Inferring User Interests from Tweet Times

Dinesh Ramasamy, Sriram Venkateswaran and Upamanyu Madhow
Electrical and Computer Engineering Department
University of California Santa Barbara
{dineshr, sriram, madhow}@ece.ucsb.edu

ABSTRACT

We propose and demonstrate the feasibility of a probabilistic framework for mining user interests from their tweet times alone, by exploiting the known timing of external events associated with these interests. This approach allows for making inferences on the interests of a large number of users for which text-based mining may become cumbersome, and also sidesteps the difficult problem of semantic/contextual analysis required for such text-based inferences. The statistic that we propose for gauging the user's interest level is the probability that he/she tweets more frequently at certain times when this topic is in the "public eye" than at other times. We report on promising experimental results using Twitter data on detecting whether or not a user is a fan of a given baseball team, leveraging the known timing of games played by the team. Since people often interact with others who share similar interests, we extend our probabilistic framework to use the interest level estimates for other users with whom a person interacts (by referring to them in his/her tweets). We demonstrate that it is possible to significantly improve the detection probability (for a given false alarm rate) by such information pooling on the social graph.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Time series analysis—*Poisson Processes*; H.2.8 [**Database Applications**]: Data mining

Keywords

Twitter; metadata; online social networks; Bayesian inference

1. INTRODUCTION

The culture of Twitter, with its brief tweets, encourages users to express their *current* thoughts. In this paper, we explore whether the *timing* of a user's tweets tells us something about her/his interests, by comparing it against the known timing of external events associated with a particular interest. As an example, consider two groups of users: (i) Fans of the San Francisco Giants baseball team (the SFGiants), with "ground truth" based on analysis of the text of their tweets, and (ii) randomly picked users, presumed to be non-fans. In Figure 1, we plot histograms of the ratio of #tweets/hour during times when the SFGiants played a game to the #tweets/hour at other times for these two sets of users. This data was collected over a one month window. We see that a higher proportion of fans tweet more often during game times (the red curve due to fans is more to the right of the ratio = 1 line). It is clear, therefore, that there is information to be mined from the tweet times of a user. In this paper, we propose a statistical framework for doing so, and report on promising preliminary results on inferring baseball "fandom" for a given team.

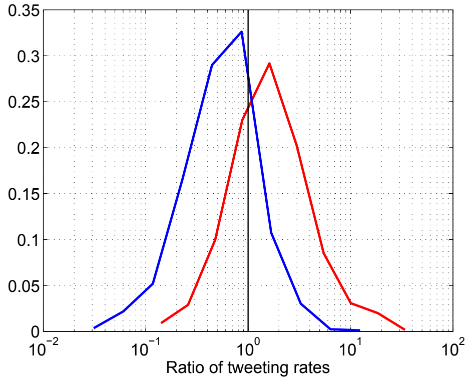

Figure 1: Histograms of the ratio of #tweets/hour during games to #tweets/hour at other times for fans (red) and randomly picked users (blue)

Our model, motivated by the empirical findings such as those in Figure 1, is simple: a fan is likely to tweet at a higher rate in a window around game times than at other times. This leads to a statistical measure for a user's fandom which is the Bayesian posterior probability, based on measured tweet times, of the user's tweet rate during games being higher than at other times. Under our model, this probability only depends only on the *numbers* of tweets (rather than on their exact timing) during games and during other times. This makes the statistic attractive for inference in large-scale systems, both in terms of measurement and computation.

COSN'13, October 7–8, 2013, Boston, Massachusetts, USA.

ACM 978-1-4503-2084-9/13/10 ...$15.00.
http://dx.doi.org/10.1145/2512938.2512960.

The proposed approach extends naturally to incorporate information from a given user's "neighbors," defined on the Twitter graph as follows. Most twitter accounts are public and users often label their tweets using hashtags. These hashtags bring tweets to the attention of other users who are interested in the *content* of the tweet, even when they do not necessarily follow the user who authored the tweet. In this manner, Twitter encourages conversation among individuals who share a common interest. We define the neighbors of a Twitter user as those who are mentioned in his/her tweets (this information is available in the tweet metadata, and does not require parsing of the tweet). We show that pooling measurements from neighbors enhances the reliability of detecting fandom.

