Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors

Daejin Choi
Seoul National University
djchoi@mmlab.snu.ac.kr

Jinyoung Han
University of California, Davis
rghan@ucdavis.edu

Taejoong Chung
Seoul National University
tjchung@mmlab.snu.ac.kr

Yong-Yeol Ahn
Indiana University
yyahn@indiana.edu

Byung-Gon Chun
Seoul National University
bgchun@snu.ac.kr

Ted "Taekyoung" Kwon
Seoul National University
tkkwon@snu.ac.kr

ABSTRACT

It becomes the norm for people to communicate with one another through various online social channels, where different conversation structures are formed depending on platforms. One of the common online communication patterns is a threaded conversation where a user brings up a conversation topic, and then other people respond to the initiator or other participants by commenting, which can be modeled as a tree structure. This paper seeks to investigate (i) the characteristics of online threaded conversations in terms of volume, responsiveness, and virality and (ii) what and how content properties and user participation behaviors are associated with such characteristics. To this end, we collect 700 K threaded conversations from 1.5 M users in Reddit, one of the most popular online communities allowing people to communicate with others in the form of threaded conversations. Using the collected dataset, we find that 'social' words, difficulties of texts, and document relevancy are associated with the volume, responsiveness, and virality of conversations. We also discover that large, viral conversations are mostly formed by a small portion of users who are reciprocally communicate with others by analyzing user interactions. Our analysis on discovering user roles in conversations reveal that users who are interested in multiple topics play important roles in large and viral conversations, whereas heavy posting users play important roles in responsive conversations. We expand our analysis to topical communities (i.e., subreddits) and find that news-related, image-based, and discussion-related communities are more likely to have large, responsive, and viral conversations, respectively.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; H.4.3 [**Communications Applications**]: Bulletin boards; J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Reddit; Online Communication; Threaded Conversation; Comment; Subreddits; User Behavior; Virality;

COSN'15, November 2–3, 2015, Palo Alto, California, USA.

DOI: http://dx.doi.org/10.1145/2817946.2817959.

1. INTRODUCTION

The advances in information technology over a couple of decades have been revolutionizing how people communicate with one another. *Online* communication channels, such as messengers, online social networks (OSNs), or social media, have become important in everyday life. These online digital channels of communications are not only facilitating the communications among people, 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" [20].

From the old message boards such as USENET or BBS to the recent OSNs like Twitter and Facebook, there have been many computational data-driven studies that provide valuable insights into communication patterns on various online spaces [14, 18, 22–24, 26]. One of the common communication patterns is a *threaded conversation* that can effectively capture how people communicate with each other for a given particular topic in a structural fashion. In a threaded conversation, a person (or an initiator) brings up a conversation topic (by uploading a write-up or posting her opinion), and then other people (or participants) respond to the initiator or other participants recursively, which can be modeled as a tree structure.

This paper first seeks to model and characterize a threaded (online) conversation from three perspectives: (i) **volume** — how big the conversation is, (ii) **responsiveness** — how fast conversation participants react to the conversation initiator or other participants, and (iii) **virality** — how many participants elicit others to join the conversation. Note that the virality mostly refers to how content (or information) is spread by word-of-mouth mechanisms, or more specifically, how many nodes in a cascade are responsible for attracting other nodes. In this sense, we explore *the virality of a threaded conversation* that can quantify how participants' comments elicit others' responses, which signifies a multi-

generative property of the threaded conversation. We also investigate what factors (e.g., participant or content properties) are associated with the volume, responsiveness, and virality of a threaded conversation, which might be the key to modeling online conversation patterns. Such a model has a great utility in predicting future online conversation patterns, which can provide valuable implication on content providers, opinion leaders, or marketers.

To this end, we collect and analyze posts and comments on `Reddit`, one of the most popular online communities where users can communicate with one another in a threaded conversation for sharing their topical interests. In Reddit, there are various topical communities, so-called *"subreddits"*, each of which provides an independent space for users who are interested in any particular topic, e.g., game, politics, or sports. According to *Reddit*[1], as of May 2015, more than 169 M unique users (from 209 countries) have visited more than 7.5 B pages in Reddit. As of May 2015, Reddit is the 25th and 11st most popular web site in the world and the United States, respectively [1]. A user in Reddit can (i) visit a topical community (subreddit) to browse content, (ii) submit either self-writing or URL link content, (iii) write a comment on such content, and (iv) write a comment to another comment.

Using the collected dataset, we seek answers to the following questions by investigating the patterns of threaded conversations observed in Reddit: How can we characterize an online threaded conversation in terms of volume, responsiveness, and virality? What are the main drivers (e.g., user participation behaviors or content properties) that determine large, responsive, or viral conversations? How do conversations in different topical communities (or subreddits) show similar or different patterns?

To answer such questions, we analyze a large dataset (700 K posts and 18 M associated comments generated by 1.5 M users) that we collected from March 13 to April 18, 2014 from Reddit. By keeping track of all the newly-uploaded posts and their follow-up comments, we extract 700 K threaded conversations, each of which is represented as a *comment tree* model. That is, since users can leave comment(s) for all posts and comments, each post or comment can have follow-up comments in a nested fashion, forming a tree structure for a conversation[2]. Based on the comment tree model, we investigate the volume, responsiveness, and virality of each online conversation, and explore how the content properties and user participation behaviors are associated with them.

