Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose

Fred Morstatter
Arizona State University
699 S. Mill Ave.
Tempe, AZ, 85281

Jürgen Pfeffer
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA, 15213

Huan Liu
Arizona State University
699 S. Mill Ave.
Tempe, AZ, 85281

Kathleen M. Carley
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA, 15213

Abstract

Twitter is a social media giant famous for the exchange of short, 140-character messages called "tweets". In the scientific community, the microblogging site is known for openness in sharing its data. It provides a glance into its millions of users and billions of tweets through a "Streaming API" which provides a sample of all tweets matching some parameters preset by the API user. The API service has been used by many researchers, companies, and governmental institutions that want to extract knowledge in accordance with a diverse array of questions pertaining to social media. The essential drawback of the Twitter API is the lack of documentation concerning what and how much data users get. This leads researchers to question whether the sampled data is a valid representation of the overall activity on Twitter. In this work we embark on answering this question by comparing data collected using Twitter's sampled API service with data collected using the full, albeit costly, Firehose stream that includes every single published tweet. We compare both datasets using common statistical metrics as well as metrics that allow us to compare topics, networks, and locations of tweets. The results of our work will help researchers and practitioners understand the implications of using the Streaming API.

Introduction

Twitter is a microblogging site where users exchange short, 140-character messages called "tweets". Ranking as the 10th most popular site in the world by the Alexa rank in January of 2013[1], the site boasts 500 million registered users publishing 400 million tweets per day. Twitter's platform for rapid communication is said to be a vital communication platform in recent events including Hurricane Sandy[2], the Arab Spring of 2011 (Campbell 2011), and several political campaigns (Tumasjan et al. 2010; Gayo-Avello, Metaxas, and Mustafaraj 2011). As a result, Twitter's data has been coveted by both computer and social scientists to better understand human behavior and dynamics.

Social media data is often difficult to obtain, with most social media sites restricting access to their data. Twitter's policies lie opposite to this. The "Twitter Streaming API"[3] is a capability provided by Twitter that allows anyone to retrieve at most a 1% sample of all the data by providing some parameters. According to the documentation, the sample will return at most 1% of all the tweets produced on Twitter at a given time. Once the number of tweets matching the given parameters eclipses 1% of all the tweets on Twitter, Twitter will begin to sample the data returned to the user. The methods that Twitter employs to sample this data is currently unknown. The Streaming API takes three parameters: keywords (words, phrases, or hashtags), geographical boundary boxes, and user ID.

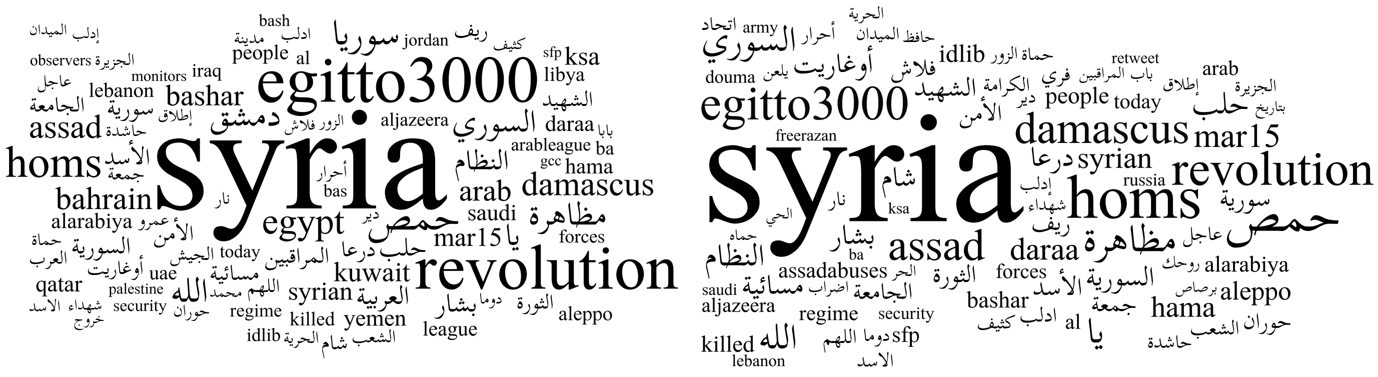
One way to overcome the 1% limitation is to use the Twitter Firehose—a feed provided by Twitter that allows access to 100% of all public tweets. A very substantial drawback of the Firehose data is the restrictive cost. Another drawback is the sheer amount of resources required to retain the Firehose data (servers, network availability, and disk space). Consequently, researchers as well as decision makers in companies and government institutions are forced to decide between two versions of the API: the freely-available but limited Streaming, and the very expensive but comprehensive Firehose version. To the best of our knowledge, no research has been done to assist those researchers and decision makers by answering the following: How does the use of the Streaming API affect common measures and metrics performed on the data? In this article we answer this question from different perspectives.

We begin the analysis by employing classic statistical measures commonly used to compare two sets of data. Based on unique characteristics of tweets, we design and conduct additional comparative analysis. By extracting topics using a frequently used algorithm, we compare how topics differ between the two datasets. As tweets are linked data, we perform network measures of the two datasets. Because tweets can be geo-tagged, we compare the geographical distribution of geolocated tweets to better understand how sampling affects aggregated geographic information.

[1] http://www.alexa.com/topsites

[2] http://www.nytimes.com/interactive/2012/10/28/nyregion/hurricane-sandy.html

[3] https://dev.twitter.com/docs/streaming-apis



(a) Firehose

(b) Streaming API

Figure 1: Tag cloud of top terms from each dataset.

