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

Feng Chen* and Daniel B. Neill

University at Albany, SUNY

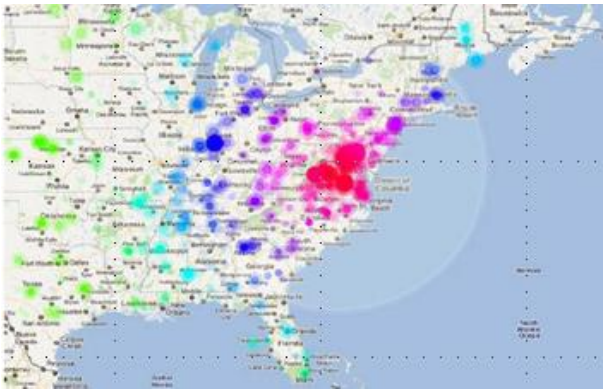
6-18-2014

Why Can We Detect & Forecast Events from Social Media?

2012 July-14, Mexico Protest

2012 Washington D.C. Traffic

Tweet Map for 2011 VA Earthquake



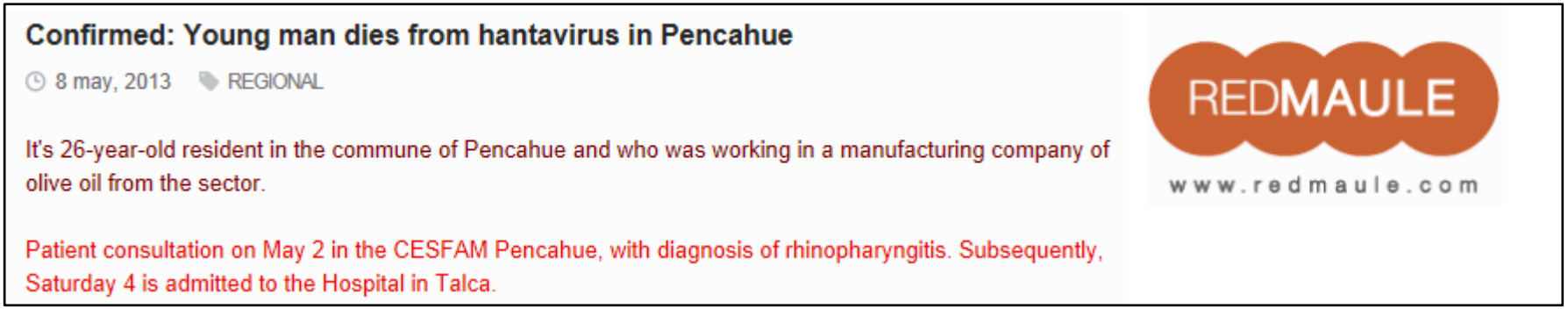
- **Event = Large-scale population behavior**
- **Social media is a real-time "sensor" of large population behavior**
- **Event Detection vs. Forecasting**
  - **Sense of public discussions about ongoing events vs. trigger events using social media**

Disease Event Signals on Twitter

People are dying from hantavirus in Osorno hydroelectric government workers do not report Camila I beg help @ camila_vallejo

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue

Confirmed: Young man dies in Pencahue Hanta: This is a 26-year residence in the commune of http://t.co/5lkD0CZDmf



Confirmed: Young man dies from hantavirus in Pencahue

8 may, 2013 REGIONAL

It's 26-year-old resident in the commune of Pencahue and who was working in a manufacturing company of olive oil from the sector.

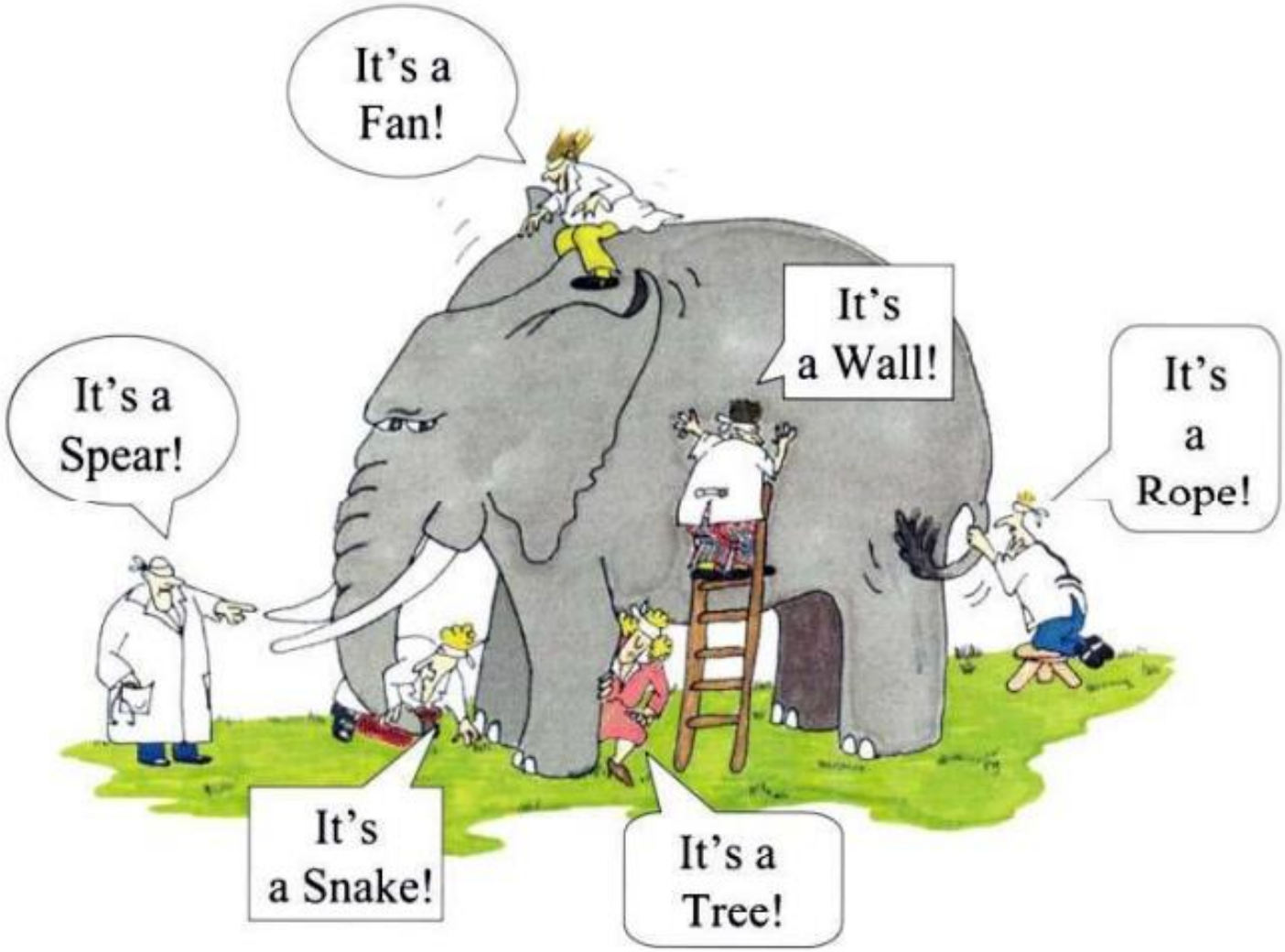
Patient consultation on May 2 in the CESFAM Pencahue, with diagnosis of rhinopharyngitis. Subsequently, Saturday 4 is admitted to the Hospital in Talca.

