Rolling through Tumblr: Characterizing Behavioral Patterns of the Microblogging Platform

Jiejun Xu, Ryan Compton, Tsai-Ching Lu, David Allen
HRL Laboratories
3011 Malibu Canyon Road
Malibu, CA 90265
{jxu, rfcompton, tlu, dallen}@hrl.com

ABSTRACT

Tumblr, a microblogging platform and social media website, has been gaining popularity over the past few years. Despite its success, little has been studied on the human behavior and interaction on this platform. This is important as it sheds light on the driving force behind Tumblr's growth. In this work, we present a quantitative study of Tumblr based on the complete data coverage for four consecutive months consisting of 23.2 million users and 10.2 billion posts. We first explore various attributes of users, posts, and tags in detail and extract behavioral patterns based on the user generated content. We then construct a massive *reblog* network based on the primary user interactions on Tumblr and present findings on analyzing its topological structure and properties. Finally, we show substantial results on providing location-specific usage patterns from Tumblr, despite no built-in support for geo-tagging or user location functionality. Essentially this is done by conducting a large-scale user alignment with a different social media platform (e.g., Twitter) and subsequently propagating geo-information across platforms. To the best of our knowledge, this work is the first attempt to carry out large-scale measurement-driven analysis on Tumblr.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.4 [**Computer Applications**]: Social And Behavioral Sciences; C.2 [**Computer-Communication Networks**]: General

General Terms

Measurement, Human Behavior

Keywords

Tumblr, Online Social Network, Quantitative Methods, Location-based Patterns

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.

ACM 978-1-4503-2622-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2615569.2615694 .

1. INTRODUCTION

Microblogging has become one of the most popular mediums for users to create and share information. Some of the most well-known microblogging platforms include Twitter™, Tumblr™, and Sina Weibo™[1]. Generally these platforms allow users to generate short-form, mixed-media (e.g., text, image) posts on any topics that are of interest to the users. These posts are then exchanged and propagated either through public broadcast or within a social network based on the connected users. While much research has been carried out to study user behaviors and social networks on different platforms, such as Twitter [20, 19, 36, 23] and Sina Weibo [13, 15, 33], little work has been done for Tumblr despite its rapid growth and popularity. As of the writing of this article, there are a total of 171.7 million blogs and 76.9 billion posts on Tumblr [1]. According to the data from the web traffic analysis company Alexa[2], Tumblr is ranked 17 in Unitied States and 32 globally based on a combination of average daily visitors and pageviews to the site over the past 3 month. On average, a user spends six and a half minutes on Tumblr daily. An online user study [12] shows that Tumblr is ranked second (after Facebook™) in terms of average time US visitors spent on all social media sites in a month, and it is four times more than Twitter. A survey of American internet users [28] also reveals that Tumblr is the favorite social network site for younger social media users in the age groups of 13-18 (teens) and 19-25 (young adults). One of the main design goals of Tumblr is to let everyone create and share anything effortlessly. It allows users to publish short text, photos, and other form of posts to the platform similar to other microblogging platforms. At the same time, it allows longer posts and provides extensive support for users to customize HTML page as in more traditional blogs. In a way, Tumblr is considered as a hybrid of Twitter and Wordpress™ as it combines parts from both [25]. The content-rich nature of Tumblr attracted not only individuals but also organizations to participate with different intentions ranging from business marketing [3] to government campaigns. The sheer volume of users and activities makes Tumblr an appealing platform to study human behavior of using social media as well as web-based large-scale social interaction. To this end, we conducted an in-depth measurement-driven analysis on Tumblr with 100% data covering four consecutive months

[1]The most popular microblogging platform in China.
[2]www.alexa.com/siteinfo/tumblr.com
[3]http://moz.com/blog/how-to-use-tumblr-for-seo-and-social-media-marketing

(between June 1^{st} 2013 to September 30^{th} 2013). To our knowledge, this is the first attempt to conduct such a large-scale study on Tumblr. The main contributions of our work are as follows:

- We explore different attributes on Tumblr (e.g., users, posts, tags) in detail and characterize their patterns based on the user generated content.
- We construct a massive *reblog* network based on the primary user interaction on Tumblr and present findings on analyzing its topological structure and properties.
- We introduce a practical solution to extract location-specific usage patterns from Tumblr by conducting a large-scale user alignment between Tumblr and Twitter.

The paper is organized as follows. Section 2 reviews related work on different microblogging and social media platforms. Section 3 provides background information about Tumblr as well as details of the data set. Section 4 describes the findings of human behavioral and usage patterns from the data. In Section 5, the Tumblr *reblog* network is discussed in-depth. In Section 6, we present the detail of obtaining location information for Tumblr users. Finally, Section 7 concludes the paper with insights and discussions.

2. RELATED WORK

Prior studies on human behavior and social interaction have been carried out quite extensively on existing microblogging platforms. Java et al. [20] analyzed the topological and geographical properties of the Twitter social network to determine individual user's intention in using such a platform. They found people predominantly use microblogging service to talk about their daily activities and to seek or share information. Kwak et al. [23] conducted an analysis on the follower-following topology of the Twitter social network, and found a significant deviation from known characteristics of human social networks in terms of the non-power-law follower distribution, a short effective diameter, and low reciprocity. Furthermore, the authors proposed three measures to rank influential users. Similar study on influential users was carried out in [5]. The authors compared users with different set of measures, and they found that the most followed users were not necessary the highest in other measures. Wu et al. [31] studied a few longstanding questions in media communication in the context of Twitter. They first developed a mechanism to distinguish between elite users versus ordinary users. Subsequently they found that a strong concentration of attention, in that half of the URLs consumed are generated by a small percentage of elite users. In addition, a significant homophily was observed in their work, which means contact between similar people occurs at a higher rate than among dissimilar people. Yu et al. [33] examined the key topics that trend on Sina Wiebo, and compared them with the observations on Twitter. They found a vast contrast between the two, in that trends in China are more centered around content such as jokes, images and video. Hutto et al. [18] conducted a longitudinal study on Twitter to an attempt to understand what factors lead to more followers. They concluded that variables for message content, social behavior, and network structure should all be given equal consideration when attempting to predict followers. Park et al. [27] took an unique angle by studying the emoticons on Twitter communication. They found that emoticons not only convey specific emotions, but also reflect socio-culture norms which vary depending on the identity of the speaker.