Related Work

Twitter's Streaming API has been used throughout the domain of social media and network analysis to generate understanding of how users behave on these platforms. It has been used to collect data for topic modeling (Hong and Davison 2010; Pozdnoukhov and Kaiser 2011), network analysis (Sofean and Smith 2012), and statistical analysis of content (Mathioudakis and Koudas 2010), among others. Researchers' reliance upon this data source is significant, and these examples only provide a cursory glance at the tip of the iceberg. Due to the widespread use of Twitter's Streaming API in various scientific fields, it is important that we understand how using a sub-sample of the data generated affects these results.

From a statistical point of view, the "law of large numbers" (mean of a sample converges to the mean of the entire *population*) and the Glivenko-Cantelli theorem (the unknown distribution X of an attribute in a population can be approximated with the observed distribution x) guarantee satisfactory results from sampled data when the randomly selected sub-sample is big enough. From network algorithmic (Wasserman and Faust 1994) perspective the question is more complicated. Previous efforts have delved into the topic of network sampling and how working with a restricted set of data can affect common network measures. The problem was studied earlier in (Granovetter 1976), where the author proposes an algorithm to sample networks in a way that allows one to estimate basic network properties. More recently, (Costenbader and Valente 2003) and (Borgatti, Carley, and Krackhardt 2006) have studied the affect of data error on common network centrality measures by randomly deleting and adding nodes and edges. The authors discover that centrality measures are usually most resilient on dense networks. In (Kossinets 2006), the authors study global properties of simulated random graphs to better understand data error in social networks. (Leskovec and Faloutsos 2006) proposes a strategy for sampling large graphs to preserve network measures.

In this work we compare the datasets by analyzing facets commonly used in the literature. We start by comparing the top hashtags found in the tweets, a feature of the text commonly used for analysis. In (Tsur and Rappoport 2012), the authors try to predict the magnitude of the number of tweets mentioning a particular hashtag. Using a regression model trained with features extracted from the text, the authors find that the content of the idea behind the tag is vital to the count of the tweets employing it. Tweeting a hashtag automatically adds a tweet to a page showing tweets published by other tweeters containing that hashtag. In (Yang et al. 2012), the authors find that this communal property of hashtags along with the meaning of the tag itself drive the adoption of hashtags on Twitter. (De Choudhury et al. 2010) studies the propagation patterns of URLs on sampled Twitter data.

Topic analysis can also be used to better understand the content of tweets. (Kireyev, Palen, and Anderson 2009) drills the problem down to disaster-related tweets, discovering two main types of topics: informational and emotional. Finally, (Yin et al. 2011; Hong et al. 2012; Pozdnoukhov and Kaiser 2011) all study the problem of identifying topics in geographical Twitter datasets, proposing models to extract topics relevant to different geographical areas in the data. (Joseph, Tan, and Carley 2012) studies how the topics users discuss drive their geolocation.

Geolocation has become a prominent area in the study of social media data. In (Wakamiya, Lee, and Sumiya 2011) the authors try to classify towns based upon the content of the geotagged tweets that originate from within the town. (De Longueville, Smith, and Luraschi 2009) studies Twitter's use as a sensor for disaster information by studying the geographical properties of users tweets. The authors discover that Twitter's information is accurate in the later stages of a crisis for information dissemination and retrieval.

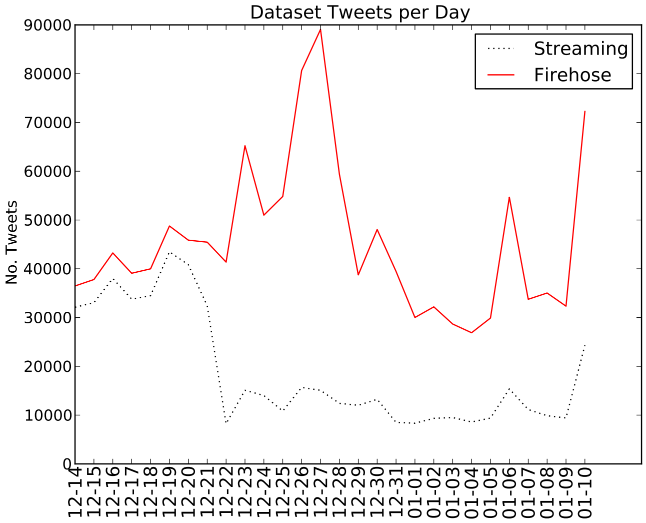


Figure 2: Raw tweet counts for each day from both the Streaming API and the Firehose.

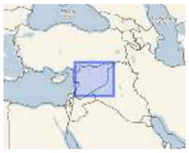
The Data

From December 14th, 2011 - January 10th, 2012 we collected tweets from the Twitter Firehose matching any of the keywords, geographical bounding boxes, and users in Table 1. During the same time period, we collected tweets from the Streaming API using TweetTracker (Kumar et al. 2011) with exactly the same parameters. During the time we collected 528,592 tweets from the Streaming API and 1,280,344 tweets from the Firehose. The raw counts of tweets we received each day from both sources are shown in Figure 2. One of the more interesting results in this dataset is that as the data in the Firehose spikes, the Streaming API coverage is reduced. One possible explanation for this phenomenon could be that due to the Western holidays observed at this time, activity on Twitter may have reduced causing the 1% threshold to go down.

