Non-Parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs

Feng Chen
Department of Computer Science
State University of New York at Albany
fchen5@albany.edu

Daniel B. Neill
H.J. Heinz III College
Carnegie Mellon University
neill@cs.cmu.edu

ABSTRACT

Event detection in social media is an important but challenging problem. Most existing approaches are based on burst detection, topic modeling, or clustering techniques, which cannot naturally model the implicit heterogeneous network structure in social media. As a result, only limited information, such as terms and geographic locations, can be used. This paper presents Non-Parametric Heterogeneous Graph Scan (NPHGS), a new approach that considers the entire heterogeneous network for event detection: we first model the network as a "sensor" network, in which each node senses its "neighborhood environment" and reports an empirical p-value measuring its current level of anomalousness for each time interval (e.g., hour or day). Then, we efficiently maximize a nonparametric scan statistic over connected subgraphs to identify the most anomalous network clusters. Finally, the event represented by each cluster is summarized with information such as type of event, geographical locations, time, and participants. As a case study, we consider two applications using Twitter data, civil unrest event detection and rare disease outbreak detection, and present empirical evaluations illustrating the effectiveness and efficiency of our proposed approach.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Non-Parametric Scan Statistics; Event Detection and Forecasting; Social Media; Heterogeneous Graphs

1. INTRODUCTION

Social microblogs such as Twitter and Weibo are experiencing explosive growth, with billions of users globally sharing their daily observations and thoughts online. Unlike traditional channels, where collection of information such as patient data, crimes, and financial transactions is costly and time consuming, social media provides the vast amount of data available in real time on the internet at almost no cost. Social media also helps spread information earlier and faster than traditional media. For example, Twitter first leaked credible word of Osama bin Laden's death before President Obama's announcement, and there were a half million tweets (and only 800 news mentions) one hour after the event [1]. As a social "sensor" which can identify emerging patterns in sentiments and opinions, the use of social media holds great promise for detection and forecasting of significant societal events.

However, the size and complexity of social media datasets create a number of technical challenges. First, the language used in social media is highly informal, ungrammatical, and dynamic, and thus traditional natural language processing (NLP) techniques cannot be directly applied. Second, social media is naturally structured as a heterogeneous graph, with entities such as user, post, geographic location, term, hashtag, and link; and relationships such as follower, friendship, reply, retweet, and spatial neighborhood. In addition, the attributes of each entity type could be heterogeneous as well. For example, a user may have daily attributes such as the numbers of active followers, posts, and retweets, while a tweet may have attributes such as the number of terms and the sentiment score. Finally, the size of data necessitates development of new, scalable detection methods.

This paper focuses on the problem of domain-specific event detection and forecasting, for events such as disease outbreaks, civil unrests, and financial crises. Most existing approaches to event detection can be classified into three categories, including burst detection, geographic topic modeling, and clustering. **Burst detection**-based approaches search for space-time regions where the aggregated counts of some predefined terms are abnormally high compared with the counts outside the regions [9, 10, 13]. Sakaki et al. consider spatial-temporal Kalman filtering, which is similar to space-time burst detection, to track the geographical trajectory of hot spots of tweets related to earthquakes [19]. **Geographic topic model**-based approaches estimate language distributions (over a predefined vocabulary) that are distinct in some geographic regions [25, 11, 8]. **Clustering**-based approaches search for novel clusters of documents or terms using predefined similarity metrics, such as cosine similarity and social similarity for documents [22], or autocorrelations [24] and co-occurrences [20, 23] for terms. Similarly, [21] uses features related to text content and link information to cluster tweets. For each cluster identified, the re-

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623619.

lated documents may have different geographical locations, which can be combined by weighted voting [22].

We note that these domain-specific event detection approaches differ in focus from general-domain event detection methods such as RW-Event [3], which attempt to distinguish events from non-event patterns (such as memes) rather than identifying events of a specific type. Such methods do not use content of tweets but only features such as temporal trends of term volume, rely on a large amount of labeled training data (as opposed to the unsupervised problem we consider here), and require extensive parameter tuning.

Each of the aforementioned methods only exploits partial information from social media that is useful for event detection. However, there is very limited work that is able to model the entire social media graph for event detection, due to the computational challenge of modeling the complicated heterogeneous relationships between entities and attributes, and the risk of overfitting [6]. Our approach, described below, incorporates three types of heterogeneity in the social media graph, including heterogeneous 1) entity types; 2) entity attributes; and 3) entity relationships. In addition, the highly informal, ungrammatical, and dynamic language used in social media motivates our use of nonparametric statistical models to provide more accurate detection and forecasting [18].

To address the above technical challenges, we propose a Non-Parametric Heterogeneous Graph Scan (NPHGS) approach for event detection and forecasting using social media data. We attempt to consider all potentially useful information in social media in a unified nonparametric statistical framework, to facilitate the early detection and accurate forecasting of societal events. Specifically, we first model a heterogeneous graph, in which: 1) each node can be of different types, such as user, tweet, geographic location, and hashtag; 2) the relationships between nodes can be of different types, such as retweet, reply, and follower; and 3) each node type can have different attributes, such as the numbers of tweets and users for a given geographic location; the numbers of followers, tweets, and retweets for a given user; and the number of terms and the sentiment score for a given tweet. Second, we further model the network as a "sensor" network, in which each node senses its "neighborhood environment" and reports an empirical p-value measuring the current anomalousness levels of various neighborhood-related attributes. Third, we efficiently maximize a nonparametric scan statistic over connected subgraphs to identify the most anomalous network clusters. Each cluster is returned as the indicator of an ongoing or upcoming event, and is summarized with information such as type of event, geographic locations, time, and participants. The main contributions of our study are summarized as follows:

- **Formulation of the NPHGS framework**. To the best of our knowledge, this is the first work that models the whole heterogeneous social media graph as a "sensor" network, enabling the use of novel, nonparametric graph scan statistics for accurate and scalable event detection and forecasting.
- **Design of two-stage empirical p-value calibration process.** The heterogeneity of different node types and node attributes, and the correlations between different attributes of a node, are well addressed by calibrating all node types and attributes on the same scale using the proposed two-stage empirical calibration process.
- **Development of an approximate algorithm for non-parametric graph scanning**. The nonparametric scan statistic over connected subgraphs is approximately maximized by iterative subgraph expansion and linear time subset scanning, with time complexity $O(|\mathcal{V}| \log |\mathcal{V}|)$, where $|\mathcal{V}|$ refers to the total number of graph nodes.
- **Evaluation of theoretical properties**: Our proposed approximate algorithm is guaranteed to find the globally optimal solution if the data contains no "breaktire" entities (see Subsection 3.3), and is equivalent to percolation-based graph scan under certain simplifying assumptions.
- **Comprehensive experiments to validate the effectiveness and efficiency of the proposed techniques**. NPHGS was evaluated by extensive experiments on real Twitter datasets. The results demonstrate that NPHGS outperforms existing representative techniques for both event detection and forecasting, increasing detection power, forecasting accuracy, and forecasting lead time while reducing time to detection.