Besides typical microblogging platforms, behavioral studies have been carried out in more general social media platforms. Cha et al. [7] collected and analyzed large-scale traces of information dissemination in the Flickr™ social network. They found that even the most popular photos do not spread widely and quickly, which is contrary to common marketing belief. Ugander et al. [29] studied the structure of social graph of active Facebook users. The authors observed a clear degree assortativity patterns in the graph by studying the demographic and network property of users. They also observed a strong effect of age on friendship preferences as well as a globally modular community structure driven by nationality. A followup study on the Facebook social graph [4] reported a 4.74 degree of separation (i.e., average number of intermediaries on the path) between active users. Hochman et al. [17] applied Cultural Analytic techniques on large collection of Instagram™ photos to identify collective recurring visual patterns that provide insights into the study of different cultural practices. A more recent behavioral study on photo-sharing websites was conducted in [8]. The authors performed analysis of a large Flickr user logs to examine the navigation patterns between photo streams. Based on the observation, a stream transition graph was constructed to analyze common stream topic transitions. The graph was later incorporated in a collaborative filtering scheme for photo recommendations. Lately, less prevalent social medias platforms have also been explored. Coscia et al.[11] performed an empirical approach to study the behavior of internet memes, which are defined as specific fundamental cultural traits. The authors proved that in Quickmeme.com there are actual memes as they compete and collaborate, and sometimes cluster in large ensembles. Their work differed from main stream studies, in that they proposed a perspective without the use of network effects. Wang et al. [30] described findings of a detailed analysis on a social media-based question and answer site Quora™, based on three connection networks derived from the site. Their results showed a diversity in the user and question graphs are significant contributors to the quality of Quora's knowledge base. Finally, Mitta et al. [24] analyzed the Pinterest™ social network to extract general user behavior and common characteristics of different attributes in that platform. Two recent works which focused more on specific aspects of Pinterest can be found in [35] and [26]. Our work on Tumblr marks the latest addition to the large-scale social media analytic domain.

3. BACKGROUND

Tumblr is a content-rich microblogging and social media platform, where users create and share posts that are of interest to them. Each Tumblr user owns a *blog*, which contains all the posts from the user, and serves as the gateway to follow others. Tumblr primarily works with seven types of post: *text*, *photo*, *quote*, *link*, *chat*, *audio* and *video*. Note that there is another less common post type known as *answer*. Basically this allows a Tumblr user to add a box

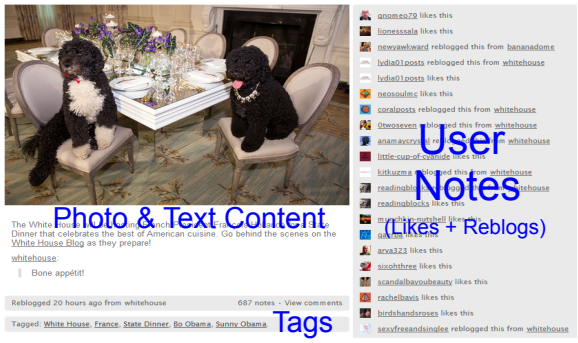


Figure 1: An example of a Tumblr *photo* post with caption. User activities (*Notes*) include both *likes* and *reblogs*

section on his *blog* for visitors to ask questions. Once the question is answered, it will turn into a regular post and displayed on the owner's Tumblr *blog*. Once a post is created, it could be published to Tumblr using one of the privacy control options. Unless the post is set to private, it is visible to other users. There are two types of actions (*like* and *reblog*), which can be applied on the post. The former is similar to Twitter *favorites* and marks what a user likes. The latter is similar to Twitter *retweet* and will clone the post to the *blog* of the acting user. The two actions are commonly referred as *notes* on Tumblr. Figure 1 shows an example of a Tumblr post. Users who "liked" and "rebloged" this post are displayed on the right pane.

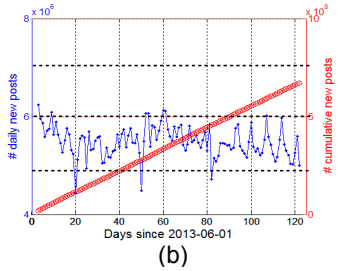
As can be seen from Figure 1, Tumblr users can also assign tags to their posts. These tags make it easier for other users to find posts about a specific topic. In fact, the original search mechanism on Tumblr only applies to tags, which means there is no way to retrieve a post from the Tumblr search engine if it is not tagged. This search function was improved in late 2013 to provide more comprehensive search results by checking other textual information in post contents and image captions.

The *follow* functionality in Tumblr provides a convenient way for users to get the latest updates (e.g., new posts) from other users. Similar to Twitter, the *follow* relation here is directional. This means a user U_a can follow user U_b without explicit permission, and U_b's updates will automatically appear in U_a's activity feeds (known as *dashboard* in Tumblr). By default, *follow* is considered as private data, thus not displayed publicly.

3.1 Tumblr Dataset

Time Period	2013-06-01 to 2013-09-30
# Posts	10,152,776,493
# Unique Posts	694,999,518
# Unique Users	23,228,627
# Unique Tags	230,898,675

(a)



(b)

Figure 2: (a) Statistics for the four-month data. (b) Number of original posts growing per day.

Photo	8,147,730,433	Answer	81,898,088
Text	1,406,432,506	Chat	73,191,208
Quote	246,930,034	Audio	63,099,299
Video	88,097,336	Link	45,397,589

Table 1: Number of Tumblr posts for each post type.

We obtain the Tumblr corpus over a 122 day period between June 1^{st}, 2013 to September 30^{th}, 2013 via GNIP[4] "firehose" (the complete stream of all posts). During the period, every public activity on Tumblr posts (including *publish*, *like* and *reblog*) is delivered to our system in real time encapsulated as a JSON record. Besides a record id, there is also a special identifier known as the *reblogkey* associated with each record. This is to indicate the origin of the post. For instance, if a new post is "liked" and "rebloged" after it is "published", we will receive a total of three records indicating the three activities, and these records all share the same *reblogkey*. The total size of data collected is roughly 10.6 TB with bzip2 compression. Data is stored via Hadoop Distributed File System (version 0.20.2-cdh3u3) deployed across a multi-node multi-core cluster.

