Effects of Sampling on Twitter Trend Detection

Andrew Yates[1], Alek Kolcz[2], Nazli Goharian[1], Ophir Frieder[1]
[1]Information Retrieval Lab, Department of Computer Science, Georgetown University
[2]Twitter, Inc.
{andrew,nazli,ophir}@ir.cs.georgetown.edu

Abstract

Much research has focused on detecting trends on Twitter, including health-related trends such as mentions of Influenza-like illnesses or their symptoms. The majority of this research has been conducted using Twitter's public feed, which includes only about 1% of all public tweets. It is unclear if, when, and how using Twitter's 1% feed has affected the evaluation of trend detection methods. In this work we use a larger feed to investigate the effects of sampling on Twitter trend detection. We focus on using health-related trends to estimate the prevalence of Influenza-like illnesses based on tweets, and use ground truth obtained from the CDC and Google Flu Trends to explore how the prevalence estimates degrade when moving from a 100% to a 1% sample. We find that using the public 1% sample is unlikely to substantially harm ILI estimates made at the national level, but can cause poor performance when estimates are made at the city level.

Keywords: LR Infrastructures and Architectures, Social Media, Trend Detection

1. Introduction

Much work has focused on detecting health-related Twitter trends and, in particular, trends related to the prevalence of Influenza-like illnesses (ILI). Such trends are useful because they can provide prevalence estimates more quickly than formal reporting systems. The majority of this trend detection research has been conducted on a 1% sample of Twitter's public feed, however, and it is unclear whether this has biased evaluation results. In this work we use a complete feed of all public tweets to investigate the effects of using the 1% sample. We evaluate the performance of a trend detection method using progressively larger sample sizes to characterize how large the sample must be before evaluation results begin to converge.

Influenza-like illnesses are unusual in that they follow a seasonal pattern, affect a relatively large percentage of the population, and are closely monitored by organizations such as the ECDC (European Centre for Disease Prevention and Control) and CDC (the United States' Centers for Disease Control and Prevention). Publicly available prevalence data has been widely used to evaluate how well trend detection methods can estimate ILI prevalence, making ILI prevalence estimation well-suited for evaluating the performance of trend detection methods.

We use CDC and GFT (Google Flu Trends) (Ginsberg et al., 2009) ILI prevalence data to evaluate how methods for detecting health-related Twitter trends (Yates et al., 2014) perform when used with varying sample sizes. Our contributions are:

- an exploration of how sampling affects flu detection, and health-trend detection in general, on Twitter;
- an analysis of the degree to which inaccuracies introduced by sampling can be explained by the number of tweets being analyzed; and
- suggestions for reducing the effects of sampling when using a sampled Twitter feed.

2. Related Work

Several researchers have investigated how Twitter chooses the tweets in its 1% sample. Joseph et al. (2014) investigate whether different clients receive the same 1% samples. They find that there is, on average, 96% overlap between the samples received by different clients. Twitter's documentation has since been updated to state that the "Tweets returned by the default access level are the same, so if two different clients connect to this endpoint, they will see the same Tweets."[1]. Kergl et al. (2014) reverse engineer the tweet IDs returned by Twitter's API, and discover that the IDs contain a timestamp and information about the Twitter infrastructure that processed the tweet. They determine that tweets are chosen for the 1% sample based on their timestamp, with all tweets made in a specific 10ms interval of each second appearing in the sample.

Ghosh et al. (2013) explore the utility of sampling users rather than tweets. They identify 500,000 expert users, retrieve their tweets, and compare their tweets to Twitter's 1% sample. They find that the tweets from expert users contain less spam, cover a wider range of topics, and report breaking news slightly faster than those tweets in the 1% sample. They note that sampling users does not result in a representative sample, however, and is thus not a replacement for sampling tweets directly.

Morstatter et al. (2013) compare data from Twitter's Streaming API, which returns tweets matching a query (e.g., keywords and geographic area) but is capped at 1% of all tweets, to Twitter's Firehose API (i.e., all public tweets). They find that the two APIs differ in terms of their hashtag distribution, the topics detected in tweets, and the social network structure. They also find that Twitter's 1% sample is similar to the Firehose, however, indicating that most of the bias they observed is introduced by the Streaming API. While the Streaming API is a useful tool for collecting all future tweets matching a query, it does not return historical tweets and cannot be used as a substitute for the

[1]dev.twitter.com/streaming/reference/get/statuses/sample

Sampling API in all scenarios, such as for the real-time detection of new topics. Morstatter et al. (2014) further investigate the issue of bias in the Streaming API. They find that the Sampling API can be used to identify periods during which the Streaming API was biased, and they confirm that the distribution of hashtags in the Sampling API is similar to the hashtag distribution found in the Firehose. Given that the Streaming API's bias appears to be correlated with the number of tweets it returns, Sampson et al. (2015) investigate methods for splitting the Streaming API's query into multiple disjoint queries. By reducing the number of tweets returned by each Streaming API query, they are able to reduce the API's bias. We focus on randomly sampled tweets (i.e., tweets from the Sampling API or from a random sampling of the Firehose) rather than on tweets that satisfy a query.