REDMAULE

www.redmaule.com

RT @ RADIOPALOMAFM: ISP confirmed case of hantavirus nvo rural sector in Linares. Woman, 38, who died May 11 at the UCI via @ SeremiSaludM

Elephant And The Blind Men



Elephant And The Blind Men

Hantavirus Disease Outbreak

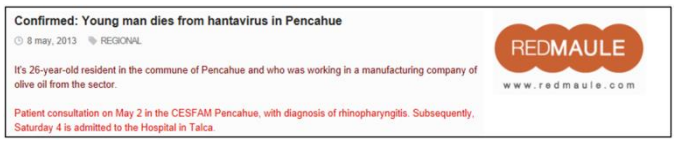
"#VIRUSHANTA" mentioned 100 times

Keyword "Hantavirus" Mentioned 90 times

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue

re-tweeted 50 times

Araucania State has 15 active users and 100 tweets



http://t.co/5lkD0CZDmf mentioned 10 times

Influential User "SeremiSaludM" (1497 followers) posted 8 tweets

Elephant And The Blind Men

Hashtag

"#VIRUSHANTA" mentioned 100 times

Hantavirus Disease Outbreak

Tweet

RT @SeremiSaludM: Se confirmó primer caso de hantavirus en el Maule y con consecuencia fatal. Se trata de un joven de 25 años de Pencahue

re-tweeted 50 times

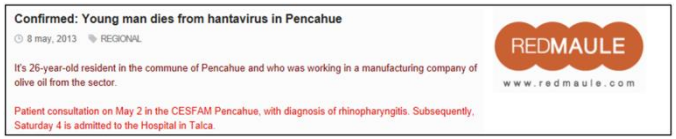
Keyword

Keyword "Hantavirus" Mentioned 90 times

Araucania State has 15 active users and 100 tweets

Location

Link

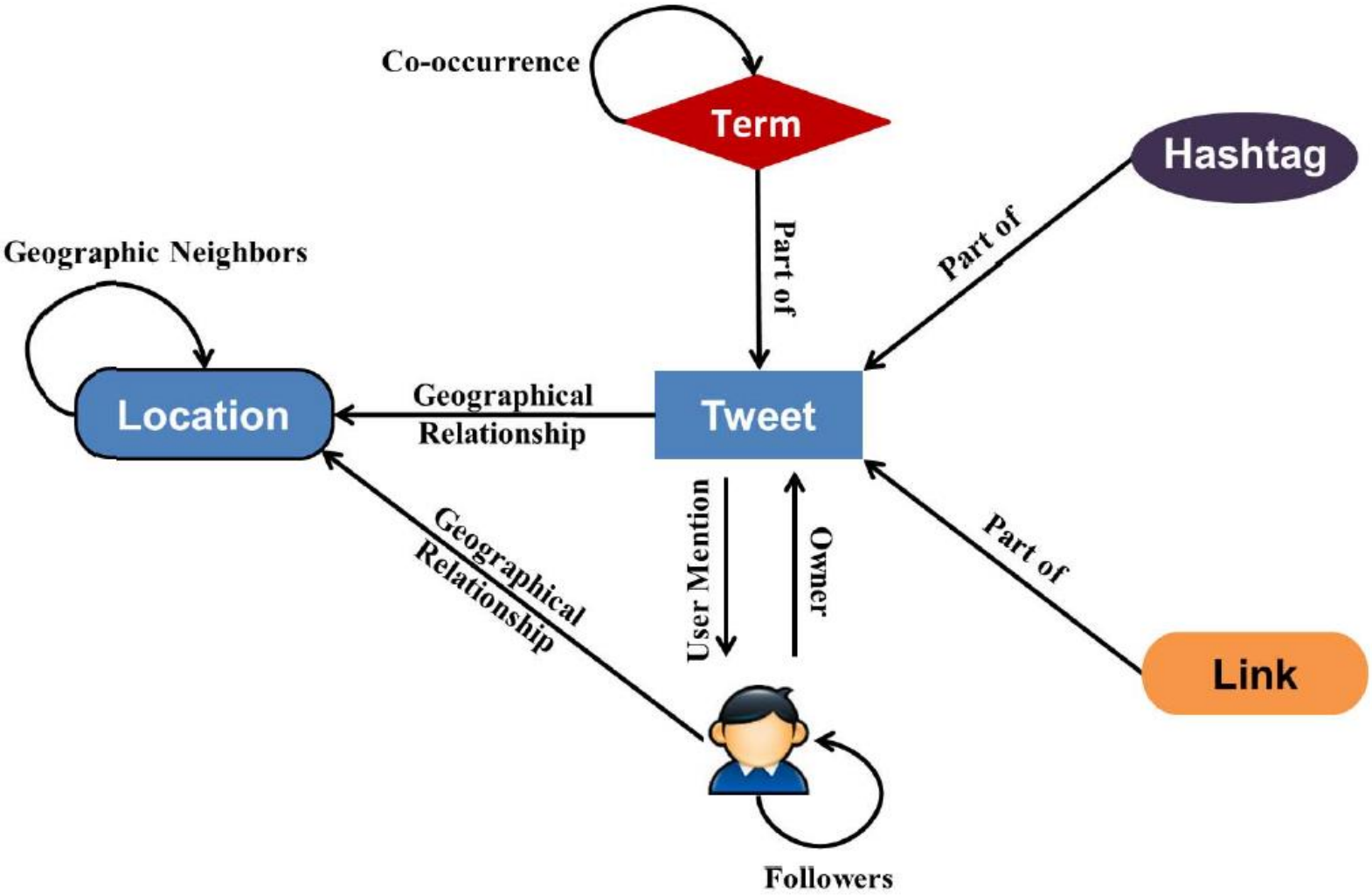


http://t.co/5lkD0CZDmf mentioned 10 times

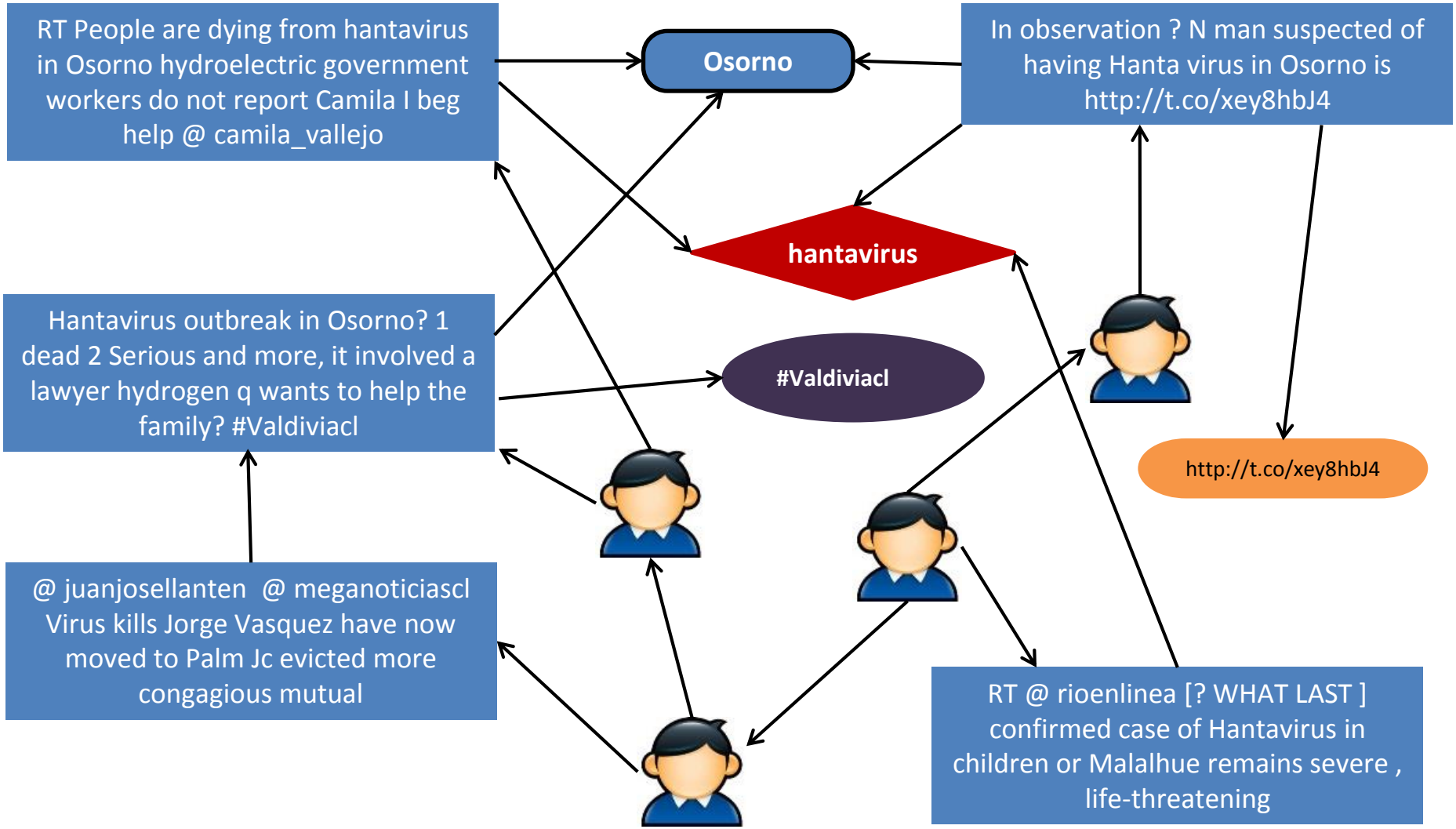
User

Influential User "SeremiSaludM" (1497 followers) posted 8 tweets

Twitter Heterogeneous Network



Twitter Heterogeneous Network (Example)