We highlight the main contributions of our work as follows:

- **Measurement:** To our knowledge, this is the first large-scale, extensive measurement study that characterizes the online conversation patterns in Reddit in terms of volume, responsiveness, and virality, based on 700 K comment trees shared by 1.5 M users. We make our anonymized dataset online at: `http://mmlab.snu.ac.kr/traces/reddit`.
- **Content Properties of Conversations:** We explore whether content properties, e.g., sentiment or difficulty, are associated with the volume, responsiveness, and virality of conversations. We find that large, responsive, and viral conversations tend to have (i) social words and (ii) high document relevancy between parent and child comments. 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.
- **Users' Participation in Conversations:** We investigate how characteristics of users' participation behaviors in a conversation (i.e., a comment tree) are associated with the volume, responsiveness, and virality of the comment tree. We find that a large and viral comment tree is often generated from a small portion of users who reciprocally communicate to each other in the tree. Interestingly, users who are interested in multiple topics play important roles in large and viral conversations, whereas heavy posting users play more roles in responsive conversations.
- **Conversations in Different Topical Communities (Subreddits):** We explore the conversation patterns in different topical communities (i.e., subreddits) in Reddit. We find that the news-related and image-based subreddits are more likely to have large and responsive conversations, respectively. On the other hand, we observe that the conversations in discussion-related subreddits tend to be viral, implying that discussions are likely to elicit many other users to join the conversations.

The rest of this paper is organized as follows. We present the background of Reddit and review the related work in Section 2. We explain our measurement and analysis methodology in Section 3. We start our analysis by investigating characteristics of comment trees in Section 4. We then analyze how the comment trees show different patterns across topical communities in Section 5, followed by concluding remarks in Section 6.

2. BACKGROUND

2.1 Reddit

Reddit allows users to share news, articles, and opinions with each other on 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" who are responsible for moderating and policing conversations among members. 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 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".

2.2 Related Work

Online conversation: Online communications with diverse forms (e.g., messengers, social media) have begun to dominate everyday social interactions. This has spurred

[1]http://www.reddit.com/about

[2]In this paper, we consider a (threaded) conversation as a set of communications among participants associated to a post (i.e., the root).

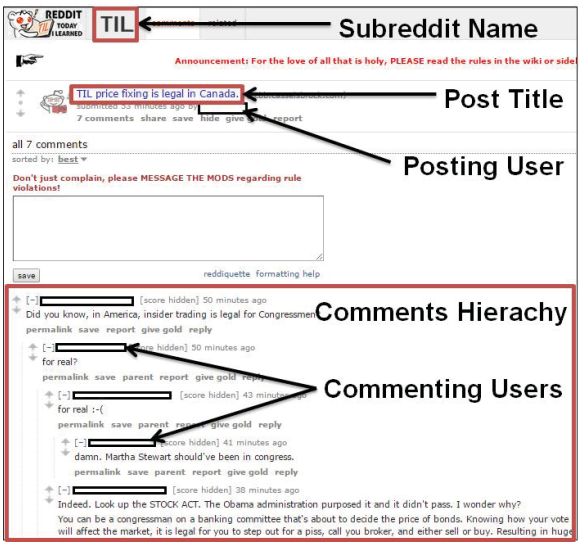


Figure 1: A post with its associated comments in the subreddit "Today I Learned (TIL)" is illustrated.

studies on online communication behavior across different systems such as online chatting [23], online communities [14, 18, 22], and OSNs [24, 26]. Mayfield *et al.* investigated a way to disentangle the conversation threads from multi-part chatting [23]. 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 [18]. Marcoccia investigated conversation threads in USENET newsgroups [22], 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 [14], which is 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. Rossi *et al.* analyzed conversations generated from specific hashtags in Twitter, and found that a choice of hashtags could make a conversation between users with non-reciprocal friendship [24]. Wang *et al.* proposed a model to predict the volume of conversations in `Digg.com`, and also applied the model to different platforms such as Twitter and Reddit [26]. We focus on analyzing what 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 and understanding online conversation patterns.

Reddit – popular online communities: Reddit has recently received a great attention as it becomes one of the largest online communities where people can communicate with one another sharing a variety of interests [11, 25]. Recently, many researchers have investigated user behaviors [9, 17], commenting patterns [10, 27], and content popularity [11, 19, 21, 25] on Reddit. Singer *et al.* investigated the user preferences for different topics shared among Reddit users [25]. The two case studies on "Hurricane Sandy" [21] and "duplicated image submissions" [19] 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 [11]. Weninger *et al.* analyzed top-scoring posts and their comments in Reddit, and showed that comments in closer positions in a comment tree are topically more similar than the ones in farther positions [27]. Choudhury *et al.* investigated texts of the posts and comments that contain self-discourse about mental health in Reddit, and found that posts with higher emotional intensity tend to receive more comments [10]. In this paper, we characterize conversations in terms of volume, responsiveness, and virality, and explore what and how characteristics of content and user participation are associated with such criteria.