The basic statistics of our dataset is summarized in Figure 2(a). As can be seen, the total number of posts collected over the four-month period is over 10 billion. However, the number of unique posts (i.e., excluding *reblog* posts) is only about 695 million. This suggests that large portion of Tumblr posts are simply duplicates of a small percentage of original contents (6.8%). Figure 2(b) shows the number of original posts observed on each day, as well as the cumulative number of posts observed during the same period. On average, there are $5,696,717$ original posts and $77,522,762$ *reblogs* generated everyday on Tumblr. Over the period of the four months, the total number of posts appears to grow linearly. We also analyze the distribution of different types of Tumblr posts (see Table 1). As can be seen, *photo* content dominates the microblogging platform by contributing 80% of the total posts. During our investigation, we also found that motion gif images are particularly popular on Tumblr. Finally, we have observed more than 23 million unique users and 230 millions unique tags in the data corpus. All the experiments in this work are implemented using a combination of standard Java MapReduce code and Apache Pig [5].

4. BEHAVORAL PATTERNS

We begin by exploring various Tumblr attributes in detail and extracting their associated patterns.

Posts and Users: We first analyze the relationship between posts and users. Figure 3 plots the distribution of number of posts per users over the four months period in a log-log grid. As *photo* posts consist of the primary content in Tumblr, we first focus our attention on its corresponding distribution (plotted in the top-left). This curve mirrors a typical heavy-tailed distribution, which occurs commonly in the social media domain. If a power law distribution is fitted to the curve, the parameter of α is found to be 2.17. It shows that for large majority of the users, each user publishes only a small number of posts during the period, while a small

[4]http://gnip.com/sources/tumblr/
[5]http://pig.apache.org

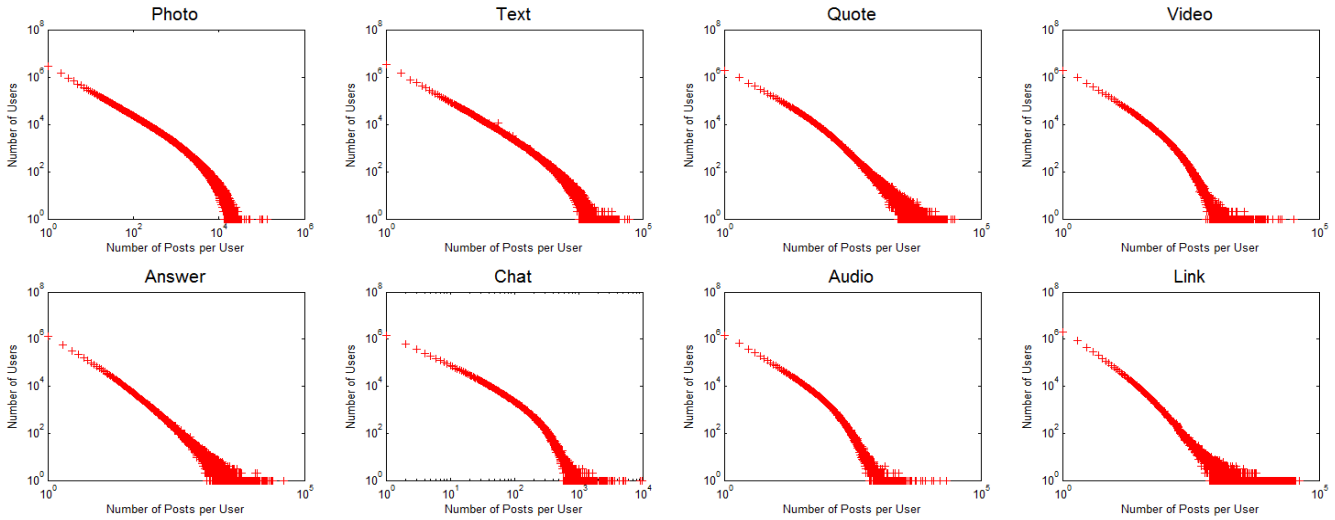


Figure 3: Number of posts per users (for different types of Tumblr posts). Plots are ordered based on the popularity of the corresponding post types from top-left to bottom right.

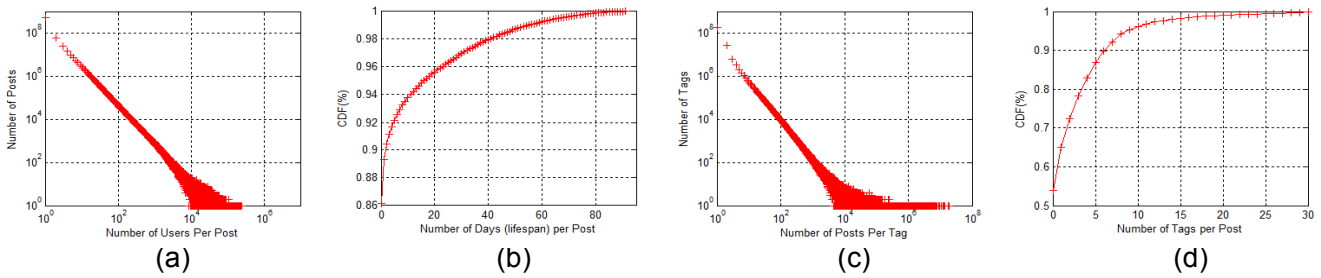


Figure 4: Various distributions of posts, users, and tags.

portion of users publish much more than the average. Similar distribution can be observed from the rest of the plots in Figure 3. The heavy-tail distribution in these plots indicate a high variability in user posting patterns. Note that longer tails are observed for less popular post types.

Next we study how users interact with posts. Specifically we compute the number of users who have "liked" or "reblogged" particular Tumblr posts. This computation can be conveniently implemented in a MapReduce paradigm, where the (key, value) pair consists of Tumblr *reblogkey* and user ID (or *uid*). The *Map* procedure takes in the data corpus and extract the specified (*reblogkey*, *uid*) pairs. Then the *Reduce* procedure takes in the collected pairs and combines them by counting the number of distinct *uids* for each key. In order to see the overall distribution, we invoke a final grouping step to sum up the number of keys at a each count level. Figure 4(a) plots the resulting distribution. This suggests that large majority of the original contents only attract a small percentage of users, while a small portion of popular contents are responsible for attracting most of the user interactions.

We are also interested in knowing the lifespan of original posts on Tumblr. We consider a post "alive" as long as it is still being "liked" or "rebloged". Since each original post in Tumblr is identified by a *reblogkey*, we can simply compute the lifespan of an original post by finding the first and last occurrence of its *reblogkey* in our data corpus. In order to compensate posts which are generated in late September and stayed "alive" beyond the coverage of our dataset, we only consider original posts which are first published between June and August 2013. The distribution of number of days a Tumblr post stays "alive" is shown in Figure 4(b). We observe that 90% of original posts has lifespan less than 2 days since its first creation. On the other hand, there is about 5% of original posts stay "alive" even 20 days after its creation, and about half of those goes beyond 40 days. The average lifespan of Tumblr posts appears to be longer than their counterparts in other micorblogging or social media platforms. Similar finding was reported in an online article regarding to Tumblr content[6].

