Topical Semantics of Twitter Links

Michael J. Welch[*]
Yahoo! Inc.
Sunnyvale, CA 94089
mjwelch@yahoo-inc.com

Uri Schonfeld
UCLA Computer Science Dept
Los Angeles, CA 90095
shuri@shuri.org

Dan He
UCLA Computer Science Dept
Los Angeles, CA 90095
danhe@cs.ucla.edu

Junghoo Cho
UCLA Computer Science Dept
Los Angeles, CA 90095
cho@cs.ucla.edu

ABSTRACT

Twitter, a micro-blogging platform with an estimated 20 million unique monthly visitors and over 100 million registered users, offers an abundance of rich, structured data at a rate exceeding 600 tweets per second. Recent efforts to leverage this social data to rank users by quality and topical relevance have largely focused on the "follow" relationship. Twitter's data offers additional implicit relationships between users, however, such as "retweets" and "mentions". In this paper we investigate the semantics of the follow and retweet relationships. Specifically, we show that the transitivity of topical relevance is better preserved over retweet links, and that retweeting a user is a significantly stronger indicator of topical interest than following him. We demonstrate these properties by ranking users with two variants of the PageRank algorithm; one based on the follows sub-graph and one based on the implicit retweet sub-graph. We perform a user study to assess the topical relevance of the resulting top-ranked users.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms

Algorithms, Experimentation

Keywords

Twitter, Web graph, link semantics, modeling, ranking

[*] Work completed while author was a graduate student in the UCLA Computer Science Department.

WSDM'11, February 9–12, 2011, Hong Kong, China.
Copyright 2011 ACM 978-1-4503-0493-1/11/02 ...$10.00.

1. INTRODUCTION

`Twitter` is a micro-blogging site which currently[1] ranks 10^{th} world wide in total traffic according to `Alexa`[2], and according to `compete.com`, has over 28 million unique monthly visitors[3]. Twitter has often demonstrated itself as a leading provider of information for breaking news events, and its rapidly growing influence and reach have inspired many researchers to look deeper into the nature of the information it captures [26, 12, 15].

The Web is constantly evolving, and the changes go beyond advances in the presentation technology (Flash, AJAX, new HTML standards, and so on). Sites such as Twitter are making large quantities of potentially useful, highly structured information available through publicly accessible interfaces. Simply modeling this information as a graph of independent, text-based pages (nodes) connected by hyperlinks (directed edges), will fail to capture all that this information has to offer, and thus produce less than ideal results.

In this paper, we present a rich graphical model for Twitter with multiple semantic edges, and demonstrate the key principle that "not all edges are created equal." We explore how the Twitter graph compares with the Web graph and, as a result of this discussion, we will uncover some of the semantic differences between the various types of links represented in Twitter. More specifically, we explore the relationship between users and topics with respect to two types of edges. A *follow* link indicates that one user is reading what the other is writing. A *retweet* link is formed when one user reposts what another user posted.

The importance of understanding the semantics of a link becomes even more apparent when using them in a ranking algorithm. The problem of finding topic-specific influential Twitter users has recently been examined [26], showing that *follow* links are correlated with topical similarity of user interests. Our research complements this work by showing that retweet links are better suited for inferring a user's topical relevance than follow links.

We demonstrate this property by identifying relevant users for a particular topic. Given a seed set of users relevant to a topic, we compute "topic sensitive" PageRank [21] on two sub-graphs of the Twitter graph, one based only on follow links and another based on retweet links. We then evaluate the topical relevance of the resulting high-ranked users. Our

[1] Data as of November 14, 2010

[2] http://www.alexa.com/siteinfo/twitter.com

[3] http://siteanalytics.compete.com/twitter.com/

experiments indicate that the high ranking users based on retweets are more likely to remain topically relevant then the high ranking users based on follow links. In other words, the act of repeating a user's post carries a stronger indication of topical relevance.

To better understand why follow links are less suited for determining topical relevance, we explore the notion of a user's dual role on Twitter. That is, each user acts both as a content consumer, or *reader* interested in what other users post, and as a content producer, or *writer* by publishing new posts. A user follows other users because she is interested in *reading* the topic(s) they write about. On the other hand, other users follow her because of the topic(s) she *writes* about, which may differ from what she reads. Our experiments highlight the potential disconnect between a user's topical relevance as a *reader* and as a *writer*, and clarifies the transitivity of topical relevance over links on Twitter.

The rest of this paper is structured as follows. We begin with a discussion of some influential related work in Section 2. Section 3 briefly introduces basic, important aspects of Twitter and defines some of the terminology used throughout the paper. In Section 4 we describe how common Web modeling techniques can be adapted or augmented to better represent the structure of Twitter. In Section 5 we analyze a large dataset collected from Twitter and compare it to the Web. In Section 6 we take a closer look at global rankings over Twitter and comment on some of the interesting aspects of the data. The results of our experiments for ranking users with topic sensitive PageRank are described in Section 7. We conclude and describe some future work in Section 8.

2. RELATED WORK

One of the goals of this paper is to provide insights towards a better understanding of the overall structure of Twitter. Comprehensive studies of the Web graph [17, 4] have exposed significant underlying features of the link structure, including confirmation of power-law distributions of inlinks and outlinks and studies of connected components. These studies have had tremendous influence on key areas of Web research, such as crawling strategies and ranking algorithms.

Researchers have also studied the structure and growth of particular sites in the past. Almeida et al. [22] study user behavior on Wikipedia, observing trends such as an exponential increase in contributing users over time and a power-law distribution for users making edits to pages. Weng et al. present analysis of follow relationships and posting frequency for a small set of approximately 6800 Twitter users in Singapore [26]. Our work includes similar analysis on a larger dataset of approximately 1.1 million users.

Blogs and social networks have received significant attention in recent years as well. Studies have investigated reading and posting behavior [25], blogger influence on the public [8, 9, 19, 20], and the general structure of Web communities [7] or the blogosphere as a whole [16].