Many researchers have considered the problem of estimating the flu's real-world prevalence from Twitter data. (Achrekar et al., 2011; Paul et al., 2014; Chew and Eysenbach, 2010; Gesualdo et al., 2013; Aramaki et al., 2011; Ji et al., 2012; Culotta, 2010; Achrekar et al., 2012; Sadilek et al., 2012; Lamb et al., 2013; Yates et al., 2014; Parker et al., 2015; Nagel et al., 2013; de Quincey and Kostkova, 2010; Lampos and Cristianini, 2010; Szomszor et al., 2012; Nagar et al., 2014; Chakraborty et al., 2014) Yates et al. (2016a) model relationships between drug, symptom, and medical condition Twitter mentions with the goal of improving health-trend detection models. Most researchers have focused on training supervised methods to estimate flu prevalence, but some work has also considered the performance of unsupervised methods (Yates et al., 2014; Parker et al., 2015). Supervised methods generally perform well when used to estimate ILI prevalence based on Twitter; such methods can also be used to augment the performance of existing prevalence estimation models, such as Google Flu Trends, and can be trained using ILI prevalence data provided by the CDC or Google Flu Trends as ground truth. (Chakraborty et al., 2014; Paul et al., 2014)

3. Methodology

Many methods have been proposed for estimating the flu's prevalence based on Twitter activity. While supervised models have performed best at this task (Lamb et al., 2013; Aramaki et al., 2011; Achrekar et al., 2011), keyword-based approaches have been shown to perform acceptably (Aramaki et al., 2011) and have formed the basis for the U.S. Department of Health and Human Services' Now Trending Challenge [2]. To avoid biasing our results towards a supervised model designed only for flu detection, we use a keyword-based sliding window approach from a framework for Twitter public health surveillance (Yates et al., 2014) to match the ILI terms in the Now Trending Challenge's Illness Term Taxonomy. In short, we count the number of flu-related terms from the Illness Term Taxonomy in an n-term sliding window, and normalize each daily count by the total number of tweets posted on that day. We utilize a 2-term sliding window since the majority of the Illness Terms are single terms, but some may be tokenized into two consecutive terms.

[2]http://www.nowtrendingchallenge.com/

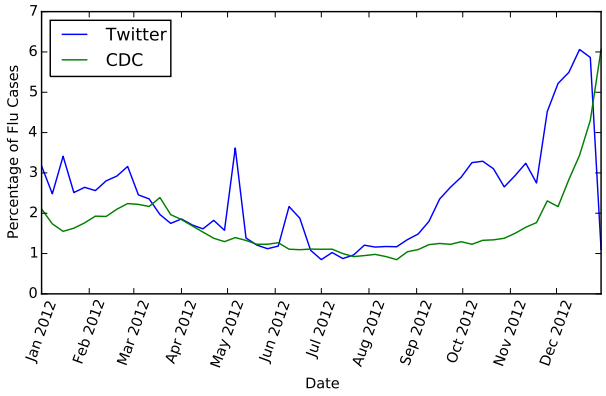

Figure 1: Flu prevalence estimates from Twitter's full feed of public tweets using an unsupervised method. Even when using an unsupervised method, the Twitter prevalence estimates are strongly correlated with CDC data and Google Flu Trends estimates, with Spearman's $\rho = 0.79$ for Google Flu Trends and $\rho = 0.73$ for the CDC data.

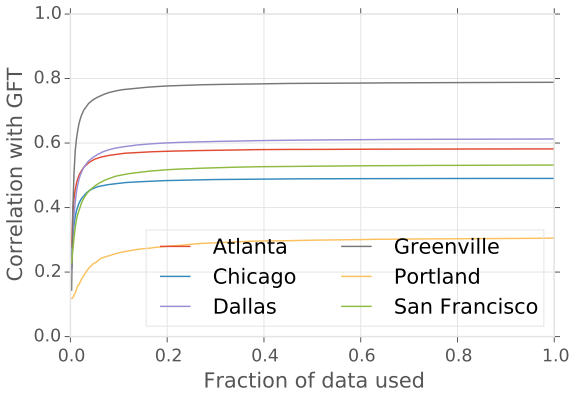

Figure 2: Correlation with Google Flu Trends for various cities as the Twitter sample size increases from a 1% sample to the full feed. Every city's correlation sharply increases until the sampling level reaches about 5%, at which point the correlations begin to increase slowly.

