JMLR: Workshop and Conference Proceedings 53:1–17, 2016 BIGMINE 2016

Disease Propagation in Social Networks: A Novel Study of Infection Genesis and Spread on Twitter

Manan Shah MANAN.SHAH.777@GMAIL.COM
The Harker School, San Jose, CA

Abstract

The CDC (Centers for Disease Control and Prevention) currently diagnoses millions of cases of infectious diseases annually, generating population disease distributions that, while accurate, are far too delayed for real-time monitoring. The ability to instantly compile and monitor such distributions is critical in identifying outbreaks and facilitating real-time communication between health authorities and health-care providers. This task, however, is made challenging due to the lack of instantly available public health information, creating a need for the analysis of disease spread on frequently updated social media websites. We introduce a novel pipeline based model to generate a real-time, accurate depiction of infectious disease propagation using Twitter data. Our approach, an amalgam of natural language processing and supervised machine learning, is invariant to mass media hype and significantly reduces the noise introduced by the use of tweets. The correlation coefficient between the Twitter disease distribution obtained via our approach and CDC data from mid-2013 to mid-2014 was 0.983, improving upon the best model published for the 2012-13 flu season. Our model further correlates well with theoretical models of infection spread across airport networks, verifying its robustness and applicability in the public sphere.

Keywords: infection spread; natural language processing; machine learning; disease propagation; diffusion models; data mining; big data

1. Introduction

The widespread adoption of social media as a tool for daily communication has opened the door for novel developments in big data computational epidemiology. With an estimated 113 million people in the United States alone using the Internet to access health-related information, the relationship between search activity and underlying disease trends remains confounded without adequate contextual information (Bodnar and Salathé, 2013). Research in the amalgamation of data science and disease spread has primarily been conducted in the realms of social networks such as Twitter, Facebook, and Tumblr.

Twitter is of particular interest due to its widespread use as a microblog and as a tool for mobile communication. Although recent studies have observed that a substantial portion of the "Twitter stream" consists of simple discussions and high levels of noise, Twitter users often provide relevant information regarding human behavior (Analytics, 2009). Due to the 140 character limit enforced upon each tweet, most information is sent from handheld devices on location, conveying a sense of urgency (Signorini et al., 2011).

Prior studies have utilized Twitter data to analyze textual sentiment, public anxiety regarding stock market prices, and opinions of restaurants and movies (Pak and Paroubek, 2010; Basari et al., 2013; Bollen et al., 2011). However, few investigations have been con-

ducted in the identification of disease propagation within such social networks. To date, proposed methodologies have either presented a keyword-based Tweet distribution to approximate CDC curves or formulated a regression problem, employing supervised machine learning techniques to model disease spread over time. Prior approaches, however, fail to adequately eliminate irrelevant tweets, posing significant issues to learning-based predictors that subsequently train using irrelevant data. Such algorithms are further prone to news and media hype regarding rare diseases such as Ebola and Zika, presenting severe problems to distributions that aim to characterize influenza-like illnesses (ILI). Finally, many prior methods are unable to plot real-time ILI distributions, rendering them unable to provide early-warning benefits for health care providers.

In this work, we attempt to holistically characterize disease spread using Twitter, with the aim of ascertaining the efficacy of the social media platform in modeling infectious illness frequency. Our method is distinguished from prior approaches in its multi-step classification procedure, whereby tweets are categorized into distinct subsets from which only relevant tweets are considered. We further develop random forest and support vector machine classifiers to cull spam and identify tweets regarding infectious diseases, generating a real-time ILI distribution exclusively from Twitter data. We evaluate the effectiveness of our model by comparing our Twitter-generated disease distribution with both the CDC ILI curve and SEIR (*susceptible, exposed, infected, recovered*) disease spread simulation distribution (Yang et al., 2011).

Overview of results. Our model performed exceptionally well, achieving a Pearson's correlation coefficient of 0.983 with the CDC ILI distribution for the 2013-14 flu season. Our model additionally reported a correlation coefficient of 0.947 with the theoretical SEIR infection spread model, validating its holistic structure. Our approach can be readily deployed to the public health and informatics sector, is the first to discard and manage noise prevalent on large scale social networks, and may provide a tool to epidemiologists for faster response to unknown infectious diseases.

In summary, the contributions of our work are the following:

- A novel infectious disease model premised on real-time Twitter data that incorporates a multi-step approach to identify "disease-linked" relevant tweets.
- A correlation with the CDC ILI distribution ($r = 0.983$) representing an improvement over current state-of-the-art Twitter-based methodologies across one year.
- Proof of robustness of our approach to external noise as signified by its correlation coefficient of 0.947 with mathematical disease simulations.
- Applications of our pipeline to international disease surveillance, including the recommendation of quarantine zones (an impossible task without global data).

We begin by detailing the CDC ILI distribution and prior approaches that aim to model the curve with social media data. We next discuss the intuition and methodology involved in our Twitter pipeline, delving into each stage in the multi-step process. We further characterize our SEIR infection spread simulation and depict its generated disease curves. The subsequent experiments section compares the Twitter-based distribution to the CDC ILI curve,

qualitatively and quantitatively analyzing each curve's characteristics. We conclude with a foray into the international applications of our pipeline and further avenues for research.

2. Background

In this section, we introduce the CDC ILI distribution and provide qualitative graphical analysis for sample curves from years 2003—2015 (Thompson et al., 2010). We further discuss earlier attempts of disease distribution modeling using social media and state their achieved correlations with the CDC distribution.