Node Attributes

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

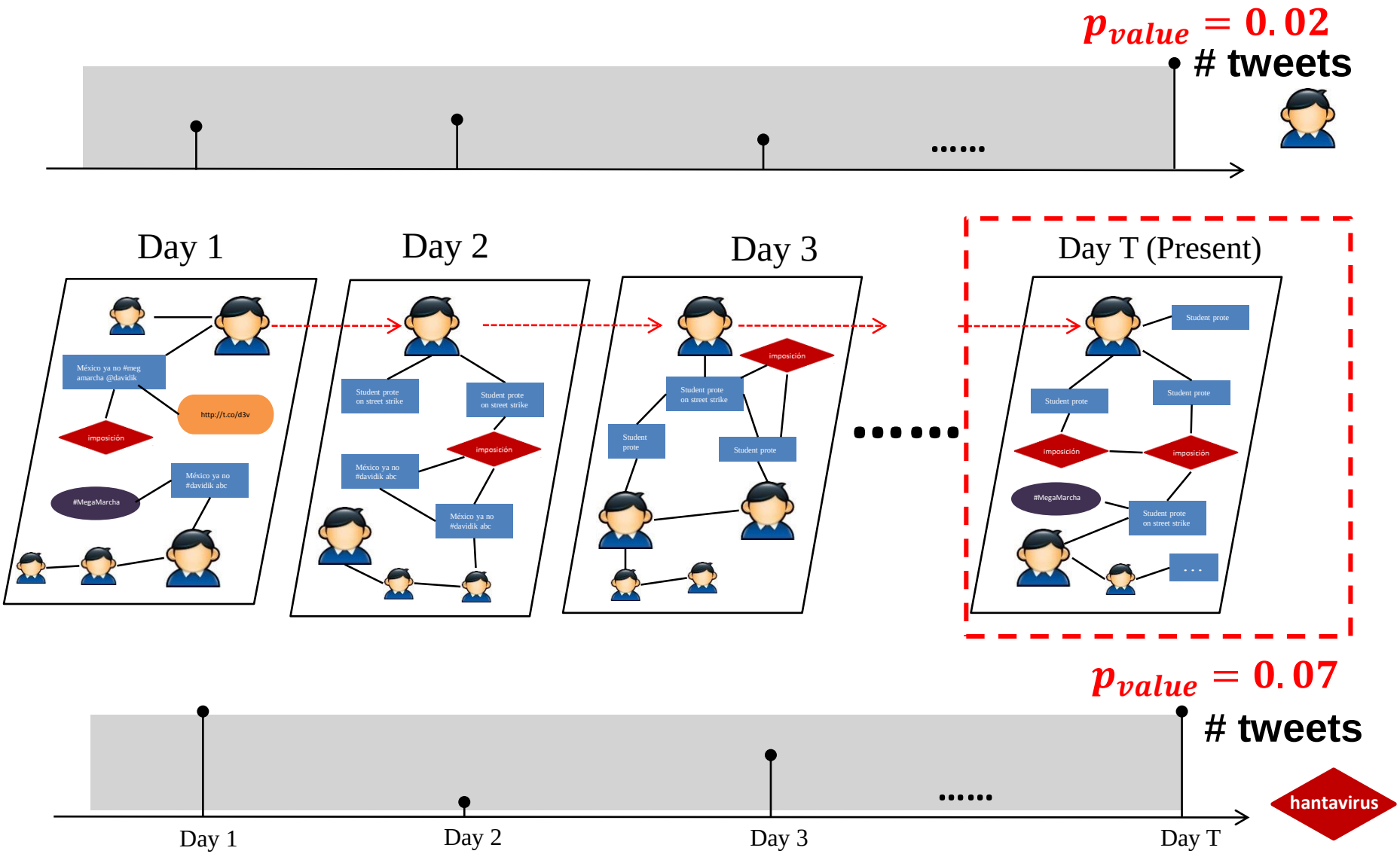
Research Questions

- **A heterogeneous graph (HG) is composed of nodes, attributes, and relations that could be of multiple different types.**
- **The following questions are answered:**
- **Q1: How to define an appropriate scan statistic for a given connected subgraph (“window”) of HG?**
- **Q2 How to efficiently find connected subgraphs in HG that have the largest scan statistic scores?**

Summary of Our Major Contributions

- **Q1: How to define an appropriate scan statistic for a given connected subgraph ("window") of HG?**
  - **First, we propose a two-stage empirical calibration process to calculate an empirical p-value for each node of HG**
  - **Second, we propose a non-parametric scan statistic for a given connected subgraph of HG based on node-level empirical p-values**
- **Q2 How to efficiently find connected subgraphs that have the largest scan statistic scores?**
  - **We design an efficient algorithm to approximately maximize the proposed nonparametric scan statistic over connected connected subgraphs with the time complexity ($O(|V| \log |V|)$), where $|V|$ refers to the total number of nodes in G**

Two Stage Empirical Calibration Process



Two Stage Empirical Calibration Process

- **Question: Each node has multiple p-values. How can we aggregate multiple p-values in to a single p-value?**
  - **Step 1: Calculate the minimal of the multiple p-values**
  - **Step 2: Estimate the single p-value based on empirical calibration of minimal p-values in historical data**

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

Two Stage Empirical Calibration Process

- **THEOREM: The empirical p-value (p(v)) calculated using two-stage empirical calibration process follows a uniform distribution on [0, 1] under the assumption that the current multivariate observations for a single node are exchangeable within the reference set given the null hypothesis that no events of interest are occurring.**

Nonparametric Scan Statistics

Number of nodes in S with p values $\leq \alpha$

Sub-graph

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

Significance level

Number of nodes in S

Berk-Jones (BJ) Statistic

$$\phi_{BJ}(\alpha, N_\alpha(S), N(S)) = N(S) K\left(\frac{N_\alpha}{N}, \alpha\right)$$

Kullback-Liebler Divergence

$$K(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y},$$

Berk-Jones (BJ) Statistic

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], where the alternative hypothesis assumes a piecewise constant distribution with probability density function

$$f(x) = \begin{cases} f_1 & \text{for } 0 \leq x \leq \alpha \\ f_2 & \text{for } \alpha \leq x \text{ with } f_1 \geq f_2 \end{cases}$$

Berk and Jones (1979) demonstrated this test statistic fulfills several optimality properties and has greater power than any weighted Kolmogorov statistic.

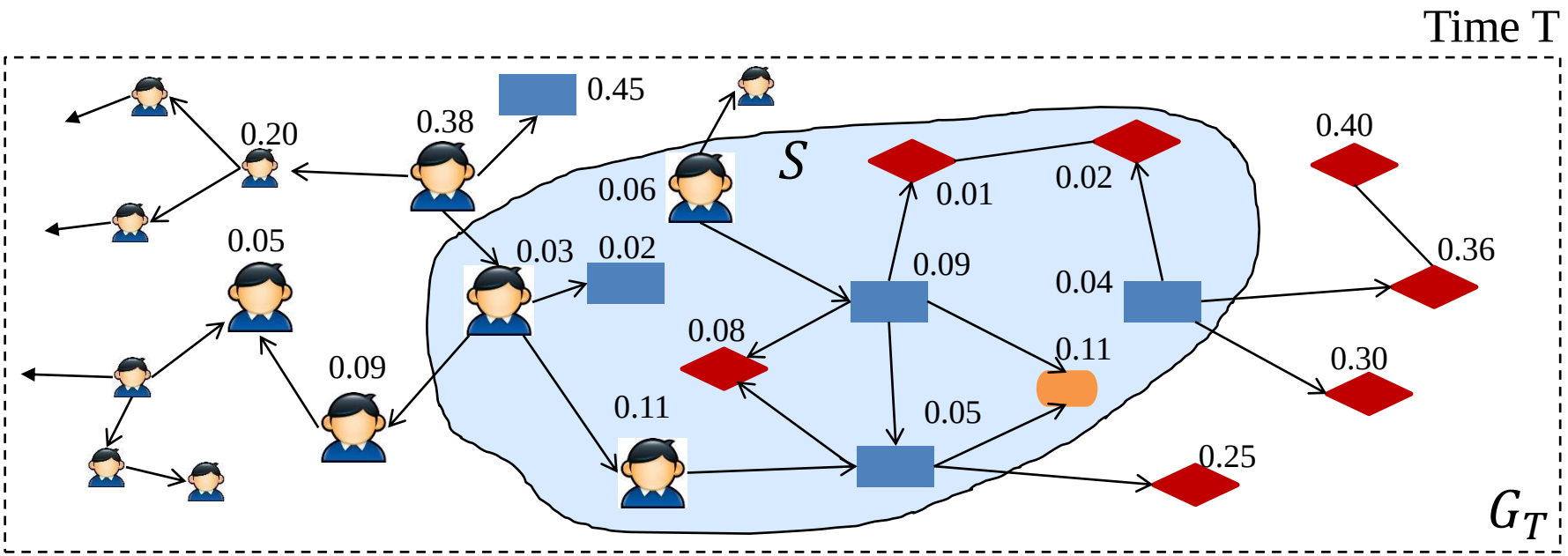
Berk-Jones (BJ) Statistic

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], where the alternative hypothesis assumes a piecewise constant distribution with probability density function

$$f(x) = \begin{cases} f_1 & \text{for } 0 \leq x \leq \alpha \\ f_2 & \text{for } \alpha \leq x \text{ with } f_1 \geq f_2 \end{cases}$$

Berk and Jones (1979) demonstrated this test statistic fulfills several optimality properties and has greater power than any weighted Kolmogorov statistic.

