Non-Parametric Scan Statistics for Disease Outbreak Detection on Twitter

Feng Chen and Daniel B. Neill

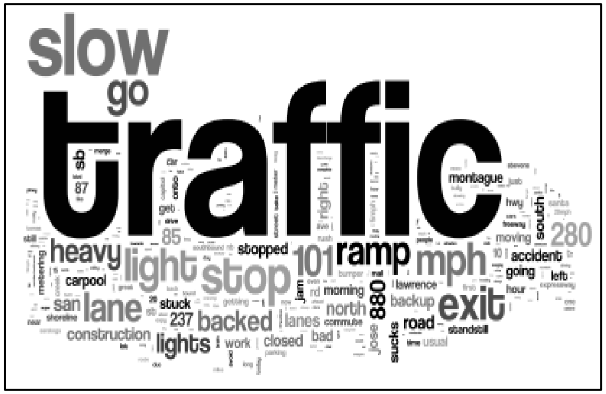
Carnegie Mellon University

12-12-2013

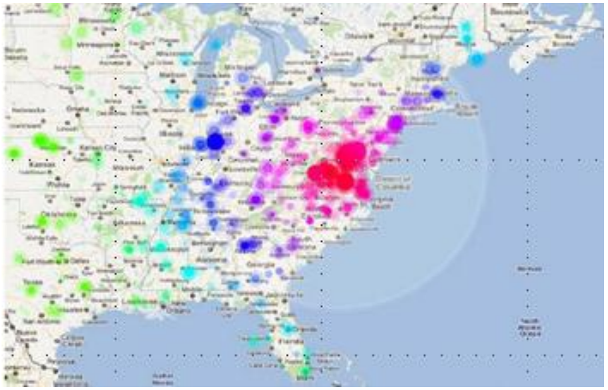
Why Can We Detect Events from Social Media?

2012 July-14, Mexico Protest

2012 Washington D.C. Traffic



Tweet Map for 2011 VA Earthquake



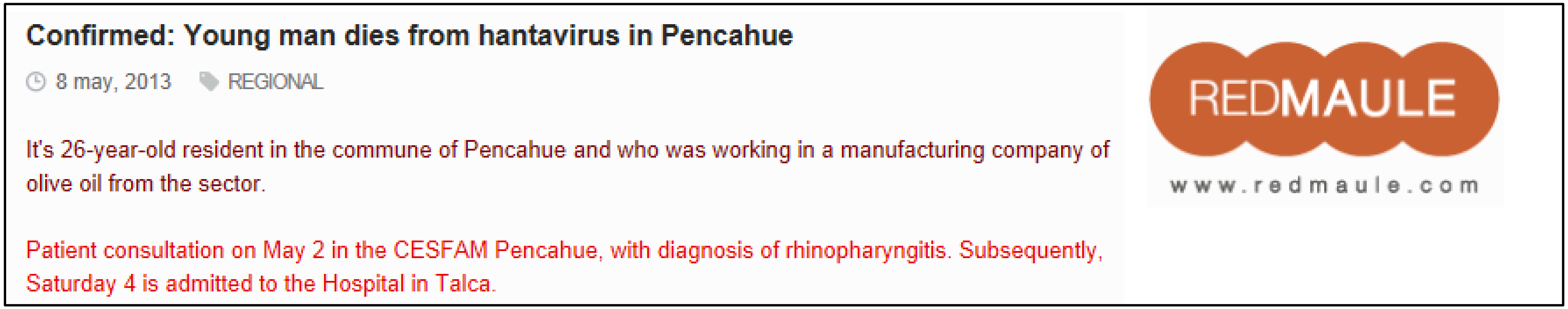
- **Event = Large-scale population behavior**
- **Social media is a real-time "sensor" of large population behavior**
- **Event Detection vs. Forecasting**
  - **Sense of public discussions about ongoing events vs. trigger events using social media**

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

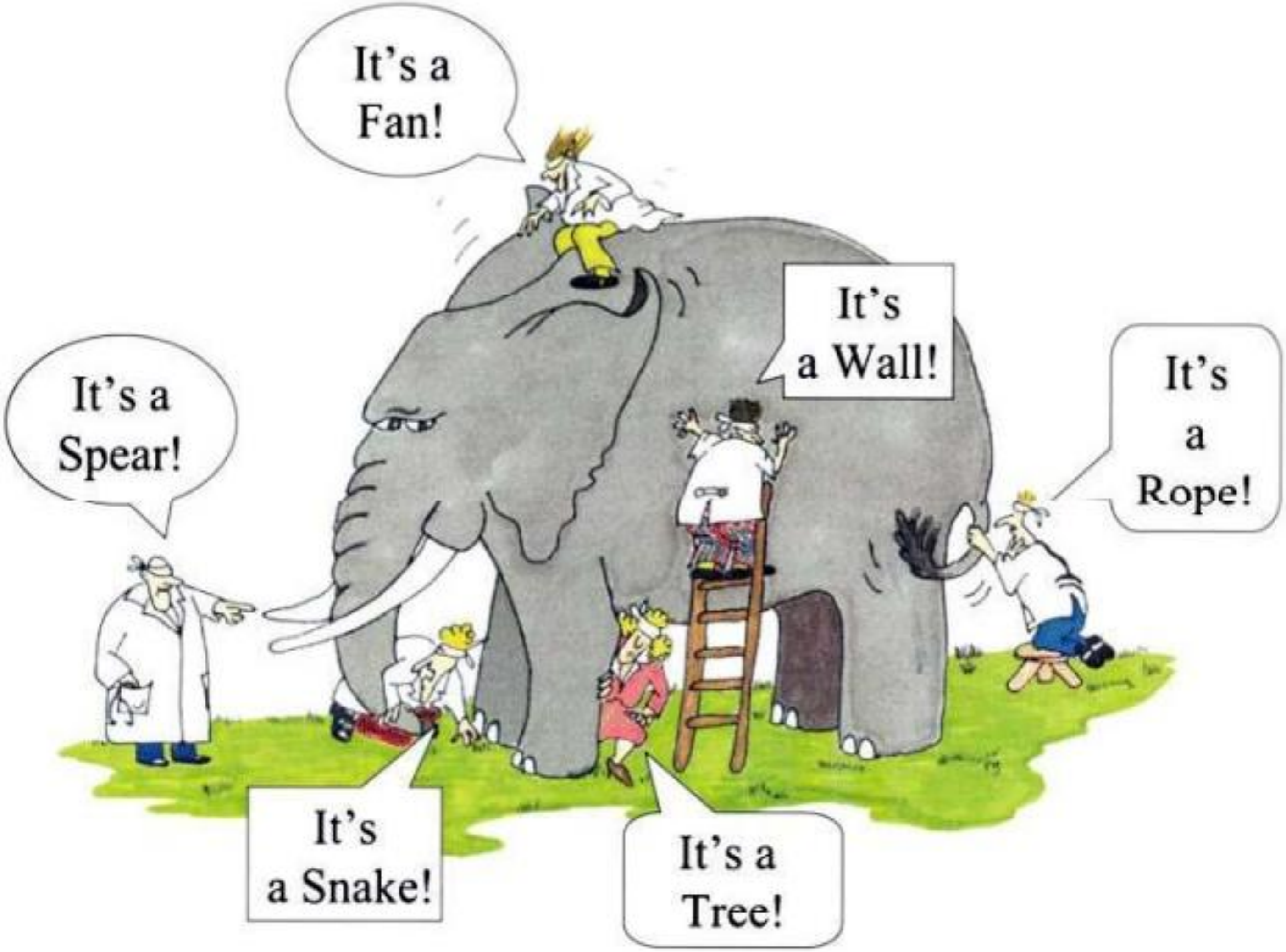
8 may, 2013 REGIONAL

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.

