SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources

Theodoros Rekatsinas[*], Saurav Ghosh[†], Sumiko R. Mekaru[‡], Elaine O. Nsoesie[‡]
John S. Brownstein[‡], Lise Getoor[§], Naren Ramakrishnan[†]

Abstract

Rapidly increasing volumes of news feeds from diverse data sources, such as online newspapers, Twitter and online blogs are proving to be extremely valuable resources in helping anticipate, detect, and forecast outbreaks of rare diseases. This paper presents SourceSeer, a novel algorithmic framework that combines spatio-temporal topic models with source-based anomaly detection techniques to effectively forecast the emergence and progression of infectious rare diseases. SourceSeer is capable of discovering the location focus of each source allowing sources to be used as experts with varying degrees of authoritativeness. To fuse the individual source predictions into a final outbreak prediction we employ a multiplicative weights algorithm taking into account the accuracy of each source. We evaluate the performance of SourceSeer using incidence data for hantavirus syndromes in multiple countries of Latin America provided by HealthMap over a timespan of fifteen months. We demonstrate that SourceSeer makes predictions of increased accuracy compared to several baselines and is capable of forecasting disease outbreaks in a timely manner even when no outbreaks were previously reported.

1 Introduction

There has been a growing interest in developing statistical models for detecting infectious disease outbreaks to enable effective control measures to be taken in a sufficiently timely fashion. Most early approaches relied on highly specialized data, including medical records or environmental time series [28, 29]. Recently, however, there has been a growing interest in monitoring disease outbreaks using publicly available data on the Web, including news articles [5, 13], blogs [6], search engine logs [9] and micro-blogging services, such as Twitter [7, 16, 17]. Due to their volume, ease of availability, and citizen participation, such *open source indicators* have been shown to be effective at monitoring disease emergence and progression. Most prior work focuses on detecting outbreaks of common diseases, such as influenza, by discovering temporal patterns over predefined groups of keywords. However, many infectious diseases are *rare* with only a few incidences being reported in open sources. Forecasting outbreaks of rare diseases raises several challenges. We use a real-world scenario to illustrate these challenges.

1.1 Challenges

We focus on Hantavirus, a rare infectious disease. We examine incidences in Latin America analyzing a corpus of public news articles from 798 different sources (source here refers to the publisher of the article) referring to multiple diseases over a timespan of 15 months.

The first challenge is that keyword based techniques have significant limitations at forecasting outbreaks of rare diseases, such as hantavirus. As incidences are rare, disease related keywords may be scarce over time or totally unavailable in the available past data. Therefore, it is difficult for keyword-based techniques to identify temporal patterns and predict new outbreaks in a timely manner. Next, we provide evidence on why keyword based techniques can be ineffective and present a detailed evaluation in Section 5.

EXAMPLE 1. *We focus on Chile, Argentina, Brazil and Uruguay. No incidences were reported in other countries. We compare the number of mentions over time for the set of hantavirus specific keywords {"hanta", "hantavirus", "roedores", "ratones", "cardiopulmonar"} and the actual timeline of hantavirus incidences for each country. The actual hantavirus incidences were extracted by a third-party gold standard (Section 4). Figure 1(a) shows the timeline of hantavirus incidences in the four countries, while Figure 1(b) and Figure 1(c) show the timeline of word mentions for the aforementioned keyword set. There are cases where despite having an increased number of hantavirus incidences the number of keyword mentions is low. Also, the two timelines are not aligned with spikes in the keyword timeline appearing with a delay after spikes in the actual incidences timeline.*

The second challenge is that different data sources may exhibit different delays at reporting rare disease incidences, and using their data for predicting outbreaks may lead to predictions of significantly different accuracies.

[*]Dept. of Computer Science, University of Maryland, College Park
[†]Dept. of Computer Science, Virginia Tech
[‡]Children's Hospital Informatics Program, Boston Children's Hospital
[§]Dept. of Computer Science, University of California, Santa Cruz

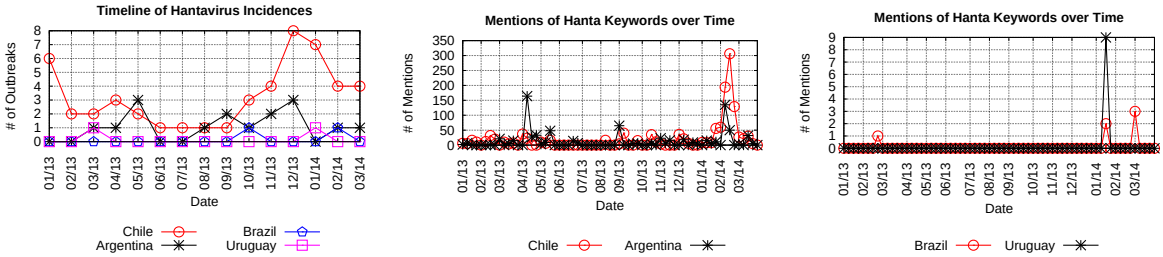


Figure 1: Timeline of hantavirus outbreaks from January 2013 to March 2014 for Chile, Argentina, Brazil and Uruguay.

EXAMPLE 2. *For the previous scenario, we consider using each data source in isolation for predicting hantavirus outbreaks in Chile, Argentina, Brazil and Uruguay. Figure 2 shows the source accuracy histograms for Chile and Brazil. As shown, the accuracy levels of different data sources vary significantly. Similar results were observed for Argentina and Uruguay but omitted due to space limitations. The model used for predicting outbreaks is described in Section 4.*

1.2 Contributions Motivated by these examples we study the problem of forecasting disease rare outbreaks across different locations by analyzing a dynamic corpus of publicly available news articles updated at fixed intervals. We introduce SourceSeer, a novel rare disease outbreak forecast framework that consists of two major components: (a) analysis of past data to detect disease spatio-temporal patterns and (b) prediction of future outbreaks.