The rest of this paper is organized as follows. Section 2 discusses heterogeneous graph modeling for social media, and considers Twitter data as a case study. Section 3 proposes nonparametric scan statistics for heterogeneous graphs. Experiments on real Twitter datasets are presented in Section 4, and Section 5 describes future work.

2. HETEROGENEOUS GRAPH MODELING

A heterogeneous graph is composed of nodes, attributes, and relations that could be of multiple different types. The formal definition of a heterogeneous graph is as follows:

Definition 1 (Heterogeneous Graph). *A heterogeneous graph is defined as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f, \psi)$, where $\mathcal{V} = \{\mathcal{V}_1 \cup \cdots \cup \mathcal{V}_C\}$, $\mathcal{V}_c$ refers to the set of entities of type c, C refers to the total number of entity types, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ refers to the set of edges, $f = \{f_1 \dots f_C\}$ is a set of C mapping functions, $f_c : \mathcal{V} \to \mathcal{R}^{D_c}$ defines a D_c-dimensional feature vector ($f_c(v)$) for each node v of type c, ψ refers to a mapping function such that each $e \in \mathcal{E}$ belongs to a particular type of relation $\psi(e) \in \{1 \dots Q\}$, and Q refers to the number of different relation types.*

Throughout the paper, we consider the detection of civil unrest events and rare disease outbreaks using Twitter data as a case study. For this application, we assume that a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f, \psi)$ has been extracted from the Twitter data. Additionally, we assume historical data $f_c(v^{(t)})$ (for $t = 1 \dots T$), corresponding to each feature vector $f_c(v)$, which will be used to estimate the anomalousness of the current feature values. Given these data, we will return the most anomalous connected sub-graphs $\max_{S \subseteq \mathcal{V}} F(S)$, where F is the nonparametric heterogeneous graph scan statistic defined below. Our goal is to identify subsets S corresponding to events of interest, as given by a separate gold standard dataset and as measured by the performance metrics defined in Section 4.

The selected set of entity types includes User, Location, Term, Tweet, Link, and Hashtag. The selected entity relation types and entity attributes are summarized in Figure 1 and Table 1. The sentiment (polarity) score was calculated using the python sentiment analysis package named "Pattern" [27]. The klout score [28] is an overall measure of the tweet author's influence on a scale from 1 to 100. It is assumed that domain-specific content filtering has been conducted as a preprocessing step, and the majority of unrelated tweets have been removed. In our study, domain-specific dictionaries for civil unrest events and hantavirus outbreaks were obtained from domain experts, and tweets that match less than three terms in the dictionary were removed. More complicated content filtering techniques can be applied, such as the training of a SVM classifier [19]. We note that identical pre-processing steps were applied for all methods in our experiments below, and we do not expect the relative performance of methods to be strongly dependent on this pre-processing.

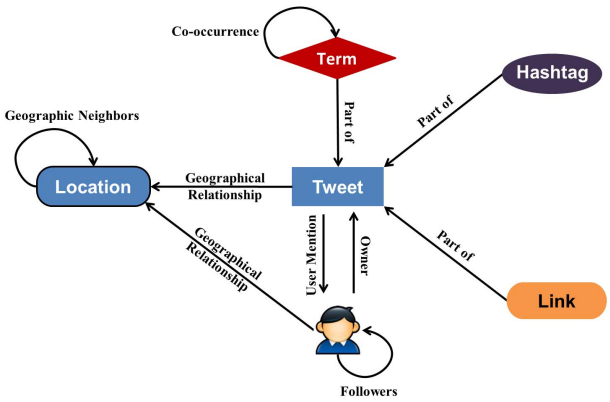


Figure 1: Entity Diagram for Twitter Data Modeling

Node Type	Features Used
User	#tweets, #retweets, #active-followers, #active-followees, #mentions, #replies
Tweet	klout, sentiment, replied-by-graph-size, reply-graph-size, retweet-graph-size, retweet-graph-depth
State	#tweets, #active-users
Term	#tweets
Link	#tweets
Hashtag	#tweets

Table 1: Twitter Node Attributes

3. NON-PARAMETRIC SCAN STATISTICS FOR HETEROGENEOUS GRAPHS

This section presents non-parametric scan statistics for heterogeneous social media graphs. Specifically, Subsection 3.1 discusses the modeling of the heterogeneous graph as a "sensor" network by estimating an empirical p-value for each graph node; Subsection 3.2 presents nonparametric scan statistics; and Subsection 3.3 presents an efficient approximate algorithm that maximizes the nonparametric scan statistic over connected subgraphs to identify the most anomalous graph clusters, which can be regarded as indicators of ongoing or new events.

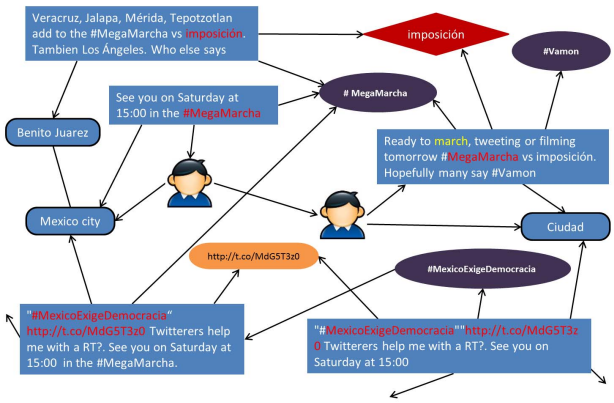


Figure 2: An Example of Twitter Heterogeneous Network

3.1 Two-Stage Empirical Calibration Process

To deal with the heterogeneity of node attributes for different entity types, we propose to model the heterogeneous graph as a "sensor" network. Each entity senses its local "neighborhood" in the graph and reports measurements of some predefined features. For event detection, it is necessary to estimate a baseline distribution for each attribute that characterizes its behavior when there is no event occurring. Given these baseline distributions, it is then possible to estimate an *empirical p-value* representing the degree of anomalousness for each node. This empirical p-value can be regarded as the signal strength of the node's current attribute values as an indicator of some ongoing or newly emerging event.

