Mass Movements and Their Adoption in Social Media

Fang Jin

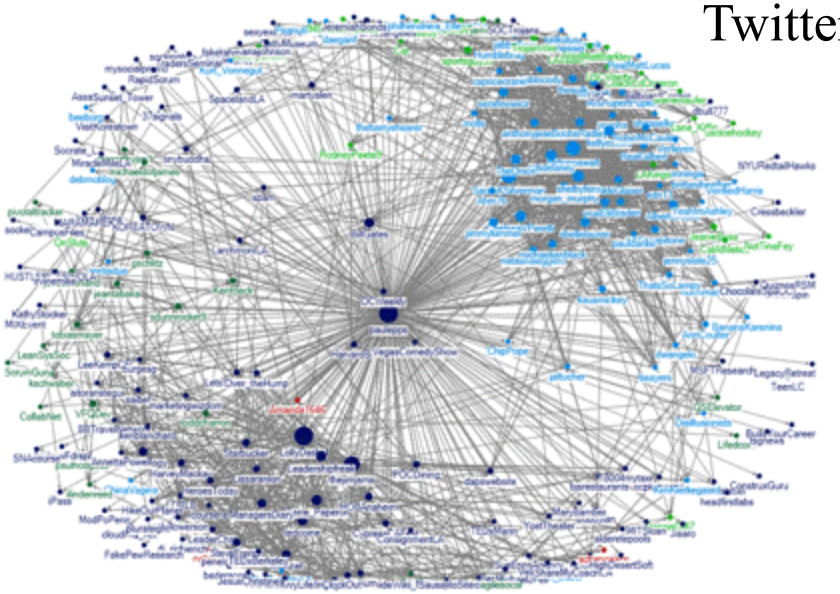
Assistant Professor

Department of Computer Science, Texas Tech University

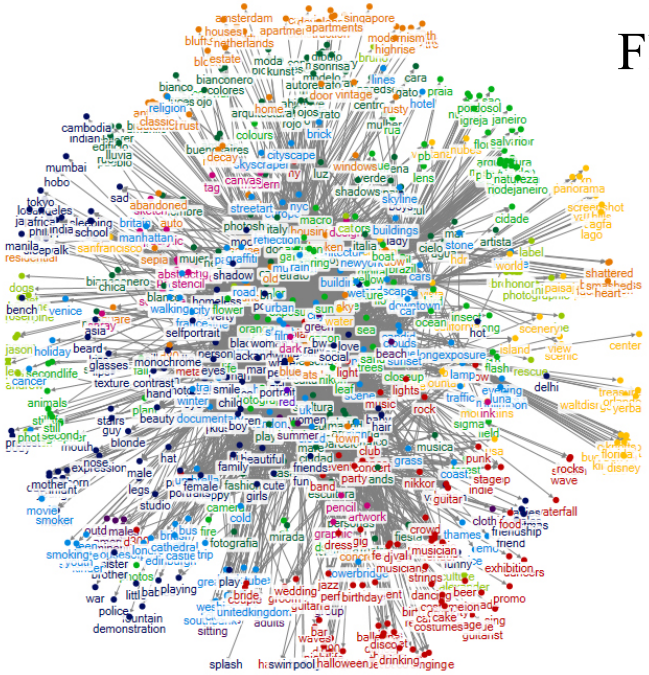
Ubiquity of social media

Facebook

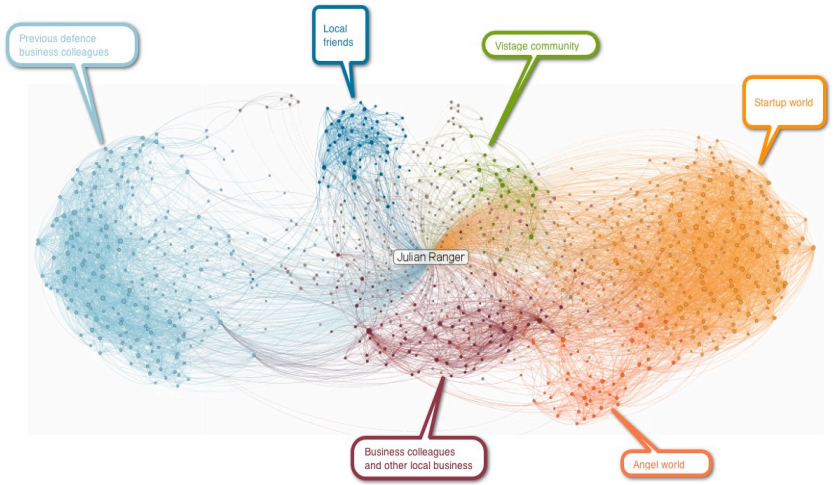
Twitter users



Flickr tags



LinkedIn network



Big data research on social networks

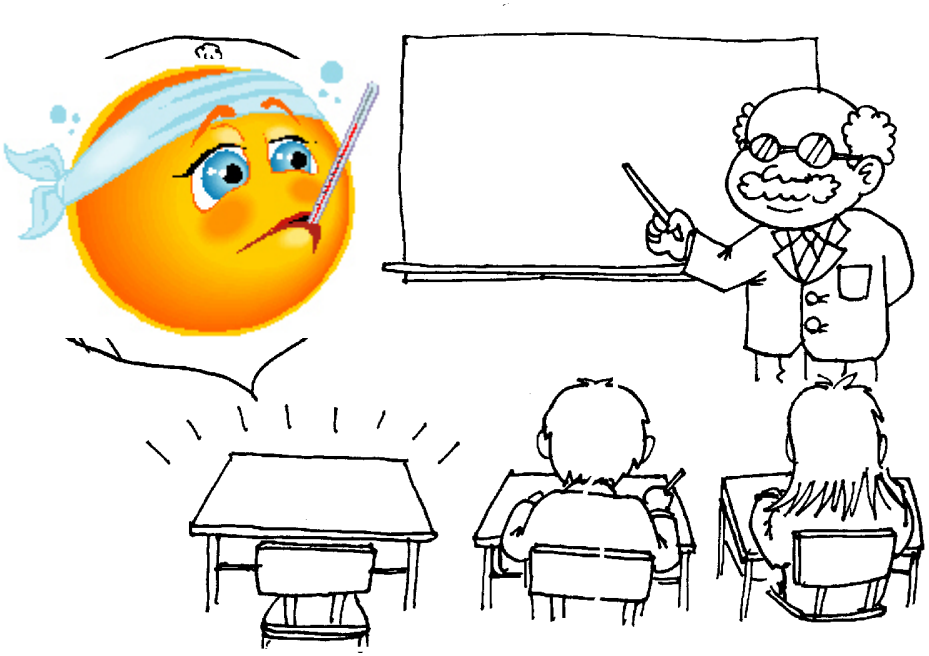
1. How do we identify group anomaly?

2. How do we detect civil unrest events in social networks?

3. How do we distinguish rumors from real news?

Group Absenteeism
as a basis for Event Detection

Motivation

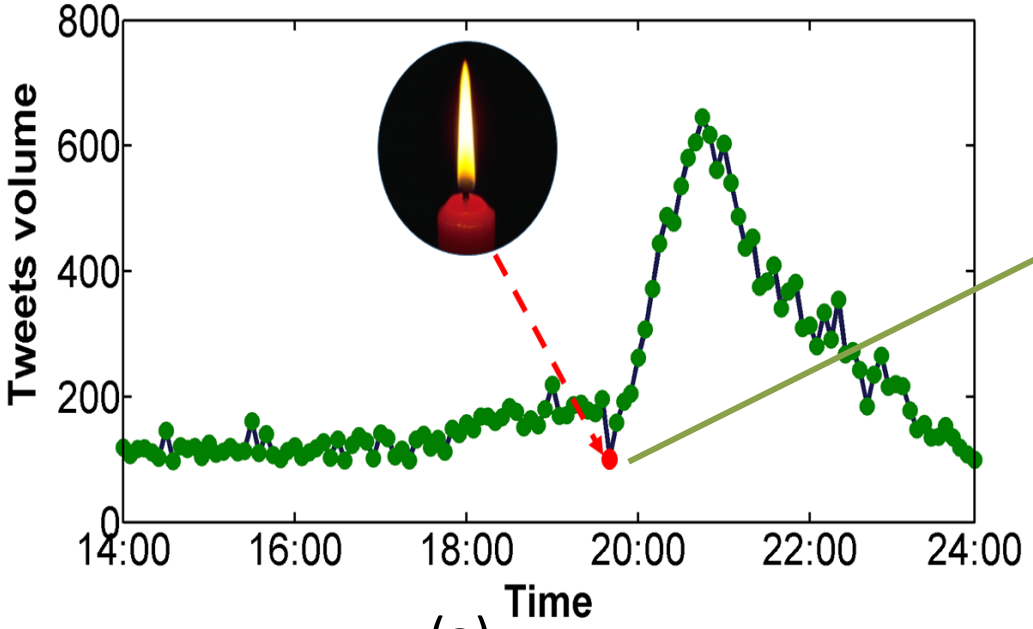


Student absent

Information absenteeism

How to detect group absenteeism on Twitter?

Why study absenteeism?

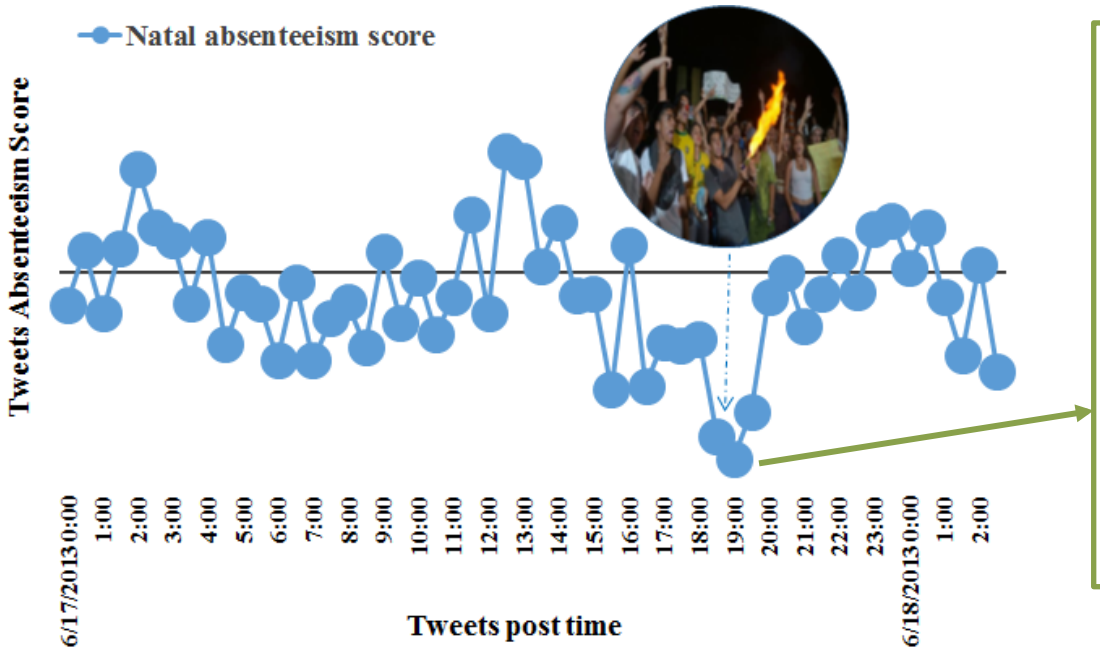


(a)

(b)

Caracas, Venezuela power cut on 2013-12-02, 8:00 PM

Why study absenteeism?

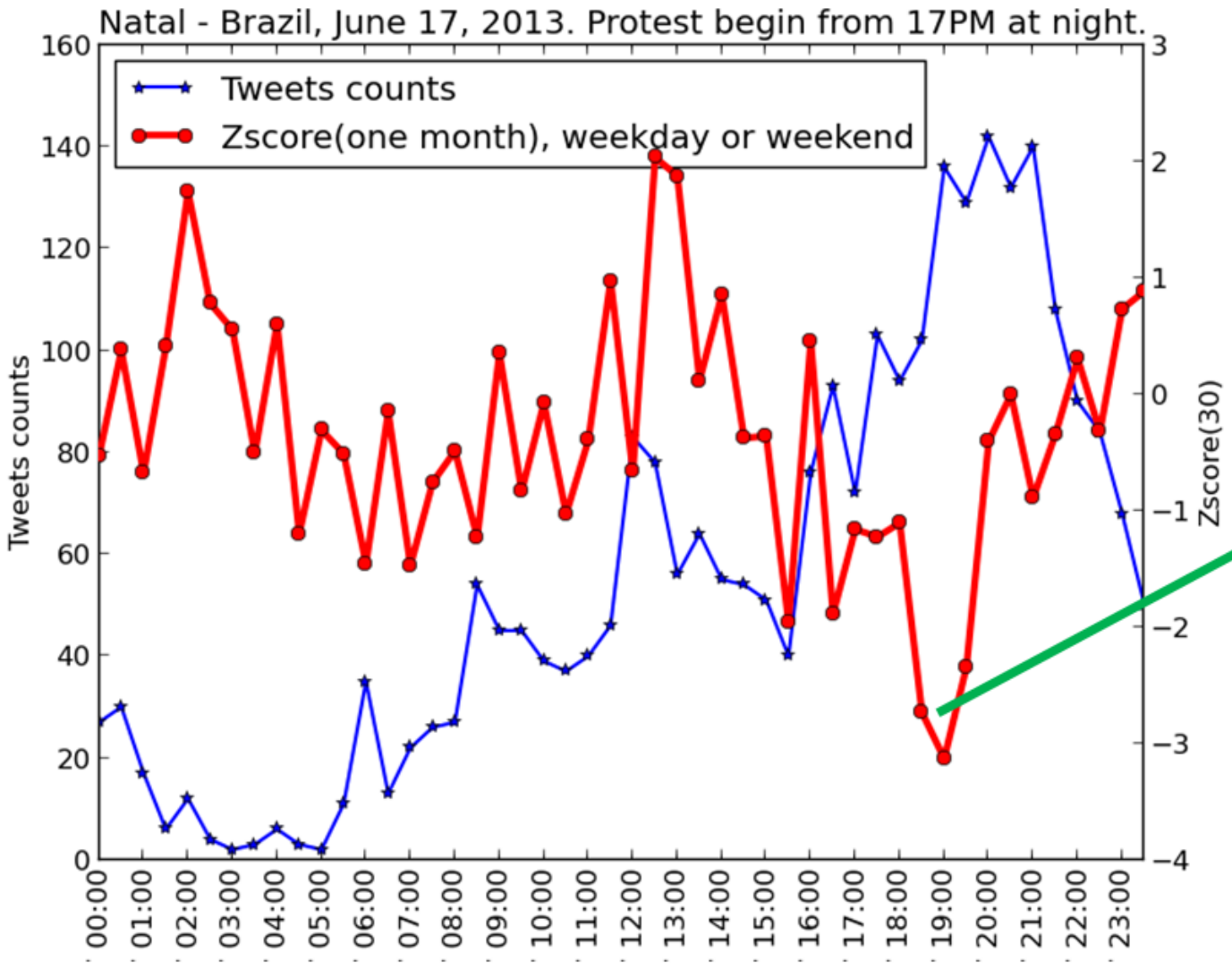


(a)

