How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?

Munmun De Choudhury[†] **Yu-Ru Lin**[†] **Hari Sundaram**[†]
K. Selçuk Candan[†] **Lexing Xie**[‡] **Aisling Kelliher**[†]
[†]Arizona State University, Tempe, AZ 85281, USA
[‡]IBM TJ Watson Research Center, Hawthorne, NY 10532, USA
[†]{munmun.dechoudhury, yu-ru.lin, hari.sundaram, candan, aisling.kelliher}@asu.edu
[‡]xlx@us.ibm.com

Abstract

Platforms such as Twitter have provided researchers with ample opportunities to analytically study social phenomena. There are however, significant computational challenges due to the enormous rate of production of new information: researchers are therefore, often forced to analyze a judiciously selected "sample" of the data. Like other social media phenomena, information diffusion is a social process–it is affected by user context, and topic, in addition to the graph topology. This paper studies the impact of different attribute and topology based sampling strategies on the discovery of an important social media phenomena–information diffusion.

We examine several widely-adopted sampling methods that select nodes based on attribute (random, location, and activity) and topology (forest fire) as well as study the impact of attribute based seed selection on topology based sampling. Then we develop a series of metrics for evaluating the quality of the sample, based on user activity (e.g. volume, number of seeds), topological (e.g. reach, spread) and temporal characteristics (e.g. rate). We additionally correlate the diffusion volume metric with two external variables–search and news trends. Our experiments reveal that for small sample sizes (30%), a sample that incorporates both topology and user-context (e.g. location, activity) can improve on naïve methods by a significant margin of ∼15-20%.

Introduction

Over the past forty years, traditional methods of studying social processes such as information diffusion, expert identification or community detection have been focused on longitudinal studies of relatively small groups. However, the widespread proliferation of several social websites such as Facebook, Twitter, Digg, Flickr and YouTube has provided ample avenues to researchers to study such processes at *very large scales*. This is because electronic social data can be acquired and stored over extended time intervals, and for very large populations. The result is that study of social processes on a scale of million nodes, that would have been inconceivable a decade ago, is becoming routine.

Consider the particular social process of information diffusion. The pervasive use of social media has made the cost involved in propagating a piece of information to a large audience extremely negligible, providing extensive evidences of large-scale social contagion. As a result researchers, today, are able to conduct massive empirical studies on diffusion, such as involving blog postings (Gruhl et al. 2004), Internet chain-letter data (Liben-Nowell and Kleiberg 2008), social tagging (Anagnostopoulos, Kumar, and Mahdian 2008), Facebook news feed (Sun et al. 2009), online games (Bakshy, Karrer, and Adamic 2009) and so on.

The attention paid to data volume, however, has overshadowed seemingly less obvious but two equally important challenges: namely, *data acquisition bottleneck* and *information analysis complexity*. For example, the social network Facebook currently features more than 350M users, while the social media Twitter has a rate of approximately 17,000 posts (tweets) per minute. Under such circumstances, firstly, typical data acquisition tools as provided by the publicly available APIs (Application Programming Interfaces) are often not sufficient to track all the data that is being generated–note, the rate limit of API calls for Twitter is only 20,000 per hour[1]–hence creating an acquisition bottleneck. Second, there is extensive resource cost involved in storage of data of this scale, and also, thereby considerably high complexity in analyzing the data itself.

These challenges necessitate the need for collecting a *sample* of the social data that spans over a diverse set of users. Typically, researchers rely on some judiciously chosen sampling practice (e.g. random sampling or snowballing (Frank 1978)) that can recover the topological characteristics of the particular social graph independent of the particular application in question (Leskovec, Kleinberg, and Faloutsos 2005), (Leskovec and Faloutsos 2006). However, in order to study complex dynamic social processes such as diffusion, apart from topology of the social graph, there is a need to consider the nature of the shared information content as well as the rich social context–users (nodes) in a social network are associated with various attributes (e.g. location, age, profession, etc.) and the relationship (edges) may have various properties (e.g. friendship may have duration, or may be asymmetric).

In this paper, we formally study how the choice of different sampling strategies impacts discovery of the partic-

[1] The default rate limit for API calls is 150 requests per hour; a whitelisted account or IP is allowed 20,000 requests per hour.

ular social phenomenon, diffusion. Diffusion has found extensive potential in addressing the propagation of medical and technological innovations (Newman 2002), cultural bias (Zachary 1977), (Bakshy, Karrer, and Adamic 2009) and understanding information roles of users (Kempe, Kleinberg, and Tardos 2003), (Watts and Dodds 2007).

Our approach comprises two steps. First, we utilize several popularly used sampling techniques such as random sampling, degree of user activity based sampling, forest-fire and location-attribute based sampling to extract subgraphs from a social graph of users engaged in a social activity. Second, these subgraphs are used to study diffusion characteristics with respect to the properties of the users (e.g. participation), structural (e.g. reach, spread) and temporal characteristics (e.g. rate) as well as relationship to events in the external world (e.g. search and news trends).

We have conducted extensive experiments on a large-scale dataset collected from Twitter to understand, up to what extent the results of diffusion analysis obtained from different types of samples are affected by the corresponding sampling methods. Our experiments reveal that methods that incorporate both network topology and user-context such as activity, or attributes related to "homophily" (e.g. location) are able to explain diffusion characteristics better compared to naïve methods (e.g. random or activity based sampling) by a large margin of $\sim$ 15-20%. Besides, for moderately small sample sizes (30%), these methods can explain the metrics computed on unsampled graph better than pure attribute or topology based strategies.