Most prior work on mining user interests from Twitter employs text analysis on their tweets (we mention some selected references shortly). This is significantly more expensive than our approach in terms of computation, and hence more difficult to scale to large numbers of users. We view our minimalistic approach as complementary to such text-based approaches; for example, user interests predicted by our approach could be verified by more detailed text-based analysis. It is worth noting, however, text-based analysis is by no means an infallible gold standard. The brief (limited to 140 characters) and ephemeral nature of tweets forces upon them a context-dependent language, making text analysis difficult. For example, people can talk about baseball in their tweets by mentioning the stand, usher, pitcher, bat, ball, etc. All of these words have broad and in some cases multiple meanings. It is therefore difficult to interpret such words without context, and it is difficult to build context from the few words in a tweet. Thus, even if there were no computational bottlenecks, there may be considerable value to hybrid techniques that use dynamics, as we do, along with text-based analysis, to enhance the reliability of mining user interests.

Prior work

Prior work on mining Twitter feeds has mostly been fed by text analysis. TwitterStand [7] maintains a news stand by parsing through different tweet feeds. The timing of tweets has been used here to help in the clustering of tweets into different news groups. The authors in [6] build a system that can locate events such as an earthquake in space and time from tweets (using tweet location and times). However, unlike the solution proposed herein, both[7, 6] rely mainly on text analysis, with tweet times being used only in the later stages. PET[3] tracks the evolution of events, and users' interest in them, as a function of time. Unlike our approach, PET uses text analysis, and does not use the specific tweet time or its relation to external events (PET analyzes tweets collected daily to infer the evolution of topics from day to day). A method of training a classifier to do sentiment analysis of individual tweets is proposed in [5]. Here smileys are used in a bootstrapping mechanism to build a corpus of words along with an associated sentiment (positive or negative) for each word. The preceding references do not explicitly aim to mine for the interests of *a user*, which is the focus of our work. A system that employs Wikipedia as an external corpus to do word associations is proposed in [4] for mining *broad* interests on a *per user* basis. In [1], the authors observe that in identifying political affiliation of a 1000 hand-labeled users, the structure of the re-tweet graph is more useful than the text in the tweets themselves. They arrive at this conclusion by implementing a text based classifier and comparing it with the results obtained merely by identifying the community structure in the re-tweet graph.

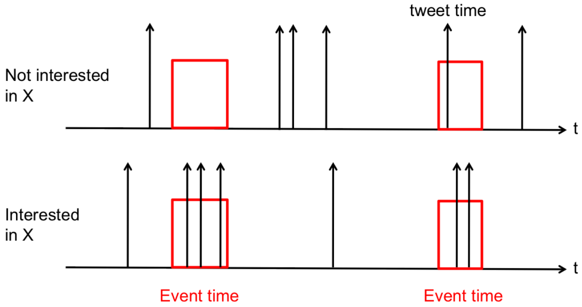

Figure 2: Tweet times of the user marked by arrows. Event times are marked in red. All other times are non-event times. Top: Tweeting behavior of a person not interested in X. Bottom: A person interested in X

2. TWEET TIMES MODEL

In this section we present a probabilistic model for tweet times of a user over an observation time window (this need not consist of contiguous intervals) . Our basic premise is the following: a user who is interested in topic X (say the SFGiants baseball team) tweets more often at times when X is in the "public eye" (SFGiants play a baseball game) than at other times. Thus, we partition the observation window into two complementary sets:

1. **Event times** are times within the observation window when X is in the "public eye" (which, according to our hypothesis, stimulates users interested in X to engage in conversations on Twitter).

2. **Non-event times:** All other times over the observation window.

This partitioning, along with the behaviors we expect for users who are interested (or not) in the topic X, is shown in figure 2.

The tweet times of a user are modeled as a homogeneous Poisson process of rate λ_1 tweets per unit time during event times and an *independent* homogeneous Poisson process of rate λ_0 tweets per unit time during non-event times. As depicted in figure 2, we expect that $\lambda_1 > \lambda_0$ for users interested in topic X.

