Engine** and **Social Network** are significantly high (42 and 14, respectively). Considering that average weighted in-degrees of the two categories are higher than others (as shown in Figure 6.7(a)), this implies that substantial numbers of the domains of the two categories play crucial roles as publishing spaces. Note that only a few **Computer & Electronics** domains are used as heavy publishing sources while most of the domains in the **Computer & Electronics** are used only as referrers in general. Similarly, the relative ratios of out-degrees of **File Sharing**, **Search Engine**, and **Streaming** are low while their average out-degrees are high (as shown in Figure 6.7(a)), meaning that only a small number of domains in these categories play important roles as publishing sources. Note that the relative ratio of out-degrees of **News & Media** are substantially high (16 times more than the average), implying that these domains typically play the role of content providers.

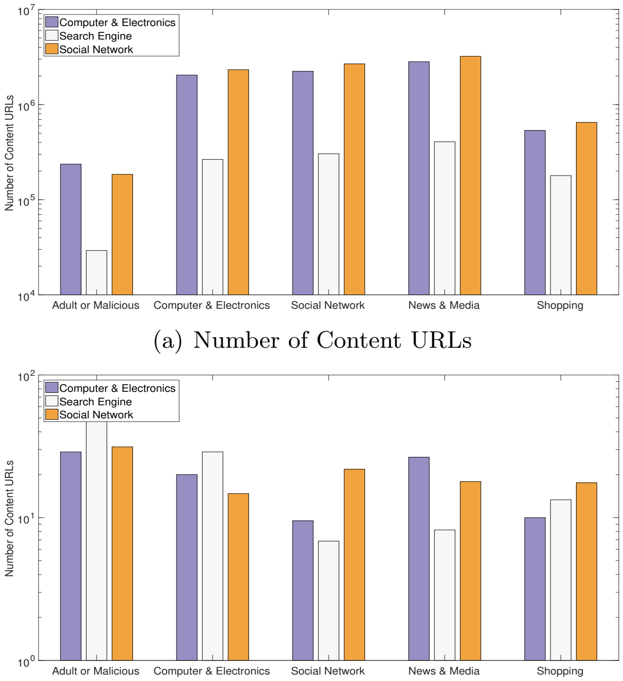
6.3 Referrer Analysis

6.3.1 Referrer Preference

We examine how content pages are published and requested from different referral domains. Here, we focus on the three referrer domains, **Computer & Electronics**, **Search Engine**, and **Social Network** domains, each with high in-degrees (see Figure 6.7(a)), which signifies significant roles in sharing content. In particular, we investigate how content URLs created in the domains of five representative categories (in terms of number of requests as shown in Table 6.4) are published in the above three referrer categories: (i) **Adult or Malicious**, (ii) **Computer & Electronics**, (iii) **Social Network**, (iv) **News & Media**, and (v) **Shopping**.

Figure 6.8 shows the number of published content URLs (in the form of short URLs) and average number of requests for each content URL shared through **Computer & Electronics**, **Search Engine**, and **Social Network** categories. As shown in Figure 6.8(a), content URLs created from the five categories are likely to be published through **Computer & Electronics** and **Social Network** referrer domains rather than **Search Engine**. This indicates that the content URLs created from the five categories tend to be published mostly through the **Computer & Electronics** and **Social Network** referrer domains, while the content URLs created from other categories (e.g., **Streaming**, **File Sharing**) are likely to be published through the **Search Engine** referrer domains. Note that we showed that **Search Engine** domains are heavily used as the publishing spaces of short URLs (see Section 6.2.3).

Interestingly, when we look at Figure 6.8(b), content access patterns are



(a) Number of Content URLs

(b) Average Number of Requests

Figure 6.8 The number of short URLs and average number of requests per short URL across the categories are shown.

disparate from content publishing ones (see Figure 6.8(a)). For example, the average number of requests of the content URLs for **Adult or Malicious** and **Computer & Electronics** published through **Search Engine** referrer domains is higher than the ones through other referrer domains. That is, users tend to access **Adult or Malicious** and **Computer & Electronics** content through the **Search Engine** referrer domains rather than the **Computer & Electronics** or **Social Network** referrer domains. The **Social Network** and **Shopping** content pages tend to be requested more through the **Social Network** referrer domains, which indicates that people interested in **Social Network** and **Shopping** are likely to request such content though in **Social**

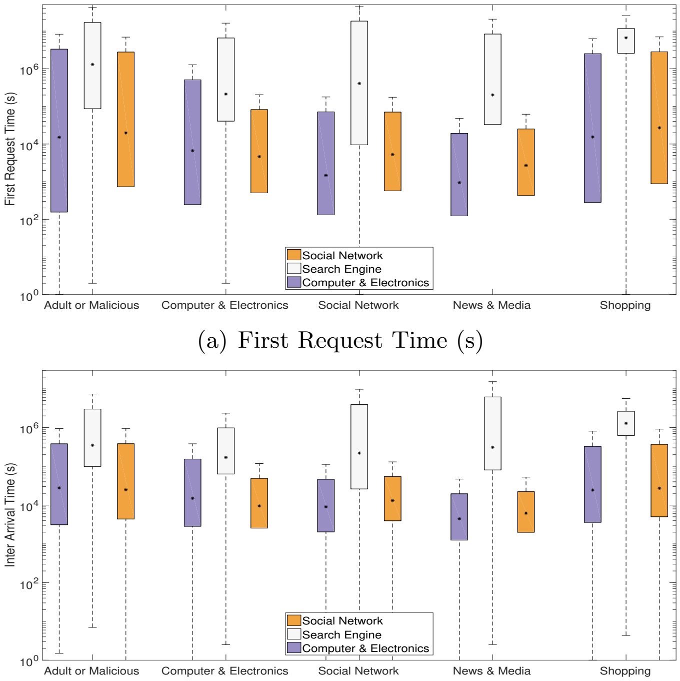
Network domains. The **News & Media** content pages are largely requested through the **Computer & Electronics** referrer domains, meaning that the **Computer & Electronics** domains are major channels for **News & Media** content.

In summary, three popular referrer domains (as publishing spaces) play different roles in sharing content from different categories. In other words, there exist effective spaces (i.e., referrer domains) that can attract users' requests for different content types, which sheds important insights for content publishers who wish to elicit more user responses or attention.

6.3.2 Referrer Responsiveness

We next investigate the access patterns of content URLs through the three referrer domains from a temporal perspective. To this end, we measure two metrics which reflect user responsiveness to content: (i) *first request time* of a URL as the time difference between the URL creation time and its first requested time, and (ii) *inter-arrival time* of a URL, which is defined as the average time between two consecutive requests for the URL from the first request to the last request. Note that we take into account only URLs that are requested at least twice.