Nonparametric Graph Scanning

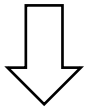


$$S^{\star} = \underset{S \in V_T, S \text{ is connected}}{\operatorname{argmax}} F(S)$$

We propose approximate algorithm with time cost $O(|V_T| \log |V_T|)$.

Non-Parametric Graph Scanning

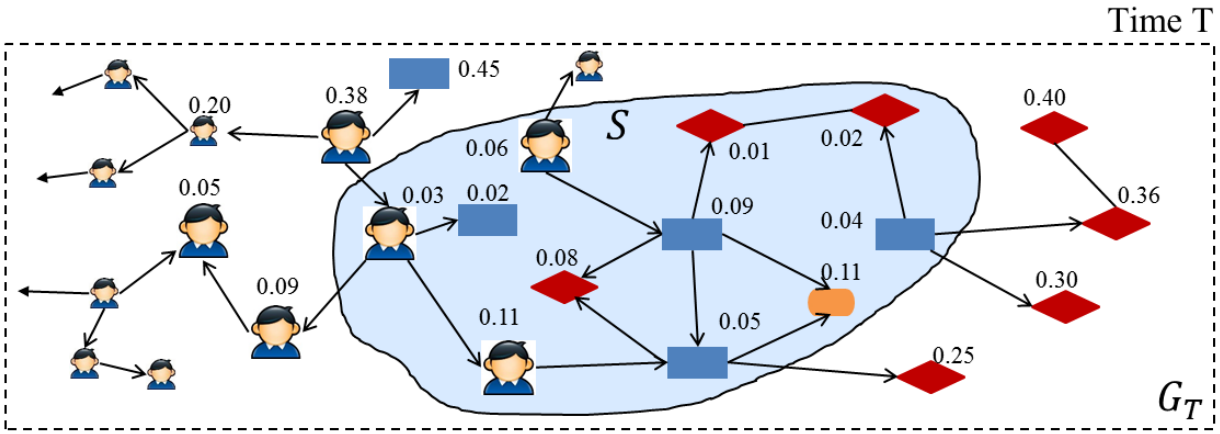
$$\max_{S \subseteq V, S \text{ is connected}} \max_{\alpha \leq \alpha_{max}} \phi(\alpha, N_\alpha(S), N(S))$$



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

The union of $\{\alpha_{max}\}$ and the set of distinct p-values less than α_{max} in V

Non-Parametric Graph Scanning

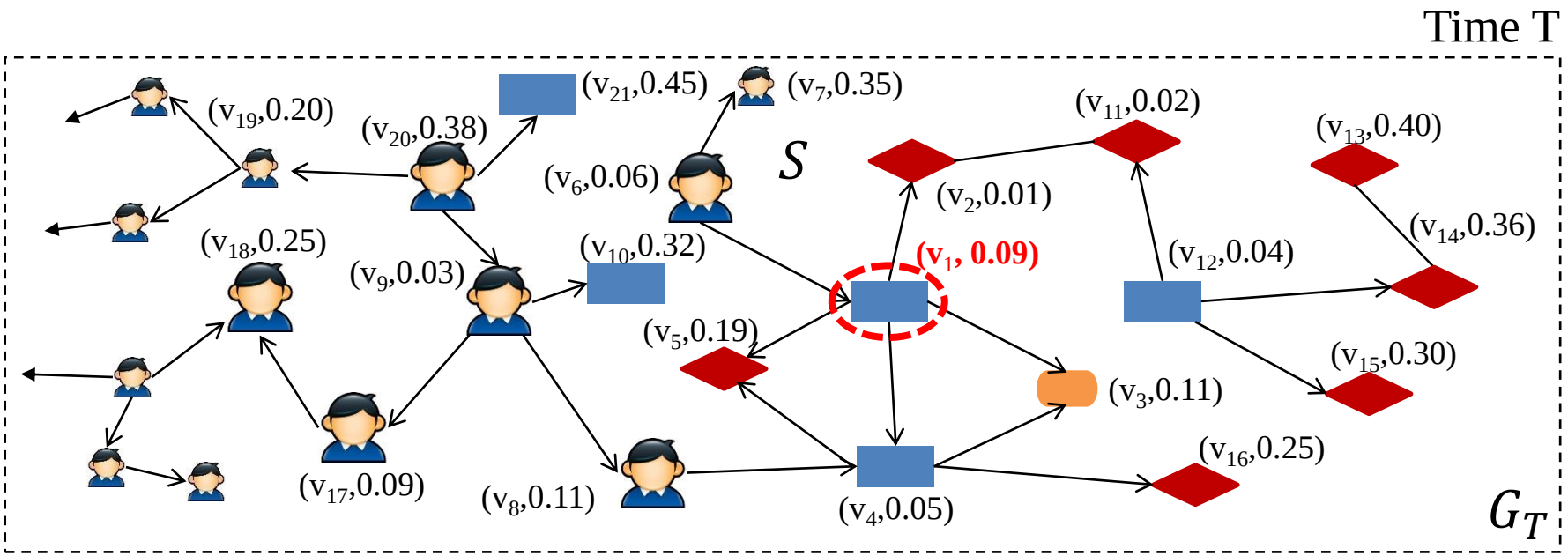


$$\max_{S \subseteq V, S \text{ is connected}} \quad \max_{\alpha \leq \alpha_{max}} \phi(\alpha, N_\alpha(S), N(S))$$

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

The union of $\{\alpha_{max}\}$ and the set of distinct p-values less than α_{max} in V

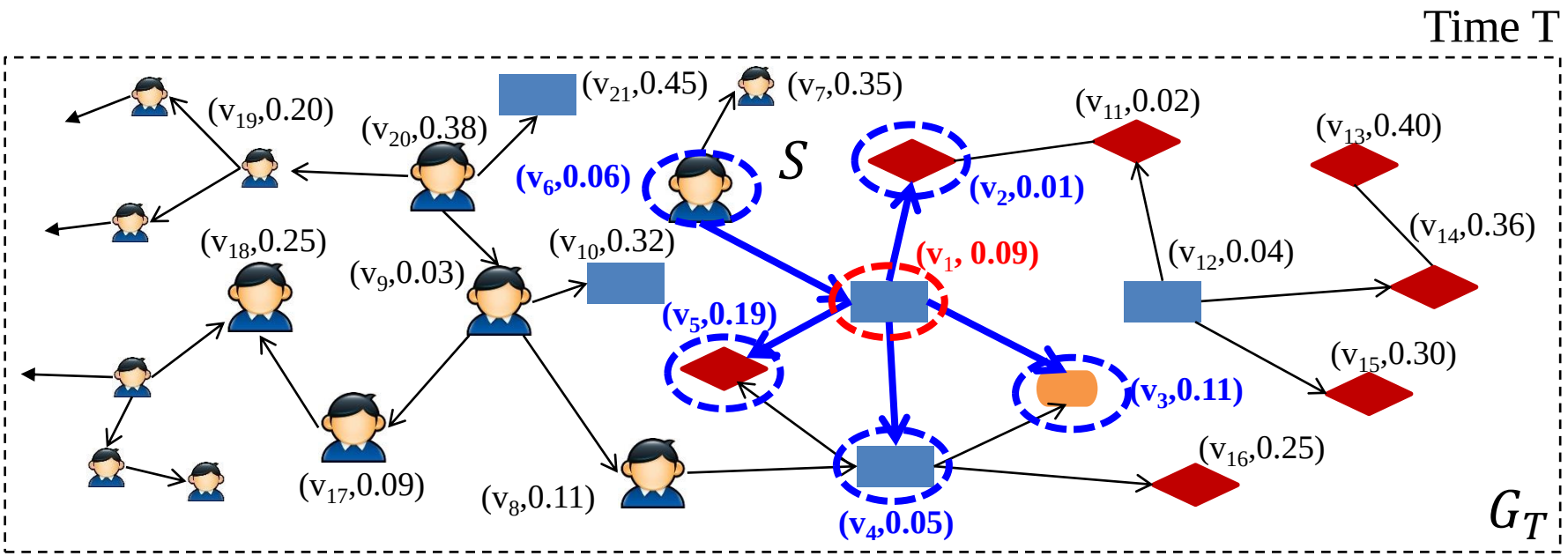
Non-Parametric Graph Scanning



Consider each node as a candidate cluster center (or start point)

In this example, we start from the seed set $S^{(0)} = \{v_1\}$.