We use CDC Influenza Surveillance[3] and Google Flu Trends (Ginsberg et al., 2009) flu prevalence data as ground truth to investigate how the full Twitter feed compares to more readily-available samples, such as the 1% of tweets available from the Sampling API. To validate our keyword-based flu detection methodology, we compute the correlation between the full feed and each ground truth using Spearman's ρ, and find both correlations to be reasonable, though lower than those generally obtained with supervised methods designed for ILI prevalence estimation (Lamb et al., 2013; Paul et al., 2014). Using an unsupervised model avoids biasing our results towards a supervised model designed only for this domain; our unsupervised approach is not dependent on the flu detection domain, because we simply count relevant keywords rather than utilizing a super-

[3]http://www.cdc.gov/flu/weekly/overview.htm

	Fraction of tweets sampled										
	0.005	0.010	0.015	0.020	0.050	0.100	0.200	0.300	0.400	0.500	1.000
Mean	0.09	0.18	0.22	0.25	0.33	0.37	0.40	0.41	0.42	0.42	0.43
Median	0.09	0.17	0.22	0.24	0.33	0.36	0.40	0.40	0.41	0.42	0.42
Top 25% of cities	0.32	0.42	0.46	0.49	0.57	0.61	0.63	0.64	0.64	0.65	0.65
Top 10% of cities	0.42	0.54	0.59	0.62	0.68	0.71	0.72	0.73	0.73	0.73	0.74

Table 1: The Pearson correlation between Google Flu Trends and Twitter ILI prevalence estimates at different sampling levels. The 0.01 level corresponds to Twitter's public 1% API. The mean correlation, median correlation, correlation when only the top 25% of cities are considered, and correlation when only the top 10% of cities are considered are shown.

	Fraction of tweets sampled										
	0.005	0.010	0.015	0.020	0.050	0.100	0.200	0.300	0.400	0.500	1.000
Weekly mean (cities)	0.22	0.42	0.52	0.59	0.77	0.87	0.93	0.96	0.97	0.98	1.00
Monthly mean (cities)	0.51	0.64	0.72	0.77	0.88	0.94	0.97	0.98	0.99	0.99	1.00
Weekly mean (US)	0.98	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Monthly mean (US)	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: The fraction of the maximum correlation between Google Flu Trends and Twitter ILI prevalence estimates at different sampling levels and at different aggregation levels (i.e., by week and by month). The 0.01 level corresponds to Twitter's public 1% API, which reaches 42% of its maximum correlation with weekly aggregation and 64% of its maximum correlation with monthly aggregation. Regardless of the aggregation level, a 10% sample comes close to the correlation obtained using the full Twitter feed (i.e., the 1.0 sampling level). The correlation comes to within 98% of its maximum value when only 50% of the tweets are used (i.e., the 0.5 sampling level).

vised model. Figure 1 shows the CDC data and estimated prevalence based on all public tweets made in 2012 and aggregated by week. Spearman's $\rho = 0.79$ for Google Flu Trends data (Pearson's $r = 0.52$) and $\rho = 0.73$ for the CDC data. We use this methodology to compare trend detection results given different sampling levels (e.g., the 1% sample vs. the full Twitter feed).

4. Experiments

We use our trend detection methodology and corpus of all public tweets made in 2012[4] to evaluate:

- how sampling affects the correlation between Twitter Influenza-like illness (ILI) prevalence estimates and Google Flu Trends estimates,
- the impact of aggregating trends over different timespans (e.g., weekly vs. monthly) on prevalence estimation performance, and
- the impact of the number of ILI-related tweets prevalence estimation performance.

4.1. Google Flu Trends

We investigate the effects of sampling on Twitter flu detection by randomly sampling from our Twitter corpus and computing correlations with Google Flu Trends using the sampled data. That is, we randomly select $n\%$ of the tweets from the full Twitter feed, use our methodology to estimate the ILI prevalence based on these tweets, and compute the correlation between our prevalence estimate and Google Flu Trends' estimate. We evaluate each sampling level from 0.002 to 1.00 in increments of 0.002. For each level, we sample $5,000$ times and calculate the sampling level's mean correlation. We aggregate Twitter activity by week to correspond with Google Flu Trends estimates, which are released at the week level. We use only geolocated tweets (i.e., tweets that have either a geotag or other location information associated with them) and compute prevalence estimates over each geographic region. We consider only the 70 cities that appear in both Google Flu Trends and our Twitter data[5].

4.1.1. Correlation

The Pearson correlations between Google Flu Trends and our Twitter prevalence estimates at different sampling levels are shown in Table 1. Correlations are computed across all 70 cities. The mean correlation (i.e., 0.33) reaches 75% of its maximum correlation (i.e., 0.42) when only 5% of the Twitter data are used and 86% of its maximum when 10% of the data are used. This is substantially more data than

[4] In the *Tweet Count Correlation* section we report the number of ILI-related tweets associated with different geographic areas. The exact number of tweets made in 2012 is proprietary information, however, so we do not report it.