Tags and Posts: Tags generally indicate specific topics, here we study the relation between tags and posts. Figure 4(c) shows the distribution of number of posts per tag. We observe a heavy-tail distribution: millions of user-defined tags are used in less than 10 posts, while a small set of tags are responsible for annotating the majority of the posts. Data points that lie on the right end of the curve correspond to most commonly used tags, and data points that lie on the left end of the curve correspond to rare tags. The long tail typically results from unconstrained tagging, where a user can put in any arbitrary text.

[6]http://unionmetrics.tumblr.com/post/45919888558/

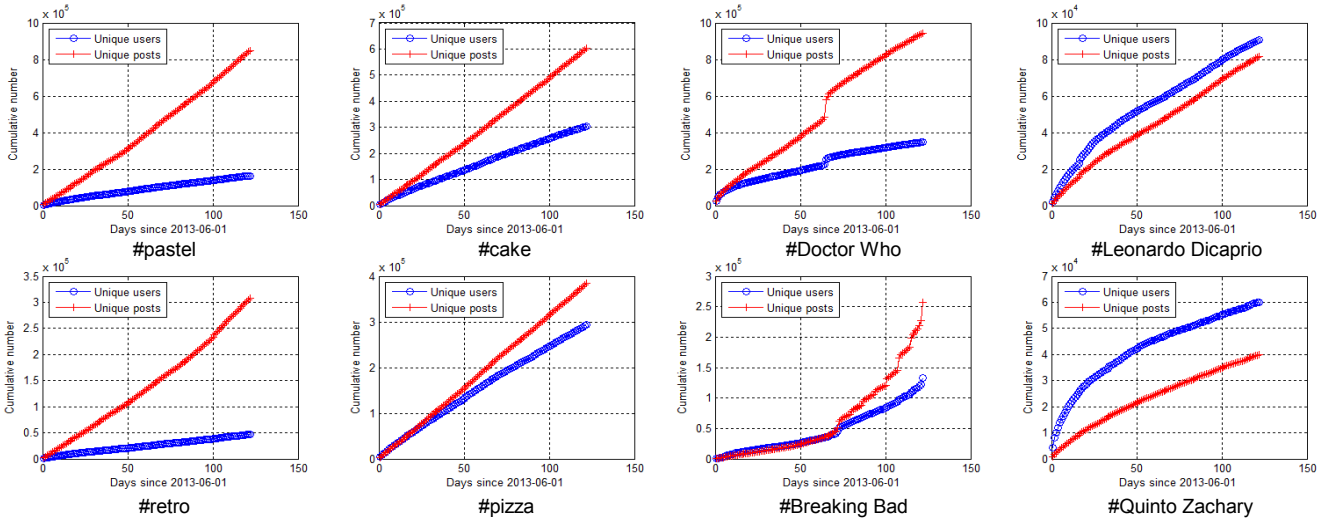


Figure 6: Cumulative numbers of unique users and posts (with mentions of specific tag) over time. Plots in the same column are considered to be in the same category.

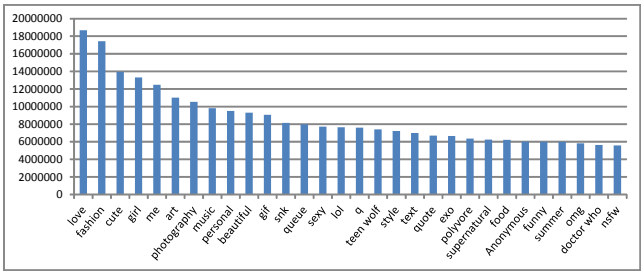


Figure 5: Most mentioned tags in Tumblr.

Knowing that a small number of tags are responsible for most posts, it is interesting to see what exactly the top tags are in Tumblr. Figure 5 shows the most frequently used tags in the platform. Clearly, Tumblr posts are highly biased towards certain topics. For example, "love" is used in more than 18 million Tumblr posts (about 1% of all), followed by "fashion" which is used in a similar magnitude. Based on the observation of the top-10 used tags, it again hints that Tumblr is largely visual-driven and the platform is centered around user hobbies and interests.

Next we study the average number of tags associated with Tumblr posts. We focus the study with respect to original contents, because by default *reblogs* do not include the tags from the original posts and *reblog* users are not likely to add them back. In our implementation, we first identify the list of original posts. Then for each post, we compute the maximum number to tags occurred in the *reblog* chain. The resulting distribution is plotted in Figure 4(d). It shows that more than 53% of the posts contain zero tags. On the other hand, there are about 4% posts which have more than 10 tags. We expect the curve to shift upward if we were to include all *reblog* posts.

Tags and Users: Here we investigate the popularities of different topics on Tumblr by examining the aggregated growth patterns of new users and posts for each topic. In particular, we focus on the following categories: fashion, food, TV show and celebrity. As topics are not explicitly defined in Tumblr, we carry out the experiment based on the representative tags from each topic. Similar method has been shown to be effective in existing literature [31, 32]. For each topic, we select two frequently used tags according to the summaries in the official Tumblr Year-in-Review webpage[7]. The final selections of tags are "pastel", "retro", "cake", "pizza", "Doctor Who", "Breaking Bad", "Leonardo Dicaprio" and "Quinto Zachary". Figure 6 shows the growth patterns for each tag. Plots located on the same column are from the same topic.