The rapidly increasing popularity of Twitter has sparked the interest of many researchers in recent years. Huberman et al. observe that a user's true "friends" are typically a very small subset of those they follow [11]. Krishnamurthy et al. [15] study the growth and the usage of patterns in Twitter, including examining the source of posts (SMS, Web clients, Facebook, and so on) and geographic locales. Ramage et al. [23] first categorize Twitter posts into four broad categories using a survey of Twitter users, then filter tweets into these categories according to their topics. Sarma et al. [6] try to improve the ranking accuracy for the "Twitter-like" postings in forums with a comparison-based mechanism, such as thumb and star ratings. Ritter et al. [24] propose an unsupervised approach to model dialogue in Twitter, which aims to identify strong topic clusters within noisy conversations. Java et al. [12] study the Twitter graph and applications of the HITS algorithm [14] for detecting user intent. Their observations of power-law distributions for links agrees with our findings.

Some of the closest work to our own [26] extends PageRank to Twitter, making use of follow relationships in the Twitter graph as well as topical similarity derived from the user's tweets to find influential users for various topics. Our work is complementary, examining the semantics of follow links more closely and proposing retweet links as an additional or alternative source of information.

In the Web ecology project [3], Twitter users are ranked with different criteria such as number of followers, average content spread per tweet, average conversation activity per tweet, and so on. None of these ranking mechanism utilizes retweet information. TunkRank [2] proposes a ranking mechanism to identify the most influential Twitter users. They describe a notion of a user's "influence" as the expected number of users who will read a tweet from them, whether directly as a follower or via a retweet. This influence is propagated in the graph among user, similar to the work in [26], where the weights are propagated along the follow links. Their work acknowledges that retweets are important for probabilistically determining how far a user's post will propagate, and is thus factor in measuring their overall influence. They do not focus on topic sensitive ranking, however, and only propagate influence over follow links.

3. TWITTER 101

Twitter is a blogging platform which allows registered users to publish small articles of text though multiple interfaces, including Web, SMS, and instant messaging. Each post, or *tweet*, is limited to a maximum of 140 characters, leading to the description of Twitter as a *micro-blogging* platform.

On Twitter every user has dual roles, both as a publisher of posts, or *writer*, and as a subscriber, or *reader* of other's posts. As a reader, a user may choose to *follow* another user's posts. The set of users you follow are referred to as your *friends* (this terminology can be a bit confusing as, despite its symmetric-sounding nature, the relationship is only one way), and the set of users who follow you are called your *followers*. All of the posts from a user's friends are accessible via a private stream, sorted by their publication timestamp. For the vast majority of users, their friends, followers, and posts are publicly accessible through both a Web interface and through published REST APIs (the Twitter API[4]).

The basic structure of friend and follower relationships described above is enforced through Twitter's implementation and the APIs they make available to third-party developers. However, there are also interesting structural phenomenon resulting from social conventions which have "naturally" evolved in the Twitter community. For example,

[4] http://dev.twitter.com/

when publishing a post which references another user, that user is referred to by their username prefixed with the '@' character. This is called a *mention* of the user by the author of the new tweet. A more specific type of mention occurs when a user chooses, for the benefit of his followers, to repeat another user's post. In this case, he prefixes the content of that post with "RT @" followed by the username of the original author of the post. This type of post is referred to as a *retweet.* To reiterate, these last two examples began as social conventions amongst Twitter users and, until recently, were not explicitly representable in the Twitter API.

Over the past year, Twitter added an explicit retweet mechanism, though the style of interaction differs somewhat from the "old style" retweets. A user cannot modify the text of the post she is repeating when using the newer explicit retweet method. Also, third party websites are increasingly placing "retweet" buttons on their articles, which allows a user to click and generate a "retweet" with a link to the page. Despite this new API, the original style of retweeting remains common. As their semantics (e.g. forwarding interesting content to followers) are similar, in this paper we refer to both the original and new styles of retweeting simply as *retweets.*

A second feature on Twitter, added in late 2009, allows users to construct and organize a group of users referred to as a *list.* The users followed by a list are referred to as the list *members.* Lists help a user to, for example, focus on the posts of certain subsets of their friends or follow a group of users en masse. Roughly speaking, lists on Twitter fall into two broad categories: topical lists and classification lists. Topical lists are generally centered around the discussion of common interests or subjects, such as "politics," while classification lists are generally formed to group users who share a common trait, such as "celebrities" or "professional athletes". As a side effect, lists generate meaningful manually-created categorizations of users.

In the next section we more formally define the content and structure of Twitter. We define the Full Twitter graph and introduce a simplified Twitter graph, and compare them to the Web graph.

4. MODELING TWITTER

In this section we introduce a graph representation of Twitter information and compare it with the Web graph, which has been extensively studied in the past [17, 4]. In the graph model of the Web, pages are represented as nodes and the hyperlinks connecting them are represented as directional edges. This model enables the application of many graph analysis techniques, such as inlink and outlink distributions and the PageRank [21] algorithm.

4.1 The Full Twitter Graph Model

The Web graph is commonly represented as an n by n matrix M, where n is the number of pages on the Web. M_{ij} is equal to $\frac{1}{c_j}$ if page j contains a link to page i and has a total of c_j outgoing links.

The Twitter graph is inherently more complex. First, there are at least two types of entities which could be represented as nodes: users and tweets. Second, there are at least four types of relationships between these nodes which would be represented as directional edges: follows, publishes, retweets, and mentions.

	User	Tweet
User	Follow	Publish
Tweet	Mention	Retweet

Table 1: Twitter Graph Edges

A *follow edge* from user u_a to user u_b exists if u_a follows the posts of u_b. A *publish edge* from user u_a to post p_a indicates authorship of the post. Both of these edge types stem from the enforced structure of Twitter, and the graph containing just these two edge types we call the *Enforced* Twitter graph.