In order to estimate the baseline distribution of each attribute for each entity, the key component is to collect a good training sample for distribution estimation. We first define an appropriate time granularity for event detection (e.g., hourly, daily, or weekly). We then collect a set of historical observations for each entity and attribute. We consider three scenarios:

- **Entities with sufficient observations**: For Twitter data, these include locations, regular users, and existing keywords, hashtags, and links.
- **Entities with insufficient observations**: For Twitter data, these include new users and newly occurring keywords, hashtags, and links.
- **Entities with single observations**: For Twitter data, the related entities mainly include tweets.

For the first category of entities, it is possible to collect enough historical records for that specific entity to estimate its baseline distribution. For example, if a given user, who has used Twitter for more than a month, tweeted 100 times today, that value would be compared to her daily numbers

of tweets in historical data. For the second and third categories, there are not sufficient historical observations available for each entity. In this situation, we consider all historical records of the same entity type as the training sample, calibrated based on time since first occurrence. For example, if a given tweet is retweeted 50 times on the third day since its creation, that value would be compared to the numbers of retweets for *all* tweets on their third day.

Once the set of historical observations $\{v^{(1)}, \ldots, v^{(T)}\}$ is defined for a given node v, we can then compute the empirical p-values $p_d(v)$ for each attribute of node v, and an overall empirical p-value $p(v)$ for node v, by comparing the current and historical attribute values. Here we assume one-tailed p-values, which answer the question: under the null hypothesis that no events of interest are occurring, what is the probability that a randomly selected sample would have an observed value greater than or equal to the current observation. (The proposed approach can be easily extended to incorporate two-tailed p-values as well.) For *empirical* p-values, we assume that the set of historical observations represent the null distribution of interest, thus testing the null hypothesis of exchangeability of past and current observations. The empirical p-value of a specific feature d for a node v of type c is defined as:

$$p_d(v) = \frac{1}{T}\sum_{t=1}^{T} I\left(f_{c,d}(v^{(t)}) \geq f_{c,d}(v)\right), d = 1 \ldots D_c, \quad (1)$$

where $f_{c,d}(v)$ refers to the d-th component of the feature vector $f_c(v)$. The empirical value $p_d(v)$ defined above can be interpreted as the proportion of historical observations $f^{(t)}_{c,d}(v)$ when there was no event occurring with observed values that are greater than or equal to the current observation $f_{c,d}(v)$. The empirical p-value of node v is then defined as:

$$p(v) = \frac{1}{T}\sum_{t=1}^{T} I\left(\min_{d=1\ldots D_c} p_d(v^{(t)}) \leq \min_{d=1\ldots D_c} p_d(v)\right). \quad (2)$$

The proposed two-stage empirical p-value p(v) has the nice theoretical property of uniformity as shown in Theorem 1:

THEOREM 1 (UNIFORMITY OF P(V)). *The two-stage empirical p-value p(v) defined by Equations (1) and (2) follows a uniform distribution on [0,1] under the assumption that the current multivariate observations for a single node are exchangeable with the reference set given the null hypothesis that no events of interest are occurring.*

PROOF. The assumption of exchangeability of multivariate observations for a single node implies that each feature's observations are exchangeable with the reference set, so that the first-stage p-values are uniform on [0,1]. Moreover, the features of a node are assumed to have the same correlation structure as the reference set, so the minimum of the first-stage p-values is exchangeable with the corresponding minima in the reference set; and thus the second-stage p-values are uniform on [0,1]. □

The challenge of heterogeneity of different node types and node attributes is well addressed from the following three perspectives: First, it deals with network heterogeneity by calibrating all node types on the same scale, such that all p-values $p(v)$ are drawn uniformly on [0,1] under the null hypothesis that no events of interest are occurring. Second, it allows us to consider multiple attributes for a single User, Tweet, or State node without knowing *a priori* which ones will be most indicative of the events of interest. Finally, it accounts for correlation between the first-stage empirical p-values $p_d(v)$ when computing the overall empirical p-value $p(v)$.

To better understand the advantages of our proposed two-stage solution, we briefly compare it to two alternative approaches. First, we could have simply used a one-stage calibration process where the feature-level p-values $p_d(v)$ are passed directly into the nonparametric scan statistic described below. The nonparametric scan computes the score $F(S)$ of a subgraph S as a function of the number of p-values in S which are significant at level α and the total number of p-values in S. However, we expect the p-values for the various features of a given node to be highly correlated. As a result, the one-stage calibration process would be biased toward detecting nodes with more features. For example, suppose that nodes of type 1 have 100 redundant (fully correlated) features, while nodes of type 2 have only a single feature. In this case, a node of type 1 with all 100 p-values equal to .05 would have a much higher score (given one-stage calibration) than a node of type 2 with p-value equal to .05. The two-stage calibration correctly accounts for the correlation structure and would give both nodes the same score.

A second alternative approach would be to define $p(v)$ as the minimum p-value $\min_{d=1\ldots D_c} p_d(v)$ without re-calibrating the significance of $p(v)$ using the historical data. Clearly this naive approach does not account for multiple hypothesis testing; it can be readily proved that this estimator does not follow a uniform distribution under the null and is biased, tending to underestimate the empirical p-value. For example, if 100 p-values for a node were independently drawn from [0,1], we would expect the minimum p-value to be less than $\alpha = .05$ with probability $1 - (1 - 0.05)^{100} = 0.994$. Moreover, the naive approach would again be biased toward giving higher scores to nodes with more features, since the minimum p-value can only be decreased (made more significant) by adding features. Re-calibrating using the historical data in our two-stage process correctly adjusts for this bias. Finally, we note that our two-stage process is sufficiently flexible so that other p-value combination methods (such as Fisher's method) could easily have been used in Equation 2 instead of the minimum p-value, while still satisfying Theorem 1.

3.2 Non-Parametric Scan Statistics

As described above, we obtain a "sensor" network $\mathcal{H} = (\mathcal{V}, \mathcal{E}, p)$ that is the same as the heterogeneous graph $\mathcal{G}$, except that the mapping function $p : \mathcal{V} \rightarrow [0, 1]$ now defines a single empirical p-value corresponding to each node v of the heterogeneous network. Note that the mapping function ψ has been removed: all relation types will be treated identically in the discussion below.