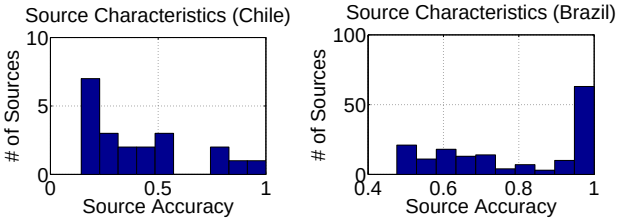


Figure 2: Source accuracy histograms for Chile and Brazil.

Since analyzing keyword mentions over time is not sufficient to discover the temporal patterns rare disease outbreaks may exhibit, we use topic models to discover the word co-occurrence patterns in the available news articles in an automated fashion. We model data sources as *evolving documents* over time, and introduce a new spatio-temporal topic model (Section 3) that explicitly models time and location jointly with word mentions. To exploit the fact that different sources exhibit different levels of accuracy when predicting rare disease outbreaks, we combine the proposed topic models with source-based anomaly detection techniques considering each source as an individual expert and fuse the individual source predictions in a single final prediction. We use anomaly detection techniques since for many locations no outbreaks may be reported in the available news articles, and hence, detecting unknown patterns is crucial.

The specific contributions of our approach are:

- *Effectiveness:* SourceSeer operates on large collections of news articles and can clearly rare disease topics and their corresponding spatio-temporal patterns.
- *Diversity:* Our model enables rare-disease forecasting for a diverse set of locations with significantly different outbreak patterns under a unified scheme.
- *Accuracy:* As we illustrate in an extensive experimental evaluation, considering the spatial focus and accuracy of each data source offers improved accuracy in forecasting disease outbreaks as opposed to analyzing the input of all sources for a specific location in a collective manner.
- *Forecasting instead of detecting:* SourceSeer is able to forecast outbreaks several days before they occur with a significant *lead-time* over reporting in news media.

2 Forecasting Disease Outbreaks With Many Sources

We assume a continuously updated collection of time-stamped event articles from a collection of data source S, referring to a set of locations L, and containing words from a vocabulary V. We consider a discretization of time and assume that new data entries are added in batches over intervals of fixed time. For example, these time intervals may correspond to a specific day or week. For the remainder of the paper we consider a time granularity of one week, defined as the 7-day period from Sunday to Saturday referred to as an *epidemiological week*, or *epi-week* for short.

We assume an input over a fixed discretized time window up to time point T including data entries associated with a single time point in $\{1, \ldots, T\}$. It is convenient to convert this input to a collection of tuples of the form *(source, location, word, time point; count)* where the count corresponds to the total number a specific word was mentioned in all articles associated with the source, location and time point in the tuple. For example, a tuple ("www.biobiochile.cl", ("Los Lagos", "Chile"), "hanta", "28"; 35) means that the word "hanta" was mentioned 35 times in all articles referring to the state of Los Lagos in Chile over the epi-week 28 provided by source "www.biobiochile.cl".

Given a time point T, we partition the data across different sources in S and view each source $s \in S$ as a *time-evolving document* consisting of a collection N_s of time-stamped tuples, each associated with a certain *latent topic*.

Finally, let $\mathcal{X}$ denote the set of all tuple collections N_s for all sources $s \in S$ until time point T. Assuming a set of tuple collections $\mathcal{X}$ that get updated with time, our goal is to predict potential disease outbreaks for all locations present in the input data for the future time point $T+1$. Finally, we assume access to a gold-standard report (GSR) providing ground truth information for disease outbreaks at locations in L for time points $t \prec T$.

3 Spatio-temporal Topic Models

The first component of SourceSeer deals with the topic and pattern discovery problem. We introduce a topic model that explicitly models time and location, jointly with the word co-occurrence patterns over news articles from multiple data sources. This is done by incorporating both spatial and temporal component into the basic Latent Dirichlet Allocation (LDA) framework [4].

Our proposed spatio-temporal topic model uses location and topic specific distributions to model the generation of words and timestamps. Topic discovery is influenced not only by word co-occurrences, but also spatial and temporal information. Our notation is summarized in Table 1, and the graphical model representation of the model is shown in Figure 3. The generative process for the word and time point of each entry corresponding to an observed location is:

Table 1: Notation used in this paper.

Symbol	Description
K	Number of topics
S	Number of sources
V	Number of words
T	Number of discrete time-points
L	Number of locations
N_s	Number of entries in each source s
θ_l	Topic multinomial distr. for location l
ϕ_z	Word multinomial distr. for topic z
ξ_z	Time point multinomial distr. for topic z
z_{si}	Topic of the ith entry from source s
l_{si}	Location of the ith entry from source s
w_{si}	Word of the ith entry from source s
t_{si}	Time point of the ith entry from source s

STAT generative process

1. Draw K multinomials $\phi_z \sim \text{Dir}(\beta)$ for each topic z
2. Draw K multinomials $\xi_z \sim \text{Dir}(\gamma)$ for each topic z
3. Draw L multinomials $\theta_l \sim \text{Dir}(\alpha)$ for each location l
4. For each source $s \in S$ and entry $i \in N_s$ with l_{si}:
   (a) Draw a topic z_{si} from the multinomial $\theta_{l_{si}}$
   (b) Draw a word w_{si} from multinomial $\phi_{z_{si}}$
   (c) Draw a time-point t_{si} from multinomial $\xi_{z_{si}}$

Each source entry is associated with a location $l_{si} \in L$ and we consider a distribution $\theta_{l_{si}}$ over topics that is randomly sampled from a Dirichlet with parameter α. To generate each entry $i \in N_s$ for source s, first, a topic z_{si} is chosen from the topic distribution $\theta_{l_{si}}$, and then, a word w_{si} and time-point t_{si} are generated by randomly sampling from the topic-specific multinomial distributions $\phi_{z_{si}}$ and $\xi_{z_{si}}$. In our experiment we assume a fixed number of topics K.