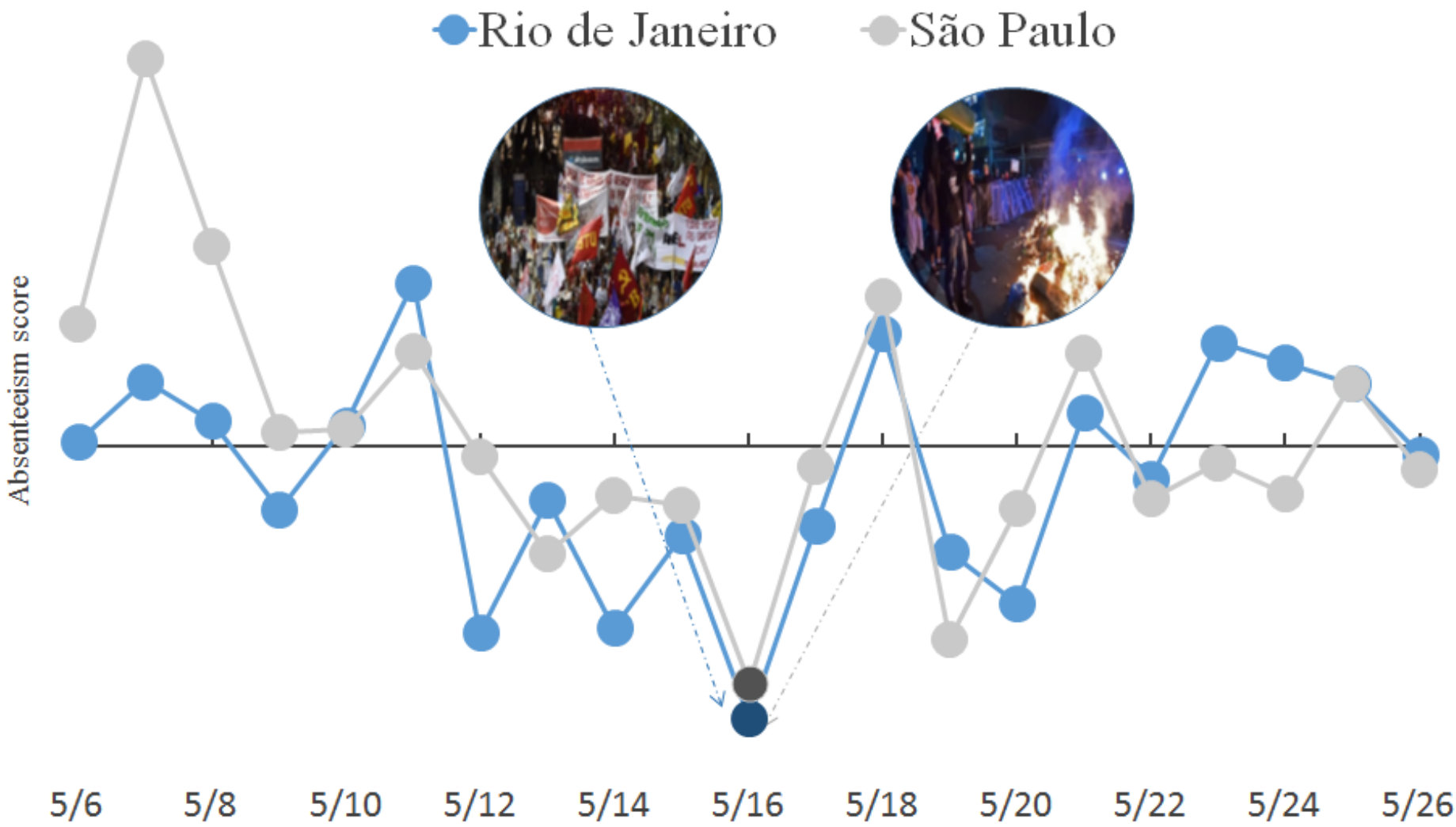
(b)

Natal, Brazil protest on 2013-06-17, 18:00 – 20:00

Why study absenteeism?

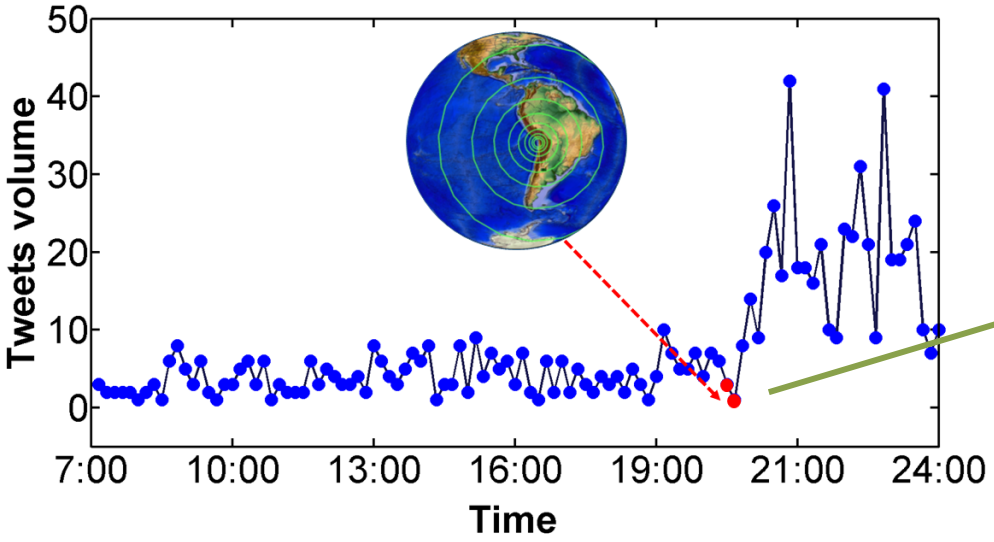


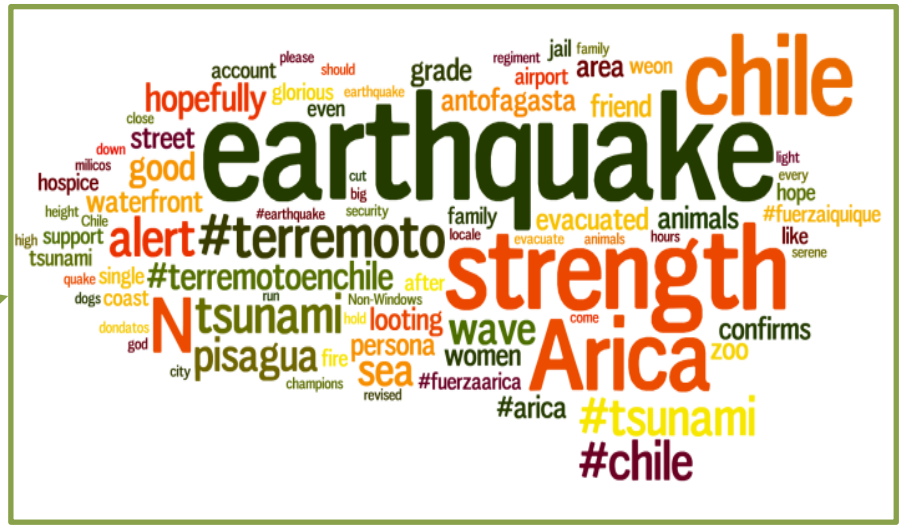
Why study absenteeism?



Protests in Brazil against world cup, 2014

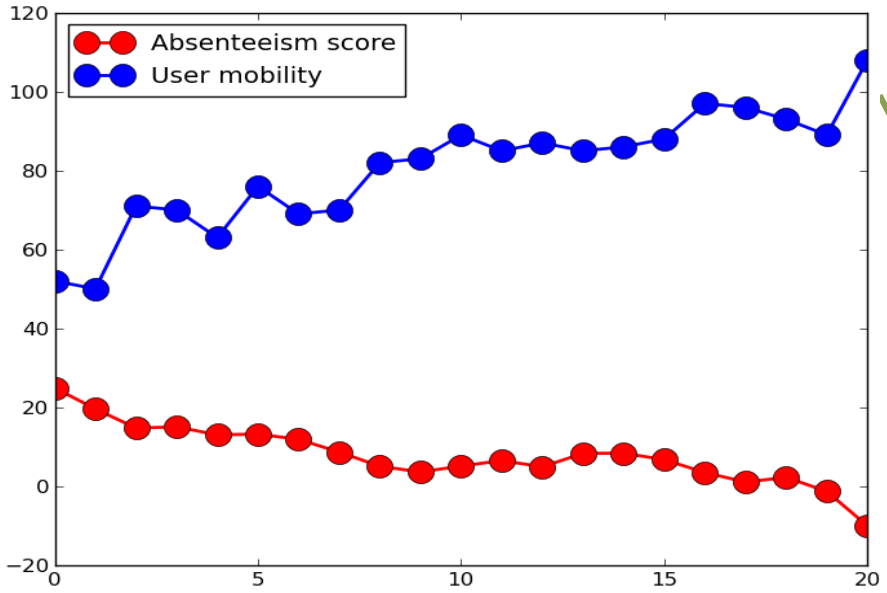
Why study absenteeism?





Chile Iquique earthquake on 2014-04-01

Why study absenteeism?

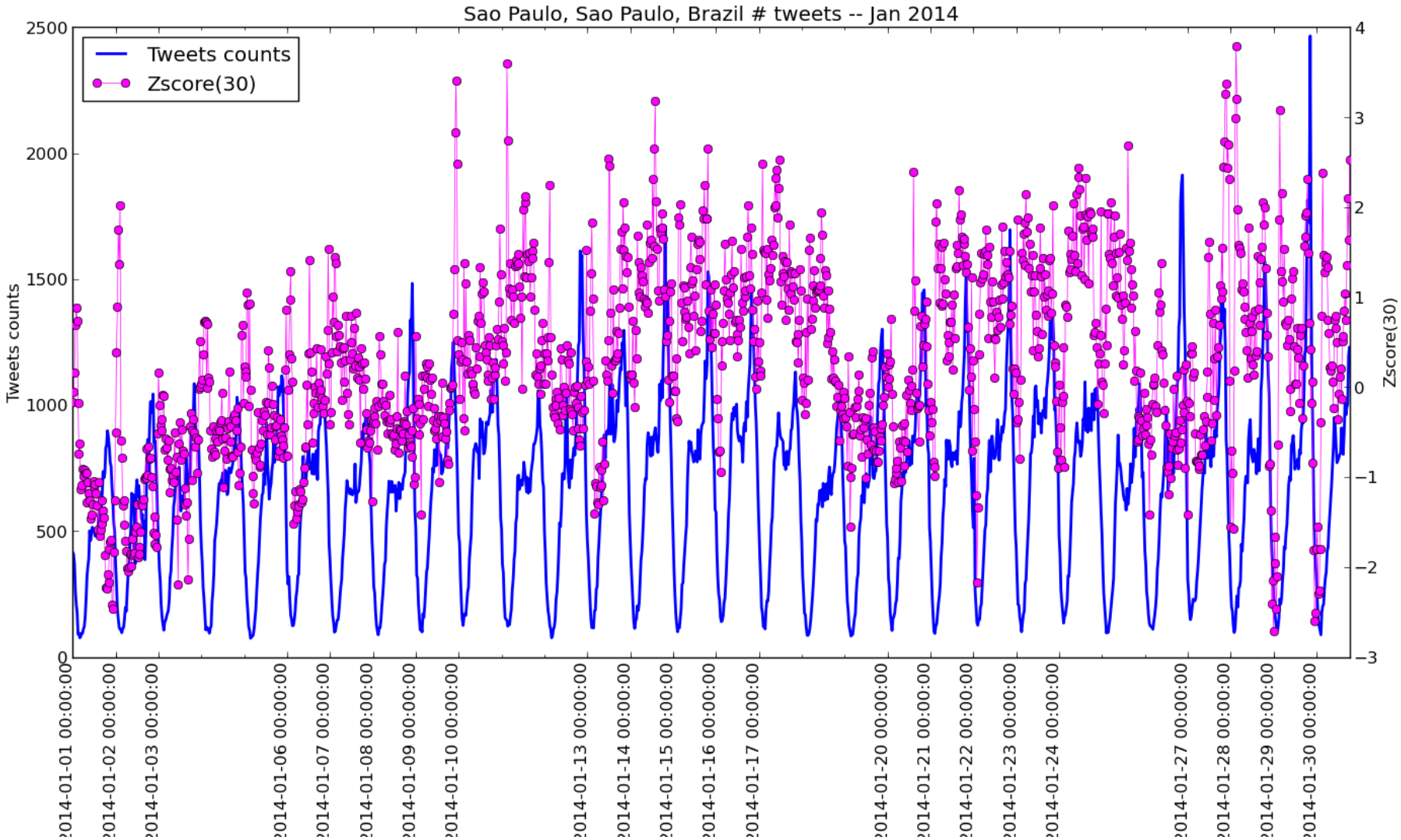


(a)

