We Know Who You Followed Last Summer: Inferring Social Link Creation Times in Twitter

Brendan Meeder
Carnegie Mellon University
Pittsburgh, PA, USA
bmeeder@cs.cmu.edu

Brian Karrer
University of Michigan
Ann Arbor, MI, USA
karrerb@umich.edu

Amin Sayedi
Carnegie Mellon University
Pittsburgh, PA, USA
ssayedir@andrew.cmu.edu

R. Ravi
Carnegie Mellon University
Pittsburgh, PA, USA
ravi@cmu.edu

Christian Borgs
Microsoft Research
Cambridge, MA, USA
borgs@microsoft.com

Jennifer Chayes
Microsoft Research
Cambridge, MA, USA
jchayes@microsoft.com

ABSTRACT

Understanding a network's temporal evolution appears to require multiple observations of the graph over time. These often expensive repeated crawls are only able to answer questions about what happened from observation to observation, and not what happened before or between network snapshots. Contrary to this picture, we propose a method for Twitter's social network that takes a single static snapshot of network edges and user account creation times to accurately infer when these edges were formed. This method can be exact in theory, and we demonstrate empirically for a large subset of Twitter relationships that it is accurate to within a few hours in practice.

We study users who have a very large number of edges or who are recommended by Twitter. We examine the graph formed by these nearly 1,800 Twitter celebrities and their 862 million edges in detail, showing that a single static snapshot can give novel insights about Twitter's evolution. We conclude from this analysis that real-world events and changes to Twitter's interface for recommending users strongly influence network growth.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement, Theory

Keywords

online social networks, network evolution, graph analysis, large-scale data collection, user behavior

1. INTRODUCTION

Twitter is a popular social networking website that enables users to send and receive short messages of at most 140 characters, which are also called *tweets*. Tweets are not highly directed messages like email, but are instead broadcast to all of a user's *followers*. Following is the sole social connection in Twitter; a user's primary view in Twitter is a reverse chronological stream of tweets from accounts that user is following. Academic studies of Twitter typically represent the user population as a directed graph (or directed network) because the following relationship can be, and often is, asymmetric.

Information about Twitter can be gathered from the open Twitter Application Programming Interface (API) [1] which provides access to a broad range of information including both tweet content and the current social graph. Despite providing timing information on most other accessible data, Twitter does not provide the time when a user u starts following another user v, which we call the *follow time* of u (v will be clear from the context). However, the Twitter API does return a user's followers *in the reverse order in which they started following that user*. We specifically exploit this extra structure to estimate when edges were created.

Unlike smaller social networks for which recrawling or continuous observation of the social graph is feasible, even a single crawl of a relatively small fraction of Twitter can be a time-consuming enterprise. A major contribution of this work is a simple method to estimate follow times that only requires one static social network snapshot and the time at which certain user accounts were created. For any edge in the network, the method assigns a time that is a *lower bound* for the time the edge was created. Despite only using one crawl, we show that for users who rapidly gain followers the process of assigning times, which we call *timestamping*, can be extremely accurate both in theory and in practice. We emphasize that for any follow-relationship in the graph, our method outputs lower bounds for the actual creation time of that edge. The inferred timestamps can always have arbitrarily large error; however, as we show theoretically and validate empirically, the error of inferred timestamps for popular users is quite small.

Fortunately, Twitter has many interesting users who rapidly gain followers. Most users only have a few followers but some accounts on Twitter have garnered an enormous number of followers. These popular accounts, which can gain thousands of new followers per day, include real-life celebrities such as Lady Gaga and Justin Bieber, politicians such as President Barack Obama and former vice president Al Gore, and news media such as CNN Breaking News and The New York Times. Twitter also promotes several hun-

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

dred accounts through its *suggested users list*. New users are encouraged to follow these suggested users as an introduction to Twitter, and the lucky users placed on the suggested users list gain elevated numbers of followers per day.

How does the rate of accumulation of followers change over time for these prominent users in Twitter? What are the key factors that influence these changes? What is the pattern of users following celebrities in relation to their account creation times and can this pattern for existing users tell us anything about the importance of celebrities to the Twitter graph? For all these questions, we need accurate temporal information about edge formation that is not available from Twitter's API. In Section 3 we explain the simple timestamping method that estimates edge creation times in Twitter, proving its good theoretical properties and demonstrating explicitly how the error should decrease as a function of follow rate. We validate this method in practice on the members of the suggested users list and Twitter celebrities, finding it to be very accurate and robust to link deletions over time in Section 4.

We use the inferred times to answer many of the above questions through a detailed study of the temporal properties of this Twitter subgraph. Our analysis reveals the importance of the Twitter interface in driving followers to the subgraph and case studies indicate the qualitative magnitude of these effects during different phases of the interface. Examining the distribution of inferred timestamps reveals that more than half of the edges in this subgraph, which constitutes a non-negligible fraction of the total number of edges in Twitter's social network, formed within one month of the following user joining Twitter. Finally, we demonstrate through several examples that real-world events can correlate strongly with the attractiveness of a celebrity to followers. These results, all temporal in nature, are captured by a single network snapshot in combination with the timestamping method.

2. BACKGROUND

Online social networks have attracted much attention as topics for academic study [12, 16]. Many recent papers have demonstrated that online social networks have some of the typical characteristics of real-world networks [18] including short path-lengths, clustering, and heavy-tailed distributions in the number of connections. These distributions are often claimed to follow a power-law, although this requires careful study [4].

Twitter, as one of the major online social media websites, has not escaped scrutiny. Communities in Twitter's social network who tweet about similar topics and interests were studied by [9]. Huberman et. al. [8] studied the interaction patterns underlying the social network, suggesting that only a portion of the edges matter for communication over Twitter. Parts of the graph collected under three separate methodologies were analyzed and compared in [11]. More recently, a network analysis based on data collected through breadth-first search was performed by [13] who found a non-power-law follower distribution and low following reciprocity. None of these studies captured the whole of Twitter's social network though, and a discussion of whether network measures are robust under imperfect data is contained in [3].

The interest in online social networks goes well beyond static network analysis. Questions regarding the dynamical evolution of a social network are often very interesting, but also difficult to answer. The dynamical social networks of Flickr and Yahoo! 360 were studied in [12] which had access to precise event times, like those we wish to recover for Twitter. Learning the time intervals in which events occur only from repeated crawling can result in bias for studying certain influence models over social networks [5]. In [15], several networks were shown to densify over time, with the number of edges growing superlinearly with the number of vertices, and average distances shrunk with network size. These novel insights, which contradicted standard views, were not possible without temporal data.