Information/Content cascade in OSNs: As OSNs have become popular platforms in sharing information or content, there have been great efforts in investigating the patterns of information (or content) cascades in OSNs [4, 6, 7, 13, 16]. Cha *et al.* analyzed photo propagation patterns in Flickr and showed that photos do not spread widely and quickly in Flickr [6]. Goel *et al.* [13] analyzed the cascades of URLs in Yahoo! and Twitter, and found that the majority of the diffusions occur within one hop from a seed node. Han *et al.* analyzed the cascades of pins (i.e., images) in Pinterest, and showed that pin propagation in Pinterest is mostly driven by pin's properties such as its topic, not by user's characteristics like the number of followers [16]. 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 [7]. These studies have analyzed how information (or content) such as images or URLs is reshared, thereby generating a cascade, and what factors drive such information/content cascades in OSNs. On the other hand, we consider a threaded conversation cascade among participants associated to a post (i.e., the root). We further explore how user participation or content properties issued by a post lead to large, responsive, and viral conversation cascades on Reddit.

3. METHODOLOGY

In this section, we detail our measurement methodology for data collection, and describe the dataset used in this paper. We then characterize a conversation in terms of volume, responsiveness, and virality, with *a comment tree* model.

3.1 Data Collection

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 and analysis as shown in Figure 2. 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[3] provided by Reddit. We utilize 'Python Reddit API Wrapper (PRAW)[4]' package.

To monitor all the posts and their follow-up comments, we developed two key submodules in the core module: the post observer and comment observer. Once in every minute, the post observer monitors and fetches all new posts in

[3]Reddit provides public APIs, through which third party applications such as crawlers and readers are supported.
[4]http://praw.readthedocs.org/en/v2.1.16/

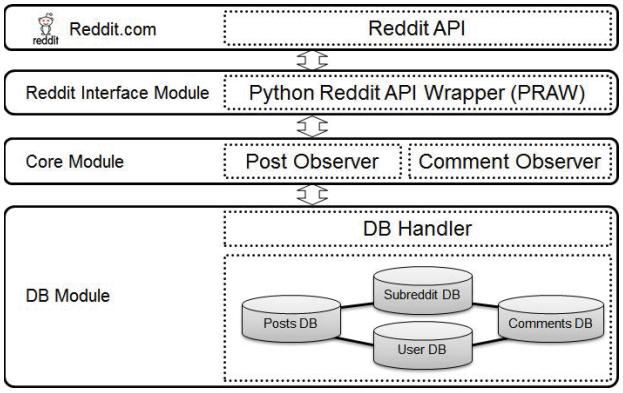


Figure 2: The architecture of the Reddit measurement system is depicted.

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 not to miss newly-uploaded posts. Whenever the post observer identifies a new post, the comment observer begins to keep track of all 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 since the observed maximum number of messages per minute was 722, which did not exceed the collected message 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 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 contains 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.

3.2 Comment Tree

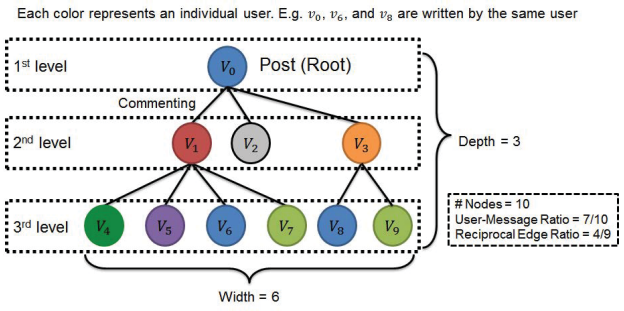


Figure 3: A comment tree is illustrated for a post that has 9 comments.

To model a conversation thread from a given post and its follow-up comments, we define a *comment tree* as an undirected 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 illustrates a comment tree that has one post and nine comments.

We characterize 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 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 inverses of time differences, *responsiveness* R_T, which indicates the average number of messages generated in a comment tree (during a minute). 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) [7,12], seeks to quantify the average range of a node's effect on the conversation in terms of connectivity. 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 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. The WIs are calculated for the four 10-node comment trees in Figure 4 for illustration purposes.

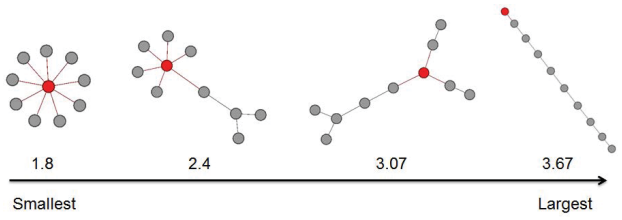


Figure 4: Virality values are calculated for 10-node comment trees ($N_T = 10$).

