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

Article Published by Chen, Feng & B. Neill, Daniel. (2014)

Paper Review in Data Analytics I Seminar by Eduardo Salvador Rocha (2018)

Abstract

The “Data Analytics Seminar I” of the Winter Semester 2017/2018 consisted in an overview of different algorithms used for real-life Event Detection on social media microblogs based on user’s reports.

This paper tries to summarize the work done by Chen F. and Neill D. on the usage of Non-Parametric Heterogeneous Graph Scan (NPHGS) for modeling the implicit heterogeneous network structure in twitter. The NPHG Scan uses the information about tweets, users and their relation in order to construct a network in which each node report an empirical p-value of anomalousness against its surroundings. The algorithm uses the Berk Jones test to determine which nodes have a statistically different p-value, and in conjunction with a scanning algorithm, returns a subgraph (cluster) which is defined to be the detected event of interest. The paper also shows two applications using Twitter data on civil unrest and rare disease outbreaks detection, and present empirical evaluations illustrating the effectiveness and efficiency of the proposed algorithm against homogeneous version of the networks and more popular techniques like burst detection, topic modeling and clustering (methods that are delimited to only certain information such as terms and geographic locations).

Introduction

Social microblogs became relevant for investigation due billions of users globally sharing their daily observations and thoughts in real time at almost no cost. In many cases such posts carries information about real-life events that is spreaded earlier and faster than in traditional media.

This type of information can be used either for constantly update an ongoing event (such as natural disasters, traffic, accidents) or to identify patterns in sentiments and opinions about specific topics and in some cases forecast about their final results (such as elections, stock market indexes movements, “uprising” of civil unrests, manifestations). This type of analysis is known as Event Detection.

Due the huge amount of data, event detection from Twitter streams have drawn techniques from different fields, including machine learning and data mining, natural language

processing, information extraction, text mining, and information retrieval[1] in order to provide a wider toolbox for analysis.

Related work

Different algorithms can be classified based on the type features used for the event detection. General-domain event detection methods 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 and require extensive parameter tuning.

Domain-Specific methods on the other hand, tries to identify events using directed dictionaries (terms related to the wanted event) and twitter content (location, text, date). In this area, there are three major approaches taken including burst detection, geographic topic modeling and clustering. Burst detection searches for space-time regions aggregated by the counts of predefined terms that are abnormally high compared with the counts outside the regions. Geographic topic model-based approaches estimate language distributions (over a predefined vocabulary) that are distinct in some geographic regions. Clustering-based approaches use features related to text content and link information to search for novel clusters of documents or terms using predefined similarity metric, social similarity for documents, or autocorrelations and co-occurrences for terms.

TABLE 2. Summary of Detection Techniques and Feature Representations.

References	Detection techniques	General features	Twitter-specific features
Sankaranarayanan et al. (2009)	Naive Bayes classifier and online clustering	Term vector	Hashtags and timestamps
Phuvipadawat and Murata (2010)	Online clustering	Term vector, proper nouns (conventional NER)	Hashtags, #followers, #retweets and timestamps
Petrović et al. (2010)	Online clustering (based on locality sensitive hashing)	#tweets, #users and entropy of messages	–
Becker et al. (2011a)	Online clustering and support vector machine classifier	Term vector	Hashtags, multi-word hashtags with special capitalization, retweets, replies and mentions.
Long et al. (2011)	Hierarchical divisive clustering	Word frequency and entropy	Probability of word occurring in hashtags
Weng and Lee (2011)	Discrete wavelet analysis and graph partitioning	Individual words	–
Cordeiro (2012)	Continuous wavelet analysis and latent Dirichlet allocation	–	Hashtag occurrences
Popescu and Pennacchiotti (2010)	Gradient boosted decision trees	Correlation of target events (or entities) with the Web and traditional news media	Proportion of nouns, verbs, questions, bad words, etc.; #tweet, #retweets, #replies, #tweets per user, hashtags; proportion of tweets and hashtags involving buzziness, sentiment, controversy