We define the *Full* Twitter graph by including two additional edge types, representing relationships inferred from the posts when assuming the social conventions discussed earlier. We create a *retweet edge* from post p_a to post p_b if p_a is a retweet of p_b, and a *mention edge* from post p_a to user u_b if p_a mentions u_b. These four edge types are summarized in Table 1.

Interestingly, the type of edge in this model is uniquely identified by the types of vertices it connects. As a result, no special distinction is needed for the edge type in the graph. The graph is thus a simple directed graph whose vertices can be divided into two disjoint sets: U and P, corresponding to the *users* and *posts*, respectively.

The matrix representation of the Twitter graph can be modeled identically to the Web graph using T, a $|U| + |P|$ by $|U| + |P|$ matrix where $|U|$ is the number of users and $|P|$ is the number of posts. A non-zero value in T_{ij} represents an edge between node i and node j, the semantics of which are defined in Table 1.

4.2 Additional Twitter Information

There are three important pieces of information that are not captured in this graph representation, which we briefly mention here.

4.2.1 Time

The most notable omission from the graph relates to the temporal nature of the data. Twitter includes timestamp information for when each post was written as well as when accounts were created. There is no explicit way to determine when a follow link was created using the public API, though going forward these can be approximated with repeated crawling. Time data would prove valuable for studying factors such as evolution of the graph [18] or charting popularity over time, but was omitted here for clarity and as it is not necessary for the focus of this paper.

4.2.2 Hyperlinks

The second type of information excluded from this graph are standard hyperlinks embedded in the posts. Extending the graph to include hyperlinks would require crossing boundaries between Twitter and the Web at-large, complicating the analysis we wish to discuss here. As an intermediate step, our model could be augmented with a third node type representing a Web page, uniquely identified by a URL. A directed edge indicating a reference to the page's URL would exist from any post p_a to the "Web page node" whose URL appears in its content.

A minor difficulty modeling hyperlinks in Twitter is the common use of URL shortening services. Links or URLs are typically shortened using services like `TinyURL`[5] and `bit.ly`[6] to stay within the post size limit. This prevents making use of keywords or other interesting artifacts the URL may contain directly, and makes additional processing of the data necessary. Whenever link analysis for URLs is done, URL normalization would likely be required.

4.2.3 Post Content

We use the content of a post primarily to extract metadata: username mentions and the identification of retweets for generating *mention* and *retweet* edges, respectively. The remaining textual content of a post can potentially be useful toward determining the topics of interest to a user as well, though the small size of the posts introduces several difficulties, including sparsity of data and tokens resulting from frequent use of nonstandard shorthand notation.

4.3 The Simplified Twitter Graph

The Full Twitter graph G_F attempts to represent all the important entities and relationships in Twitter. We will now describe a simplified Twitter graph G_S which only includes user nodes, while still capturing the most important information from the original representation as it pertains to the users.

The user-user follow links remain as they are from the Full Twitter graph. However, for every retweet edge from post p_a, written by user u_a, to post p_b, written by user u_b (the original post's author), we add a *retweet edge* in G_S from user u_a to u_b. For simplicity, we omit mention links in G_S. While simplifying the graph by removing the post nodes altogether, this new representation requires us to explicitly note the different edge types, as the properties and semantics of follow and retweet edges differ. For the experiments presented later in this paper, we use the simplified G_S graph.

5. ANALYSIS OF THE GRAPH

In this section we take a closer look at the data itself. All the figures, unless specified otherwise, are derived from a dataset we collected between October 2009 and January 2010. The dataset includes over 1.1 million Twitter users, with more than 273 million follow edges and over 2.9 million retweet edges. The data was collected by beginning with an initial seed set of the top 1000 users in `twitterholic.com`[7] and crawling in a BFS manner, traversing the follow links in a forward direction.

5.1 Link Distributions

We look deeper at the Twitter user's graph by investigating the inlink and outlink distributions of the different edge types. We begin by looking at follow edges.

5.1.1 Follow Edges

Figure 1 shows a log-log scale of the inlink and outlink frequencies. Both plots, as in the Web graph inlink and outlink distributions analyzed by Broder et al. [4], show a generally power-law distribution.

Figure 1(a) shows the inlink distribution, or how users are followed as writers. The graph only includes users who have at least one follower, and relatively few users have over 12,000 followers. Manual inspection of these highly followed users indicates the majority are related to professional organizations (CNN), or celebrities and public figures (Ashton Kutcher, Barack Obama).

The outlink distribution is plotted in Figure 1(b), showing how users act as readers. Again, the plot shows several interesting traits, including a large spike around the 20-friend region. An unusually high number of users follow exactly 20 other users, which is very likely a side effect of Twitter's account signup process, which provides an initial a set of 20 "recommended" users to follow. Another large spike appears exactly on the 2000-friend mark, which is likely due to the restrictions Twitter places on following more than 2000 users[8].

We computed the power-law exponents for the inlink and outlink distributions of retweet and follow links, using a nonlinear least-squares (NLLS) algorithm by Marquardt-Levenberg implemented as part of the gnuplot package. Follow inlinks have a power-law exponent of -1.6237 with an asymptotic error of 0.001977 (0.1218%). This value is lower than the values measured for the Web [4] and those for earlier studies on Twitter [12], which were found to be around 2.1 and 2.4, respectively. As a result of the noisy outlink data, although the data appears power-law, we were unable to find a reliable fit for the exponent.

5.1.2 Retweet Edges

The two distributions relating to retweets appear in Figure 2. The retweet inlinks distribution, or number of unique users who retweeted at least one post of the user, is shown in Figure 2(a). This distribution is similar to the other power-law distributions we have seen, with a power-law exponent of -2.01108 with a high asymptotic standard error of 0.001167 (0.05804%), closer to typical values for the Web. The retweet inlinks following a distribution similar to hyperlinks on the Web might be more than just a coincidence. A retweet edge is similar in nature: a link from one author to another, typically considered an indication of relevance or quality of the content, much like a hyperlink between two pages on the Web.