The rest of the paper is organized as follows. In the next section we present a discussion on related work. We present our problem definition in the following section, and then discuss different sampling techniques. The following section deals with our evaluation metrics, used to estimate diffusion bias under sampling. Thereafter we present experimental studies over Twitter, present a discussion of our work and finally conclude with our major contributions.

Related Work

We discuss related prior work from two different perspectives: first, sampling of large-scale graphs, and second, information diffusion in social media and networks.

Graph Sampling

Our work deals with extracting information from large-scale social networks, which is closely related to the problem of "subgraph sampling." A subgraph sampling method commonly used in sociology studies is *snowball sampling* (Frank 1978); another well-known method is *random walk sampling* (Klovdahl et al. 1977). Recent work has investigated sampling of large-scale graphs, with a focus on recovering topological characteristics such as degree distribution, path length etc. (Rusmevichientong et al. 2001) as well as analyzing the impact of missing data on social network properties (Kossinets 2006). For example, Leskovec et al. in (Leskovec, Kleinberg, and Faloutsos 2005), (Leskovec and Faloutsos 2006) focused on empirically observed static and dynamic graph properties such as densification and shrinking diameter. They studied different sampling methods, including random node/edge selection, random walk etc for recovering solely these topological properties. They also introduced *forest fire sampling*, which randomly selected a subset of neighbors of current traversed node to form a sample according to a forwarding probability.

Social Diffusion Analysis

The analysis of social information diffusion has been of interest to researchers from various domains ranging from social sciences, epidemiology, disease propagation, physics and economics (Zachary 1977), (Newman 2002), (Watts and Dodds 2007). There has been prior work on modeling and predicting pathways of diffusion of information in social networks useful for several applications, ranging from recommendation systems, online advertising, user behavior prediction and disease containment (Kempe, Kleinberg, and Tardos 2003), (Gruhl et al. 2004), (Song et al. 2006), (Anagnostopoulos, Kumar, and Mahdian 2008), (Kossinets, Kleinberg, and Watts 2008), (Liben-Nowell and Kleiberg 2008).

In an early work (Kempe, Kleinberg, and Tardos 2003), the authors propose solution to the optimization problem of selecting the most influential nodes in a social network which could trigger a large cascade of further adoptions. In a recent work, Bakshy et al (Bakshy, Karrer, and Adamic 2009) study how "gestures" make their way through an online community–Second Life. In another work, Sun et al (Sun et al. 2009) study the diffusion patterns on the Facebook "News Feed" and conclude that in online social media, diffusion dynamics are often triggered by the collision of short chains of information trigger.

Although these prior work provide useful insights into the subgraph sampling problem as well as into characteristics of diffusion in social media separately, they suffer from limitations in the context of this work:

- *Information content:* Today's online social media feature extensive activity that is dependent on the information content being shared (i.e. the topic) as well as have been historically observed to exhibit correlation with external events (Gruhl et al. 2005). Hence, pure topology-based sampling might not be suitable to study social processes that depend on the relationship between the shared content and external user actions and events.
- *Social context:* Most of the prior research does not consider the *contextual information* of the users in the social graph, such as the geographical location, or how quickly the user changes her status (e.g. the rate of social activity). Such contextual information in crucial in studying social phenomena like diffusion, whose impact will be investigated in this paper.

Problem Definition

We now introduce the social graph model, the key concepts and then define our research problem.

Social Graph Model. Our social graph model is based on the social media Twitter. Twitter features a micro-blogging service that allows users to post short content, known as "tweets", often comprising URLs usually encoded via bit.ly, tinyurl, etc. The particular social action of posting a tweet is popularly called "tweeting". Users can also "follow" other users; hence if user u follows v, Twitter allows u to subscribe to the tweets of v via feeds. Two users are denoted as "friends" on Twitter if they "follow" each other.

We now define a social graph $G(V, E)$ that is directed and where V is the set of users and $e_{uv} \in E$ if and only if users u and v are "friends" of each other. Note that, using the bi-directional link is more useful in the context of Twitter compared to the uni-directional "follow" link because the former is more likely to be robust to spam–a normal user is less likely to follow a spam-like account.

Topic Diffusion. We define diffusion with respect to a particular topic as the flow of information from one user to another via the social graph $G(V, E)$, also called "social contagion". Given two users v and u sharing a "friend" link, topic diffusion on Twitter can manifest itself through three types of evidences: (1) users posting tweets using the same URL, (2) users tweeting with the same hashtag (e.g. #Election2008) or a set of common keywords, and (3) users using the re-tweet (RT) tag. We utilize all these three cases of diffusion in this paper.

Diffusion Series. In order to study diffusion characteristics, we now define a topology called a *diffusion series* that summarizes diffusion via social contagion in the social graph G for a given topic θ. Note, a diffusion series is similar to a diffusion tree as in (Liben-Nowell and Kleiberg 2008), (Bakshy, Karrer, and Adamic 2009), however we call it a "series" since it is constructed progressively over a period of time and allows a node to have multiple sources of diffusion from more than one contact and at more than one time interval.