As shown in Figure 6.9, the first request time and inter arrival time of content URLs (in the five categories) are different across different referrer domains. That is, users' responses are different temporally across different publishing spaces. Overall, both the first request time and inter arrival time of content URLs published in the **Search Engine** referrer domains are higher than those published through the **Computer & Electronics** and **Social Network** referrer domains. Note that the gaps between the **Search Engine**



(a) First Request Time (s)

(b) Inter-Arrival Time of Requests

Figure 6.9 First request time and inter arrival time for five categories are plotted.

and other referrer domains become relatively larger for the **News & Media** and **Shopping** content than other content, which indicates that, if news or shopping content pages are published through **Computer & Electronics** and **Social Network**, users tend to access the content quickly and virally.

Figure 6.9 also reveals that user access patterns for different content categories are various even though they are published in the same referrer domains. For example, as shown in Figure 6.9(a), the first request times of **News & Media** and **Social Network** content published through the **Computer & Electronics** and **Social Network** referrer domains are lower than those of other content categories. This implies that news or SNS-related content tend to be requested quickly. Note that the lengths of boxes (i.e., range from 25% to 75%) of the **Adult or Malicious** and **Shopping** content published in **Computer & Electronics** and **Social Network** referrer domains are longer than others, implying that the users' first responses spread more in these cases.

Chapter 7

Predicting Large/Viral Conversations and Image Cascades

7.1 Predicting Large/Viral Conversations

Our measurement-based characterization have revealed that there exists a set of distinctive factors (e.g., content properties, users' participating behaviors) that are associated with the volume, responsiveness, and virality of a conversation. In this section, based on lessons learned, we propose leaning-based models to predict large or viral conversations. In particular, we consider two scenarios where our proposed models can be applied: (i) when only post information is available and (ii) when an initial conversation (i.e., post and a few early comments) can be observable. To this end, we first formulate the prediction task as a function of the features extracted from a post and initially observed comments. We then develop learning-based models to predict whether a comment tree would be large or viral.

7.1.1 Problem Formulation

To predict whether a comment tree would be large or viral, we first extract the sub-tree t_m, consisting of a post and initial m comments, from the given comment tree t in T where T is the set of all comment trees. We then compute the input features of the t_m, denoted by f_{t_m}. Based on the f_{t_m}, we identify whether the t_m would belong to the set of comment trees T^k_{vol} or T^k_{vrl}, where T^k_{vol} and T^k_{vrl} indicate the top $k\%$ of comment trees in terms of volume and virality, respectively. In other words, our prediction problem for a given comment tree can be defined as finding the function g, formulated as follows:

$$g_{type} : f_{t_m} \rightarrow \delta^k_{type} \tag{7.1}$$

where type is one of target variables (either volume or virality in our case) and δ^k_{type} is 1 if t is in T^k_{type}, otherwise 0. Note that we consider the following two scenarios where our models can be applied: (i) when only post information is available (i.e., $m = 0$) and (ii) when an initial conversation can be observable (i.e., $m > 0$).

7.1.2 Experiment Setup

Input Features: Sections 4.1.1 and 4.1.2 have shown that content properties and users' participating behaviors are related to the volume and virality of conversations, meaning that all these features can be used for the prediction task. In addition to those high-level features (analyzed in the previous sections), we further use the *word embedding* to extract the low-level features from the text itself. To this end, we separate a given text into a set

Feature Group	Feature Name	Feature Description
Content (Semantic) (Section 4.1.1)	diff_p	difficulty of a post.
	diff_t	difficulty of an initial tree
	emo_p	sentiment scores of a post
	emo_t	sentiment scores of an initial tree
	tfidf_r	average TF-IDF similarity values of comments to its root
	tfidf_p	average TF-IDF similarity values of comments to its parent
Content (Text)	wv_p	mean word vectors of a title of post
	wv_t	mean word vectors of all messages in an initial tree
User (Section 4.1.2)	$gini_{inc}$	Gini coefficient of trees (including root node)
	$gini_{exc}$	Gini coefficient of trees (excluding root node)
	um	user-message ratio
	$recip$	reciprocal edge ratio
	rcv_{post}	received comment ratios by U_{post}
	rcv_{cmt}	received comment ratios by U_{cmt}
	rcv_{rcvcmt}	received comment ratios by U_{rcvcmt}
	rcv_{uni}	received comment ratios by U_{uni}
Community (Section 4.2)	-	one-hot encoded subreddit vector
All	-	Combination of content and user feature groups.

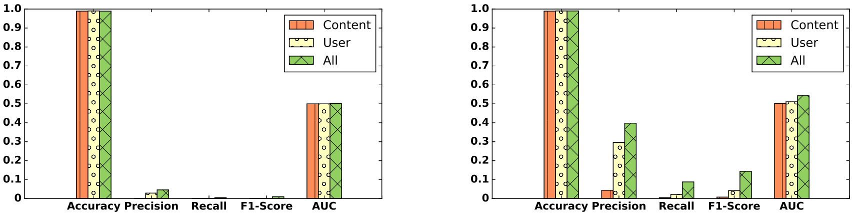
Table 7.1 The features used for the prediction task are summarized.

of words, filter out stopwords, and encode each word with the corresponding word vector using the Glove [61], a set of pre-trained word vectors based on Wikipedia. We finally compute the mean word vectors for the given text (denoted by wv). All the features used in our model are summarized in Table 7.1. Note that we use the features extracted from only posts (i.e., features with the subscript p) in the first scenario (when only post information is available, $m = 0$) while all the post and initial features in Table 7.1 are used in the second scenario (when an initial conversation can be observable, $m > 0$).

Hyperparameters: In our experiment, there are two hyperparameters, m and k, each of which represents the number of initial comments and the percentage of selected (top) comment trees, respectively. We first choose $m = 0, 1, 2, 3$ to examine the effect of the amount of observable information on the prediction. That is, the smaller m assumes the case that predicts large or viral conversations in earlier time, which implies less informative. Note that we only consider the comment trees that include more than 2 comments in our experiment. We set $k = 1$ for predicting extremely large or viral (i.e., top 1%) conversations.

Classifiers: We build the prediction model based on four well-known classifiers: SVM, Random Forest, Neural Network, and Logistic Regression with L1 penalty function. Since the performances of the four classifiers are mostly similar, we only report the results of the Logistic Regression classifier, which slightly outperforms others.

Sampling: Identifying the top k% comment trees naturally leads to a *class imbalance problem* [12, 44]. That is, the classifier may be biased towards the major class, which may result in low performance. We first report the performance results (in terms of accuracy, precision, recall, F1-score, and AUC) based on the original population (i.e., 1:99 for the top and non-top trees, respectively) in Figure 7.1. As shown in Figure 7.1, the models show significantly poor performance.



(a) Volume

(b) Virality

Figure 7.1 The performance results of prediction models based on the original population are plotted.

To cope with such a problem, we adopt a sampling technique to balance the number of instances between the top (i.e., minor class) and non-top (i.e., major class) trees. In particular, we apply the SMOTE [12] technique to oversample instances from the minor class (i.e., top trees) and randomly sample instances from the major class (i.e., non-top trees) with the same number of instances sampled from the minor class.