Non-Parametric Graph Scanning

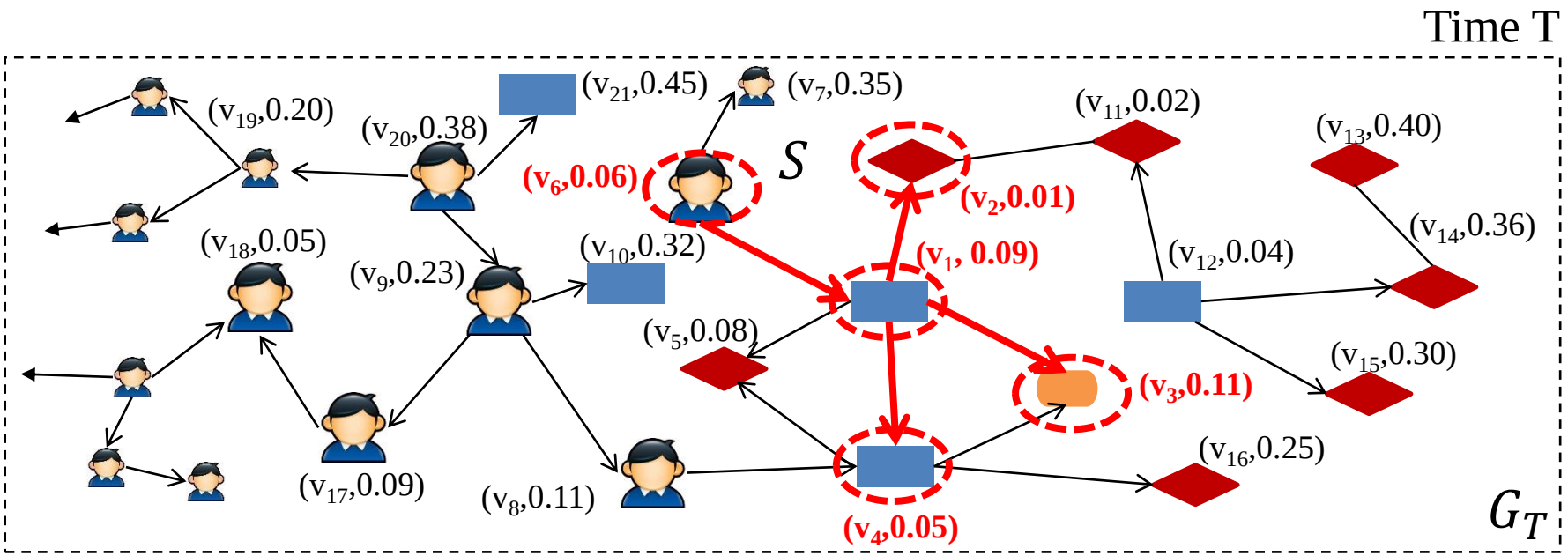


Expand $S^{(0)}$ by adding positive neighbor nodes:

$$\mathcal{N}(S^{(0)}) = \{v_2, v_3, v_4, v_5, v_6\}$$

$$S^{(1)} = S^{(0)} \cup \arg \max_{\alpha \in U(S^{(0)} \cup S, \alpha_{max})} \left\{ \max_{S \in \mathcal{N}(S^{(0)})} NK\left(\frac{N_\alpha(S^{(0)} \cup S)}{N(S^{(0)} \cup S)}, \alpha \right) \right\}$$

Non-Parametric Graph Scanning



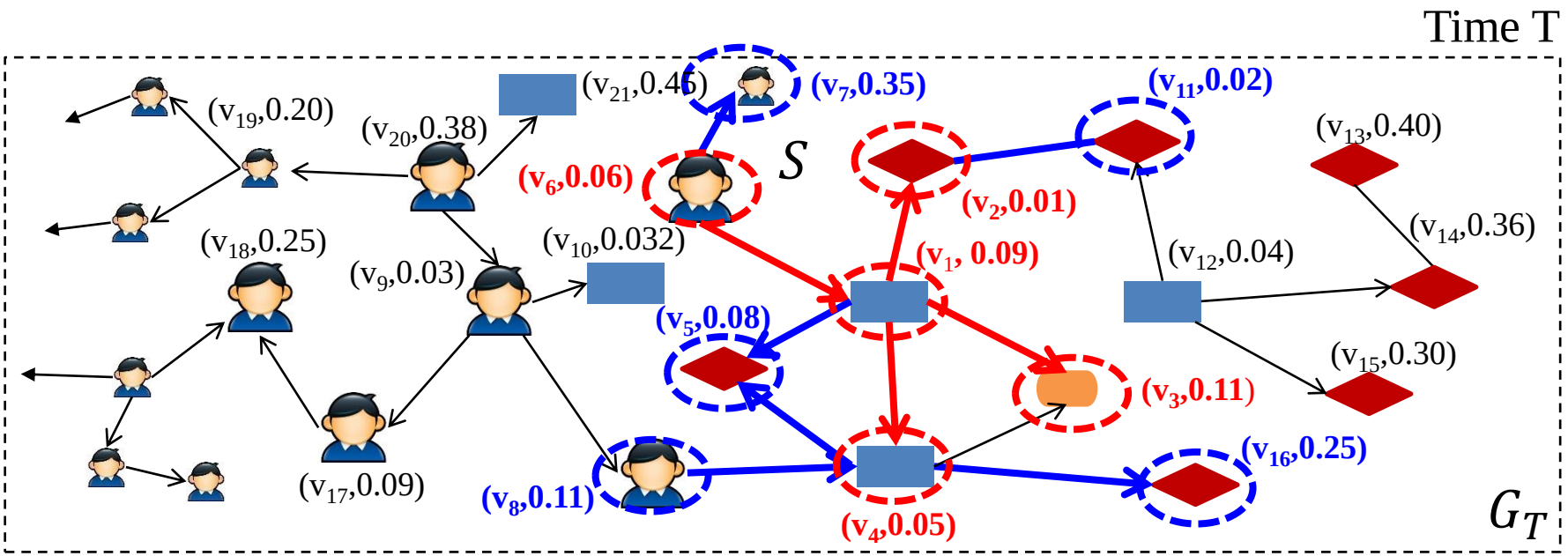
Expand $S^{(0)}$ by adding positive neighbor nodes:

$$\mathcal{N}(S^{(0)}) = \{v_2, v_3, v_4, v_5, v_6\}$$

$$S^{(1)} = S^{(0)} \cup \arg \max_{\alpha \in U(S^{(0)} \cup S, \alpha_{max})} \left\{ \max_{S \in \mathcal{N}(S^{(0)})} NK\left(\frac{N_\alpha(S^{(0)} \cup S)}{N(S^{(0)} \cup S)}, \alpha\right) \right\}$$