The retweet outlinks, or number of unique users whose posts were retweeted by a given user, is shown in Figure 2(b). This distribution, unlike the others (including the number of friends, or users a person follows), does not seem to follow the power-law distribution. That is, while the number of friends one has is generally power-law, the number of users one finds truly interesting (or worth repeating) does not appear to scale in a similar fashion. We considered a log-normal distribution as well, which also did not prove to be a good fit.

5.1.3 Posting Frequency

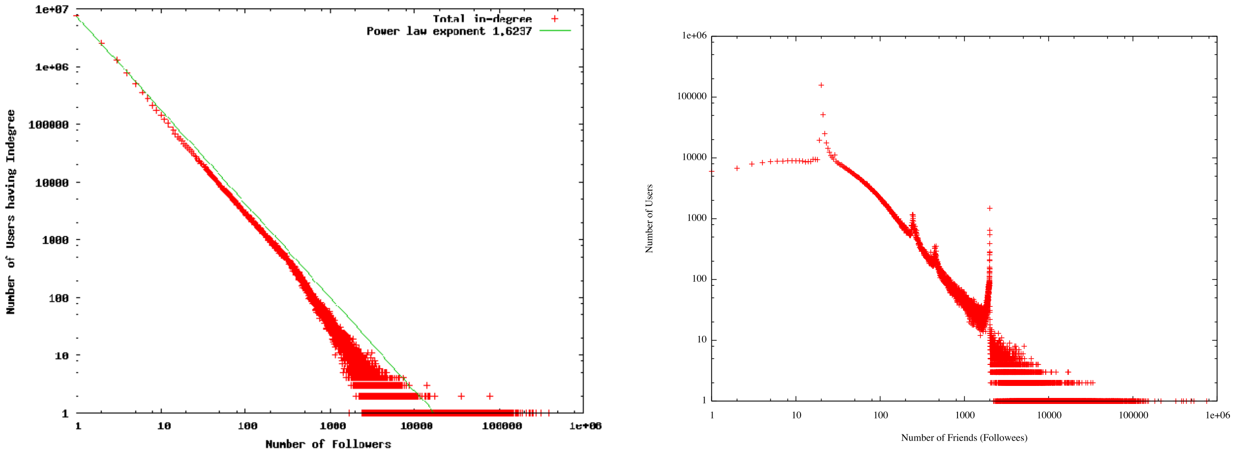
We next examine the posting frequency behavior of users. Figure 3 shows, on a log-log scale, the number of posts published vs. the number of users writing that many posts for 417,613 users during the one month period of December 2009. From our data set we excluded the 683,347 users who did not publish any posts during the month from the graph, and due to API restrictions limiting us to collecting only the most recent 200 posts per user, we also excluded the 5,760

[5] http://tinyurl.com/

[6] http://bit.ly/

[7] http://twitterholic.com/

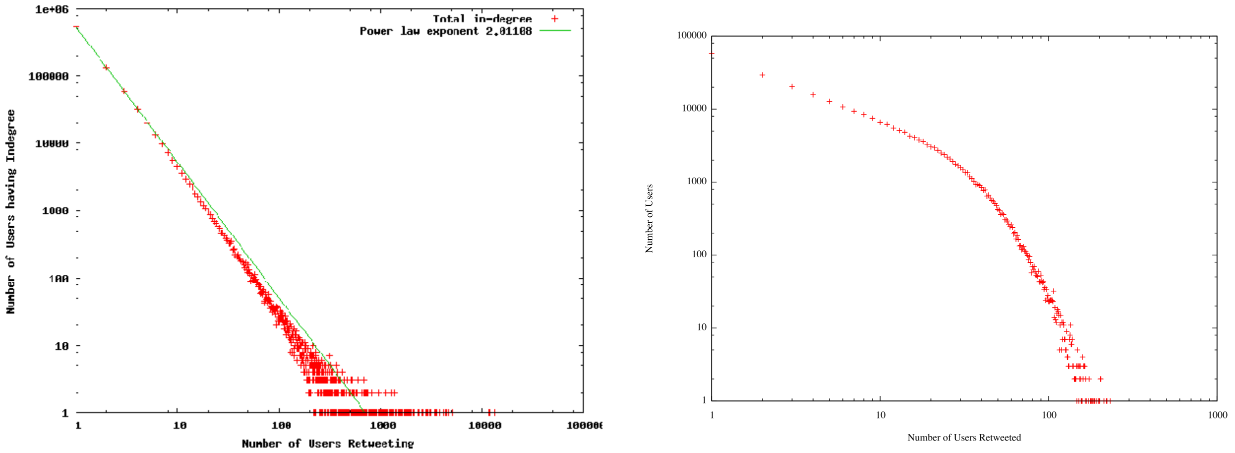
[8] http://help.twitter.com/forums/10711/entries/68916



(a) Inlinks (followers)

(b) Outlinks (friends)

Figure 1: Follow Inlink and Outlink Distributions



(a) Retweet Inlinks

(b) Retweet Outlinks

Figure 2: Retweet Inlink and Outlink Distributions

users whose first collected post occurred after December 1 and last post occurred on or before December 31.

The plot shows a few interesting trends, including a large group of over 58,000 users who published only a single post during the month. It also shows a large number of users wrote more than 100 posts in the 31 day span.