One of the key questions we ask in this work is how the amount of coverage affects measures commonly performed on Twitter data. Here we define coverage as the ratio of data from the Streaming API to data from the Firehose. To better understand the coverage of the Streaming API for each day, we construct a box-and-whisker plot to visualize the distribution of daily coverage, shown in Figure 3. In this period of time the Streaming API receives, on average, 43.5% of the data available on the Firehose on any given day. While this is much better than just 1% of the tweets promised by the Streaming API, we have no reference point for the data in the tweets we received.

The most striking observation is the range of coverage rates (see Figure 3). Increase of *absolute* importance (more global awareness) or *relative* importance (the overall number of tweets decreases) result in lower coverage as well as fewer tweets. To give the reader a sense for the top words in both datasets, we include tag clouds for the top words in the Streaming API and the Firehose, shown in Figure 1.

Table 1: Parameters used to collect data from Syria. Coordinates below the boundary box indicate the Southwest and Northeast corner, respectively.

Keywords	Geoboxes	Users
#syria, #assad, #aleppovolcano, #alawite, #homs, #hama, #tartous, #idlib, #damascus, #daraa, #aleppo, #سوريا[*], #houla	 (32.8, 35.9), (37.3, 42.3)	@SyrianRevo

[*] Arabic word for “Syria”

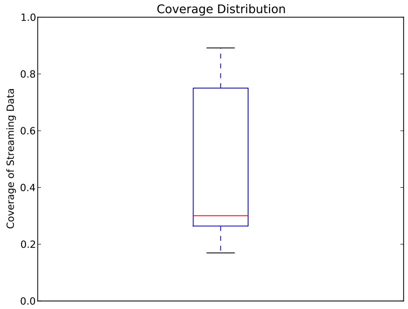


Figure 3: Distribution of coverage for the Streaming data by day. Whiskers indicate extreme values.

Statistical Measures

We investigate the statistical properties of the two datasets with the intent of understanding how well the characteristics of the sampled data match those of the Firehose. We begin first by comparing the top hashtags in the tweets for different levels of coverage using a rank correlation statistic. We continue to extract topics from the text, matching topical content and comparing topical distribution to better understand how sampling affects the results of this common process performed on Twitter data. In both cases we compare our streaming data to random datasets obtained by sampling the data obtained through the Firehose.

Top Hashtag Analysis

Hashtags are an important communication device on Twitter. Users employ them to annotate the content they produce, allowing for other users to find their tweets and to facilitate interaction on the platform. Also, adding a hashtag to a tweet is equivalent to joining a community of users discussing the same topic (Yang et al. 2012). In addition, hashtags are also used by Twitter to calculate the trending topics of the day, which encourages the user to post in these communities.

Recently, hashtags have become an important part of Twitter analysis (Efron 2010; Tsur and Rappoport 2012; Recuero and Araujo 2012). For both the purpose of community formation and trend analysis it is important that our Streaming dataset convey the same importance for hashtags as the Firehose data. Here we compare the top hashtags in

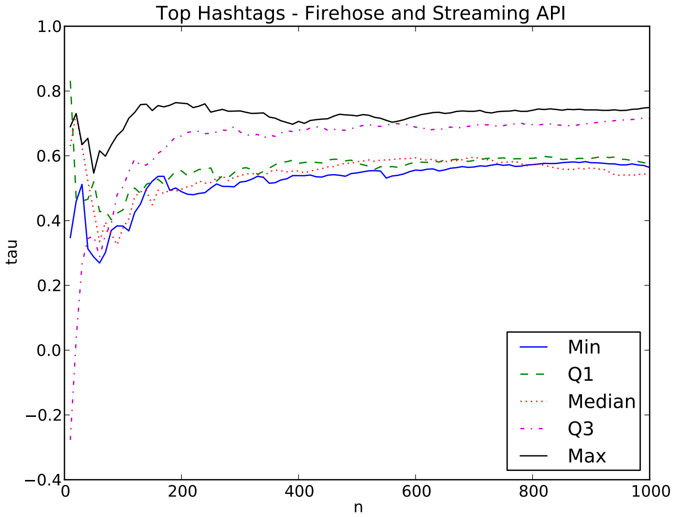


Figure 4: Relationship between n - number of top hashtags, and the correlation coefficient, τ_β.

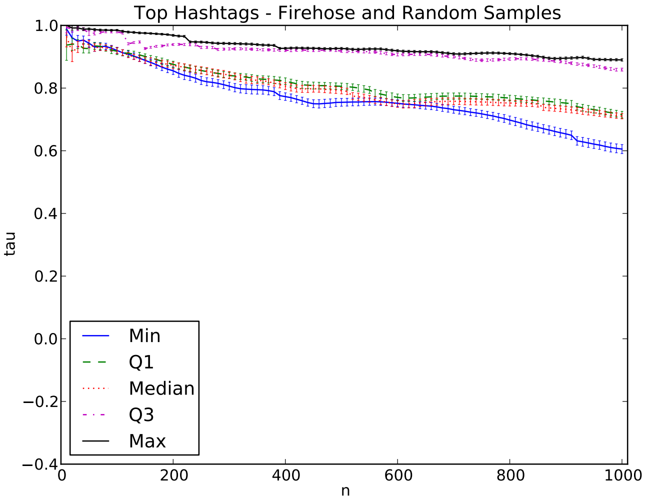


Figure 5: Random sampling of Firehose data. Relationship between n - number of top hashtags, and τ_β - the correlation coefficient for different levels of coverage.

the two datasets using Kendall's τ rank correlation coefficient (Agresti 2010).