Figure 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 5(a), while

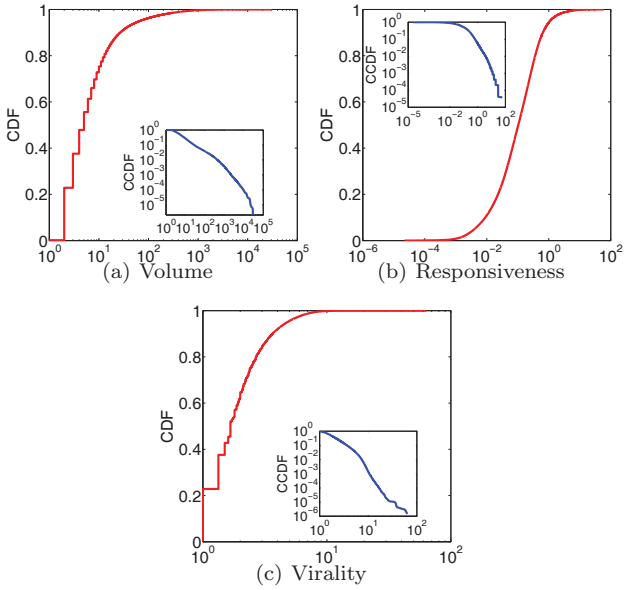


Figure 5: Distributions of volume, responsiveness, and virality of comment trees are plotted for all the comment trees.

72.8% of trees consist of less than 10 nodes, top 0.1% of the trees attracted more than 2,211 messages, indicating a large deviation among 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 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 virality values of the comment trees is 2.09, which implies a comment in a tree are 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, which implies that an average inter-comment time is around 3 minutes. In addition, the responsiveness values of the top 5% of trees are more 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.

4. 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 we find that all the p-values are smaller than 0.05.

4.1 Content Perspectives

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

4.1.1 *Semantic Characteristics*

We first perform a semantic 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 the 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.

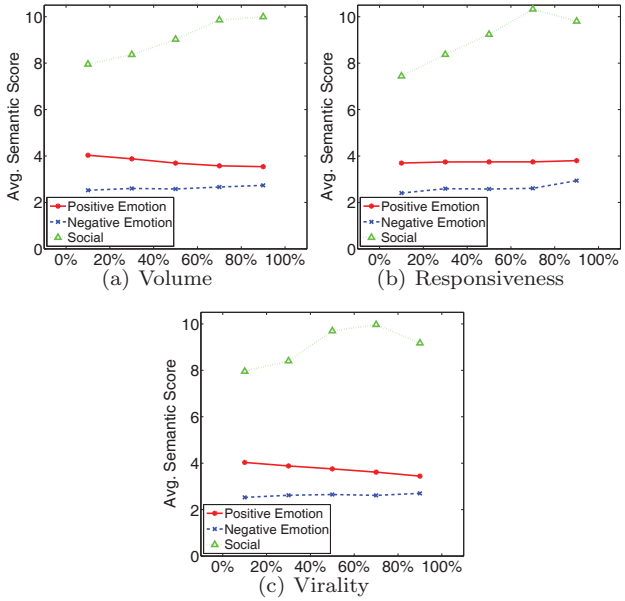


Figure 6: The distributions of emotional scores of posts are plotted.

Figures 6 and 7 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 [2]. There are no significant differences in the two emotional scores as the volume, responsiveness, and virality increase. On the other hand, the social scores of the titles tend to be high in larger, more responsive, and viral trees. The plotting of social scores of titles implies that a post whose title contains more social words is likely to be able to generate large, responsive, and viral trees to a certain degree.

4.1.2 *Document Difficulty*

We next measure whether the (readability) difficulties of titles and texts of trees are related 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 [15]. 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

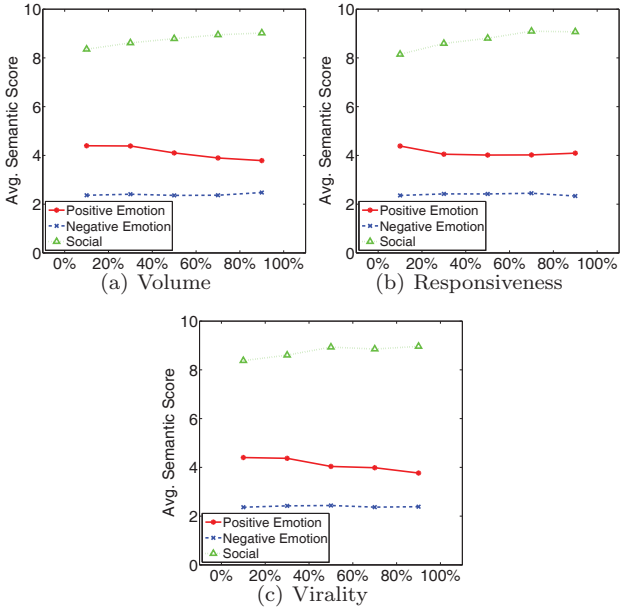


Figure 7: Semantic scores of texts in conversation trees are plotted.