2.1. The CDC ILI Distribution

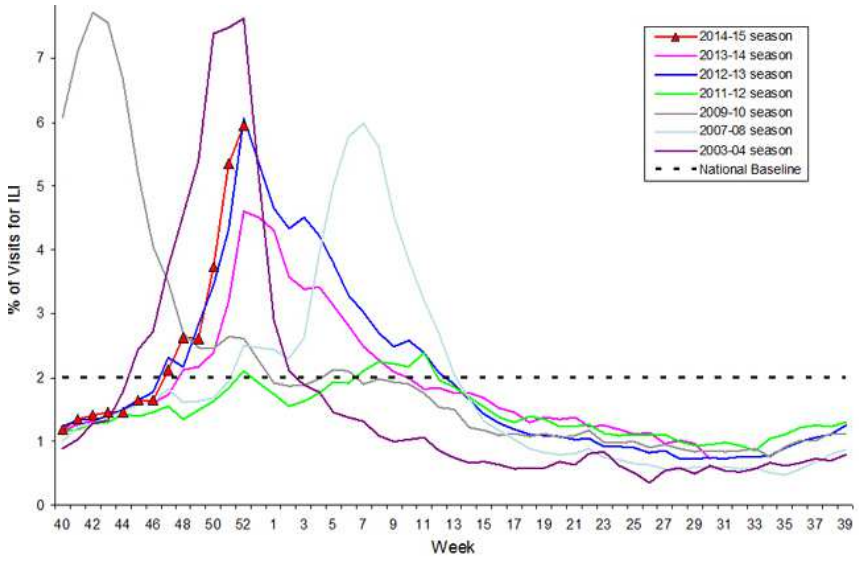

Figure 1: Percentage of visits for ILI as reported by ILINet (weekly national summary from 2003—mid-2015)

The ILI distribution (Figure 1) depicts the percentage of visits for influenza-like illness reported to the CDC by the US outpatient ILI surveillance network. Note the right-skewed nature of the curve, depicting the infection frequency increasing between months of November and January. An approximate three week delay is incurred in the generation of the disease distribution due to the time-consuming process of aggregating national patient reports. The methodology proposed in our work remedies this issue, using disease-related tweets to provide an accurate, real-time representation of the annual curve. Specifically, we test our Twitter model against the CDC ILI distribution for 2013—2014.

2.2. Prior Approaches

Prior work in the field of disease distribution modeling in social networks has been sparse and limited (Culotta, 2010; Paul and Dredze, 2011; Lampos and Cristianini, 2012; Signorini et al.,

2011; Sadilek et al., 2012; Lamb et al., 2013; Nagar et al., 2014). Bodnar and Salathé (2013) provide a comprehensive summary of these methods, using over 240 million tweets in their analysis. Their work concludes that the inclusion of "seemingly irrelevant" tweets in a support vector machine multivariable regressor yields correlations as high as 0.783, suggesting that methods reporting lower r-values have failed to properly learn information from tweets, potentially fitting the data due to other associated factors. The authors additionally develop a Twitter-based model for the 2012-13 flu season utilizing keyword-based tweet topic modeling, reporting a correlation coefficient of 0.877 with the ILI distribution.

While such approaches have detailed the benefits of Twitter-derived information in influenza forecasting, their proposed techniques fail to categorically eliminate tweets on premises other than hashtag analysis. With the expansion of big data and the ever-increasing flow of information from social network websites, it is crucial to be able to eliminate vast selections of irrelevant data, especially from a noise-riddled network such as Twitter, and successfully model the disease distribution with the resulting salient information. Our unique pipeline hopes to ameliorate this issue, developing a staged process towards identifying critical tweets and achieving a high level of noise invariance as a result. Following the paradigm that analysis is only as good as the data upon which it is based, our work both achieves higher correlation coefficients than those currently cited and better approaches the problem of ascertaining information from tweets, allowing for a reduced runtime with improved results.

3. Pipeline Description

In this section, we discuss and develop intuition for the multi-step pipeline based approach used to generate a real-time ILI distribution from input tweets.

3.1. Tweet Category Definitions

In order to develop a robust and viable model of the CDC ILI distribution, we differentiate between three unique categories of tweets: *self-reported, non self-reported,* and *spam*. Note that a tweet may only be placed in one of the three categories, and that each individual tweet must reside in a given category. We only consider self-reported tweets in our pipeline; in this process, we eliminate anomalies in our generated curve due to mass media coverage of rare diseases. We additionally distinguish individuals who have a disease from those who are worried about another's ailments, with the former affecting the resultant distribution.

- **Self-Reporting Tweets.** Self-reporting tweets are those that originate from either an infected individual or someone associated with an infected individual. Tweets in this category signify that the author is likely to have a direct influence on the ILI curve.

- **Non Self-Reporting Tweets.** Non self-reporting tweets encompass tweets posted by news networks and concerned citizens not immediately affected by a sickness. Tweets in this category, although they provide pertinent information regarding massive outbreaks, do not affect the ILI distribution. If included, such tweets would inflate

portions of the generated distribution due to media hype, resulting in an incorrectly augmented output.

- **Spam.** As in all social networks, spam messages drastically increase distribution noise and provide no saliency when generating the ILI distribution. In this work, we consider as spam all tweets that do not refer to disease.