We make a few key observations from Figure 6 above. First, all topics do show steady rise in terms of number of users and posts over the four months period. However, the growth patterns appear to be rather different across topics. For instance, there are noticeble "jumps" in popularity for the TV show category. We suspect that they were triggered by real life events. In fact, the big jump in the number of users and posts for "Doctor who" was related to the announcement regarding to the actor of the show in June 2013. Similarly the continuous jumps for "Breaking Bad" correspond to the weekly broadcast dates. The correlations to real-world events may suggest that Tumblr is another effective social media for TV and movie box office predictions [3]. Another observation is that the large increase in the number of posts do not seem to bring in many new users for the fashion topic. We suspect that the active user group who drives the growth of the topic is highly focused and specialized. In this case, users are likely to be advocates of the particular fashion trends. Similar number of posts was accumulated for the food category as for fashion. However, this appears to be a more general topic as the growth in the number of posts is more aligned with the number of new users. Finally, the growth pattern is most different for the celebrity category, in that there are more users than the number of posts. The large number of users most likely consist of fans of the

[7]http://yearinreview.tumblr.com/2013

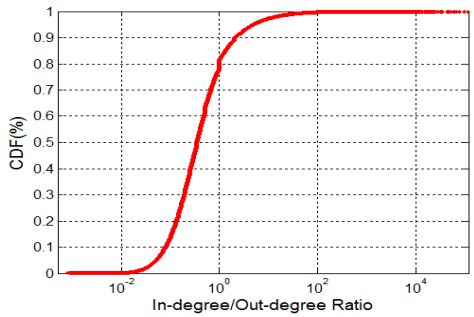


Figure 7: CDF(%) of the user in-degree and out-degree ratio.

celebrities. The sharp increase for "Quinto Zachary" at the beginning of the curve happens to align with the release of the latest "Star Trek" movie.

5. REBLOG NETWORK

Numerous network can be derived from Tumblr based on different user attributes. In this section, we study the *reblog* network as it is the primary online interaction on Tumblr.

5.1 Reciprocity

We build a social network, the *reblog* network, with users as vertices and weighted directed edges corresponding to *reblogs*. That is, if user i "reblogs" user j for w_{ij} times, then we have a directed edge from i to j with weight w_{ij}. Note that in our network users who generate popular content will have a high in-degree while users who primarily share content will have a high out-degree.

Our data contains $7,062,528,912$ *reblogs*. From these we build a network consisting of $18,367,173$ users with $999,548,135$ directed edges between them. Based on the constructed network, we observe that only $87,832,337$ (8.8%) of the edges were reciprocated (i.e. user i "reblogged" user j and user j "reblogged" user i).

The low reciprocation rates happen elsewhere in the social media platform Twitter [23]. In addition to the popular *retweet* phenomenon, Twitter users often "@mention" each other by appending an "@" symbol to the mentioned user's name. This two activities are often used as the mediums to study reciprocity. To put in perspective the network structure of Tumblr, we directly compare it with the network structure of Twitter. We use the same Twitter data as we used in [10], which consists of 10% sample of public tweets from April 2012 to January 2014 obtained through the GNIP Decahose[8]. The same Twitter corpus is also used for large-scale user alignment and it will be discussed in Section 6.1. The full dataset amounted to 67.2TB of uncompressed JSON data and a total of $22,455,584,506$ @mentions of any type. From *retweets* only, we built a weighted and directed network of $137,269,098$ users with $4,493,285,385$ edges. Filtering edges for reciprocation left us with in $248,104,403$ edges, i.e. a 5.5% chance of *retweet* reciprocation. From nonretweet @mentions only, we built a weighted and directed network of $198,755,741$ users and $3,954,866,992$ edges. Here, we found $738,295,950$ reciprocated edges, i.e. an 18.6% chance of nonretweet @mention reciprocation. Our results suggest that *reblogs* on Tumblr are stronger indicators of social ties than their Twitter counterpart of *retweets*, but weaker indicators than Twitter @mentions.

[8]http://gnip.com/sources/twitter/

Model	In-dgree	Out-degree
Power law	$\alpha = 1.26$	$\alpha = 1.13$
Truncated power law	$\alpha = 1.18, \lambda = 1.23 \times 10^{-4}$	$\alpha = 1.00, \lambda = 2.78 \times 10^{-4}$
Lognormal	$\mu = 2.40, \sigma = 2.90$	$\mu = 3.35, \sigma = 2.85$
Stretched exponential	$c = 0.248, \lambda = 0.092$	$c = 0.319, \lambda = 0.016$

Table 2: Fitted parameters for degree distributions of the Tumblr *reblog* network

We also investigate the distribution of the ratio of a user's incoming and outgoing degrees in the *reblog* network. Results are shown in Figure 7. In our dataset, only $11,259,743$ users have both nonzero in-degree and out-degree, and our experiment is carried out among these users. Overall, $2,486,406$ (22%) users have higher in-degree than out-degree. If we loosely assume that the number of in-degree is proportional to the number of "follower", these users have higher "follower-followee" ratios. A very small portion (2.76%) have 10 times more in-degree than out-degree.

5.2 Activity vs Degree

We observe that node degree distributions of the Tumblr *reblog* network do not closely follow standard power laws. A similar result was obtained for other blogging platforms in [16]. To be precise, we fit degree distributions to the following models:

- Power law: $x^{-\alpha}$
- Exponentially truncated power law: $x^{-\alpha}e^{-\lambda x}$
- Lognormal: $\frac{1}{x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$
- Stretched exponential: $x^{c-1}e^{-\lambda x^c}$

We utilized the software package provided by [2] for precise model fittings, and the parameters are obtained via the method of [9]. The complete networks for both degree cases are used in the fitting process. For computational efficiency, we only fit once over the full degree range (i.e., $x_{min} = 1$ and x_{max} = the max weighted node degree in each network). Results are summarized in Table 2, plots of the data with various fits are shown in Figure 8.

Pairwise comparison of different models is achieved via the loglikelihood-ratio test. Truncated power laws fit better than standard power laws in both situations (for in-degrees: loglikelihood ratio 46461.17, $p < 0.0001$; for out-degrees: loglikelihood ratio 2103152.75, $p < 0.0001$). On the in-degree network, a lognormal model outperforms the truncated power law (loglikelihood ratio 19800.16, $p < 0.0001$) and appears to outperform the stretched exponential in the tail. Overall fit, however, shows that the difference between the lognormal and stretched exponential fits is not statistically significant (loglikelihood ratio 31.91, $p = 0.96$). Similarly, for out-degrees, the truncated power law models the tail well, but this result is not statistically significant when compared against the stretched exponential over the full range (loglikelihood ratio 0.92, $p = 0.99$). Ultimately, understanding the true underlying processes of Tumblr network formation requires a detailed sociological model (such as the work found in [6]) which is outside the scope of this work.