To determine which connected subgraphs are most anomalous, we generalize the nonparametric scan statistic [17], which extends Kulldorff's spatial scan [12] and was originally proposed for modeling spatial-temporal count data, to heterogeneous graphs. The nonparametric scan has also been used for anomalous pattern detection in general categorical datasets [14], but we note that the present work is the first work to generalize nonparametric scan statistic to het-

erogeneous graphs, requiring both the novel two-stage calibration procedure described above (to obtain p-values) and the novel graph scan algorithm described below (to identify subgraphs with surprisingly high numbers of low, significant p-values). Ignoring the graph constraints as in [14] leads to the identification of unconnected subsets consisting of unrelated, individually anomalous nodes from different parts of the heterogeneous network, resulting in substantially reduced detection performance.

The general form of the proposed Non-Parametric Heterogeneous Graph Scan (NPHGS) statistic is defined as:

$$F(S) = \max_{\alpha \leq \alpha_{max}} F_\alpha(S) = \max_{\alpha \leq \alpha_{max}} \phi(\alpha, N_\alpha(S), N(S)), \quad (3)$$

where $S \subseteq \mathcal{V}$ refers to a connected set of nodes, $N_\alpha(S) = \sum_{v \in S} I(p(v) \leq \alpha)$ is the number of p-values significant at level α, and $N(S) = \sum_{v \in S} 1$ is the total number of p-values in subset S. The significance level α can be optimized between 0 and some constant $\alpha_{max} < 1$. The function $\phi(\alpha, N_\alpha(S), N(S))$ refers to a nonparametric scan statistic, i.e., a function that compares the observed number of p-values N_α that are significant at level α to the expected number of significant p-values $E[N_\alpha(S)] = \alpha N(S)$, under the null hypothesis that p-values are uniformly distributed on $[0, 1]$. In this work, we explore the use of one nonparametric scan statistic $\phi(\alpha, N_\alpha(S), N(S))$: the Berk-Jones (BJ) statistic [4]. The BJ statistic is defined as:

$$\phi_{BJ}(\alpha, N_\alpha(S), N(S)) = N \times \mathrm{KL}\left(\frac{N_\alpha}{N}, \alpha\right), \quad (4)$$

where KL is the Kullback-Liebler divergence between the observed and expected proportions of p-values less than α:

$$\mathrm{KL}(a, b) = a \log\left(\frac{a}{b}\right) + (1 - a) \log\left(\frac{1 - a}{1 - b}\right).$$

The BJ statistic can be interpreted as the log-likelihood ratio statistic for testing whether the empirical p-values follow a uniform or piecewise constant distribution. Berk and Jones [4] demonstrated that this statistic fulfills several optimality properties and has greater power than any weighted Kolmogorov statistic.

It is important to consider a range of α in NPHGS, rather than a single threshold for significance. For a fixed α such as $\alpha = 0.05$, the resulting statistic may lose the power to detect a small number of highly anomalous p-values (much smaller than 0.05) or a larger number of subtly anomalous p-values (slightly greater than 0.05). The selection of α_{max} is in practice slightly greater than typical significance levels predefined by users; here we use $\alpha_{max} = 0.15$.

We note that the subsets of p-values identified by our algorithm are affected by multiple testing on two dimensions: we maximize $F(S)$ over subgraphs S and over thresholds $\alpha \leq \alpha_{max}$. To adjust for multiple testing and correctly measure the significance of the detected clusters, we could apply a permutation test, shuffling the temporal component of the data (assuming the null hypothesis of exchangeability), identifying the maximum subgraph score for each permuted sample, and finally comparing the detected cluster score to the distribution of maximum cluster scores for the permuted samples to obtain the p-value of the detected cluster.

3.3 Efficient Non-Parametric Scanning

Based on the proposed NPHGS statistic, the detection of the most anomalous connected subgraph from $\mathcal{V}$ can be formalized as the following optimization problem:

$$\max_{S \subseteq \mathcal{V}: S \text{ is connected}} \max_{\alpha \leq \alpha_{max}} \phi(\alpha, N_\alpha(S), N(S)). \quad (5)$$

It can be shown that the time cost of exact solution to the above optimization problem (5) is exponential in the total number of graph nodes $|\mathcal{V}|$ in the worst case. Therefore, it is necessary to develop approximate solutions. We first observe that it is possible to solve a relaxed version of this problem efficiently by removing the connectivity constraint. Note that we will use this efficient unconstrained optimization as a building block to solve the optimization problem with connectivity constraints. The relaxed problem is formalized as follows:

$$\max_{S \subseteq \mathcal{V}} \max_{\alpha \leq \alpha_{max}} \phi(\alpha, N_\alpha(S), N(S)), \quad (6)$$

which is equivalent to the problem:

$$\max_{\alpha \in U(\mathcal{V}, \alpha_{max})} \max_{S \subseteq \mathcal{V}} \phi(\alpha, N_\alpha(S), N(S)), \quad (7)$$

where $U(S, \alpha_{max})$ refers to the union of $\{\alpha_{max}\}$ and the set of distinct p-values less than α_{max} in S. Because the BJ scan statistic satisfies the linear time subset scanning (LTSS) property [16], the subproblem

$$\max_{S \subseteq \mathcal{V}} \phi(\alpha, N_\alpha(S), N(S)) \quad (8)$$

can be solved in $O(|\mathcal{V}|)$ time, assuming that the entities $v \in \mathcal{V}$ have already been sorted by priority. Specifically, the LTSS property guarantees that the only subsets S with the potential to be optimal are those consisting of the top-n highest priority nodes $\{v_{(1)}, \cdots, v_{(n)}\}$, for some n between 1 and $|\mathcal{V}|$. In this case, a lower (more significant) p-value corresponds to a higher priority. Therefore, the relaxed version (6) of the original problem (5) can be solved in the time complexity $O(|\mathcal{V}| \times |U(\mathcal{V}, \alpha_{max})| + |\mathcal{V}| \log |\mathcal{V}|)$, where the additional $|\mathcal{V}| \log |\mathcal{V}|$ is required to sort the entities by priority. Based on the computational efficiency of the relaxed problem, we propose an efficient approximate algorithm by targeted seeding, iterative subgraph expansion, and relaxation. The proposed algorithm, described in Algorithm 1, will return a connected subgraph of the heterogeneous graph that approximately maximizes the proposed non-parametric scan statistic.