3.2. Social Network Analysis Pipeline

Figure 2 details our model pipeline to its fullest extent, noting each relevant process. The pipeline accepts as input either a list of hashtags or auto-inferred terms from prior analysis (determined via linguistic term association). Our model leverages exhaustive uninformative tweet elimination to allow for the identification of anomalies and unique disease outbreaks, thus providing prognostic significance. The key steps involved are as follows:

1. **Hashtag Specification**: As our pipeline accepts keywords as input to search for relevant tweets, we initially obtain hashtags linked to specific diseases (such as *#influenza, #dengue, #zika*, etc.) by ascertaining the popularity of disease related hashtags currently in use.

2. **Linguistic Term Association**: We use linked n-grams in order to obtain additional hashtags and keywords aside from those directly linked to disease, such as *#sick* and *#nyquil*.

3. **Term Corpus Topic Modeling**: We assign numeric feature vectors to collected tweets utilizing TF-IDF (term frequency–inverse document frequency) vectorization within corpora of hashtags.

4. **Tweet Clustering**: Using the TF-IDF features ascertained in Step 3 and a mixed euclidean-cosine similarity measure, we cluster tweets according to minimal cluster RSS value via the centroid-based k-means approach.

5. **Salient Tweet Isolation**: We train and apply a linguistic attribute-based random forest classifier to randomly selected subsets of each cluster, rejecting an entire cluster if its chosen subset contains a sufficiently large number of *non self-reported* tweets.

6. **ILI Analog Frequency Distribution**: We subsequently plot the frequency distribution of relevant tweets over time in order to model the CDC ILI curve.

3.3. Hashtag Specification

The developed pipeline accepts as input a list of hashtags and keywords with which candidate tweets are obtained. Hence, it is imperative to determine which terms best characterize individual ailment or illness. We initially curated a list of relevant expressions of common infectious diseases (Hay et al., 2013). As social networks are not predisposed to informative discussions about specific illnesses, we ascertained the "popularity" of each disease keyword

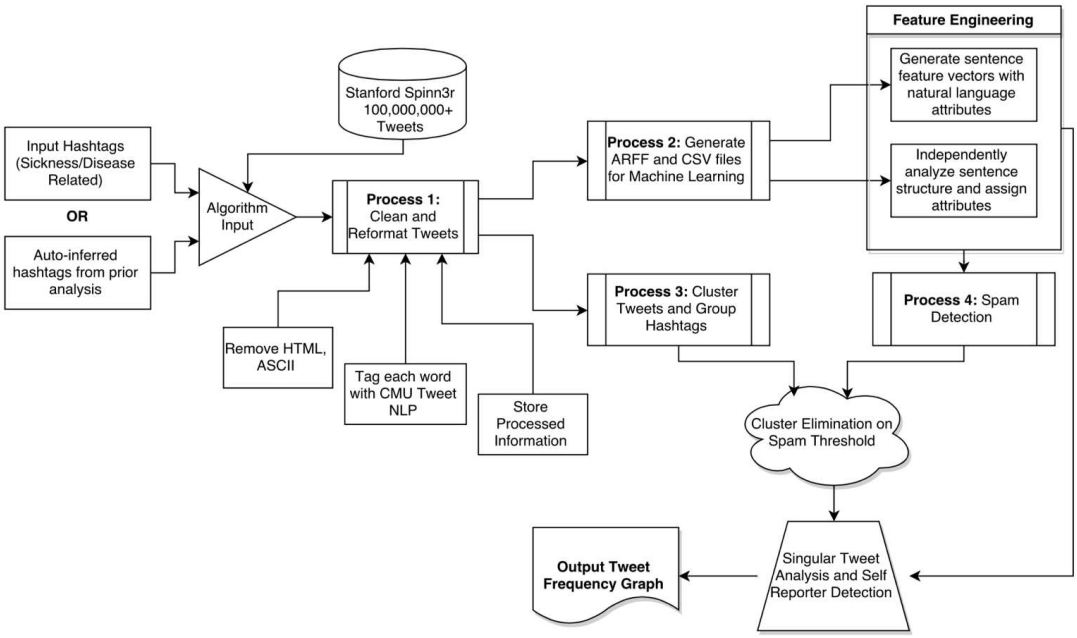

Figure 2: A comprehensive depiction of the model pipeline used to obtain disease distribution ILI curves from input tweets.

by analyzing recent tweet frequency and user variation. We define the popularity $\mathcal{P}$ of a search term S as

$$\mathcal{P}(S) = \text{unique}(U) \times \prod_{n=1}^{3} \left(\frac{1}{1 + \#(G_n) - \text{unique}(G_n)} \right)^n \tag{1}$$

where U is the set of users, G_n is the set of the top fifteen n-grams of the collected tweets, and unique(Q) represents the number of unique elements in set Q. Intuitively, $\mathcal{P}$ is directly proportional to the number of non-unique users and inversely proportional to the number of unique phrases used. We exponentially weight repeated higher-order n-grams as such occurrences are found with significantly diminished frequency and indicate repetition of similar messages among tweets. We selected 63 terms with the highest $\mathcal{P}$-metrics in a fixed period of time as salient for analysis; the remainder either consisted of excessively repetitive tweets or lacked enough unique users for ILI discrimination.

