Ebola data from the Internet: An Opportunity for Syndromic Surveillance or a News Event?

Elad Yom-Tov
Microsoft Research
13 Shenkar st.
Herzeliya 46733, Israel
eladyt@microsoft.com

ABSTRACT

Syndromic surveillance refers to the analysis of medical information for the purpose of detecting outbreaks of disease earlier than would have been possible otherwise and to estimate the prevalence of the disease in a population. Internet data, especially search engine queries and social media postings, have shown promise in contributing to syndromic surveillance for influenza and dengue fever. Here we focus on the recent outbreak of Ebola Virus Disease and ask whether three major sources of Internet data could have been used for early detection of the outbreak and for its ongoing monitoring. We analyze queries submitted to the Bing search engine, postings made by people using Twitter, and news articles in mainstream media, all collected from both the main infected countries in Africa and from across the world between November 2013 and October 2014. Our results indicate that it is unlikely any of the three sources would have provided an alert more than a week before the official announcement of the World Health Organization. Furthermore, over time, the number of Twitter messages and Bing queries related to Ebola are better correlated with the number of news articles than with the number of cases of the disease, even in the most affected countries. Information sought by users was predominantly from news sites and Wikipedia, and exhibited temporal patterns similar to those typical of news events. Thus, it is likely that the majority of Internet data about Ebola stems from news-like interest, not from information needs of people with Ebola. We discuss the differences between the current Ebola outbreak and seasonal influenza with respect to syndromic surveillance, and suggest further research is needed to understand where Internet data can assist in surveillance, and where it cannot.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Search process; J.3 [**Life and Medical Sciences**]: Health

DH'15, May 18–20, 2015, Florence, Italy.

ACM 978-1-4503-3492-1/15/05 ...$15.00.
http://dx.doi.org/10.1145/2750511.2750512.

General Terms

Experimentation

Keywords

Ebola,Syndromic surveillance,Query logs

1. INTRODUCTION

The Centers for Disease Control (CDC) define syndromic surveillance as methods that are tasked to "detect outbreaks of disease earlier than would otherwise be possible with traditional public health methods" [6]. Research in syndromic surveillance has attempted to detect outbreaks of disease earlier than would have been possible otherwise [5] and also to estimate the number of cases of the disease in a population [7].

Data generated by users of the Internet has long been claimed to be effective for the detection of outbreaks of viral disease, specifically, Influenza and Dengue fever. Search engine queries [8], online advertisements [3] and social media postings [2] have all been demonstrated as a source for such detection and tracking. In all cases, these data are used to model disease incidence through the appearance of specific textual terms.

Recently, however, Google Flu Trends, one of the most widely-used systems for Internet-based surveillance of Influenza, was shown to overestimate influenza rates [1] in two consecutive years. This was partly due to intense media interest in Influenza, which drove users to search for information on the disease. The resulting increase of terms related to Influenza in Google searches caused an overestimate of Influenza rates.

An outbreak of Ebola Virus Disease (henceforth Ebola) began in December 2013 in Guinea, and spread to neighboring Sierra Leone and Liberia. Several cases were detected in other neighboring countries, including Nigeria, Mali, and Senegal [9]. Sporadic cases (mostly as a result of medical treatment given to Ebola patients) were detected in the USA and Spain. To date, this is the largest Ebola outbreak ever documented.

In the case of the current Ebola epidemic, the index case occurred in a remote region of Guinea. Between February and mid-March, health authorities began noticing a hemorrhagic fever which infected 35 people. This disease was confirmed to be Ebola on March 22nd.[1] Therefore, an early warning for Ebola would have to occur before March 22nd.

[1] `http://www.nbcnews.com/news/africa/guinea-confirms-ebola-has-killed-59-n59686`

However, given the limited health infrastructure in the infected countries, monitoring the number of cases would also provide useful information to aid agencies. Therefore, in this paper we discuss both aspects of syndromic surveillance.

We note that the 2013-2014 Ebola outbreak occurred in a region of the world where the Internet has relatively limited penetration: According to the International Telecommunications Union[2] as of 2013, only 1.6% of people in Guinea, 16.5% in Liberia, and 1.7% in Sierra Leone have access to the Internet. This limits the number of people who generate data from the most infected regions, though the absolute number of people is still large enough to obtain useful aggregate statistics.

Here we consider two competing hypotheses: One that the Ebola outbreak was similar to Influenza, in that information generated on the Internet, including queries to search engines, postings on Twitter, and media articles, could have been used to alert and track the spread of Ebola. We contrast this with an alternative explanation, which is that Ebola on the Internet was, by and large, a news event. This explanation, if true, means that Internet data did not provide any information over that already known to news organization, and by extention, health authorities.