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

Technical Challenges



Technical Challenges

Hantavirus Disease Outbreak

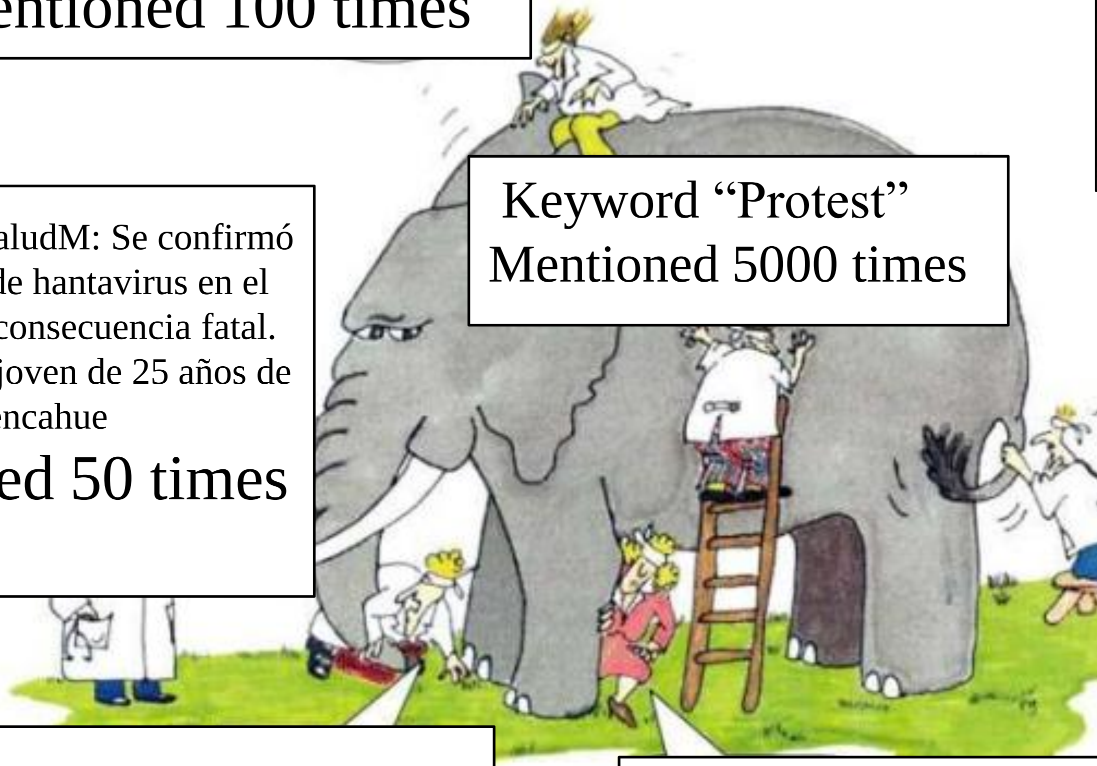
"#VIRUSHANTA"
mentioned 100 times

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

Keyword "Protest"
Mentioned 5000 times

Araucania State
has 15 active users
and 100 tweets

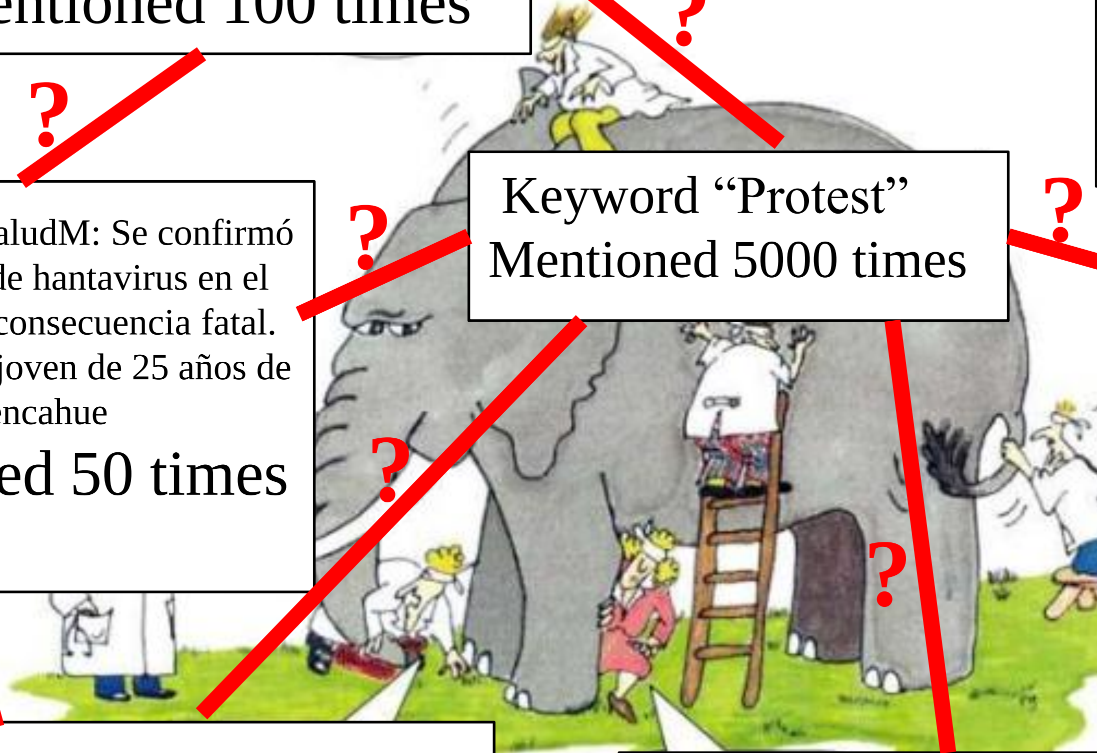


http://t.co/5lkD0CZDmf
mentioned 10 times

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

Technical Challenges

Hantavirus Disease Outbreak



"#VIRUSHANTA"
mentioned 100 times

Keyword "Protest"
Mentioned 5000 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

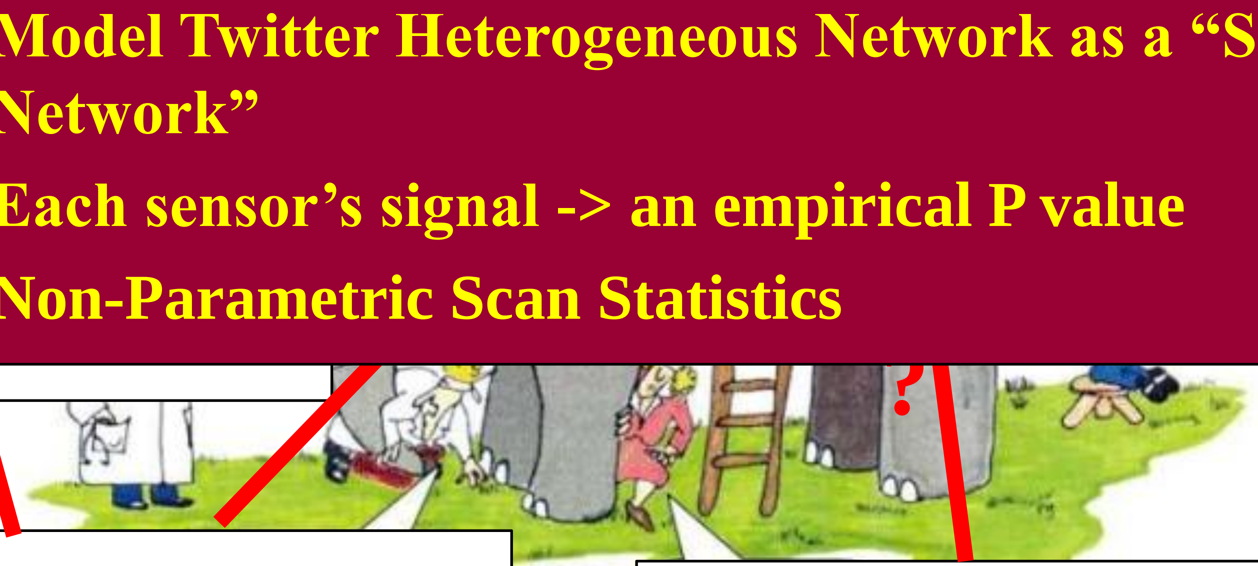
Influential User "SeremiSaludM"
(1497 followers) posted 8 tweets

Technical Challenges

Hantavirus Disease Outbreak

"#VIRUSHANTA" mentioned 100 times

?



RT @
prin
Mau
Se tr

re-

e
ers
s

?

?

?