$$= \{v_1, v_2, v_3, v_4, v_6\}$$

Nonparametric Graph Scanning Algorithm

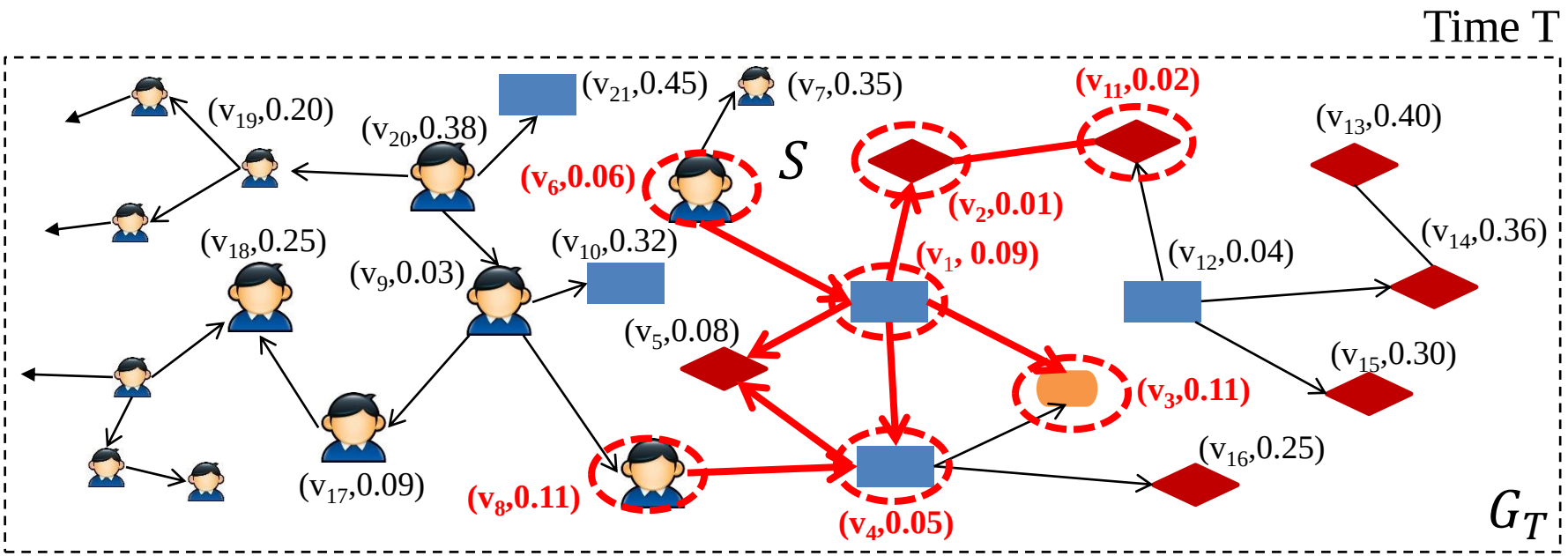


Expand $S^{(1)}$ by adding positive neighbor nodes:

$$\mathcal{N}(S^{(1)}) = \{v_5, v_7, v_8, v_{11}, v_{16}\}$$

$$S^{(2)} = S^{(1)} \cup \arg \max_{\alpha \in U(S^{(1)} \cup S, \alpha_{max})} \left\{ \max_{S \in \mathcal{N}(S^{(1)})} NK\left(\frac{N_\alpha(S^{(1)} \cup S)}{N(S^{(1)} \cup S)}, \alpha \right) \right\}$$

Nonparametric Graph Scanning Algorithm



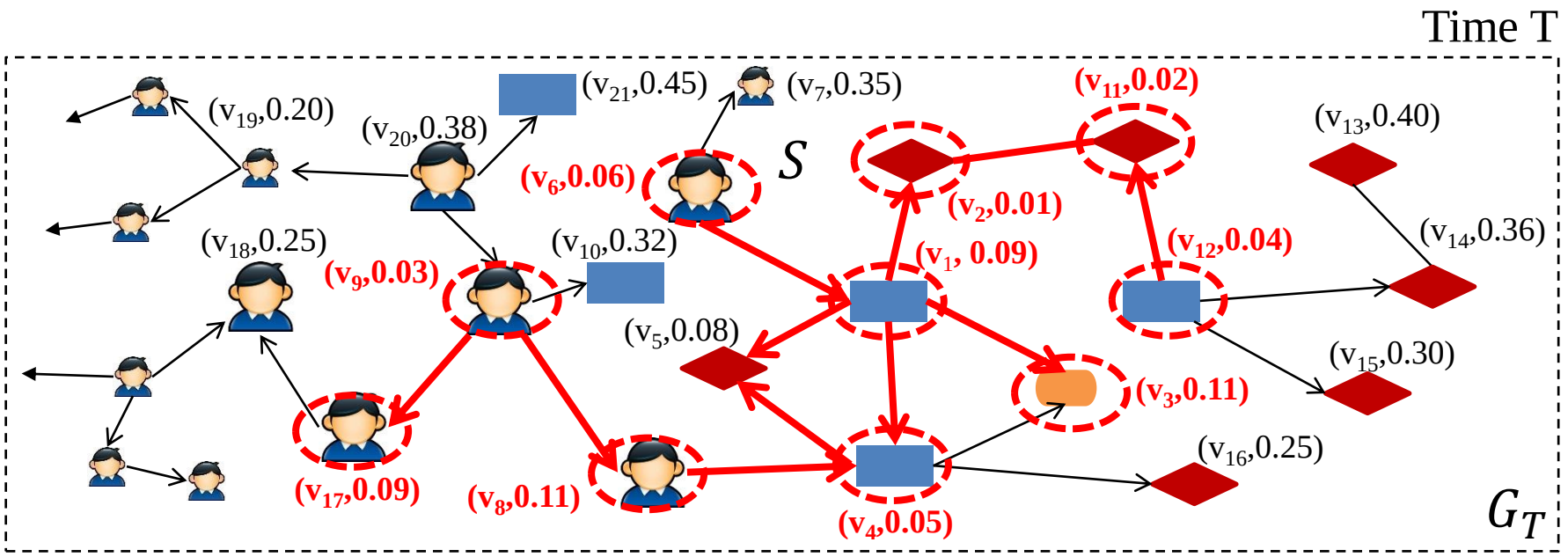
Expand $S^{(1)}$ by adding positive neighbor nodes:

$$\mathcal{N}(S^{(1)}) = \{v_5, v_7, v_8, v_{11}, v_{16}\}$$

$$S^{(2)} = S^{(1)} \cup \arg \max_{\alpha \in U(S^{(1)} \cup S, \alpha_{max})} \left\{ \max_{S \in \mathcal{N}(S^{(1)})} NK\left(\frac{N_\alpha(S^{(1)} \cup S)}{N(S^{(1)} \cup S)}, \alpha\right) \right\}$$

$$= \{v_1, v_2, v_3, v_4, v_6, v_8, v_{11}\}$$

Nonparametric Graph Scanning Algorithm

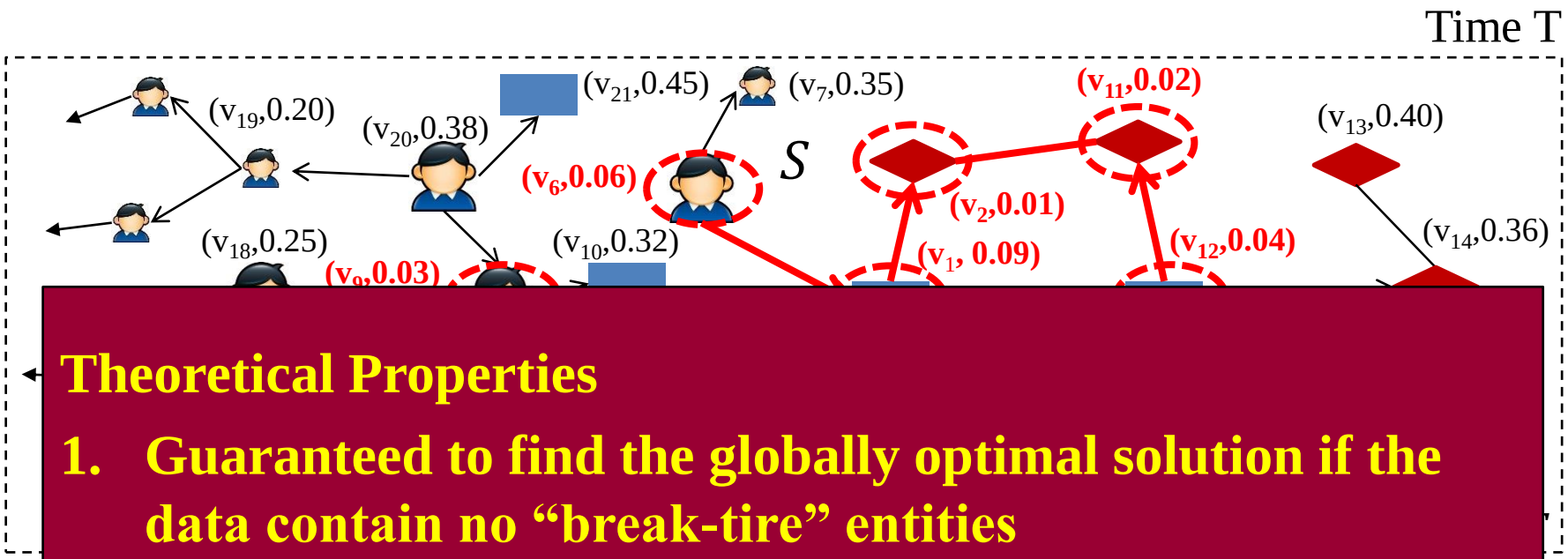


Consider each node as a candidate cluster center (or start point)