3.4. Linguistic Term Association

In order to expand our list of relevant keywords beyond disease names, we employed n-gram based linguistic analysis to identify additional terms that may be linked to infectious diseases. Specifically, we obtained the unigrams and bigrams that appeared with highest frequency among the 63 chosen hashtags, as denoted in Algorithm 1. Note that our approach maintains an algorithmic complexity of $\mathcal{O}(H^2)$, with H denoting the number of hashtags;

this computation is only required once to provide us with a sufficiently large list of terms to process. Interesting results obtained by use of this approach include the keywords *dayquil, nyquil, sleepy, drowsy*, and *upset*, all critical terms that may have been overlooked had we exclusively used disease names and common hashtags (such as *sick, headache, influenza*, etc.).

Algorithm 1: Identifying Associated Keywords

Input: A set of disease-related hashtags H and the level of k-grams to search
Output: A set of unique additional terms associated with H

$L \leftarrow [\,]$
$S \leftarrow$ stopwords
for i *in range* $(0,$ *length*$(H))$ **do**
  for j *in range* $(i+1,$ *length*$(H))$ **do**
    $L \leftarrow L + \text{topgrams}(S, H(i), H(j), k)$
  end
end
$L \leftarrow \text{unique}(L)$
Procedure `topgrams`(S,A,B,k)
  $U_A \leftarrow k\text{-grams}(A) \cup \bar{S}$
  $U_B \leftarrow k\text{-grams}(B) \cup \bar{S}$
  $U \leftarrow U_A \cap U_B$
  return U

3.5. Term Corpus Topic Modeling

Our hashtag and keyword determination methodologies seek out potential candidates for disease related tweets; we next consider approaches to eliminate irrelevant tweets as defined in Section 3.1. To this end, we categorize tweets using TF-IDF feature vectors (Ramos, 2003). A numerical statistic that aims to reflect the importance of a word in a text corpus, TF-IDF was used to weight tweets for k-means clustering. Specifically, we have

$$\text{tf}(t,d) = 1 + \log f_{t,d} \tag{2}$$

$$\text{idf}(t,D) = \log\left(1 + \frac{|D|}{n_t}\right) \tag{3}$$

$$\text{tfidf}(t,d,D) = \text{tf}(t,d) \times \text{idf}(t,D) \tag{4}$$

where t is a term in document d within corpus D. To be precise, $f_{t,d} = |t \in d|$ and $n_t = |\{d \in D : t \in d\}|$. Each tweet is denoted as a document d_j within its hashtag corpus D_i, and matrices of TF-IDF features across unigrams, bigrams, and trigrams are generated for each tweet to effectively characterize their respective corpora.

3.6. Tweet Clustering

With each tweet represented as a matrix of pertinent features, it is possible to cluster tweets by their pairwise similarity. In order to minimize the residual sum of squares metric in our

	Tweet Example
Self-reporting	ive never been more sick in my life than i am right now. Throat swollen, body aches, flu like symptoms and i cant sleep ?
Non self-reporting	#Flu Myth: flu vaccine gives the flu. NO! Dead virus is used. Flu mist has live but engineered to remove parts that cause sickness #smedtips
Spam	Try Swahili ones RT@FactHive "Sixth sich sheik's sixth sheep's sick" is the world's hardest tongue twister according to Guinness Wld Records

Table 1: Representation of automated tweet clustering on raw tweets

k-means clustering approach, we opted to use a mixed distance metric between each tweet and cluster centroids, defined as the difference between the cosine and euclidean distances respectively. We enforced a limit on the number of clusters proportional to number of processed tweets to ensure that the resulting distribution of tweets among clusters would remain dense. Table 1 depicts an example of such clustering where $k = 3$. For $k > 3$, each defined category is divided into multiple unique components, which are retained or eliminated by the same criteria.

3.7. Salient Tweet Isolation

3.7.1. Categorical Classification

In order to ascertain the salience of each cluster of tweets in modeling the ILI distribution, we trained supervised classifiers to distinguish between self-reported, non self-reported, and spam tweets. To perform this task, we characterized each tweet as a representative feature vector with twenty linguistic attributes. Sample features calculated included the number of conjunctions, average sentence length, and the number of emoticons in each tweet. Utilizing a manually annotated training set of 200 examples derived from various hashtags, we trained support vector machine and random forest classifiers to distinguish between the three defined classes. Our linguistic attribute-based machine learning model performed remarkably well, reporting a quadratic weighted kappa statistic of 0.872 and a classification accuracy of 87% when tested using 10-fold cross-validation. We optimized our random forest with 100 trees, each constructed considering five random features, and we incorporated a Gaussian radial basis function kernel for our support vector classifier. An analysis of each classifiers' weights yielded the point of view of a tweet (first, second, or third person), the number of slang words, and average word length as the most salient inter-class differentiators.

3.7.2. Cluster Elimination

Maintaining model scalability when working with large-scale datasets is imperative; for practical use, our pipeline must successfully compute disease distributions of a large (and constantly updating) database of input data. Such a task calls for a more efficient manner of irrelevant tweet elimination than classification of the relevance of each individual tweet; we therefore classify N random samples from each cluster selected with probability p. If the number of "bad" tweets (defined as the sum of non self-reporting and spam tweets) exceeds

a given threshold T, we discard the entire cluster for further analysis. We additionally incorporate a weighting function W to augment predictions based on each tweet's medical relevance. The introduction of such a bias increases the tweet's relevance probability by a constant factor if popular medical jargon (such as the terms *stomach, tummy,* and *belly*) are present in a tweet. Although our method is certainly prone to reject relevant tweets loosely associated with "bad" ones in their clusters, the sheer volume of data obtained via Twitter allows for the elimination of false positives with minimal accuracy loss. Algorithm 2 provides a high-level depiction of the steps taken in this process.