Kendall's τ of Top Hashtags Kendall's τ is a statistic which measures the correlation of two ordered lists by analyzing the number of concordant pairs between them. Consider two hashtags, #A and #B. If both lists rank #A higher than #B, then this is considered a concordant pair, otherwise it is counted as a discordant pair. Ties are handled using the τ_β statistic as follows:

$$\tau_\beta = \frac{|P_C| - |P_D|}{\sqrt{(|P_C| + |P_D| + |T_F|)(|P_C| + |P_D| + |T_S|)}} \quad (1)$$

where P_C is the set of concordant pairs, P_D is the set of discordant pairs, T_F is the set of ties in the Firehose data, but not in the Streaming data, T_S is the number of ties found in the Streaming data, but not in the Firehose, and n is the number of pairs in total. The τ_β value ranges from -1, perfect negative correlation, to 1, perfect positive correlation.

To understand the relationship between n and the resulting correlation, τ_β, we construct a chart showing the value of τ_β for n between 10 and 1000 in steps of 10. To get an accurate representation of the differences in correlation at each level of Streaming coverage, we select five days with different levels of coverage as motivated by Figure 3: The minimum (December 27th), lower quartile (December 24th), median (December 29th), upper quartile (December 18th), and the maximum (December 19th). The results of this experiment are shown in Figure 4. Here we see mixed results at small values of n, indicating that the Streaming data may not be good for finding the top hashtags. At larger values of n, we see that the Streaming API does a better job of estimating the top hashtags in the Firehose data.

Comparison with Random Samples After seeing the results from the previous section, we are left to wonder if the results are an artifact of using the Streaming API or if we could have obtained the same results by any random sampling. Would we obtain the same results with a random sample of equal size from the Firehose data, or does the Streaming API's filtering mechanism give us an advantage? To answer this question we repeat the experiments for each day in the previous section. This time, instead of using Streaming API data, we select tweets uniformly at random (without replacement) until we have amassed the same number of tweets as we collected from the Streaming API for that day. We repeat this process 100 times and obtain results as shown in Figure 5. Here we see that the levels of coverage in the random and Streaming data have comparable τ_β values for large n, however at smaller n we see a much different picture. The random data gets very high τ_β scores for $n = 10$, showing a good capacity for finding the top hashtags in the dataset. The Streaming API data does not consistently find the top hashtags, in some cases revealing reverse correlation with the Firehose data at smaller n. This could be indicative of a filtering process in Twitter's Streaming API which causes a misrepresentation of top hashtags in the data.

Topic Analysis

Topic models are statistical models which discover topics in a corpus. Topic modeling is especially useful in large data, where it is too cumbersome to extract the topics manually. Due to the large volume of tweets published on Twitter, topic modeling has become central to many content-based studies using Twitter data (Kireyev, Palen, and Anderson 2009; Pozdnoukhov and Kaiser 2011; Hong et al. 2012; Yin et al. 2011; Chae et al. 2012). We compare the topics drawn from the Streaming data with those drawn from the Firehose data using a widely-used topic modeling algorithm, latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003). Latent Dirichlet allocation is an algorithm for the automated discovery of topics. LDA treats documents as a mixture of topics, and topics as a mixture of words. Each topic discovered by LDA is represented by a probability distribution which conveys the affinity for a given word to that particular topic. We analyze these distributions to understand the differences between the topics discovered in the two datasets. To get a sense of how the topics found in the Streaming data compare with those found with random samples, we compare with topics found by running LDA on random subsamples of the Firehose data.

Topic Discovery Here we compare the topics generated using the Firehose corpus with those generated using the Streaming corpus. LDA takes, in addition to the corpus, three parameters as its input: K - the number of topics, α - a hyperparameter for the Dirichlet prior topic distribution, and η - a hyperparameter for the Dirichlet prior word distribution. Choosing optimal parameters is a very challenging problem, and is not the focus of this work. Instead we focus on the similarity of the results given by LDA using identical parameters on both the Streaming and Firehose corpus. We set $K = 100$ as suggested by (Dumais et al. 1988) and use priors of $\alpha = 50/K$, and $\eta = 0.01$. The software we used to discover the topics is the *gensim* software package (Řehůřek and Sojka 2010). To get an understanding of the topics discovered at each level of Streaming coverage, we select the same days as we did for the comparison of Kendall's τ.

Topic Comparison To understand the differences between the topics generated by LDA, we compute the distance in their probability distribution using the Jensen-Shannon divergence metric (Lin Jan). Since LDA's topics have no implicit orderings we first must match them based upon the similarity of the words in the distribution. To do the matching we construct a weighted bipartite graph between the topics from the Streaming API and the Firehose. Treating each topic as a bag of words, we use the Jaccard score between the words in a Streaming topic T_i^S and a Firehose topic T_j^F as the weight of the edges in the graph,

$$d(T_i^S, T_j^F) = \frac{|T_i^S \cap T_j^F|}{|T_i^S \cup T_j^F|}. \tag{2}$$

After constructing the graph we use the maximum weight matching algorithm proposed in (Galil 1986) to find the best matches between topics from the Streaming and Firehose data. After making the ideal matches, we then compute the Jensen-Shannon divergence between the two topics. Treating each topic as a probability distribution, we compute this as follows:

$$JS(T_i^S || T_j^F) = \frac{1}{2}[KL(T_i^S || M) + KL(T_j^F || M)], \tag{3}$$

where $M = \frac{1}{2}(T_i^S + T_j^F)$ and KL is the Kullback-Liebler divergence (Cover and Thomas 2006). We compute the Jensen-Shannon divergence for each matched pair and plot a histogram of the values in Figure 6. We see a trend of higher divergence with lower coverage, and lower divergence with higher coverage. This shows that decreased coverage in the Streaming data causes variance in the discovered topics.