Evaluation Metrics: To evaluate the proposed model, we adopt the AUC, the Area Under the receiver operating characteristic (ROC) curve, which shows the relation between true-positive rate (TPR) and false-positive rate (FPR) [26]. If a classifier shows similar performance with random guessing for the balanced dataset, AUC is close to 0.5. On the other hands, the AUC score will be close to 1 if the classifier can correctly predict the large or viral conversations.

7.1.3 Performance Analysis

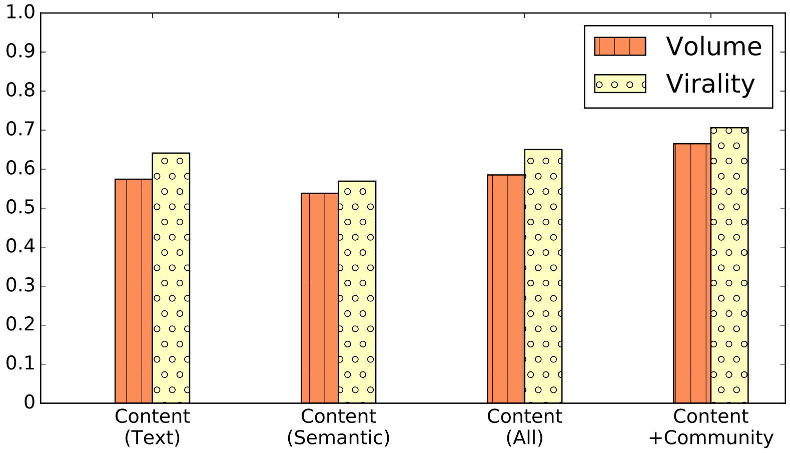


Figure 7.2 The performance results of using only post features ($m = 0$) is plotted.

Scenario 1 – when only post information is available ($m = 0$): Figure 7.2 shows the performance results of the proposed models with $m = 0$, which assumes the case when only post information is available. Note that we combine two feature groups, “Content (Text)” and “Content (Semantic)”, denoted as “Content (All)”. Note that “Content + Community” includes all the features (except the user features) described in Table 7.1. As shown in Figure 7.2, our proposed models using only post features can identify large or viral conversations with better performance than expected AUC of random guessing (0.5), meaning that the content features of posts can be a good predictor in identifying large or viral conversations. In particular, the model using text features outperforms the one using semantic features in both volume and virality cases, meaning that text features by word embedding play more important roles. The models with combined features (“Content (All)” or “Content (All) + Community”) show better performance than the models using only one feature group (e.g,. Content (Text) or Content (Semantic)), implying that those features are complementary to each other. Note that the AUCs of the virality case are higher than the ones of volume case, signifying that predicting viral conversations is relatively easier than large conversations.

Scenario 2 – when an initial conversation can be observable ($m > 0$): Figure 7.3 shows the performance results of using the features from both post and initial comments ($m = 3$). Overall, the model using both post and initial comments outperforms than the models using only post features, meaning that observing initial comments can improve the model performance. Interestingly, feature groups play different roles in predicting large and viral conversations. The AUC of the model using user features is higher than the model using text and semantic content features in the case of predicting large

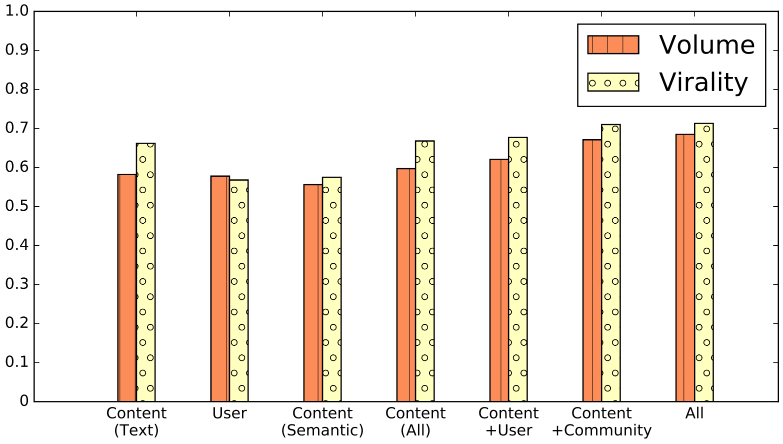


Figure 7.3 The performance results of observing both post and initial comments ($m = 3$) is described.

conversation while text content features show better performance than user and semantic content features in predicting viral conversations. This result implies that user participation behaviors are important for predicting large conversations whereas content itself is an important predictor for identifying viral conversations.

We next investigate how amount of initial observation can improve the prediction of performance. To this end, we perform three more analysis with the different numbers of initial comments. Figure 7.4 shows the AUC scores in the case of that the numbers of the initial comments are 1, 5, and 10. Intuitively, the AUC scores increase as the numbers of initial comments increase, implying that observing more information improves the performance of the prediction. Interestingly, the gap of AUC scores between 1 and 5 is larger than 5 and 10, meaning that the very initial comments are more important to predict large and viral conversations. Note that AUC scores for virality are

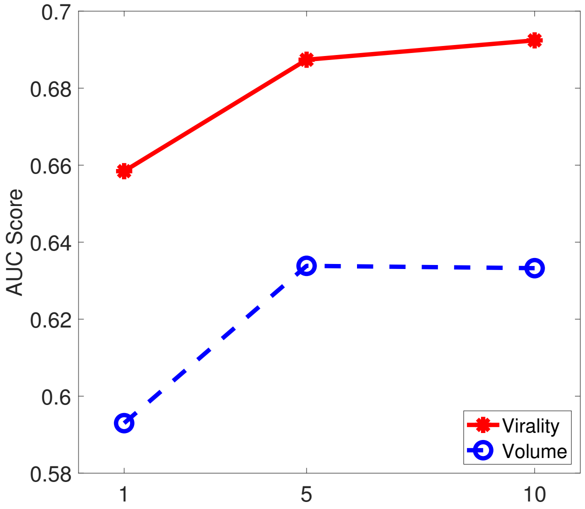


Figure 7.4 The AUC scores with different count of initial comments are plotted.

higher than the ones of volume, meaning that predicting viral conversation is relatively easier.

7.2 Popular and Viral Image Prediction

We have revealed that popular images are not necessarily viral in Pinterest, which motivates us to study whether there are distinctive features to accurately predict popular and viral images, respectively. In this section, we aim to predict popular or viral image cascades. In particular, we identify popular cascades whose volumes are higher than 230 and 74, which account for the top 0.1% and 1% of all the image cascades (in terms of volume or cascade size), respectively. We also identify 'viral' cascades whose structural virality (or WIs) are higher than 6.32 and 3.85, which account for the top 0.1% and 1% of all the image cascades (in terms of WIs), respectively. Note that only a small portion of the top popular and viral cascades are overlapped as shown in the previous section; only 17.4% of the cascades are overlapped between the volume-based top 1% cascades and the virality-based top 1% cascades.