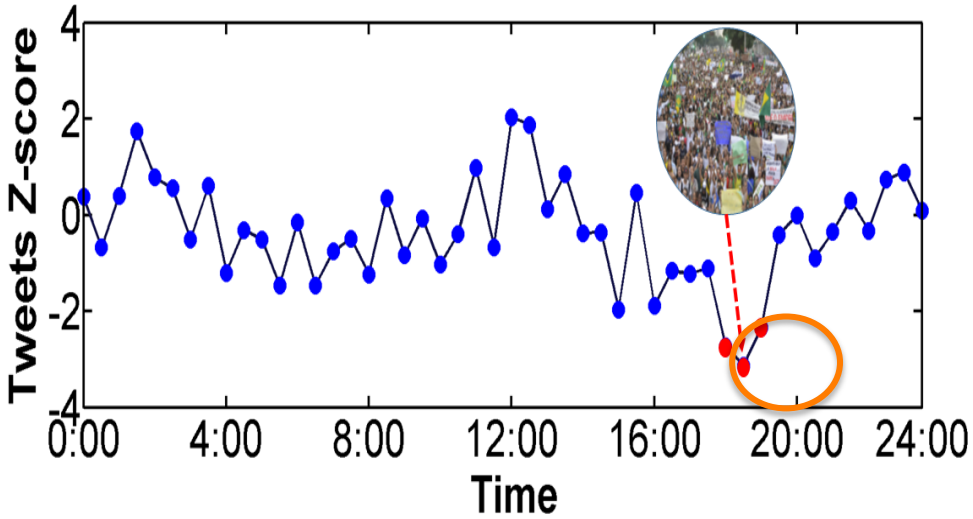
(b)

Argentina, Christmas holiday on 2015-12-25

Absenteeism score

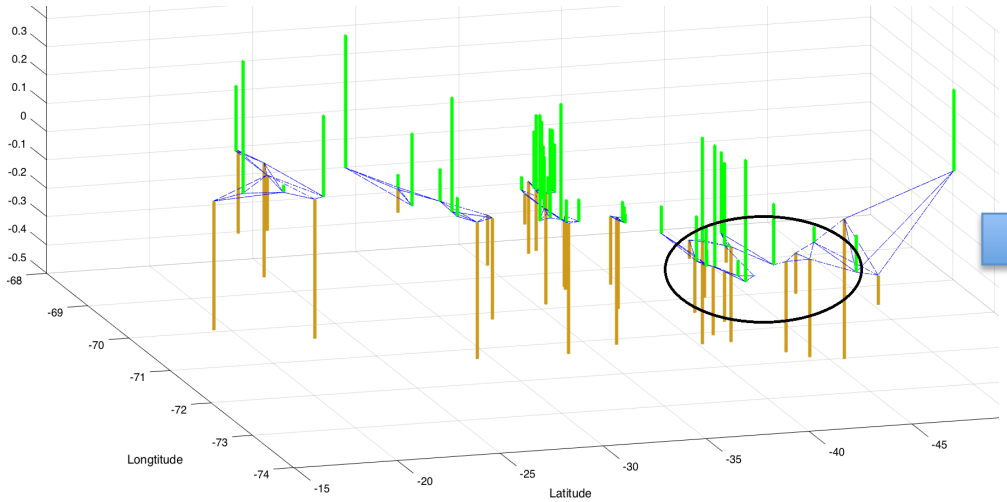


Motivation



- Absenteeism score (normalization of Tweeter volumes).
- Absenteeism score vector f(n) on graph G.

Natal, Brazil protest began at 18 PM on June 17, 2013



How to find a group of cities with uniform anomaly?

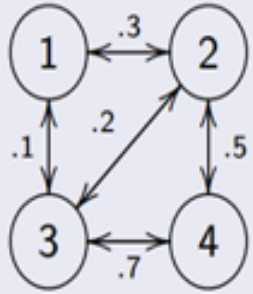
Absenteeism score distribution vector f(n) on April 1, 2014 in Chile.

Our approach

1. Graph wavelet based approach, considering both the graph structure and the vector f;
2. Define an anomaly index of f's distribution on G;
3. Identify abnormal locations using graph wavelet;

Graph spectrum

Connected, undirected, weighted graph $G(V, W; f)$



Degree matrix D: zeros except diagonals, which are sums of weights of edges incident to corresponding node

Non-normalized Laplacian: $L := D - W$

$$W = \begin{bmatrix} 0 & .3 & .1 & 0 \\ .3 & 0 & .2 & .5 \\ .1 & .2 & 0 & .7 \\ 0 & .5 & .7 & 0 \end{bmatrix}$$

Complete set of orthonormal eigenvectors and associated real, non-negative eigenvalues:

$$D = \begin{bmatrix} .4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1.2 \end{bmatrix}$$

$$L\chi_l = \lambda_l \chi_l$$

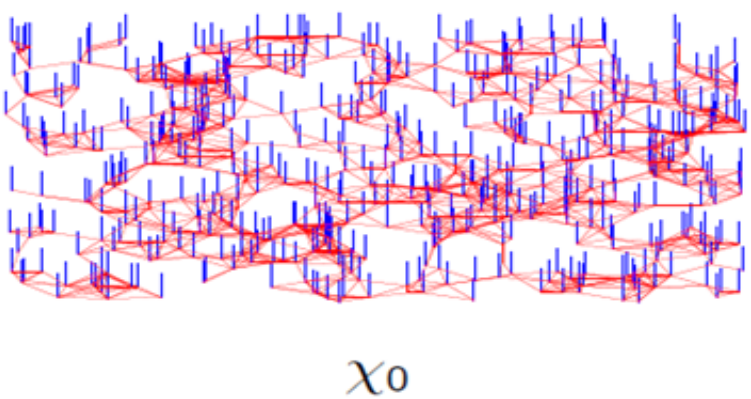
$$0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{N-1} := \lambda_{max}$$

$\sigma(\mathbf{G}) := \{(\lambda_l, \chi_l)\}_{l=0}^{N-1}$ is graph spectrum.

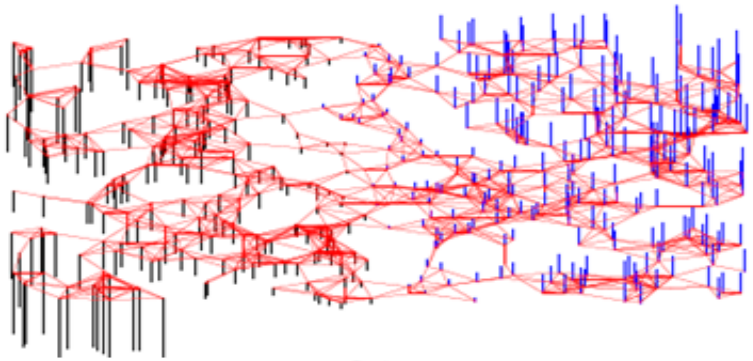
Shuman, David I., Benjamin Ricaud, and Pierre Vandergheynst. "A windowed graph Fourier transform." *Statistical Signal Processing Workshop (SSP), 2012 IEEE*. Ieee, 2012.

Eigenvalue and eigenvector property (1)

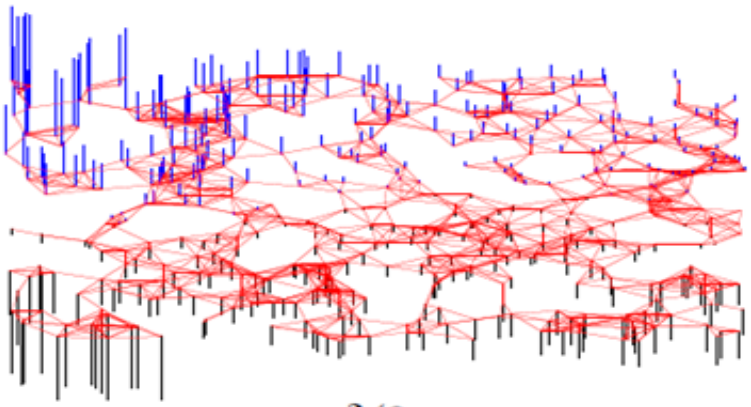
The set of eigenvector represents N types' pattern of graph G



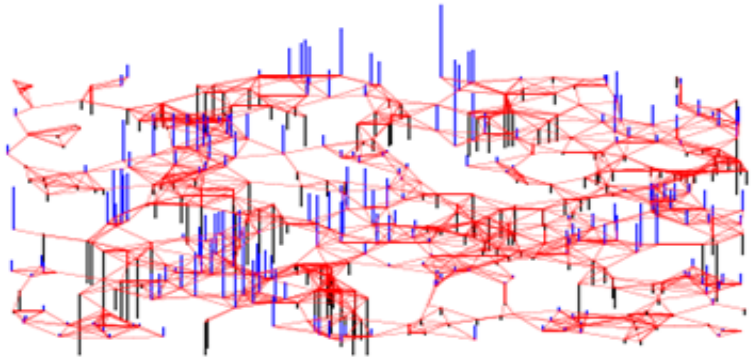
χ_0



χ_1



χ_2



χ_{50}

The larger eigenvalue corresponds to a severe fluctuation.

Shuman, David I., Benjamin Ricaud, and Pierre Vandergheynst. "A windowed graph Fourier transform." *Statistical Signal Processing Workshop (SSP), 2012 IEEE*. Ieee, 2012.

Eigenvalue and eigenvector property (2)

f decomposition on the eigenvectors χ_l:

$$f(n) = \hat{f}(0)\chi_0 + ... + \hat{f}(l)\chi_l + ... + \hat{f}(N-1)\chi_{N-1}$$

$$\hat{f} = \begin{pmatrix} < \chi_0, f > \\ < \chi_1, f > \\ < \chi_2, f > \\ ... \\ < \chi_{N-1}, f > \end{pmatrix}$$

1) $\chi_0, \chi_1, \chi_2, ..., \chi_{N-1}$: N types of pattern for any vector f defined on graph G.

2) For larger λ_l, χ_l has larger vibration.

$$\lambda_l = \chi_l^T \lambda_l \chi_l = \sum_{e_{mn} \in E} w_{mn}[\chi_l(m) - \chi_l(n)]^2$$

3) $\hat{f}(l)$ represents the similarity of χ_l and f.

Anomaly index on graph

1. Define the eigenvector χ_l anomaly index:

$$\gamma_f(l;\mathbf{G}) = \lambda_l \hat{f}^2(l) = \lambda_l < f, \chi_l >^2$$

2. Define the global anomaly index of f on G:

$$\gamma_f(\mathbf{G}) = \max_{0 \le l \le N-1} \gamma_f(l;\mathbf{G})$$

$\gamma_f(G)$ depends on two parts:

(1) the eigenvalue which reflects the deviations of χ_l;

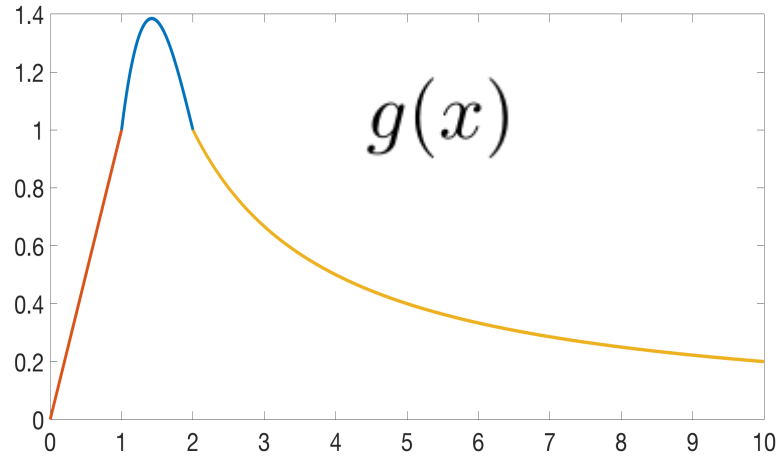
(2) the $|\hat{f}(l)|^2$ which represents the power of χ_l in f.

Graph wavelet construction

$$f(n) = \sum_{l=0}^{N-1} \hat{f}(l)\chi_l(n)$$

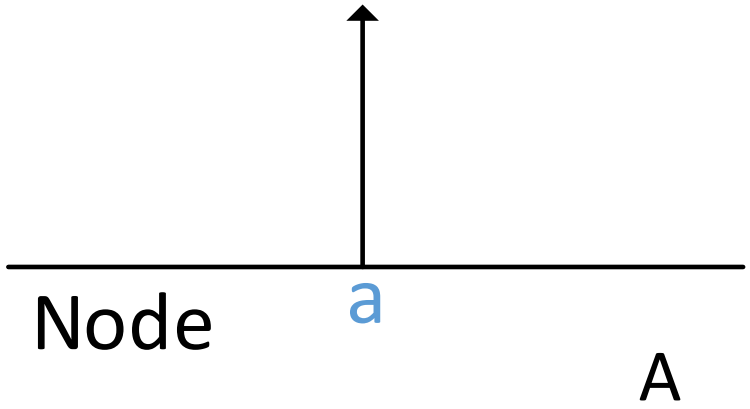
$g(x)$ is a kernel function, which can be treated as a filter, so the original $f(n)$ now become $f'(n)$:

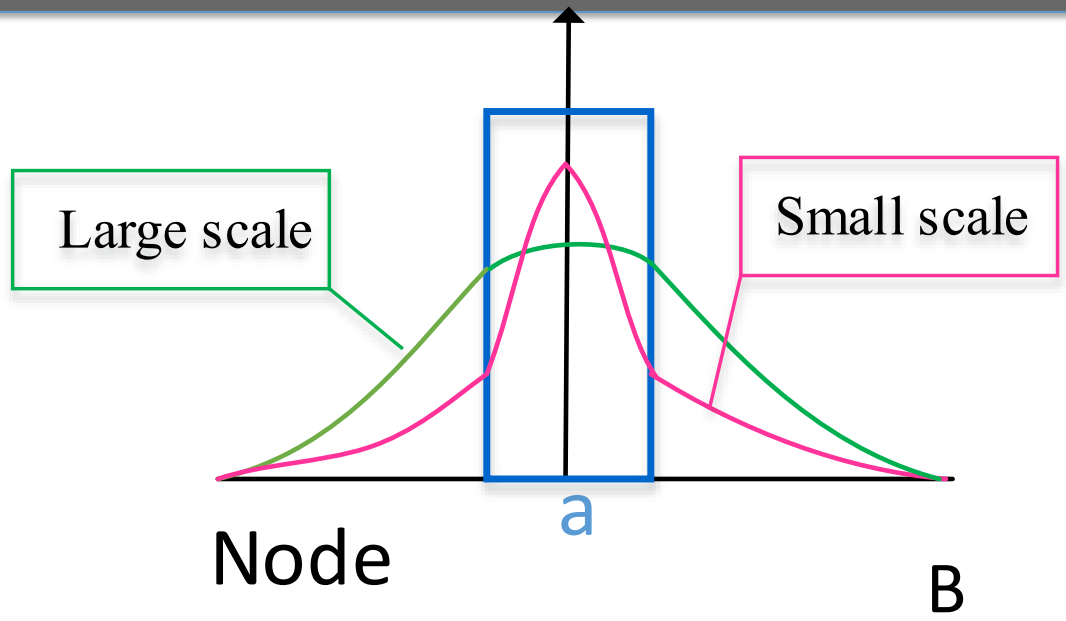
$$f'(n) = \sum_{l=0}^{N-1} g(s\lambda_l)\hat{f}(l)\chi_l(n)$$

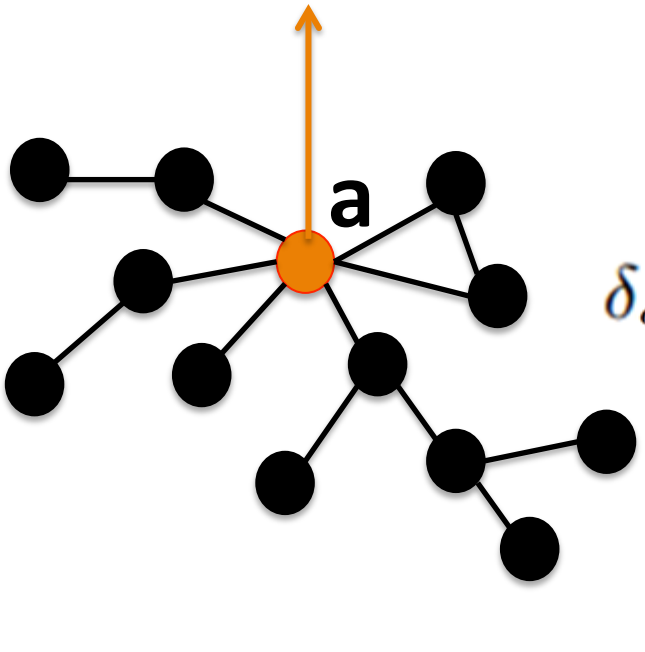


When $f(n) = \delta_a(n)$, $f'(n)$ is written as $\psi_{s,a}(n)$, we call it graph wavelet.

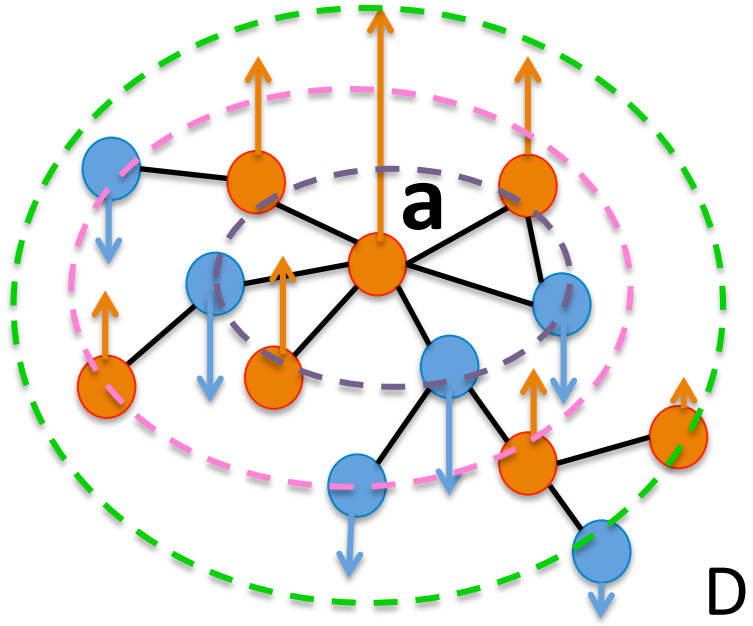
Graph wavelet property (1)







$$\delta_a \xrightarrow{s} \psi_{s,a}(n)$$



Graph wavelet coefficient

The wavelet coefficients for f can be defined as:

$$W_f(s,a) =< \psi_{s,a}, f >= \sum_{l=0}^{N-1} g(s\lambda_l)\hat{f}(a)\chi_l(n)$$

$W_f(s,a)$ is a constant, the largest $W_f(s,a)$ means high similarity between f and $\psi_{s,a}$, means f's energy focus on the area centered around v_a, with scale s.

f(n) can be recovered by the wavelet coefficients:

$$f(n) = \sum_{s=1}^{j}\sum_{a=1}^{N} W_f(s,a)\cdot\psi_{s,a}(n)$$

Graph wavelet property (2)

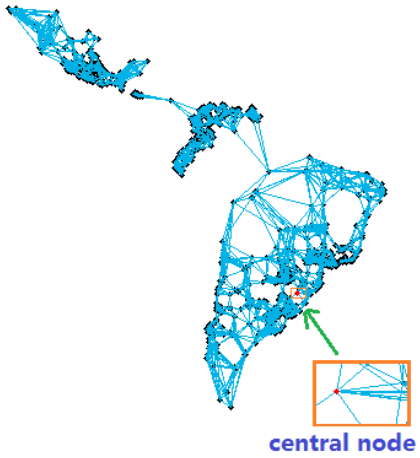
Localization in vertex domains. Given a central vertex v_a and its graph wavelet $\psi_{s,a}(n)$, let v_n be an vertex of $\mathbf{G}$ with $d_G(n,a) > K$, then there exist constants D and β, such that

$$\frac{|\psi_{s,a}(n)|}{||\psi_{s,a}||} \leq Ds$$

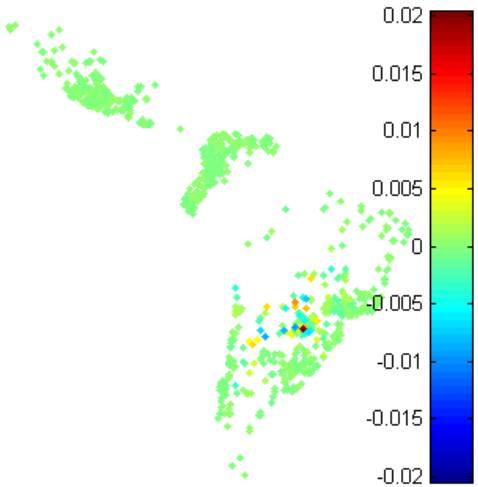
for vertex v_n which is far away form vertex v_a, its wavelet value is linearly attenuated by scale s.

g is $K+1$ times continuous differentiable.
$d_G(n,a)$ is the is the minimum number of edges that connect vertices v_n and a.

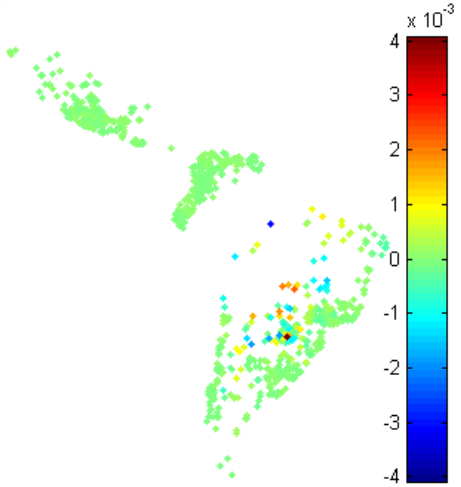
Graph wavelet scale example



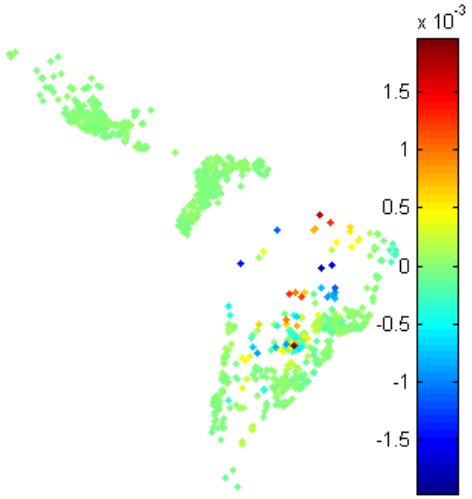
(a) Center node



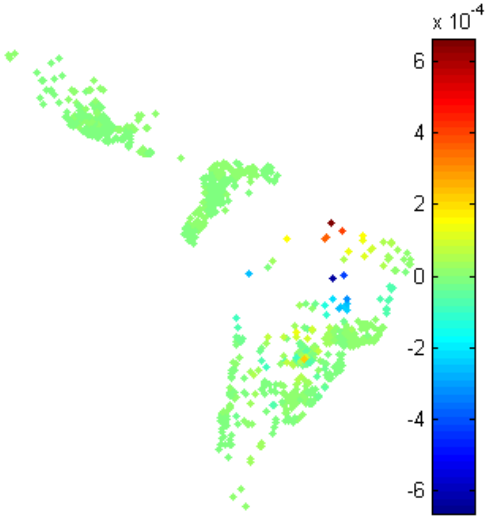
(b) scale at 8



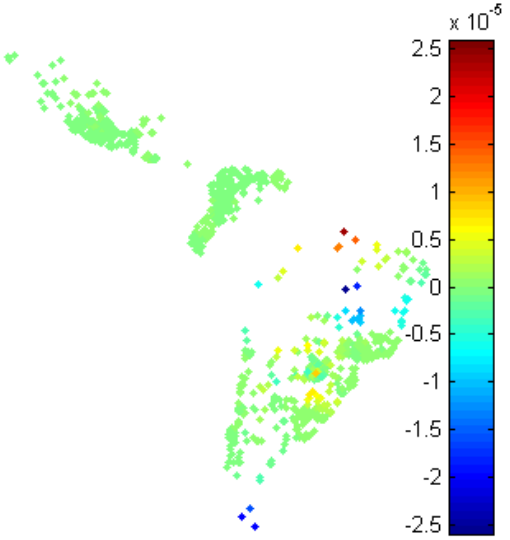
(c) scale at 18



(d) scale at 26



(e) scale at 80



(f) scale at 400

Spectral graph wavelet on South America graph.

Experiment design

Data Source

- Gold standard report (GSR) protests in Latin American countries
- 10% random sampled twitter data, from Jul. 2012 to Dec. 2014

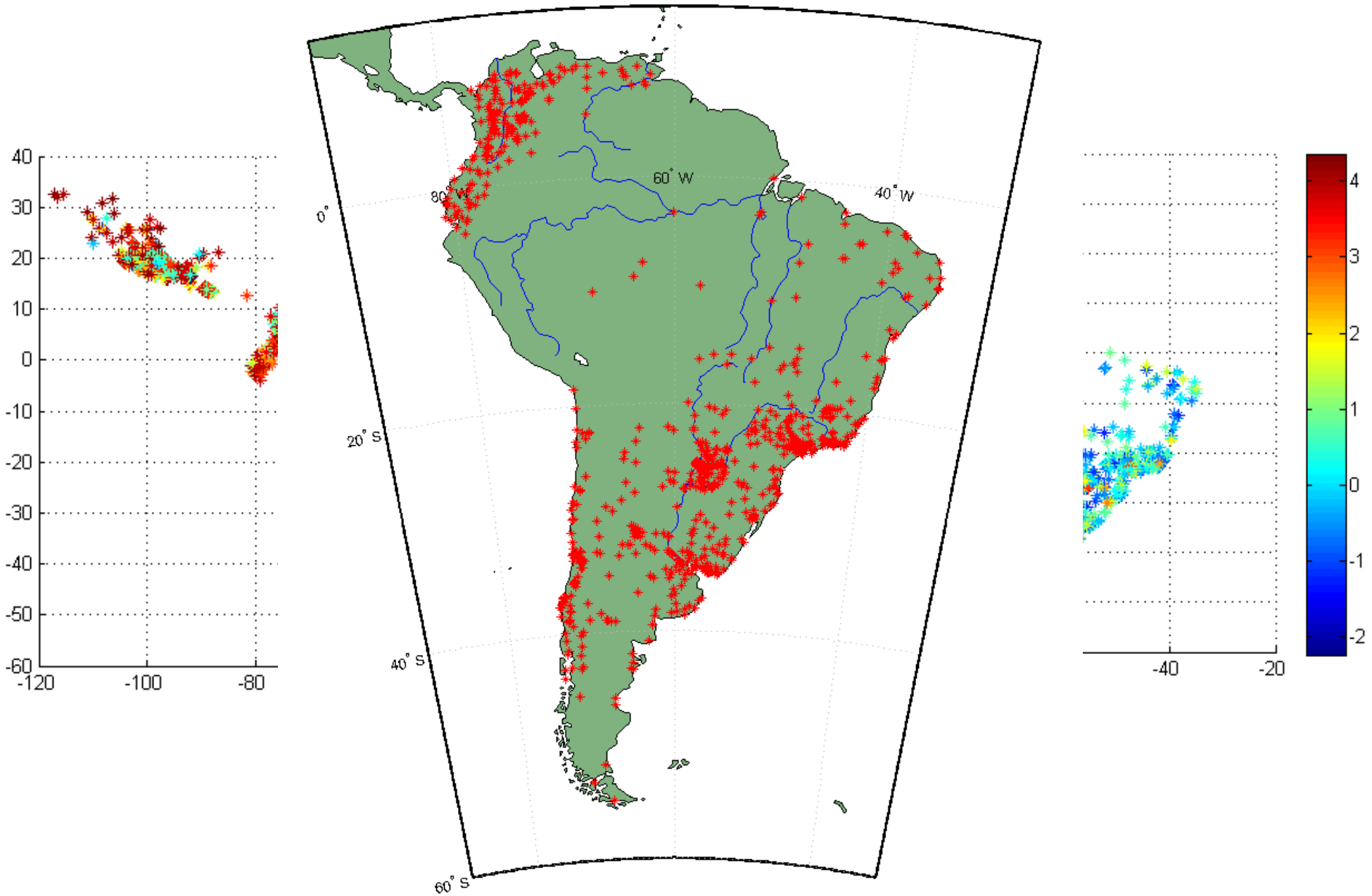
Implementation

- Build graph G for each country, based on KNN
- Compute f(n) based on each city's absenteeism score (Zscore30)
- Calculate anomaly index of f on G
- Set the wavelet coefficient threshold, find the central node and its kernel cities.