The time complexity of Algorithm 1 is dominated by the computation of the relaxed problem (6) for the updated subgraph $S \cup G$ in each iteration of graph expansion. This computation must be performed at most $KCZ = KC \log |\mathcal{V}|$ times, each requiring $O(|\mathcal{V}| \times |U(\mathcal{V}, \alpha_{max})|)$ time assuming that nodes have already been sorted by priority. Here K is the number of seed entities considered for each of the C entity types, and Z is the number of iterative subgraph expansions performed for each seed entity. Computing the empirical p-values is $O(|\mathcal{V}|T \log T)$, and sorting the nodes by priority is $O(|\mathcal{V}| \log |\mathcal{V}|)$. Therefore, the total computational complexity equals $O(KC \times |U(\mathcal{V}, \alpha_{max})| \times |\mathcal{V}| \log |\mathcal{V}| + |\mathcal{V}|T \log T)$. Furthermore, we note that $|U(\mathcal{V}, \alpha_{max})|$ can be considered a constant, since only at most $\alpha_{max}T$ distinct p-values less than α_{max} will be generated by the empirical p-value estimation method described above, and thus the algorithm scales as $O(|\mathcal{V}| \log |\mathcal{V}|)$. For our civil unrest and

Algorithm 1 Non-Parametric Heterogeneous Graph Scan

Input: $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f, \phi)$
Output: The most anomalous subgraph $S^\star$
Obtain "sensor" network $\mathcal{H} = (\mathcal{V}, \mathcal{E}, p)$ as above;
Set $\alpha_{max} = 0.15$, $K = 5$, $Z = \log|\mathcal{V}|$, and $S^\star = \emptyset$;
for $(k, c) \in [1, \cdots, K] \times [1, \cdots, C]$ **do**
  Select seed node v_0 from $\mathcal{V}_c$, where v_0 is the kth highest-priority node of that type;
  Set $S = \{v_0\}$;
  for $z \in [1, \cdots, Z]$ **do**
    Set $G = \{v \mid \exists e \in S, v \notin S, (v, e)$ or $(e, v) \in \mathcal{E}\}$;
    Obtain the highest-scoring subset $B \subseteq S \cup G$, where $S \subseteq B$, by solving the relaxed problem (6);
  end for
  if $B - S \neq \emptyset$ **then**
    Set $S = B$;
  else
    Break;
  end if
end for

rare disease detection experiments below, we have $\alpha_{max}T = (.15)(215) \approx 32$.

In addition to its computational efficiency, our proposed algorithm also has two nice theoretical properties, as follows:

THEOREM 2 (OPTIMALITY). *Consider the simplified case in which we fix α instead of allowing α to vary between 0 and $\alpha_{\max}$. Let $S^\star = \arg\max_S F_\alpha(S)$ denote the optimal connected subgraph. Assume that K is set sufficiently large to select some $v \in S^\star$ as a seed entity, and Z is set greater than the diameter of $S^\star$. If $S^\star$ satisfies the property that there is no "break-tire" entity $v \in S^\star$ (i.e., a node v with p-value greater than α and whose deletion will break the connectivity of $S^\star$), then Algorithm 1 is guaranteed to identify the optimal connected subgraph $S^\star$.*

PROOF. The BJ scan statistic satisfies three intuitive properties: 1) monotonically increasing with respect to N_α; 2) monotonically decreasing with respect to N and α; and 3) convex. Therefore, if $S^\star$ contains no "break-tire" entities, we know that two properties hold: a) $S^\star$ consists entirely of p-values less than or equal to α, and b) no neighbor of $S^\star$ has p-value less than or equal to α. Property a) holds since any leaf node with p-value greater than α could be deleted without disconnecting $S^\star$, increasing the score. Property b) holds since any neighbor with p-value less than or equal to α could be added without disconnecting $S^\star$, increasing the score. Now, when $v \in S^\star$ is selected as a seed entity, each successive graph expansion will add all and only those neighbor entities with p-values less than or equal to α. If Z is greater than the diameter of $S^\star$, the iteration continues until no remaining neighbors have p-values less than or equal to α, at which point $S = S^\star$. Therefore Algorithm 1 is guaranteed to identify the optimal connected subgraph $S^\star$. □

Note that another popular nonparametric scan statistic, the Higher Criticism (HC) statistic [7], also satisfies the above three properties and Theorem 2 still holds. Our preliminary analysis (omitted due to space limitations) demonstrates that the BJ statistic outperforms the HC statistic for heterogeneous graphs and hence we focus on the BJ statistic for the remainder of our study. We also show an interesting connection to the percolation-based scan statistic defined by Arias-Castro et al. [2], which provides nice asymptotic decision-theoretical properties.

THEOREM 3 (PERCOLATION-BASED SCAN STATISTIC). *As in Theorem 1, we consider the simplified case with fixed α, and again assume that K and Z are set sufficiently large. Following [2], let $F_\alpha(S) = |S|$ if all p-values in S are less than or equal to α, and 0 otherwise. Algorithm 1 is guaranteed to find the optimal subgraph $S^\star = \arg\max_S F_\alpha(S)$.*

PROOF. As in Theorem 2, we know that two properties hold: a) $S^\star$ consists entirely of p-values less than or equal to α, and b) no neighbor of $S^\star$ has p-value less than or equal to α. Property a) holds since the inclusion of any p-value greater than α would reduce $F_\alpha(S)$ to zero. Property b) holds since any neighbor with p-value less than or equal to α could be added without disconnecting $S^\star$, increasing $F_\alpha(S)$ by 1. The remainder of our proof proceeds identically to the proof of Theorem 2; note that we do not need the additional assumption that no "break-tire" entities exist, since this follows directly from the definition of the percolation-based scan statistic $F_\alpha(S)$. □

Note that we have chosen to use the BJ nonparametric scan statistic because of its superior detection power, in comparison with other test statistics such as the percolation-based scan. Nevertheless, this theorem shows the applicability of our work to percolation-based approaches as well. The computational cost of maximizing BJ over connected subgraphs is high, but we have proposed an efficient, approximate algorithm to address this issue. The approximation quality has also been validated by extensive experiments on real-world datasets, as described below.

4. EXPERIMENTS

This section evaluates the effectiveness and efficiency of the proposed NPHGS approach based on comprehensive experiments on four countries' Twitter data. We considered the detection and forecasting of civil unrest events such as protests and strikes, and detection of rare disease (hantavirus) outbreaks as two case study scenarios, but the proposed techniques can also be directly applied to other applications, such as the detection and forecasting of human rights violations and financial crises.