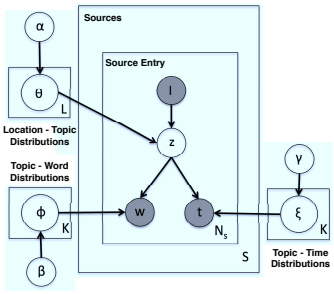


Figure 3: The proposed spatio-temporal topic model.

We use Gibbs sampling to perform approximate inference. Using a Dirichlet conjugate prior for the multinomial distributions allows us to easily integrate out θ, ϕ and ξ. To estimate the model parameters, we calculate the conditional probability distribution $\Pr(z_{si}|\mathbf{w}, \mathbf{t}, \mathbf{l}, \mathbf{z}_{-si}, \alpha, \beta, \gamma)$ where $\mathbf{z}_{-si}$ represents the topic assignments for all entries in s except the i-th entry. We have:

$$\Pr(z_{si}|\mathbf{w}, \mathbf{t}, \mathbf{l}, \mathbf{z}_{-si}; \alpha, \beta, \gamma) \propto \frac{n_{w_{si}}^{k,-(s,i)} + \beta_{w_{si}}}{\sum_{r=1}^{V} n_r^{k,-(s,i)} + \beta_r} \cdot \frac{m_{t_{si}}^{k,-(s,i)} + \gamma_{t_{si}}}{\sum_{t=1}^{T} m_t^{k,-(s,i)} + \gamma_t} \cdot \frac{o_{l_{si}}^{k,-(s,i)} + \alpha_{l_{si}}}{\sum_{l=1}^{L} o_l^{k,-(s,i)} + \alpha_l} \tag{3.1}$$

where n_r^z denotes the number of times word r was associated with topic z across all sources and entries, m_t^z denotes the number of times time-point t was associated with topic z across all sources, o_l^z denotes the number of times location l was associated with topic z across all sources and their entries, and $-si$ in the superscript indicates that the current example has been excluded by the count summations. The derivation of the Gibbs sampling algorithm is provided in the supplementary material [19]. Once the sampler has converged, the parameters for θ, ϕ, and ξ are estimated as:

$$\begin{aligned} \theta_{l,z} &= \frac{o_l^z + \alpha_l}{\sum_{z=1}^{K} o_l^z + \alpha_l} \\ \phi_{z,v} &= \frac{n_l^z + \alpha_l}{\sum_{v=1}^{V} n_v^z + \beta_v} \\ \xi_{l,z} &= \frac{m_t^z + \alpha_l}{\sum_{t=1}^{T} m_t^z + \gamma_t} \end{aligned} \tag{3.2}$$

For each entry in a set of event collection $\mathcal{X}$ we assign a hidden topic z according to Equation (3.1), and update the appropriate counts. After the sampling, we compute the distributions $\boldsymbol{\theta}$, $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ according to equation Equation (3.2).

4 Source-based Disease Outbreak Prediction

The second component of SourceSeer is responsible for forecasting outbreaks at a future time point t for each of the locations present in $\mathcal{X}$. At a high-level, for each location l, we extract an individualized prediction from each source that is relevant to l and fuse the individual predictions using weighted majority voting. We learn the corresponding weights using a multiplicative weights update algorithm.

4.1 Predicting Disease Outbreaks with a Single Source

Detecting an anomaly in the content of source s for location l, requires reasoning about the relevance of the source's content to the discovered disease topics. We view this problem as an instantiation of the *document classification* [24] problem and show how the relevance between the content of a source and a topic can be measured using *cosine similarity*.

For each topic $z \in K$, we have a distribution ϕ_z over all words in the vocabulary V. Following a similar approach to Matsubara et al. [15], we extract the average occurrence rate $\bar{x}_w$ for each word $w \in V$ across all entries and construct an *average representative document* for each topic $z \in Z$, characterized by a vector F_z that contains the expected occurrence frequency of each word $w \in V$ given the topic. We define the w-th entry of F_z corresponding to word w as $F_z, w = \bar{x}_w \cdot \phi_{z,w}$. Similarly, given a source s, a location l and a time point t, the content of a source is described with a word frequency vector $F_{s,l,t}$. Given the vectors F_z and $F_{s,l,t}$ we define the relevance of the content of source s for location l at time t to topic z as:

$$\mathsf{Relevance}(s, z; l, t) = \mathsf{CosineSimilarity}(F_{s,l,t}, F_z) \tag{4.3}$$

where the cosine similarity of two vectors A and B is:

$$\mathsf{CosineSimilarity}(A, B) = (A \cdot B)/(\|A\|\|B\|) \tag{4.4}$$

We want to predict disease outbreaks at future time points when the content of each source is not available to us. Therefore, given a source s, a location l and a future time point t, we estimate the entries of $F_{s,l,t}$ considering the expected frequency of each word. Let $\hat{F}_{s,l,t}[w]$ denote the expected frequency for word $w \in V$. To compute the expected frequency $\hat{F}_{s,l,t}[w]$, we need to consider the conditional probability of source s mentioning word w at a future time point t, denoted by $\Pr(t|s, w)$, the conditional probability of source s publishing word w in an article related to the location l, denoted by $\Pr(w|s, l)$, and the probability of word w being generated by any topic $z \in K$, given location l and time point t. We have that:

$$\hat{F}_{s,l,t}[w] = \bar{x}_w \cdot \Pr(t|s, w) \cdot \Pr(w|s, l) \cdot \sum_{z \in K} \phi_{z,w} \cdot \theta_{l,z} \cdot \xi_{z,t} \tag{4.5}$$

where $\bar{x}_w$ denotes the average rate of occurrences of word w in $\mathcal{X}$, and $\phi_{z,w}$, $\theta_{l,z}$, and $\xi_{z,t}$, can be retrieved by the output of the topic model component of SourceSeer. Given the historical data, we estimate the probability $\Pr(w|s, l)$ with its maximum likelihood as:

$$\Pr(w|s, l) = \frac{n_{w,s,l}}{\sum_{w \in V} n_{w,s,l}} \tag{4.6}$$

where $n_{w,s,l}$ denotes the number of mentions of word w from source s in location l. Notice that $\xi_{z,t}$, i.e., the probability of topic z being prominent at time t, and $\Pr(t|s, w)$ correspond to future time points and need to be estimated.

According to the problem description in Section 2, the available historical data spans up to time point $t - 1$. Thus, we estimate the probability of the source mentioning a particular word w at a future time t by considering the weighted average occurrence rate of word w in the source:

$$\Pr(t|s, w) = \frac{\sum_{\tau=1}^{t-1} \frac{1}{t-\tau} I(\tau, s, w)}{\sum_{\tau=1}^{t-1} \frac{1}{t-\tau}} \tag{4.7}$$

where $I(s, \tau, w)$ is an indicator variable equal to one if source s mentioned word w at least once at time τ, and zero otherwise. To estimate the probability $\xi_{z,t}$ with $z \in \{1, 2, \ldots, K\}$, we use the values of distribution $\xi_z, \forall z \in K$ corresponding to past time points. In particular, we use an autoregressive model over the values of topic z for the n previous time intervals, denoted by $\xi_{z,t-1}, \xi_{z,t-2}, \ldots, \xi_{z,t-n}$:

$$\xi_{z,t} = a_1 \cdot \xi_{z,t-1} + a_2 \cdot \xi_{z,t-2} + \cdots + a_n \cdot \xi_{z,t-n} \tag{4.8}$$

where $a_1, a_2, ..., a_n$ are the regression coefficients. We compute the source-topic relevance for each source-location and topic combination using the aforementioned techniques.

Since rare disease incidences are scarce over time, the source-topic relevance values for a rare disease topic will be low for most time points and high only for few time points corresponding to an outbreak. Thus, high relevance values for a rare disease topic, can be viewed as anomalous points, and anomaly detection techniques can be used.

We use one-class SVMs [22] (OCSVM) to classify the source-topic relevance values as anomalous or not. OCSVMs have successfully been used in a variety of anomaly detection tasks [14, 10, 23]. Furthermore, OCSVMs present superior performance compared to other anomaly detection techniques, such as Nearest Neighbor classification, in scenarios where a small number of anomalous example is available [12]. Finally, OCSVMs do not make any assumptions on the distribution of the data points.

To predict outbreaks for a future time point t, we train a separate OCSVM for each source-location pair (s, l) using the source-topic relevance values for all time points up to $t - 1$ as training data. The training entry for a time point $t' \prec t$ corresponds to a vector $< \mathsf{Relevance}(s, z_1; l, t'), \mathsf{Relevance}(s, z_2; l, t'), \cdots >$ containing the relevance values for all topics $z_1, z_2, \ldots$ that are relevant to the rare disease under consideration.

4.2 Fusing Multiple Predictions

To forecast an outbreak for a specific location, we fuse the predictions of all sources

into a single prediction for each location $l \in L$ at time t. We use a weighted majority voting algorithm based on the multiplicative weights update framework[2].

Given time t in the future, we focus on a location l and view each source $s \in S$ as an expert providing a prediction $d_s \in [-1, 1]$ with the value -1 corresponding to the emergence of an outbreak and 1 otherwise. We assign a weight w_s to each source, and given the predictions of all sources, we predict yes/no for an outbreak at location l by taking the majority vote $\sum_{s \in S} w_s \cdot d_s$. We learn weights w_s using the multiplicative weights algorithm shown in Alg. 1.

Algorithm 1 Multiplicative Weights Update for Sources

1: **Input:** S_l: set of sources for location l; $\mathbf{D_l}$: training points; R_{S_l}: source-topic relevance dictionary for sources in S_l and points in $\mathbf{D_l}$; O_{S_l}: one-class SVMs for S_l; ϵ: discount factor
2: **Output:** $\mathbf{W}$: weights for sources in S_l
3: Initialize all weights W to 1
4: **for all** $d \in \mathbf{D_l}$ **do**
5: **for all** $s \in S_l$ **do**
6: */*Extract the expert's vote*/*
7: $v \leftarrow O_{S_l}[s].predict(R_{S_l}[s][d])$
8: **if** v is wrong **then**
9: $W_k \leftarrow W_k \cdot \exp(-\epsilon)$ */* Decrease the weight */*
10: **else**
11: $W_k \leftarrow W_k \cdot \exp(\epsilon)$ */* Increase the weight */*
12: Normalize the weights to sum up to 1.0
13: **return** $\mathbf{W}$

Consider a location l. To construct the necessary input for the multiplicative weights update algorithm, we: (i) identify the set of sources S_l relevant to location l, i.e., sources that have published for location l, and (ii) construct the set of training points $\mathbf{D_l}$ by considering the reported outbreaks in GSR (Section 2) for location l and the disease under consideration. We populate $\mathbf{D_l}$ with tuples of the form $(timepoint, outbreak)$ for all historical time points up to the latest time point present both in $\mathcal{X}$ and GSR and set the value of $outbreak$ to -1 if an actual outbreak was reported and 1 otherwise. Finally, we use the past source-topic relevance values for the sources in S_l and the training points in $\mathbf{D_l}$.