The empirical analysis of mechanisms for network growth (cf. [6]) also requires such data and has occurred at two scales. Macroscopic observations such as that done by [10, 20] found that preferential attachment, a particular mechanism, does appear to hold in certain empirical networks. Microscopic investigations of social networks, at the scale of individual edge placement, has recently been suggested by [14] who compute the likelihood of a host of network formation mechanisms, although not for Twitter. A specific investigation of Twitter performed by [21] demonstrated the importance of triangle closure in formulating ties. At a smaller scale, triangle closure among many other tie formation mechanisms was investigated in [7]. However, the large-scale study of mechanistic explanations for Twitter's network evolution is limited by the lack of temporal edge placement data from Twitter. Triangle closure is a special case that can be studied with information from the Twitter API directly. We now show how to gather temporal edge placement data for Twitter and bypass this limitation.

3. INFERRING EDGE CREATION TIMES

In this section, we define our timestamping method to infer edge creation times. To understand the procedure, we need to describe the relevant temporal information available from Twitter. There is an API method that returns the current followers of a particular user in the reverse order in which they followed that user. So even though the time at which the network edges were created is not provided, the (local) order of their creation is known. Account creation times are also available through the API.

We consider each user individually along with their ordered list of followers and the account creation time for each follower. These user creation times along with the edge ordering for a chosen user will be the input to our procedure. Timestamping a collection of users' followers is done through repeated application of the method to each user in turn. Because we apply this method to Twitter's celebrities, for the sake of convenience, we refer to the user chosen for timestamping as a celebrity.

We estimate the edge creation time for any follower of a celebrity by positing that it is equal to the greatest lower bound that can be deduced from the edge orderings and follower creation times for that celebrity. For each follower u of the celebrity, we retrieve the account creation time of u and each user v that followed the celebrity before u did, according to that celebrity's ordered follower list. The maximum of these account creation times is our estimate for the follow time of u. To explicitly demonstrate that this is the greatest lower bound, we begin by defining a few relevant variables.

Let U be the set of all users following a particular celebrity, C_u be the account creation time of $u \in U$, and F_u be the

unknown time at which u starts following the celebrity. Naturally, $C_u \leq F_u$ for all $u \in U$. From the ordered follower list, $F_u \leq F_v$ if and only if u appears before v in the follower list. For an arbitrary user $u \in U$ we define $B(u) \subseteq U$ to be all users $v \in U$ such that $F_v \leq F_u$. The complement of this set, $A(u)$, are the users who follow the celebrity after u. Every $v \in B(u)$ (which includes u) provides a lower bound on F_u because $C_v \leq F_v \leq F_u$. Any user $v \in A(u)$, for which $F_v > F_u$, can not provide such a bound because they could have been created before or after the follow time of u. The maximum over all of the lower bounds provided by each $v \in B(u)$ is our estimate for the follow time of u. This greatest lower bound is denoted by $\hat{F}_u$, and is defined to be

$$\hat{F}_u = \max_{v \in B(u)} C_v. \quad (1)$$

We call any user v who is the argument of this maximum for user $u \in U$ a *record-breaker* for user u. If v is a record-breaker for $u \neq v$ then v is a record-breaker for itself. Finding all record-breakers can then be simply performed by a single sweep over the celebrity's followers recording every user u that has creation time greater than all preceding users. Note that a user is, or is not, a record-breaker for each celebrity that they follow independently.

Our algorithm embodied in Eq. 1 is to identify the record-breakers of the celebrity and assign each follow time to be at the creation time of the most recent record-breaker. One could also consider a scenario in which the estimated times are unrestricted and the quality of the estimate is given by a quantity such as the sum of squared errors. Under such alternatives, other methods, such as interpolating between user creation times, could provide better estimates.

3.1 Theoretical analysis

In this subsection, we demonstrate that under circumstances appropriate to Twitter's celebrities, the actual follow times are concentrated about the estimated follow times using the record-breaker users' creation times. We consider a model of following for a given celebrity: Fix creation times C_u for all users u that will follow the celebrity. For each user u, draw an independent, identically distributed non-negative random variable L_u from an arbitrary latency distribution that represents how long u waits until they decide to follow this celebrity. The probability density function of the latency distribution is given by $\ell(t)$, where $\ell(t)$ allows arbitrarily small latencies. So for each user u the actual follow time is given by $F_u = C_u + L_u$.

For simplicity, the creation times are assumed spaced uniformly with time interval λ between each user and the first user is created at time 0. The sequence of creation times is then $0, \lambda, 2\lambda$, etc. Let $P(F_u - \hat{F}_u > \delta)$ be the probability that the error in the estimated follow time for user u is greater than δ. (Remember that $\hat{F}_u \leq F_u$ so the error is always non-negative.) Our main theoretical result, proved in the appendix, shows the following error bound:

PROPOSITION 1. *Let $\epsilon > 0$. If*

$$\left(\int_{\delta/2}^{\infty} \ell(t) dt \right)^{\delta/(2\lambda)} \leq \epsilon,$$

then $P(F_u - \hat{F}_u > \delta) < \epsilon$ for any user u.

Note that as λ goes to zero, the proposition is satisfied for any δ, implying that the method becomes arbitrarily accurate in this asymptotic limit. This proves that the follow times are accurately estimated by their greatest lower bounds for sufficiently small λ (i.e. high rates of follower creation.)

It is not essential that the latency distribution be identical between users, that the spacing be given by λ, or that the distribution allow arbitrarily small latencies. Fundamentally, if the rate of new user arrival for a celebrity is high as defined by the proposition, then the error in the inferred follow times will be small. On the other hand, if the rate is moderate, then the errors could be quite large and we emphasize that the timestamping method should not be applied haphazardly. In the next section, we present a thorough validation of the method on empirical Twitter data and demonstrate negligible error on a particular subset of Twitter's edges explicitly.