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

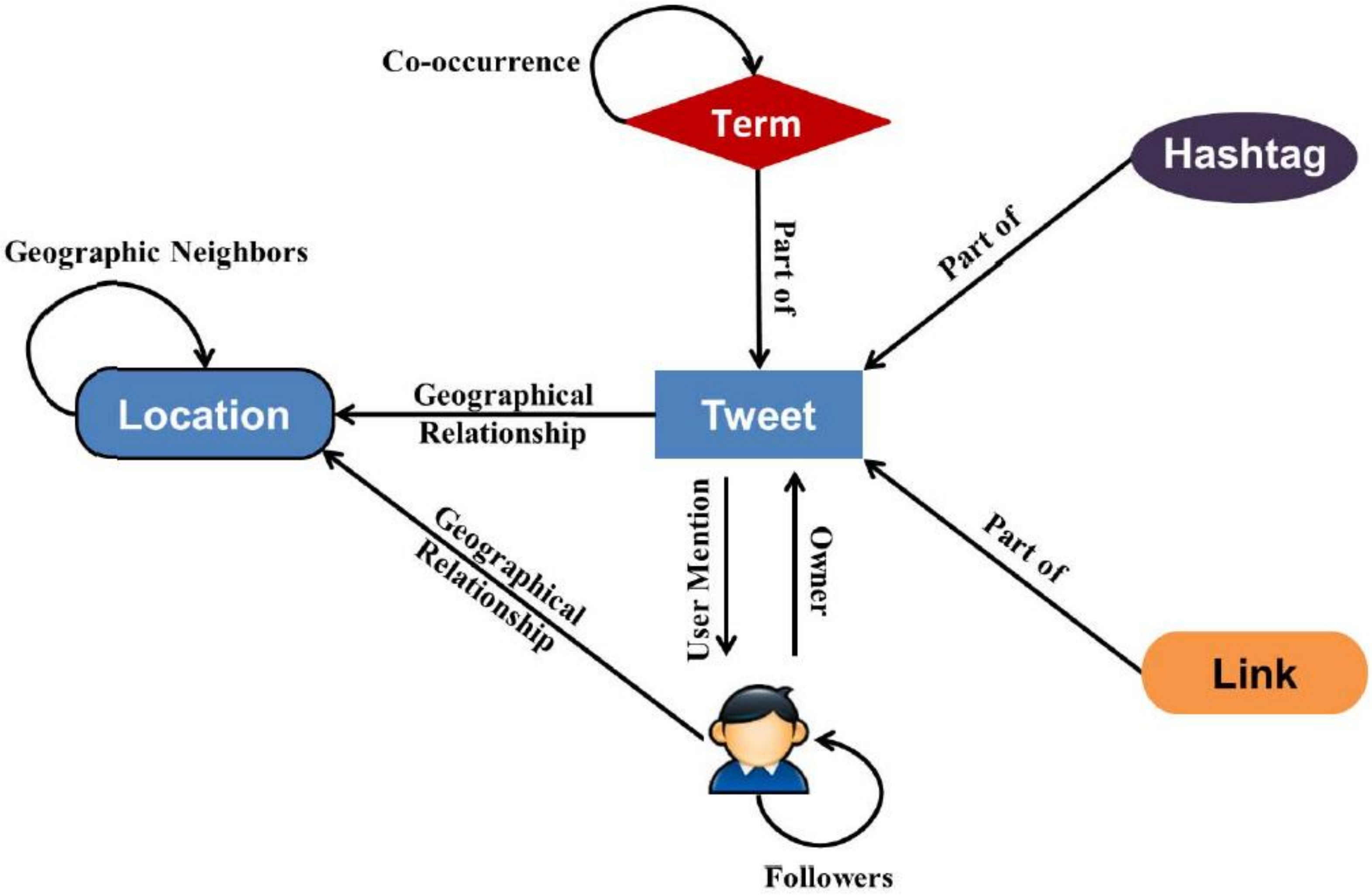
?

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

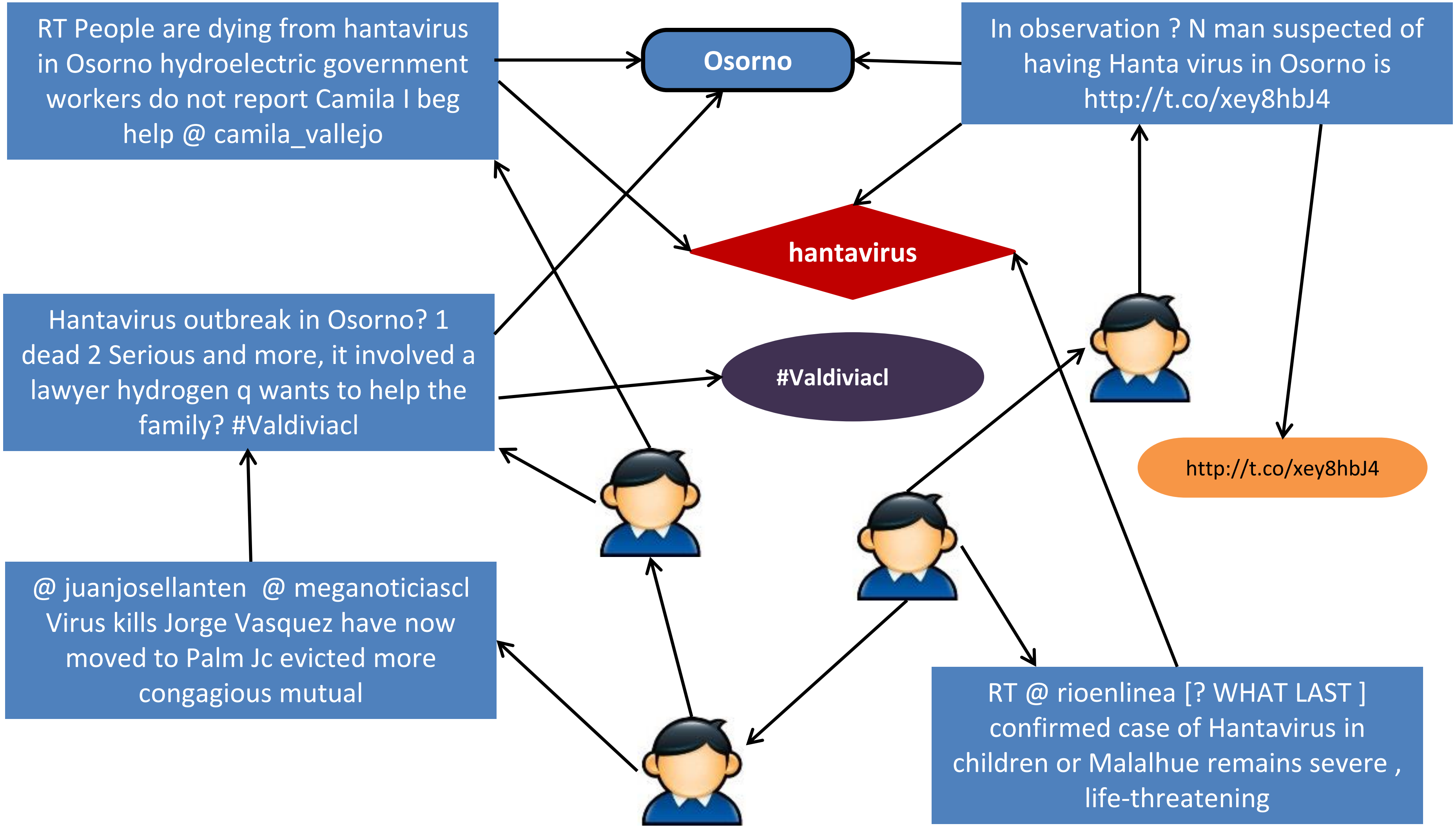
Our Solution

1. **Model Twitter Heterogeneous Network as a "Sensor Network"**
2. **Each sensor's signal -> an empirical P value**
3. **Non-Parametric Scan Statistics**

Twitter Heterogeneous Network



Twitter Heterogeneous Network (Example)



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

Osorno

In observation ? N man suspected of having Hanta virus in Osorno is http://t.co/xey8hbJ4

hantavirus

Hantavirus outbreak in Osorno? 1 dead 2 Serious and more, it involved a lawyer hydrogen q wants to help the family? #Valdiviacl

#Valdiviacl

http://t.co/xey8hbJ4

@ juanjosellanten @ meganoticiascl Virus kills Jorge Vasquez have now moved to Palm Jc evicted more congagious mutual

RT @ rioenlinea [? WHAT LAST] confirmed case of Hantavirus in children or Malalhue remains severe , life-threatening

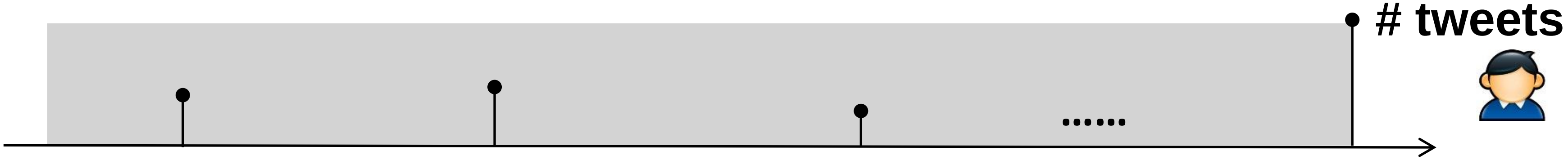
Step 1: “Sensor Network” Modeling

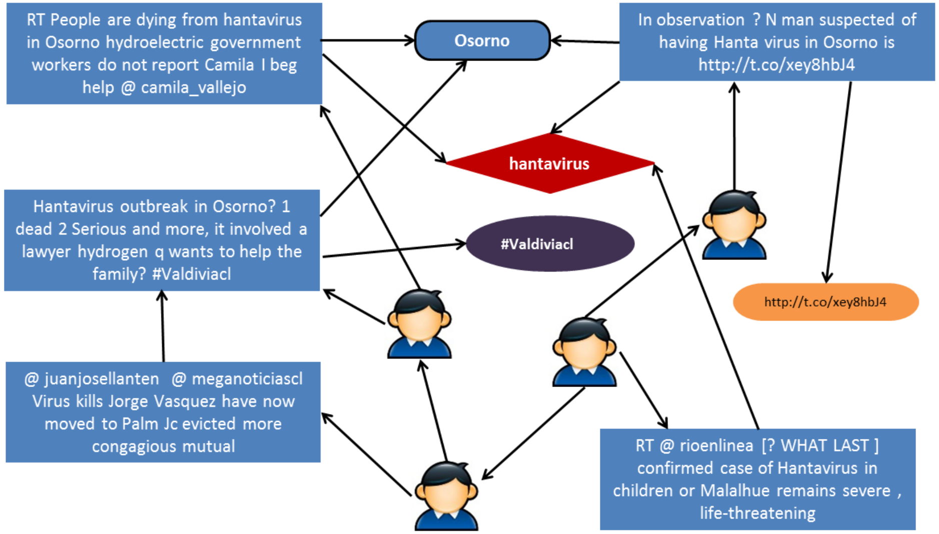
- **Model the twitter network as a "sensor" network, in which each node senses its "neighborhood environment" and reports an empirical p-value measuring the current level of anomalousness for each time interval (e.g., hour or day).**