To answer this question we examine three sources of Internet data: Queries to the Bing search engine, posts on the Twitter social network, and news articles published online. Importantly, these data represent users from over 200 countries, including the countries most affected by Ebola. Our findings indicate that even in the latter countries, user queries and Twitter posts are more strongly correlated with news media attention than with the dynamics of the disease, suggesting that Ebola may not be amenable to syndromic surveillance using Internet data.

2. DATA

We used three sources of Internet data, and one source for ground-truth information on the spread of Ebola during the studied outbreak. In the following we describe each of these data.

2.1 Search log data

We extracted all queries made to the Bing search engine between November 1st, 2013 and October 31st, 2014 which contained the word "Ebola". For each query we extracted the query text, the list of pages which were presented to the user and the ones on which she chose to click on, an anonymous user identifier, and the country from which the user made the search (based on their IP address). This resulted in approximately 25 million queries from 240 countries.

We also extracted all queries submitted from the three most affected countries (Liberia, Guinea, and Sierra Leone) from November 1st, 2013 to March 31st, 2014, and analyze them for mentions of symptoms of Ebola, as in the Results.

2.2 News data

We extracted all news articles indexed by Bing News (`news.bing.com`) which used Latin characters and were published between November 1st, 2013 and October 31st, 2014 which contained the word "Ebola" in their title. This resulted in 3,411,170 articles from 145 countries.

[2]`http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx`

Since not all news sources from the three main infected countries are indexed by Bing News, we found additional news websites from these countries listed on Wikipedia and on the Stanford University Library "Africa south of of the Sahara, selected Internet resources" [3] pages. We then found pages from these websites displayed in response to Ebola queries on Bing, and listed their publication date as the first day they appeared in search results. This added 3,517 news items to our dataset, mostly from Sierra Leone and Liberia.

2.3 Twitter data

We extracted all messages on Twitter between November 1st, 2013 and October 31st, 2014 which contained the word "Ebola" and had geographic coordinates. This resulted in 480,090 messages from 226 countries. Additionally, we extracted all 23,720 tweets with geographic coordinates within Guinea for the period between November 1st, 2013 and March 22nd, 2014.

2.4 Health authority data

Our ground truth data on the number of Ebola cases per week in each country were obtained from the World Health Organization's Ebola data and statistics page[4] Situation Summaries. These data comprise of the number of confirmed Ebola cases at the end of each week during the outbreak.

3. RESULTS

3.1 Early detection

The first news item mentioning Ebola from the infected countries is from March 20th, two days before the official announcement about the epidemic (March 22nd).

The first Tweet from the three infected countries mentioning Ebola was recorded in December 26th, 2013. However, it is difficult to associate this single tweet with the outbreak, since it may have been humerus ("This Ebola of a virus come bad pass HIV...May God help us") and was written by a user who mostly tweets about football.

The next tweet mentioning Ebola is from March 25th, and is probably about the epidemic, stating that "Guinea has banned the sale and consumption of bats to prevent the spread of the deadly Ebola virus". However, the date of this tweet is also when the Guinean government reported an outbreak of Ebola to WHO.

We searched the tweets geo-located in Guinea from the beginning of November 2013 to the end of March 2014, when Ebola was officially announced, for mentions of the symptoms of Ebola. These included vomit, diarrhea, rash, headache, bleeding, fever, sore throat, and muscle pain. Only a small number of tweets included any of these words (in English or French), and none could be directly attributed to Ebola.

We conducted a similar search for symptoms of Ebola in the query data from Guinea, Liberia, and Sierra Leone from November 2013 to the end of March 2014. Figure 1 shows the fraction of queries over time. This figure also shows three thresholds, computed according to the volume of query data

[3]`http://web.stanford.edu/dept/SUL/library/prod/depts/ssrg/africa/guide.html`

[4]`http://apps.who.int/gho/data/node.ebola-sitrep.ebola-summary?lang=en`

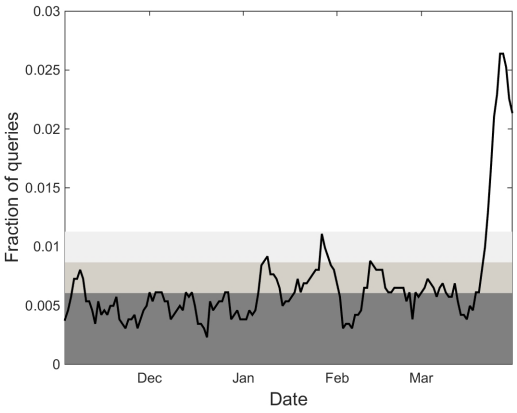

Figure 1: The fraction of Bing queries containing words related to symptoms of Ebola from Guinea, Liberia, and Sierra Leone. The grey thresholds are of the average plus one, three and five standard deviations of the baseline query rate, computed according to queries made in November 2013. All data is smoothed using a seven day window.