A diffusion series δ_θ on topic θ is a directed graph where the nodes are users tweeting on a topic over a period of time. Specifically, a node represents an occurrence of a user creating at least one tweet about a topic at a particular time, and nodes associated with the same period of time are arranged into the same level[2]. Each level l_m in the diffusion series δ_θ is defined over a day in this paper, i.e. in the context of Twitter, all the users in a particular level tweet about the information on the same day; and two consecutive levels have a time difference of one day. The edges across nodes between two adjacent levels indicate that user u in level l_m tweets about the information on the m-th day, after (via the subscribed feed) her contact (i.e. friend) v has tweeted about the same information on the previous day (at level l_{m-1}). Note there are also some other approximations on the diffusion process captures by this topology. Social diffusion is a continuous-time process, but the diffusion series are constructed over discrete time, at 24 hr increments. That is, there are no edges between nodes at the same level–a diffusion series in this work captures flow of information *across* days, and does not include possible flow occurring at the same day. However, our definition is generic enough to construct diffusion series over any arbitrary discrete temporal granularity. An example of a diffusion series on Twitter over the topic "global warming" has been shown in Figure 1. Significant news events associated with diffusion have also been annotated in the series in the figure.

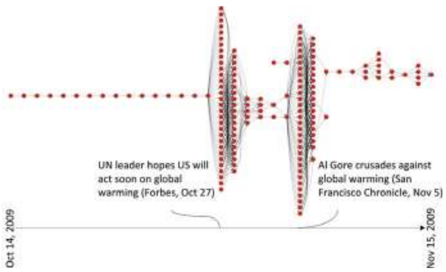
UN leader hopes US will act soon on global warming (Forbes, Oct 27)
Al Gore crusades against global warming (San Francisco Chronicle, Nov 5)
Oct 14, 2009
Nov 15, 2009

Figure 1: Example of a diffusion series from Twitter on the topic "global warming".

Since each topic θ can have multiple disconnected diffusion series δ_θ at any given time, we call the set of all diffusion series a *diffusion collection* $\mathcal{D}_\theta = \{\delta_\theta\}$.

Problem Statement. Given a topic θ and observable social actions of users in a social graph G, our goal is to: (1) utilize a set of sampling techniques $S \in \mathcal{S}$ to extract samples from the original social graph G, as given by $\widehat{G}(S)$ where $\widehat{G} \subset G$; and (2) empirically study the diffusion characteristics in the original social graph G (given by the diffusion collection $\mathcal{D}_\theta$) as well as in the samples $\widehat{G}(S)$ (given by $\widehat{\mathcal{D}}_\theta(S)$); thereby empirically estimating the robustness and effectiveness of various sampling techniques $S \in \mathcal{S}$ useful for selectively retrieving information on the topic θ.

Sampling Diffusion Data

We define a sampling strategy $S \in \mathcal{S}$ as a technique that selectively chooses nodes from the social graph G based on a certain attribute or technique. Typically the sampled graph $\widehat{G}(S)$ is considerably smaller in size in terms of number of nodes, compared to G; the size being determined based on a ratio (called the sampling ratio): $\rho = size(\widehat{G}(S))/size(G)$. In this paper for each strategy S, we have defined samples based on ρ linearly ranging between 0.1 and 1.0.

We now examine six different sampling strategies, that include three attribute-based and three topology-based techniques. For attribute based techniques, we form the sampled graph $\widehat{G}(S)$ as follows. Using the nodes obtained from the sampling method S, we examine G for existence of edges between all pairs of the selected nodes. Thereafter we associate an edge between nodes u and v in $\widehat{G}(S)$ if u and v are connected in G. For the topology-based techniques, we select a seed user set based on an attribute, and then use the topology of G to expand the sampled graph $\widehat{G}(S)$.

Attribute-based Techniques

1. *Random Sampling*: We focus on a random sampling strategy where we select users based on a uniform distribution.

2. *Activity-based Sampling*: This sampling strategy involves choosing a subgraph $\widehat{G}(S)$ that comprises the most active users (in terms of their number of tweets).

[2] Hence, the same user may be present multiple times at different levels in a series if s/he tweets about the same topic at different points in time (different days).

3. *Location-based Sampling*: For this method, we divide the users in G into different categories based on their location attribute[3]. The different locations considered in this paper are the different continents, e.g. 'North America', 'Asia' and 'Europe'. For ℓ different locations and a given ρ, we randomly select ρ/ℓ users corresponding to each location to construct the subgraph sample $\widehat{G}(S)$.

Topology-based Techniques

We adopt the forest-fire sampling method as described in (Leskovec and Faloutsos 2006). We have used three different ways to choose the seed user set in this method. In the first case, the seed user set is selected at random. We expand from this seed set based on their contacts (i.e. friends), where the contacts are chosen based on a forwarding probability p_f. This process is repeated in turn for each of the contacts and so on, until we "burn" sufficient number of users for each sample $\widehat{G}(S)$. In the second case, the seed user set is chosen based on a particular location, e.g. 'Asia', 'Europe' etc; while in the third, it is chosen in terms of measure of user activity.