4. EMPIRICAL VALIDATION

Now that we have presented the timestamping method, we evaluate its performance on real data. The method requires two inputs, a map between user identifiers and account creation times and a collection of ordered follower lists for all users that we want to timestamp. In fact, we only require a map between record-breaker user identifiers and their account creation time. Because the number of record-breakers in our data would require an unreasonable number of queries to the API (on the order of millions), we estimate the creation time of record-breakers using a reference set of users. We crawl every 250th user ID for this reference set and use these times to compute the best lower bound on every record-breaker users' creation time, given only the creation times of users in the reference set. The error introduced through this procedure is insignificant because during the period of analysis, hundreds of thousands of new user accounts are being created each day and the amount of time for 250 new users to sign up is on the order of minutes.

Gathering ordered follower lists requires selecting users. While the method provides lower bounds on follow times for any user, a high follow rate is required for these lower bounds to be accurate and therefore timestamping is generally inapplicable to most Twitter users. However, there is a collection of interesting users on Twitter for whom the method is naturally applied because they generally gain followers at a high rate. One such candidate set are those "celebrity" users who already have a large number of followers. We define a celebrity to be any of the 1,000 most followed users on Twitter, according to the website Twitaholic.com. Another candidate set are those users on the suggested user list. We include the accounts on the suggested user list and put off discussing the various implementations of the suggested users list until Section 5. Hereafter, we also refer to these users as celebrities because they are likely to gain followers at a higher rate compared to accounts with a comparable number of followers that are not on the list. Collected from these two sources on September 18, 2010, we have slightly less than 1,800 celebrity users with varying, yet relatively high, rates of following.[1] We call the collection of relation-

[1]Because the list of celebrities selected this way would vary over time, the complete list of celebrities studied here is available upon request.

ships in which any user follows any of the celebrities the *celebrity follower subgraph*.

The assumptions required in the theoretical analysis, while not particularly strong, are not necessarily met by this empirical data. For example, the process of following a celebrity may not be adequately described by the combination of account creation time and latency distribution. Moreover, it is unlikely that users select their following times independently; Romero and Kleinberg [21] have shown that triangle closure influences network formation in Twitter.

We identify two criteria to determine whether the inferred timestamps are useful in practice. First, the inferred follow times should ideally have errors on the order of hours; we do not want to have errors greater than one day. In order to be consistent with the theory, the timestamp errors should decrease as the new follower rate increases. Second, the method must be robust against follower deletions. Follower relationships can disappear because Twitter deletes a spam account that followed the celebrity, or because a user chooses to *unfollow* the celebrity or deletes their account.[2]

Our validation proceeds by testing the two criteria of accuracy and robustness against repeated crawls of the celebrity follower subgraph in Sec. 4.1. Finally, in Sec. 4.2, we determine the celebrities for which the inferred timestamps are highly likely to be accurate arbitrarily far into Twitter's past.

4.1 Evaluating timestamp errors

To measure the maximum errors of the inferred timestamps we perform a crawl of all 1,800 celebrities every thirty minutes for a 220 hour period from September 18, 2010 until September 28, 2010. As said previously, we use the public Twitter API [1] to collect follower information for each user which is returned in *pages* of 5,000 followers per page. By comparing two consecutive crawls taken at times T_1 and T_2, the users who started following a celebrity in the interval $[T_1, T_2]$ can be determined. Since we crawled with a high frequency, it was sufficient to only retrieve the first page of a celebrity's followers which contains their most recent 5,000 followers.

Thus we have a sequential list of users who started following each celebrity, as well as the time interval in which each following occurred, for this period. A total of 23,258,723 follow events occurred in this period and we compare each estimated follow time to the time interval given by the crawls.

Since we know the interval in which the edge was created, but we do not know exactly when the user started following a celebrity, we can only deduce upper bounds on the timestamp error. The upper error bounds for a follow time that happens in $[T_1, T_2]$ is given by

$$E_U(\hat{F}_u) = T_2 - \hat{F}_u. \quad (2)$$

The upper bound is always positive, and diminishes as record-breaking events occur later in each interval. In Figure 1 we plot for each celebrity the maximum and mean errors (i.e. these upper bounds) of the follow events versus the number of followers that celebrity gains during the data collection period.

[2]If a user unfollows a celebrity and follows the same celebrity later, our method would only apply to the second edge creation as the first follow event would no longer be contained in the follower list.

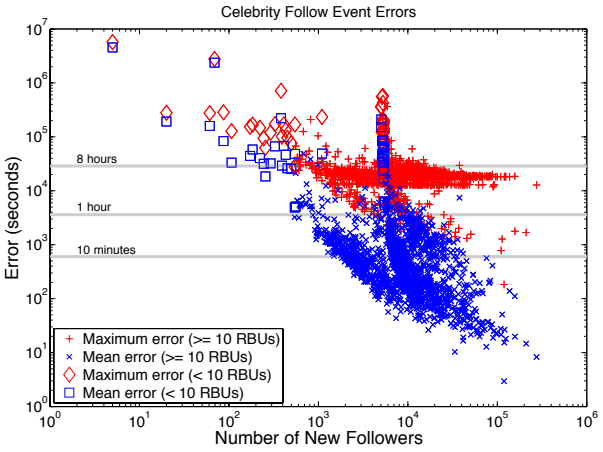

Figure 1: Mean and maximum error (upper bounds) for all celebrities.

The mean and the maximum errors decrease as the number (rate) of new followers increases, in qualitative agreement with the theoretical analysis of Sec. 3. The data is further broken down into points with less than ten record-breaker users (denoted RBU in the caption) and points with more than ten record-breakers. Note that receiving more than one record-breaker a day appears necessary to avoid large maximum errors. For these celebrities, all average timestamp errors are less than three hours, and 97% are less than 15 minutes. The figure clearly shows that the method infers event timestamps with maximum errors that are not particularly large (less than one day) for most celebrities.

The accuracy achieved by the timestamping could be far greater than the thirty-minute resolution of the repeated crawls. In order to test further, we did more rapid recrawls, but because Twitter places a limit on how many API requests can be made per hour, it was not feasible to rapidly recrawl all of the celebrity accounts. Instead we crawled the 25 users with the highest follow rate for every five minutes for 128 hours starting on October 10, 2010. For these very high rate celebrities, the maximum error was a few minutes, showing that remarkable accuracy is possible for timestamping the followers of Justin Bieber and Lady Gaga.