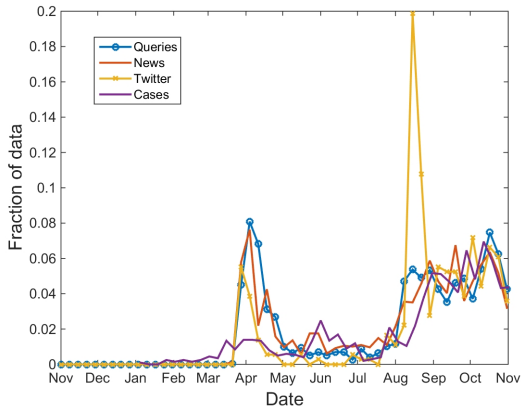

Figure 2: The fraction of Bing queries, Twitter messages, news articles, and confirmed Ebola cases for Guinea. Bing queries, Twitter messages and news articles are smoothed using a 7-day moving average.

from November 2013, before the Ebola epidemic began. The displayed thresholds are the average plus one, three and five standard deviations. As can be seen, queries on symptoms of Ebola begin to peak in late March 2014, after the official announcement on Ebola.

Interestingly, during January 2014, two peaks surpass the three standard deviation threshold. These first of these peaks contained queries which mentioned skin rash, and the second queries about typhoid fever. It is difficult to ascertain if these queries are related to Ebola, though we cannot rule out that they are.

Thus, it is unlikely that any of three data sources would have provided an alert on Ebola more than a week before WHO officially announced the existence of the epidemic. It is however possible that query log data would have alerted to unexpectedly high levels of Ebola-related symptoms, though these could have been interpreted as other diseases. This is because symptoms of Ebola are typical to many other diseases, including, for example, typhoid fever, typhus, Marburg fever, and Lassa fever.

3.2 Disease progression tracking

The fraction of Bing queries, Twitter messages, news articles, and confirmed Ebola cases for Guinea are shown in Figure 2. As the figure shows, Guinea experienced a significant number of cases from early February, but, as noted above, Internet data was only noticeable in any volume from mid-March. The second peak in case numbers, during June, is noticeably absent in all three data sources. Moreover, all three Internet sources are highly correlated among each other, and only to a lesser extent with the number of Ebola cases.

The Spearman correlation values between the four datasets are shown in Table 3.2. The correlation between all three Internet data sources and the actual number of cases is always lower than that found between Internet sources themselves. This difference in correlations suggests that the mechanism driving both search engine queries and Twitter mentions is the number of news articles, not the number of Ebola cases.

We examined whether any particular query text or individual words within queries were correlated well with the number of disease cases, since queries and query words are frequently used as the building blocks for predictive models in influenza. The correlation between the number of Ebola cases per week and the most correlated individual queries is 0.65. The same correlation with the best individual words is 0.79. Both are significantly lower than that obtained for the total number of queries. Figure 3 shows the fraction of queries per week for the two words most correlated with the number of Ebola cases in Guinea. As the figure demonstrates, the correlation during the first phase of the epidemic (March-May) is markedly lower than during the third phase (July-November). This suggests that concern about what was initially a relatively unknown disease was driving interest, rather than the actual number of people who contracted the disease. The correlation between the number of Ebola cases and the best individual word on Twitter is 0.84 (for the word "Ebola" in tweets from Liberia), but again the correlation during the first wave of the epidemic (when very few cases were recorded in Liberia) is very low.

3.3 Worldwide interest

The percentage of news articles related to Ebola published in a country (of the total news volume) is strongly correlated with its distance from the three main infected countries: Figure 4 show the correspondence between the percentage of news articles related to Ebola in a given country during the data period and the distance of that country to one of the three major infected countries. Among the 75 countries which use Latin characters in their main official language and are represented in the data, a power-law model to predict the fraction of Ebola-related news articles given the above-mentioned distance reaches a correlation of $R^2 = 0.27$ ($\alpha = -1.217$, $p < 0.0001$). Such a relationship was previously observed in the media attention given to news events [11]. This lends additional credence to the idea that Ebola was viewed mainly as a news event.