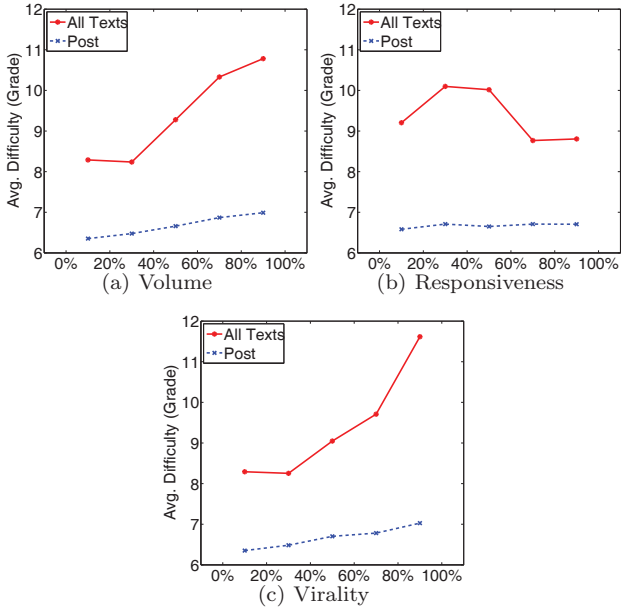


Figure 8: The average difficulties of trees and posts are plotted as the volume, responsiveness, and virality increase.

tree T is defined by:

$$G_T = 0.4[(\frac{N^T_{words}}{N^T_{sentences}}) + 100(\frac{N^T_{complex}}{N^T_{words}})] \quad (1)$$

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. As similar to the semantic analysis, we finally calculate the difficulties of comment trees for (i) title of a post and (ii) all texts of a comment tree (including its title).

Figure 8 plots the text difficulties of the title and the texts of the comment trees. As shown in Figure 8, 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 7) since a title is usually short and consists of a few keywords. Interestingly, the difficulties of both the titles of posts and the texts of comment trees increase notably as the volume increases, and more rapidly as the virality increases. This implies that a larger and more viral tree tends to consist of 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 less responsive trees, which implies using less and more plain words is positively related to the quick responsiveness.

4.1.3 *Document Relevancy*

We finally observe whether the relevance between two messages are related to volume, responsiveness, and virality. To this end, we compute the message relevance 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 area [3]. 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*. After measuring the TF-IDF score for each word, we then compute the cosine similarity of two score vectors between two documents. (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 9 shows the document similarity (1) between a post and its child comments, or (2) between a parent and its child comments. 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 9, 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 relevant 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 relevant to their parent messages.

4.2 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, reciprocal edge ratio, and user-message ratio across the five intervals. We then investigate how different roles of users are related to generating large, responsive, and viral comment trees.

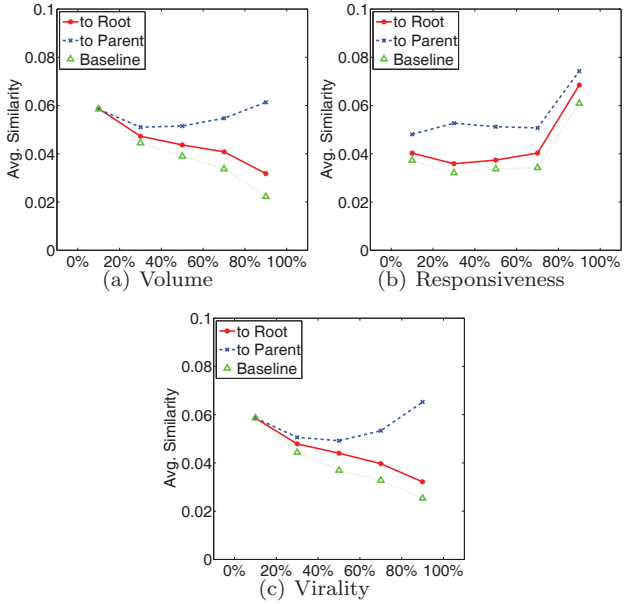


Figure 9: The relevance among messages of a comment tree is plotted.

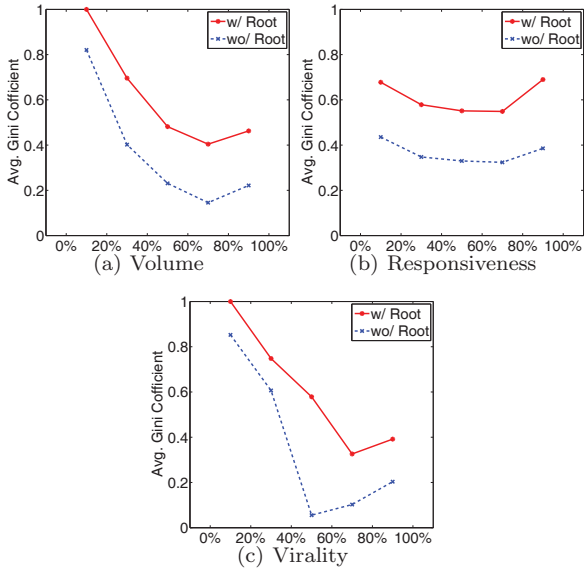


Figure 10: Average of Gini coefficients of comment trees are plotted as the volume, responsiveness, and virality increase.