Moving on to our second criteria of robustness, we consider applying our method to network snapshots gathered at two different times. Followers from the first snapshot may no longer exist in the second snapshot due to effects such as unfollowing or account deletion. Users disappearing from a celebrity's follower list could cause the inferred follow times to change. In particular, when a record-breaker user u in one network snapshot no longer follows the celebrity in a later snapshot, follow times determined by u's creation time get reassigned to a smaller creation time which means larger error. Over a long period, the aggregate effect of unfollowings could significantly diminish the method's accuracy.

To test robustness, we take each follower list collected during the coarse-scale crawl and randomly delete each follower with probability 0.5. After the deletions, 1,692 celebrities still have at least ten record-breaker users. For these users, the average increase in the maximum error is slightly more than 5,300 seconds. For all but 26 of these users, the maximum error increased by no more than six hours. The method

is thus very robust to a large amount of statistically independent edge deletions.

Because we remove such a large fraction of events, we are highly confident that the method will continue to accurately timestamp edges created during this period despite the occasional deletion occurring over time. However, it is likely that the rate at which celebrities acquire followers has been changing over time and that the present crawls are not necessarily representative of past performance. In the next section, we discuss how we can apply the timestamping method on historical edge creations for which we do not know the follow rates.

4.2 Historical accuracy

Our analysis has shown that the timestamp method is accurate and robust during time periods where there are high follow rates. However, we wish to apply the timestamping method to every follower of a celebrity, not just those for whom we can guarantee high rates through comparison to repeated crawls. This capability of the timestamping method is one of the most significant advantages over repeated crawls, beyond ease of implementation. Yet how can we judge the method's accuracy in the past?

We answer this question by computing an upper bound on the error that is partially observable from a celebrity's followers. Consider a non-record-breaker user u who is immediately after record-breaker v and immediately before record-breaker z. Then the error of the follow time assigned to user u is

$$F_u - \hat{F}_u = F_u - C_v \leq F_z - C_v = L_z + C_z - C_v. \qquad (3)$$

This upper bound consists of the unobservable latency of record-breaker z and the observable difference in the creation times of the record-breakers.

Without any assumptions on the distribution of the latencies for record-breakers, we cannot provide any guarantees about inferring historical follow times. At best, confidence could be retroactively asserted by showing that the edge creation times are reasonable by other criteria, as indeed occurs in Section 5. We can do better by assuming that the record-breakers in the validation data are characteristic of record-breakers in the past. Of the 23,258,723 follow events in the validation data, about 10% correspond to record-breakers.

We show the maximum upper bound record-breaker error and the average upper bound record-breaker error over all celebrities in Figure 2. Again, we see that like in Figure 1, record-breakers must come with sufficient frequency, at least one per day, to avoid large error.

We now only discuss the 1,748 celebrities that meet this condition (the x and cross symbols). They have maximum record-breaker errors of less than 8.1 hours and 91.6% have maximum record-breaker errors of no more than 2 hours. The average record-breaker error is less than 2.25 hours, and 99% have an average record-breaker error of no more than thirty minutes.

In order to apply Eq. 3 to historical data, we assume that a record-breaker created less than 24 hours before the next record-breaker never has a latency bigger than 20 hours and should be considered accurate. This condition for accuracy forces record-breakers to come sufficiently quickly, which we observed to be an important factor in Figures 1 and 2. With this assumption, we can explicitly evaluate our error bound

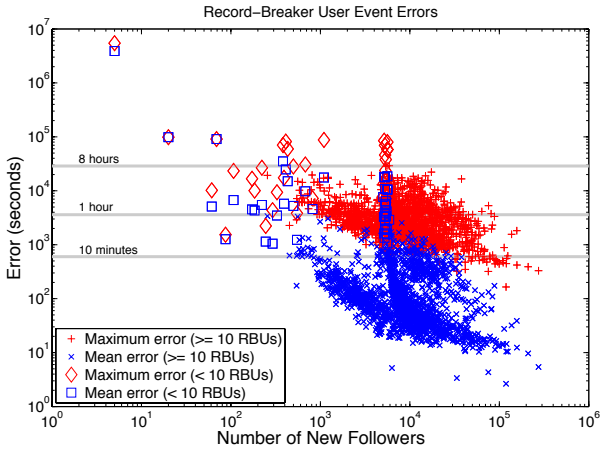

Figure 2: Mean and maximum error (upper bounds) for the record-breaker users over all 1800 celebrities.

on other follow events using Eq. 3 and $L_z = 20$ hours for these accurate record-breakers.

To achieve an error bound of less than a day, the difference in record-breaker creation times must be less than 4 hours, assuming that the later record-breaker is accurate and $L_z = 20$. As said before, an error of less than a day is considered an accurate timestamp for a created edge. Note that the accuracy applies to a particular estimated follow time and not to a celebrity as a whole. The entire procedure applied to each celebrity's edges is then as follows: 1. All record-breaker users that are created less than 24 hours before the next record-breaker are declared accurate. 2. Any non-record-breaker user that follows the celebrity between two record-breakers is accurate if the later record-breaker is accurate and the later record-breaker created their account less than 4 hours after the earlier record-breaker. 3. Any user not covered by either condition is declared inaccurate.

This procedure will denote certain follow times as accurate and others as inaccurate. Roughly speaking, these edges should form temporal regions when the celebrity was and was not gaining followers at a reasonably high rate respectively. Examining these regions could be interesting, but for the purpose of this paper, we focus on those celebrities that contain predominantly accurate timestamps. We call a celebrity's collection of timestamps accurate if it contains 95 percent accurate timestamps.

On October 12, 2010, we crawled the complete follower lists for each of the 1,800 celebrities in the validation data. Originally, we planned on timestamping every one of their followers using the accuracy procedure just described. Unfortunately, due to caching issues on Twitter, some of the follower lists have spurious data, and this necessitated we drop the earliest ten thousand users from the ordered follower list for each celebrity. All analysis hereafter will be of the remaining edges placed between users and celebrities, where these oldest edges have been removed.

Applying our accuracy procedure, we find that 1508 of the celebrities are accurate by this standard which shows that the timestamp method is eminently appropriate for these Twitter accounts.

5. TEMPORAL NETWORK ANALYSIS

In this section, we study the celebrity subgraph formed by the 1508 accurate celebrities found in Section 4.2. What insights can we now gain that would not be possible without knowing when social links were formed? We first perform a broad analysis of the celebrity subgraph in Section 5.1 and then we examine typical accounts in Section 5.2. We focus largely on temporal analyses of this subgraph as this is the novel information provided by our method.