A homogeneous Poisson process is parameterized by a single parameter, its rate λ. Such a parsimonious model for the tweet times of a user has two advantages: robustness (heterogeneity among twitter users may make more detailed usage profiles, such as allowing for tweet rates dependent on the time of day, counterproductive) and simplicity (e.g., the decision statistics we obtain require aggregate tweet counts rather than individual tweet times). For a Poisson process of constant rate λ tweets/unit time, the number of tweets N made in a time interval of length T (need not be contiguous) is a Poisson random variable with mean $\lambda \times T$. i.e., the probability that the user puts out n tweets in T time units is given by

$$\Pr\left[N = n|\,\lambda\right] = \frac{e^{-\lambda T}\left(\lambda T\right)^n}{n!},\ n = 0, 1, 2, \ldots, \infty.$$

Further, under the Poisson model, the number of tweets put out by the user in non-overlapping time intervals are independent random variables.

3. INFERRING INTEREST LEVELS FROM TWEET TIMES

We propose a statistic that measures our confidence in the assertion that the user tweets more *frequently* during event times than other times. i.e., his/her tweet rate during event times is larger than the rate at other times. This statistic is our metric for the user's interest level in the topic X. We use knowledge of the event and non-event times to estimate the probability distributions of the corresponding tweet rates λ_1 and λ_0 from the tweet times of the user, and then compute the statistic from these posterior distributions.

Under our Poisson model, the posterior distribution of λ_1 given the tweet times depends only on the total number of tweets put out by the user during event times, which we denote by N_1. The tweet times themselves do not matter. Similarly, to make probabilistic inferences on λ_0, all we need is the total number of tweets during non-event times, denoted by N_0. In the language of estimation theory, N_1 and N_0 are *minimal sufficient statistics* for estimation of λ_1 and λ_0, respectively. Let the total time span of the event times and non-event times be T_1 and T_0 respectively.

Continuing with our minimalism in modeling, we assume a *non-informative* prior on the rates λ_1 and λ_0, assuming that the prior density $p(\lambda_1, \lambda_0) \propto 1/\sqrt{\lambda_1\lambda_0}$ for all $\lambda_1 > 0, \lambda_0 > 0$ (the corresponding marginal priors are $p(\lambda_i) \propto 1/\sqrt{\lambda_i}$ for $\lambda_i > 0,\ i = 0, 1$). Of course, this prior cannot exist over an infinite support, since densities must integrate to one, but this is a standard trick in Bayesian estimation when the ground truth on priors is difficult to determine. This joint prior is the Jeffreys non-informative prior on the rate parameters (λ_1, λ_0) [2]. In our case, accurately estimating priors for each topic X would require the ground truth on the interests of a large number of users, which goes counter to our objective of mining for these interests. Furthermore, we would need to constantly revise our ground truth data set for a heterogeneous population of Twitter users with dynamically evolving interests, which is clearly infeasible.

Since we assume that the two Poisson processes corresponding to the event times and non-event times are independent, the corresponding counts N_1 and N_0 are conditionally independent given λ_1, λ_0. Putting this together with our assumption of non-informative prior on λ_1, λ_0, we obtain, using Bayes' rule, that the posterior distributions of λ_1, λ_0 also factor and are given by $p\left(\lambda_i|N_i\right) \propto \Pr\left[N_i|\lambda_i\right] p\left(\lambda_i\right) \propto {\lambda_i}^{N_i - 0.5} e^{-T_i\lambda_i}$, $\lambda_i > 0$. Normalizing the posteriors so they integrate to one (which we can do even though we employed improper priors), we obtain

$$p\left(\lambda_i = x|N_i\right) = \begin{cases} \frac{T_i(T_i x)^{N_i - 0.5} e^{-T_i x}}{\gamma(N_i + 0.5)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}, \qquad (1)$$

where $\gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ is the gamma function. The statistic which we propose to quantify the user's interest level in the topic X is $Z = \Pr\left[\lambda_1 > \lambda_0|N_1, N_0\right]$. We declare a user to be interested in X when Z exceeds a certain threshold. Thus, we conclude that the user is interested in X, when we are "confident enough" that his/her tweet rate during event times is larger than that during non-event times. Given the observations N_1, N_0, T_1, T_0, the statistic Z can be computed using the posteriors (1) as follows:

$$\begin{aligned} Z &= \Pr\left[\lambda_1 > \lambda_0|N_1, N_0\right] \qquad (2) \\ &= \iint_{x>y} p\left(\lambda_1 = x|N_1\right) p\left(\lambda_0 = y|N_0\right)\ dx\ dy. \end{aligned}$$