4.2.1 *Participation Behaviors of Users*

We first quantify the skewness (variability) of comment generation in a tree by computing the *Gini Coefficient*, a metric that is most commonly used to capture inequality of income distribution in Economics [8]. The Gini coefficient, represented in the range of [0, 1], increases as the distribution of incomes is increasingly skewed. Thus, 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 in the opposite case (i.e., only one node (i.e., 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 10 plots 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 values sharply decrease 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 many 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, and show 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.

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 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.

Figure 11 shows the reciprocal edge ratio and user-message ratio as the volume, responsiveness, and virality increase. Obviously, 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 11(b) reveals that the top 20% responsive trees have the smaller reciprocal edge ratio and the higher user-message ratio. This result is in line with Figure 10 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.

4.2.2 *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} **(or initiators)** are the top 1% of users in terms of the number of uploaded posts. We call them *initiators* as they initiate conversations by writing many posts.

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

- U_{rcvcmt} **(or attractors)** are the top 1% users identified by the number of received comments from others. These users attract many comments from others, and

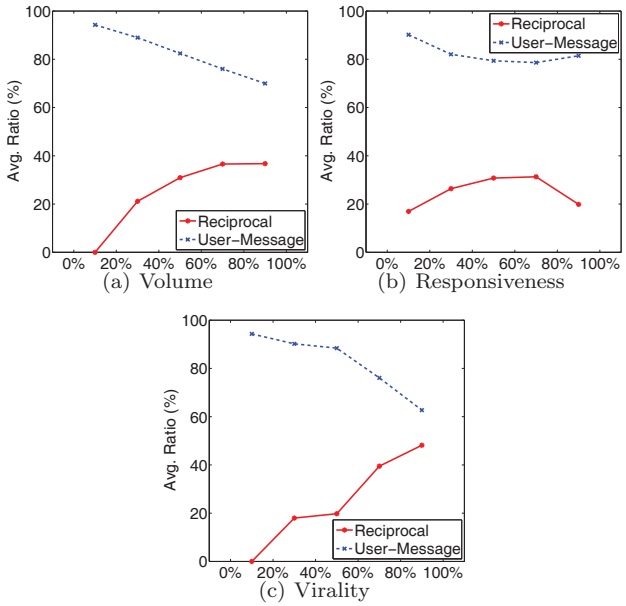


Figure 11: Reciprocal edge ratio and User-Message Ratio are plotted.

may play a major role in developing large, responsive, or viral conversations.

- U_{uni} **(or translators)** are the users who participate in a number of subreddits. These users are often known as *translators* [5] 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:

$$Entropy_u = -\sum_{m=1}^{N^u_{sub}} p^u_m \log p^u_m \quad (2)$$

where N^i_{sub} is the number of subreddits where the user u uploaded messages and p^u_m is the relative message portion of the m^{th} subreddit. We finally choose the top 1% of users based on the subreddit entropy. Since we do not normalize the subreddit entropy with 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.

We first measure how (identified) role types are overlapped by calculating the conditional probabilities of each pair of role types in Table 1. As shown in Table 1, 14% of users in U_{post} and U_{cmt} are overlapped, indicating that a small portion of users play 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 a lot also tend to receive many 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 distinctive from the users in other activity-related roles.

We now investigate how each role type contribute in large, responsive, and viral conversations, respectively. Figure 12

	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 1: Conditional probabilities among role types are described.

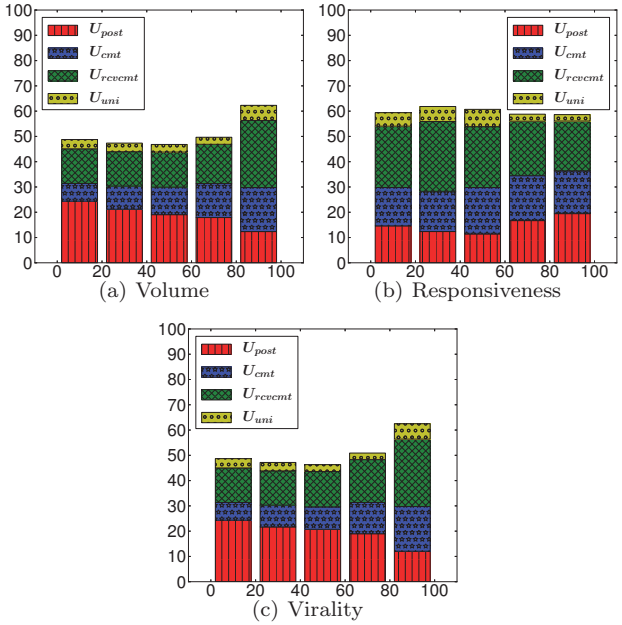


Figure 12: Contribution ratios of four user role types are plotted.

shows the portions of comments received by the users in each role type. As shown in Figure 12, 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 are larger and more viral, whereas the ones elicited by others increase, which indicates the users in the U_{post} play diminishing roles in large and viral conversations. Interestingly, U_{uni} attract more comments in the top 20% intervals both by the volume and virality, implying that translators who have broad interests are likely to more attract comments in a large or viral conversation. The responsive conversations show distinctive 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 a post content.