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

Step 2: Sensor Signals → Empirical P-values

$p_{value} = 0.02$





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

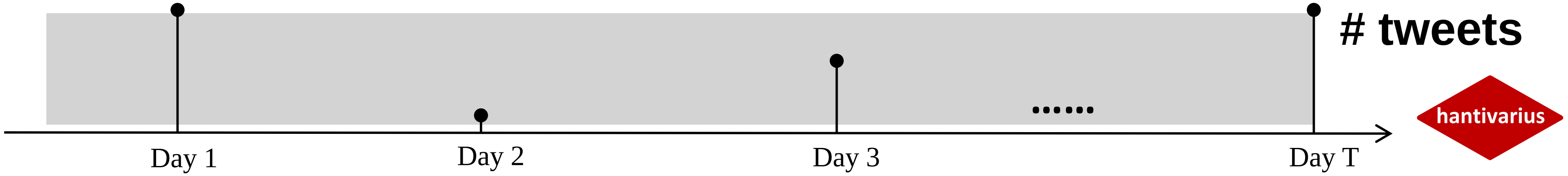
In observation ? N man suspected of having Hanta virus in Osorno is http://t.co/xey8hbJ4

Hantavirus outbreak in Osorno? 1 dead 2 Serious and more, it involved a lawyer hydrogen q wants to help the family? #Valdiviacl

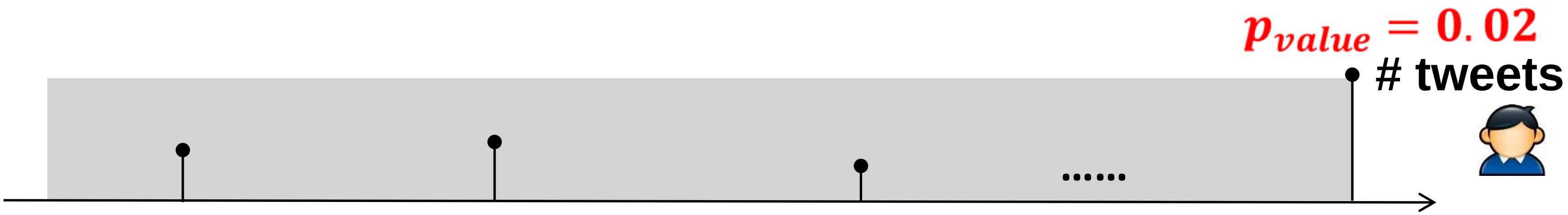
@ juanjosellanten @ meganoticiascl Virus kills Jorge Vasquez have now moved to Palm Jc evicted more congagious mutual

RT @ rioenlinea [? WHAT LAST] confirmed case of Hantavirus in children or Malalhue remains severe , life-threatening

$p_{value} = 0.07$

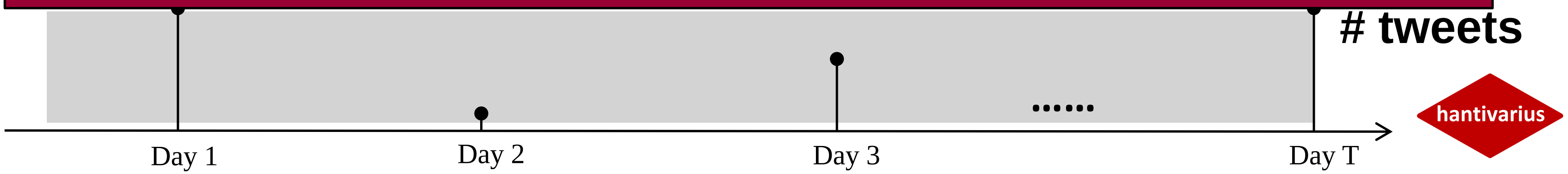


Step 2: Sensor Signals → Empirical P-values

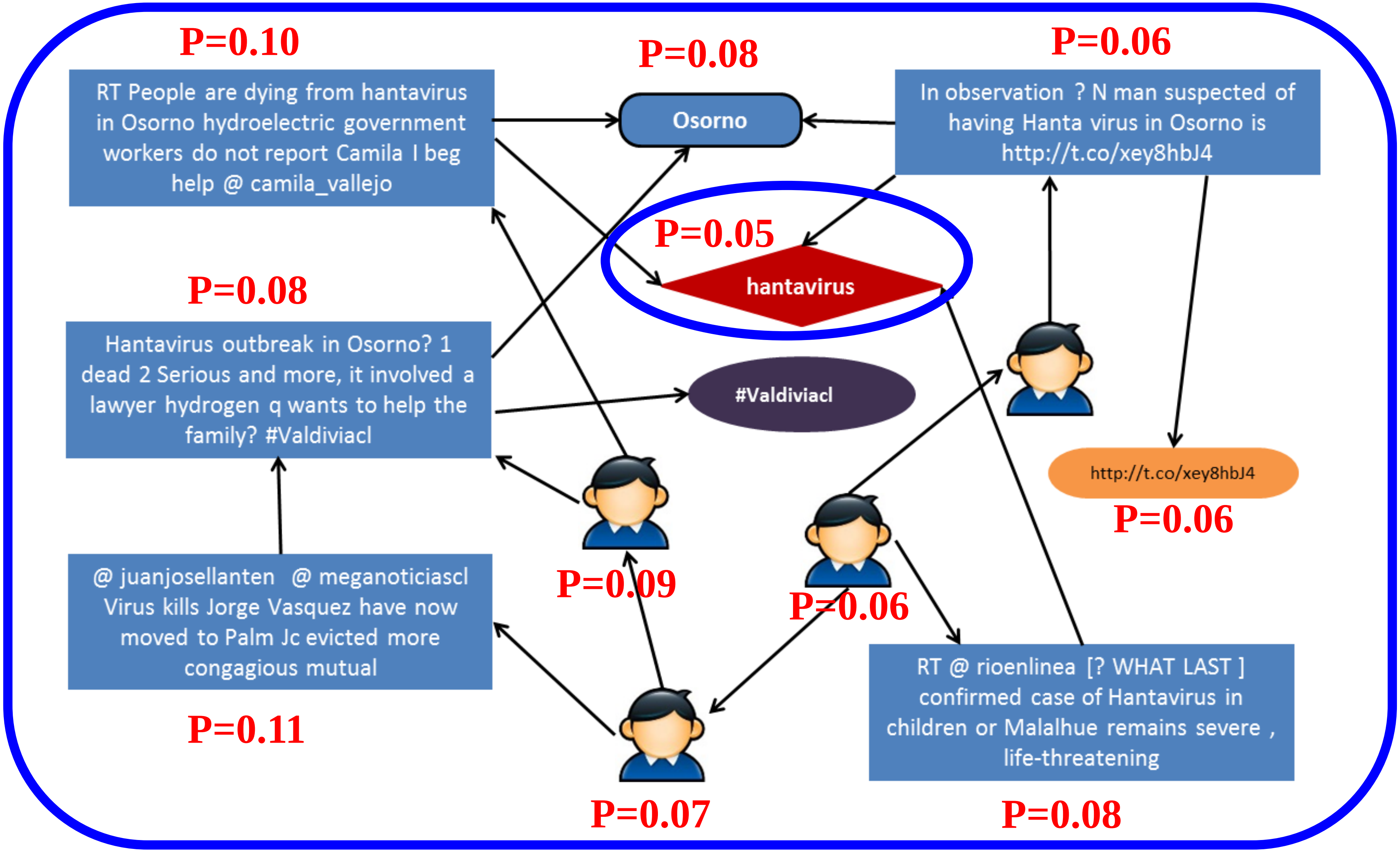


Why we calculate an empirical p-value for each entity node?