4. EXPLOITING USER INTERACTIONS

We use social networks to engage in conversation with others who share our interests. If a user has interacts with others who are interested in the topic X, we expect that the probability that he/she is also interested in X is higher than for a randomly picked user. We present a method that relies on this simple intuition to improve our estimates of the interest level of a "tagged" user using interest level estimates of other users mentioned in his/her tweets. During the observation time window, this tagged user may mention other users using their twitter handle (for example, the official SFGiants twitter handle `@SFGiants`, or another individual `@johnadams2001`) in his/her tweets. We call such users the "neighbors" of the tagged user. Since we will be combining the Z statistics of multiple users, we need to pay attention to scaling. In particular, we expect that we would weight the tagged user's Z statistic higher than that of his/her neighbors. We now describe a framework for motivating such scaling.

Notation: Let the index 0 denote the tagged user and the indices $i = 1, \ldots, M$ denote the neighbors. From the number of tweets during event times $N_1(i)$ and non-event times $N_0(i)$ of the i-th user ($\mathbf{N}(i)$ denotes the pair $(N_1(i), N_0(i))$), we arrive the statistic (2) which we denote by Z_i. Let $\lambda_1(i), \lambda_0(i)$ denote the tweet rates of the i-th user in the event and non-event times and Y_i denote the event that the i-th user tweets more frequently during event times than other times. i.e., $Y_i = 1$ if $\lambda_1(i) > \lambda_0(i)$ and $Y_i = 0$ otherwise (note that $Z_i = \Pr[Y_i = 1|\mathbf{N}(i)]$). Let C_i represent the true interest of i in the topic X (C_i takes the value 1 if this user is interested in X and 0 otherwise).

A user who is not interested in X may still happen to tweet more often during event times. Likewise, a user interested in X may happen to tweet less frequently during event times than at other times. Therefore, we first relate Y_i to C_i in a probabilistic manner to derive a function of the Z_i statistic for each user i (i.e., the tagged user and his/her neighbors) such that, when combined across users to make an inference regarding the tagged user, no one user has too big an influence. We then discuss a model for the dependence between the tagged users and his/her neighbors which motivates combining these individual statistics.

For the first step, let $p_t = \Pr[Y_k = 1|C_k = 1]$ denote the probability that a user interested in topic X is *timely* (i.e., tweets more frequently during event times than at other times), and let $p_f = \Pr[Y_k = 1|C_k = 0]$ denote the probability of *false alarm* (i.e., a user not interested in X happens to tweet more frequently during event times). We now compute the likelihood ratio of user k's interest in topic X based on its own measurements, defined as

$$\phi_k = \frac{\Pr[\mathbf{N}(k)|C_k = 1]}{\Pr[\mathbf{N}(k)|C_k = 0]},$$

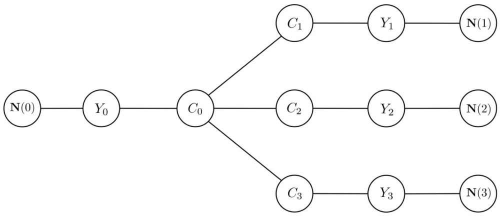

Figure 3: Markov structure of the user interests C_i, the tweet rate differentials $Y_i = \lambda_1(i) > \lambda_0(i)$ and the number of tweets $\mathbf{N}(i) = (N_1(i), N_0(i))$. The index 0 refers to the tagged user while $1, 2, 3$ denote the neighbors of this user

in terms of the statistic $Z_k = P[Y_k = 1|\mathbf{N}(k)]$, which we already know how to compute from Section 3.