We cast this problem as a supervised learning problem, where we observe a set of features of a cascade and predict whether the given cascade belongs to the top popular or viral cascades. We build a learning model based on the Random Forest ensemble algorithm [7]. We used other classifiers including support vector machine or logistic regression, but we only report the results of the Random Forest ensemble classifier as it performs better than others. We report the following performance metrics: (i) accuracy ($ACC = \frac{TP+TN}{TP+FP+FN+TN}$), (ii) true positive rate or recall ($TPR = \frac{TP}{TP+FN}$), (iii) false positive rate ($FPR = \frac{FP}{FP+TN}$), and (iv) area under the receiver operating characteristics (ROC) curve (AUC)[1] [26], where TP, FP, FN, and TN represents the true positive, false positive, false negative, and true negative, respectively. We perform a 10-fold cross-validation.

Predicting the top popular or viral cascades can be suffered from the class imbalance problem, e.g., the ratio between the minority and majority classes for identifying the top 0.1% cascades is 1:999. To remedy this issue, we apply the Synthetic Minority Over-sampling TEchnique (SMOTE) [13], which allows us to learn with over-sampled instances from the minority class (i.e., top cascades). We learn randomly under-sampled instances from the majority class (i.e., non-top cascades). We varied the sampling ratios of minority and majority classes, from 1:1 to 1:2 to 1:4 to 1:8, but we only report the results of 1:1 ratio as it shows a similar performance with others.

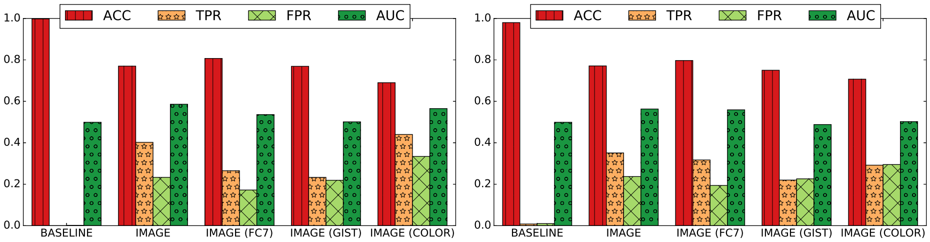
We consider different scenarios in predicting the popular or viral image

[1] AUC indicates the effectiveness of the given model. A perfect model has an AUC of 1, while a random model generates an AUC of 0.5.

cascades. In particular, we answer the following questions: (1) If only an image is given (and available), can we predict whether it becomes popular or goes viral?, (2) At the moment when an image is posted and its meta and/or poster (or pinner) information is available, can we predict whether the image becomes popular or goes viral?, and (3) If an initial image propagation pattern is observable, does it help to predict popular and/or viral image cascades? Answering these questions can give important insight into predicting popular or viral image diffusion for content providers, OSN operators, and marketers.

7.2.1 Predictive Power of Image Itself

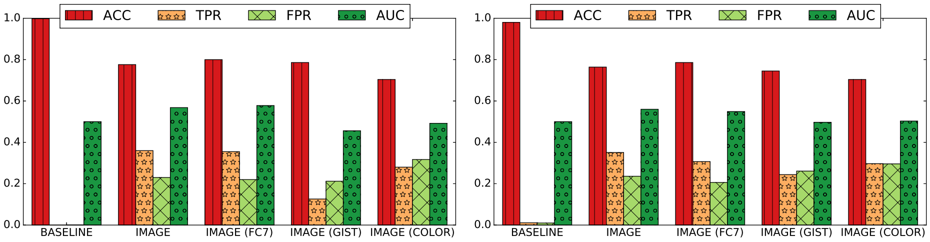
We first study the role of image content in its popularity and virality prediction without using other features such as poster or posting information. This assumes the situation where (i) pinner or posting information is not available (e.g., due to privacy issues) or (ii) the image is not yet posted by anyone. To this end, we extract features from an image using the deep learning technique. We adopt a well-known visual categorization developed for the task of image classification and feature learning, ImageNet [23], which defines 1000 image classes (mostly object classes). We use a convolutional neural network (CNN) [47], which is known as very effective for visual feature learning. Our model architecture is the VGG-16 [65] and we use a publicly available pre-trained model on the Imagenet data. For each image, we extract the final image features at 1000 category level (referred to as 'IMAGE') as well as intermediate 4096 features at the last fully-connected layer ($FC7$) (referred to as 'IMAGE(FC7)'). In addition to high-level features ('IMAGE' and 'IMAGE(FC7)'), we further consider the following low-level features: (i)



(a) Top 0.1%

(b) Top 1%

Figure 7.5 Prediction results on popular cascades using image features.



(a) Top 0.1%

(b) Top 1%

Figure 7.6 Prediction results on viral cascades using image features.

512 'gist' image features [58] which describe gradient-based (Gabor filters) scene features such as texture or edge (referred to as 'IMAGE(gist)') and (ii) the mean and standard deviation of image color in RGB (referred to as 'IMAGE(color)'). Based on the extracted image features, our classifier (i.e., the Random Forest ensemble) identifies whether the given image belongs to the top popular and/or viral cascades.

Figures 7.5 and 7.6 show the prediction results on popular and viral cascades, respectively, using image features. For a comparison purpose, we include the result of a null model, 'BASELINE'. Since the 'BASELINE' model predicts the popular or viral cascades according to their distributions, the *ACCs* are 99.9% and 99% for predicting top 0.1% and 1%, respectively. Note that the 'BASELINE' has *AUC* of 0.5. As shown in Figures 7.5 and 7.6, the models using image features ('IMAGE', 'IMAGE(FC7)', 'IMAGE(gist)', and

'IMAGE(color)') perform slightly better than 'BASELINE', but their *AUCs* are mostly lower than 0.55, implying that using image features alone is not as effective in predicting popular or viral image cascades. In other words, popular or viral image cascades are not predictable using only image features. Note that high-level features ('IMAGE' and 'IMAGE(FC7)') perform slightly better than low-level features ('IMAGE(gist)' and 'IMAGE(color)'). We find that 'IMAGE' mostly performs better than other image feature sets, hence we only consider 'IMAGE' features hereafter.

To identify the specific features that contribute most towards predicting top 1% popular and viral cascades based on 'IMAGE' features, respectively, we apply the *Chi-squared* (χ^2) statistic evaluation [52] to all of the 'IMAGE' features, which results in assigning a score to each feature. We rank the features according to the χ^2 values. The top 3 features for predicting top 1% popular cascades are 'menu', 'brassiere', and 'binoculars', while those for predicting top 1% viral cascades are 'menu', 'binoculars', and 'plate'.