5. CONVERSATIONS IN DIFFERENT COMMUNITIES

In this section, we compare subreddits based on conversation patterns captured in volume, responsiveness, and virality of comment trees. Based on these patterns, 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 in each subreddit.

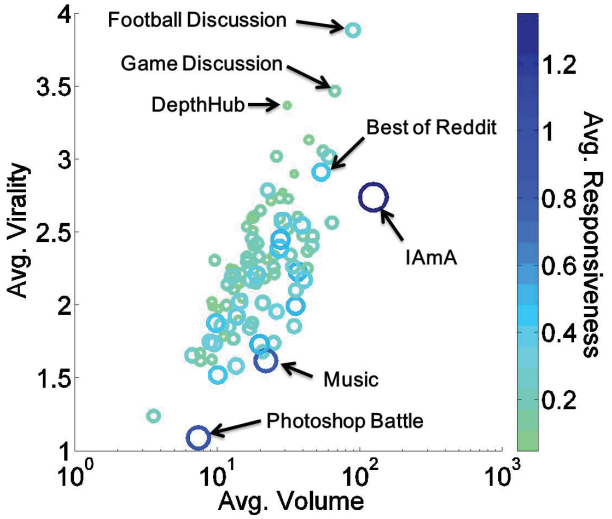


Figure 13: We map subreddits by calculating the average values of their trees in terms of the volume, responsiveness, and virality.

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 2: Top 10 subreddits in terms of volume, responsiveness, and virality are described.

5.1 Conversations in Subreddits

We investigate how conversations in different 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. We then plot *a subreddit map* in Figure 13 where each circle represents a subreddit. The position of a subreddit displays the average volume and virality, while the responsiveness is shown with colors. Note that a larger circle a subreddit is, its conversations tend to have higher 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.

Overall, Figure 13 shows that the average volume and virality of most subreddits exhibit a strong correlation, 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.

To further analyze conversation patterns in different subreddits in detail, we select the top 10 subreddits ranked by the volume, responsiveness, and virality, respectively, which are listed in Table 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 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 so 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-related 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`[5] 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 users' quick responses, while quick responses 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).

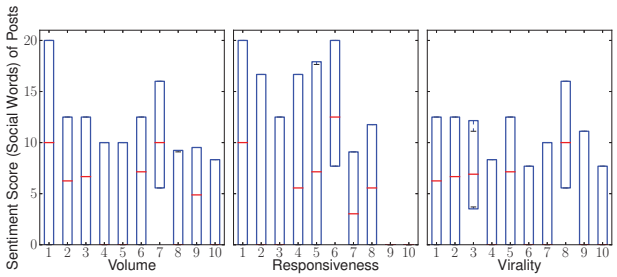
5.2 Content and User Characteristics

We now analyze how content properties and users' participation behaviors are different across the top 10 topical communities in Table 2. For the content properties, we report the three representative metrics, which turn out to be relevant ones with the large, responsive, and viral conversations in Section 4.1: (i) the semantic (social) score of a post, (ii) the document difficulty of a conversation by Gunning-Fog indexes, and (iii) the document relevancy to a post in a conversation.

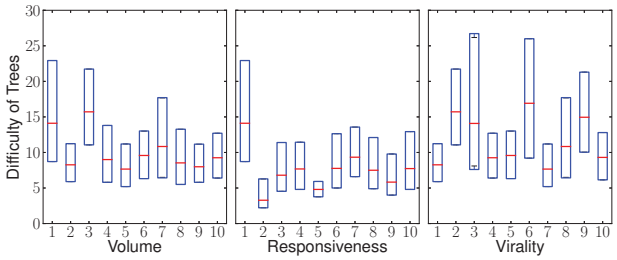
Figure 14(a) first shows the distributions of the social scores of posts across different topical communities. Note that we exclude outliers to present ones ranging from 25% to 75% of the distribution (as a box plot) to focus on the normal user behavior. For brevity, we refer to n^{th} subreddits in 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 14(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 14(b), we find that some subreddits in S_{vrl} have higher difficulties than oth-

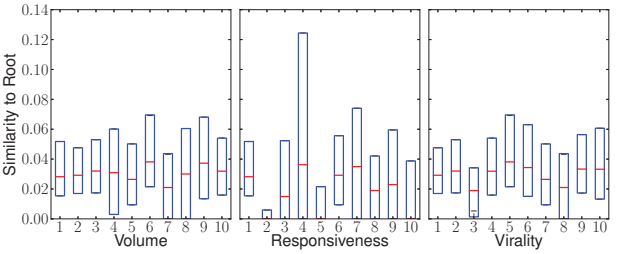
[5] `DepthHub` gathers the best in-depth submissions and discussions on Reddit.