Under our uninformative prior, it is easy to show that $\Pr[Y_k = 1] = \Pr[\lambda_1 > \lambda_0] = \frac{1}{2}$, Conditioning on Y_k and using the conditional independence of $\mathbf{N}(k)$ and C_k given Y_k (the Markov structure in figure 3):

$$\begin{aligned}\Pr[\mathbf{N}(k)|C_k] &= \Pr[Y_k = 1|C_k] \Pr[\mathbf{N}(k)|Y_k = 1] \\ &\quad + \Pr[Y_k = 0|C_k] \Pr[\mathbf{N}(k)|Y_k = 0] \\ &= 2\Pr[\mathbf{N}(k)]\Big(\Pr[Y_k = 1|C_k] Z_k \\ &\quad + \Pr[Y_k = 0|C_k](1 - Z_k)\Big),\end{aligned}$$

where we have used $Z_k = \Pr[Y_k = 1|\mathbf{N}(k)]$. Using the above, we obtain that

$$\phi_k = \frac{p_t Z_k + (1-p_t)(1-Z_k)}{p_f Z_k + (1-p_f)(1-Z_k)} = \frac{1 + p_t\left(\frac{Z_k}{1-Z_k} - 1\right)}{1 + p_f\left(\frac{Z_k}{1-Z_k} - 1\right)}.$$

This effectively corresponds to soft thresholding the raw likelihood ratio $\Pr[\lambda_1 > \lambda_0]/\Pr[\lambda_1 \leq \lambda_0] = Z_k/(1-Z_k)$ between an upper limit of $1/p_f$ and a lower limit of $1 - p_t$. Both ϕ_k and the raw likelihood ratio are monotone increasing in Z_k. Thus, for a single user (as considered in the previous section), threshold rules based on any of these statistics are equivalent. However, when combining across multiple users, the soft thresholding in ϕ_k is important for robustness, since it ensures that no one user has too large an influence on the outcome.

Let us now consider the second step: relating the interests of the tagged user and his/her neighbors. We expect that it is more likely that the neighbors are interested in X when the tagged user is interested in X than when the tagged user is not: Denoting $\Pr[C_k = 1|C_0 = 1]$ by α and $\Pr[C_k = 1|C_0 = 0]$ by β, we expect that $\alpha \gg \beta$. It is actually the difference in α and β that affects how we combine these statistics, rather than their raw values. For example, even if α is small (e.g., 0.1, so that there is only a 10% probability of the neighbor of a fan also being a fan), if $\beta = 10^{-4}$, then we still get very useful information from the neighbors' measurements.

We make a simplifying assumption on the structure of interactions among neighbors: The true interests of the neighbors, $\{C_i, i > 0\}$, are independent when conditioned on the interest status of the tagged user $C_0 : \Pr[C_1, \ldots, C_M|C_0] = \prod \Pr[C_i|C_0]$. This is illustrated via the Markov structure depicted in figure 3 (in the figure $M = 3$). This assumption is violated when a neighbor of the tagged user refers to another neighbor of the tagged user in his/her tweets (therefore introducing additional dependencies between the two neighbors). However, as we will see in the results section, this simple structure by itself gives us considerable gains over just using the interest level estimates Z_0 of the tagged user alone.

Our statistic that incorporates information from the neighbors is the following log likelihood ratio:

$$S = \frac{\Pr[\mathbf{N}(0), \mathbf{N}(1), \ldots, \mathbf{N}(M)|C_0 = 1]}{\Pr[\mathbf{N}(0), \mathbf{N}(1), \ldots, \mathbf{N}(M)|C_0 = 0]}.$$

From the Markov structure in figure 3, we observe that the true interests of the neighbors C_i given that of the tagged user C_0 are independent. This observation leads to the following simplification:

$$S = \log\frac{\Pr[\mathbf{N}(0)|C_0 = 1]}{\Pr[\mathbf{N}(0)|C_0 = 0]} + \sum_{k=1}^{M} \log\frac{\Pr[\mathbf{N}(k)|C_0 = 1]}{\Pr[\mathbf{N}(k)|C_0 = 0]}.$$