Evaluation of Diffusion Samples

Given the sampled social graphs $\widehat{G}(S)$ constructed through sampling strategies $S \in \mathcal{S}$, we construct the diffusion collections $\widehat{\mathcal{D}}_\theta(S)$. To evaluate the quality of these diffusion collections, we now propose two types of metrics: (a) *saturation metrics*: characterizing the quality of $\widehat{\mathcal{D}}_\theta(S)$ with respect to the diffusion collections $\mathcal{D}_\theta$ constructed from unsampled graph G; (b) *response metrics*: quantifying how the diffusion results obtained from $\widehat{\mathcal{D}}_\theta(S)$ corresponding to popular external activities e.g. search and news trends.

Diffusion Saturation Metrics

We describe eight different metrics for quantifying diffusion on a certain topic that are discovered via a variety of sampling techniques. The metrics are categorized through various aspects such as: properties of users involved in diffusion (volume, participation and dissemination), diffusion series topology (reach, spread, cascade instances and collection size) and temporal properties (rate):

1. *Volume*: Volume is a notion of the overall degree of contagion in the social graph. For a sampling technique $S \in \mathcal{S}$, we formally define volume $v_\theta(S)$ with respect to θ as the ratio of $n_\theta(S)$ to $N_\theta(S)$, where $n_\theta(S)$ is the total number of users (nodes) in the diffusion collection $\widehat{\mathcal{D}}_\theta(S)$, and $N_\theta(S)$ is the number of users in the sampled social graph $\widehat{G}(S)$ associated with topic θ. Note, $N_\theta(S)$ would include users who are not part of the diffusion collection, but nevertheless have tweeted about θ.
2. *Participation*: Participation $p_\theta(S)$ ((Bakshy, Karrer, and Adamic 2009)) is fraction of users involved in the diffusion of information on a particular topic who further trigger other users in the social graph to get involved in the diffusion. It is the number of non-leaf nodes in the diffusion collection $\widehat{\mathcal{D}}_\theta(S)$, normalized by $N_\theta(S)$.
3. *Dissemination*: Dissemination $d_\theta(S)$ is given by the ratio of the number of users in the diffusion collection $\widehat{\mathcal{D}}_\theta(S)$ who do not have a parent node, normalized by $N_\theta(S)$. In other words, they are the "seed users" or ones who get involved in the diffusion due to some unobservable external influence, e.g. a news event.
4. *Reach*: Reach $r_\theta(S)$ ((Liben-Nowell and Kleiberg 2008)) is conceptually defined as the extent in the social graph, to which information on a particular topic θ reaches to users. We define it formally as the mean of the number of levels in all the diffusion series $\widehat{\delta}_\theta(S) \in \widehat{\mathcal{D}}_\theta(S)$.
5. *Spread*: For the diffusion collection $\widehat{\mathcal{D}}_\theta(S)$, spread $s_\theta(S)$ ((Liben-Nowell and Kleiberg 2008)) is defined as the ratio of the maximum number of nodes at any level in $\widehat{\delta}_\theta(S) \in \widehat{\mathcal{D}}_\theta(S)$ to $n_\theta(S)$.
6. *Cascade Instances*: Cascade instances $c_\theta(S)$ is defined as the ratio of the number of levels in the diffusion series $\widehat{\delta}_\theta(S) \in \widehat{\mathcal{D}}_\theta(S)$ where the number of *new* users at a level l_m (i.e. non-occurring at a previous level) is greater than that at the previous level l_{m-1}, to L_δ, the number of levels in $\widehat{\delta}_\theta(S) \in \widehat{\mathcal{D}}_\theta(S)$.
7. *Collection Size*: Collection size $\alpha_\theta(S)$ is the number of diffusion series $\widehat{\delta}_\theta(S)$ in $\widehat{\mathcal{D}}_\theta(S)$ over a certain topic θ .
8. *Rate*: We define rate $\gamma_\theta(S)$ as the "speed" at which information on θ diffuses in the collection $\widehat{\mathcal{D}}_\theta(S)$. It depends on the difference between the median time of posting of tweets at all consecutive levels l_m and l_{m-1} in the diffusion series $\widehat{\delta}_\theta(S) \in \widehat{\mathcal{D}}_\theta(S)$. Hence it is given as:

$$\gamma_\theta(S) = 1/(1+\frac{1}{L_\delta}\sum_{l_{m-1},l_m \in \widehat{\delta}_\theta(S)} (t_\theta^m(S) - t_\theta^{(m-1)}(S))), \tag{1}$$

where $t_\theta^m(S)$ and $t_\theta^{(m-1)}(S)$ are measured in seconds, $t_\theta^m(S)$ corresponds to the median time of tweet at level l_m in $\widehat{\delta}_\theta(S) \in \widehat{\mathcal{D}}_\theta(S)$.

We now define a distortion metric in order to evaluate quantitatively the performance of each of these diffusion saturation metrics for the different sampling strategies, $S \in \mathcal{S}$. The distortion metric is defined as:

$$F_\theta(m;S) = \frac{|m_\theta(S) - \widehat{m}_\theta(S)|}{m_\theta(S)}, \tag{2}$$

where m is the particular metric under consideration. $\widehat{m}_\theta(S)$ is the measure of metric m under S and computed over the diffusion collection $\widehat{\mathcal{D}}_\theta(S)$, while $m_\theta(S)$ is the metric over $\mathcal{D}_\theta$, corresponding to the unsampled social graph.