Comparison criteria

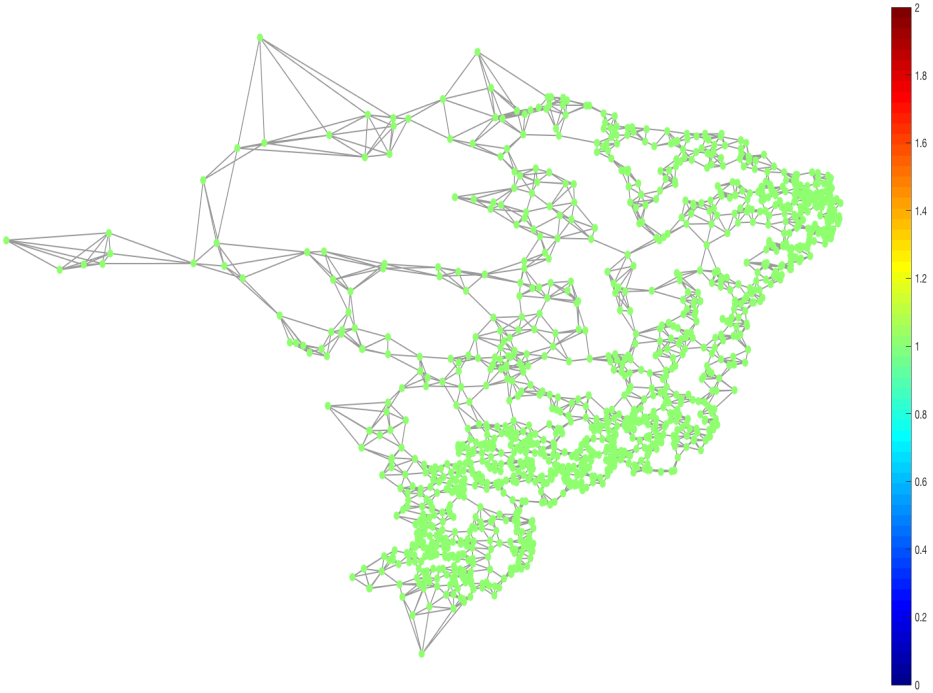
- Event date
- Location (city)
- Group size (group anomaly cities)
- Protest or not

Experiment dataset

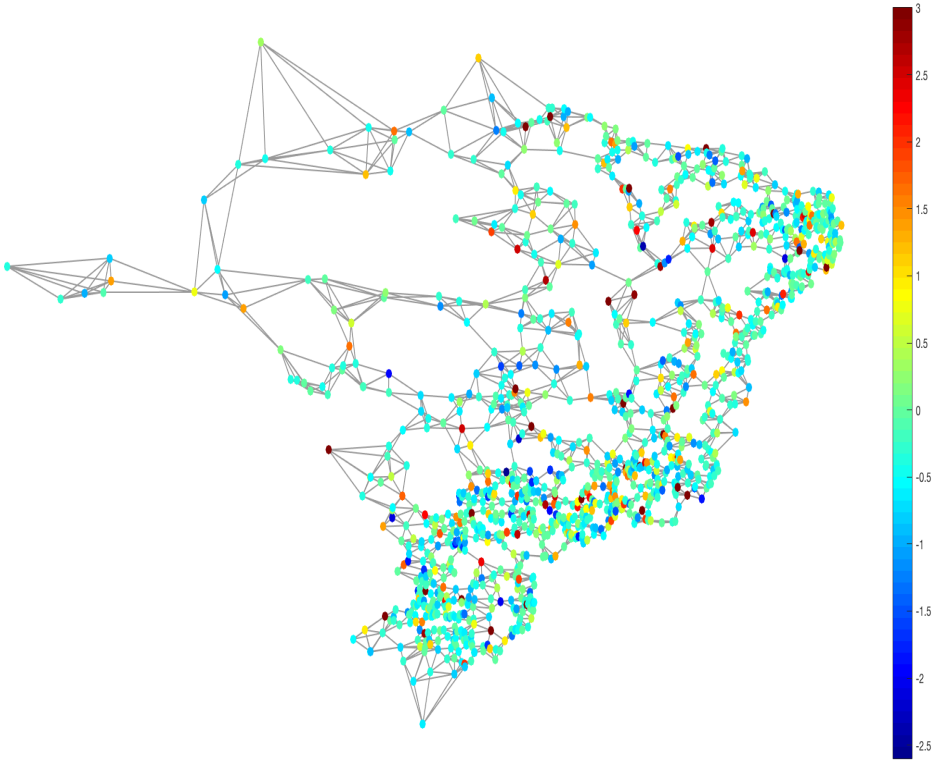


Experiment implementation (1)

1. Build graph G, based on KNN, set K = 5.
2. Compute f(n) based on each city's absenteeism score (Zscore30)



Brazil 5 nearest-neighbor Graph: 1276 cities with all edge weights are 1.

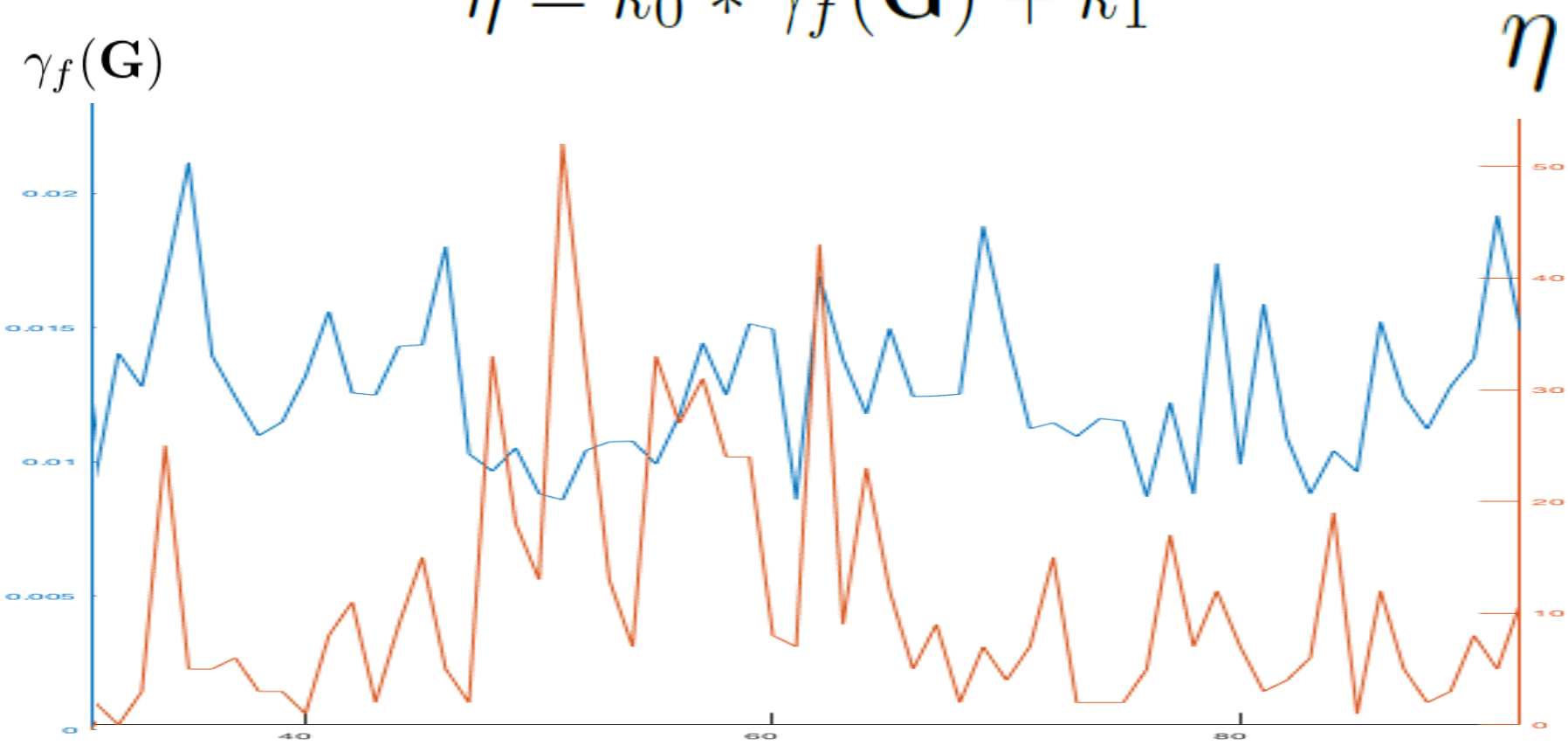


Brazil absenteeism score distribution on June 1st, 2013

Experiment implementation (2)

3. We assume anomaly index $\gamma_f(\mathbf{G})$ has linear relationship with protest events η:

$$\eta = k_0 * \gamma_f(\mathbf{G}) + k_1$$

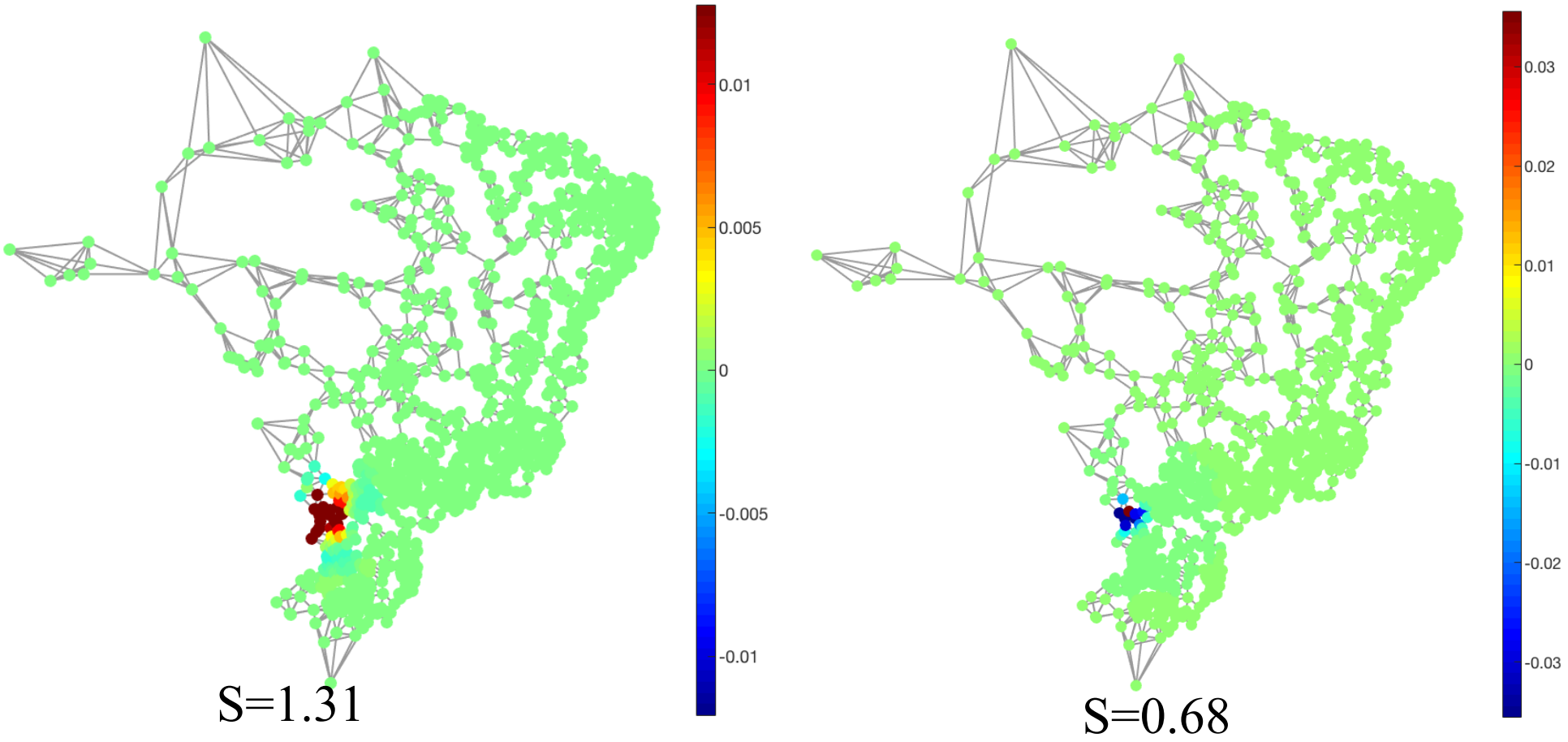


Train historical dataset to get the k_0 and k_1

Experiment implementation (3)