[5] Albany, Albuquerque, Anchorage, Atlanta, Austin, Baltimore, Baton Rouge, Birmingham, Boise, Boston, Buffalo, Charlotte, Chicago, Cleveland, Colorado Springs, Columbia, Columbus, Dallas, Dayton, Denver, Des Moines, Eugene, Fresno, Gainesville, Grand Rapids, Greensboro, Greenville, Honolulu, Houston, Indianapolis, Jackson, Jacksonville, Kansas City, Knoxville, Las Vegas, Lexington, Lincoln, Little Rock, Los Angeles, Lubbock, Madison, Memphis, Miami, Milwaukee, Nashville, New Orleans, Norfolk, Oklahoma City, Omaha, Orlando, Philadelphia, Phoenix, Pittsburgh, Portland, Providence, Reno, Richmond, Rochester, Sacramento, Salt Lake City, San Antonio, San Diego, San Francisco, Seattle, Spokane, Springfield, Tampa, Tucson, Tulsa, and Wichita

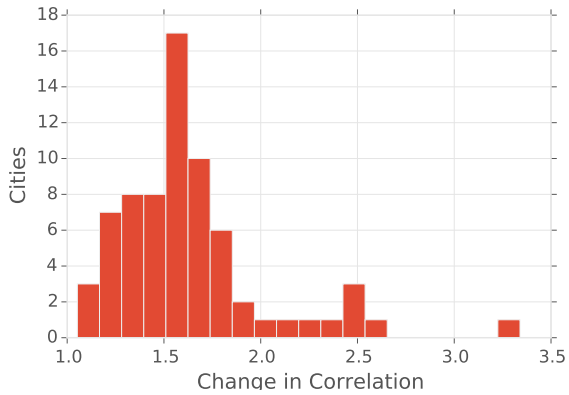

Figure 3: Increases in each of the 70 cities' correlations with Google Flu Trends when moving from a 1% sample to the full feed.

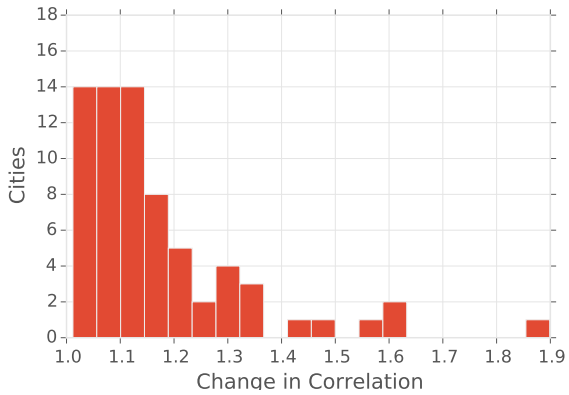

Figure 4: Increases in each of the 70 cities' correlations with Google Flu Trends when moving from a 10% sample to the full feed.

	Fraction of tweets sampled										
	0.0075	0.01	0.015	0.02	0.05	0.10	0.20	0.30	0.40	0.50	1.0
Pearson correlation	0.27	0.25	0.22	0.20	0.15	0.12	0.09	0.08	0.08	0.08	0.07
Spearman correlation	0.70	0.69	0.68	0.66	0.54	0.45	0.39	0.36	0.35	0.34	0.30

Table 3: The correlation between the tweet count and GFT Pearson correlation at each sampling level. Both the Pearson and Spearman correlation sharply decrease as the sampling level increases. At the 20% sampling level both correlations are within 30% of their correlations at the 100% sampling level and within 15% at the 50% sampling level. This suggests that poor ILI prevalence estimation performance is partially caused by a low number of tweets, but as the number of tweets increases with the sampling level other factors become responsible for the prevalence estimation's performance.

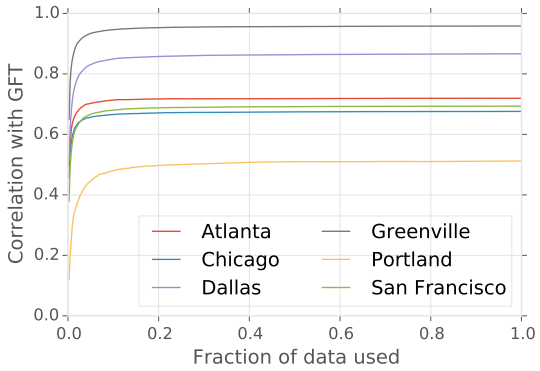

Figure 5: Correlation with Google Flu Trends for various cities when aggregating tweets at the month level. As with week-level aggregation, each city's correlation sharply increases until the sampling level reaches approximately 5%. The rate of increase is lower than that observed with week-level aggregation, however.