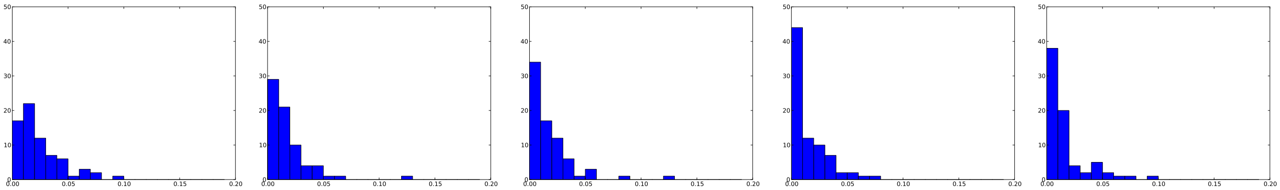
Comparison with Random Samples In order to get additional perspective on the accuracy of the topics discovered in the Streaming data, we compare the Streaming data with data sampled randomly from the Firehose, as we did earlier to compare the correlation. First, we compute the average of the Jensen-Shannon scores from the Streaming data in Figure 6, S. We then repeat this process for each of the 100 runs with random data, each run called x_i. Next, we use maximum-likelihood estimation (Casella and Berger 2001) to estimate the parameters of the Gaussian distribution from which these points originate, $\hat{\mu} = \frac{1}{100}\sum_{i=1}^{100} x_i$, and $\hat{\sigma} = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(x_i - \hat{\mu})^2}$. Finally, we compute the z-Score for S, $z = \frac{S-\hat{\mu}}{\hat{\sigma}}$. This score gives us a concrete measure of the difference between the Streaming API data and the random samples. Results of this experiment, including z-Scores are shown in Figure 7. Nonetheless, we are still able to get topics from the Streaming API that are close to those found in random data with higher levels of coverage. A threshold of *3-sigma* is often used in the literature to indicate extreme values (Filliben and Others 2002, Section 6.3.1). With this threshold, we see that overall we are able to get significantly better topics with the random data than with the Streaming API on 4 of the 5 days.

Network Measures

Because Twitter is a social network, Twitter data can be analyzed with methods from Social Network Analysis (Wasserman and Faust 1994) in addition to statistical measures. Possible 1-mode and 2-mode networks are: *User* × *User* retweet networks, *User* × *Hashtag* content networks, *Hashtag* × *Hashtag* co-occurrence networks. For the purpose of this article we focus on *User* × *User* retweet networks. Users who send tweets within a certain time period are the nodes in the network. Furthermore, users that are retweeted within this time period are also nodes in this network, regardless of the time their original tweet was tweeted. The networks created by this procedure are directed and not symmetric by design, however, bi-directional links are possible in case $a \rightarrow b$ and $b \rightarrow a$. We ignore line weight created by multiple $a \rightarrow b$ retweets and self-loops (yes, some user retweet themselves). For the network metrics, the comparison is done on both the network, and the node levels. Networks are analyzed using ORA (Carley et al. 2012).

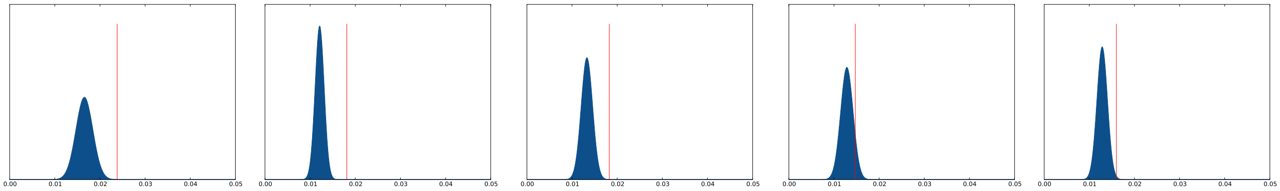
Node-Level Measures

The node-level comparison is accomplished by calculating measures at the user-level and comparing these results. We calculate three different *centrality measures* at the node level, two of which—Degree Centrality and Betweenness Centrality—were defined by Freeman as "distinct intuitive conceptions of centrality" (Freeman 1979, p. 215). Degree Centrality counts the number of neighbors in unweighted networks. In particular, we are interested in In-Degree Centrality as this reveals highly respected sources of information in the retweet network (where directed edges point to the source). Betweenness Centrality (Freeman 1979) identifies brokerage positions in the Twitter networks that connect



(a) Min. $\mu = 0.024$, $\sigma = 0.019$.
(b) Q1. $\mu = 0.018$, $\sigma = 0.018$.
(c) Median. $\mu = 0.018$, $\sigma = 0.020$.
(d) Q3. $\mu = 0.014$, $\sigma = 0.016$.
(e) Max. $\mu = 0.016$, $\sigma = 0.018$.

Figure 6: The Jensen-Shannon divergence of the matched topics at different levels of coverage. The x-axis is the binned divergence. No divergence was > 0.15. The y-axis is the count of each bin. μ is the average divergence of the matched topics, σ is the standard deviation.