5.1 Broad analysis of celebrity subgraph

There are 74, 184, 348, or about 75 million, unique users who follow at least one of the 1508 accurate celebrities. For reference, we estimate the total number of unique users on Twitter to be around 190 million. So a broad spectrum of user accounts are captured in the subgraph. Some of these unique users are themselves celebrity accounts, so the subgraph is not entirely bipartite. Celebrities do follow each other.

The accurate celebrity subgraph has a total of 835, 117, 954, or about 835 million, directed edges in it which is actually a non-negligible fraction of edges in Twitter's social graph. A recent study of Twitter as a whole, gathered by breadth-first search, collected 1.47 billion edges in total [13]. An estimate of the total number of edges by the present authors suggests there are around 7 billion edges in the present social graph.

The left window of Figure 3 displays the fraction of celebrities with greater than k followers as a function of k. Around 20% of the accurate celebrities have more than a million followers. The right window of Figure 3 displays the fraction of users following k celebrities as a function of k on a log-log scale. One feature that stands out is the existence of three peaks in the distribution at following 20, 241, and 461 celebrities.

We have been unable to precisely determine the cause of the 241 and 461 peaks, but following 20 celebrities has a simple explanation. It is due to the original formulation of the suggested users list. The suggested users list, in its original design, gave new users the opportunity to automatically follow 20 users randomly selected from a pre-selected collection of users. The default option was to follow all 20 users, but one could click this off to follow a particular subset. The motivation behind the suggested users list was to provide interesting (hand-picked by Twitter) accounts for a new user to follow. According to this article [26], the suggested users list on July 16, 2009 had 241 users on it which is probably the cause of the peak at 241 celebrities. We have been unable to determine if at some time the suggested user list had 461 accounts on it. These peaks constitute prominent evidence that Twitter's interface has dramatically affected the celebrity subgraph.

Further indications can be seen in Figure 4 where the blue curve shows the number of edges created in the accurate celebrity subgraph per hour as a function of time. We have labeled three distinct changes in this total celebrity follow rate.

These changes correspond to three distinct adjustments to Twitter's user interface. The first label (1) is the introduction of the suggested users list which occurred around February 2009 [24, 22]. Using the account creation times of the users who follow 20 celebrities suggests that the actual date was Feb. 13, 2009, when there was a large upward surge in following 20 celebrities. Label (2) shows when the

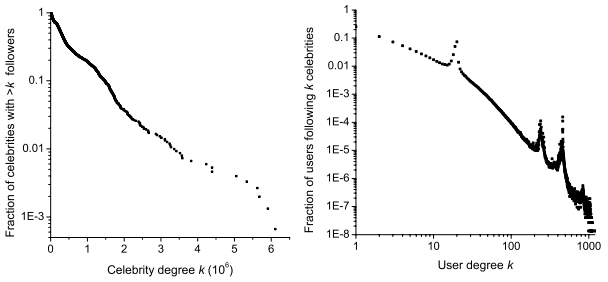

Figure 3: Left side: The complementary cumulative distribution function for the number of followers of a celebrity. Note that this is a log-linear scale. Right side: The distribution of the number of celebrities followed by a user plotted on a log-log scale. Notice the three peaks at $k = 20, 241$ and 461.

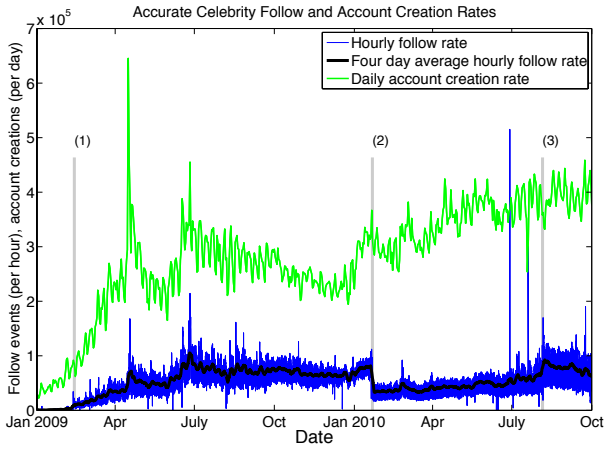

Figure 4: The total celebrity follow rate (follow events per hour) and Twitter account creation rate (accounts created per day) over time. The three labels correspond to the introduction of the suggested users list, the update to the suggested users list, and introduction of "users you may be interested in". The black smoothed curve shows a four day average of the celebrity follow rate.

old suggested user list was changed to its current format on Jan. 21, 2010 [25] at which point the number of followers drops dramatically. The updated format displays a number of categories such as science and entertainment and a new user is encouraged to follow suggested users corresponding to their interests.[3] Much of the drop in volume that occurs on Jan. 21, 2010 is due to the suggested users list no longer defaulting to follow 20 celebrities. Correspondingly, there is a sharp decline in the number of users following 20 celebrities after Jan. 21, 2010.

The last change (3) is due to the introduction of the "users you may be interested in" (or "Suggestions for You") feature which was rolled out on July 30, 2010 [23]. This feature suggests accounts to existing Twitter users that they might

[3]The suggested users list could also be reached from the Twitter homepage in both of its implementations.

want to follow. We see another upsurge in celebrity follow rate around the same time.

One possible explanation for these rapid changes is that the introduction of a feature, or change in user recommendation system, by Twitter adjusts the rate at which accounts are created. We test this hypothesis by computing the rate at which accounts were created for Twitter, shown in the green curve of Figure 4. While there is perhaps a slightly contemporaneous increase in total celebrity follow rate and account creation when the old suggested user list is introduced, the increase in user creation is not sustained. Similarly, the change in follow rate due to the switch from old to categorical suggested user list and introduction of "users you may be interested in" is not explained by changes in account creation. Since the creation rate of Twitter accounts is unable to account for the changes in celebrity follower rate due to altered Twitter features, the more plausible explanation is instead that these features altered how users discover and follow celebrity accounts.

In order to analyze these effects further, we examine several typical accurate celebrities on the suggested users list as case studies in the next section.

5.2 Impact of the Suggested Users List

Given that the overall celebrity follow rate halved when Twitter switched to the categorical suggested users list, it is clear that being on the suggested users list increases the acquisition of new followers substantially. Anil Dash, a tech blogger and entrepreneur, has written about his experiences being on the old version of the suggested users list [2] and is an illustrative example.