In this example, we start from the seed set $\hat{S} = \{v_1\}$, and after four expansions, we obtain the local optimum solution:

$$S_{v_1}^{\star} = \{v_1, v_2, v_3, v_4, v_6, v_8, v_9, v_{11}, v_{12}, v_{17}\}$$

Nonparametric Graph Scanning Algorithm



Theoretical Properties

1. **Guaranteed to find the globally optimal solution if the data contain no "break-tire" entities**
2. **Equivalent to percolation-based graph scan under certain simplifying assumptions**

In this example, we start from the seed set $S = \{(v_1, 0.09)\}$, and after four expansions, we obtain the local optimum solution:

$$S_{v_1}^{\star} = \{v_1, v_2, v_3, v_4, v_6, v_8, v_9, v_{11}, v_{12}, v_{17}\}$$

Experimental Evaluations

- **Detection and Forecasting of Hantavirus Disease Outbreaks**
- **Detection and Forecasting of Civil Unrest Events**

Experiment Settings

- **Twitter Dataset**
  - **10% random sample of public twitter data**
  - **17 rare Hantavirus disease outbreaks collected by Chilean Ministry of Health and also reported in local news reports from 2013-January to 2013-June**
- **Performance Metrics**
  - **Forecasting: Have an alert in the same state 1-7 days before the event**
  - **Detection: Did not have an alert in that state 1-7 days before the event but did have an alert in the event 07 days after the event**

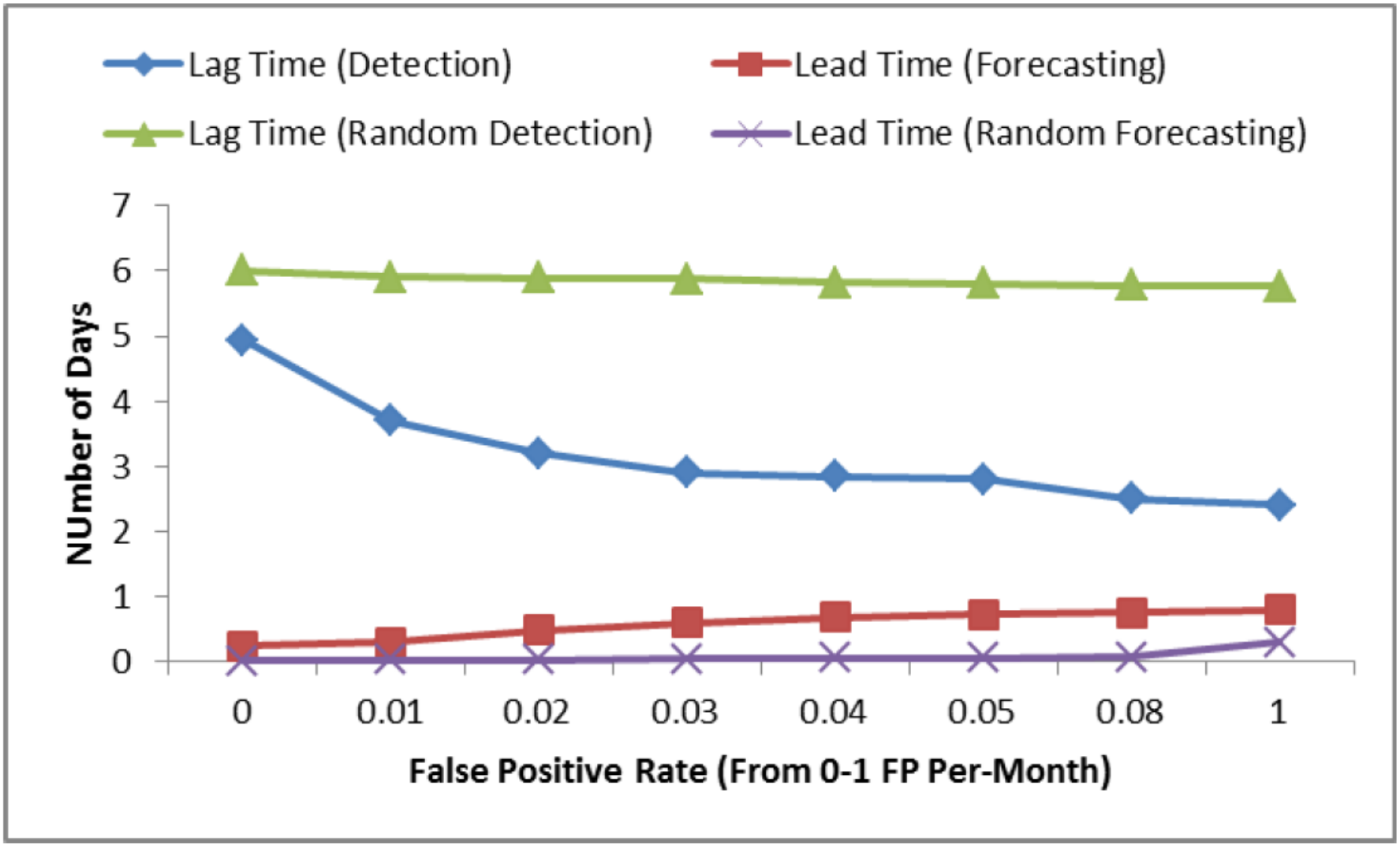
Twitter Dataset

Country	# of tweets	News source*
Chile	14 ,000,000	La Tercera; Las Últimas Notícias; El Mercurio

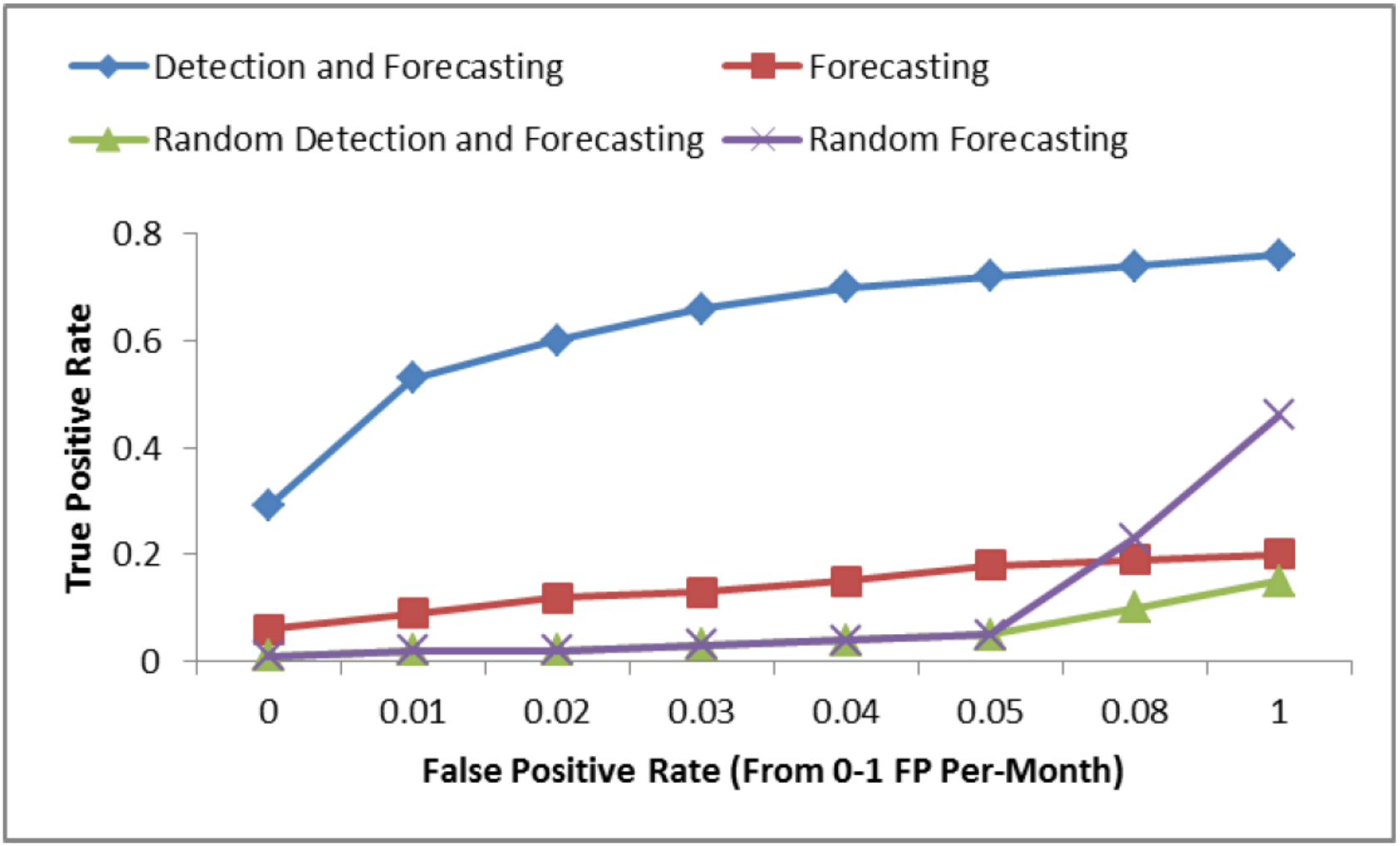
Time Period: From **2013 Jan.** to **2013 Jun.** Totally **17** Hantavirus outbreaks

Example of an event label: (PROVINCE = “La Araucanía”, COUNTRY = “Chile”, DATE = “2013-01-19”, News Title = “A 11 year old boy who was admitted to a clinic in Temuco down suspected hantavirus died during the day on Saturday.”, News Link = “http://www.biobiochile.cl/2013/01/19/muere-menor-sospechoso-de-hanta-en-temuco.shtml”).