Guinea

	News articles	Twitter	Ebola cases
Bing queries	0.911	0.859	0.792
News articles		0.815	0.803
Twitter			0.735

Liberia

	News articles	Twitter	Ebola cases
Bing queries	0.896	0.917	0.884
News articles		0.908	0.878
Twitter			0.862

Sierra Leone

	News articles	Twitter	Ebola cases
Bing queries	0.956	0.930	0.851
News articles		0.936	0.896
Twitter			0.911

Table 1: Correlation (at weekly resolution) between the number of Bing queries, news articles, Twitter activity and Ebola cases in Guinea, Liberia, and Sierra Leone. The correlation between all three Internet data sources and the actual number of cases is always lower than that found between Internet sources themselves.

3.4 Information need regarding Ebola

We analyzed the websites that users clicked on following their search for Ebola. We categorized the 20 most common websites for each of the 3 infected countries, as well as for the entire dataset, into three categories: Health, News, or others. Since Wikipedia was the single most accessed website in each of the 3 infected countries, we analyzed it as a fourth, separate, category.

We calculated the correlation in the number of page clicks per week between pairs of categories. In the three infected countries this correlation was, on average, $R = 0.86$ ($p < 10^{-4}$, range: $0.64 - 0.96$). The correlation for the entire dataset was $R = 0.94$ ($p < 10^{-4}$). Therefore, the ratio of clicks per category was relatively constant over the data period.

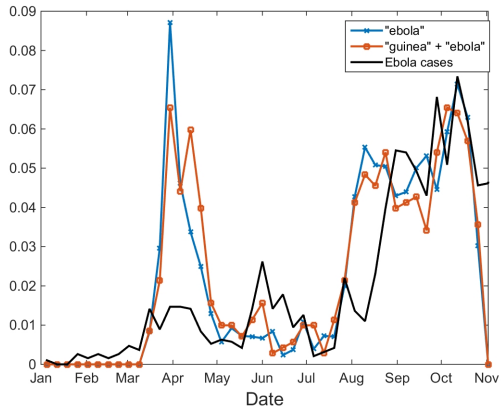

Figure 3: Fraction of searches with words most correlated with the number of Ebola cases in Guinea.

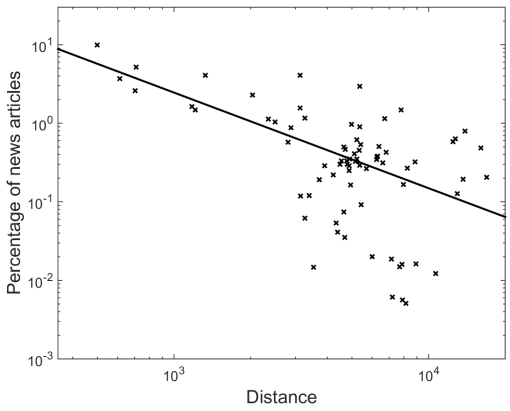

Figure 4: The percentage of news articles in a country as a function of its distance from the 3 major infected countries. Each dot represents on of 75 countries which use Latin characters in their main official language and are represented in the data. The line shows an exponential fit with a slope of $(distance)^{-1.217}$, which has an $R^2 = 0.27$.

Figure 5 shows the ratio of clicks (during the entire data period) among the different website categories in the various countries. First, it is evident that the ratio is approximately the same, indicating a similar use for the information across countries. Second, as noted above, Wikipedia is the single website with most clicks, but the largest category of clicks is to news sources. This again hints that most people, including in the most affected countries, viewed the Ebola outbreak as a news event rather than as an event of personal medical significance, that is, an event where users seem information on how to behave when infected with Ebola, because they themselves (or someone close to them) is infected with the virus.

4. DISCUSSION

User Generated Content (UGC) on the Internet, be it intentionally generated (e.g., posts on social media) or unintentionally created (e.g., queries to search engines) has been hailed as a potential source for syndromic surveillance. Evidence for the usefulness of these data came predominantly from seasonal diseases such as Influenza and Dengue Fever. The ongoing Ebola epidemic in West Africa had provided an important test case for the use of UGC to detect and track another infectious disease.

As our data shows, UGC closely tracks media attention to Ebola, not the dynamics of the disease, as apparent by the lower correlation of UGC to case numbers, compared to the number of news reports. This suggests that UGC is not advantageous as a source for syndromic surveillance in the case of Ebola.