Algorithm 2: Tweet Cluster Elimination

Input: A list of clusters C, the selection probability p, the threshold for individual tweet retention T

Output: A set of salient clusters L culled from C; that is, $L \subseteq C$

$L \leftarrow [\,]$
for i *in range* $(0,$ *length*$(C))$ **do**
  $C_i \leftarrow C[i]$
  $N \leftarrow 0$
  $R \leftarrow \text{length}(C_i) \times p$
  for j *in range* $(0, R)$ **do**
    `// Note: A larger P indicates a greater spam likelihood`
    $P \leftarrow W(\text{predict}(\text{random}(C_i)))$
    if $P > T$ **then**
      $N \leftarrow N + 1$
    end
  end
  if $N < R \times 0.5$ **then**
    $L \leftarrow L + C_i$
  end
end

On average, our tweet clustering and elimination procedure discards 73% of clusters it encounters, with the remaining high-quality data included in our resultant disease distribution. The unique free parameters in our approach (the selection probability p and the tweet retention threshold T) were initially defined as 0.25 and 0.5 respectively. A larger p ought to be selected for faster cluster elimination (specifically, when generating distributions on larger datasets), and an increase in T penalizes lower quality tweets with greater severity, resulting in a more sparse (yet potentially more accurate) distribution. For our hashtag-reduced dataset of approximately $400,000$ tweets, the aforementioned fixed values yielded a dense and salient distribution, as desired.

Additionally, note that our procedure allows for a reduced complexity of $\mathcal{O}(|C||R|)$ as opposed to $\mathcal{O}(C^2)$, a significant improvement in the limit $R \ll C$ (assuming the prediction function for a given decision tree-based classifier is $\sim \mathcal{O}(1)$). We plot a frequency distribution of remaining tweets as a function of time, resembling a real-time ILI curve (similar to that of the CDC) with the additional benefit of potential outbreak and anomaly detection.

3.8. ILI Analog Frequency Distribution

In the final steps of our pipeline, we synthesize a plot detailing the distribution of disease-linked tweets as a function of time (analogous to individuals reporting infections to the CDC). Our frequency distribution aims to be robust to news hype, spam, and irrelevant information contained in Twitter noise. In order to better characterize the smooth CDC ILI curve, we condense our daily distribution into a weekly one. We represent the frequency of each week as the mean of the daily data, excluding the minimum and maximum values, and we additionally apply sliding mean data smoothing with a window of 5 (the length of each reduced week), such that each frequency value is the average of the corresponding subset of a larger set of data points.

4. SEIR Disease Simulation

In addition to the empirical national CDC ILI distribution, disease spread within populations may be numerically modeled via a system of differential equations. Although the primary goal of our pipeline is to approximate the ILI curve, a significant similarity between the shapes of the theoretically simulated and generated distributions will further validate our approach's robustness to Twitter noise and media hype. Such a comparison will additionally allow for an analysis of the distinctions between both curves, potentially providing salient information regarding variances between theoretical contact-based models and observed outcomes. With the goal of ascertaining whether our model derived from Twitter sufficiently represents the expected spread of infectious illness, we utilized an airport-based disease network, defining airports as nodes and connecting flights as edges.

4.1. Theoretical Primer

Infectious diseases may be modeled within populations by stratifying individuals into broad categories; the simplest simulation categorizes individuals into susceptible, infectious, and recovered groups. As most common infectious diseases are not fatal, we can write $S+I+R = N$, where N is the constant population (with the degenerate assumption of equal birth and death rates) (Miller and Volz, 2013). Once such a model is developed, infection parameters of disease extent, spread, and duration may be obtained. Common infections additionally include an incubatory period in which an infected individual is not contagious. Assuming the incubation period is a random variable with an exponential distribution, we have the following system of differential equations for susceptible (S), exposed (E), infectious (I), and recovered (R) individuals, with $\dot{N} = 0$ (Heesterbeek, 2000).

$$\frac{dS}{dt} = \mu N - \mu S - \beta \frac{I}{N} S \tag{5}$$

$$\frac{dE}{dt} = \beta \frac{I}{N} S - (\mu + a) E \tag{6}$$

$$\frac{dI}{dt} = aE - (\gamma + \mu) I \tag{7}$$

$$\frac{dR}{dt} = \gamma I - \mu R \tag{8}$$

Here, $\beta, 1/\gamma$, and μ are defined as the disease contact rate, the average infectious period, and the average death rate respectively, and the average incubation period is modeled with the hyperbolic distribution $1/a$. In order to numerically compute the infection distribution at arbitrary intervals, we may represent these differential equations as functions of time, replacing the differentials with discrete (yet small) time intervals Δt. We calculate the values of S, E, I, and R at each interval, thereby generating disease distributions for each identified subcategory.

4.2. Graph Dataset Description

The dataset we use to generate nodes and edges for our simulation is published on OpenFlights.org (Patokallio, 2014), with the model structure derived from Yager and Taylor (2014). The database contains 6,977 airports spanning the globe along with their locations, and includes 5,888 airlines. 59,036 routes between 3,209 airports on 531 airlines spanning the globe are recorded; graph nodes are selected airports, and edges are those routes interconnecting multiple airports. Only airports that have entering or exiting routes are considered; the resulting graph (Figure 3A) consists of one connected component depicting an international network of travelers.