Detection Lag Time and Prediction Lead Time



Detection and Forecasting Results



Twitter Data Set & Golden Standard Reports (GSR)

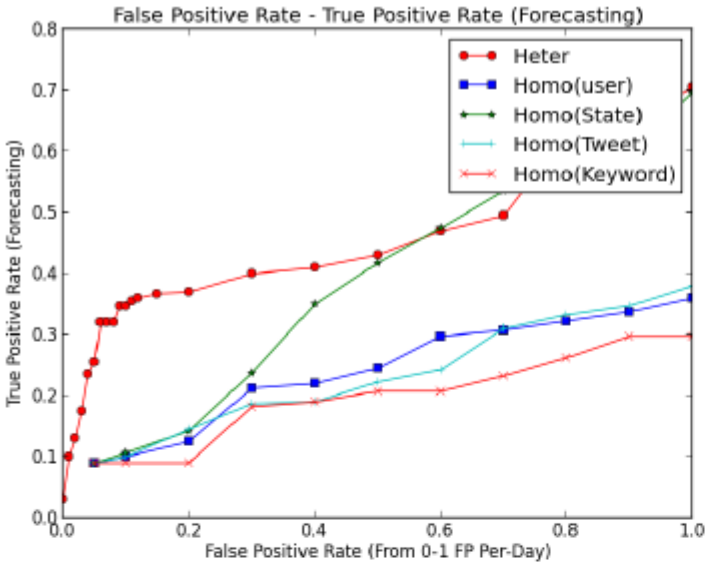
Country	# of tweets	News source*
Argentina	29 ,000,000	Clarín; La Nación; Infobae
Brazil	32 ,000,000	O Globo; O Estado de São Paulo; Jornal do Brasil
Chile	14 ,000,000	La Tercera; Las Últimas Notícias; El Mercurio
Colombia	22 ,000,000	El Espectador; El Tiempo; El Colombiano
Ecuador	6,900,000	El Universo; El Comercio; Hoy
El Salvador	3,700,000	El Diáro de Hoy; La Prensa Gráfica; El Mundo
Mexico	24 ,000,000	La Jornada; Reforma; Milenio
Paraguay	4,600,000	ABC Color; Ultima Hora; La Nacíon
Uruguay	1,800,000	El Paí; El Observador
Venezuela	31 000,000	El Universal; El Nacional; Ultimas Notícias

Time Period: From **2012 Jul.** to **2012 Dec.** Totally **1544** civil unrest events

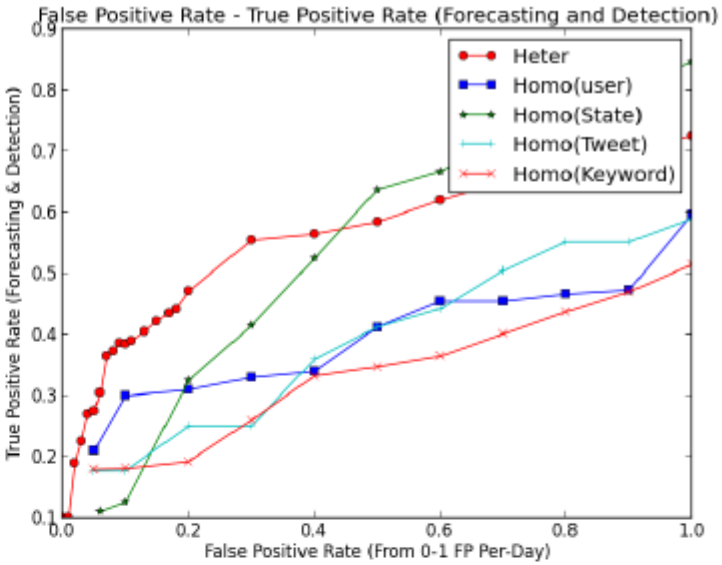
Comparison Study

- **Comparison Between Our Proposed Approach and Existing Methods**

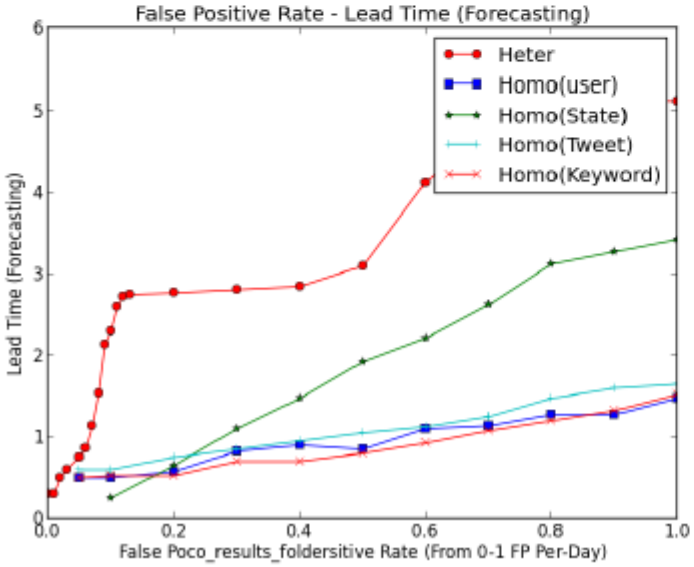
Method	FPR (FP/Day)	TPR (Forecasting)	TPR (Forecasting & Detection)	Lead Time (Days)	Lag Time (Days)	Run Time (Hours)
ST Burst Detection	0.65	0.07	0.42	1.10	4.57	30.1
Graph Partition	0.29	0.03	0.15	0.59	6.13	18.9
Earthquake	0.04	0.06	0.17	0.49	5.95	18.9
Geo Topic Modeling	0.09	0.06	0.08	0.01	6.94	9.7
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



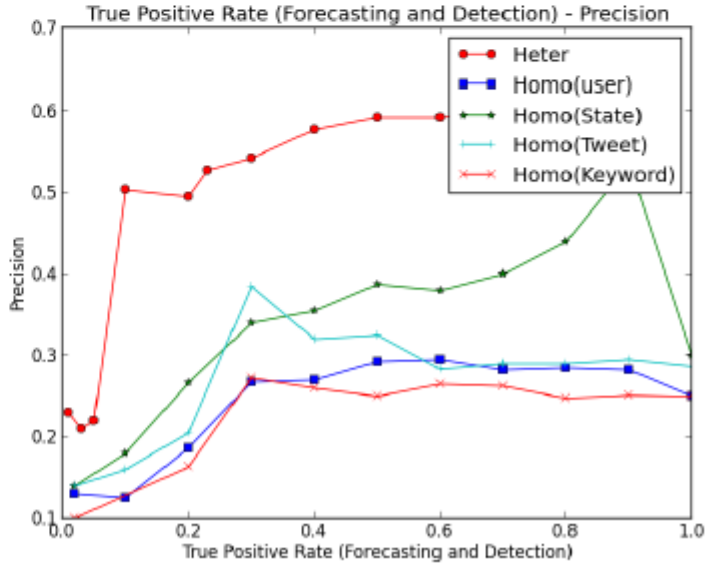
(a) FPR vs. TPR (Forecasting)



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



(c) FPR vs. Leadtime (Forecasting)



(d) TPR (Forecasting and Detection) vs. Precision

Conclusion

- **This work presents a nonparametric scan statistics approach to event detection and forecasting in heterogeneous social media graphs**
- **We argue that nonparametric methods are better suited to social media than parametric methods**
- **Extensive empirical evaluations in real world datasets demonstrate the effectiveness and efficiency of our proposed approach**

Thank you!

Questions?