4.1 Experiment Design

Datasets: We randomly collected ten percent of all the raw Twitter data from June 1, 2012 to June 30, 2013 across four countries: Argentina, Chile, Columbia, and Ecuador. Sampling was conducted at the tweet level, instead of user level. The civil unrest event labels, called Golden Standard Report (GSR), were collected and confirmed from the local newspapers that are accessible from internet. The collected tweet volume, news sources, and number of civil unrest events reported for each country are summarized in Table 2. An example of a labeled GSR event is: (PROVINCE = "El Loa", COUNTRY = "Chile", DATE = "2012-05-18", TITLE = "A large-scale march was staged by inhabitants of the northern city of Calama, considered the mining capital of Chile, who demand the allocation of more resources to copper mining cities", NEWS-LINK = "http://www.pressenza.c

om/2012/05/march-of-dignity-in-mining-capital-of-chile/"). For rare disease outbreaks, we considered hantavirus outbreaks in Chile as a case study, because there was a spread of hantavirus outbreaks there last year that greatly threatened public safety and stability. The outbreak labels were reported by Chilean Ministry of Health [26] and local news reports. Specifically, there were 17 rare Hantavirus disease outbreaks in more than eight different states from January 1, 2013 to June 30, 2013. We post-processed both the civil unrest and the disease outbreak data to create binary variables representing whether or not a GSR event occurred in each province (state) for each date.

Country	# tweets	News sources	# events
Argentina	48 million	Clarin; La Nacion; Infobae	302
Chile	25 million	La Tercera; Las Ultimas Noticias; El Mercurio	216
Colombia	37 million	El Espectador; El Tiempo; El Colombiano	251
Ecuador	12 million	El Universo; El Comercio; Hoy	149

Table 2: Description of civil unrest data by country

Data Preprocessing: After we collected the raw tweets, several preprocessing steps were conducted for our proposed approach and all the comparison methods, including:

1) **Vocabulary Generation**: We first generated a vocabulary of ∼1000 terms related to civil unrests and a vocabulary of 25 terms related to hantavirus from domain experts;

2) **Content Filtering**: Only the raw tweets that match more than two terms from the vocabulary were preserved;

3) **Tweet Geocoding**: We implemented a geocoding library for tweets based on three major rules with priorities. For each tweet, we first searched for location and landmark mentions in the tweet text, then for geotags that are available if the user enabled the geocoding function in his/her phone, and finally for location information from the user's profile. The first location information identified was returned as the geographic location of this tweet.

Comparison Methods: We compared our proposed NPHGS approach with five existing representative methods, including Spatio-Temporal (ST) Burst Detection [13], Graph Partition [24], Earthquake Detection [19], Real-World (RW) Event Identification [3], and Geographic Topic Modeling [25]. The first method returns an alert for each spatio-temporal burst that is detected. The second method applies wavelet analysis to build signals for individual words, and then clusters the signals based on their auto-correlations using modularity based graph partitioning. Each cluster is returned as an alert, with the most frequent location in the related tweets identified as the event location. The third method classifies tweets based on predefined features, and develops a probabilistic spatiotemporal model to identify the geographic center and date of the event. The fourth method considers the framework of online clustering for event detection, but designs a new similarity function in order to capture features related to time, topical coherence, and social interactions. The fifth method detects geographic topics day by day, each of which is returned as an alert.

Implementations of the first and the fifth methods were obtained from the authors, the second method was replicated under the authors' instructions, and the other two methods were implemented based on the published papers [3, 19]. We strictly followed the strategies recommended by the authors in their papers to select features and estimated the related model parameters. Specifically, the parameters of Earthquake Detection, Graph Partition, and Spatio-Temporal Burst Detection were trained using cross validation. The Twitter data from June 2012 to December 2012 were used as training data, data from January 2013 to April 2013 were considered as the test dataset for the detection and forecasting of civil unrest events, and data from January 1, 2013 to June 30, 2013 were considered as the test dataset for the detection and forecasting of rare disease outbreaks. As an unsupervised approach, the geographic topic model has two major parameters, including the numbers of geographic regions and topics. These two parameters were predefined based on our interpretation of the data distribution. Given the statistics of GSR event labels, the numbers of cities with high frequencies of civil unrest events in the four countries are mostly smaller than 20. The two parameters were set to 20 and 2, respectively. RW-Event has a number of parameters, including the number of most frequent cluster terms, the parameters related to the incremental clustering algorithm, and the parameters related to the classification model used. In our implementation, we considered linear weighted support vector machine (SVM) to handle the issue of unbalanced class labels. We used 10-fold cross validation to identify the best combination of all the related parameters.

Our NPHGS and Baseline Homogeneous Graph Scan Methods: Our proposed NPHGS is designed in a nonparametric statistical framework and the specification of parameters is hence relatively straightforward. Values of α_{max} and the number of seed entities K were set to 0.15 and 5, respectively. We observed that performance of our method is not sensitive to the settings of these two parameters. In addition to the above five comparison methods, we also compared our proposed NPHGS with four different homogeneous versions of NPHGS, including tweet, location, keyword, and user level homogeneous networks. In order to make fair comparisons, for each homogeneous graph, we defined a connection between two entities if they have direct relationships or if they share some neighbors in the heterogeneous graph. For example, two tweets are connected if they have retweet or reply relationships, or if they are connected to the same geographic location or the same terms. In each case, we assume that the system alerts when detecting a subgraph with score $F(S)$ above some threshold, allowing us to consider the tradeoffs between false positive rate and the other four performance metrics defined below by varying the alert threshold.

Performance Metrics: This study focuses on the evaluation of both event detection and forecasting for different methods. The related performance metrics include: 1) false positive rate (FPR), 2) true positive rate (TPR) for forecasting, 3) true positive rate for both detection and forecasting, 4) average lead time for forecasting, and 5) average lag time for detection. For each method, the reported alerts are structured as tuples of (date, location), where "location" is defined at the province level.

For each gold standard event, we determine whether the method: a) Had an alert in that province from 1 to 7 days before the event (such events are considered to be "success-

Method	FPR (FP/Day)	TPR (Forecasting)	TPR (Forecasting & Detection)	Lead Time (Days)	Lag Time (Days)	Run Time (Hours)
ST Burst Detection	0.65	0.07	0.42	1.10	4.57	30.1
Graph Partition	0.29	0.03	0.15	0.59	6.13	18.9
Earthquake	0.04	0.06	0.17	0.49	5.95	18.9
RW Event	0.10	0.22	0.25	0.93	5.83	16.3
Geo Topic Modeling	0.09	0.06	0.08	0.01	6.94	9.7
NPHGS (FPR=.05)	0.05	0.15	0.23	0.65	5.65	38.4
NPHGS (FPR=.10)	0.10	0.31	0.38	1.94	4.49	38.4
NPHGS (FPR= .15)	0.15	0.37	0.42	2.28	4.17	38.4
NPHGS (FPR=.20)	0.20	0.39	0.46	2.36	3.98	38.4

Table 3: Comparison between NPHGS and Existing Methods on the civil unrest datasets