7.2.2 Predictive Power of Image Meta and Pinner Information

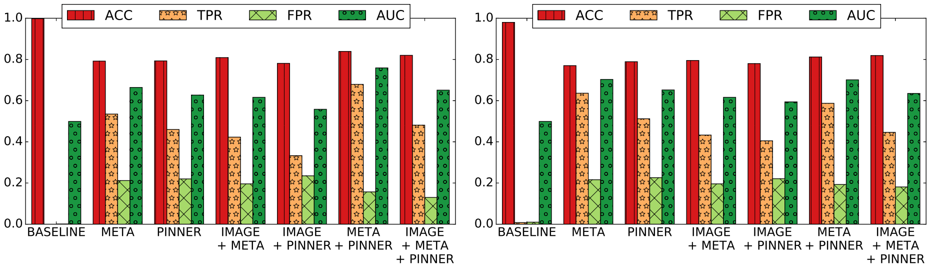
We next investigate whether image meta and/or poster (or pinner) information is useful in predicting popular or viral image cascades. For the meta information of a pin (referred to as 'META'), we consider the following features: (i) category popularity (i.e., number of pins) where the given pin belongs, (ii) source popularity (i.e., number of pins) where the given pin comes, (iii) maliciousness of the pin, and (iv) revealed sentiment from the pin's title and description. For detecting the maliciousness of a pin, we submit the source (i.e., URL where the pin comes) of the pin to a commercial URL scanner, VirusTotal [69], which scans a submitted URL over a corpus of over 60 website scanning engines. We identify each source as malicious if two or more security engines indicate it malicious. To calculate the revealed sentiment of a pin, we use LIWC (Linguistic Inquiry and Word Count), which counts words into psychologically meaningful categories [60]. We calculate the positive, negative, cognitive, and social scores of each pin's title and description using LIWC.

We also consider the following characteristics of a pinner (referred to as 'PINNER'): (i) number of pins the pinner has, (ii) number of followers who follow the pinner, (iii) number of followees the pinner follows, (iv) number of likes the pinner likes, (v) number of boards the pinner has, (vi) number of categories the pinner has, and (vii) category entropy of the pinner. A category entropy quantifies how a user's interest (pinning/repinning) is distributed across multiple categories. We calculate the category entropy of user u as follows:

$$H_{category}(u) = -\sum_{i=1}^{C_u} p_i^u ln(p_i^u) \tag{7.2}$$

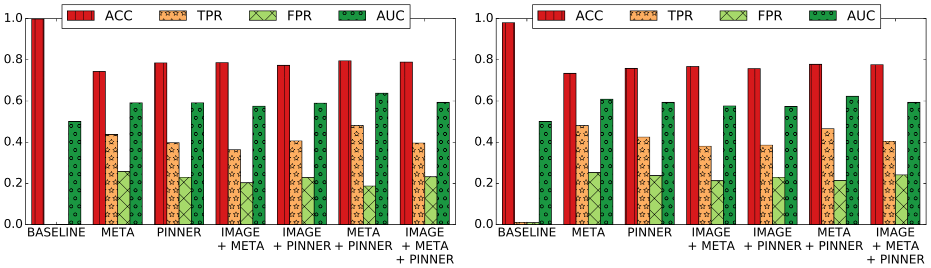
where C_u is the number of categories the user u has, and p_i^u is the portion of pins/repins in the category i by u.

Figures 7.7 and 7.8 show the prediction results on popular and viral cascades, respectively, using image meta and/or pinner information. To investigate whether there is a synergy among different feature sets, we also consider (i) image and image meta features ('IMAGE+META'), (ii) image and pinner features ('IMAGE+PINNER'), (iii) image meta and pinner features ('META+PINNER'), and (iv) image, image meta, and pinner features ('IMAGE+META+PINNER').



(a) Top 0.1% (b) Top 1%

Figure 7.7 Prediction results on popular cascades using image meta and/or pinner information.



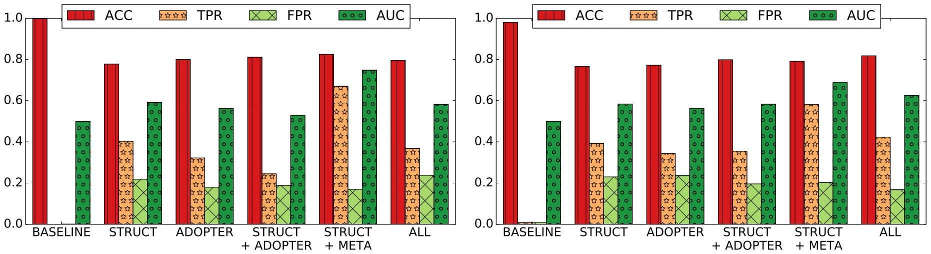
(a) Top 0.1% (b) Top 1%

Figure 7.8 Prediction results on viral cascades using image meta and/or pinner information.

As shown in Figures 7.7 and 7.8, 'META' and 'PINNER' performs better than 'BASELINE', meaning that meta information of an image as well as its poster's information are useful in predicting both popular and viral image cascades. The image meta information shows a stronger predictive power than pinner's information in predicting popular image cascades, which implies that information about the image is more important than information of a user who posts the image. On the other hand, image meta information performs slightly better (or similarly) than pinner's information in predicting viral image cascades, implying that viral image cascades are similarly associated with image meta and pinner information. If we consider both of image meta and pinner features, i.e., 'META+PINNER', it performs better than others in Figures 7.7 and 7.8, which signifies that image meta and pinner features are complementary to each other. Note that the AUC of 'META+PINNER' for predicting the top 0.1% popular cascades is 0.76, which is much higher than 'BASELINE'. The top 3 features (ranked by the χ^2 values) in 'META' for predicting top 1% popular cascades are 'social sentiment score', 'positive sentiment score', and 'maliciousness', while those for predicting top 1% viral cascades are 'maliciousness', 'source popularity', and 'cognitive sentiment score'. On the other hand, the top 3 features in 'PINNER' for predicting top 1% popular cascades are 'number of followers', 'social sentiment score', and 'number of categories', while those for predicting top 1% viral cascades are 'cognitive sentiment score', 'negative sentiment score', and 'positive sentiment score'. Interestingly, combining 'IMAGE' features (e.g., 'IMAGE+META', 'IMAGE+PINNER', 'IMAGE+META+PINNER') do not contribute much in predicting popular and viral image cascades. It is worth noting that the prediction performance of popular cascades is higher than that of viral cascades, which signifies that image meta and pinner information are more useful in predicting popular image cascades than viral image cascades.