Diffusion Response Metrics

We now describe metrics for quantifying the relationship between diffusion characteristics obtained from samples within Twitter, and the trends of the same given topic

[3] We use the timezone attribute of the users on Twitter to extract locations, e.g. Eastern Time would indicate users from New York, Massachusetts, Pennsylvania etc.

obtained from external world. We collect two kinds of external-world trends: (1) *search trends*–the search volume of a particular topical keyword over a period of time[4]; (2) *news trends*–the frequency of archived news articles about a particular topical keyword over a period of time[5]. Based on these trends, we define two diffusion response metrics:

1. *Search response*: We first compute the cumulative distribution function (CDF) of diffusion volume as $E_D(x) = \sum_{i \leq x} |l_i(\widehat{\mathcal{D}}_\theta(S))|/Q_D$, where $|l_i(\widehat{\mathcal{D}}_\theta(S))|$ is the number of nodes at the ith level in the collection $\widehat{\mathcal{D}}_\theta(S)$ obtained via sampling technique S. Q_D is the normalized term and is defined as $\sum_i |l_i(\widehat{\mathcal{D}}_\theta(S))|$. Next, we compute the CDF of search volume as $E_S(x) = \sum_{i \leq x} f_S(i)/Q_S$, where $f_S(i)$ is the search volume returned by the Google Trends API for the given time i, and Q_S is the normalization term. The search response is defined as $1 - D(E_D, E_S)$, where $D(A, B)$ is the Kolmogorov-Smirnov (KS) statistic and is defined as $max(|A(x) - B(x)|)$.
2. *News response*: Similarly, we compute the CDF of news volume as $E_N(x) = \sum_{i \leq x} f_N(i)/Q_N$, where $f_N(i)$ is the number of archived news articles available from Google News for the given time i, and Q_N is the normalization term. The news response is similarly defined as $1 - D(E_D, E_N)$, where $D(A, B)$ is the KS statistic.

Experimental Study

Twitter Dataset

We have focused on a large dataset crawled from Twitter. We have undertaken a focused crawl[6] based on a snowballing technique, over a set of quality users (∼465K), who mutually form a reasonably large connected component. First, we seeded the crawl from a set of genuine (or authoritative) users, who post about a diverse range of topics and reasonably frequently. Our seed set size is 500; and comprises politicians, musicians, environmentalists, techies and so on. These lists were collected from the popular social media blog, Mashable (http://mashable.com/2008/10/20/25-celebrity-twitter-users/). Next we expand the social graph from the seed set based on their "friend" links[7]. We finally executed a dedicated cron job that collected the tweets (and their associated timestamps) for users in the entire social graph every 24 hours. Table 1 gives some basic statistics of the crawled data that were used for studying diffusion.

Table 1: Summary of statistics of the data used for studying diffusion on Twitter.

#nodes	465,107
#edges	836,541
#nodes with time-zone attribute	385,547
#tweets	25,378,846
Time span of tweets posting times	Oct 2006–Nov 2009

Experimental Procedure

The crawled social graph, comprising the users and their tweets are now deployed in the study of diffusion. Since we are interested in studying diffusion at the granularity of a topic, we first define how we conceive of the topics. For our experiments, we focus on the "trending topics"[8] that are featured on Twitter over a two month period between Oct and Nov 2009. From the ensemble of these trending topics, a set of ~ 125 topics are selected at random; of which there are 25 hashtags and the rest, phrases or groups of words.

For the ease of analysis, we organize the different trending topics into generalized themes. For automatically assigning theme to trending topic associations, we use the popular open source natural language processing toolkit called "OpenCalais"[9]. In the context of Twitter, we filter tweets give a trending topic, and then use OpenCalais to return theme labels over those tweets. Based on this process, we associated the 125 trending topics with a total of nine themes, such as 'Business Finance', 'Sports' etc'.

Now our experimental goal is to utilize the crawled social graph to construct diffusion samples *per topic* and thereby study the impact of sampling on diffusion.

Results

We present our results from two different perspectives:

1. *What is a good sampling strategy?* i.e. how the choice of a particular sampling technique affects diffusion characteristics and their relationship to external trends.
2. *What is a good sampling ratio?* i.e. how to choose the sample size for different sampling strategies, such that it can explain well the distributions over diffusion metrics on the unsampled graph.

Analysis of Sampling Strategies

Saturation Metrics. Figure 2(a-h) gives the results of how the six different sampling methods impact the saturation metrics over the sampling ratio ρ. The results are averaged over all trending topics. Note, for the forest-fire based methods, the results are averaged over different values of the forwarding probability p_f and for all methods, 50 iterations were undertaken to ensure statistical significance.

We observe that there is significant variation in the distortion measures across the six different strategies. Primarily, for the user-based metrics–volume, participation and dissemination, we observe that the values of distortion lie in a range of 20-25% across the different methods. Also the relatively narrower range of distortion across strategies shows that these three metrics seem to be less sensitive to different sampling methods. Nevertheless, note that the forest-fire with activity based seeds sampling outperforms others.

[4] Google Trends: (http://www.google.com/intl/en/trends/about.html).

[5] Google News: (http://news.google.com/).

[6] http://apiwiki.twitter.com/

[7] Note that the social graph crawled in this work is a static snapshot made only once at the time of the crawl.

[8] Trending topics are Twitter-generated list of popular topics. Note they can either be hashtags (i.e. words or phrases preceded by the # symbol), or could be groups of words.