[1] Atefeh F., and Khreich W. (2015), A Survey of Techniques for Event Detection in Twitter, Computational Intelligence, 31, 132–164, doi: 10.1111/coin.12017

References	Detection techniques	General features	Twitter-specific features
Popescu et al. (2011)	Gradient boosted decision trees	Part-of-Speech tagging and regular expressions (in addition to the features used by Popescu et al. (2011))	Relative positional information, length of snapshot, category, language (in addition to the features used by Popescu et al. (2011))
Benson et al. (2011)	Factor graph model and conditional random fields	Term vectors for artist names (extracted from Wikipedia) and for city venue names.	Word shape, patterns for emoticons, time references, venue types
Lee and Sumiya (2010)	Statistical modeling of normal crowd behavior	–	#Tweet, #Crowd, #MovingCrowd based on geotags
Sakaki et al. (2010)	Support vector machine classifier	–	#Words, #keywords and the words surrounding users query
Becker et al. (2011)	Recursive query construction	Term frequency and co-location	Hashtags and URL
Massoudi et al. (2011)	Generative language modeling		Emoticons, post length, shouting, hyperlinks, capitalization, recency, #reposts and #followers
Metzler et al. (2012)	Temporal query expansion technique		Burstiness score based on the frequency of query term occurrence
Gu et al. (2011)	Event modeling (ETree)	Term vector and n-gram models	Replies to tweets

References	Type of event		Detection method		Detection task		Application
	Specified	Unspecified	Supervised	Unsupervised	NED	RED	
Sankaranarayanan et al. (2009)		x	x	x	x		Breaking-news detection
Phuvipadawat and Murata (2010)		x		x	x		Breaking-news detection
Petrović et al. (2010)		x		x	x		General (unknown) event detection
Becker et al. (2011a)		x	x	x	x		General (unknown) event detection
Long et al. (2011)		x		x	x		General (unknown) event detection
Weng and Lee (2011)		x		x	x		General (unknown) event detection
Cordeiro (2012)		x		x	x		General (unknown) event detection
Popescu and Pennacchiotti (2010)	x		x		x		Controversial news events about celebrities
Popescu et al. (2011)	x		x		x		Controversial news events about celebrities
Benson et al. (2011)	x		x			x	Musical event detection
Lee and Sumiya (2010)	x			x	x		Geosocial event monitoring
Sakaki et al. (2010)	x		x		x		Natural disaster events monitoring
Becker et al. (2011)	x		x			x	Query-based event retrieval
Massoudi et al. (2011)	x			x		x	Query-based event retrieval
Metzler et al. (2012)	x			x		x	Query-based structured event retrieval
Gu et al. (2011)	x			x		x	Query-based structured event retrieval

Each of the aforementioned methods only exploits partial information from user's posts and do not include the inherent heterogeneity of a social media entities and their inter-relationships or even the entity's attributes (which are also heterogeneous).

This type of relations can be modeled through graphs which are mathematical structures used to model pairwise relations between objects. In this context, a graph is made up of vertices, nodes, or points which are connected by edges, arcs, or lines[2] (Fig 1 a). A graph may be undirected, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be directed from one vertex to another. [In the case] of twitter, their heterogeneous network as a directed graph is exemplified in figure 1b.