7.2.3 Predictive Power of Initial Propagation Pattern

We finally examine whether the observation of initial image propagation pattern helps to predict popular or viral image cascades. That is, we observe the first k repins of an image cascade, and predict whether the cascade will belong to the top popular or viral cascades in the future. Note that higher k shows better performance, but we only report $k = 5$ here since our goal is to observe the propagation pattern in the very early stage of a cascade. We consider two perspectives of initial propagation of a cascade: (i) how the



(a) Top 0.1%

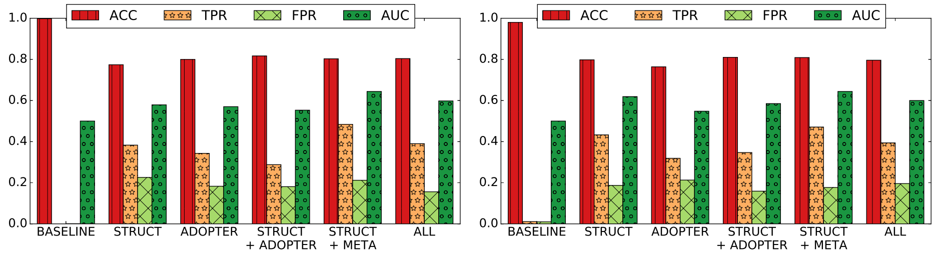
(b) Top 1%

Figure 7.9 Prediction results on popular cascades using initial propagation pattern.

cascade initially looks like (referred to as ‘STRUCT’) and (ii) who are the early adopters in the cascade (referred to as ‘ADOPTER’).

The ‘STRUCT’ features consist of (i) max width, (ii) max depth, (iii) wiener index based on Equation 3.1 (when $N = 6$), (iv) width entropy quantifies how the distribution of widths in a cascade is even or skewed (calculated similarly with Equation 7.2), (v) inter-repin times for the five repins, and (vi) positive, negative, cognitive, and social sentiment scores for each repin’s description using LIWC. The ‘ADOPTER’ consists of each repinner’s following characteristics: (i) number of pins the repinner has, (ii) number of followers who follow the repinner, (iii) number of followees the repinner follows, (iv) number of likes the repinner likes, (v) number of boards the repinner has, (vi) number of categories the repinner has, (vii) category entropy of the repinner (Equation 7.2), and (viii) positive, negative, cognitive, and social sentiment scores of the repinner’s introduction text.

Figures 7.9 and 7.10 show the prediction results on popular and viral cascades, respectively, using the initial propagation pattern. To investigate whether there is a synergy among different feature sets, we also consider (i) ‘STRUCT+ADOPTER’, (ii) ‘STRUCT+META’, and (iii) ‘ALL’ that in-



(a) Top 0.1%

(b) Top 1%

Figure 7.10 Prediction results on viral cascades using initial propagation pattern.

cludes all the features. As shown in Figures 7.9 and 7.10, ‘STRUCT’ performs better than ‘ADOPTER’, meaning that the initial propagation shape of the cascade is a stronger predictor than the information of users who initially participate in the cascade. The ‘STRUCT+ADOPTER’ performs worse than ‘STRUCT’, meaning that ‘ADOPTER’ may not contribute much in predicting popular and viral image cascades. Note that the top 3 features (ranked by the χ^2 values) in ‘STRUCT’ for predicting top 1% popular cascades are ‘wiener index’, ‘social sentiment score’, and ‘positive sentiment score’, while those for predicting top 1% viral cascades are ‘width entropy’, ‘max depth’, and ‘repin time’. On the other hand, if we combine the ‘STRUCT’ and ‘META’ features, it performs better than others in Figures 7.9 and 7.10, which implies that ‘META’ and ‘STRUCT’ features are complementary to each other. Note that ‘STRUCT+META’ are not mostly about *‘human factors’* but more about *‘content factors’*, implying that content factors are important predictors in predicting popular and viral image cascades.

In summary, ‘META+PINNER’ is the strongest predictor in predicting popular image cascades while ‘STRUCT+META’ is the strongest predictor in predicting viral image cascades. This implies that we can forecast popular image cascades using the image meta and pinner information at the moment when an image is posted. Also, if we observe initial propagation pattern of an image cascade, we can predict whether the image goes viral based on its meta information and initial propagation pattern. It is worth to note that 'META' information is commonly useful for predicting both popular and viral image cascades.

Chapter 8

Conclusion

This thesis presents a comprehensive large-scale measurement study on three different datasets collected from Reddit, Pinterest, and Bitly through the analysis architecture and methodology introduced in Chapter 3. Using the comment tree model, the online threaded conversations are characterized in terms of volume, responsiveness, and virally, and how content, user behavior, and community factors are associated with the characteristics are investigated (Chapter 4). As analysis result, this thesis reports that difficulties of texts, portion of reciprocal communications, types of communities (e.g., multimedia-related, or discussion-encouraged communities) are highly associated with the large, responsive, or viral conversations. Interestingly, the difficulty of content texts is an important indicator that can differentiate large/viral and responsive conversations; a large/viral conversation is likely to have difficult texts, whereas a responsive conversation tends to have plain texts. The further analysis on characteristics of comment trees across different topical communities indicates that the news-related and image-based subreddits are more likely to have large and responsive conversations, respectively, and the conversations in discussion-driven subreddits tend to be viral, The characteristics of image cascades are investigated in Chapter 5. In particular, this thesis focuses on how micro-level virality (i.e., Weiner Index) is different from volume, revealing that popular images are not necessarily viral and structurally viral image cascades are propagated more quickly, compared to broadcast-shape large image cascade. To understand how web content URLs such as image, video, or web pages is shortened, published, and accessed across different types of domains, the large-scale URL request logs are analyzed in Chapter 6. By modeling the relations among websites, this thesis shows that domains play different roles in publishing and sharing short URLs. For example, adult or malicious content tend to be requested from search engines, shopping content is primarily accessed through online social networks, and news content are usually clicked through computer & electronics websites. This thesis also revealed that news or shopping content, published through online social networks, tend to be requested quickly and virally. The learning-based model proposed in Chapter 7 can predict whether a conversation or an image cascade would be large or viral in very early time, with high performance.

The findings revealed in this thesis are expected to give an important insight for online marketing or novel platform design. For example, the evidence that image-only features are not effective to predict whether the give image will be viral may drive online marketers who want to propagate her content to crowds to use not only image alone, but text information, through influential users in the platform. Furthermore, the machine learning-based application proposed in this thesis shows the possibility of improvement of platform of interpretable AI in practice; the platform designer may use the key features (e.g., meta or community information) reported in this thesis for more effective access (e.g., caching) of the uploaded content that is expected to be viral.

Although the exploration in this thesis is expected to give an important insight for content providers, OSN operators, and marketers in predicting popular or viral content diffusion, this thesis has several limitations. Firstly, generalizing the findings and results to other OSNs or specific subreddits should be cautioned since the results of the analysis in this thesis are involved in three online services. However, the methods used in this thesis are encouraged to leverage understanding user behavior and content propagation patterns in other online services. Secondly, although the proposed model in this thesis indicates micro-level relation of content propagation and online conversation, the characteristics analyzed in this thesis are somewhat macro-level numeric metric. (e.g., volume, virality). Comprehensive analysis of more specific micro-level analysis such as characterizing and predicting *path* of content dissemination or user-to-user interactions can give important implication to better understand user behaviors or content propagation patterns. Lastly, the state-of-art machine learning techniques such as deep learning can improve the performance of the proposed prediction model and even more challenging prediction problems can be resolved through the techniques. As part of addressing these limitations, the further (path-level) analysis on content propagation or online conversations are studied and deep-learning-based application to infer next path of content propagation will be designed in the future work.