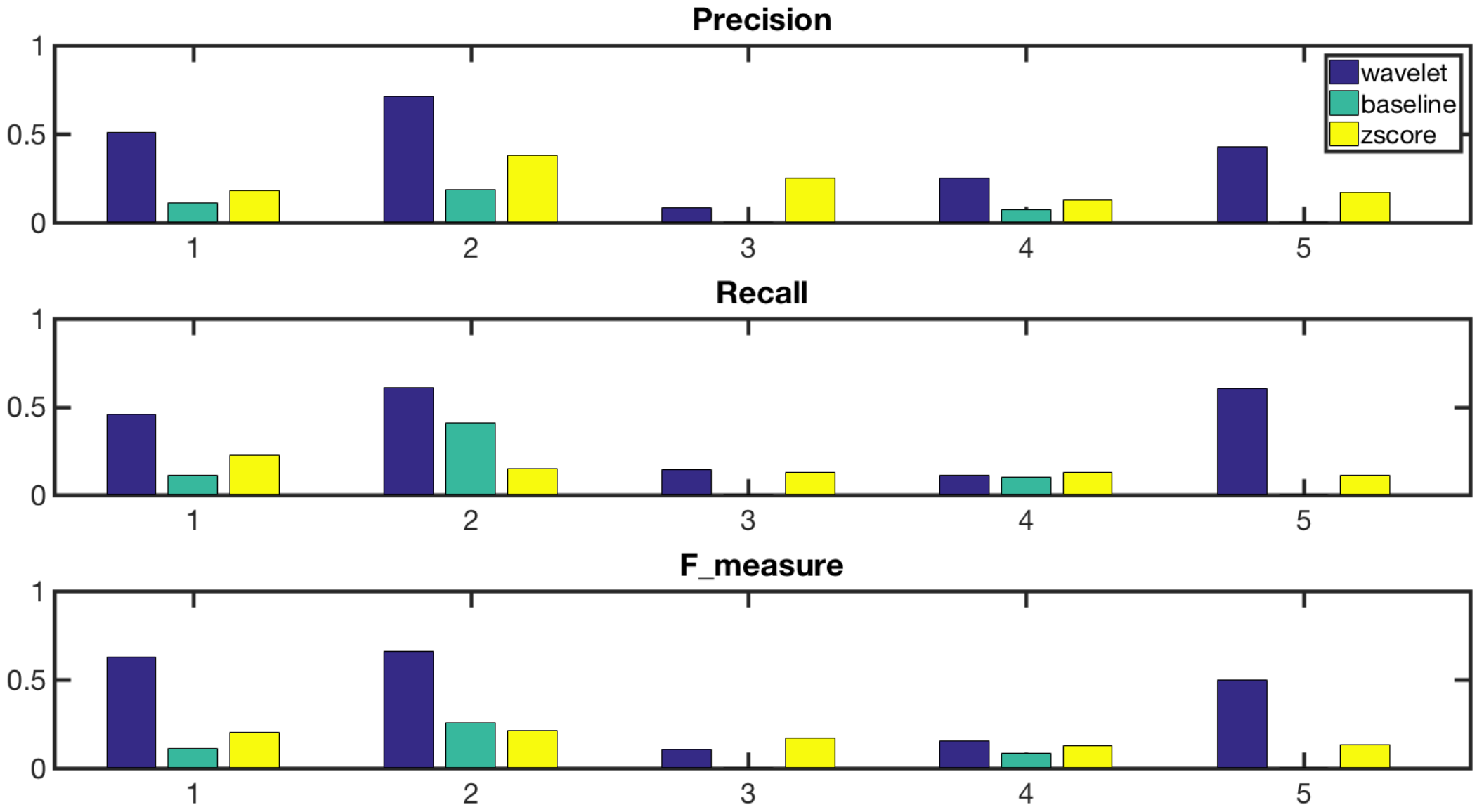
4. Calculate wavelet coefficient $W_f(s, a)$ for each node a with different scale s.

5. Select top η wavelet coefficient with scale s, and center a.



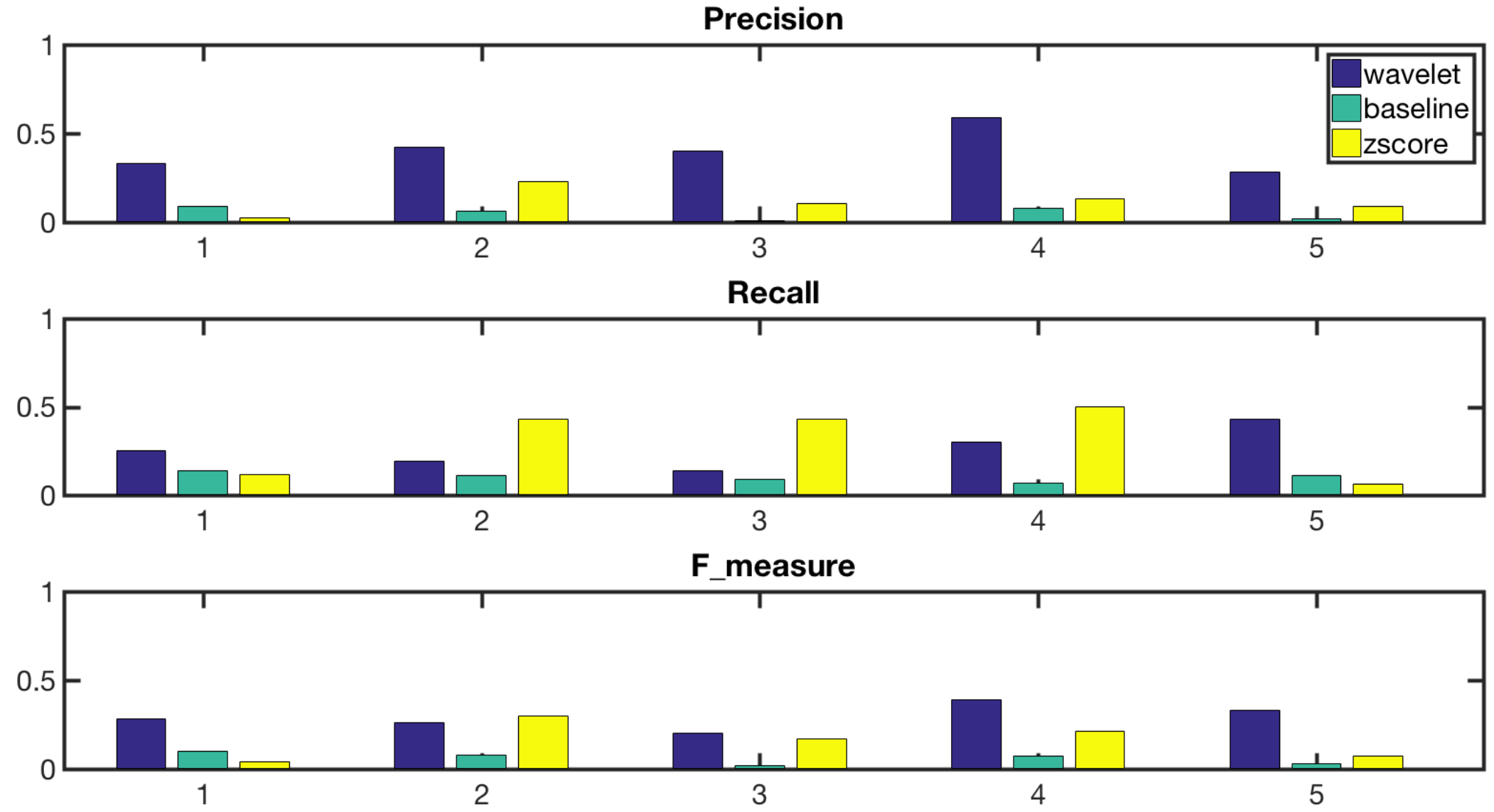
Two graph wavelet with different scale s