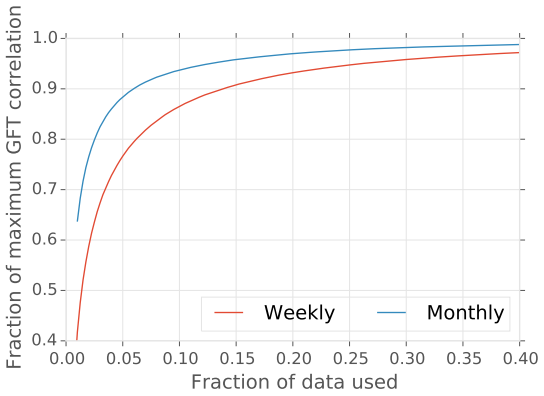

Figure 6: Correlation with Google Flu Trends as the fraction of the city's maximum correlation. While the correlation sharply increases with both aggregation levels, the increase is sharper with monthly aggregation. Monthly aggregation reaches 90% of its maximum correlation once the sampling level reaches about 6%, whereas weekly aggregation does not reach 90% of its maximum correlation until the sampling level reaches about 13%.

Twitter's 1% API provides, however, and when using a 1% sample the mean correlation is only 40% of its maximum value. Sampling disproportionately affects the cities that perform poorly; when only the top 10% of the cities are considered, the mean correlation reaches 73% of its maximum value at a 1% sampling level.

Figure 2 shows the correlation between Google Flu Trends and our flu prevalence estimates for six cities as the sample size increases. While the maximum correlations vary, the trend appears similar for each city. In each case, the correlation increases substantially from the 1% sample to the full feed. Flu detection performance increases much more slowly once the sample size reaches approximately 5%, however, indicating that little is gained by adding the

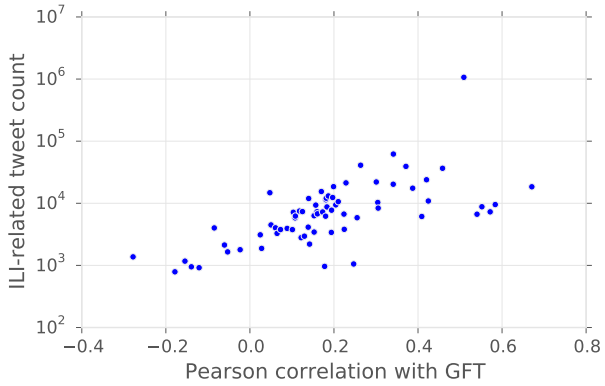

Figure 7: The relationship at the 1% sampling level between the number of flu tweets in a city and the Pearson correlation between Twitter flu prevalence estimates and Google Flu Trends' estimates. The tweet count and GFT correlation are fairly correlated, with Pearson's $r = 0.25$ and Spearman's $\rho = 0.69$, suggesting a strong monotonic relationship and weak linear relationship.

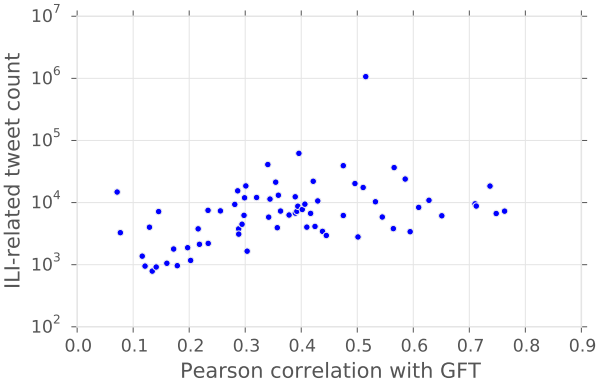

Figure 8: The relationship at the 10% sampling level between the number of flu tweets in a city and the Pearson correlation between Twitter flu prevalence estimates and Google Flu Trends' estimates. The tweet count and GFT correlation are much less correlated than at the 1% sampling level, with Pearson's $r = 0.12$ and Spearman's $\rho = 0.45$.

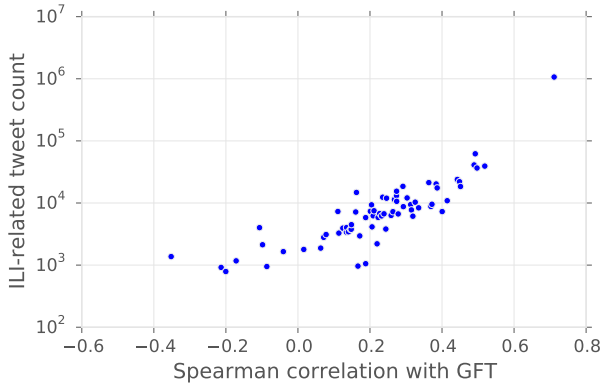

Figure 9: The relationship at the 1% sampling level between the number of flu tweets in a city and the Spearman correlation between Twitter flu prevalence estimates and Google Flu Trends' estimates. The tweet count and GFT correlation are relatively highly correlated, with Pearson's $r = 0.37$ and Spearman's $\rho = 0.86$.