Bibliography

[1] A. Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39(1):45–65, 2003.

[2] D. Antoniades, I. Polakis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis. we.b: The web of short URLs. *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*, 2011.

[3] S. Aral and D. Walker. Identifying Influential and Susceptible Members of Social Networks. *Science*, 337(6092):337–341, 2012.

[4] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. *Proceedings of the 21st International World Wide Web Conference (WWW 2012)*, 2012.

[5] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[6] S. Barbosa, D. Cosley, A. Sharma, and R. M. C. Jr. Averaging gone wrong: using time-aware analyses to better understand behavior. *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, 2016.

[7] L. Breiman. Random Forests. *Springer Machine learning*, 45(1):5–32, 2001.

[8] C. Budak, D. Agrawal, and A. El Abbadi. Where the blogs tip: connectors, mavens, salesmen and translators of the blogosphere. *The Workshop on Social Media Analytics*, 2010.

[9] J. Campbell. Top 12 websites in the world - then and now: 2012, 2007, 2001. `https://www.socialtalent.com/blog/recruitment/top-12-websites-in-the-world-then-and-now-2012-2007-2001`, 2012. Online; accessed 20-Oct-2016.

[10] Can ben silbermann turn pinterest into the world's greatest shopfront? `http://www.fastcodesign.com/1670681/ben-silbermann-pinterest`.

[11] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. *Proceedings of the 18th International World Wide Web Conference (WWW 2009)*, 2009.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

[14] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec. Do cascades recur? *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[15] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? *Proceedings of the 23rd International Conference on World Wide Web Conference (WWW 2014)*, 2014.

[16] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/$ocial: the phishing landscape through short urls. *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, 2011.

[17] D. Choi, J. Han, T. Chung, Y.-Y. Ahn, B.-G. Chun, and T. T. Kwon. Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors. *ACM Conference on Online Social Networks (COSN)*, 2015.

[18] D. Choi, J. Han, T. Chung, Y.-Y. Ahn, B.-G. Chun, and T. T. Kwon. Characterizing conversation patterns in reddit: from the perspectives of content properties and user participation behaviors. *Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN 2015)*, 2015.

[19] M. D. Choudhury and S. De. Mental health discourse on reddit: self-disclosure, social support, and anonymity. *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.* 2014.

[20] C. Dagum. The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquée*, 33(2), 1980.

[21] P. T. Danish Y. D. Discovering response-eliciting factors in social question answering: a reddit inspired study. *The 10th International AAAI Conference on Web and Social Media*, 2016.

[22] S. Das and A. Lavoie. The effects of feedback on human behavior in social media: an inverse reinforcement learning model. *The International Conference on Autonomous Agents and Multi-agent Systems*, 2014.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *IEEE CVPR*, 2009.

[24] A. Deza and D. Parikh. Understanding Image Virality. *IEEE CVPR*, 2015.

[25] DGTraffic. Indonesia internet users. `http://www.dgtraffic.com/indonesia-internet-users/`, 2012. Online; accessed 20-Oct-2016.

[26] T. Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[27] B. Gelley and A. John. Do i need to follow you?: examining the utility of the pinterest follow mechanism. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work (CSCW 2014)*, 2015.

[28] B. Gelley and A. John. Do I Need To Follow You?: Examining the Utility of The Pinterest Follow Mechanism. *ACM CSCW*, 2015.

[29] E. Gilbert. Widespread underprovision on reddit. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 2013.

[30] E. Gilbert, S. Bakhshi, S. Chang, and L. Terveen. "i need to try this!": a statistical overview of pinterest. *ACM CHI*, 2013.

[31] S. Goel, A. Anderson, J. Hofman, and D. J. Watts. The Structural Virality of Online Diffusion. *Management Science*:1–17, 2015.

[32] S. Goel, A. Anderson, J. Hofman, and D. J. Watts. The structural virality of online diffusion. *Management Science*, 2015.

[33] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. *ACM Conference on Electronic Commerce (EC 2012)*, 2012.

[34] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. *Proceedings of the 17th International Conference on World Wide Web Conference (WWW 2008)*, 2008.

[35] M. Guerini, J. Staiano, and D. Albanese. Exploring Image Virality in Google Plus. *IEEE International Conference on Social Computing*, 2013.

[36] R. Gunning. *The Technique of Clear Writing.* McGraw-Hill, 1952.

[37] N. Gupta, A. Aggarwal, and P. Kumaraguru. Bit.ly/malicious: deep dive into short URL based e-crime detection. *CoRR*, abs/1406.3687, 2014.

[38] J. Han, D. Choi, A.-Y. Choi, J. Choi, T. Chung, T. ‘ Kwon, J.-Y. Rha, and C.-N. Chuah. Sharing Topics in Pinterest: Understanding Content Creation and Diffusion Behaviors. *ACM Conference on Online Social Networks (COSN)*, 2015.

[39] J. Han, D. Choi, A.-Y. Choi, J. Choi, T. Chung, T. T. Kwon, J.-Y. Rha, and C.-N. Chuah. Sharing topics in pinterest: understanding content creation and diffusion behaviors. *Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN 2015)*, 2015.

[40] J. Han, D. Choi, B.-G. Chun, T. Kwon, H.-C. Kim, and Y. Choi. Collecting, organizing, and sharing pins in pinterest: interest-driven or social-driven? *Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS 2014)*, 2014.

[41] J. Han, D. Choi, J. Joo, and C.-N. Chuah. Predicting popular and viral image cascades in pinterest. *International AAAI Conference on Web and Social Media (ICWSM)*, 2017.

[42] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[43] G. Hsieh, Y. Hou, I. Chen, and K. N. Truong. “welcome!”: social and psychological predictors of volunteer socializers in online communities. *Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work*, 2013.

[44] N. Japkowicz and S. Stephen. The class imbalance problem: a systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002.

[45] A. Khosla, A. D. Sarma, and R. Hamid. What Makes an Image Popular? *WWW*, 2014.

[46] F. Klien and M. Strohmaier. Short links under attack: geographical analysis of spam in a url shortener network. *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT 2012)*, 2012.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems (NIPS)*, 2012.

[48] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, 2010.

[49] H. Lakkaraju, J. McAuley, and J. Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[50] 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. V. Alstyne. Social science: computational social science. *Science*, 323(5915):721–723, Feb. 2009.

[51] A. Leavitt and J. A. Clark. Upvoting hurricane sandy: event-based news production processes on a social news site. *ACM CHI*, 2014.