From Bayes' rule, for the neighbors,

$$\begin{aligned}\Pr[\mathbf{N}(k)|C_0] &= \Pr[\mathbf{N}(k), C_k = 1|C_0] \\ &\quad + \Pr[\mathbf{N}(k), C_k = 0|C_0] \\ &= \Pr[\mathbf{N}(k)|C_k = 1] \Pr[C_k = 1|C_0] \\ &\quad + \Pr[\mathbf{N}(k)|C_k = 0] \Pr[C_k = 0|C_0].\end{aligned}$$

Using the above, we obtain that:

$$\frac{\Pr[\mathbf{N}(k)|C_0 = 1]}{\Pr[\mathbf{N}(k)|C_0 = 0]} = \frac{\alpha\phi_k + (1-\alpha)}{\beta\phi_k + (1-\beta)}.$$

Therefore, the statistic S depends only on the likelihood ratios ϕ_k of the tagged user and his/her neighbors, as follows:

$$S = \log\phi_0 + \sum_{k=1}^{M} \log\frac{1 + \alpha(\phi_k - 1)}{1 + \beta(\phi_k - 1)}.$$

While we can tune the parameters α and β to get good performance with this statistic, in practice, we have found the following modified rule, using a single parameter to scale down the sum of the neighbors' log likelihood ratios, to work well:

$$\tilde{S} = \log\phi_0 + \kappa \textstyle\sum_{k=1}^{M} \log\phi_k. \qquad (3)$$

In our numerical results, therefore, we report on the performance of this modified statistic, with $\kappa = 1/6$ (found to work well empirically).

5. NUMERICAL RESULTS

In this section, we test our statistical framework by trying to identify whether a user is a fan of the San Francisco Giants (SFGiants) baseball team from the user's tweet times (we also briefly report on analogous results for the NY Yankees). The times when SFGiants played Major League Baseball (MLB) games are used as a natural candidate for event times. We also include a 15 minute window on either side of each game in our definition of event times to account for the buzz before and after each game when fans are expected to tweet heavily.

Dataset description: The data set is a 10% random sampling of all *public* tweets over a month (May-June) in the summer of 2011. In this one month window, SFGiants played 29 games. Each tweet, apart from its brief text, is tagged with an user ID, the time when this tweet was made and the user IDs of twitter handles mentioned in the tweet (if any).

Ground truth: In order to characterize the effectiveness of the statistic that we propose, we need to know the fandom of users on whose tweet times we apply the statistic. For this purpose, we searched the text of all tweets (in our dataset) that were made in the first and last 10 minutes of all SFGiants games for keywords associated with this baseball team. The keywords that we used were: `sfgiants`, `#sfgiants`, `rowand`, `#rowand`, `lincecum` and `#lincecum`. We identified 640 users in this manner. We assume that these users who used the keywords associated with the SFGiants baseball team are indeed their fans. We also picked a random set of 1000 users who appear in our dataset (they tweeted at least once in this one month window). None of these randomly picked users used the preceding keywords in their tweets and we assume that they are not fans of SFGiants.

For all of the above users (fans and non-fans) we keep a list of the times at which they put out tweets in this one month window. We use these times to evaluate the statistic (2) for these users. We also keep a list of user IDs for each of these users and this list gives our per user neighbor list. The entries in this list are the users who are mentioned in the tweets of the tagged user over the one month time window (his/her neighbors). In order to compute the statistic (3) which uses estimates of the interest levels of the neighbors, we also compile a list of the tweet times of the neighbors of every user.

Interests from user times: We evaluate the statistic Z in (2) from the tweet times of the 640 fans and 1000 non-fans. When computing Z, we account for an average of ten hours of sleep daily. We do this by scaling the total one month time window $T_1 + T_0$ by 14/24 and computing the total sleep compensated non-event times via $T_0' = (14/24) \times (T_1 + T_0) - T_1$. We assume that the user is awake during event times (thus leaving T_1 as it is). Let $\hat{\lambda}_i = N_i/T_i$ denote the empirical estimate of λ_i. We threshold the Z statistic at different values and plot the number of correctly detected fans versus the false alarms (number of randomly picked users misclassified as fans) in figure 4 (blue curve, top). Contrast this with naive ratio of empirical tweet rate estimates $\hat{\lambda}_1/\hat{\lambda}_0 = (N_1 T_0)/(N_0 T_1)$ that is plotted in black. When we are interested in small false alarm rates, $\hat{\lambda}_1/\hat{\lambda}_0$ metric is not useful: for a false alarm rate of 10/1000 we detect a mere 51/640 fans when we use $\hat{\lambda}_1/\hat{\lambda}_0$, whereas, we are able to detect 137/640 fans using the statistic Z. However, when we are willing to tolerate more false alarms ($> 40/1000$), we see that the performance of Z is comparable to that of empirical tweet rate ratios $\hat{\lambda}_1/\hat{\lambda}_0$.