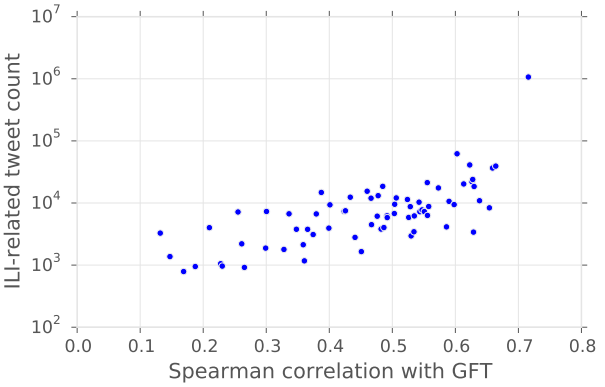

Figure 10: The relationship at the 10% sampling level between the number of flu tweets in a city and the Spearman correlation between Twitter flu prevalence estimates and Google Flu Trends' estimates. The tweet count and GFT correlation are fairly correlated, with Pearson's $r = 0.26$ and Spearman's $\rho = 0.69$.

remaining 90% of data. This is significant because Twitter's Streaming API can return a sample containing more than 1% of ILI tweets under some circumstances (i.e., the Streaming API returns public tweets that match a query; it returns up to 1% of all public tweets, which can be more than 1% of tweets matching the query).

Figure 3 and Figure 4 show the distribution of the changes in correlation when moving from a 1% sample to the full feed and from a 10% sample to the full feed, respectively. With a 10% sample the vast majority of cities receive a correlation increase of 15% or less when the full feed is used, whereas the majority of cities receive a correlation increase of over 50% when the 1% sample is used. These results further illustrate that moving from a 1% feed to a 10% sample can substantially improve results, but the difference between a 10% sample and the full feed is relatively small.

These results do not hold true for national ILI trends; when all US tweets are considered, a 1% sample comes within 96% of the Google Flu Trends correlation obtained using the full Twitter feed. In the *Tweet Count Correlation* section we investigate the degree to which differences in tweet counts explain this discrepancy.

4.1.2. Impact of Aggregation

We investigate how tweet aggregation interacts with sampling. While Google Flu Trends does not provide the data necessary to compare prevalence estimates at the daily-level, we can aggregate both our Twitter ILI prevalence estimates and Google Flu Trends' estimates at the monthly level. Figure 5 shows the relationship between the GFT correlation and the sampling level when tweets are aggregated by month. The increase in correlation appears simi-

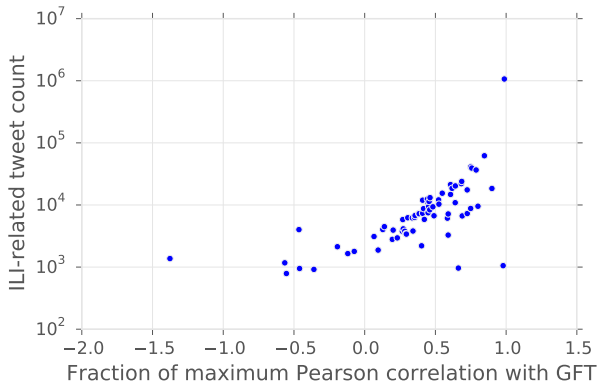

Figure 11: The relationship between the number of flu tweets in a city and the fraction of the maximum GFT Pearson correlation achieved at the 1% sampling level (i.e., $corr_{0.01}/corr_{1.0}$). There is a weak Pearson correlation ($r = 0.23$) and strong Spearman correlation ($\rho = 0.75$) between the two.

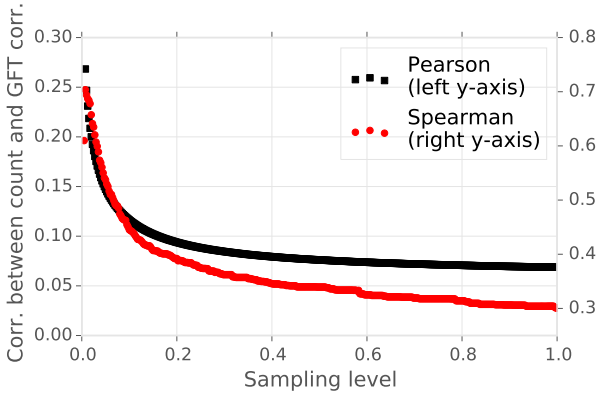

Figure 12: The correlation between the tweet count and GFT Pearson correlation at each sampling level. The correlation decreases sharply as the sampling level increases, with Pearson's $r = 0.25$ at the 1% sampling level decreasing to $r = 0.12$ at the 10% sampling level before reaching its final value of $r = 0.07$ at the 100% sampling level.

lar to that in Figure 2, which shows the relationship when tweets are aggregated by week. The relative performance of the six cities change, however, which suggests that the aggregation is not only boosting the average correlation.