[9] http://www.opencalais.com/

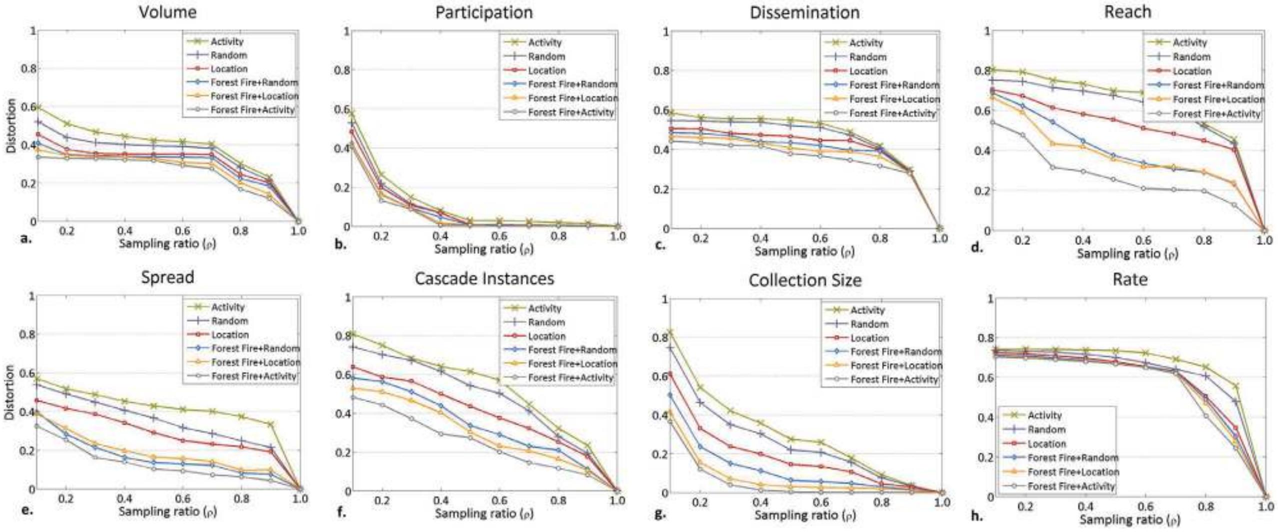

Figure 2: Distortion of different diffusion metrics for different choices of sampling ratio ρ, averaged over all the nine themes (i.e. $\sim$ 125 trending topics from Twitter).

In the case of the diffusion series based topology metrics–reach, spread, cascade instances and collection size, the distortion measures across the different strategies seem to be comparatively more widely-ranged, i.e. the variations range between 30-35%. For example, the three forest-fire based sampling techniques perform significantly better compared to just the activity based, random or location based methods–reinforcing the fact that incorporating topology in the sampling process gives less distortion in terms of the diffusion series topology.

The last metric, rate (that is time-based) also exhibits monotonic increase in distortion for decreasing ρ. Interestingly, the range of variation for different methods is narrowly around 10% in this case; indicating that the choice of the particular sampling strategy does not seem to impact much the measurement of the rate.

Now we present some results that illustrate how diffusion on each topic is affected by the choice of the sampling strategies, averaged across all values of the sampling ratio ρ. Figure 3(a-c) gives the mean measures of distortion for the nine themes and for different sampling techniques, shown for the three categories of diffusion metrics. Across these three metric categories, we observe the following variations over different themes:

- Context such as demographics (location) seem to perform well in yielding quality samples for themes that are 'local' in nature, e.g. 'Sports' comprising topics such as 'NBA', 'New York Yankees', 'Chargers', 'Sehwag' and so on–each of them being of interest to users respectively from the US, NYC, San Diego and India.
- Pure topology based sampling (i.e. forest-fire with random seeds) seem to perform well for themes that are of global importance, such as 'Social Issues' that subsumes topics like '#BeatCancer', 'Swine Flu', '#Stoptheviolence' and 'Unemployment'.
- Incorporating both context (i.e. activity) and topology (i.e. forest-fire) in sampling seems to perform well for themes that relate to external events or issues, e.g. 'Technology-Internet' comprising topics like 'Android 2', 'Google Wave' and 'Windows 7'. Similar observations can be made for 'Politics' that subsumes topics like 'Tiger Woods', 'Healthcare' and 'Afghanistan'–all of which were associated with important external happenings during the period of our analysis.

Response Metrics. Now we present analysis of diffusion in terms of its response to external variables: search and news trends. Figure 4 shows the impact of sampling ratio ρ on response. We observe that for search response, the range of variation of response (based on the KS statistic) lying between $\rho = 0.1$ and 1.0 is larger compared to that of news. This indicates that for smaller values of ρ, the news trends are more responsive to diffusion characteristics compared to search trends. We conjecture that it is because diffusion processes on Twitter are heavily related to external news-related events.

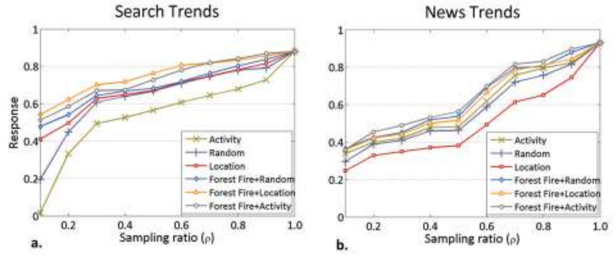

Figure 4: Response behavior (using the KS statistic) with respect to search and news trend for different choices of sampling ratio ρ, averaged over all themes.