Incorporating neighbor tweet times: From the tweet times of the neighbors of the 640 fans and the 1000 randomly picked users, we compute their interest level statistic Z (again accounting for a per day average of ten hours of sleep). We then use the interest level estimates of the tagged user and his/her neighbors to compute the statistic $\tilde{S}$ in (3). To compute ϕ_k from the individual interest levels Z_k, we choose $p_t = 0.9$ and $p_f = 10^{-20}$. We threshold the statistic $\tilde{S}$ at different values and plot as before the number of correctly detected fans versus the false alarms in figure 4 (red curve, top). From the figure, we see that for any fixed false alarm rate, we are able to detect more fans via the consolidated statistic $\tilde{S}$ than the interest level Z_0 of the tagged user alone. For example, for a false alarm rate of 10 in a 1000, we are able to improve the detection accuracy for SFGiants from 138/640 using Z_0 alone, to 233/640 using the consolidated statistic $\tilde{S}$ with $\kappa = 1/6$.

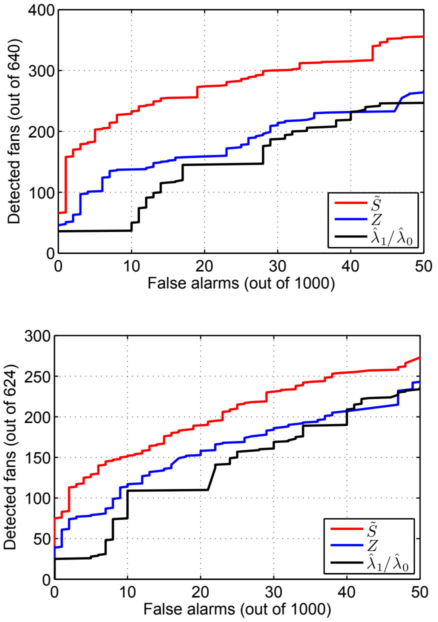

Figure 4: Number of correctly detected fans plotted versus the number of randomly picked users misclassified as fans for the statistics $Z, \tilde{S}$ and $\hat{\lambda}_1/\hat{\lambda}_0$. Top: SFGiants and Bottom: Yankees

We run an identical analysis for 623 fans of the New York Yankees baseball team (identified in a manner similar to the SFGiants fans). These results are plotted in figure 4 (bottom). We see the same trend with the Yankees, with $\tilde{S}$ outperforming Z.

When interpreting the results summarized in figure 4, we must bear in mind the importance of operating at low false alarm rates. The proportion of "fans,", or users interested in any particular topic, is expected to be small. For example, suppose 10% of the overall user population are fans. Then, for a moderately large false alarm rates of 10%, the number of misclassified non-fans is 9% of the user pool. This can overwhelm the pool of correctly classified fans, which is at most 10% for our example. This is the well known *multiple comparisons* problem, for which the natural regime of interest is low false alarm rates. From figure 4, we see that we are able to detect a significant fraction of fans for false alarm rates as small as 1%.

6. ASSUMPTIONS AND LIMITATIONS

While the numerical results on baseball fandom demonstrate the promise of the proposed approach, it is important to clearly outline its assumptions and limitations. Detecting interest in topic X from tweet times alone relies on two key assumptions: (i) Users interested in the topic X are *timely* in their tweeting habits: they respond to either the external stimulus which defines the event times (such as the baseball game played by their favorite team) or the increased chatter about X among their peers during event times, by *tweeting during event times.* (ii) Event times for topic X should not overlap significantly with event times for another interest, say Y, over the observation time interval. Otherwise, we would not be able to attribute the increased tweet rate during event times of X to interest in X *alone* but to interest in *either* X or Y. Such ambiguities get exacerbated if there are many interest groups which share the event windows.