To determine whether monthly aggregation is changing the rate at which prevalence estimates converage, for each sampling level we compute how close an estimates' correlation with GFT is to the correlation's maximum value. That is, we compute $corr_s/corr_{1.0}$ where $corr_s$ is the correlation with GFT at sampling level s. The results over all 70 cities are shown in Figure 6. While estimates at both aggregation levels quickly approach their maximum correlations, the monthly aggregation level results in a faster increase. When using a 10% sample, the monthly aggregation level's correlation is 94% of its maximum, whereas the weekly aggregation level's correlation is 87% of its maximum. Similarly, when using a 1% sample the monthly aggregation's correlation is 64% of its maximum, whereas the weekly aggregation's correlation is 42% of its maximum. Table 2 shows the results averaged over the 70 cities in our dataset and the results when all US tweets are used. The correlation quickly reaches its maximum correlation when all US tweets are used, with a 99% weekly correlation and 100% monthly correlation at the 1% sampling level.

Our analysis suggests that when researchers must utilize Twitter's 1% sample, they can reduce the effects of sampling by increasing the aggregation level they use. While monthly ILI prevalence estimates are much less useful than weekly estimates, a monthly aggregation level can still be helpful in assessing the quality of different trend detection algorithms. Changing the aggregation level makes little difference for ILI prevalence estimation at the US national level, however, since both weekly and monthly aggregation perform well with a 1% sample.

4.2. Tweet Count Correlation

Our previous experiments have shown that city-level ILI prevalence estimation performance increases as larger samples are used (e.g., the average Pearson correlation at the 1% sampling level is 42% of the maximum and 93% of the maximum at the 20% level), but we did not observe the same trend for estimates made over the entire United States (e.g., the Pearson correlation at the 0.5% sampling level is 98% of the maximum correlation). In this section we investigate how well this performance increase correlates with the total number of ILI tweets used in the analysis. That is, we investigate whether poor preformance at low sampling levels is caused by a low number of tweets. Figures 7 and 8 show the relationship between the ILI tweet count and ILI prevalence estimation performance (i.e., the ILI prevalence estimate's Pearson correlation with Google Flu Trends' estimate) at the 1% and 10% sampling levels, respectively. Tweet count and estimation performance have a low Pearson correlation ($r = 0.25$) and high Spearman correlation ($\rho = 0.69$) at the 1% sampling level, suggesting that there is a monotonic relationship between the two (i.e., more tweets result in better performance), but the relationship is not linear (i.e., performance does not increase linearly with the number of tweets). Both correlations decrease substantially at the 10% sampling level (i.e., Pearson's $r = 0.12$ and Spearman's $\rho = 0.45$), illustrating the diminishing returns of increasing the sampling level. As shown in figures 9 and 10, similar trends occur when we consider the Spearman correlation rather than the Pearson correlation between our Twitter prevalence estimates and GFT's estimates.

To correct for cities that always perform poorly, we can measure each city's fraction of its maximum GFT correlation at the 1% sampling level rather than measuring its GFT correlation at the 1% sampling level directly. That is, we compare the tweet count to $corr_{0.01}/corr_{1.0}$, where $corr_s$ is the Pearson correlation between Twitter and GFT at the s sampling level. The comparison is shown in Figure 11. As in the previous comparisons at the 1% sampling level, there is a weak Pearson correlation and a strong Spearman correlation between the two.

To determine how strongly the sampling level is associated with the correlation between tweet count and Twitter prevalence estimation performance (i.e., correlation with Google Flu Trends), we plot their relationship in Figure 12 and show several points in Table 3. The performance correlation (i.e., the correlation between the tweet count and prevalence estimation performance) quickly decreases as the sampling level increases, further illustrating the diminishing returns of increasing the sampling level after about 10%. These results suggest that low numbers of tweets can cause poor prevalence estimation performance, but the impact of a low tweet count quickly diminishes as the sampling level increases; poor performance at sampling levels above approximately 10% is not highly correlated with the tweet count, and may be influenced by other factors such as the noise added by non-experiential tweets written in response to prominent news articles (Yates et al., 2016b). We observe that cities which had at least 10^4 ILI-related tweets in 2012 tend to have better prevalence estimation performance even at the 1% sampling level, so it may be helpful to require a minimum number of relevant tweets when trend detection is performed on a 1% sample. We note that this does not apply to ILI prevalence estimation at the national level, where even a 0.5% sample performs well.