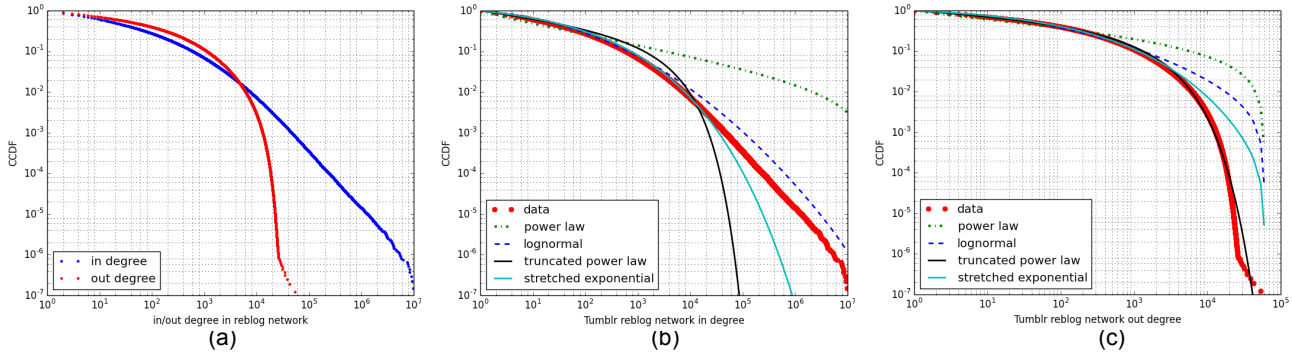


Figure 8: (a) Log-scale plots of in/out-degree distributions of the Tumblr *reblog* network. (b)-(c) Log-scale plots of in/out-degree distributions of the *reblog* network with various statistical fits.

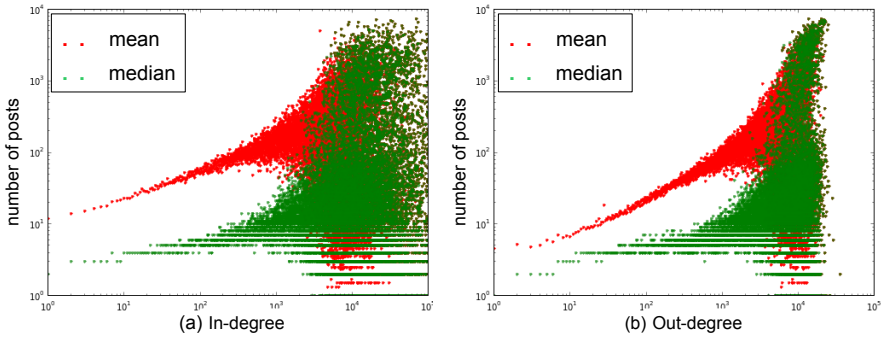


Figure 9: (a) Activity level v.s. in-degree. (b) Activity level v.s. out-degree.

In order to investigate the correlation between user node degrees (unweighted) and that of user activities (i.e.,published posts), we plot the number posts against the number of in/out-degree a user has in Figure 9(a) and Figure 9(b). As can be seen, user activity is an increasing function of both degree types when $x \leq 100$. This suggests that high-degree users are more apt to post often. In addition, we observe that the average number of posts again the number of degrees per user is above the median in most degree range. Towards larger degrees, extremely active outlying users are becoming more prevalent as evidenced by the discrepancy between mean and median in both degree cases.

5.3 Ranking Tumblr Blogs

The popularity of Tumblr users (or their corresponding blogs) can be estimated by various measures. Here we consider two aspects: the total number of *reblogs* of a user and the (unweighted) in-degree of the user in the *reblog* network. We compute and compare the results of the top-20 users from both measures in a side-by-side manner. For privacy concern, we don't explicitly disclose the details of individual users. Based on our observation, 14 out of 20 users are common in both rankings, and the top-4 users are identical in both cases. The top-ranked user according to *reblog* counts has a total of $26,030,925$ *reblogs*, which is an order of magnitude more than the next top-ranked user. The rest of the 19 users in the list have *reblog* counts ranging between 4 million to 9 millions during the four months period. In terms of ranking by in-degree, the top-ranked user has more than $500,000$ in-coming nodes, and the rest of the users have between $250,000$ to $470,000$ in-coming nodes. A closer look on each blog reveals that majority of the blogs belong to individuals. This is radically different from the top-ranked users (with the same measures) in Twitter, as most of them come from news organizations or celebrities [23]. The majority contents from the top-ranked Tumblr blogs consist of fashion images, jokes and other topics that are most popular among younger social media users.

6. USER LOCATIONS

Another interesting study is to dissect the Tumblr social network based on users' geolocations and investigating how users' geolocation impacts their participation in the platform. However, Tumblr does not support any forms of geotaggging or provide any user location information in its native platform. Inspired by the success of our prior work on geocoding Twitter users [21, 10], we propose a practical solution to obtain geolocation for Tumblr users by first conducting a large-scale user alignment with the Twitter platform and subsequently propagating user geolocations across.

6.1 User Alignment

User alignment across social media has been a topic of growing interest [34, 22]. However, most existing approaches rely on complicated correlation analysis and modeling, which impose strong constraints on scalability. In this work, we take a simpler extractive-based approach to align social media users by utilizing the additional Twitter data we obtained (see Section 5.1). Specifically, we search for two types of user linkages from Twitter as follows.

Explicit Self-reported Links: Due to the increasing popularity of social media, many active social media users have their virtual identities in multiple social media platforms. For instance, it is common for Twitter users to provide alternative social media accounts in their profile in order to promote their online presence. For our case, we are interested in detecting explicit mentions of the Tumblr user accounts. To do that, we simply search for every user profiles in our Twitter corpus with the regular expression "`http://[www.]*[a-zA-Z0-9-_].tumblr.com`". This gives us a total of $5,672,374$ pairs of aligned users between Twitter and Tumblr.

Implicit Cross-Links: Many existing social media sites support content synchronization in order to reduce end user effort. Basically, this allows users to submit a post from one platform, and the content of the post will be automatically published to all other social media platforms under the same user. For the case of Twitter, there is usually a URL appended at the end of a tweet to indicate its origin. For example, the top image in Figure 10 shows a sample tweet which encodes a link to the Tumblr platform (e.g., http://tmblr.co/ZVxw1y15H_Go3). In other words, the tweet content was original published in Tumblr, and it was automatically synchronized to Twitter. Bottom image in Figure 10 shows the referenced original Tumblr post.