We now give examples for which the preceding assumptions are not easily met:
Movie release times: One possible approach for identifying fans of a particular kind of movie (e.g., Sci-Fi) is to employ event windows around movie releases. However, it may take a few days for even an avid fan to find the time to watch a newly released movie. Thus, interests such as these may not elicit a timely response from fans. This may not necessarily make it difficult to employ the proposed statistic to identify fans of Sci-Fi flicks. It may just mean that we need to use a large window (e.g., a few days) around each movie release when defining event times. However, a large event window may require a large observation window, in order to collect statistics over enough event and non-event windows. This is because many unrelated events may transpire over a window of few days, hence we may require many event windows (movie releases) in order to "average out" the effects due to unmodeled interests which "interfere" with the task of inferring interest in Sci-Fi movies.
Television showtimes: Unlike movie buffs, it is reasonable to expect a timely response from fans of a TV show (say X). However, for TV shows it is possible that the air times for another TV show (say Y) overlap significantly with those of X over the observation time interval. This makes it difficult to know whether a user who tweets more often during X's air-times is indeed interested in X or whether his increased activity is due to interest in the show Y. We may be able to resolve this ambiguity if we identify sufficient additional events in the observation time window when the fans of one of the two TV shows tweet aggressively, but not the other: One example of which could be announcements regarding plans for the next season for the show X. Thus, an important topic for future work is to understand how different the times of events corresponding to two interest groups must be in order to disambiguate them.

On the other hand, for our baseball example in the previous section, both assumptions (i) and (ii) are met: it is reasonable to assume that users who are fans of a baseball team talk about the game and/or engage in conversation with other fans mostly during games (and are therefore timely) as they are expected to watch/follow the games when they are live-on-air. The timing of baseball games does not follow any specific pattern such as the daily/weekly patterns exhibited by TV shows. Therefore, *all* of the times when a particular baseball team plays its games are very unlikely to be the event times for another interest.

7. CONCLUSIONS & FUTURE WORK

We have demonstrated that significant information about a user's interest can be mined from his/her tweet times alone, by correlating these with the timing of appropriately chosen events in the external world. The Bayesian framework that we develop for extracting this information is shown to be effective in detecting baseball fandom from the tweet times of users over a one month period. Measurements from "neighbors" (in the sense of Twitter mentions) provides additional performance gains, with improvements of about 50% in detection accuracy for a false alarm rate of 1%.

We view the results in this paper as a small first step towards a broader investigation of the information that can be gleaned from spatial and temporal dynamics on social networks, and how this information can be best fused with traditional content-based analysis, while accounting for computational and privacy constraints. A Bayesian approach such as the one used here provides a natural framework for such information fusion. For the problem considered here, there are three research directions of immediate interest. As we have discussed, in order to detect interests from tweet times, users have to be timely in their tweets. Thus, an important direction for future research is to identify which interest groups exhibit such timeliness, and how to disambiguate between interest groups that share recurrent event windows. Another important direction is a deeper investigation of the problem of information pooling on the social interaction graph. In our present work, we have assumed that the interests of the neighbors of a tagged user are independent, given the interests of the tagged user. However, in practice, we expect heavy overlap among the friends of every user, so that further performance gains may be available by revisiting the independence assumption. Finally, we have assumed here that event times are known beforehand for the topic of interest. One direction for future work is to mine for these event *times* themselves from a bag of aggregate (network-wide) feeds such as Twitter's trending topics list or Google Trends. Since such an event/non-event times demarcation algorithm is topic-specific and not user-specific, it can employ sophisticated methods including text analysis on these aggregate feeds. Ideas along the lines of those in [3] can potentially be used to identify event times and this needs further study.

8. REFERENCES

[1] M. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the Political Alignment of Twitter Users. In *2011 IEEE third international conference on social computing (SOCIALCOM) and Privacy, Security, Risk and Trust (PASSAT)*, 2011.

[2] H. Jeffreys. *Theory of probability*. Oxford University Press, 1998.

[3] C. X. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, 2010.

[4] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, 2010.

[5] A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, may 2010.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, 2010.

[7] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, 2009.