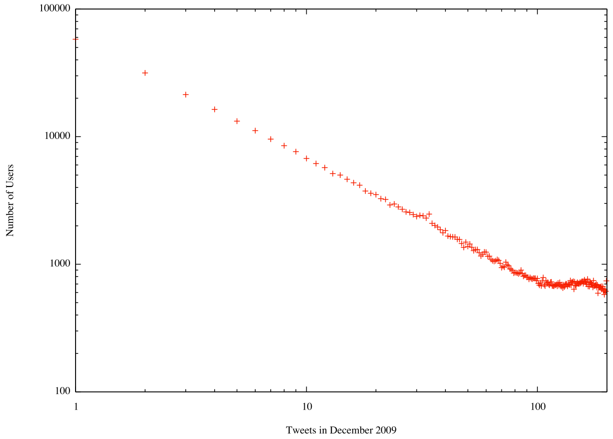


Figure 3: Tweet Frequency

5.2 Graph Formation

In the previous section we examined the general link distributions and observed that most tend to follow a power-law distribution, similar to those found in the traditional Web graph. In this section we take a deeper look at the connections between the inlinks and outlinks, as well the connection between the different edge types.

5.2.1 Readers and Writers

An interesting aspect to consider is the overall posting behavior of a user, and possible connections between the user as a *reader* and the user as a *writer*. Three potential scenarios are: (1) a user acts primarily as a reader (sink) with little or no posts, (2) a user frequently retweets posts of interest but writes little to no original content, acting primarily as a filter of their friend's content, and (3) a user contributes significant new content.

Figure 4 shows the differences between user's reading and writing behavior for a period of 31 days starting mid-October 2009[9]. In this figure, each dot represents a unique user. The x-axis denotes the combined number of posts written by the user's friends, and the y-axis the number of posts published by the user, both of which are log scales. The size of each dot indicates the user's PageRank based on follow-edges (we will discuss more in Section 6 how this is computed), and the shade indicates the "originality" of their posts. The lighter shades indicate less "original" content, meaning a larger percentage of the users posts are in fact retweets instead of new, original content.

The general trend shows that, for users who post very frequently (the upper portion of the graph), a larger fraction of their posts are actually retweets. We also observe an interesting phenomenon: many users retweeted at least one post which they did not read from one of their friends.

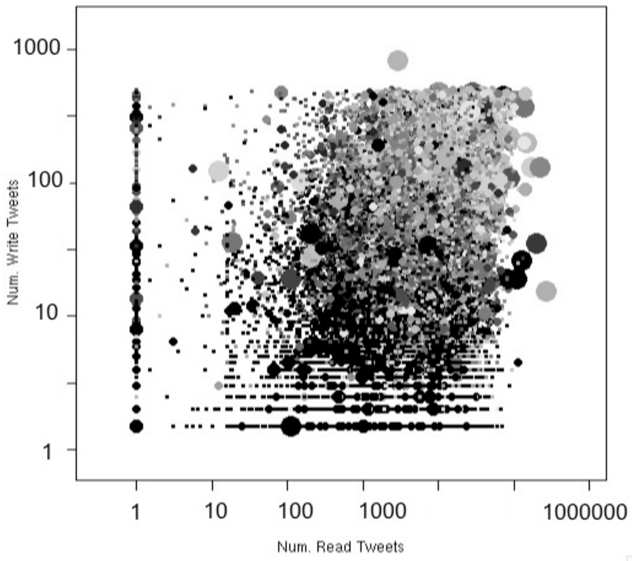


Figure 4: Reading, Writing, PageRank, and Originality

For example, the vertical line at $x = 0$ are users whose friends collectively wrote zero posts, yet many of these users published retweets. This suggests that, despite the explicit friendship links available in the site structure, it is still not possible to know exactly what a user reads. For example, many websites (such as the WSDM2011 homepage[10]) are adding modules which display Twitter results.

6. EXPLORING LINK SEMANTICS

The Twitter graph model we presented in Section 4 includes two main types of entities: users and posts. Exploring the ranking of either would be of interest. In this paper, however, we focus on investigating the semantics of *follows* and *retweets* edges, highlighting their suitability for ranking users for topical relevance. In this section we present a closer look at the semantics for the two types of edges.

6.1 Link Semantics

On the Web, a link from page a to page b signifies an endorsement of the quality of page b, and to some extent its relevance to page a. In the simplified Twitter graph there are two distinct link types, and they each carry different significance.

Similar to a link on the Web, a follow link on Twitter from user a to user b can be understood as an endorsement of quality or interest. However, the semantics of the link actually state that user a, acting as a *reader*, is interested in user b acting as *writer*. This distinction is especially significant for any attempt at a recursive definition of importance which requires link transitivity: if the topics a user writes about are important to you, how important to you, if at all, are the topics which they read?

Compare this with the semantics of a retweet link. A retweet link is also expected to signify an endorsement of

[9] This graph was produced from a smaller crawl of $120k$ users collected between October and November of 2009

[10] http://www.wsdm2011.org/

quality, however in different roles. User a will retweet the posts of user b if he either is interested in writing about the topic or expects his readers to be interested in this post. Thus a retweet edge signifies a connection from user a as *a writer* to user b as *a writer*. We expect this link to carry both an endorsement of quality and that of relevance, and thus carries a stronger topical signal.

6.2 Retweet vs. Follow based Ranking

Based on the above intuition we expect that PageRank computed over different edges will produce significantly different results. On the one hand, follow links (when viewed in a recursive sense) are primarily an indication of importance or "trustworthiness". Retweet links, however, are a more direct indication of *topical* importance or writing "interesting" posts.

In order to gain further intuition of the different semantics that retweet links and follow links carry, we computed PageRank over the following two sub-graphs of the simplified Twitter graph introduced in Section 4.3: a graph consisting of retweet links only, and a graph consisting of follow links only. Their distributions are plotted in Figure 5. The retweet-based ranking displays a relatively simple power-law distribution with a drop around the 7^{th} ranked user. The follow-based ranking distribution also has a drop, around the 14^{th} ranked user, but it appears to consist of two different segments with different power-law coefficients. It is difficult to determine the exact cause of such an artifact.