5. Conclusion

We investigated the impact of sampling Twitter's 100% API stream on Influenza-like illness (ILI) prevalence estimation by comparing prevalence estimates at different sampling levels to the estimates provided by Google Flu Trends. Our results show that sampling has little impact on ILI prevalence estimation when performed at the national level, but sampling levels below 5-10% significantly degrade prevalence estimation performed at the city level. We examined the correlation between the number of ILI-related tweets associated with a city and the city's ILI prevalence estimation performance, and found that low number of tweets are only correlated with poor estimation performance at low sampling levels (i.e., sampling levels below 10%). Futhermore, we found that aggregating tweets by month rather than by week reduces the effects of sampling, and we observed that cities which had at least 10^4 ILI-related tweets were less likely to perform poorly at low sampling levels. Based on these findings, we suggest that researchers excercise caution when using Twitter's public 1% sample to identify trends in small geographic areas at the daily or weekly aggregation level; researchers may be able to reduce the impact of using the sample API by aggregating tweets by month or by performing their analysis over a larger geographic area.

6. Acknowledgments

This work was partially supported by the US National Science Foundation through grant CNS-1204347.

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707, April.

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2012). Twitter improves seasonal influenza prediction. In Emmanuel Conchon, et al., editors, *HEALTHINF*, pages 61–70. SciTePress.

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chakraborty, P., Khadivi, P., Lewis, B., Mahendiran, A., Chen, J., Butler, P., Nsoesie, E. O., Mekaru, S. R., Brownstein, J. S., Marathe, M., et al. (2014). Forecasting a moving target: Ensemble models for ili case count predictions. In *SIAM international conference on Data Mining*.

Chew, C. and Eysenbach, G. (2010). Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11):1–13, 11.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

de Quincey, E. and Kostkova, P. (2010). Early warning and outbreak detection using social networking websites: The potential of twitter. In Patty Kostkova, editor, *Electronic Healthcare*, volume 27 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 21–24. Springer Berlin Heidelberg.

Gesualdo, F., Stilo, G., Agricola, E., Gonfiantini, M. V., Pandolfi, E., Velardi, P., and Tozzi, A. E. (2013). Influenza-like illness surveillance on twitter through automated learning of nave language. *PLoS ONE*, 8(12), 12.

Ghosh, S., Zafar, M. B., Bhattacharya, P., Sharma, N., Ganguly, N., and Gummadi, K. (2013). On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1739–1744, New York, NY, USA. ACM.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014. doi:10.1038/nature07634.

Ji, X., Chun, S., and Geller, J. (2012). Epidemic outbreak and spread detection system based on twitter data. In Jing He, et al., editors, *Health Information Science*, volume 7231 of *Lecture Notes in Computer Science*, pages 152–163. Springer Berlin Heidelberg.

Joseph, K., Landwehr, P., and Carley, K. (2014). Two 1twitters streaming api. In WilliamG. Kennedy, et al., editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 8393 of *Lecture Notes in Computer Science*, pages 75–83. Springer International Publishing.

Kergl, D., Roedler, R., and Seeber, S. (2014). On the endogenesis of twitter's spritzer and gardenhose sample

streams. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 357–364, Aug.

Lamb, A., Paul, M. J., and Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. *Proceedings of NAACL-HLT*, pages 789–795.

Lampos, V. and Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416, June.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *International AAAI Conference on Weblogs and Social Media*.

Morstatter, F., Pfeffer, J., and Liu, H. (2014). When is it biased?: Assessing the representativeness of twitter's streaming api. In *WWW (Companion Volume)*, pages 555–556.

Nagar, R., Yuan, Q., Freifeld, C., Santillana, M., Nojima, A., Chunara, R., and Brownstein, J. (2014). A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research*, 16(10):e236, 10.

Nagel, A. C., Tsou, M.-H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., Yang, J.-A., Han, S., Peddecord, K. M., Lindsay, S., et al. (2013). The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *Journal of medical Internet research*, 15(10):e237.

Parker, J., Yates, A., Goharian, N., and Frieder, O. (2015). Health-related hypothesis generation using social media data. *Social Network Analysis and Mining*, 5(1).

Paul, M. J., Dredze, M., and Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Sadilek, A., Kautz, H. A., and Silenzio, V. (2012). Modeling spread of disease from social interactions. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*.

Sampson, J., Morstatter, F., Maciejewski, R., and Liu, H. (2015). Surpassing the limit: Keyword clustering to improve twitter sample coverage. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 237–245, New York, NY, USA. ACM.

Szomszor, M., Kostkova, P., and de Quincey, E. (2012). #swineflu: Twitter predicts swine flu outbreak in 2009. In Martin Szomszor et al., editors, *Electronic Healthcare*, volume 69 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 18–26. Springer Berlin Heidelberg.

Yates, A., Parker, J., Goharian, N., and Frieder, O. (2014). A framework for public health surveillance. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Yates, A., Goharian, N., and Frieder, O. (2016a). Learning the relationships between drug, symptom, and medical condition mentions in social media. In *International AAAI Conference on Weblogs and Social Media*.

Yates, A., Joselow, J., and Goharian, N. (2016b). The news cycles influence on social media activity. In *International AAAI Conference on Weblogs and Social Media*.