Additionally, we also observe that the sampling technique that yields maximum response in case of search trends is the forest-fire with location based seeding technique. Moreover, pure location based sampling appears to perform better compared to attributes such as activity based (that performs quite poorly). This implies that since search behavior of-

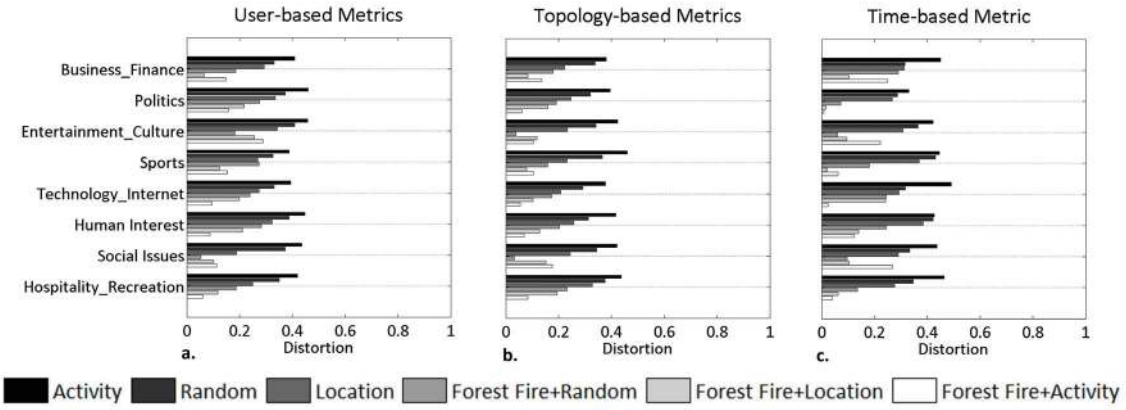

Figure 3: Distortion of diffusion metrics across themes, averaged over sampling ratio ρ. Metrics – user-based: volume, participation and dissemination; topology-based: reach, spread, cascade instances and collection size; time-based: rate.

ten heavily relies on user demographics, diffusion samples drawn based on attributes like location yield good response measures. While for news trends, best performance is given by the forest-fire technique seeded based on activity.

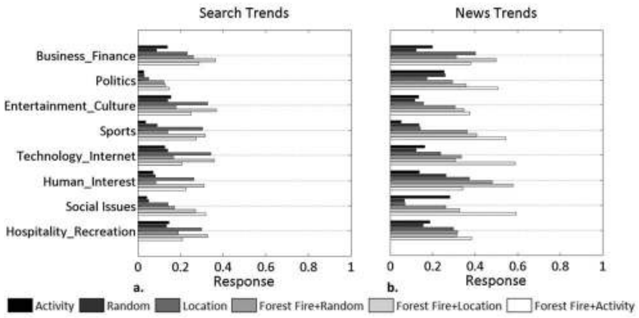

Figure 5: Response behavior (using the KS statistic) with respect to search and news trend across different themes, averaged over sampling ratio ρ.

Finally, in Figure 5, we present the results of response for search and news over the nine different themes averaged across ρ. In the case of search trends, the results indicate that location based sampling and forest-fire seeded based on location perform considerably well in comparison to other techniques; especially for themes that heavily reflect user interest and are aligned along certain demographic attributes, e.g. 'Entertainment-Culture' (example topics are 'Chris Brown', 'Eagles'), 'Sports' and 'Technology-Internet'. While for news trend, we again observe that for several themes, activity based sampling yields good performance. For example, themes such as 'Politics' (subsuming topics like 'Healthcare' and 'Afghanistan'), 'Sports' and 'Technology-Internet' being associated with external events, diffusion samples for these themes drawn using user activity seem to be highly responsive to news trends.

We summarize performance of the sampling techniques over the saturation and the response metrics in Table 2. Note, the results have been found to be statistically significant based on the student t-test statistic and using 50 independent runs of each sampling method.

Table 2: Summary of performance of different sampling techniques over the saturation and response metrics.

Method	Saturation Distortion	Response
Random	0.41	0.68
Activity	0.44	0.64
Location	0.35	0.66
Forest-fire+Random	0.28	0.73
Forest-fire+Location	0.27	0.76
Forest-fire+Activity	0.22	0.78

Analysis of Sampling Ratio

In Figure 2 we had observed that a moderate sampling ratio of ρ=0.3 yields low distortion for user-based and topology-based metrics. To account for choice of ρ more concretely, we present the performance of different strategies for ρ=0.3. Figure 6 shows the topic distribution for diffusion volume and spread–X-axis representing values of the metric while Y-axis, the number of topics having a particular value of the metric. We compare three sampling methods against the distributions on unsampled graph. It appears that, for a moderately small ρ=0.3, forest-fire method seeded on location outperforms pure attribute or topology-based techniques in explaining these distributions, with mean error of ~8.5%.