1. **P-value is uniformly distributed between 0 and 1 under null even the true distribution is unknown**
2. **Entities of different types can be evaluated consistently based on their p-values**
3. **Empirical p-value is a nonparametric and computationally convenient approach to estimate p-value**



Step 2: Sensor Signals → Empirical P-values



P=0.10

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

P=0.08

Osorno

P=0.06

In observation ? N man suspected of having Hanta virus in Osorno is http://t.co/xey8hbJ4

P=0.05

hantavirus

P=0.08

Hantavirus outbreak in Osorno? 1 dead 2 Serious and more, it involved a lawyer hydrogen q wants to help the family? #Valdiviacl

#Valdiviacl

http://t.co/xey8hbJ4

P=0.06

@ juanjosellanten @ meganoticiascl Virus kills Jorge Vasquez have now moved to Palm Jc evicted more congagious mutual

P=0.11

P=0.09

P=0.06

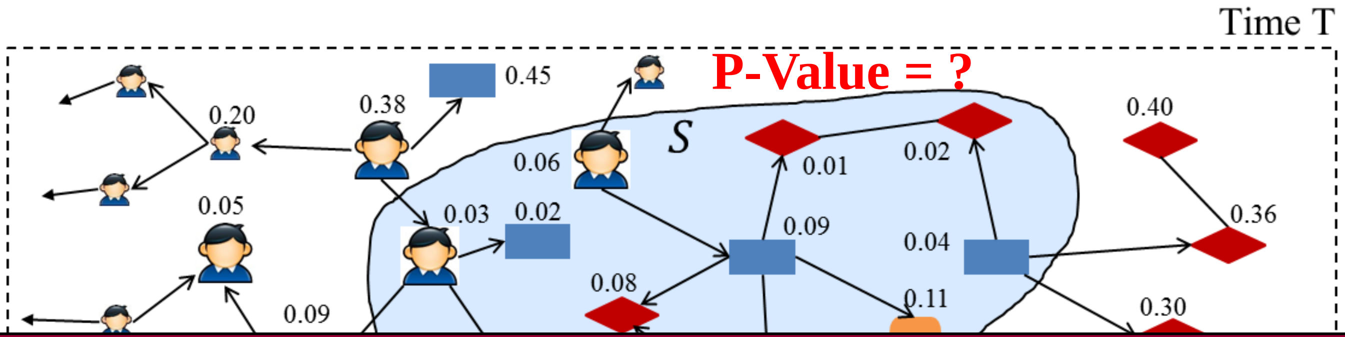
P=0.07

RT @ rioenlinea [? WHAT LAST] confirmed case of Hantavirus in children or Malalhue remains severe , life-threatening

P=0.08

As a group, what is the p-value? (< 0.05)

Step 2: Sensor Signals → Non-Parametric Statistics



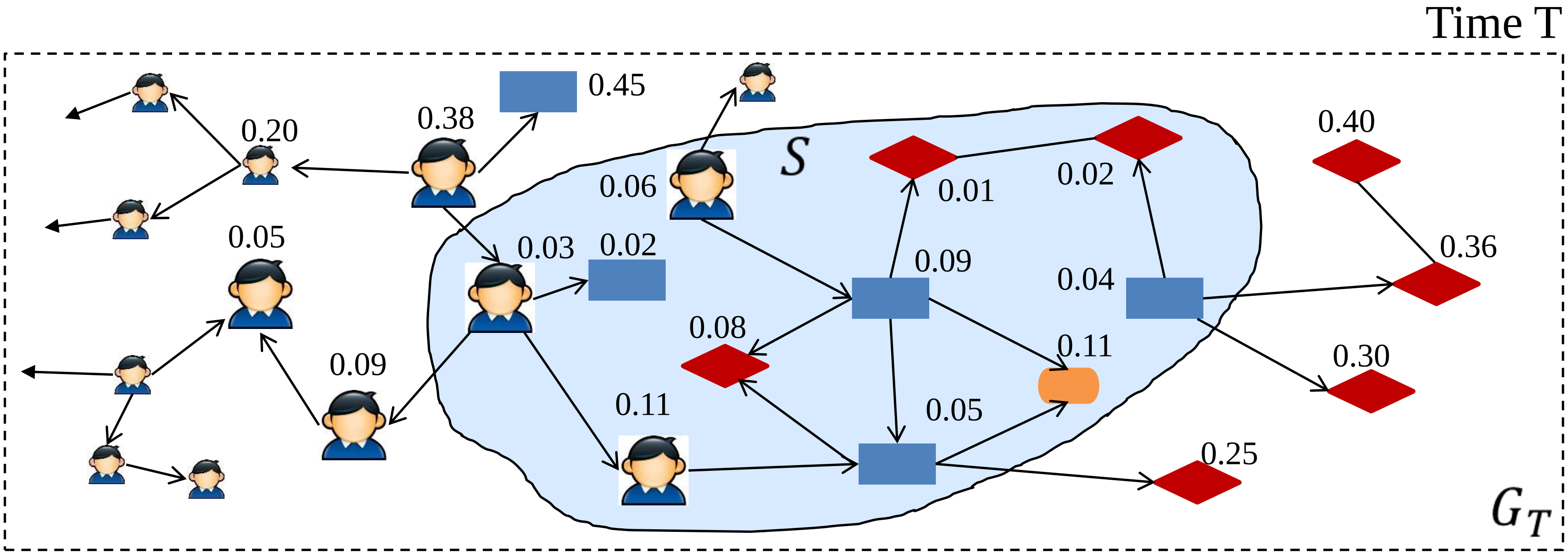
Why we consider non-parametric statistics?

1. A score function to measure a group of interesting nodes
2. Computationally very efficient
3. Asymptotic convergence to the true group p-value
4. Special cases:
    1. Burst detection of keyword volume
    2. Spatial Event Detection based on tweet counts in spatial regions

Sig

→ Kullback-Leibler Divergence

Step 3: Nonparametric Scan on “Sensor Network”

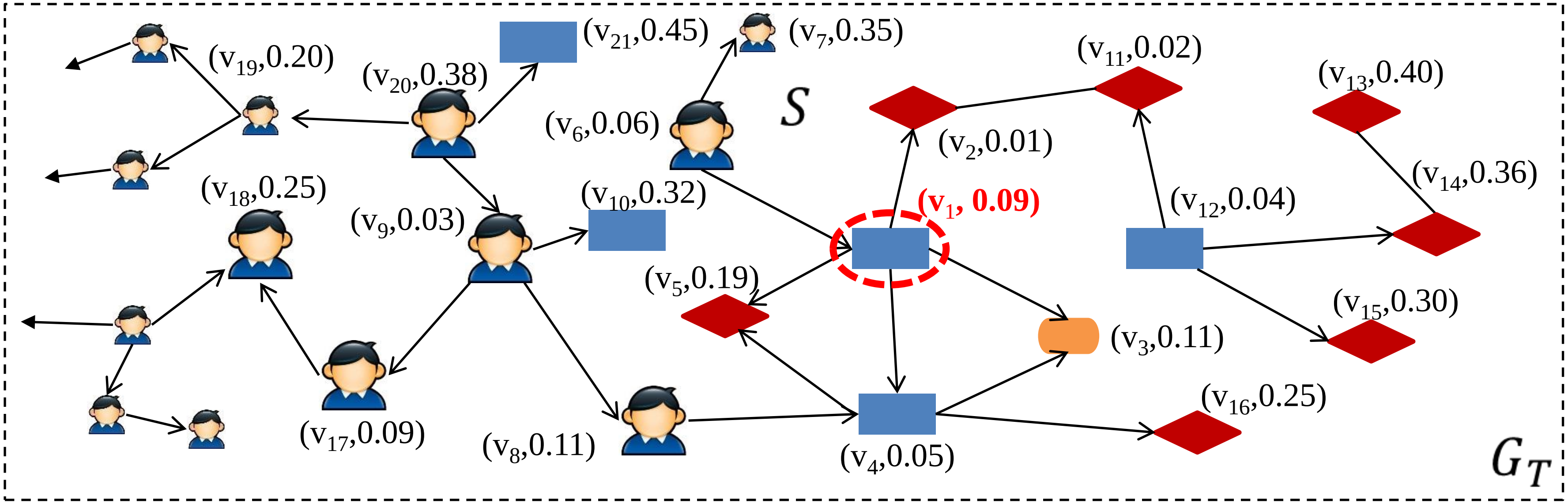


$$S^\star = \underset{S \in V_T, S \text{ is connected}}{\operatorname{argmax}} \boxed{F(S)}$$

We propose novel nonparametric scan statistics for connected sub-graphs, and an approximate algorithm with time cost $O(|V_T| \log |V_T|)$.