[52] H. Liu and R. Setiono. Chi2: Feature Selection and Discretization of Numeric Attributes. *IEEE International Conference on Tools with Artificial Intelligence*, 1995.

[53] F. Maggi, A. Frossi, S. Zanero, G. Stringhini, B. Stone-Gross, C. Kruegel, and G. Vigna. Two years of short urls internet measurement: security threats and countermeasures. *Proceedings of the 22nd International Conference on World Wide Web (WWW 2013)*, 2013.

[54] M. Marcoccia. On-line polylogues: conversation structure and participation framework in internet newsgroups. *Journal of Pragmatics*, 36(1):115–145, 2004.

[55] E. Mayfield, D. Adamson, and C. P. Rosé. Hierarchical conversation structure prediction in multi-party chat. *The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012.

[56] A. Newman. Bitly helps the red cross get to hope.ly - 2014, 2014. `https://www.nytimes.com/2014/12/02/business/media/bitly-helps-the-red-cross-get-to-hopely.html?_r=0`.

[57] N. Nikiforakis, F. Maggi, G. Stringhini, M. Z. Rafique, W. Joosen, C. Kruegel, F. Piessens, G. Vigna, and S. Zanero. Stranger danger: exploring the ecosystem of ad-based url shortening services. *Proceedings of the 23rd International Conference on World Wide Web Conference (WWW 2014)*, 2014.

[58] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[59] R. Ottoni, J. P. Pesce, D. Las Casas, G. Franciscani Jr, W. Meira Jr, P. Kumaraguru, and V. Almeida. Ladies first: analyzing gender roles and behaviors in pinterest. *ICWSM*, 2013.

[60] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The development and psychometric properties of liwc2007, 2007. `http://www.liwc.net/LIWC2007LanguageManual.pdf`.

[61] J. Pennington, R. Socher, and C. D. Manning. Glove: global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[62] M. R. Rahman, J. Han, and C.-N. Chuah. Unveiling the Adoption and Cascading Process of OSN-based Gifting Applications. *IEEE INFOCOM*, 2015.

[63] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC 2011)*, 2011.

[64] L. Rossi and M. Magnani. Conversation practices and network structure in twitter. *ICWSM*, 2012.

[65] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[66] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier. Evolution of reddit: from the front page of the internet to a self-referential community? *Proceedings of the 23rd International Conference on World Wide Web*, 2014.

[67] C. Tan and L. Lee. All who wander: on the prevalence and characteristics of multi-community engagement. *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[68] L. C. Totti, F. A. Costa, S. Avila, E. Valle, W. Meira J., and V. Almeida. The Impact of Visual Attributes on Online Image Diffusion. *ACM Conference on Web Science*, 2014.

[69] VirusTotal. VirusTotal - Free Online Virus, Malware and URL Scanner, 2016. `https://www.virustotal.com`.

[70] C. Wang, M. Ye, and B. A. Huberman. From user comments to online conversations. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012)*, 2012.

[71] D. Wang, S. B. Navathe, L. Liu, D. Irani, A. Tamersoy, and C. Pu. Click traffic analysis of short url spam on twitter. *Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, 2013.

[72] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabási. Information spreading in context. *WWW*, 2011.

[73] T. Weninger. An exploration of submissions and discussions in social news: mining collective intelligence of reddit. *Social Network Analysis and Mining*, 4(1), 2014.

[74] Why pinterest is 2012's hottest website. `http://edition.cnn.com/2012/02/06/tech/web/pinterest-website-cashmore`.

[75] H. Wiener. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society*, 69(1):17–20, 1947.

[76] C. Zhong, D. Karamshuk, and N. Sastry. Predicting Pinterest: Automating a Distributed Human Computation. *WWW*, 2015.

초록

사회 관계망 서비스, 소셜 미디어, 게시판 등 다양한 온라인 서비스의 발달로 한 사람이 다른 사람들과 다양한 채널을 통해 의사소통을 하는 것이 일반화 되었다. 이러한 온라인 디지털 채널들이 사용자들의 의사소통에 관련된 많은 데이터를 축적해 옴에 따라, 데이터에 기반하여 사람들의 행동이나 의사소통 방식을 모델링, 분석하고 예측하는 연구가 가능하게 되었다. 본 학위 논문에서는 이러한 연구의 한 부분으로 다음과 같은 데이터 기반 분석을 수행한다.: (i) 사용자 행동, 콘텐트, 사용자 집단 특성에 기반한 온라인 대화 패턴 분석, (ii) 인기있고 전염성 높은 (viral) 이미지 전파 특성 분석 및 예측, (iii) 온라인 콘텐트의 게시 및 소비 등 유통 흐름에 대한 분석. 이를 위해, (i) 약 150만 명의 레딧 유저로부터 생성된 70만개의 온라인 대화, (ii) 핀터레스트 내에 유포된 약 33만 개의 이미지 및 전파 데이터, (iii) Bitly를 통해 게시된 약 8천만개의 짧은 URL 및 42억개의 요청 데이터셋을 수집하고 분석한다. 이러한 분석들을 통해, 콘텐트, 사용자의 행동특성 및 집단적 특성이 각각 크고, 반응적이고, 전염적인 온라인 대화와 관련이 있음을 밝혀내었으며, 핀터레스트 데이터셋에 기반한 분석을 통해 이미지 전파에서 구조적 전염도 (Structural virality)가 단순히 큰 전파와 전파 모양 측면에서 차이가 있음을 밝혀내었다. 또한, Bitly 데이터셋에 기반하여 콘텐트와 리퍼러 (Referrer) 도메인 간의 관련성을 모델링함으로써, 서비스 별 특성 (뉴스피드, 스트리밍, 온라인 쇼핑 등) 에 따라 콘텐츠 게시 및 소비 패턴이 다름을 입증하였다. 이러한 발견들에 기반하여, 최종적으로 하나의 온라인 대화나 이미지 콘텐트가 커질지 혹은 전염적으로 확산될지를 예측하기 위한 기계학습 기반 모델을 제안하였다. 본 논문에서 제안된 모델은 최초에 관측된 코멘트 혹은 이미

지 전파 패턴, 사용자의 행동 특성, 콘텐트의 특성을 모두 활용하여 높은 확률로 크거나 전염성이 높은 대화 및 이미지 전파를 예측할 수 있었다. 본 학위 논문을 통해 발견된 현상 및 예측 모델은 온라인 사회 관계망 서비스 제공자, 마케터, 콘텐트 제공자 등 정보나 콘텐츠의 확산을 목적으로 하는 사람들은 물론, 전파 패턴이나 확산 규모 등에 대한 해석가능한 인공지능 모델을 개발하는데 있어서 큰 기여를 할 수 있을 것으로 기대한다.

주요어: 온라인 대화, 콘텐트 전파, 기계 학습, 데이터 과학, 레딧, 핀터레스트, 비틀리

학번: 2012-23243