At the time of our data collection, Mr. Dash had 332699 followers in total. In figure 5, we show the fraction of Mr. Dash's follow events per day using the inferred timestamps.

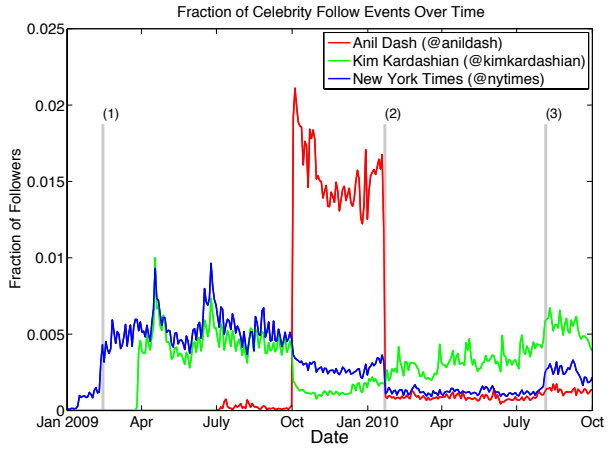

Figure 5: The fraction of follow events for each celebrity per day as a function of time. The three labeled grey lines are the times of the interface changes described in Sec. 5.1.

Very shortly after being put on the old suggested user list on Oct. 2, 2009, Mr. Dash's rate of gaining followers increased greatly. During his time on the old suggested user list, he gained around 2,500 new followers per day compared to his previous average of about 50 per day. When Twitter transitions to the categorized suggested user list, his following rate drops significantly to around 100 followers per day. Interestingly, this is still higher than before his presence on the old suggested user list. We consider two possible explanations for this continued popularity. Many models of network formation assume that edges are "sticky" in the sense that gaining followers increases the rate at which you will gain followers in the future. It is reasonable that the large number of followers gained from being on the old suggested users list had this effect for Mr. Dash. Alternatively, his account could have been present immediately in the categorized suggested users list and this mechanism could account for the additional followers. Mr. Dash is (as of October 20, 2010) in the technology category of the suggested users list, but as the list changes over time, we cannot say if he was on the list in January. A smaller, but still evident, increase in follower rate to around 200 followers per day on average occurs during the introduction of the "users you may be interested in" feature. This increase is not nearly the boost given by the old suggested users list, but it is certainly non-negligible.

Also shown on the figure are the corresponding curves for the New York Times and Kim Kardashian. The New York Times account was created before the old suggested users list and immediately benefits from its introduction at label (1). Kim Kardashian apparently was placed onto the list shortly after her account was created as her curve tracks the New York Times fairly closely during the time of the old suggested users list. In October, when Mr. Dash is placed onto the suggested user list, both @nytimes and @kimkardashian drop in their follow rate. It could be that the suggested users list expanded (perhaps to 461 from 261 accounts) or they were removed from the suggested users list. Judging by the sharp decline in @nytimes fraction at (2), it was likely on the suggested users list with Mr. Dash. Then finally the introduction of "users you may be interested in" benefited @nytimes and @kimkardashian, although again not as much as the old suggested users list. These case studies illustrate that a wide range of different Twitter celebrities experienced similar follow behavior due to the interface.

Besides knowing when edges are created, we are also interested in how long users wait to follow celebrities after they join Twitter.

5.3 Measuring following latency

In our theoretical analysis, users' following behavior is determined by a latency distribution. We examine the actual latency of users, the differences between their account creation time and following time. Because our data only contains users who have followed the celebrities when the network snapshot is taken, early users may exist who will follow the celebrities in the future and have long latencies. Ignoring these users, and their long latencies, would bias any attempt to empirically determine the latency distribution, especially because we cannot identify which users will ever decide to follow a celebrity.

So instead we measure the conditional probability that a user waits t seconds to follow the celebrity given that they follow the celebrity within a month of account creation. In Figure 6, this unnormalized probability is estimated by the number of follow events derived from users created more than a month ago on a log-linear scale in hourly bins. The large concentration at zero latency is caused by the set of record-breaker users. Of those users who follow within a

month, 86 percent follow within 24 hours and 90 percent follow within six days. If a user is going to follow a celebrity within a month of joining Twitter, they are most likely nearly immediately after joining.

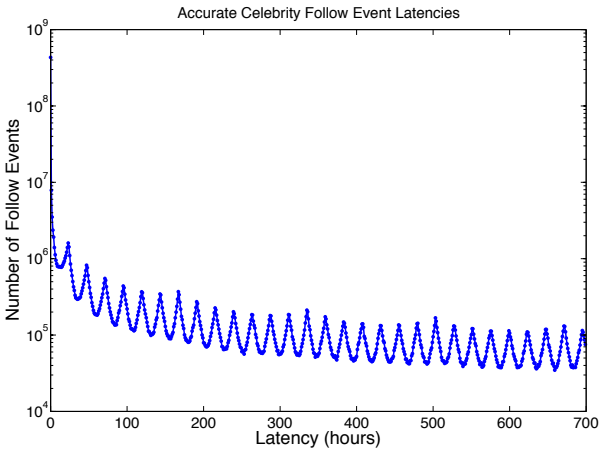

Figure 6: The number of follow events binned by hour as a function of latency for the follow events of users created before September 1, 2010.

The periodicity in the distribution occurs over 24 hour intervals and this diurnal cycle is likely due to the users logging in within several hours of when they created their account.

We check this interpretation in Figure 7 which is a heatmap on a log-scale showing the number of follow events created during the hour on the y-axis for users created during the hour on the x-axis. We only include latencies greater than a day to eliminate the large contribution due to the record-breakers. This figure is consistent with our interpretation of the latency distribution as it is runs along the diagonal with variations of a few hours in either direction. Moreover, the peak along the diagonal indicates that 4-10 pm EST is a popular time to both follow celebrities and create accounts, reflecting that the population of Twitter users are largely focused in the United States.

The fraction of each celebrity's followers who followed the celebrity within a month of joining Twitter varies widely over the celebrities with an average of 65% and a standard deviation of 18%. This large fraction of each celebrity's followers translates into nearly 580 out of the 835 million edges with latency less than a month. If we change the scale from a month to a day, on average 48% of a celebrity's followers followed them within a day. Again translated into edges, about 451 million edges have latency less than a day. In fact, about 140 million of the edges are due to record-breakers and hence are given a latency of zero. While old users do follow celebrities with occasionally large latency, low latency edges are dominant.