Step 3: Nonparametric Scan Algorithm

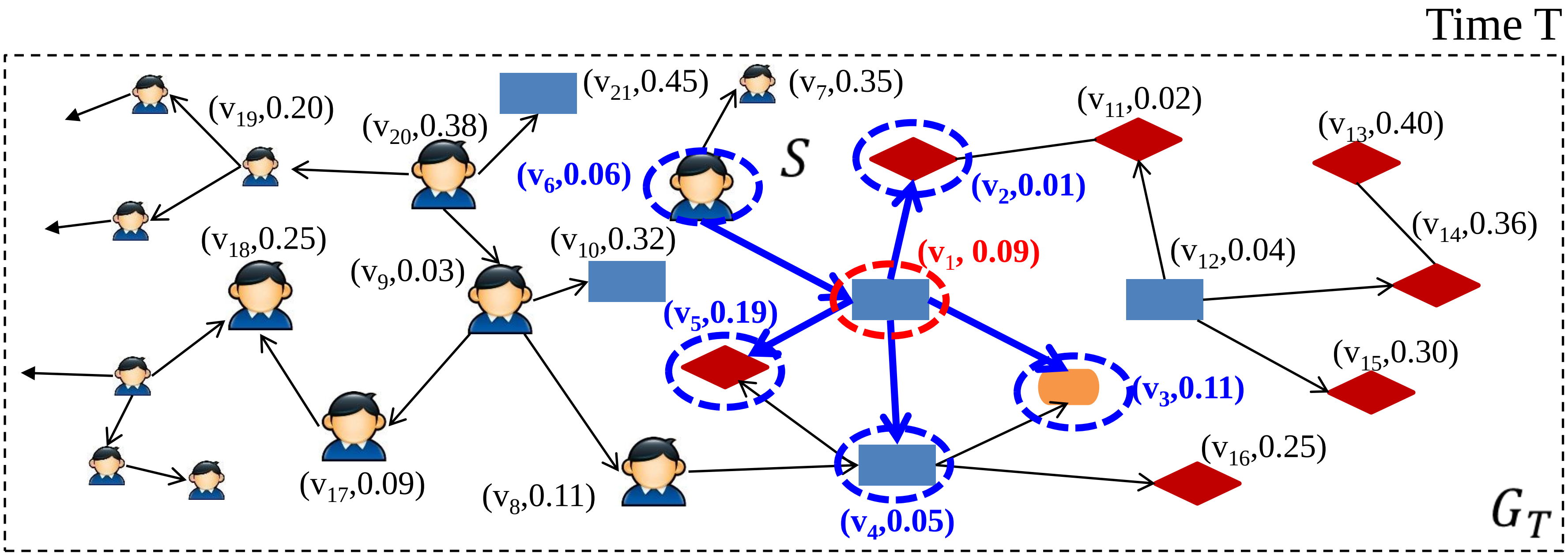
Time T



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

In this example, we start from the seed set $\hat{S} = \{(v_1, 0.09)\}$.

Step 3: Nonparametric Scan Algorithm



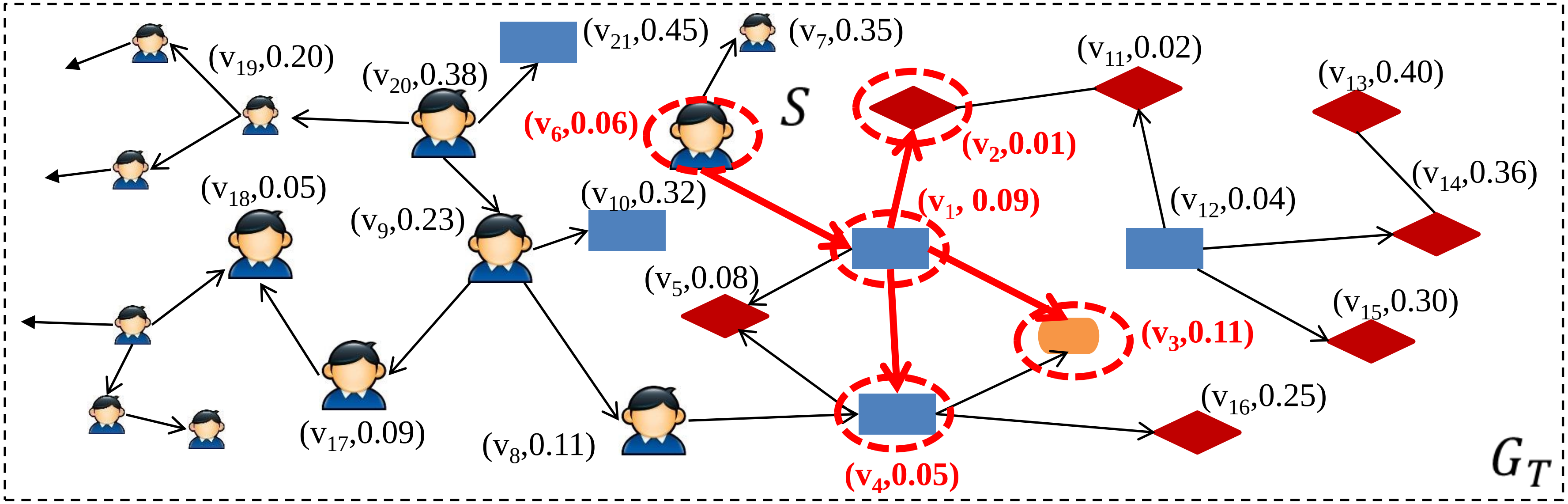
Expand Ŝ by adding the neighbor nodes:

$$\hat{S} = \{(v_1, 0.09), (v_2, 0.01), (v_3, 0.11), (v_4, 0.05), (v_5, 0.19), (v_6, 0.06)\}$$

$$\mathrm{S}^\star = \arg\max_{S\subset\hat{S}} F(S) = \arg\max_{S^\star\subset S}\left\{\max_{\alpha\le\alpha_{max}} NK\left(\frac{N_\alpha}{N}, \alpha\right)\right\}$$

Step 3: Nonparametric Scan Algorithm

Time T



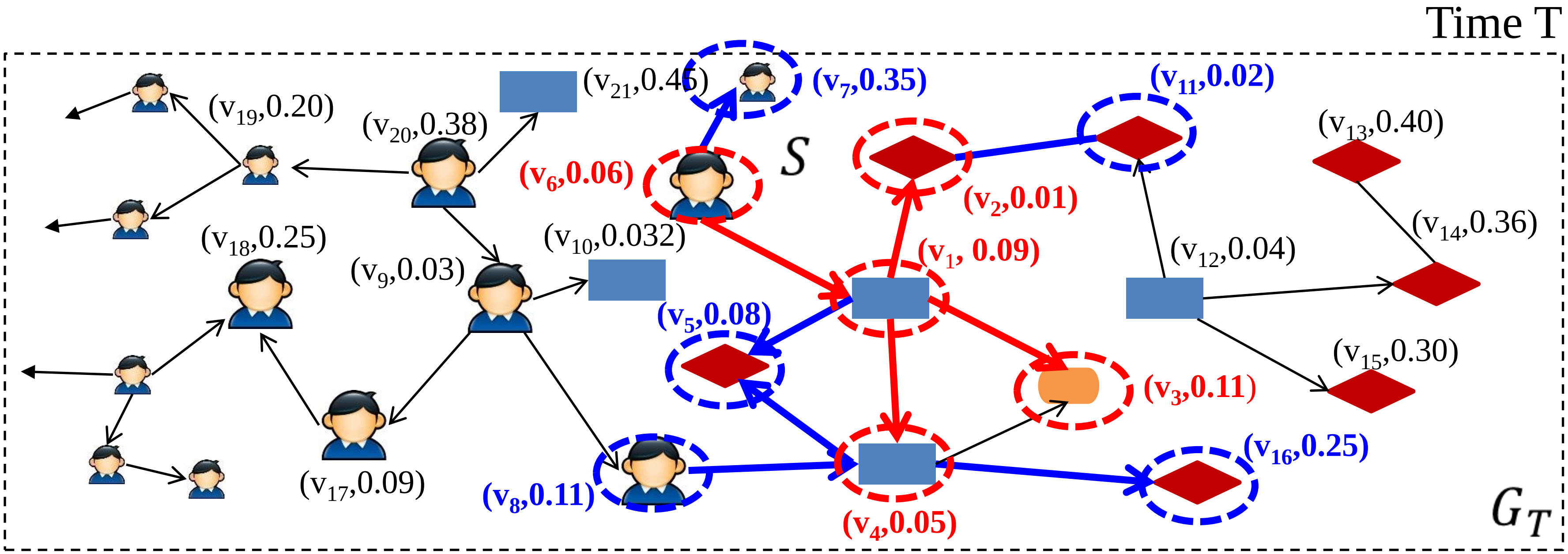
Expand Ŝ by adding the neighbor nodes:

$$\hat{S} = \{(v_1, 0.09), (v_2, 0.01), (v_3, 0.11), (v_4, 0.05), (v_5, 0.19), (v_6, 0.06)\}$$

$$S^{\star} = \arg\max_{S \subset \hat{S}} F(S) = \arg\max_{S^{\star} \subset S} \left\{ \max_{\alpha \leq \alpha_{max}} NK\left(\frac{N_\alpha}{N}, \alpha\right) \right\}$$