Given the input described above, the algorithm proceeds in an iterative fashion updating the weights of the sources considering the accuracy of their predictions. More precisely, the algorithm iterates over all training points in $\mathbf{D_l}$ (Ln. 4). At each iteration, it examines all available sources (Ln. 5) and extracts their prediction corresponding to a specific training point from the past (Ln. 6-7). If the expert is mistaken, it's corresponding weight is reduced in a multiplicative fashion (Ln. 9), otherwise its weight is increased (Ln. 11). Finally, the algorithm outputs the normalized weights, which are later used to fuse the individual source predictions for future time points. The process is repeated as more ground-truth data are becoming available in GSR.

Finally, we associate each outbreak prediction for location l with a *confidence score*. Let S be the set of relevant sources for location l and S_{-1} be the subset of sources predicting an outbreak. Moreover, let $a_l(s)$ be the overall *accuracy* of a source $s \in S_l$ considering its past predictions for location l. The accuracy of source s is defined over the available past time window as $a_l(s) = \frac{\#\text{ correct predictions}}{\#\text{total prediction}}$ and corresponds to the probability of s giving a correct prediction. Combining the above, the confidence score is:

$$\text{ConfScore} = \prod_{s \in S_{-1}} a_l(s) \cdot \prod_{s \in S_l \setminus S_{-1}} (1 - a_l(s)) \tag{4.9}$$

Given the confidence score of each outbreak prediction, one can use a threshold mechanism to select the final outbreak predictions, and balance the trade-off between precision and recall as we discuss in Section 5. In particular, one can select to report an outbreak prediction only if it is in the 95% confidence interval. Fusing the predictions of individual sources, we predict *if* a disease outbreak will happen during a specific epi-week. To predict the exact day of the incidence, we adopt a standard relative date within the epi-week to be the date at which the rare disease incidence will occur, and tune it using cross-validation.

5 Experimental Evaluation

We evaluate the proposed framework using real-world data focusing on Hantavirus outbreaks in Latin America.

5.1 Experimental Setup We first provide a description of our experimental setup.

Data: We use a dataset corresponding to a corpus of public health-related news articles extracted from HealthMap [8], a prominent online aggregator of news articles and tweets for disease outbreak monitoring and real-time surveillance of emerging public health threats. Since, we focus on countries in Latin America the vocabulary we consider consists of Spanish and Portuguese words and does not contain only disease related words. Traditional IR pre-processing such as stop-word removal and term frequency modeling is performed over a fixed vocabulary of words. The dictionary contains words that are either commonly associated with diseases (e.g.,"contagious") or words associated with a specific disease (e.g.,"rodents","hanta" for hantavirus). Finally, each article is associated with a data source and a location corresponding to a country-state pair.

When predicting for an epi-week t we use historical (weekly) data from June 2012 up to the previous week $t-1$ to discover the topics using the model presented in Section 3 and estimate the source-topic relevance values for t at each available location. As we progress to prediction for forthcoming weeks, we gather the estimated source-topic relevance values corresponding to past weeks and align them with the gold standard report (Section 4.1) to form the necessary training points for the OCSVMs. We evaluate the performance of our proposed techniques from January 2013 to March 2014. The size of the input data varies over time, as

new articles are added every epi-week. The number of words ranges from 20,908 to 48,700, the number of locations from 74 to 144 and the number of data sources from 381 to 798.

GSR: We make use of a gold standard report (GSR) which gives ground truth determinations of whether a disease incidence (hantavirus) happened in a given location. The GSR is determined by analysts of MITRE [1] considering multiple news sources and studying bulletins issued by health reporting organizations such as ProMED [1].

Models: We evaluate the following models:

- SourceSeer: Our source-based framework introduced in Sections 3 and 4 coupled with a thresholding mechanism where for a week and country accepts only the predictions with confidence scores (Equation (4.9)) in the top-k percentile of all prediction scores for that country.
- LocSeer: A variation of SourceSeer that uses the topic model component to identify disease related topics but integrates this with a *location-only* anomaly detection approach. We follow an approach similar to the one introduced in Section 4.1. For each location we calculate the location-topic relevance values for future time points and use an OCSVM to detect anomalous points. To calculate the location-topic relevance, we estimate each entry of the location's word frequency vector as $\hat{F}_{l,t}[w] = \bar{x}_w \Pr(t|l,w) \sum_{z \in K} \phi_{z,w} \cdot \theta_{l,z} \cdot \xi_{z,t}$, where $\Pr(t|l,w)$ is defined similarly to Equation (4.7). Intuitively, LocSeer integrates news articles from multiple data sources ignoring the accuracy of individual sources. We use a thresholding mechanism similar to that of SourceSeer considering the accuracy of each state-based OCSVM.
- KeyWord: A *keyword* based prediction technique that monitors the mentions of the Hantavirus related keyword set {"hanta", "hantavirus", "roedores", "ratones", "cardiopulmonar"} and uses an OCSVM to predict future outbreaks based on past mentions of words. This word-set reflects the fact that Hantaviruses have almost entirely been linked to human contact with rodent excrement and their symptoms affect the heart and the lungs.
- BRM: A *base rate model* that assumes a fixed rate for the occurrence of rare disease outbreaks for each location and for each month. To determine this rate, the model extracts the average frequency of outbreak occurrences reported over a past time window of four months. BRM reports disease outbreaks for that location at a frequency equal to the extracted rate. Alerting dates are assigned to the beginning of each month while event dates are assigned uniformly at random to a day within the corresponding month. We take the average performance over 25 independent runs.