5.4 Celebrity popularity and real-world events

We have seen that the rate at which accurate celebrities gain followers is plausibly changed by adjustments to Twitter's interface. In this section, we examine whether the rate at which a celebrity receives followers could also plausibly be changed by real-world events.

For our first demonstration of a plausible real-world event

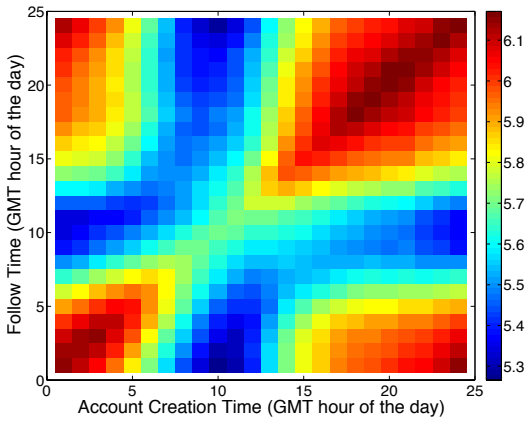

Figure 7: A heatmap of the creation time versus follow time over all celebrities with latencies greater than one day on a log-scale. The hours represent the GMT timezone.

that changed Twitter, during the Iran election in late June 2009, Twitter became a vehicle of communication among Iranian internet users planning protests and rallies. Twitter was popularized by the mainstream media at this time, and we witness a sharp increase in the number of new accounts in July 2009. Conveniently for our purposes, celebrities often show up in the national news for particular events such as political rallies, concerts, or sporting events. Are such single day events important to the temporal evolution of Twitter's celebrity follower subgraph?

It is not effective to analyze absolute follow rates to answer this question because the absolute rate depends on the total rate of user account creation which varies substantially as shown in Figure 4 with occasionally sharp changes. To compensate for such overall variation, we consider whether the relative rate, which we call relative popularity, of a celebrity changes due to real-world events.

The relative popularity $f_i(t)$ is an estimate of the probability that a user who follows a celebrity at time t decides to follow celebrity i. This relative popularity is normalized so that $\sum_i f_i(t) = 1$, where the sum is over all celebrities and the relative popularity is zero for a nonexistent celebrity at time t. We compute it using the following sliding window:

$$f_i(t) = \frac{|\text{Connections to } i \text{ within } t-\Delta \text{ and } t+\Delta|}{|\text{Edges created within } t-\Delta \text{ and } t+\Delta|}, \quad (4)$$

where the variation of $f_i(t)$ is assumed to be at a longer time-scale than window width Δ. We checked several values Δ to ensure consistent results and decided to use a window width equal to a week with t samples spaced per day. A useful comparison is the relative popularity if followers were placed randomly, which is simply $1/n(t)$ where $n(t)$ is the number of celebrities that exist at time t.

We computed these curves utilizing the top 50 celebrities and in Figure 8, we display the resulting relative popularity values for five of the most popular celebrities. These values are clearly varying over time, and are far from the predictions of random attachment represented by the black

line.[4] The behavior of the relative popularity when a new celebrity joins Twitter differs widely. Oprah Winfrey and Ellen DeGeneres (not shown), for example, have a quick spike upwards in relative popularity, but Justin Bieber begins with a small relative popularity that gradually increases over time. The relative popularity shows large variations, including several prominent peaks and drops that are not due to Twitter's interface. One such drop is near June 25, 2010 (arrow 5) where the rapper Soulja Boy (not shown), gained roughly half a million followers over a few days, garnering a relative popularity value of nearly twenty-five percent. A search of blog posts and news articles reveals that Soulja Boy deleted his Twitter account called @SouljaBoyTellEm and switched to an account called @SouljaBoy. One explanation is that these users followed Soulja Boy from his previous popular account. Alternatively, the hashtag #IfSouljaBoyWasARapper was a trending topic, which means that tweets containing the phrase #IfSouljaBoyWasARapper were extremely popular on Twitter around June 25. While the humor was decidedly unfavorable to Soulja Boy, these tweets may have had a positive effect on his relative popularity.

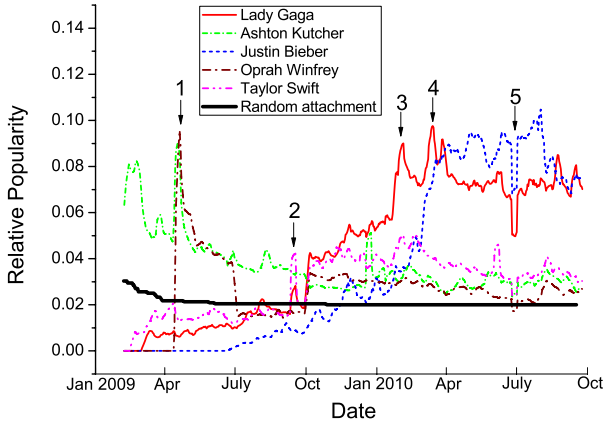

Figure 8: The relative popularity as a function of time for five celebrities. The random attachment prediction is shown in bold. Labeled arrows correspond to events discussed in the text.

For many other cases, we can also identify spikes in relative popularity as corresponding to real-world events that plausibly explain increased Twitter popularity. For example, Lady Gaga performed at the Emmy's on Feb. 1st, 2010 (arrow 3) and released her music video "Telephone" on March 13, 2010 (arrow 4). Even more interestingly, the peaks that occur simultaneously for several celebrities appear to be due to events involving them together. On Friday April 17, 2009 (arrow 1) Ashton Kutcher, who just succeeded in reaching one million Twitter followers before CNN Breaking News, appeared on Oprah's TV show, during which she joined Twitter [19]. They both received large boosts in relative popularity from this event and we suspect that Ashton and Oprah are collectively responsible for the largest gain in Twitter accounts ever that occurred on this day in April (see Figure 4). Lady Gaga also performed at the MTV Video Music Awards on Sept. 12, 2009 (arrow 2) along with Taylor Swift and Katy Perry (not shown). [17] All three of them show an increase in relative popularity at this time. Unfortunately, Kanye West, who was involved in an infamous incident with Taylor Swift that evening, was not on Twitter at the time.