2

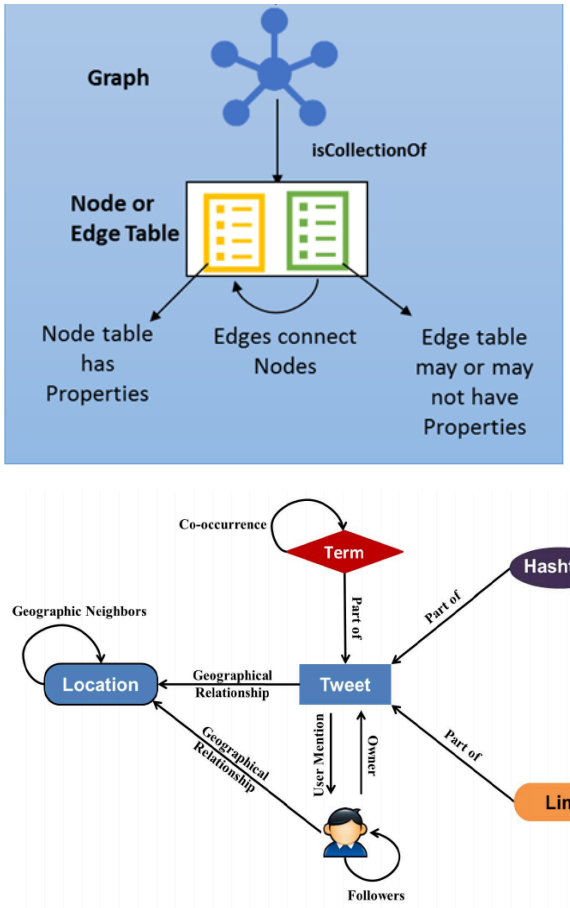


Image 1. Example of a Graph (SQL). Entity diagram for twitter data Modeling

Nevertheless, modeling the entire social media graph for event detection is computationally expensive and the time to scan the entire graph can be exponential to the total number of graph nodes, and at the end, it can lead to overfitting.

Another aspect to consider while analyzing social media is that the language used is highly informal, ungrammatical and dynamic, and this limits the use of traditional natural language processing (NLP) techniques, and thus motivates the use of a nonparametric statistical framework to provide more accurate detection and forecasting. A non-parametric structure does not infer a priori about the distribution of the information, and the model is determined directly from the input data; it does not imply a complete lack of parameters.

Considering all the previous mentioned, the authors proposed to model the heterogeneous directed graph of the twitter data for event detection. They use a two level calibration step which returns a p-value of “abnormality” of each node versus its neighborhood, and then uses a statistical test and a given threshold, in order to determine if it is “abnormal enough” to be considered a detected event.

The information considered from twitter is depicted bellow:

Object Type	Features
User	# tweets, # retweets, # followers, #followees, #mentioned_by, #replied_by, diffusion graph depth, diffusion graph size
Tweet	Klout, sentiment, replied_by_graph_size, reply_graph_size, retweet_graph_size, retweet_graph_depth
City, State, Country	# tweets, # active users
Term	# tweets
Link	# tweets
Hashtag	# tweets

Methodology

The task proposed is to analyze the heterogeneous graph from the Twitter data, and identify sub-graphs[3] (**S**) of events of interest using a nonparametric scan statistic **F**(S).

$$\max_{S \subseteq \mathcal{V}: S \text{ is connected}} : \max_{\alpha \leq \alpha_{max}} F_\alpha(S)$$

This task is achieved in the paper by the following steps:

1. Convert all nodes and node's features of the heterogeneous graph into proportions (probabilities)
2. Define an statistical measure to determine the probability that a selected node represent an event of interest
3. Use an algorithm that scans the converted network, test each node against its neighborhood using the defined statistical test, and return the identified sub-graphs as the detected event.

[3] ***S*** is defined as a subset of vertices and edges from the Graph ***G***

A sensor network “H” drawn from heterogeneous network “G”

The sensor network H = (V, E, p) consists of a set of nodes, edges and p-values obtained from converting the heterogeneous graph into proportions with uniform values in the range of [0,1]. This procedure allows to consider and compare on the same scale 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. The calibration (obtention of p-values) is performed in a two-step procedure which is preferred over a one-stage calibration process due the later would be biased toward detecting nodes with more features.

In order to calculate the p-values, it is necessary to estimate a baseline distribution for each attribute that characterizes its behavior when there is no event occurring, or in other words, define what is the “normality” of an attribute when no event is occuring.