There are several differences between the current Ebola epidemic and seasonal Infleunza and Dengue epidemics, and their effect on syndromic surveillance is difficult to untangle. These differences elude to both the nature of the disease and the area most affected by it. Ebola differs from Influenza and Dengue in its lethality and its transmission parameters. The mortality rate of Ebola is estimated at around 70% for

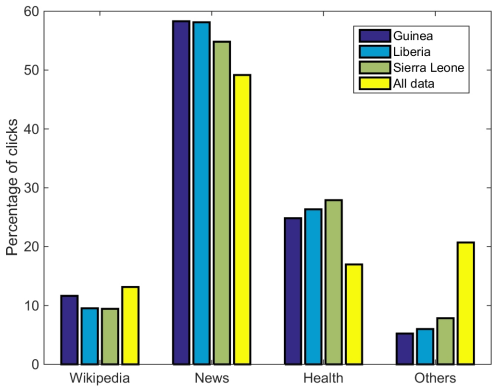

Figure 5: The percentage of clicks from different website categories in the three infected countries, as well as the entire data.

the current outbreak, compared to less than 1% in typical Influenza outbreaks. Ebola is much more infectious, with an R0 of 2.2, compared to 1.3 for Influenza [10].

As noted above, Internet penetration in the three most affected countries is relatively low: only 1.6% of people in Guinea, 16.5% in Liberia, and 1.7% in Sierra Leone have access to the Internet. This limits the volume of UGC created and the coverage of UGC for the purpose of syndromic surveillance, and consequently restricts the ability to perform early detection of the disease, as well as to track its progression. However, the percentage of cases in the population (around 0.2% in Liberia and Sierra Leone, and 0.03% in Guinea) are not dramatically different than the percentage of influenza cases reaching medical care (estimated at 0.5% [4]). In the latter, good correlation between Internet data and case numbers was found in multiple studies, as noted in the Introduction. Therefore, Internet penetration cannot be the only factor influencing the low correlation between UGC and case numbers.

The mortality rate associated with Ebola means that disease cases are more likely to be reported to medical authorities, compared to Influenza. In the latter, only a small minority of people with Influenza visit a healthcare provider [4]. Similar effects have been observed with regards to adverse drug reactions, where acute reactions are more likely to be reported to health authorities, and less severe adverse effects more likely to be mentioned in web search [12]. The high mortality rate also probably drives media interest, which is drawn to rare, dramatic events [13]. This attention, in turn, correlates with elevated interest in the epidemic by members of the public [1], as observed in recent influenza epidemics.

The implications of our findings are that a deeper understanding of the role of Internet data for syndromic surveillance is needed. It is important to characterize which diseases are amenable to syndromic surveillance via Internet data, and (perhaps more importantly) which are not. Future work will focus on characterizing diseases according to parameters outlined above, to provide an apriory assessment of the likelihood for successful syndromic surveillance for each disease.

5. REFERENCES

[1] D. Butler. When google got flu wrong. *Nature*, 494(7436):155, 2013.

[2] A. Culotta. 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, New York, USA, 2010. ACM Press.

[3] G. Eysenbach. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *AMIA 2006 Symposium Proceedings*, pages 244–248, 2006.

[4] A. C. Hayward, E. B. Fragaszy, A. Bermingham, L. Wang, A. Copas, W. J. Edmunds, N. Ferguson, N. Goonetilleke, G. Harvey, J. Kovar, et al. Comparative community burden and severity of seasonal and pandemic influenza: results of the flu watch cohort study. *The Lancet Respiratory Medicine*, 2(6):445–454, 2014.

[5] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, D. Weiss, et al. Syndromic surveillance in public health practice, new york city. *Emerg Infect Dis*, 10(5):858–864, 2004.

[6] K. J. Henning. What is syndromic surveillance? *Morbidity and Mortality Weekly Report*, pages 7–11, 2004.

[7] A. Hulth, G. Rydevik, and A. Linde. Web queries as a source for syndromic surveillance. *PloS one*, 4(2):e4378, 2009.

[8] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.

[9] W. E. R. Team. Ebola virus disease in west africaâĂŤthe first 9 months of the epidemic and forward projections. *N Engl J Med*, 371(16):1481–95, 2014.

[10] The Henry Kaiser Family Foundation. Ebola characteristics and comparisons to other infectious diseases. http://files.kff.org/attachment/ebola-characteristics-and-comparisons-to-other-infectious-diseases-fact-sheet.

[11] E. Yom-Tov and F. Diaz. The effect of social and physical detachment on information need. *ACM Transactions on Information Systems (TOIS)*, 31(1):4, 2013.

[12] E. Yom-Tov and E. Gabrilovich. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research*, 15(6), 2013.

[13] M. E. Young, G. R. Norman, and K. R. Humphreys. Medicine in the popular press: the influence of the media on perceptions of disease. *PLoS One*, 3(10):e3552, 2008.