Method	FPR (FP/Day)	TPR (Forecasting)	TPR (Forecasting & Detection)	Lead Time (Days)	Lag Time (Days)
ST Burst Detection	0.57	0.25	0.63	1.13	3.81
Graph Partition	0.57	0.06	0.19	0.19	6.10
Earthquake	0.92	0.13	0.19	0.75	5.69
RW Event	0.40	0.19	0.41	0.43	4.91
Geo Topic Modeling	0.43	0.19	0.50	0.62	4.31
NPHGS (FPR=.05)	0.05	0.20	0.78	0.71	2.44
NPHGS (FPR=.10)	0.10	0.22	0.85	0.76	1.90
NPHGS (FPR= .15)	0.15	0.25	0.93	0.80	1.36
NPHGS (FPR=.20)	0.20	0.29	0.94	0.82	1.24

Table 4: Comparison between NPHGS and Existing Methods on the Hantavirus dataset

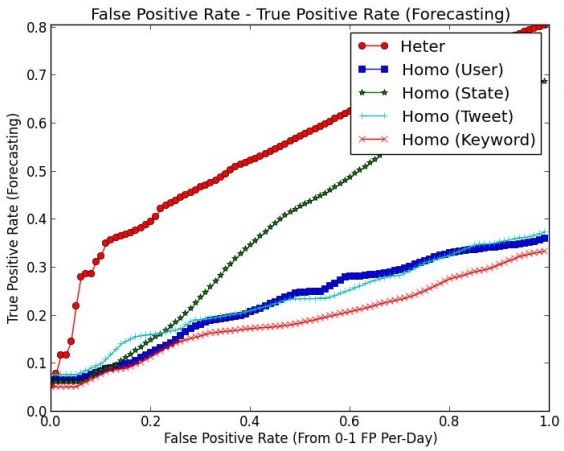
fully predicted" at the given threshold). In this case, we record the number of days of lead time for that event based on the earliest such alert; b) Did not have an alert in that province from 1 to 7 days before the event, but did have an alert in that province from 0 to 7 days after the event (such events are considered to be "successfully detected" at the given threshold). In this case, we record the number of days of lag time for that event based on the earliest such alert; or c) Did not have an alert in that province between 7 days before and 7 days after the event (such events are considered to be "undetected" at the given threshold).

Based on the preceding results, we compute how many alerts were triggered that were not within the 7-day window before and after any event (this is the number of "false positives" at the given threshold). Now, as a function of the number of false positives (we scale this by time, e.g., "1 FP per day"), we can determine: 1) Proportion of gold standard events that were successfully predicted; 2) Proportion of gold standard events that were successfully predicted or detected (this is one minus the proportion of undetected gold standard events); 3) Average lead time for all gold standard events: higher is better. Note that we average in a "0" lead time for each event that is not successfully predicted; and 4) Average lag time for all gold standard events: lower is better. Note that we average in a "0" lag time for each event that is successfully predicted OR is detected on the event day, and for undetected events we average in a "7" day lag time. Finally, we note that the maximum false positive rate that we consider, 1 FP per day, is non-trivial because we consider each unique combination of a province and a date as one binary variable, and thus the number of potential false positives per day could be up to the total number of provinces in a given country.

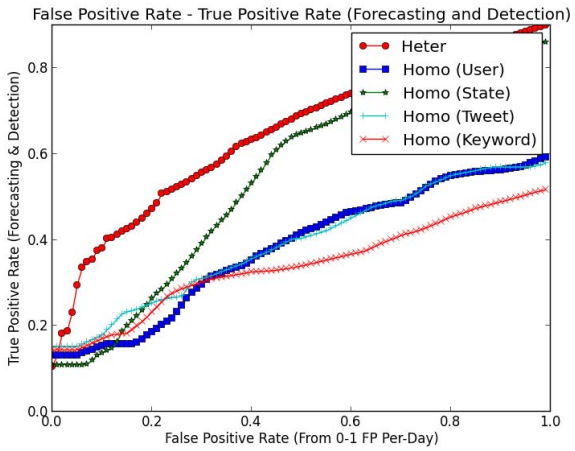
4.2 Comparison between NPHGS and Existing Methods

Table 3 presents the comparison between the proposed NPHGS approach and five competing methods for the task of forecasting civil unrest events. All measurements were averaged over the results of the four tested countries. For NPHGS, we show the performance metrics at various false positive rates. For comparable false positive rates, NPHGS achieved much higher forecasting TPR and detection TPR than all competing methods. The average lead time of NPHGS was at least one day greater than the other methods, and average lag time was consistently smaller than the other methods by 1 to 2 days. Run time of NPHGS was comparable to other methods but slightly higher because it considers all node types in the heterogeneous network rather than just tweets. We note that the true positive rates of all tested methods were lower than 50%, perhaps because some GSR events did not produce strong signals in the noisy Twitter data, or because an alert was only considered "correct" if it matched both the date and location of a GSR event.

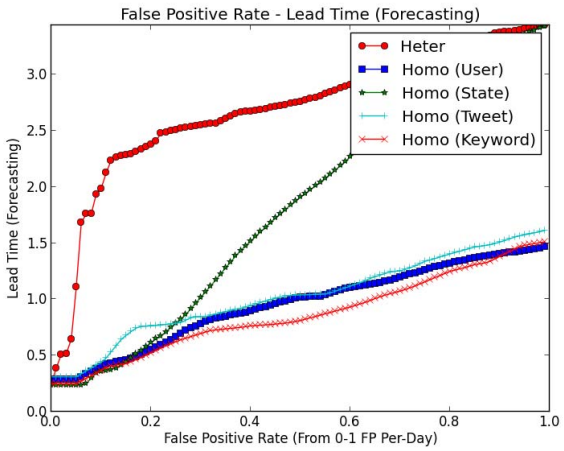
Table 4 presents the comparison results for the task of detecting hantavirus outbreaks. The results indicate consistent patterns as observed in Table 3. For comparable false positive rates, NPHGS outperformed the competitive methods in all the metrics. Specifically, NPHGS achieved 10% to 30% higher forecasting TPR and detection TPR than all competing methods. The lead time of NPHGS was ten percent to twenty percent (0.2 to 0.5 days) greater than those of competitive methods, and the lag time of NPHGS was 50% (1 to 2 days) lower than those of competing methods. By comparing Table 3 and Table 4, we observe that the true positive rate of NPHGS for forecasting alone on the Hantavirus disease data was 10% lower than that of NPHGS on the civil unrest data. Consistently, the lead time of NPHGS