4.3. Simulation Execution and Evaluation

We may visualize our multi-nodal network in Figure 3, with blue representing a normal state, red representing infectious or exposed transmission, and black representing recovery. Our simulation propagates disease starting randomly from ten airports (with a higher probability of inception in airports with more connecting flights). The airports themselves act as proxies for disease spread among individuals located at each airport, with the assumption that travelers are able to leave residual infection via permanent workers. Edges are weighted to represent the probability of infected individuals in transit according to the degree of the source and destination airports. We modeled our specific disease after influenza A, such that $\beta = 7$, $\gamma = 3$, and $\mu \approx 0$. The basic reproductive rate R_0 of our infection was therefore 2.33; in other words, approximately 2.33 secondary infections are expected from every unique primary case.

Figure 3A is a representation of the network at time $t = 28$ days, with less frequented airports beginning to develop infection and those with the longest exposure to disease sufficiently cured. As represented in the infection curve in Figure 3B, the disease spread is beginning to decline, with black areas indicating recovered nodes.

5. Experiments

5.1. Dataset Description

We tested our pipeline's efficacy in modeling the CDC ILI distribution using the Stanford Spinn3r dataset, a collection of over 100 million tweets from 2013—2014 from which we obtained disease-linked subsets for analysis. The dataset was obtained using a Gardenhose stream consisting of a 10% random sample of all public statuses. As detailed in Section 3.4, tweets were initially selected using both disease hashtags and illness-related terms, allowing for a more salient input to process.

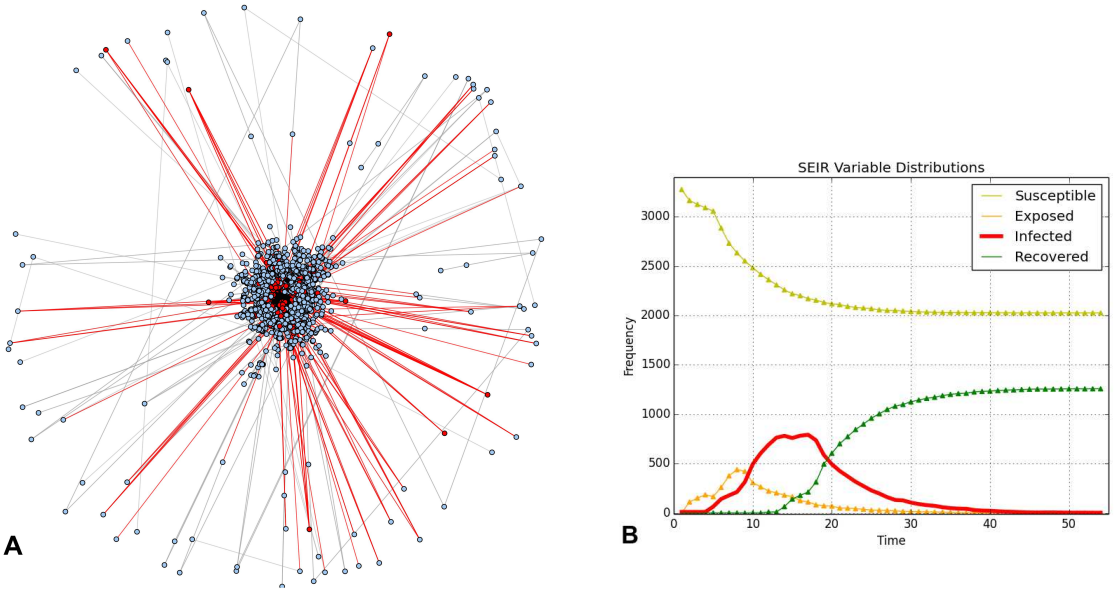

Figure 3: (**A**) A visual depiction of the structure of the disease network at time $t = 28$. (**B**) An illustration of disease propagation across the entire network for variables S, E, I, and R.

5.2. Comparative Distribution Analysis with CDC Data

Proceeding according to the methodology discussed in Section 4, we obtained a 42-week distribution from 2013 to 2014 that we compared to the analogous CDC distribution. The results of our analysis are depicted in Figure 4A; the impact of smoothing on distribution correlation is readily observed. Note that the frequency measure on the y-axis is not absolute; that is, the estimated ILI mean line was vertically shifted to provide a visual depiction of the similarity between the curves.

We additionally compared our estimated distribution with the infection propagation distribution generated via the SEIR model. Figure 4B displays all three distributions in tandem alongside a baseline distribution generated using tweets selected solely by hashtag criteria. Although the curves seem similar in shape and skewness, the simulation distribution is distinctively bimodal, while both the estimated and ground truth curves are unimodal. Furthermore, the simulation mean line predicts a more severe drop-off than the estimated or ILI curves, and flattens out towards the end of the season (as opposed to both other distributions, which seem to be slowly increasing, albeit non-monotonically).