We view the two rankings against each other in Figure 6. The highest ranked user, according to the follow links, is the current President of the United States, Barack Obama. While his status as a significant public figure alone is sufficient to explain the high ranking, his appearance in Twitter's recommended users list, a list that used to appear during the signup process, likely contributed. This would be another manifestation of the "rich get richer" phenomenon found on the Web [5]. Another interesting thing to note, however, is that according to the retweet-based ranking, he is only the 33^{rd} highest ranked user.

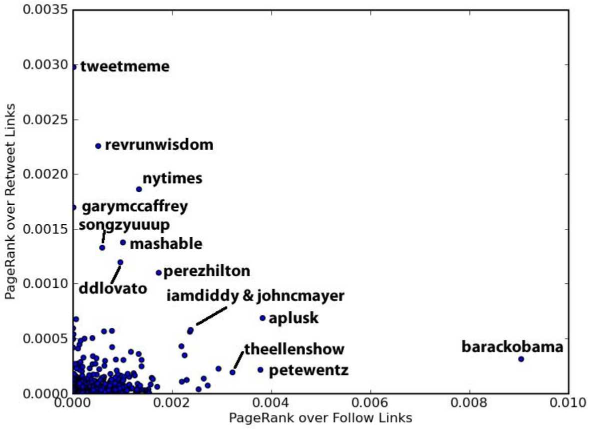


Figure 6: PageRank over Retweet links vs. Follow links

The top user according to retweet-based PageRank is tweetmeme, a site similar in nature to the social bookmarking site digg. Users can easily retweet stories with a single click, and the most retweeted stories are presented on the front page. Tweetmeme also allows embedding of retweet buttons on other sites, further facilitating retweeting of their posts. This helps explain why the site is ranked number one in the retweet rankings while still being relatively low in follow-based PageRank at rank 4, 610.

username	Follow-based	Retweet-based
barackobama	1	32
aplusk	2	9
petewentz	3	54
theellenshow	4	57
the_real_shaq	5	51
mrskutcher	7	87
johncmayer	9	12
iamdiddy	10	15

Figure 7: Top 10 Follow-based and Top 100 Retweet-based

username	Retweet-based	Follow-based
nytimes	3	30
mashable	5	60
ddlovato	7	72
perezhilton	8	15
aplusk	9	2

Figure 8: Top 10 Retweet-based and Top 100 Follow-based

Next we look closer at users who rate highly under both rankings. Figure 7 lists the ranking of the top 10 users according to follow-based PageRank that are also in the top 100 according to retweet-based PageRank. We can see that nearly all of them can be considered "high ranking" public figures or celebrities.

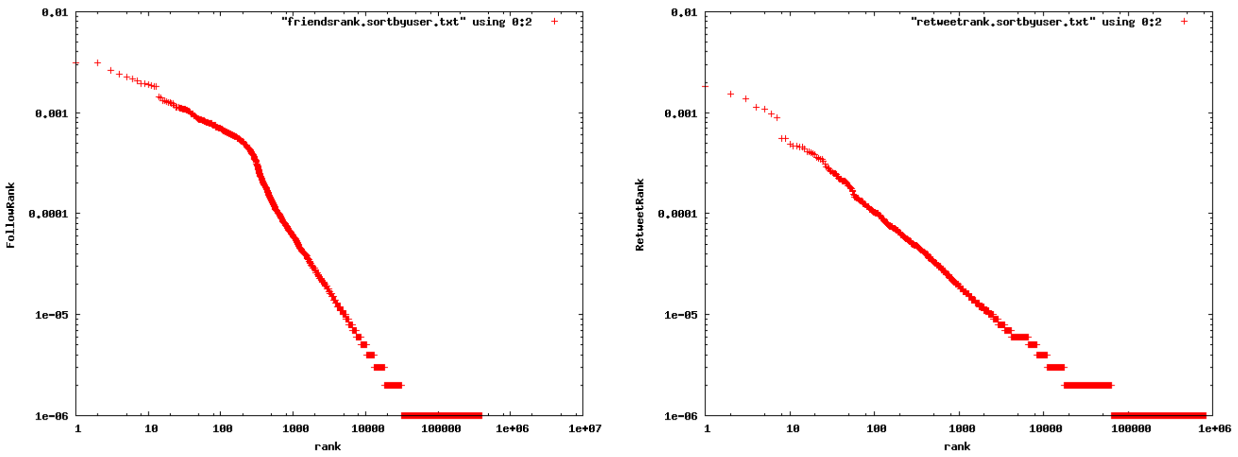
Compare this to the "opposite" side of the figure, users ranking among the top 10 based on retweet links and among the top 100 by follow links. These rankings are summarized in Figure 8. Three of the five: mashable, the New York Times, and Perez Hilton can be classified as news generating entities. Ashton Kutcher is the only user who appears in the top 10 for both rankings.

Finally, there are cases where the rankings appear affected by spam or "marketing" techniques. When examining ddlovato (actress and singer Demi Lovato), the retweets by her followers seems to suggest at least some are in fact set up by marketers. For example, the account ddlovatoRT states that its purpose is to simply retweet all posts mentioning Demi. While spam will always a concern, Twitter's research team estimates that less than 1% of Tweets are now spam[1].

6.3 Link "Virality"

From an empirical analysis of these two different rankings, it appears that follow links capture the quality of a user being popular or well known, while retweet links capture the quality of being influential or producing newsworthy or topically relevant posts. Next we wish to gain some insights into possible relationships between the two link types by examining possible interactions between them.