[1]The Mitre Corporation is a not-for-profit company that operates multiple federally funded research and development centers.

All models are implemented in Python and the evaluation is performed on an Intel(R) Xeon(R) CPU E7- 4870 @2.40GHz/64bit/1TB machine.

Parameter Setup: The OCSVM parameters are tuned using leave-one-out cross-validation. For the topic model, we set the parameters of the Dirichlet priors to $\alpha = 2/K$, $\beta = 0.01$ and $\gamma = 0.01$ where K is the number of topics. We evaluated the topic model with $K = \{8, 12, 15\}$ and found that $K = 12$ results in more meaningful topics.

Metrics: We adopt five key measures of performance. Given our predictions, we compute the precision, recall and F1-score at a country level, grouping together prediction for locations in the same country. We also compute an average warning quality for each country. Each prediction for a location in the country under consideration is assigned a quality score $Q = \frac{4}{3}(1 + a_{loc} + a_{date})$, where a_{loc} and a_{date} denote the location and date accuracy of the prediction. To calculate a_{loc} we use a two-level topology, considering the country, and state corresponding to the location of a warning. A partial score of 0.5 is assigned to a warning if it matches the country of an outbreak correctly and an additional score of 0.5 is assigned if the warning matches the state correctly. The date specific accuracy a_{date} is calculated as:

$$a_{date} = 1 - \frac{min(|\text{predicted date} - \text{actual date}|, 7)}{7} \tag{5.10}$$

Finally, we consider the lead time of our predictions, which is calculated as the time between the date of alerting and the actual date of reporting the outbreak (not the incidence date of the outbreak). Notice that lead time is different from the date accuracy described above.

Mapping Warnings to Events: Since there could be multiple events (and/or alerts) in a given month, a strategy is necessary to map events to alerts. We conduct a maximum bipartite matching between events and alerts where (i) an edge exists if the alert was issued prior to the reporting date of the event, (ii) the weight on the edge denotes the quality score.

5.2 How effective is the proposed topic model in discovering disease topics and their spatio-temporal patterns?

The HealthMap corpus contains mentions to both common and rare diseases over multiple countries in Latin America. The most prevalent diseases mentioned in the dataset are avian flu (i.e., type h5n1), dengue fever, swine flu (i.e., h1n1 flu), the hantavirus pulmonary syndrome (HPS) and the hantavirus hemorrhagic fever with renal syndrome (HFRS) [11].

We evaluate the topics discovered by our topic model. Six out of the twelve topics are related to the diseases mentioned above, while the rest are background topics related to non-disease aspects of the news articles. We focus only on the disease related topics. To evaluate the disease topics, we consider a vocabulary of 184 health-related words. For each topic, we examine the most likely words based on the health-related vocabulary and their per-topic probabilities.

Table 2 shows three topics related to hantavirus and their most likely words based on the health-related vocabulary.

Table 2: Three discovered topics that are related to hantavirus. The top words with their probability in each topic are shown.

HPS		HFRS		Hanta Transmission	
virus	0.0468	vacuna	0.0057	paciente	0.0220
epidemia	0.0443	campos	0.0031	transmissor	0.0133
enfermos	0.0066	provincial	0.0028	lixo	0.0099
hanta	0.0068	hantavirus	0.0024	criaderos	0.0088
viral	0.0038	tosse	0.0022	respiratorias	0.0061
territorio	0.0027	nariz	0.0019	manos	0.0056
pneumonia	0.0014	estornudar	0.0011	boca	0.0047
sangre	0.0014	abdominal	0.0008	rural	0.0038
ratones	0.0006	lluvia	0.0008	musculares	0.0028
cardiopulmonar	0.0002	renal	0.0005	roedores	0.0022

The first topic refers to the HPS syndrome with words such as "pneumonia", "sangre" (blood), and "cardiopulmonar" being ranked higher. We see that the proposed topic model is able to retrieve the correlation between words "hanta" and "ratones" (mice) successfully. The second topic focuses on the HFRS syndrome with words as "nariz" (nose), "estornudar" (sneeze), "renal" being more prevalent. Finally, the third topic focuses on the hantavirus transmission routes with words as "lixo" (garbage), "criaderos" (breeding places), "manos" (hands) and "roedores" (rodents) being ranked higher than others. According to Jonsson et al. [11] HPS is the main syndrome observed in the Americas while HFRS cases are mainly observed in Eurasia. Thus, observing the HFRS topic seems unexpected. However, after analyzing the actual articles in our corpus, we found that articles reporting hantavirus incidents usually mention both hantavirus syndromes for informational purposes. The top words for the other disease topics are provided in the supplementary material [19] due to space limitations.

Next, we examine the temporal patterns discovered for the hantavirus topics. Given a time point, we define the prominence of a topic as the temporal distribution value for that topic at that time point. The prominence histograms are shown in Figure 4. The HFRS topic shows small fluctuations across time. However, the HPS topic follows a trend similar to that of the hantavirus incidence timeline (Figure 1(a)).

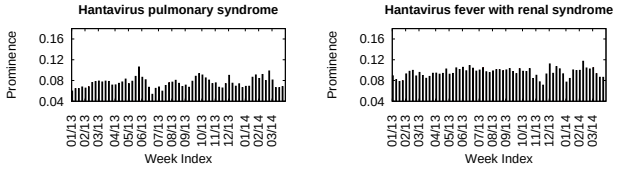


Figure 4: The topic prominence timeline for HPS and HFRS.