This kind of cross-referencing turns out to be very useful in terms of identifying the same user across the two platforms. The key here is the shorten URL with the prefix pattern of "tmblr.co". This shorten URL is automatically generated by the Tumblr server, and it would only be triggered by the synchronization process between Tumblr and Twitter. Since synchronization only happens when a user owns both accounts, this should be reliable way to identify same users across platforms. Thus we scan our entire Twitter corpus for URL mentions in the form of "`http://tmblr.co/(\\S{4,20})`" We detect a total of $35,907,479$ such cross-linking instances. By a post processing step to resolve the shorten URLs, we are able to identify a total of $1,444,447$ unique pairs of aligned Twitter and Tumber user accounts. Figure 11(a) shows the raw cross-linking frequencies from the identified users. Note that users are ordered based on the frequencies in the x-axis. Figure 11(b) plots the same data in a log-log scale. The curve shows typical characteristic of a Zipf curve.

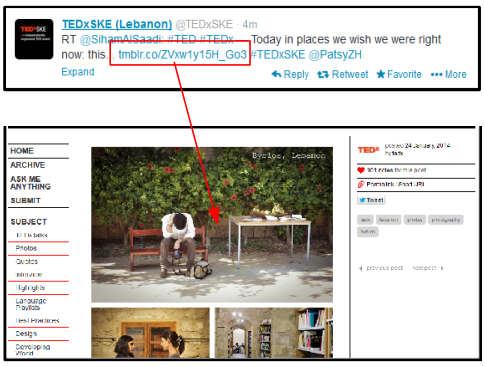


Figure 10: An example of implicit cross-linking between Twitter and Tumblr.

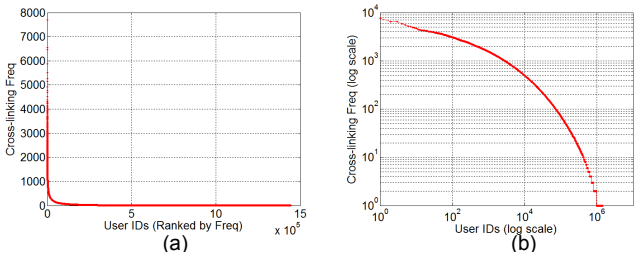


Figure 11: (a) Implicit Twitter-Tumblr cross-linking distribution by users. (b) log-log plot of the same distribution.

We perform an preliminary evaluation on the aligned users in order to verify the accuracy of the implicit cross-linking approach. Basically we sample at random 40 aligned users at different cross-linking frequency range for manual inspection. That is 40 users with at most one cross-linking instance, 40 with between 1 and 10, 40 with between 10 and 100, 40 with between 100 and 1000, and finally 40 with over 1000 cross-linking instances. This is essentially sampling at different spectrum of the log-log curve in Figure 11(b). During visual inspection, a human user determines if the linked pair is correct based on various heuristics such as user names, profile images, and others. Figure 12 summarizes the accuracy of this approach. As can be seen, the alignment accuracy remains at 100% as long as the cross-linking frequency between two accounts is greater than 1. Error only occurs in the lowest cross-linking frequency range. We suspect these rare cases are due to direct "copy-and-paste" of tweets. In order to maintain the good accuracy, we use the frequency value as a threshold and discard all aligned pairs with less than 5 cross-linking instances.

	Ture	False	N/A (Account deactivated)	Unsure
>1000	**36**	**0**	3	1
>100, <1000	**36**	**0**	4	0
>10, <100	**32**	**0**	8	0
>1, <10	**35**	**0**	5	0
=1	**23**	**9**	3	5
Overall	**162**	**9**	23	6
	94.74%	**5.26%**		

Figure 12: Evaluation on implicit cross-linking.

To summarize, we have obtained a total of $6,549,937$ unique pairs of aligned users with both explicit self-report links and implicit cross-links.

6.2 Geocoding

The basic idea of our prior works [21, 10] is to formulate the Twitter user geocoding problem as a graph-based optimization problem. Nodes in the network represent users, and edges represent geodesic distances between users weighted by their number of reciprocated @mentions. Given a small set of initial nodes (users) with geolocation information, the goal is to propagate this information to the rest of the network through the intrinsic structure of the network. Essentially the optimization is to seek a stable state of the network such that the sum over all geographic distances between connected users is minimized. The intuition behind

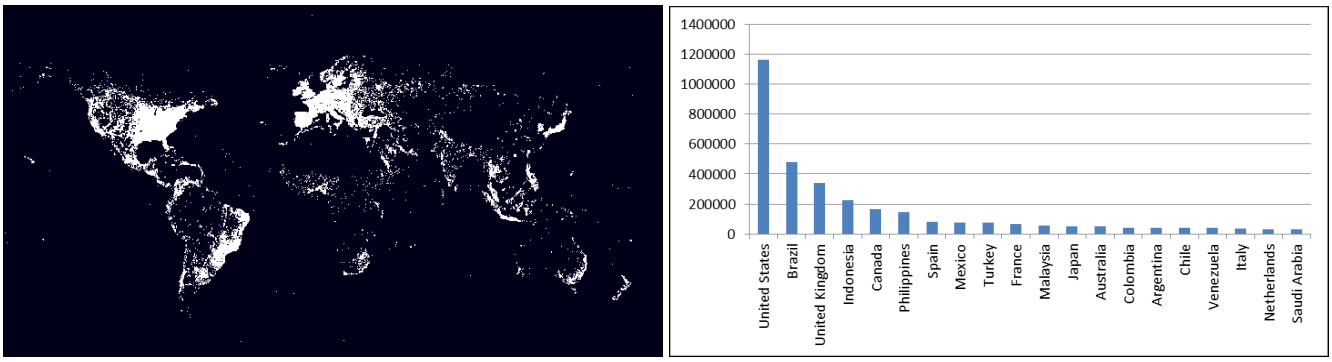


Figure 13: (Left) Distribution of geocoded Tumblr users. (Right) Top-20 countries with most geocoded users.

such an approach is that users who have strong social ties in online community are likely to live in a nearby proximity.

The works in [21, 10] successfully inferred the locations for 89% of the users in our Twitter corpus with a median error of 6.65 km. Subsequently, based on the overlapping users identified from the Twitter and Tumblr alignment, we are able to provide geolocation information for a total of 4,449,990 Tumblr users. Figure 13 shows the global distribution of geocoded Tumblr users, as well as the top-20 countries. Substantial coverage is obtained in north and south American, Europe, and south-east Asia.