Define $RoF(u)$ ("Retweeted by Friends") as the set of unique users whose posts have been retweeted by one of the



(a) Follow-based PageRank distribution

(b) Retweet-based PageRank distribution

Figure 5: Follow-based and Retweet-based Ranking

friends of user u. That is, $RoF(u)$ are the users who u has seen at least one post from via a retweet. Let $Fr(u)$ be the set of users whom user u follows. We define *Retweet Virality* as:

$$r_v(u) = \frac{RoF(u) \cap Fr(u)}{RoF(u)}$$

Similarly, define the $FoF(u)$ ("Friends of Friends") as the set of users the friends of u follow. That is, $FoF(u)$ is the set of all users who are reachable by traversing *exactly* two directed follow edges in the Twitter graph, starting from u. *Follow Virality* is then defined as:

$$f_v(u) = \frac{FoF(u) \cap Fr(u)}{FoF(u)}$$

Retweet Virality measures the probability that a follower of user u_a is following user u_b, given that user u_a retweeted a post from u_b. Follow Virality measures the probability that a follower of u_a is following user u_b given that u_a follows u_b.

Figure 9 plots the values of Retweet Virality and Follow Virality for 500 randomly selected users from our dataset. The results show that Retweet Virality is consistently higher than Follow Virality. While this comparison does not measure what direct influence observing retweets might have on a user, the consistently higher Retweet Virality suggests that retweets demonstrate a stronger notion of importance or influence to users. In particular, it suggests that users are more likely to follow people they see retweeted than those who are merely "Friends of Friends". Determining what factors lead to generation of a follow link, and in particular whether observation of retweets plays a direct role, is an intersting area for further research.

7. EXPERIMENTS ON LINK SEMANTICS

We now describe the results of our experiments which demonstrate the topical relevance of links in the Twitter graph. These experiments will show that follow links, even from a set of topically similar users, quickly diffuse into a broad range of topics. Retweet links, meanwhile, remain more concentrated on the original topic. The data used for the graph in these experiments is the same as described in Section 4, with over 1.1 million users, 273 million follow edges, and 2.9 million retweet edges in the graph.

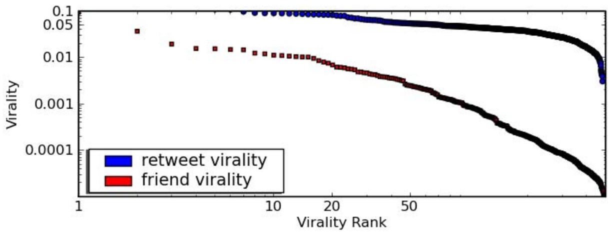


Figure 9: Virality of Retweet and Follow Relationships

7.1 Empirical Results

We performed a preliminary, empirical evaluation on a small data set to give some insight into the characteristics of the links. Starting from a seed set of users who are members of the same topical list (see Section 3), we generate two sets of users. The first set comprises all users who are exactly one *follow* edge away from any of the seed members (that is, at least one seed member follows them). The second set contains the users who are exactly one *retweet* edge away from the seed members (that is, at least one seed member has retweeted one of their posts).

We selected a random sample of 25 users from each of these sets and manually assessed them for topical relevance. We repeated this experiment for two lists, one focused on "photography" and the other on "design". The results show that, while the number of relevant users in the follow-generated samples were 4 and 5, the number of relevant users in the retweet-generated samples were 19 and 20, respectively.

7.2 Topic Sensitive PageRank

The PageRank algorithm, first proposed in [21], describes a recursive ranking formula which, on a high-level, proposes that a page is as important as the pages pointing to it. This ranking can also be explained using the *Random Surfer*

Model, which describes the notion of an abstract surfer who starts on a random Web page p with some pre-set probability t_p. He repeatedly clicks on outgoing links with uniform probability d until he gets "bored". At each step, instead of clicking on a link, the surfer (with probability $1 - d$), chooses to jump to any random page p with probability t_p. The PageRank of a page p is then the probability that a random surfer is on page p at any given time.

The significance of altering or biasing these "jump probabilities" were demonstrated for the purposes of both personalizing the ranking [13] and combating Web spam [10]. The main idea behind these works was that, if you are more likely to start each new surfing session with trusted pages, or pages relating to a certain topic of interest, the highest-ranking pages are more likely to be trustworthy or topically relevant pages, respectively.

The TwitterRank algorithm [26], which makes use of *follow* links, is compared against other ranking strategies including classic PageRank and personalized or "topic sensitive" PageRank (TSPR) for making user recommendations. The authors note that the results are somewhat mixed, and the improvements offered by TwitterRank are not significant in most of their evaluations. Thus for simplicity, we use topic sensitive PageRank to quantify the difference in topical relevance carried by follow and retweet links.

For our second experiment, we use topic sensitive PageRank for ranking users relative to a particular topic. Beginning with a *topical* Twitter list, we compute topic sensitive PageRank over the Twitter graph for both follow and retweet edges individually. Intuitively, if the links carry the "topicality" well, the high-ranking users are likely to be topically relevant to the original seed topic. We evaluate the resulting highest ranked users for relevance to the original topic with a user survey.

7.2.1 Experimental Setup

For our evaluations we manually collected 9 topical lists from `listorious.com`, a directory of popular lists on Twitter. These lists were selected to cover a broad range of topics and tags, such as politics, technology, and economic issues. The lists varied in size from 19 to 437 users, with an average of 155 and median of 49 users. These seed users had an average of $14,284$ followers. For each of these lists we computed personalized PageRank over the two different edge sets for the Twitter graph, using the list members as the trusted seed set. We selected the 30 highest ranking non-seed users for each graph variation.

To evaluate the relevance of these top ranked users to the original topic, we conducted a survey. Participants were shown a topic description along with the 30 highest ranked users for either a follow-based or a retweet-based PageRank, ordered randomly and mixed with a random set of 10 of the seed users for that topic. They were instructed to make a binary judgment of each user's relevance to the topic by inspecting the content of their tweets, biography, username, and any external websites listed on their profile. A total of 12 people participated in the survey. Each list was evaluated by at least 2 people.