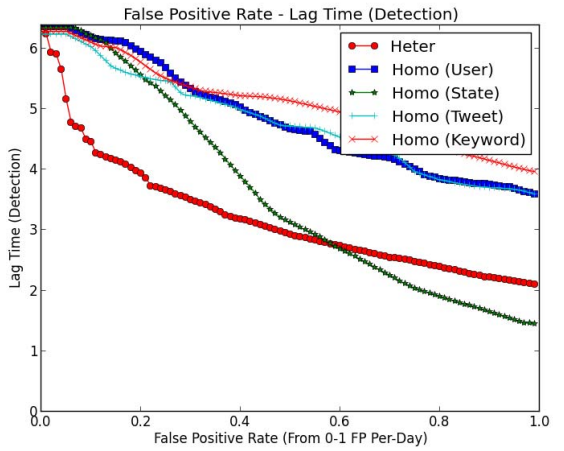
(a) FPR vs. TPR (Forecasting)



(b) FPR vs. TPR (Forecasting and Detection)



(c) FPR vs. Lead Time (Forecasting)



(d) FPR vs. Lag Time (Detection)

Figure 3: Comparison between heterogeneous and homogeneous graph scans. True positive rates for forecasting and detection, forecasting lead time, and detection lag time, all measured as a function of the false positive rate (from 0 to 1 false positive per day).

on the Hantavirus disease data is 1 day less than that of NPHGS on the civil unrest data. One potential interpretation is that civil unrest events tend to have stronger signals in Twitter data leading up to the event, since inflamed public sentiments and emotions, as well as advance planning and organization of strikes and protests, may be visible in the Twitter data. Nevertheless, our results demonstrate that we can achieve very early detection of emerging rare disease outbreaks. Table 4 shows that at a reasonably low false positive rate of 0.2 FP/day, NPHGS has detection lag time of only 1.24 days, which is better than the typical 3 to 4 days lag time using traditional public health surveillance data.

4.3 Comparison between NPHGS and homogeneous graph scans

This section compares NPHGS and different versions of homogeneous graph scan methods on the civil unrest datasets, with results shown in Figure 3. We applied the same framework of NPHGS to homogeneous networks of different entity types as the baseline methods. Hence, we label NPHGS as "Heter" in this case, and label the baseline methods as "Homo-(entity type)". The results in Figure 3 clearly demonstrate that NPHGS consistently outperforms all the homogeneous graph scan methods for all performance metrics. When the false positive rate was low (e.g., between 0 and 0.2 FP per day), NPHGS achieved huge (∼30%) absolute improvements in TPR, provided two days of additional lead time for forecasting, and detected events two days earlier.

5. CONCLUSIONS

This paper presents a nonparametric approach to the problem of event detection and forecasting for heterogeneous social media graphs. The direct statistical modeling of heterogeneous relationships between entities and attributes is very complicated and computationally challenging. Our work avoids this complicated modeling process by transforming the heterogeneous graph into a "sensor" network, where we convert the heterogeneous entities and attributes into empir-

ical p-values (using a novel, two-stage empirical calibration procedure). We then use a novel graph scan algorithm to maximize a non-parametric scan statistic over subgraphs, enabling early detection and advance forecasting of emerging societal events. In addition to evaluation of the theoretical properties of our method, we perform extensive experiments on real Twitter data. Our empirical results demonstrate that we can effectively forecast civil unrest events and achieve very early detection of rare disease outbreaks, outperforming competing methods by a substantial margin for both detection (power and timeliness) and forecasting (accuracy and lead time). For future work, we plan to extend NPHGS to do storytelling and causality analysis, since NPHGS is able to provide rich information related to ongoing or new events, such as the users, geographical locations, and key terms involved. In addition, we plan to extend NPHGS to a Bayesian framework so that rich domain knowledge can be naturally integrated.

6. ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation grants IIS-0916345, IIS-0911032, and IIS-0953330. Additional support was provided by the John D. and Catherine T. MacArthur Foundation.

7. REFERENCES

[1] How fast the news spreads through social media. In *http://blog.sysomos.com/2011/05/02/how-fast-the-news-spreads-through-social-media/*, 2012.

[2] Ery Arias-Castro and Geoffrey R. Grimmett. Cluster detection in networks using percolation. *Bernoulli*, vol. 19, no. 2, pages 676–719, 2013.

[3] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on Twitter. In *AAAI-ICWSM*, 2011.

[4] Robert H. Berk and Douglas H. Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Mathematical Reviews*, pages 47–59, 1979.

[5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on Twitter. In *WWW*, pages 675–684, 2011.

[6] Lise Getoor. Tutorial: Representation, inference and learning in structured statistical models. In *NIPS*, 2012.

[7] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, vol. 32, no. 3, pages 962–994, 2004.

[8] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287, 2010.

[9] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. Parameter free bursty events detection in text streams. In *VLDB*, pages 181–192, 2005.

[10] Jon Kleinberg. Bursty and hierarchical structure in streams. In *KDD*, pages 91–101, 2002.

[11] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas. Tsioutsiouliklis. Discovering geographical topics in the Twitter stream. In *WWW*, pages 769–778, 2012.

[12] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, vol. 26, issue 6, pages 1481–1496, 1997.

[13] Theodoros Lappas, Marcos R. Vieira, Dimitrios Gunopulos, and Vassilis J. Tsotras. On the spatiotemporal burstiness of terms. In *PVLDB*, vol. 5, issue 9, pages 836–847, 2012.

[14] Edward McFowland III, Skyler Speakman, and Daniel B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, vol. 14, pages 1533-1561, 2013.

[15] Daniel B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, vol. 8, no. 20, 2009.

[16] Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B*, vol.74, pages 337–360, 2012.

[17] Daniel B. Neill and Jeff Lingwall. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance*, 2007.

[18] Stanislav Nikolov and Devavrat Shah. A nonparametric method for early detection of trending topics. In *WIDS*, 2012.

[19] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[20] Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. Event detection and tracking in social streams. In *AAAI-ICWSM*, 2009.

[21] Charu C. Aggarwal and Karthik Subbian Event Detection in Social Streams In *SDM*, pages 624–635, 2012.

[22] Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Newsstand: a new view on news. In *GIS*, no. 18, pages 1–10, 2008.

[23] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *CIKM*, pages 2541–2544, 2011.

[24] Jianshu Weng and Bu-Sung Lee. Event detection in Twitter. In *AAAI-ICWSM*, 2011.

[25] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256, 2011.

[26] http://epi.minsal.cl/vigilancia-epidemiologica/enfermedades-de-notification-obligatorial/vigilancia-hantavirus

[27] http://www.clips.ua.ac.be/pattern

[28] http://en.wikipedia.org/wiki/Klout