[4]We also compared to a preferential attachment model and it did not capture the observed relative popularity either. We omit these figures because the predicted values vary by celebrity.

6. CONCLUSIONS

We have devised a simple and effective method for inferring follow times in the Twitter social network that has several distinct advantages over other ways of recovering this information. We are able to accurately and robustly infer link creation times using only a single crawl of the social network and user creation times. Furthermore, we are able to recover follow times arbitrarily far into Twitter's history. For the most popular users in Twitter's social network, the method was accurate to within several minutes.

Using the timestamp information, we recreated the evolution of the Twitter celebrity subgraph and gained temporal insights to user following behavior including the distribution of latencies, the importance of the Twitter interface, and the possible influence of real-world events. Overall, our approach gave us a much deeper insight into the structure and evolution of a significant and large subgraph of the Twitter social network.

Our work opens several possible avenues for future investigations. The interaction of user interface with Twitter network structure deserves detailed investigation beyond the results provided here. For example, we neglected to consider the categorization of the users in the current suggested users list. Another important factor could be where a suggested user is listed for a given category. More speculatively, is it possible to confirm the influence of external real-world events on Twitter's network structure? If this influence is strong, any network evolution mechanism that is completely internal to the network would likely fail to describe the Twitter network fully.

7. ACKNOWLEDGEMENTS

The authors thank Ray Reagans and the anonymous referees for useful discussions and comments. Part of this research was performed while the first four authors were at Microsoft Research New England during summer 2010. Brendan Meeder is supported by a National Science Foundation Graduate Research Fellowship.

APPENDIX

A. PROOF OF PROPOSITION

Proof: Pick an arbitrary user u to compute the error probability. We start from

$$P(F_u - \hat{F}_u > \delta) = \int_0^\infty P(F_u - \hat{F}_u > \delta | L_u = t)\ell(t)dt.$$

Since $\hat{F}_u \geq C_u$, the error is at most equal to t for fixed $L_u = t$, we can change the bottom limit of integration to δ to get

$$P(F_u - \hat{F}_u > \delta) = \int_\delta^\infty P(F_u - \hat{F}_u > \delta | L_u = t)\ell(t)dt.$$

Consider a fixed value t and define $N_1(u)$ as the set of users v such that $F_u - \delta/2 < C_v < F_u$ and $N_2(u)$ to be the set of

users v such that $F_u - \delta \leq C_v \leq F_u - \delta/2$. The probability $P(F_u - \hat{F}_u > \delta | L_u = t)$ is equal to the probability that all these users have $F_v > F_u$. If that happens, all of these users are in $A(u)$ and therefore, $\hat{F}_u = \max_{v \in B(u)} C_v < F_u - \delta$. The condition $F_v > F_u$ for fixed $L_u = t$ is met if $L_v > C_u - C_v + t$. Let $Q_{uv}(t) = C_u - C_v + t$ and $P(L > x) = \int_x^\infty \ell(t)dt$. Then we can express the conditional error probability as

$$P(F_u - \hat{F}_u > \delta | L_u = t) = \prod_{v \in N_1(u) \bigcup N_2(u)} P(L > Q_{uv}(t)).$$

We upper-bound this expression as follows:

$$P(F_u - \hat{F}_u > \delta | L_u = t) \leq \prod_{v \in N_2(u)} P(L > Q_{uv}(t)).$$

Note that $Q_{uv}(t) \geq \delta/2$ for $v \in N_2(u)$. Then

$$\begin{aligned} P(F_u - \hat{F}_u > \delta | L_u = t) &\leq P(L > \delta/2)^{|N_2(u)|} \quad (5) \\ &= P(L > \delta/2)^{\lfloor \delta/(2\lambda) \rfloor}. \end{aligned}$$

This bound no longer depends on t. So

$$\begin{aligned} P(F_u - \hat{F}_u > \delta) &\leq \int_\delta^\infty P(L > \delta/2)^{\lfloor \delta/(2\lambda) \rfloor} \ell(t)dt \quad (6) \\ &< P(L > \delta/2)^{\lfloor \delta/(2\lambda) \rfloor + 1} \\ &< P(L > \delta/2)^{\delta/(2\lambda)}, \end{aligned}$$

which completes our proof.

B. REFERENCES

[1] About Twitter API. `http://dev.twitter.com`.

[2] Anil Dash's experience on the suggested users list. `http://dashes.com/anil/2009/12/life-on-the-list.html`.

[3] S. P. Borgatti, K. M. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28:124–136, 2006.

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

[5] D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, X. Lan, and S. Suri. Sequential influence models in social networks. In *ICWSM*, 2010.

[6] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079–1187, 2002.

[7] S. A. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Proceedings of the Second IEEE International Conference on Social Computing*, 2010.

[8] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *Social Science Research Network Working Paper Series*, December 2008.

[9] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[10] H. Jeong, Z. Néda, and A. L. Barabási. Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(4):567–572, 2003.

[11] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.

[12] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, 2006.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media. In *WWW'10: Proceedings of the 19th internation conference on World Wide Web*, pages 591–600, 2010.

[14] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.

[15] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.

[16] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[17] MTV Video Music Awards 2009. `http://www.mtv.com/ontv/vma/2009/`.

[18] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[19] Oprah Tries Twitter, Crowns Ashton King of It. `http://blogs.wsj.com/digits/2009/04/17/oprah-tries-twitter-crowns-ashton-king-of-it/`.

[20] S. Redner. Citation statistics from 110 years of Physical Review. *Physics Today*, 58:49–54, 2005.

[21] D. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proc. 4th Internation AAAI Conference on Weblogs and Social Media*, 2010.

[22] The Newest Way to Game Twitter Fake Followers. `http://brooksbayne.com/post/79132853/the-newest-way-to-game-twitter-fake-followers\#comment-6353220`.

[23] Twitter blog post: Discovering Who To Follow. `http://blog.twitter.com/2010/07/discovering-who-to-follow.html`.

[24] Twitter blog post: Suggested Users. `http://blog.twitter.com/2009/03/suggested-users.html`.

[25] Twitter blog post: The Power of Suggestion. `http://blog.twitter.com/2010/01/power-of-suggestions.html`.

[26] Who Does Twitter Love? Breaking Down The Twitter Suggested Users List. `http://searchengineland.com/who-does-twitter-love-breaking-down-the-twitter-suggested-users-list-22640`.