7.2.2 Precision of Top Ranked Users

The first metrics we consider evaluate the accuracy of the highly ranked users with respect to the original topic. We define the Precision and Relevance for a set of users U as:

Link	Precision	Relevance
Follow	0.451	0.548
Retweet	**0.601**	0.704

Table 2: Precision and Relevance by Link Type

$$\text{Precision}(U) = \frac{1}{k}\sum_{k}\frac{|R_k(U) \cap U|}{|U|} \quad (1)$$

$$\text{Relevance}(U) = \frac{|\bigcup_k R_k(U)|}{|U|} \quad (2)$$

$R_k(U)$ is the set of users from U judged relevant in evaluation k of a particular list. Precision measures the average relevancy of a set of users, while *Relevance*[11] measures the fraction of users who were judged relevant by at least one survey taker. Table 2 shows the Precision and Relevance for follow links and retweet links, averaged over the 9 different topical lists.

The results show that the overall topical precision of top ranked users can be improved by over 30% by simply using retweet links instead of follows links in the topic sensitive PageRank computation. To verify these results, we computed the statistical significance of the difference in Precision between follow and retweet links with a two-sample t-test. The resulting p-value $p = 0.0446$ suggests that the results are in fact statistically significant with 95% confidence.

7.2.3 Cohesiveness of Seeds

To verify the seed users were an accurate reflection of the intended topic, we included 10 randomly selected seed users for each evaluation. The sample seed users had an average Precision of 0.931 across all 9 topics, with a minimum of 0.838 and a maximum of as 1.0. This indicates that the seed users represented their topics well, and that our survey takers understood and agreed upon the topic definitions.

7.2.4 Popularity of Relevant Users

We observed that the relevant users discovered by retweet links have, on average, fewer followers than those discovered by follows links. While this is a somewhat expected result, we wanted to quantify the difference. We computed the average number of followers for all users identified as relevant by at least one survey taker. For the relevant users discovered by follow-based links, the average number of followers was $257,088$. For retweet-based links, the average was $75,851$. We again computed the statistical significance of the difference between the two, with $p = 0.0011$.

The number of followers a user has is not directly related to their relevance for a particular topic. Topically "interesting" users are not necessarily the most or least popular. It is interesting to note, however, that for both follows and retweets, the average number of followers for the relevant users is higher than the average seed user.

7.3 Discussion

Our initial experiment shows that traversing even a single follow link dramatically reduces the probability of topical

[11] While the metric appears similar to *recall*, we avoid using that terminology here because the complete set of relevant users is not fully known.

relevance. Propagating the topical influence of a user over these follow links is thus problematic, as weight is quickly assigned to irrelevant users. Our experiment with topic sensitive PageRank is effectively a recursive version of this simpler evaluation. With the follow-link structure of Twitter users easily accessible, it is tempting to directly exploit it in any user-ranking or recommendation task. The end result of our experiments show, however, that by propagating a specific type of weight (in this case, a user's topical relevance) over links which are not very likely to carry that type of relationship, we wind up with a less accurate ranking.

8. CONCLUSION

Twitter's importance stems not only from its high traffic ranking, but also the amazingly rich structure it provides and realtime information it makes available. In this paper we have described a detailed model of Twitter as a graph, described key statistics about the graph, and provided some initial insights as to how the graph forms. We have demonstrated important distinctions between edge types in the graph, noting that the varying semantics and properties of these edges will have significant implications on graph algorithms such as PageRank. In particular, we have shown that *retweet* edges preserve topical relevance significantly better than *follow* edges.

In a sense, Twitter could be considered a sign of things to come. We can expect the trend of expanding site-specific, rich, structured data to continue from other sources. For example, large scale RDF data sets such as DBpedia contain millions of nodes and billions of edges conveying a wide variety of semantic relationships. For most ranking applications, these edges should probably not be treated equally. In this paper we have demonstrated the importance of adapting the "tried and true" algorithms, which have proven effective on the Web, to the specific semantics captured in the structured data they are being applied to.

9. ACKNOWLEDGEMENTS

This work is partially supported by NSF grants, IIS-0534784 and IIS-0347993. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institutions.

10. REFERENCES

[1] State of twitter spam: http://blog.twitter.com/2010/03/state-of-twitter-spam.html.

[2] Tunkrank: http://tunkrank.com/.

[3] The web ecology project: http://www.webecologyproject.org/.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[5] J. Cho and S. Roy. Impact of search engines on page popularity. In *WWW '04*, pages 20–29, 2004.

[6] A. Das Sarma, A. Das Sarma, S. Gollapudi, and R. Panigrahy. Ranking mechanisms in twitter-like forums. In *WSDM '10*, pages 21–30, 2010.

[7] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *KDD '00*, pages 150–160, 2000.

[8] K. E. Gill. How can we measure the influence of the blogosphere? In *WWW '04*, May 2004.

[9] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501, 2004.

[10] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB '04*, pages 576–587, 2004.

[11] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *CoRR*, 2008.

[12] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07*, pages 56–65, 2007.

[13] G. Jeh and J. Widom. Scaling personalized web search. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 271–279, New York, NY, USA, 2003. ACM Press.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[15] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08*, pages 19–24, 2008.

[16] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03*, pages 568–576, 2003.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Computer Networks*, pages 1481–1493, 1999.

[18] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM*, 2007.

[19] Y.-M. Li, C.-Y. Lai, and C.-W. Chen. Identifying bloggers with marketing influence in the blogosphere. In *ICEC '09*, pages 335–340, 2009.

[20] R. L. P. Melville, V. Sindhwani. Social media analytics: Channeling the power of the blogosphere for marketing insight. In *WIN '09*, 2009.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[22] J. C. R. Almeida, B. Mozafari. On the evolution of wikipedia. In *ICWSM '07*, 2007.

[23] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM '10*, 2010.

[24] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *HLT-NAACL*, 2010.

[25] K. Sia, . C. J, K. Hino, Y. Chi, S. Zhu, and B. Tseng. Monitoring rss feeds based on user browsing pattern. In *ICWSM '07*, 2007.

[26] J. Weng, E.-p. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *WSDM '10*, 2010.