(a) Semantic (social) Scores of a Post



(b) Document Difficulty of Tree

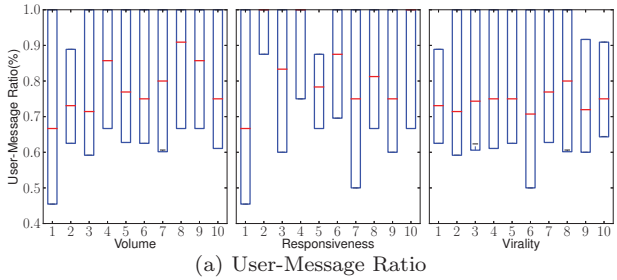


(c) Document Relevancy to a Post

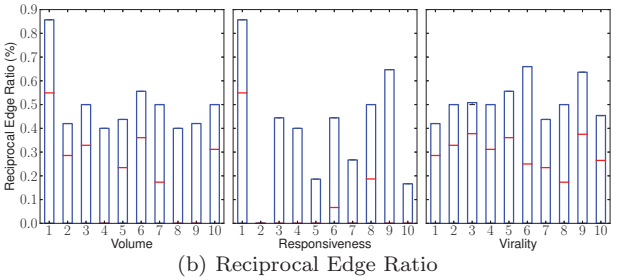
Figure 14: Three content properties across different subreddits are plotted in terms of semantic scores of a post, document difficulty of a tree, and document relevancy to a post.

ers. 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-related 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 results in responsive conversations. The average document difficulty of `IAmA` ($S_{vol}(1)$ or $S_{rsp}(1)$) are 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 difficult.

Figure 14(c) next shows the document relevancy to a post across different subreddits. We find that the document relevancy values of subreddits in S_{vol} and S_{vrl} are almost similar. However, we observe that subreddits in S_{rsp} show different patterns of the document relevancy. 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 on themselves (not on comments, but on posts).



(a) User-Message Ratio



(b) Reciprocal Edge Ratio

Figure 15: User-message and reciprocal edge ratio values are plotted across subreddits.

We next investigate users' participation behaviors across the top 10 topical communities in Table 2 with two user metrics plotted in Figure 15: (i) user-message ratio and (ii) reciprocal edge ratio. We find that the user-message ratio values of the most subreddits in S_{vol} and S_{vrl} are relatively lower than the ones in S_{rsp}; the reciprocal edge ratio values 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 4.2 that revealed large and viral conversations are 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 ratio values are high but their reciprocal edge ratio values are low, which implies that participants in those subreddits for news-related topics tend to submit a small number of comments and not to reciprocally communicate with others. 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 values of `AskReddit` ($S_{rsp}(6)$) and `Game of Thrones` ($S_{vol}(8)$) show 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 `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.

6. CONCLUSIONS

We have conducted a large-scale and comprehensive measurement study on online conversation patterns in Reddit. Using the collected dataset, we characterized online conversation patterns in terms of volume, responsiveness, and virality, and explored what and how content properties and

user participation behaviors are associated with the three metrics. We found that large, responsive, and viral conversations tend to have high document relevancy between parent and child comments. In addition, a large/viral conversation is likely to have difficult texts whereas a responsive conversation tends to have plain texts. We also discovered that large/viral conversations are built by reciprocal communications from relatively a small portion of users. As to user types in Reddit, we found users with wide interests play important roles of eliciting others' comments, which leads to large/viral trees, while heavy posting users tend to attract comments in responsive trees. We then expand our analysis to subreddits (topical communities), and learned that each community shows different characteristics; news-related, image-based, and discussion-related communities are more likely to have large, responsive, and viral conversations, respectively. We believe our analyses provide valuable insights to understand online conversations, e.g., content providers who want people to talk about their contents, or opinion leaders who want to obtain fast responses.

7. ACKNOWLEDGEMENTS

This work was supported in part by Basic Science Research Program through the "National Research Foundation of Korea(NRF)" funded by the Ministry of Science, ICT & future Planning (2013R1A2A2A01016562), Seoul National University Big Data Institute through the Data Science Research Project 2015, and Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (B0190-15-2013, Development of Access Technology Agnostic Next-Generation Networking Technology for Wired-Wireless Converged Networks).

8. REFERENCES

[1] Alexa - the web information company. `http://www.alexa.com`.

[2] LIWC statistics. `http://www.liwc.net/descriptiontable3.php`.

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

[4] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW*, 2012.

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

[6] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, 2009.

[7] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *WWW*, 2014.

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

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

[10] M. De Choudhury and S. De. Mental Health Discourse on reddit: Self-disclosure, Social Support, and Anonymity. In *ICWSM*, 2014.

[11] E. Gilbert. Widespread underprovision on reddit. In *ACM CSCW*, 2013.

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

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

[14] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW*, 2008.

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

[16] 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? In *ACM SIGMETRICS*, 2014.

[17] G. Hsieh, Y. Hou, I. Chen, and K. N. Truong. "welcome!": Social and psychological predictors of volunteer socializers in online communities. In *ACM CSCW*, 2013.

[18] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *ACM KDD*, 2010.

[19] H. Lakkaraju, J. McAuley, and J. Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *ICWSM*, 2013.

[20] 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, February 2009.

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

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

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

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

[25] 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? In *WWW Companion*, 2014.

[26] C. Wang, M. Ye, and B. A. Huberman. From user comments to on-line conversations. In *ACM KDD*, 2012.

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