5.3. Numerical Evaluation of Distribution Similarity

We evaluated the similarity between the determined distributions utilizing Pearson's correlation coefficient and the Kullback-Leibler divergence. The correlation coefficient r represents a "normalized" covariance between random variables X and Y, defined as the covariance of X and Y scaled by their respective standard deviations. The Kullback-Leibler (KL)

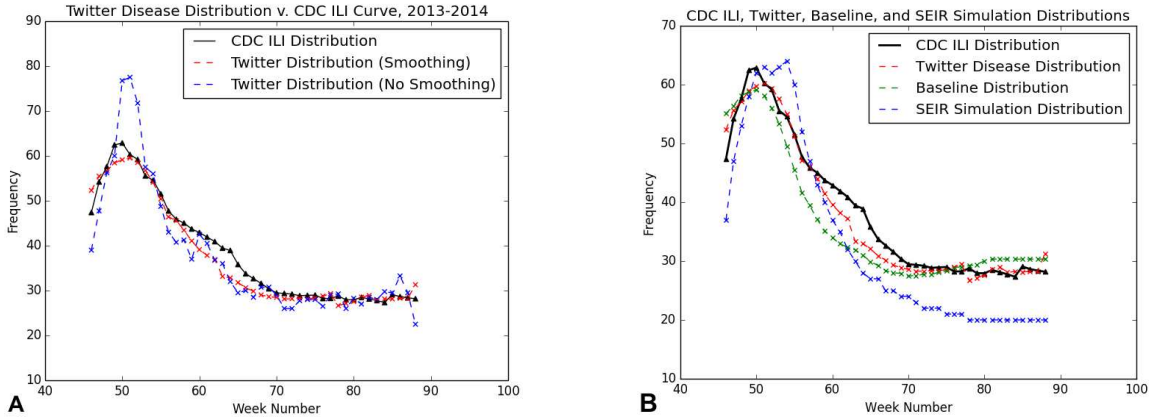

Figure 4: (**A**) A comparison of the Twitter-derived distribution with the CDC ILI curve. (**B**) A side-by-side plot of the Twitter, Baseline, SEIR, and CDC distributions.

	CDC	**Twitter**	**Simul**	**Base**	**CDC**	**Twitter**	**Simul**	**Base**
CDC ILI	–	0.983	0.931	0.938	–	0.003	0.014	0.005
Twitter	0.983	–	0.947	0.972	0.003	–	0.018	0.001
Simulation	0.931	0.947	–	0.898	0.014	0.017	–	0.025
Baseline	0.938	0.972	0.898	–	0.005	0.001	0.025	–

Table 2: Statistical measures of distribution similarity. The first three columns list correlation coefficient values; the next three list KL-divergence values.

divergence of Y from X, denoted $\mathcal{K}(X||Y)$, represents the amount of information lost when X is used to approximate Y. More precisely, the metric may be interpreted as the penalty incurred in encoding X using a Huffman code optimized for Y. It is important to note that $\mathcal{K}$ is non-symmetric and operates on normalized distributions.

Table 2 lists our obtained values representing the similarity between each proposed distribution. The "baseline" model represents results obtained solely utilizing the proposed hashtags and medical terms to cull tweets from our database; clustering and additional processing premised on tweet saliency are excluded. Note that the Pearson correlation coefficient between the official ILI distribution and our pipeline's result was 0.983, with a low KL divergence of 0.003 representing the robustness and accuracy of our method. The KL divergences and correlation coefficients of both Twitter-based approaches (the baseline model and the complete pipeline) outperform the SEIR simulation, successfully accounting for the tail end of the distribution. The clustering approach, as visualized in Figure 4B, better models the elongated infection decline over time than the generic hashtag approach, yielding superior correlation coefficients and divergence metrics with both the simulation and the CDC distribution. However, the KL divergence between the Twitter distributions is quite small, indicating that, when normalized, little variation is observed between the cluster-based and simple hashtag approach.

6. Worldwide Disease Modeling

Our network analysis pipeline allows us to address new types of problems, such as the identification of Twitter users infected by a certain illness. We applied our pipeline to the problem of modeling the global spread of disease using Twitter user relationships. In particular, we utilize Algorithm 2 and our hashtag list to generate a list of 10,000 disease-linked tweets, which we associate with their respective users. We utilize Microsoft Bing Maps' reverse-geocoding API to obtain an approximate location for each unique user, and subsequently obtain a random sample of each user's potentially infected followers premised upon the infection level of each followers' prior tweets. With such information, we generated a directed graph with countries as nodes and connections between individuals and followers as edges (Figure 5). As expected, the most prominent nodes (those with largest in and out degrees) represent Mexico, the United States, Spain, Italy, and Pakistan respectively; populous nations and popular tourist sites were frequently expressed.

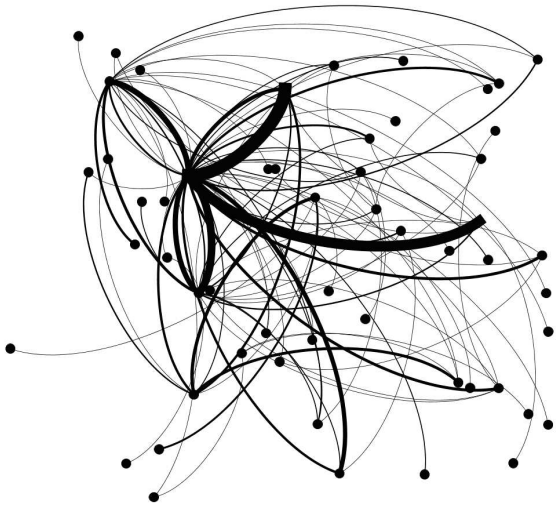

Figure 5: Depiction of the international disease relationship graph, with edges representing connections between infected individuals and infected followers.