Figure 14 shows the total number of posting activities from different countries based on geocoded users. For simplicity, only countries fall into one timezone are selected. As can be seen from the figure, clear daily recurring patterns are observed in all countries. However, each country does exhibit a distinct posting characteristics in a finer hourly scale. For instance, Japan has more activities during the day, and Colombia has more activities later in the evening. While detailed country-level analysis is beyond the scope of this work, user localization on Tumblr makes it possible to conduct similar study as in [14].

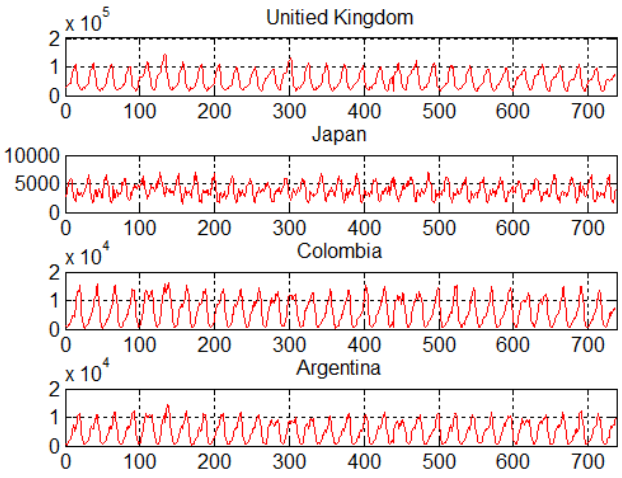


Figure 14: Country-specific hourly activity (# post) since 2013-09-01T07:00. Hours are normalized with respect to GMT+00:00.

7. CONCLUSIONS

In this paper, we presented a large-scale quantitative study to analyze the collective microblogging patterns on Tumblr. We found that 1) The contents in Tumblr are mostly centered around users' interests and hobbies, and the majority of users appear to fall in the younger demographic group; 2) Most of the Tumblr posts are essentially relogs, and original contents consist of less than 10% of the overall posts; 3) Tumblr posts typically have longer lifespan, however, the popularity growth of posts in different topics varies significantly; 4) Social ties derived from the Tumblr *reblog* network is weaker than the @mention network in Twitter. Furthermore, we provided substantial results on localizing Tumblr users based on a large-scale user alignment with the Twitter platform. We believe that our work is the first step towards better understanding and exploring the full potential of Tumblr.

In the future, we are interested in extending our work to several directions. First, we would like to study the patterns of information dissemination on Tumblr in a fine scale under different topics or categories. Second, we would like to extend our user alignment work to cover different social media platforms. Third, given the user linkages across platforms, we are interested in constructing large-scale multiplex networks and study their structures as well as the corresponding cascading behaviors.

8. ACKNOWLEDGMENTS AND DISCLAIMER

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI / NBC) Contract Number D12PC00285. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the author(s) and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

9. REFERENCES

[1] Tumblr official page. http://www.tumblr.com/about. Accessed: 2014-02-20.

[2] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PLoS ONE 9(1): e85777*, 2013.

[3] S. Asur and B. A. Huberman. Predicting the future with social media. *CoRR*, abs/1003.5699, 2010.

[4] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *WebSci*, pages 33–42, 2012.

[5] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International Conference on Weblogs and Social Media*, 2010.

[6] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, New York, NY, USA, 2007.

[7] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[8] L. Chiarandini, P. A. Grabowicz, M. Trevisiol, and A. Jaimes. Leveraging browsing patterns for topic discovery and photostream recommendation. In *ICWSM*, 2013.

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

[10] R. Compton, D. Jurgens, and D. Allen. Geotagging one hundred million twitter accounts with total variation minimization. *arXiv:1404.7152*, 2014.

[11] M. Coscia. Competition and success in the meme pool: A case study on quickmeme.com. In *ICWSM*, 2013.

[12] J. Delaney, N. Salminen, and E. Lee. Infographic: The growing impact of social media. *http://www.sociallyawareblog.com*, Nov. 2012.

[13] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu. A comparative study of users' microblogging behavior on sina weibo and twitter. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, pages 88–101, Berlin, Heidelberg, 2012.

[14] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333:1878–1881, 2009.

[15] W. Guan, H. Gao, M. Yang, Y. Li, H. Ma, W. Qian, Z. Cao, and X. Yang. Analyzing user behavior of the micro-blogging website sinaweibo during hot social events. *CoRR*, abs/1304.3898, 2013.

[16] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 369–378, New York, NY, USA, 2009. ACM.

[17] N. Hochman and R. Schwartz. Visualizing instagram: Tracing cultural visual rhythms. In *ICWSM*, 2012.

[18] C. Hutto, S. Yardi, and E. Gilbert. A longitudinal study of follow predictors on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 821–830, 2013.

[19] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11), Nov. 2009.

[20] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Workshop on Web Mining and Social Network Analysis*, pages 56–65, 2007.

[21] D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *ICWSM*, 2013.

[22] M. Korayem and D. J. Crandall. De-anonymizing users across heterogeneous social computing platforms. In *ICWSM*, 2013.

[23] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, 2010.

[24] S. Mittal, N. Gupta, P. Dewan, and P. Kumaraguru. The pin-bang theory: Discovering the pinterest world. *CoRR*, abs/1307.4952, 2013.

[25] M. Novak. Unwrapping tumblr. *http://www.slideshare.net/myklnovak*, Jan. 2013.

[26] R. Ottoni, J. P. Pesce, D. B. L. Casas, G. F. Jr., W. M. Jr., P. Kumaraguru, and V. Almeida. Ladies first: Analyzing gender roles and behaviors in pinterest. In *ICWSM*, 2013.

[27] J. Park, V. Barash, C. Fink, and M. Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *ICWSM*, 2013.

[28] C. Smith. Tumblr offers advertisers a major advantage: Young users, who spend tons of time on the site. *http://www.businessinsider.com/tumblr-and-social-media-demographics-2013-12*.

[29] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *CoRR*, 2011.

[30] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: an analysis of quora. In *WWW*, 2013.

[31] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 705–714, New York, NY, USA, 2011.

[32] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 177–186, 2011.

[33] L. L. Yu, S. Asur, and B. A. Huberman. What trends in chinese social media. *CoRR*, abs/1107.3522, 2011.

[34] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In *SIGKDD*, pages 41–49, 2013.

[35] M. A. Zarro, C. Hall, and A. Forte. Wedding dresses and wanted criminals: Pinterest.com as an infrastructure for repository building. In *ICWSM*, 2013.

[36] D. Zhao and M. B. Rosson. How and why people twitter: The role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, 2009.