Experimental results: Mexico protests



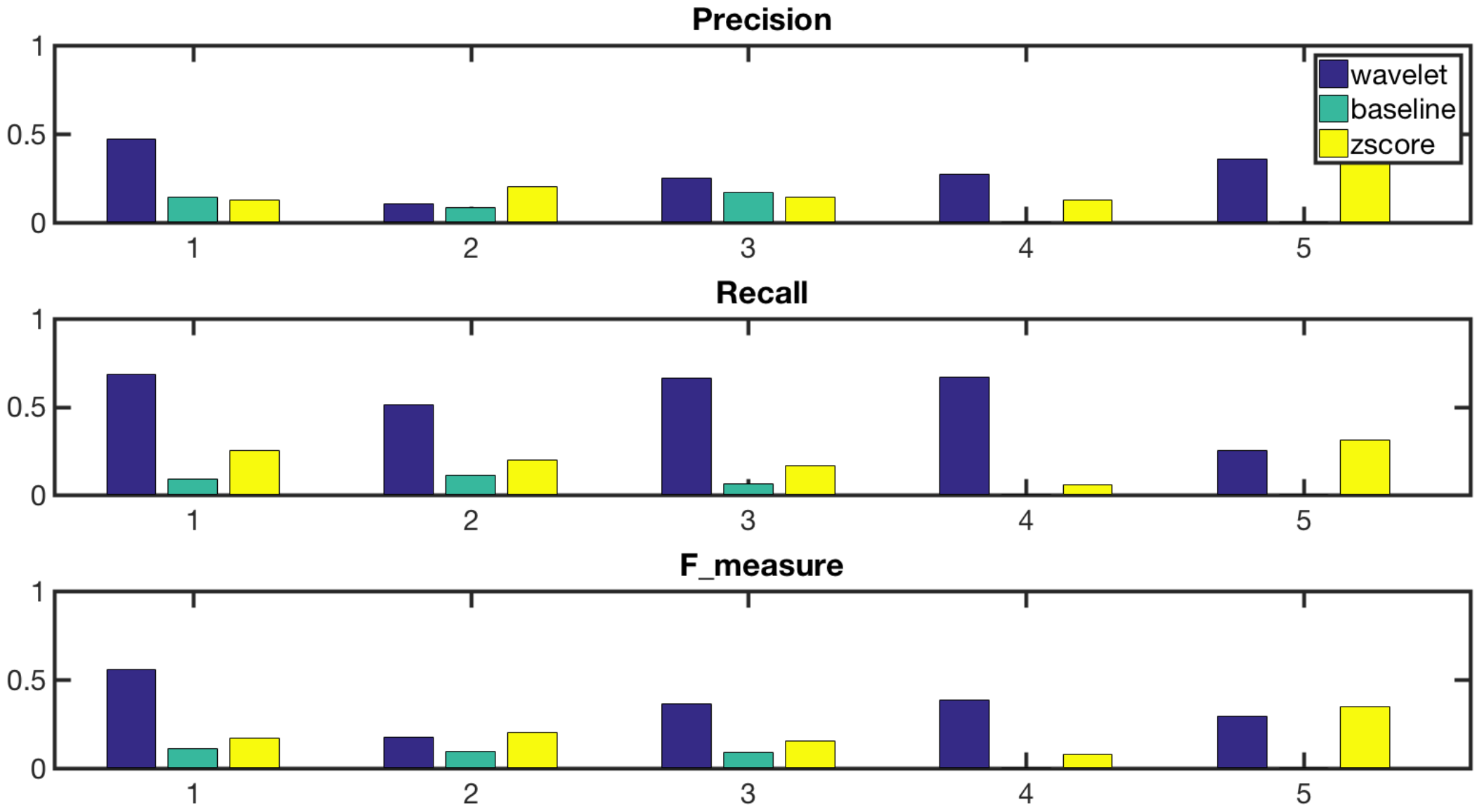
Mexico protest detection performance

Experimental results: Brazil protests



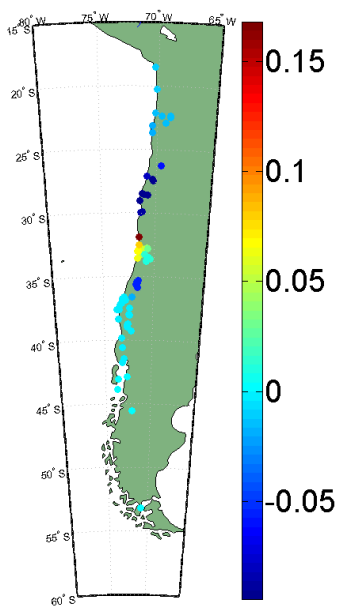
Brazil protest detection

Experimental results: Venezuela protests

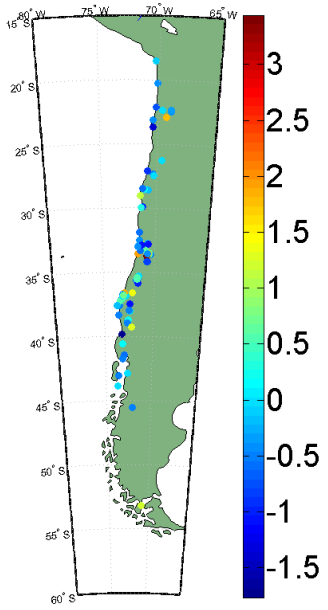


Venezuela protest detection performance

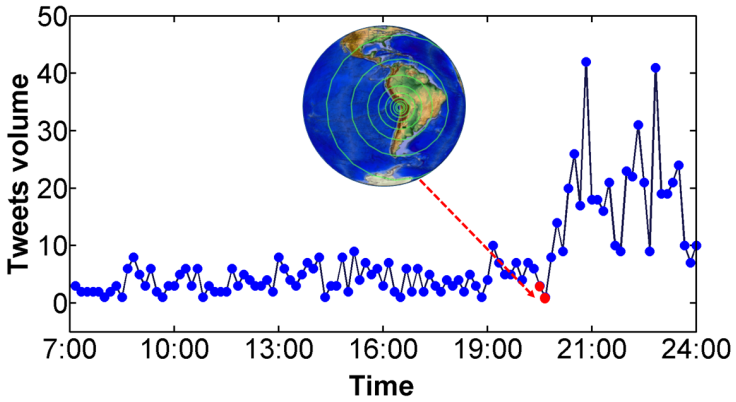
Case study: Chile Earthquake

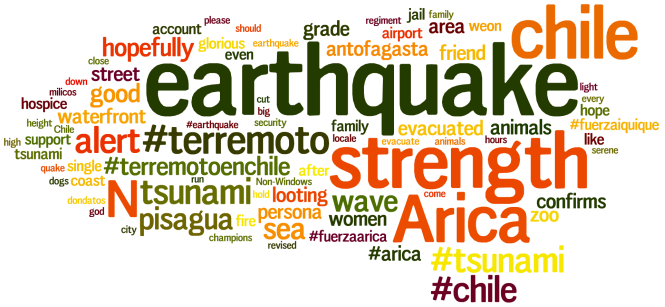


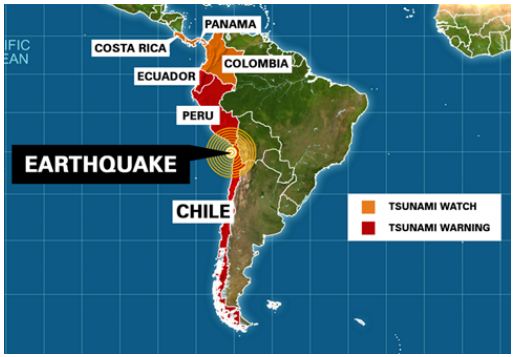
(a) absenteeism score



(b) wavelet coefficient

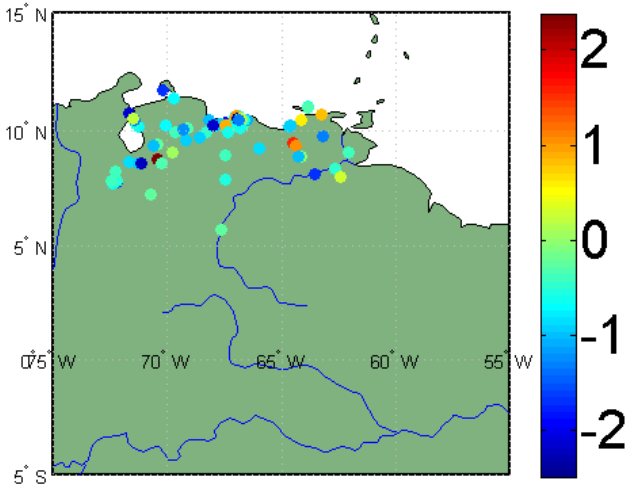




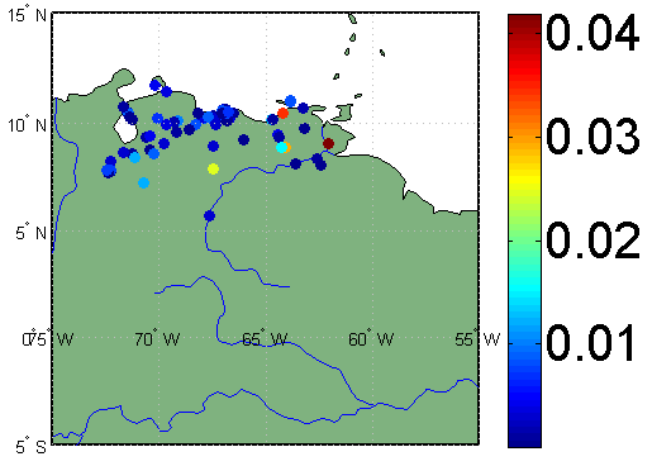


Iquique Earthquake, Chile. April 1, 2014.

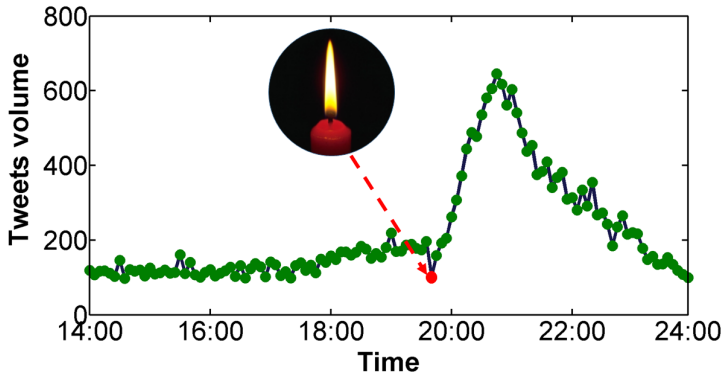
Case study: Venezuela Power Outage

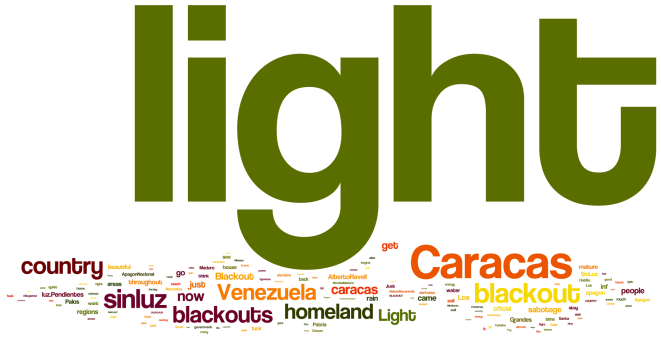


(a) absenteeism score



(b) wavelet coefficient

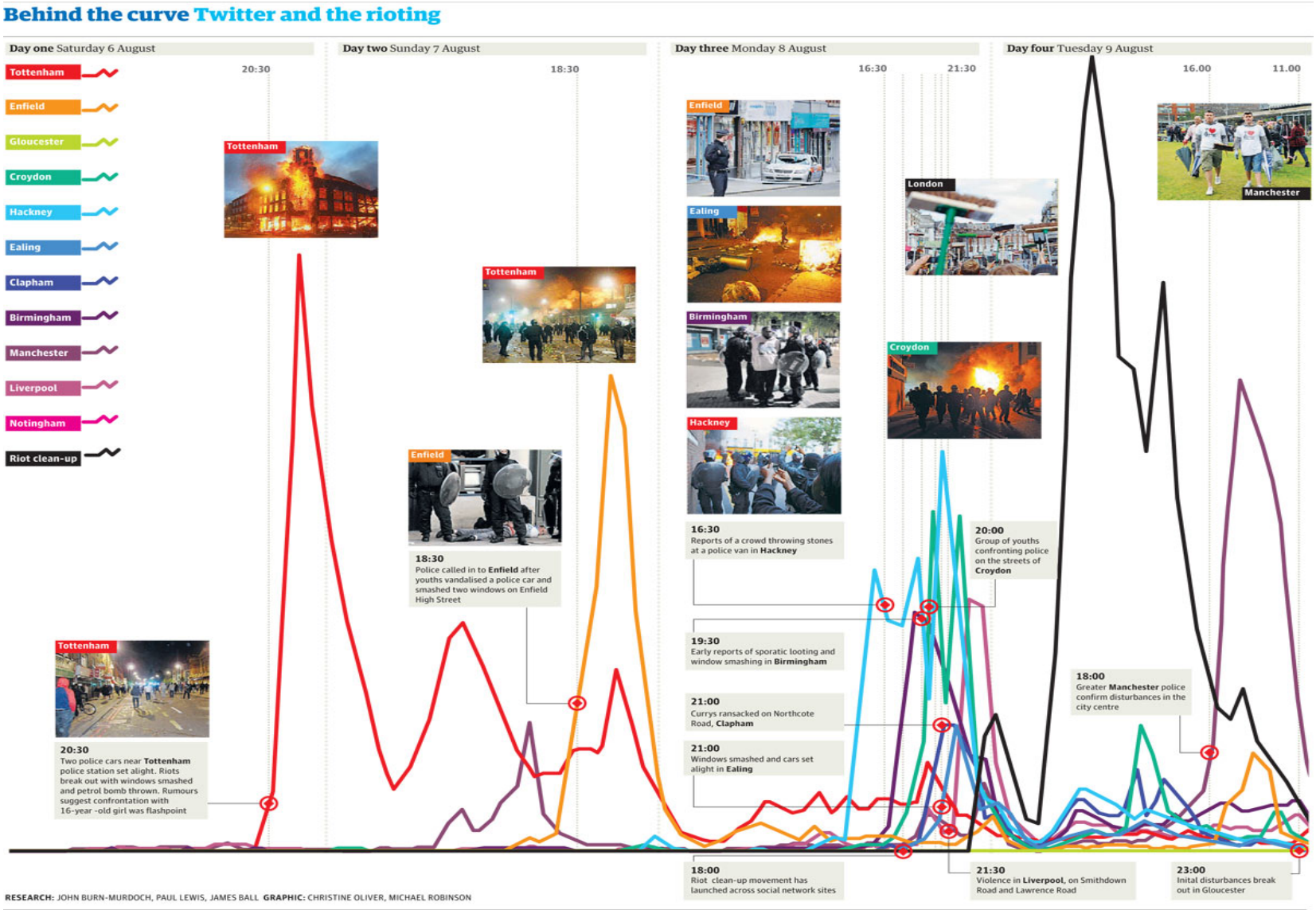




Venezuela power outage. Dec 2, 2013.

Civil Unrest Forecasting

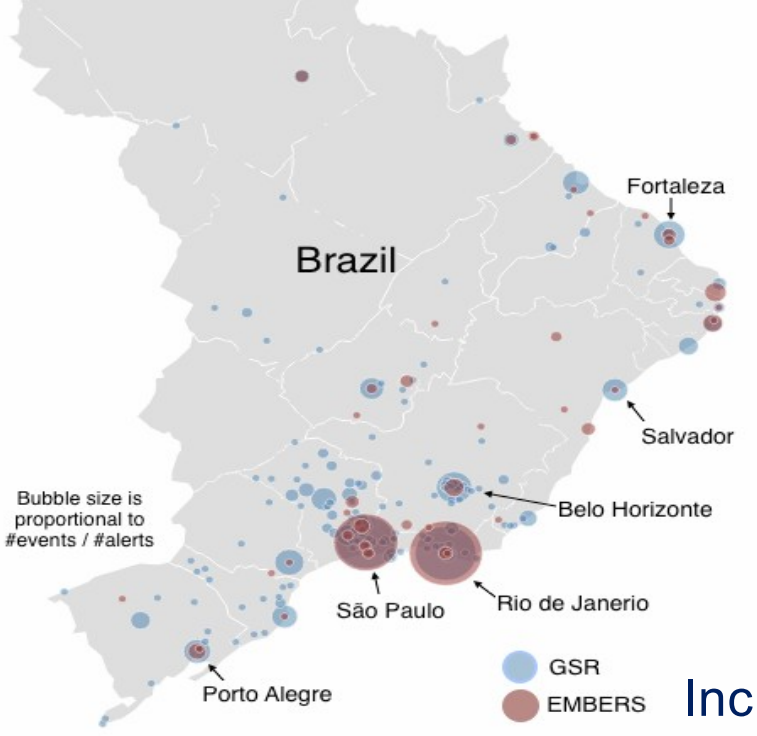
Twitter and the rioting



Protest forecasting

- Focus on 10 Middle and South American countries
- Forecast who, where, when and why

In **June 2013** countrywide protests erupted in Brazil, also known as the **Vinegar Movement**



Reasons:

Increase in bus fares, corruption, health & education costs

Distribution of civil unrest events in Latin America (Nov'12 -- Aug'14) as per Gold Standard Report*

How to forecast protests?

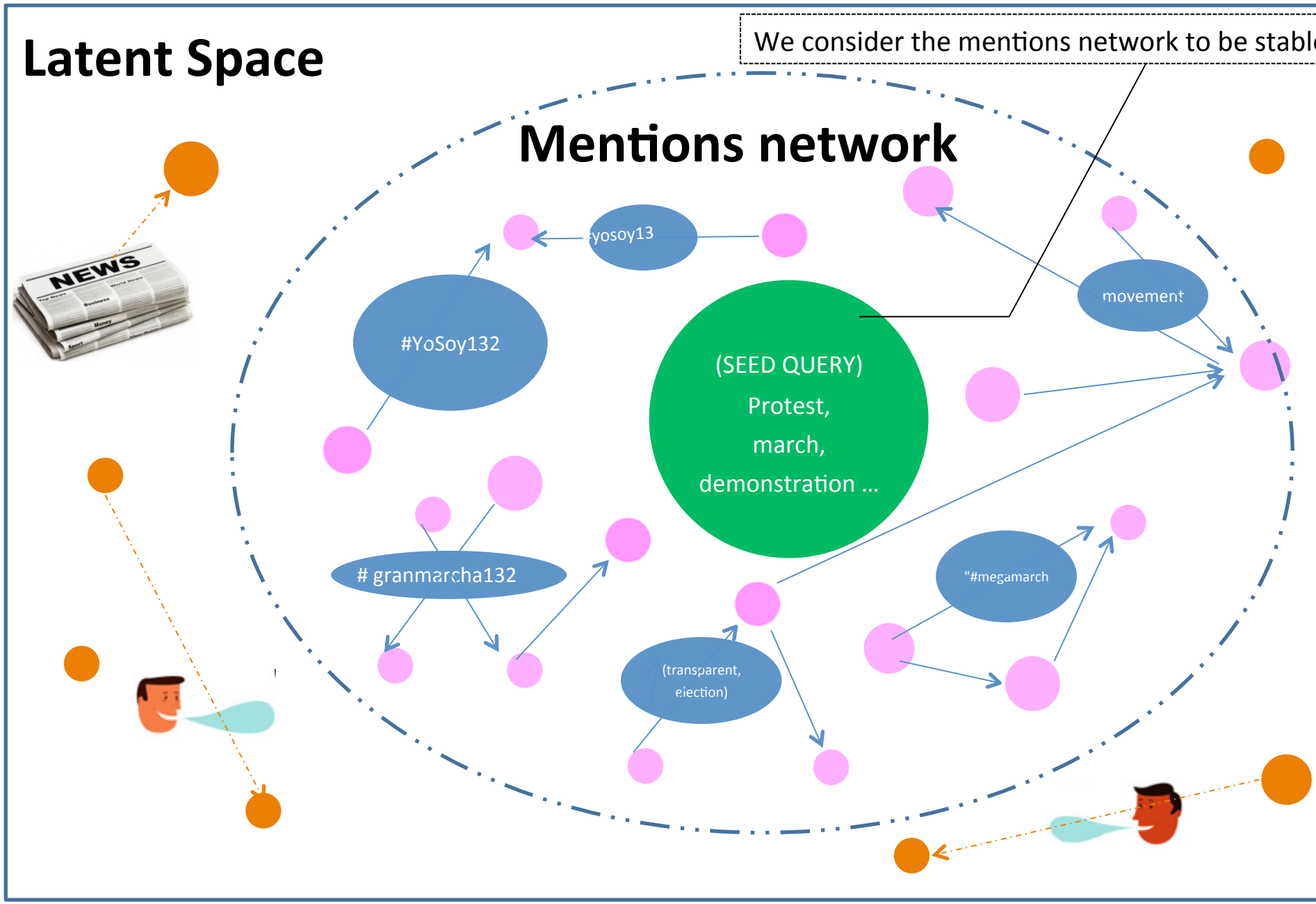
#Yosoy132 Protest – Mexico, 2012

How to forecast protest?