To construct the baseline, a collection of historical observations are required to be used as training samples. For entities with sufficient observations (locations, regular users, and existing keywords, hashtags, and links) such baseline is easy to set, but entities with insufficient (new users, newly occurring keywords, hashtags, and links) or single observations (mainly include tweets) need to be calibrated based on all historical records of the same entity 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 are defined for a given node, they are tested under the null hypothesis to determine the probability that a randomly selected node would have an historical node value (v^t) greater than or equal to the current node observation (v):

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

$p_d(v)$ = empirical value
$f_c(v)$ = feature vector
$f_{c,d}(v)$ = *d-th* component of the feature vector

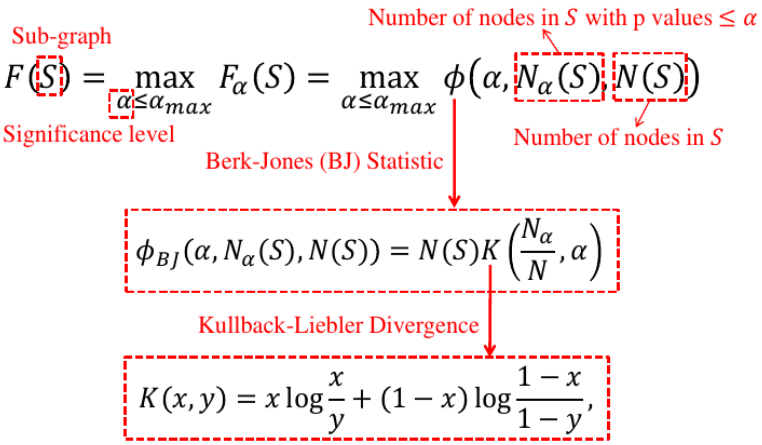
As each node has multiple p-values, a single p-value is obtained by calculating the minimal of all the p-values versus historical data:

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

There is mentioned that this two-stage process is sufficiently flexible so that other p-value combination methods (such as Fisher’s method) could easily have been used instead of the minimum p-value, while still returning values bounded to [0,1]

Non-Parametric Testing

To determine an anomalous connected subgraphs, a spatial scan (Kulldorff's [12]) over the graph H is performed using the non-parametric test F(S) o Berk Jones (statistic that is known to be optimal for detecting rare and weak signals when factors in a set are independent[4]):



In the paper, the level of α in the NPHGS is treated as an hyperparameter to be optimized, and they ended up using an α= 0.15

[4] The BJ statistic can be described as the log-likelihood ratio statistic for testing whether the empirical p-values are uniformly distributed on [0, 1]

Scan Algorithm over Graph H

Based on the proposed statistic, the detection of the most anomalous connected subgraph can be formalized as:

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

This optimization task contains an exhaustive search over all evolving subgraphs and the corresponding attributes and therefore is computationally infeasible, scaling exponentially with the number of subgraphs and attributes. An relaxed problem can be derived by assuming that the entities v∈V have been sorted by priority (a lower p-value corresponds to a higher priority), and by removing the connectivity constraint):

$$\max_{\alpha \in U(\mathcal{V}, \alpha_{max})} \max_{S \subseteq \mathcal{V}} \phi(\alpha, N_\alpha(S), N(S))$$ [5]

The scan algorithm then is required to:

Return a connected subgraph (S) by transforming the heterogeneous graph (G) in a sensor graph (H) with p-values, and approximately maximize the proposed nonparametric scan statistic using K number of seed entities for each of the C entity types, during Z iterative subgraph expansions performed for each seed entity.