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

Step 3: Nonparametric Scan Algorithm

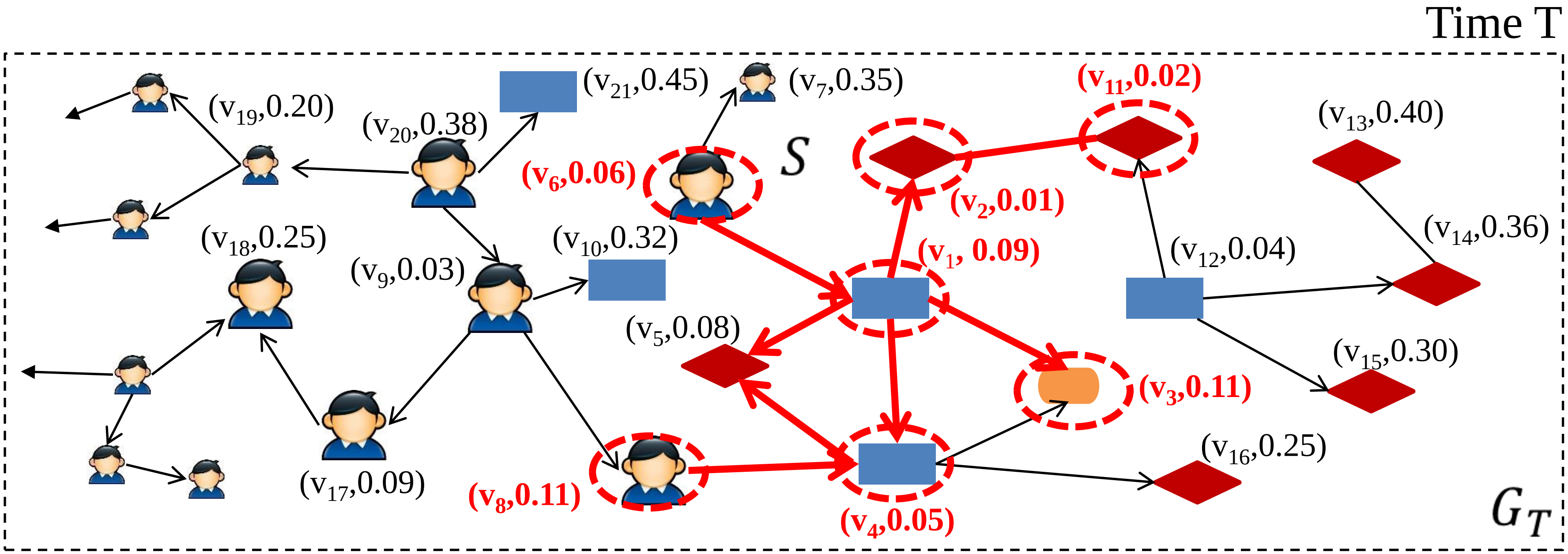


Expand $\hat{S}$ by adding the neighbor nodes:

$$\hat{S} = \{(v_1, 0.09), (v_2, 0.01), (v_3, 0.11), (v_4, 0.05), (v_6, 0.06), (v_5, 0.19), (v_7, 0.35), (v_8, 0.11), (v_{11}, 0.02), (v_{16}, 0.25)\}$$

$$S^\star = \arg\max_{S \subset \hat{S}} F(S) = \arg\max_{S^\star \subset S} \left\{ \max_{\alpha \le \alpha_{max}} NK\left(\frac{N_\alpha}{N}, \alpha\right) \right\}$$

Step 3: Nonparametric Scan Algorithm



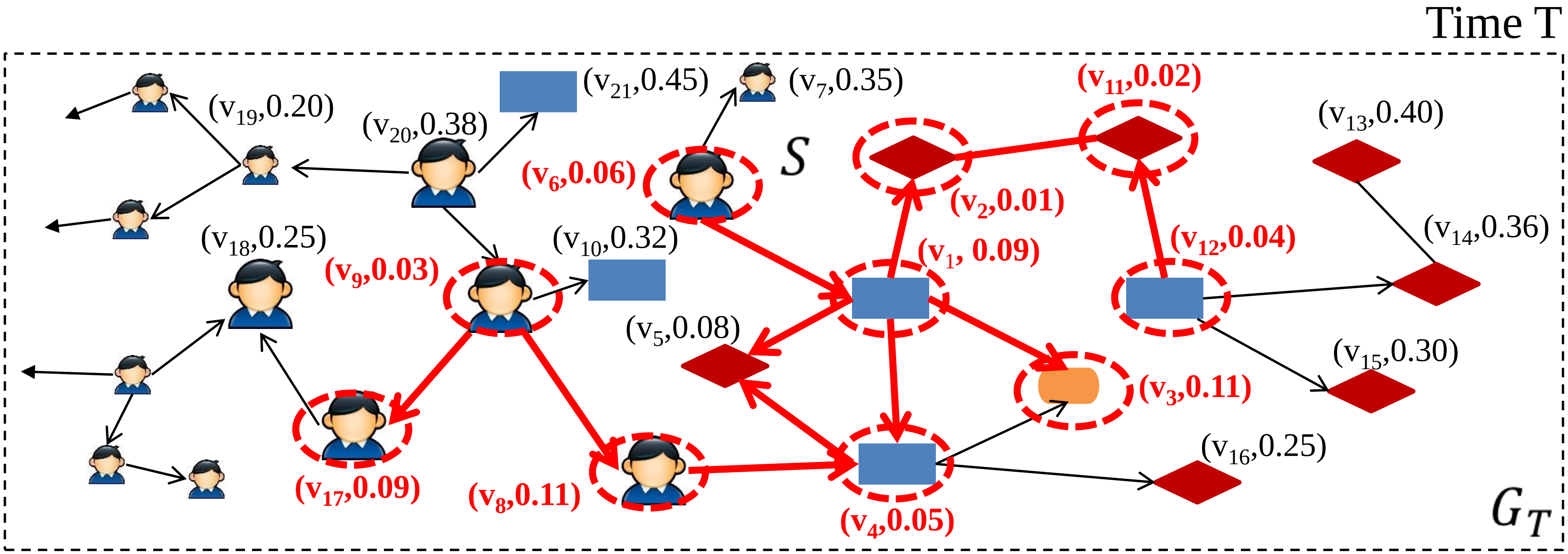
Expand $\hat{S}$ by adding the neighbor nodes:

$$\hat{S} = \{(v_1, 0.09), (v_2, 0.01), (v_3, 0.11), (v_4, 0.05), (v_6, 0.06), (v_5, 0.19), (v_7, 0.35), (v_8, 0.11), (v_{11}, 0.02), (v_{16}, 0.25)\}$$

$$S^\star = \arg\max_{S \subset \hat{S}} F(S) = \arg\max_{S^\star \subset S} \left\{ \max_{\alpha \leq \alpha_{max}} NK\left(\frac{N_\alpha}{N}, \alpha\right) \right\}$$

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

Step 3: Nonparametric Scan Algorithm

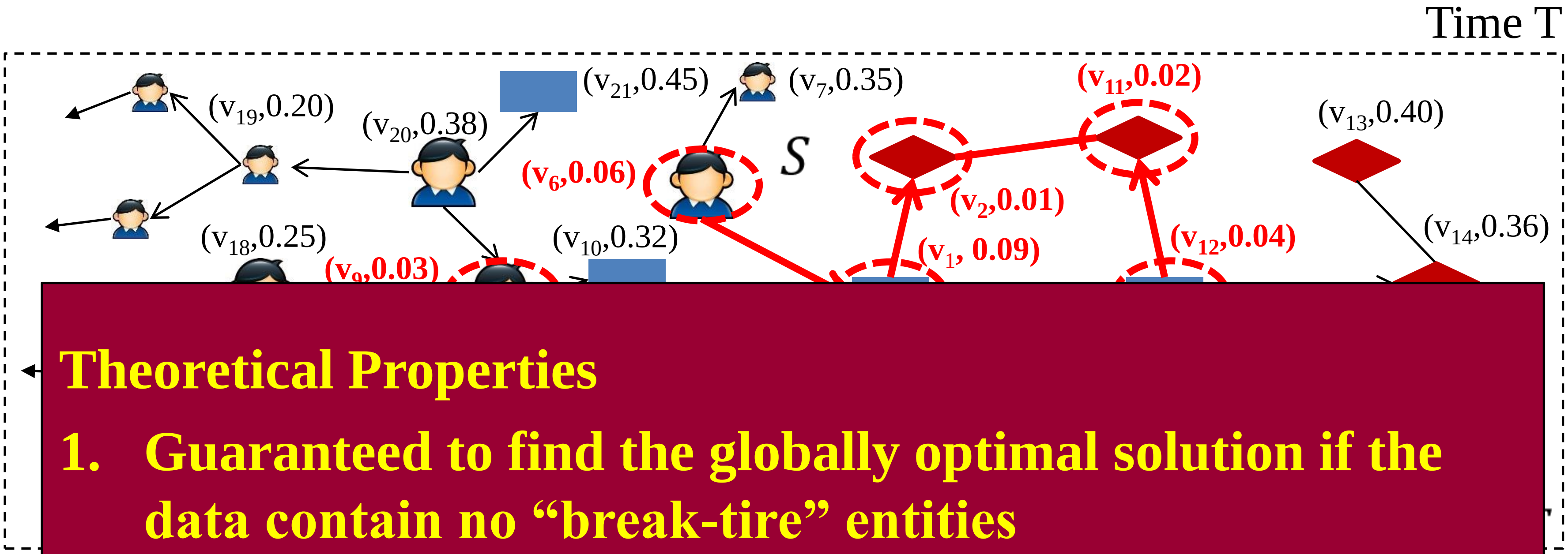


Consider each node as a candidate cluster center (or start point)