(a) Min. $S = 0.024$, $\hat{\mu} = 0.017$, $\hat{\sigma} = 0.002$, $z = 3.500$.
(b) Q1. $S = 0.018$, $\hat{\mu} = 0.012$, $\hat{\sigma} = 0.001$, $z = 6.000$.
(c) Median. $S = 0.018$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 5.000$.
(d) Q3. $S = 0.014$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 1.000$.
(e) Max. $S = 0.016$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 3.000$.

Figure 7: The distribution of average Jensen-Shannon divergences in the random data (blue curve), with the single average obtained through the Streaming data (red, vertical line). z indicates the number of standard deviations the Streaming data is from the mean of the random samples.

different communities with each other or funnel different information sources. Furthermore, we calculate the *Potential Reach* which counts the number of nodes that are reachable in the network weighted with the path distance. In our Twitter networks this is equivalent to the inverse in-distance of reachable nodes (Sabidussi 1966). This approach results in a metric that finds sources of information (users) that potentially can reach many other nodes on short path distances. Before calculating these measures, we extract the main component and delete all other nodes (see next sub-section). In general, centrality measures are used to identify important nodes. Therefore, we calculate the number of top 10 and top 100 nodes that can be correctly identified with the Streaming data. Table 2 shows the results for the average of 28 daily networks, the *min-max* range, as well as the aggregated network including *all* 28 days.

Although, we know from previous studies (Borgatti, Carley, and Krackhardt 2006) that there is a very low likelihood that the ranking will be correct when handling networks with missing data, the accuracy of the daily results is not very satisfying. When we look at the results of the individual days, we can see that the matches have, once again, a broad range as a function of the data coverage rate. In (Borgatti, Carley, and Krackhardt 2006) the authors argue that network measures are stable for denser networks. Twitter data, being very sparse, causes the network metrics' accuracy to be rather low in the case when the data sub-sample is smaller. However, identifying ∼50% key-players correctly for a single day is reasonable, and accuracy can be increased by using longer observation periods. Even more, the Potential Reach metrics are quite stable for some days in the aggregated data.

Table 2: Average centrality measures for Twitter retweet networks for 28 daily networks. "All" is all 28 days together.

Measure	$k =$	Top$-k$ *(min-max)*	All
In-Degree	10	4.21 (0–9)	4
In-Degree	100	53.4 (36–82)	73
Potential Reach	100	59.2 (32–83)	80
Betweenness	100	54.8 (41–81)	55

Network-Level Measures

We complement our node-level analysis by comparing various metrics at the network level. These metrics are reported in Table 3 and are calculated as follows. Since retweet networks create a lot of small disconnected components, we focus only on the size of the largest component. The size of the main component and the fact that all smaller components contain less than 1% of the nodes justify our focus on the main component for this data. Therefore, we reduce the networks to their largest component before we proceed with the calculations. To describe the structure of the retweet networks we calculate the clustering coefficient, a measure for local density (Watts and Strogatz 1998). We do not take all possible triads of directed networks into account, but treat the networks as undirected when calculating the clustering coefficient. $D_{in} > 0$ shows the proportion of nodes in the largest component that are retweeted and $max(D_{in})$ shows the value of the highest unscaled In-Degree value, i.e., number of unique users retweeting the same single user. The final three lines of Table 3 are network centralization indexes based on the node-level measures that have been introduced in the previous paragraph. Freeman (Freeman 1979) describes the centralization C_X of a network for any given metric as the difference of the value $C_X(p*)$ of the most central node to all other node values compared to the maximum possible difference:

$$C_X = \frac{\sum i = 1n[C_X(p*) - C_X(p_i)]}{max \sum i = 1n[C_X(p*) - C_X(p_i)]} \quad (4)$$

High centralization indicates a network with some nodes having very high node-level values and many nodes with low values while low centralization is the result of evenly distributed node-level measures.

We do not discuss all details of the individual results but focus on the differences between the two data sources. First, the coverage of nodes and links is similar to the coverage of tweets. This is a good indicator that the sub-sample is not biased to the specific Twitter user (e.g. high activity). The smaller proportion of nodes with non-zero In-Degree for the Firehose shows us that the larger number of nodes includes many more peripheral nodes. A low Clustering Coefficient implies that networks are hierarchical rather than interacting communities. Even though the centralization indexes are rather similar, there is one very interesting result when looking at the individual days: The range of values is much higher for the Streaming data as a result of the high coverage fluctuation. Further research will analyze whether we can use network metrics to better estimate how sufficient the sampled Streaming data is.

Geographic Measures

The final facet of the Twitter data we compare is the geolocation of the tweets. Geolocation is an important part of a tweet, and the study of the location of content and users is currently an active area of research (Cheng, Caverlee, and Lee 2010; Wakamiya, Lee, and Sumiya 2011). We study how the geographic distribution of the geolocated tweets is affected by the sampling performed by the Streaming API.

The number of geotagged tweets is low, with only 16,739 geotagged tweets in the Streaming data (3.17%) and 18,579 in the Firehose data (1.45%). We notice that despite the difference in tweets collected on the whole we get 90.10% coverage of geotagged tweets. We start by grouping the locations of tweets by continent and can find a strong Asian bias due to the boundary box we used to collect the data from both sources, shown in Table 1. To better understand the distribution of geotagged tweets we repeat the same process, this time excluding tweets originating in the boundary box set in the parameters. After removing these tweets, more than 90% of geotagged Tweets from both sources are excluded from the data and the Streaming coverage level is reduced to 39.19%. The distribution of tweets by continent is shown in Table 4. Here we see a more even representation of the tweets' locations in Asia and North America.