Finally, we examine the correlations between the discovered topics and the countries in Latin America under consideration. Figure 5 shows the prominence of each topic for Brazil, Chile, Uruguay and Argentina. As expected, we observe that in Chile, HPS and HFRS are more prominent, while in Brazil Dengue topic is prominent as Brazil is prone to dengue outbreaks throughout the year.

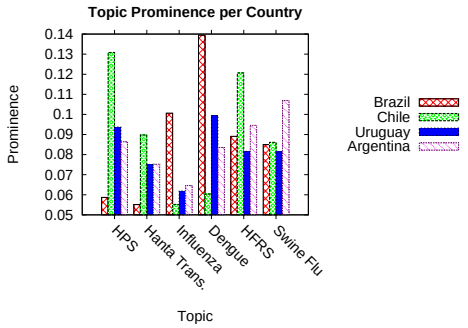
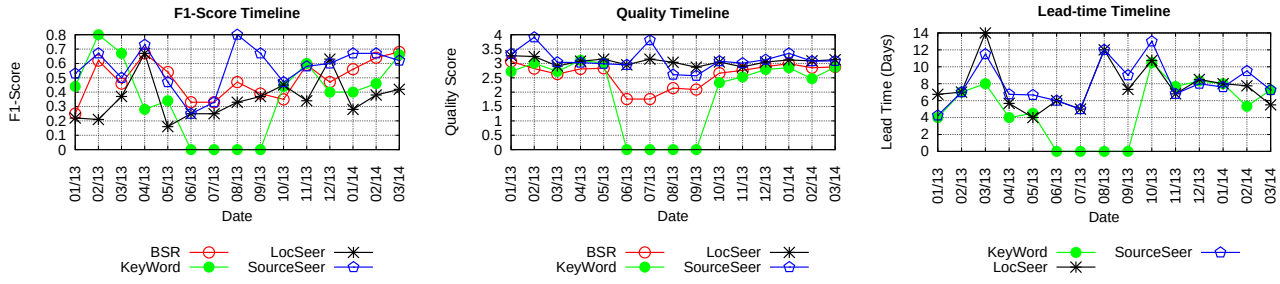


Figure 5: The country specific topic prominence for different diseases averaged over states.

5.3 How efficient is SourceSeer at forecasting disease outbreaks? We evaluate the performance of the various disease outbreak forecasting algorithms focusing on hantavirus incidences at the country level considering the predicted outbreaks for Argentina, Chile, Uruguay and Brazil. We apply BSR, KeyWord, SourceSeer and LocSeer. We evaluate the performance of SourceSeer and LocSeer with $k \in \{5, 10, 20, 30, 40, 50, 70\}$. We use the three hantavirus topics described above to construct the necessary feature vectors for SourceSeer and LocSeer.

Figure 6(a) shows the F1 score of the four approaches from January 2013 to March 2014 aggregated over all countries. As shown, SourceSeer obtains the best F1-score for most of the months. The F1 score of BSR is lower as its recall is significantly lower compared to that of SourceSeer. The latter is expected as BSR can only predict outbreaks for states where a sufficient number of outbreaks has occurred in the past. In fact, due to its design BSR fails completely to forecast outbreaks for states or countries where no outbreaks have been observed in the past (e.g., the outbreak in Brazil for October 2013 and the outbreak in Uruguay for March 2013). However this mechanism limits the number of false positives significantly, and thus, for many months we observe slightly higher or comparable precision scores for BSR with those of SourceSeer. The F1 score of LocSeer is significantly lower compared to SourceSeer due to its significantly lower precision scores. The reason for this behavior is the increased number of false positives returned by LocSeer even after the thresholding mechanism was employed. Finally, KeyWord performs reasonably well when there is an increase in the number of outbreaks in previous weeks leading to increased keyword counts. However, the model performs poorly in the presence of low keyword counts. KeyWord failed to forecast the outbreaks in August and September 2013 as only one was reported in July. A detailed evaluation on the precision, recall and F1 score is provided in the supplementary material [19].

Is the performance gain of SourceSeer significant? To obtain a clearer understanding of SourceSeer's performance gain, we perform the Wilcoxon signed-rank [27] test com-



(a) (b) (c)

Figure 6: (a) F1-score timeline (b) quality score timeline for BSR, KeyWord, LocSeer and SourceSeer on forecasting hantavirus outbreaks. (c) Lead-time timeline for LocSeer and SourceSeer on predicting hantavirus outbreaks.

paring the performance of BSR with SourceSeer, KeyWord with SourceSeer and LocSeer with SourceSeer for precision, recall, and F1-score across all months. In Table 3 we report the corresponding test statistic scores W and the z-scores. We consider a baseline confidence level of $\alpha = .05$. As shown, the performance difference between BSR and SourceSeer is statistically significant for recall and F1 (with SourceSeer outperforming BSR) while the difference for precision is not statistically significant. The same behavior was observed for KeyWord and SourceSeer. For LocSeer and SourceSeer, we see that the performance gain of SourceSeer for precision and F1 is statistically significant while the difference for recall is not. We did not observe significant differences in the performance of LocSeer and SourceSeer for different values of k.

We further analyze the performance of the four models by comparing the quality score cross all months under consideration. Figure 6(b) shows the average prediction quality score obtained by each model from January 2013 to March 2014. A higher quality score is an indicator that a model can predict outbreaks correctly at the state and not only at the country level. As shown, both LocSeer and SourceSeer outperform BSR and KeyWord significantly. This is expected since BSR relies only on past reported events to predict future outbreaks and KeyWord on increased keyword counts, hence, by design both cannot predict outbreaks in states with no reported incidents. Moreover, we observe that SourceSeer obtains higher quality scores for most of the months compared to LocSeer. This is due to its capability of weighting the predictions of difference sources based on their accuracy for each specific state.