Algorithm 1 Non-Parametric Heterogeneous Graph Scan

Input: $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f, \phi)$
Output: The most anomalous subgraph S^*
Obtain "sensor" network $\mathcal{H} = (\mathcal{V}, \mathcal{E}, p)$ as above;
Set $\alpha_{max} = 0.15$, $K = 5$, $Z = \log |\mathcal{V}|$, and $S^* = \emptyset$;
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 $G = \{v \mid \exists e \in S, v \notin S, (v, e) \text{ or } (e, v) \in \mathcal{E}\}$;
    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

[5] U(S, α_{max}) refers to the union of α and the set of distinct p-values less than α in S

Experiments

Data Gathering

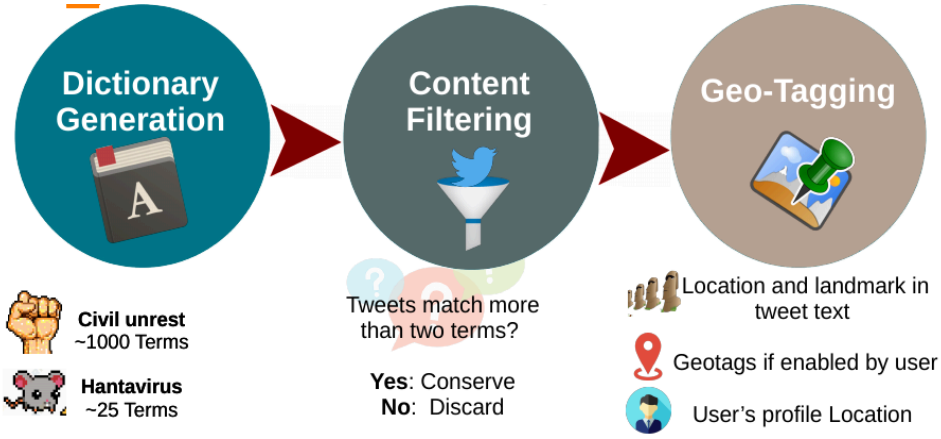
To test the accuracy of the algorithm, it was set to detect events of disease outbreaks and civil unrests. In order to obtain the correct dates of such incidents, information from local newspapers that are accessible from internet were collected and labeled (referenced as Golden Standard Reports)
Afterwards, data was gathered from Twitter in the interval of June 1, 2012 to June 30, 2013 across four countries: Argentina, Chile, Colombia, and Ecuador. The sampling was conducted at the tweet level, instead of user level. In order to obtain the events of outbreaks and civil-unrests on those countries over the period of time gathered.

Country	#Tweets	News Sources	# Events
Argentina	48 miion	Clarin; La Nacion; Infobae	302
Chile	25 milion	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

Data Pre-processing

The preprocessing steps conducted for the proposed approach and all the comparison methods, include:

- Vocabulary Generation: Approximately 1000 terms related to civil unrests and a vocabulary of 25 terms related to hantavirus from domain experts
- Content Filtering: Only the raw tweets that match more than two terms from the vocabulary were preserved
- Tweet Geocoding: Based on location and landmark mentions in the tweet text, geotags that are available if the user enabled the geocoding function in his/her phone, or the location information from the user's profile.



Comparison Methods

Alternative Algorithms

The NPHGS was compared against five existing representative methods (supervised and unsupervised), which were given directly for the authors or implemented directly from the papers (features and related model parameters were trained using cross validation):

- **Spatio-Temporal Burst Detection** [13]: returns an alert for each spatio-temporal burst that is detected
- **Graph Partition** [24]: wavelet analysis to build signals for individual words and use of modularity graph partitioning to cluster them
- **Earthquake Detection** [19]: classifies tweets based on predefined features, and develops a probabilistic spatiotemporal model to identify the geographic center and date of the event
- **Real-World Event Identification** [3]: online clustering with a new similarity function in order to capture features related to time, topical coherence, and social interactions
- **Geographic Topic Modeling** [25]: detects geographic topics day by day, each of which is returned as an alert. (unsupervised)

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.