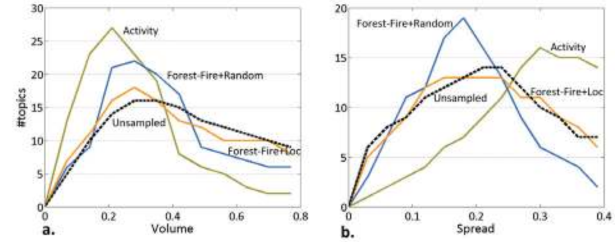

Figure 6: Comparative distribution of (a) volume and (b) spread over different strategies for sampling ratio ρ=0.3.

Discussion

Our primary observation from the results is that sampling impacts the discovery of dynamic social processes, such as diffusion, in a non-trivial manner. Contrary to prior empirical observations that topology alone can yield good subgraph samples (Leskovec and Faloutsos 2006), we have found evidence that sampling techniques that incorporate user context, e.g. activity or location along with the graph topology seem to perform better in discovery of diffusion. Interestingly also note, pure context based techniques, such as location appear to perform reasonably well–better than activity based or random sampling. We conjecture that it is related to the concept of user "homophily" (Mcpherson, Lovin, and Cook 2001); that explains that users engaged in a social activity seem to be associated more closely with ones who are "similar" to them along a certain (contextual) dimension, such as location, age, political view or organizational affiliation, compared to ones who are "dissimilar".

We also observe that diffusion characteristics are widely varied across the different themes; hence content has great impact on the quality of the sample. For example, studies of diffusion related to a political event of importance in the US would benefit more from samples chosen based on location than on pure graph topology. Or if the interest is related to a recent technological event, such as release of an electronic gadget, one can benefit more from sampling techniques based on both topology and activity.

Our results are promising, however are limited by the scope of our dataset, which itself is based on a crawl. Hence the observations on diffusion are likely to be only approximate, because we do not quantify the inherent bias in our initial snowball sample of the Twitter population. Moreover, the non-uniformities within each sampling process also have not been considered, nor have we evaluated the bias in each strategy using any form of sampling bias estimators (Kolaczyk 2009). Additionally note that in this paper we have focused on only one social process: diffusion. Nevertheless, we believe that our empirical observations are extensible to other phenomena as well, e.g. community discovery; because most social processes are affected by both topology and context. Finally, we acknowledge that alternative sampling techniques are also possible. For example, a viral marketeer intending to maximize the flow of information in a very short span of time might be interested in sampling that chooses nodes based on time-varying properties of the edges (Kossinets, Kleinberg, and Watts 2008).

Conclusions and Future Work

We have empirically studied the impact of attribute and topology based sampling methods on discovery of information diffusion in data from Twitter. Our main conclusion is that methods that incorporate both network topology and user-contextual attributes such as activity estimate information diffusion with lower error, for the same sample size, when compared to naïve methods (e.g. random or activity based sampling). The improvements are significant: $\sim$15-20%. Our results also show that for a reasonably small sample size ($\sim$30%) these methods can explain well the topic distributions for diffusion metrics on the unsampled graph.

There are several promising future research directions. We plan to extend this study to different social media datasets (e.g. Digg, Flickr), and social phenomena (e.g. community discovery) to see the effects of sampling.

References

Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *KDD '08*, 7–15.

Bakshy, E.; Karrer, B.; and Adamic, L. A. 2009. Social influence and the diffusion of user-created content. In *EC '09*, 325–334. New York, NY, USA: ACM.

Frank, O. 1978. Sampling and estimation in large social networks. *Social Networks* 1(91):101.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW '04*, 491–501. New York, NY, USA: ACM.

Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; and Tomkins, A. 2005. The predictive power of online chatter. In *KDD '05*, 78–87. New York, NY, USA: ACM.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *KDD '03*, 137–146. New York, NY, USA: ACM.

Klovdahl, A.; Dhofier, Z.; Oddy, G.; O'Hara, J.; Stoutjesdijk, S.; and Whish, A. 1977. Social networks in an urban area: First canberra study. *Journal of Soc.* 13(2):169–172.

Kolaczyk, E. D. 2009. *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated.

Kossinets, G.; Kleinberg, J.; and Watts, D. J. 2008. The structure of information pathways in a social communication network. In *KDD '08*, 435–443.

Kossinets, G. 2006. Effects of missing data in social networks. *Social Networks* 28(3):247–268.

Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *KDD '06*, 636. ACM.

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05*, 187. ACM.

Liben-Nowell, D., and Kleiberg, J. 2008. Tracing information flow on a global scale using internet chain-letter data. *Proc. National Academy of Sciences* 105(12):4633–4638.

Mcpherson, M.; Lovin, L. S.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.

Newman, M. E. J. 2002. Spread of epidemic disease on networks. *Physical Review E* 66(1):016128+.

Rusmevichientong, P.; Pennock, D.; Lawrence, S.; and Giles, C. 2001. Methods for sampling pages uniformly from the world wide web. In *AAAI Fall Symposium on Using Uncertainty Within Computation*, 121–128.

Song, X.; Tseng, B. L.; Lin, C.-Y.; and Sun, M.-T. 2006. Personalized recommendation driven by information flow. In *SIGIR '06*, 509–516.

Sun, E.; Rosenn, I.; Marlow, C.; and Lento, T. 2009. Gesundheit! modeling contagion through facebook news feed. In *ICWSM '09*. San Jose, CA: AAAI Press.

Watts, D. J., and Dodds, P. S. 2007. Influentials, networks, and public opinion formation. *Journal of Consumer Research* 34(4):441–458.

Zachary, W. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33:452–473.