What is the lead-time gain of SourceSeer? Finally, we analyze the average lead-time of KeyWord, LocSeer and SourceSeer to examine if the proposed models can forecast outbreaks in a timely manner. Figure 6(c) shows the lead-time timeline of the three models from January 2013 to March 2014. We observe that both models have a significant lead-time advantage when compared against the mention of the outbreak in news sources and also outperform KeyWord.

Discussion From our experiments we see that SourceSeer can effectively discover rare-disease topics and their spatio-temporal patterns. We also observed that exploiting the different authoritativeness levels of news sources enables us to forecast outbreaks more accurately even when no outbreaks were reported in the past. Finally, SourceSeer can forecast outbreaks ahead of news media with an average lead-time of 8 days.

Table 3: Wilcoxon signed-rank statistical significance test on SourceSeer's performance gain. H_0: The median performance difference between the pairs is zero. Reject H_0: $\|z\| \geq 1.645$ or $W \geq 15$ when z not applicable. Baseline confidence level of $\alpha = .05$. Bold fonts denotes statistically significant differences.

Metric	Score	SourceSeer **v.s. BSR**	SourceSeer **v.s.** LocSeer	SourceSeer **v.s.** KeyWord
Prec.	W	-51	81	36
	z	-1.463	**2.966**	1.349
Rec.	W	114	3	76
	z	**3.223**	-	**2.961**
F1	W	61	101	100
	z	**1.899**	**3.154**	**2.825**

6 Related Work

To the best of our knowledge most existing topic models focus on the temporal or spatial trends in isolation and do not analyze both types of trends jointly. A number of methods have been proposed for analyzing the time evolution of topics in document collections, such as the topics over time (TOT) model [26], the dynamic topic model (DTM) [3], and TriMine model [15]. TOT handles time-windows of fixed size and uses a Beta distribution to model the evolution of a topics. DTM focuses on a time-window of fixed size but uses Kalman filters, and TriMine is able to analyze windows of variate size is able to find cyclic time patterns with different timescales, which enables predicting future events.

A different line of work, Spatial Latent Dirichlet Allocation (SLDA) [25], focuses on discovering spatial patterns jointly with the word co-occurrences. While the model focuses on computer vision applications where documents are comprised by visual words the proposed techniques can be trivially extended to regular text documents. A similar approach was introduced by Ramage et al. [18] for labeled documents where the labels can correspond to locations.

Finally, although previous approaches [17, 20, 21] have considered topic models for forecasting disease outbreaks,

they focus on common diseases, like influenza, for which large amounts of data or specialized medical records are available. None of these models considers the accuracy of different data sources when predicting outbreaks.

7 Conclusions

We studied the problem of rare disease outbreak prediction when analyzing dynamic news sources providing an evolving corpus of news articles. We introduced SourceSeer, a framework that combines spatio-temporal topic models with source-based anomaly detection techniques for forecasting rare disease outbreaks at fine spatial granularity by considering the accuracy of each individual news source. Experimental results show the effectiveness of our proposed framework and illustrate how taking the accuracy of data sources into account leads to higher quality predictions.

Acknowledgements

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] International society for infectious diseases. `http://www.promedmail.org/`.

[2] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1), 2012.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. ICML, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3, 2003.

[5] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Med.*, 5(7), 2008.

[6] C. Corley, D. Cook, A. Mikler, and K. Singh. Text and structural data mining of influenza mentions in web and social media. *IJERPH*, 7(2), 2010.

[7] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. SOMA, 2010.

[8] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *JAMIA*, 15, 2008.

[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457, 2009.

[10] K. A. Heller, K. M. Svore, A. D. Keromytis, and S. J. Stolfo. One class support vector machines for detecting anomalous windows registry accesses. In *Workshop on Data Mining for Computer Security*, 2003.

[11] C. B. Jonsson, L. T. M. Figueiredo, and O. Vapalahti. A global perspective on hantavirus ecology, epidemiology and disease. *Clinical Microbiology Review*, 23(2), 2010.

[12] S. S. Khan and M. G. Madden. One-class classification: Taxonomy of study and review of techniques. *CoRR*, abs/1312.0049, 2013.

[13] J. P. Linge, R. Steinberger, T. P. Weber, R. Yangarber, and E. van der Goot. Internet surveillance systems for early alerting of health threats. *Eurosurveillance*, 14(13), 2009.

[14] L. M. Manevitz and M. Yousef. One-class svms for document classification. *JMLR*, 2, 2002.

[15] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa. Fast mining and forecasting of complex time-stamped events. KDD, 2012.

[16] J. Parker, Y. Wei, A. Yates, O. Frieder, and N. Goharian. A framework for detecting public health trends with twitter. ASONAM, 2013.

[17] M. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health, 2011.

[18] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. EMNLP, 2009.

[19] T. Rekatsinas, S. Ghosh, S.R. Mekaru, E.O. Nsoesie, J.S. Brownstein, L. Getoor, and N. Ramakrishnan. SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources - Supplementary Material. `http://www.cs.umd.edu/~thodrek/uploads/sourceseer_sup.pdf`.

[20] A. K. Rider and N. V. Chawla. An ensemble topic model for sharing healthcare data and predicting disease risk. BCB, 2013.

[21] A. Sadilek, H. A. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. AAAI, 2012.

[22] B. Schoelkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *NEURAL COMP*, 13, 1999.

[23] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *JMLR*, 6, 2005.

[24] A. Strehl, E. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *WebSearch*, AAAI, 2000.

[25] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[26] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. KDD, 2006.

[27] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), 1945.

[28] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. AAAI, 2002.

[29] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. Bayesian network anomaly pattern detection for disease outbreaks. ICML, 2003.