Homogeneous Networks variants

In addition NPHGS was also compared against their homogeneous versions by tweet, location, keyword, and user level networks (four in total). Connections were made based on shared neighbors (retweet or reply relationships), or if they are connected to the same geographic location or the same terms.

Performance Metrics

For each labeled event by province, it was determined if an event was:

- “successfully predicted” if had an alert up to 7 days before the event (the time lapse before the date and the detection was denominated as *lead time*)
- “successfully detected” if had an alert up to 7 days after the event (the time lapse after the date and the detection was denominated as *lag time*)
- “undetected” if did not have an alert between 7 days before and 7 days after the event

Based on the previous findings, the following evaluations were determined:

- False positive rate (FPR) (“undetected” events scaled up to 1 FP per day)
- True positive rate (TPR) for forecasting
- True positive rate for both detection and forecasting
- Average lead time for forecasting (higher is better)
- Average lag time for detection (lower is better)

Results

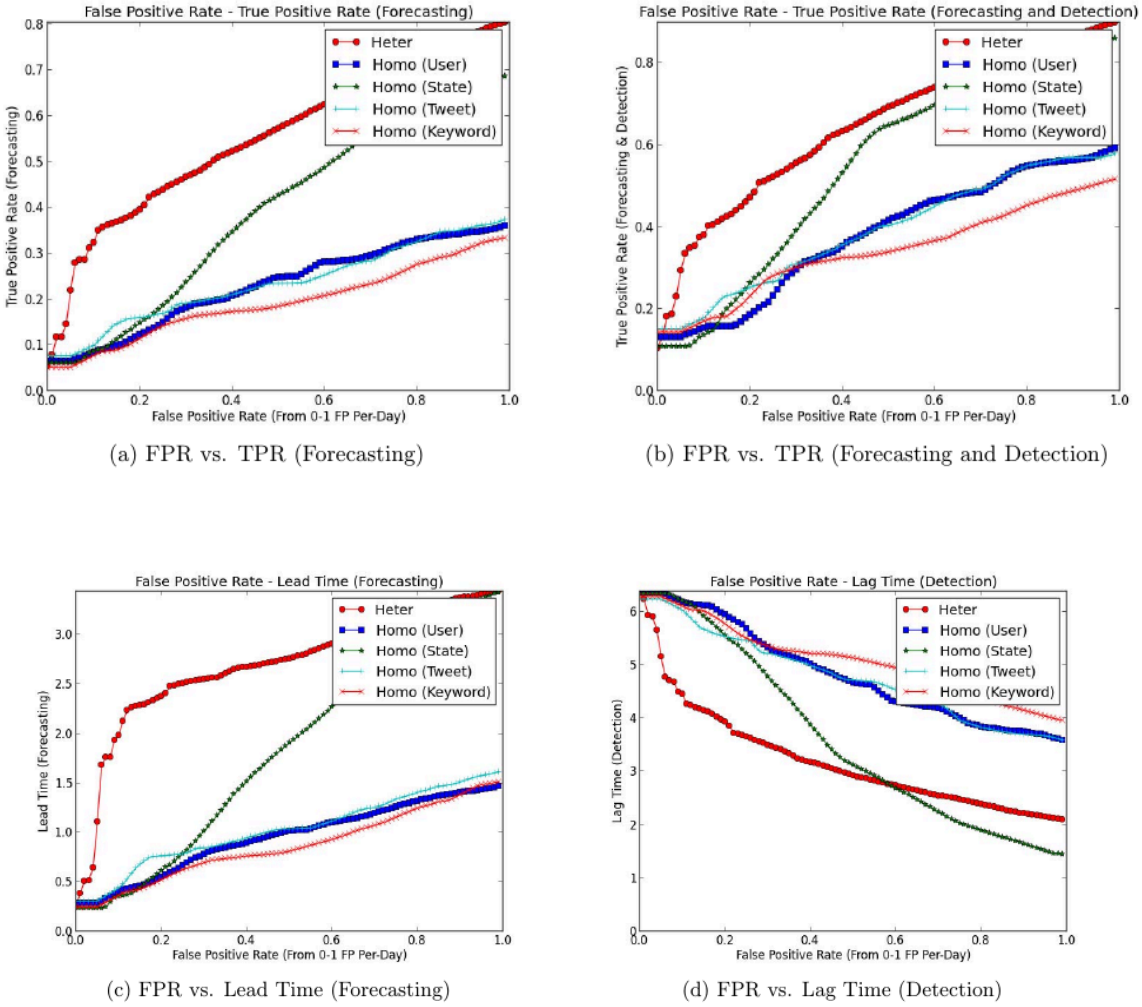
Method	FPR (FP/Day)	TPR (Forecasting)	TPR (Forecasting & Detection)	Lead Time (Days)	Lag Time (Days)	Run Time (Hours)
ST Burst Detection	0.65	0.07	0.42	1.10	4.57	30.1
Graph Partition	0.29	0.03	0.15	0.59	6.13	18.9
Earthquake	0.04	0.06	0.17	0.49	5.95	18.9
RW Event	0.10	0.22	0.25	0.93	5.83	16.3
Geo Topic 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