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

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

Step 3: Nonparametric Scan 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:

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

Experiment Settings

- Twitter Dataset
  - 10% random sample of public twitter data
  - 17 rare Hantavirus disease outbreaks collected by Chilean Ministry of Health [1] 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

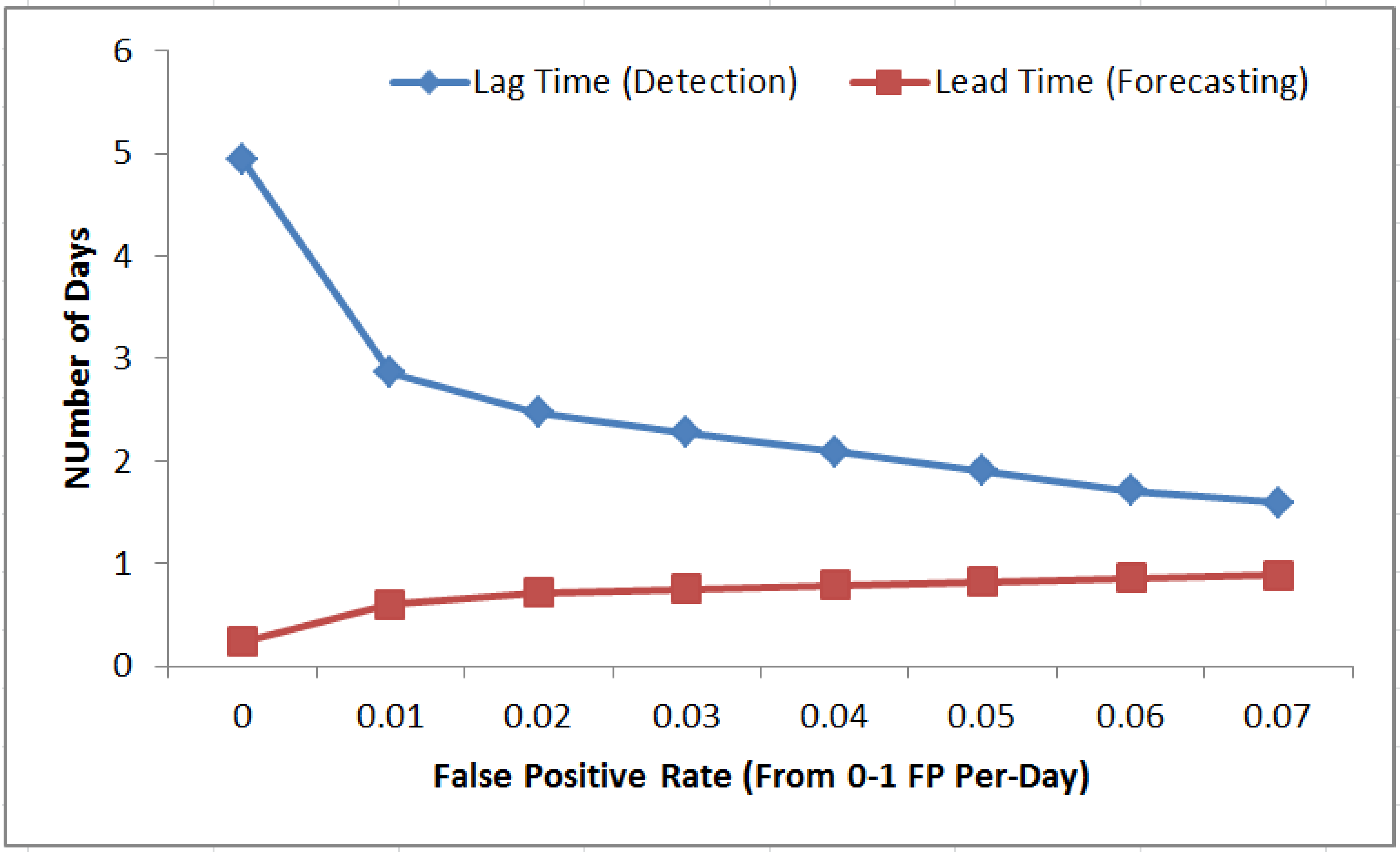
Twitter Dataset

Country	# of tweets	News source*
Chile	14 ,000,000	La Tercera; Las Últimas Notícias; El Mercurio
Colombia	22 ,000,000	El Espectador; El Tiempo; El Colombiano
Ecuador	6,900,000	El Universo; El Comercio; Hoy

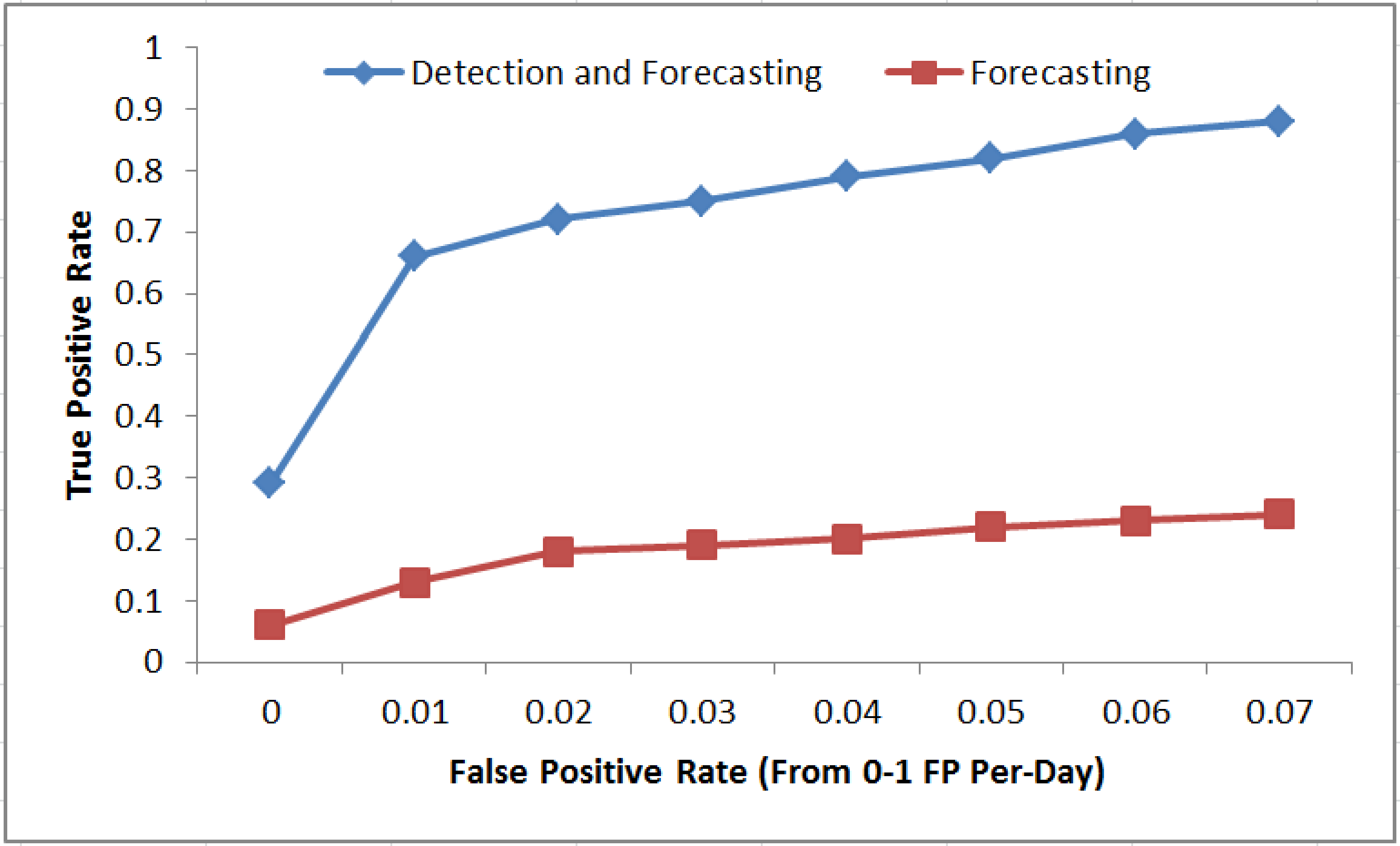
Time Period: From **2012 Jul.** to **2012 Dec.** Totally **918** civil unrest events

Example of an event label: (PROVINCE = “El Loa”, COUNTRY = “Chile”, DATE = “2012-05-18”, DESCRIPTION = “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”, FIRST-REPORTEDLINK = “http://www.pressenza.com/2012/05/march-ofdignity-in-mining-capital-of-chile/”).

Detection Lag Time and Prediction Lead Time



Detection and Forecasting Results



Conclusion

- Social media is real-time, very informal, and dynamic
- We argue that nonparametric methods are better suited to social media than parametric methods
- We propose a nonparametric graph scan statistics approach to the forecasting and detection of disease outbreaks using social media

Thank you!

Questions?