We further ascertained the most prevalent relationships between countries, determining the listing of top connections depicted in Table 3. As our graph was directed, certain paths may appear twice; however, as edges were weighted in each direction, highest weighted paths are listed first. This preliminary analysis indicates potential quarantine locations if an infectious disease is found in a certain location. Our method enabled us to identify some surprising connections, such as a strong relationship between users in India and those in Mexico. For example, if India were to develop an epidemic, it may not be immediately intuitive to suggest a quarantine of Mexico; however, our results indicate a significant connection between the two countries. Such an anomaly may be explained by the rising popularity of Mexico as a tourist destination for Indian residents, with the number of Indian tourists expected to increase by over 200,000 from 2014 to 2020 (Bisaria, 2014). Our model may thus be used both as a national metric to ascertain outbreaks and anomalies in the

ILI curve and as a method for analyzing international disease connections, making it a prominent tool in the development of more effective responses.

Source Node	**Destination Node**
Mexico	Spain
Mexico	United States
South Africa	United States
United States	Mexico
Australia	United States
Mexico	Argentina
Canada	United States
Italy	United States
Mexico	India
India	United States

Table 3: Highest-ranked connections within the international disease network (Figure 5).

7. Conclusions

In this work, we improve upon current methodologies used for determining disease spread by developing a pipeline that can accurately plot disease distributions in real time. Our work is unique in its conflation of topic modeling, machine learning, and natural language processing to eliminate tweet noise and irrelevant information, allowing for a robust characterization of the CDC ILI distribution. Our model can further scale to massive datasets, and is robust to news and media hype regarding rare (but not infectious) diseases. We verified our model by determining its correlation with both the 2013-14 CDC ILI distribution and an SEIR disease simulation, obtaining correlation coefficients of 0.983 and 0.947 respectively. To our knowledge, our model is the first in the field to achieve such high correlation coefficient values when compared to the CDC distribution over an entire flu season.

We further demonstrate the real-world applicability of our model in ascertaining important quarantine locations premised on connections between infected Twitter users and their followers. Our model thus provides a real-time disease distribution tracker with the ability to identify infectious outbreaks and facilitates international disease spread analysis at an unprecedented level.

Future Work. We hope to leverage our pipeline-based methodology in areas of spatial disease location, cascade prediction, and international modeling. As the ability to pinpoint the regional spread of certain diseases is crucial for local outbreak analysis and identification of the propagation point of disease, we initially plan on using our disease-linked tweet dataset to ascertain how diseases are distributed amongst populations of Twitter users. We additionally hope to develop a framework for cascade prediction within the Twitter disease network sub-space in order to identify how long a certain disease will last and the rate of its progression. Furthermore, as motivated in the discussion regarding Figure 5 and Table 3, we have provided a proof-of-concept use of our model in determining potential quarantine sites and international disease networks; we are excited to further investigate these avenues in future work.

Acknowledgments

The author would like to thank Rok Sosic for mentorship and guidance, Steve Eglash and Andrej Krevl for database support, Jure Leskovec for his feedback on this research, the SNAP group at Stanford University for support and access to historic Twitter data, and Eric Nelson for his feedback. Selections of source code for this work are published at `https://github.com/mananshah99/diseasenetworks`.

References

Pear Analytics. Twitter study–august 2009. *San Antonio, TX: Pear Analytics*, 2009.

Abd Samad Hasan Basari, Burairah Hussin, I Gede Pramudya Ananta, and Junta Zeniarja. Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53:453–462, 2013.

Sanjiv Bisaria. Mexico soon to become a popular tourist destination for indian travellers, 2014.

Todd Bodnar and Marcel Salathé. Validating models for disease detection using twitter. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 699–702. International World Wide Web Conferences Steering Committee, 2013.

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.

Simon I Hay, Katherine E Battle, David M Pigott, David L Smith, Catherine L Moyes, Samir Bhatt, John S Brownstein, Nigel Collier, Monica F Myers, Dylan B George, et al. Global mapping of infectious disease. *Phil. Trans. R. Soc. B*, 368(1614):20120250, 2013.

JAP Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons, 2000.

Alex Lamb, Michael J Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795, 2013.

Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4): 72, 2012.

Joel C Miller and Erik M Volz. Incorporating disease and population structure into models of sir disease in contact networks. *PloS One*, 8(8):e69162, 2013.

Ruchit Nagar, Qingyu Yuan, Clark C Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and John S Brownstein. 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, 2014.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

Jani Patokallio. Airport, airline, and route data, 2014. URL http://openflights.org/data.html.

Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272, 2011.

Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.

Adam Sadilek, Henry A Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *ICWSM*, 2012.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

MG Thompson, DK Shay, H Zhou, CB Bridges, PY Cheng, E Burns, JS Bresee, NJ Cox, et al. Estimates of deaths associated with seasonal influenza-united states, 1976-2007. *Morbidity and Mortality Weekly Report*, 59(33):1057–1062, 2010.

Nicholas A. Yager and Matthew Taylor. Edge-based control of disease propagation through the world-wide airport network. https://github.com/nicholasyager/airport-disease-modeling, 2014.

Yong Yang, Peter M Atkinson, and Dick Ettema. Analysis of cdc social control measures using an agent-based simulation of an influenza epidemic in a city. *BMC infectious diseases*, 11(1):1, 2011.