Both tables compare the NPHGS and five competing methods for the task of forecasting civil unrest events. All measurements were averaged over the results of the four tested countries. It can be observed that NPHGS outperformed the competitive methods in all the metrics (10% to 30% higher forecasting TPR and detection TPR) and that the average lead time achieved was at least one day greater than the other methods; also in the case of the average lag time it was consistently smaller than the other methods by 1 to 2 days.



(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).

In the comparison between NPHGS and different versions of homogeneous graph scan methods on the civil unrest dataset it is shown that NPHGS consistently outperforms all them 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.

Discussions and Conclusions

NPHGS have better event detection ratio than five existing representative methods, and homogeneous version of the heterogeneous network; but on the other hand the computational time is almost double. NPHGS also have a higher detection window (i.e. better forecasting time and lag time), but should be noted that in the case of civil unrests, the true positive rate is less than 50% for all the algorithms, suggesting that the signals for those type of events is weaker than the ones for disease outbreaks.

In terms of evaluation, it is not really clear how it is scaled up the false positive ratio (undetected events) to one per day, as the initial intuition would be that the proportions of detected and undetected events should always add up to one, which is not observed on the table results. Also, there is no clarification if the algorithm operates only on a static data set and thus, not meant to work on a continuous stream, an approach that other algorithms take into consideration

Compared with the other papers presented in the Seminar, the work done by Cheng and Neil is the only one that incorporates graph theory and heterogeneous networks for event detection. One of the intuitions of why this work outperforms other algorithms in the given design of the experiment, is that the use of all the relationships between users, tweets and locations with the support of a more directed dictionary increases the capability of discovering anomalous behaviors within the social media graph while also preventing the detection of trending topics as events (issue that the topic/event models based solely on words ignoring the correlations among phrases, tend to label as events).

In later papers[6], one of the improvements for this algorithm was the incorporation of an additional scan statistic function in order to measure the joint significance of evolving subgraphs and subsets of attributes to work with online/ongoing/forthcoming event in dynamic heterogeneous networks (where attributes evolve over time) and modifying the scan algorithm using a Lagrangian relaxation and a dynamic programming based on tree-shaped priors (Steiner Tree), that can efficiently find an approximation solution for the dynamic network.

References

Chen, Feng & B. Neill, Daniel. (2014). Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 10.1145/2623330.2623619.

[6] Minglai Shao, Jianxin Li, Feng Chen, Hongyi Huang, Shuai Zhang, and Xunxun Chen. 2017. An Efficient Approach to Event Detection and Forecasting in Dynamic Multivariate Social Media Networks. In *Proceedings of the 26th International Conference on World Wide Web* (WWW '17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1631-1639. DOI: https://doi.org/10.1145/3038912.3052588