Table 3: Comparison of Network-Level Social Network Analysis Metrics.

	Firehose		Streaming API	
Metrics	avg.day	28 days	avg.day	28 days
nodes	6,590	73,719	2,466 (37.4%)	30,894 (41.9%)
links	10,173	204,022	3,667 (36.0%)	76,750 (37.6%)
$D_{in} > 0$	25.1%	19.3%	32.4%	20.5%
$max(D_{in})$	341	2,956	167.3	1,252
main comp.	5,609	70,383	2,069	28,701
main comp. %	84.6%	95.5%	82.5%	92.9%
Clust.Coef.	0.029	0.053	0.033	0.050
DC_{in} Centr.	0.059	0.042	0.085	0.043
BC Centr.	0.010	0.053	0.010	0.050
$PReach$ Centr.	0.130	0.240	0.156	0.205

Table 4: Geotagged Tweet Location by Continent. Excluding boundary box from parameters.

Continent	Firehose	Streaming	Error
Africa	156 (5.74%)	33 (3.10%)	-2.64%
Antarctica	0 (0.00%)	0 (0.00%)	±0.00%
Asia	932 (34.26%)	321 (30.11%)	-4.15%
Europe	300 (11.03%)	139 (13.04%)	+2.01%
Mid-Ocean	765 (28.12%)	295 (27.67%)	-0.45%
N. America	607 (22.32%)	293 (27.49%)	+5.17%
Oceania	54 (1.98%)	15 (1.41%)	-0.57%
S. America	3 (0.11%)	2 (0.19%)	+0.08%
Total	2720 (100.00%)	1066 (100.00%)	±0.00%

Conclusion and Future Work

In this work we ask whether data obtained through Twitter's sampled Streaming API is a sufficient representation of activity on Twitter as a whole. To answer this question we collected data with exactly the same parameters from both the free, but limited, Streaming API and the unlimited, but costly, Firehose. We provide a methodology for comparing the two multifaceted sets of data and results of our analysis.

We started our analysis by understanding the coverage of the Streaming API data, finding that when the number of tweets matching the set of parameters increases, the Streaming API's coverage is reduced. One way to mitigate this might be to create more specific parameter sets with different users, bounding boxes, and keywords. This way we might be able to extract more data from the Streaming API.

Next, we studied the statistical differences between the two datasets. We used a common correlation coefficient to understand the differences between the top n hashtags in the two datasets. We find that the Streaming API data estimates the top hashtags for a large n well, but is often misleading when n is small. We also employed LDA to extract topics from the text. We compare the probability distribution of the words from the most closely-matched topics and find that they are most similar when the coverage of the Streaming API is greatest. That is, topical analysis is most accurate when we get more data from the Streaming API.

The Streaming API provides just one example of how sampling Twitter data affects measures. We leverage the Firehose data to get additional samples to better understand the results from the Streaming API. In both of the above experiments we compare the Streaming data with 100 datasets sampled randomly from the Firehose data. We compare the statistical properties to find that the Streaming API performs worse than randomly sampled data, especially at low coverage. We find that in the case of top hashtag analysis, the Streaming API sometimes reveals negative correlation in the top hashtags, while the randomly sampled data exhibits very high positive correlation with the Firehose data. In the case of LDA we find a significant increase in the accuracy of LDA with the randomly sampled data over the data from the Streaming API. Both of these results indicate some bias in the way that the Streaming API provides data to the user.

By analyzing retweet *User* × *User* networks we were able to show that we can identify, on average, 50–60% of the top 100 key-players when creating the networks based on one day of Streaming API data. Aggregating some days of data can increase the accuracy substantially. For network level measures, first in-depth analysis revealed interesting correlation between network centralization indexes and the proportion of data covered by the Streaming API.

Finally, we inspect the properties of the geotagged tweets from both sources. Surprisingly, we find that the Streaming API almost returns the complete set of the geotagged tweets despite sampling. We attribute this to the geographic boundary box. Although the number of geotagged tweets is still very small in general ($\sim$1%), researchers using this information can be confident that they work with an almost complete sample of Twitter data when geographic boundary boxes are used for data collection. When we remove the tweets collected this way, we see a much larger disparity in the tweets from both datasets. Even with this disparity, we see a similar distribution based on continent.

Overall, we find that the results of using the Streaming API depend strongly on the coverage and the type of analysis that the researcher wishes to perform. This leads to the next question concerning the estimation of how much data we actually get in a certain time period. We suggest that we found first evidence in different types of analysis that can help us to estimate the Streaming API coverage. Uncovering the nuances of the Streaming API will help researchers, business analysts, and governmental institutions to better ground their scientific results based on Twitter data.

Looking forward, we hope to find methods to compensate for the biases in the Streaming API to provide a more accurate picture of Twitter activity to researchers. Provided further access to Twitter's Firehose, we will determine whether the methodology presented here will yield similar results for Twitter data collected from other domains, such as natural disaster, protest, and elections.

Acknowledgements

This work is sponsored in part by Office of Naval Research grants N000141010091 and N000141110527. We thank Lei Tang for thoughtful advice and discussion throughout the development of this work.

References

Agresti, A. 2010. *Analysis of Ordinal Categorical Data*, volume 656. Hoboken, New Jersey: Wiley.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Borgatti, S. P.; Carley, K. M.; and Krackhardt, D. 2006. On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28(2):124–136.

Campbell, D. G. 2011. *Egypt Unshackled: Using Social Media to @#:) the System*. Amherst, NY: Cambria Books.

Carley, K. M.; Pfeffer, J.; Reminga, J.; Storrick, J.; and Columbus, D. 2012. ORA User's Guide 2012. Technical Report CMU-ISR-12-105, Carnegie Mellon University, School of Computer Science, Institute for Software Research, Pittsburgh, PA.

Casella, G., and Berger, R. L. 2001. *Statistical Inference*. Belmont, CA: Duxbury Press.

Chae, J.; Thom, D.; Bosch, H.; Jang, Y.; Maciejewski, R.; Ebert, D. S.; and Ertl, T. 2012. Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. Brighton, UK: IEEE Conference on Visual Analytics Science and Technology.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In *Proceedings of The 19th ACM International Conference on Information and Knowledge Management*, 759–768. Toronto, Ontario, Canada: International Conference on Information and Knowledge Management.

Costenbader, E., and Valente, T. W. 2003. The stability of centrality measures when networks are sampled. *Social networks* 25(4):283–307.

Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory*. Hoboken, New Jersey: Wiley InterScience.

De Choudhury, M.; Lin, Y.-R.; Sundaram, H.; Candan, K. S.; Xie, L.; and Kelliher, A. 2010. How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media. In *Proc. of the 4th Int'l AAAI Conference on Weblogs and Social Media*, 34–41. Washington, DC, USA: AAAI.

De Longueville, B.; Smith, R. S.; and Luraschi, G. 2009. "omg, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, 73–80. New York, NY, USA: ACM.

Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Deerwester, S.; and Harshman, R. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, 281–285. New York, NY, USA: ACM.

Efron, M. 2010. Hashtag retrieval in a microblogging environment. In *Proc. of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, 787–788. New York, NY, USA: ACM.

Filliben, J. J., and Others. 2002. NIST/SEMTECH Engineering Statistics Handbook. *Gaithersburg: www.itl.nist.gov/div898/handbook, NIST.*

Freeman, L. C. 1979. Centrality in Social Networks: Conceptual clarification. *Social Networks* 1(3):215–239.

Galil, Z. 1986. Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.* 18(1):23–38.

Gayo-Avello, D.; Metaxas, P. T.; and Mustafaraj, E. 2011. Limits of electoral predictions using twitter. In *Proceedings of the International Conference on Weblogs and Social Media*, volume 21, 490–493. Barcelona, Spain: AAAI.

Granovetter, M. 1976. Network Sampling: Some First Steps. *American Journal of Sociology* 81(6):1287–1303.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in twitter. In *Proc. First Workshop on Social Media Analytics*, SOMA '10, 80–88. NYC, NY, USA: ACM.

Hong, L.; Ahmed, A.; Gurumurthy, S.; Smola, A. J.; and Tsioutsiouliklis, K. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 769–778. New York, NY, USA: ACM.

Joseph, K.; Tan, C. H.; and Carley, K. M. 2012. Beyond "local", "categories" and "friends": clustering foursquare users with latent "topics". In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, 919–926. New York, NY, USA: ACM.

Kireyev, K.; Palen, L.; and Anderson, K. 2009. Applications of Topics Models to Analysis of Disaster-Related Twitter Data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, volume 1.

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

Kumar, S.; Barbier, G.; Abbasi, M. A.; and Liu, H. 2011. Tweettracker: An analysis tool for humanitarian and disaster relief. In *Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona Spain: AAAI.

Leskovec, J., and Faloutsos, C. 2006. Sampling from Large Graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 631–636.

Lin, J. Jan. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on* 37(1):145–151.

Mathioudakis, M., and Koudas, N. 2010. Twittermonitor: trend detection over the twitter stream. In *Proc. of the 2010 ACM SIGMOD Int'l Conference on Management of data*, SIGMOD '10, 1155–1158. New York, NY, USA: ACM.

Pozdnoukhov, A., and Kaiser, C. 2011. Space-time dynamics of topics in streaming text. In *Proc. of the 3rd ACM SIGSPATIAL Int'l Workshop on Location-Based Social Networks*, LBSN '11, 1–8. New York, NY, USA: ACM.

Recuero, R., and Araujo, R. 2012. On the rise of artificial trending topics in twitter. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, HT '12, 305–306. New York, NY, USA: ACM.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Sabidussi, G. 1966. The centrality index of a graph. *Psychometrika* 69:581–603.

Sofean, M., and Smith, M. 2012. A real-time architecture for detection of diseases using social networks: design, implementation and evaluation. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, HT '12, 309–310. New York, NY, USA: ACM.

Tsur, O., and Rappoport, A. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, 643–652. New York, NY, USA: ACM.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International Conference on Weblogs and Social Media*, 178–185. Washington, DC, USA: AAAI.

Wakamiya, S.; Lee, R.; and Sumiya, K. 2011. Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proc. of the 3rd ACM SIGSPATIAL Int'l Workshop on Location-Based Social Networks*, LBSN '11, 77–84. New York, NY, USA: ACM.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, MA: Cambridge University Press.

Watts, D. J., and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440–442.

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q. 2012. We know what @you #tag: does the dual role affect hashtag adoption? In *Proc. of the 21st int'l conference on World Wide Web*, WWW '12, 261–270. New York, NY, USA: ACM.

Yin, Z.; Cao, L.; Han, J.; Zhai, C.; and Huang, T. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, WWW '11, 247–256. New York, NY, USA: ACM.