Objective:

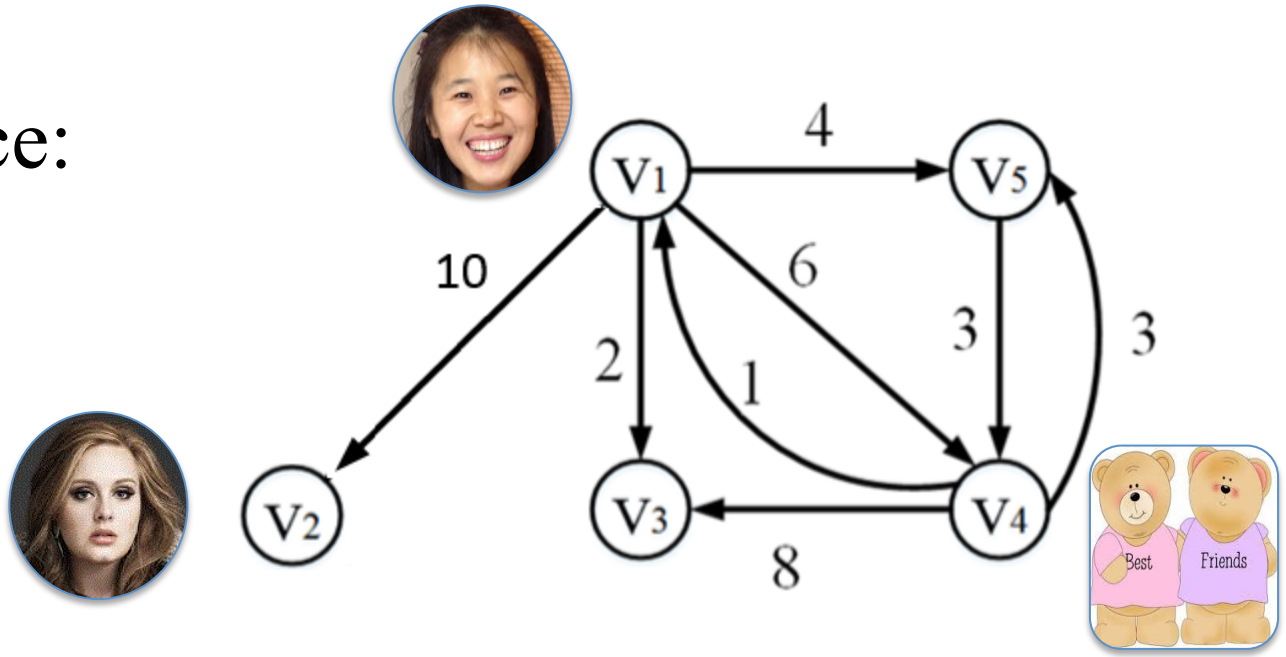
- Model the recruitment of protest participants within social networks
- Capture the underlying social network and structural dynamics
- Forecast the speed and scale of civil unrest events

Approach: Bi-space model



Propagation in the mentions network (1)

Brownian Distance:

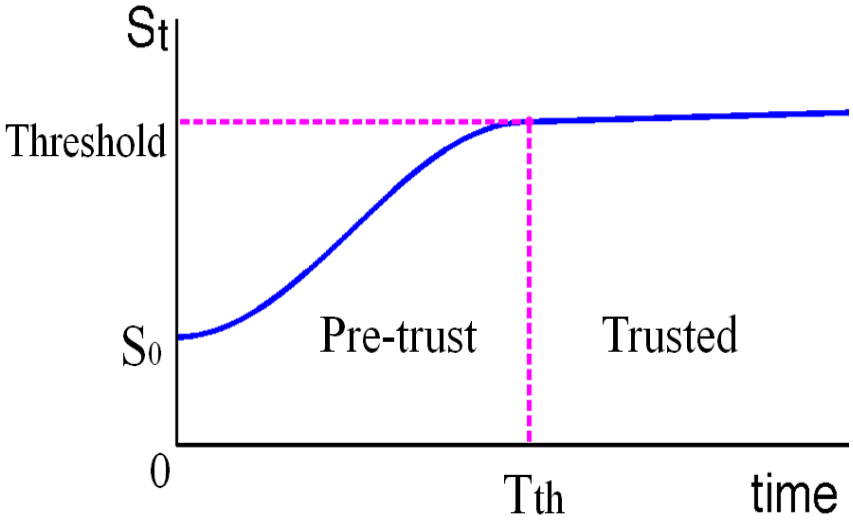


$$d_{ij} = \frac{1}{(1+\omega_{ij})(1+\omega_{ji})^{\gamma}(1+\eta_{ij})}$$

- ω_{ij}: frequency v_i mentions v_j
- $\gamma \geq 1$: higher weight of ω_{ji}
- η_{ij}: common neighbors

Propagation in the mentions network (2)

- **Trust function $S^{ij}(t)$:**
  How much does user v_j trust user v_i at time t ?



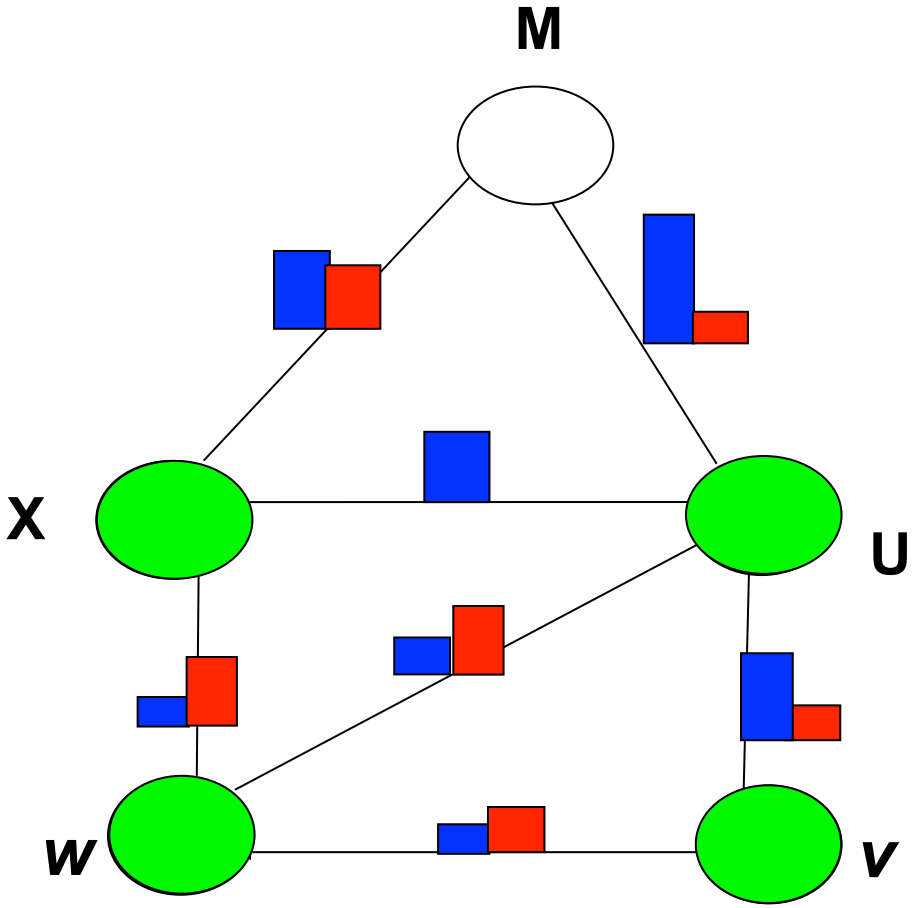
Geometric Brownian motion (GBM)

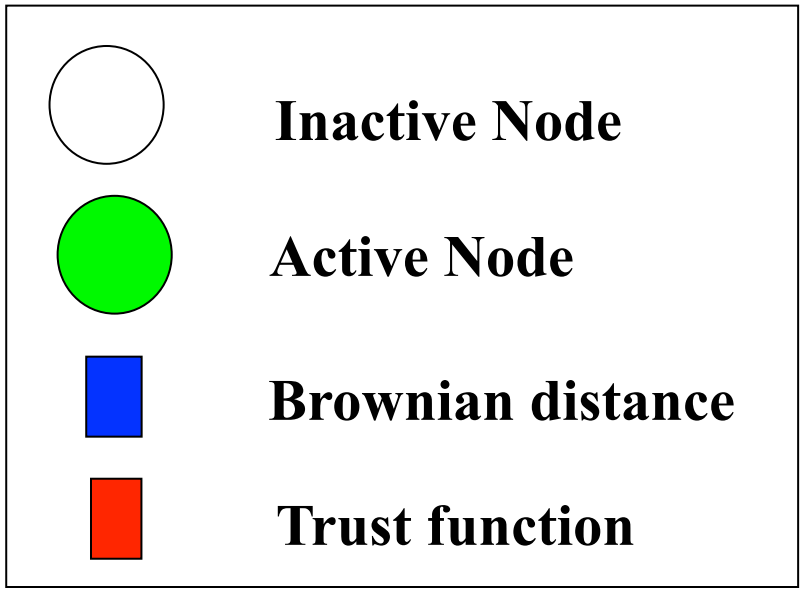
$$S_t = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right)$$

- $S^{ij}(t) \sim GBM(\mu, \sigma^2, t)$

- **Trust condition:** $ln[S^{ij}(t)] \geq d_{ij}$

Propagation in the mentions network (3)



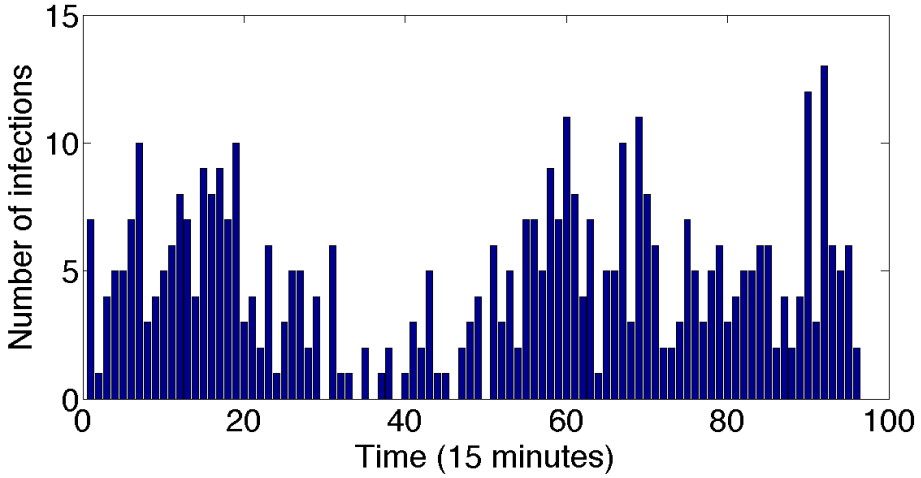


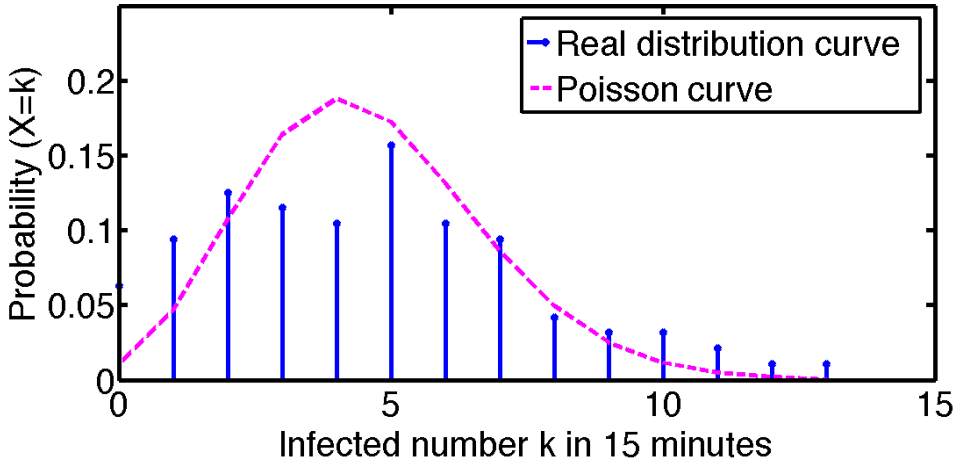
Stop!

Trust condition: $ln[S^{ij}(t)] \geq d_{ij}$

Latent space: Poisson distribution

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$





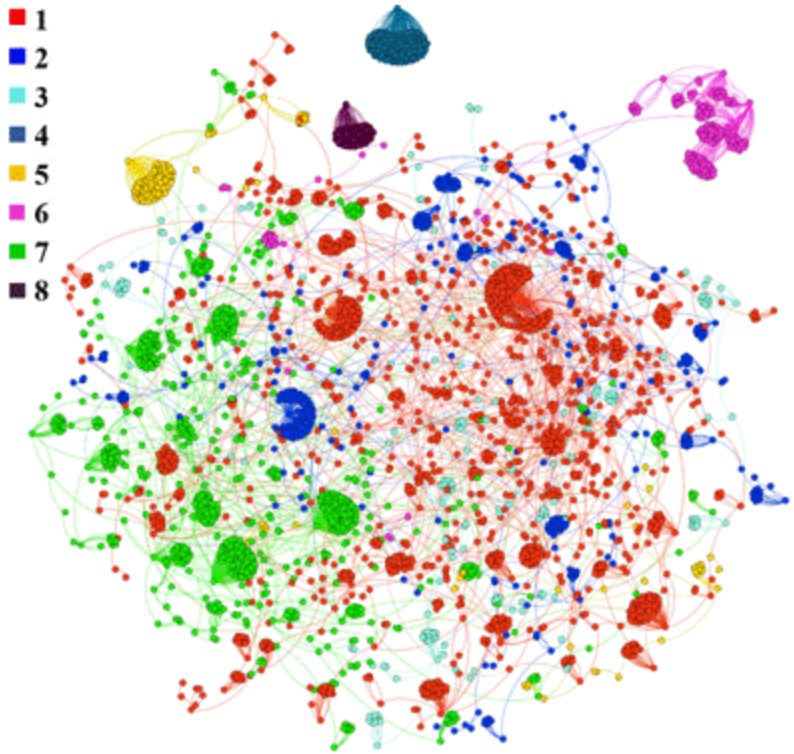
Infected nodes in latent space

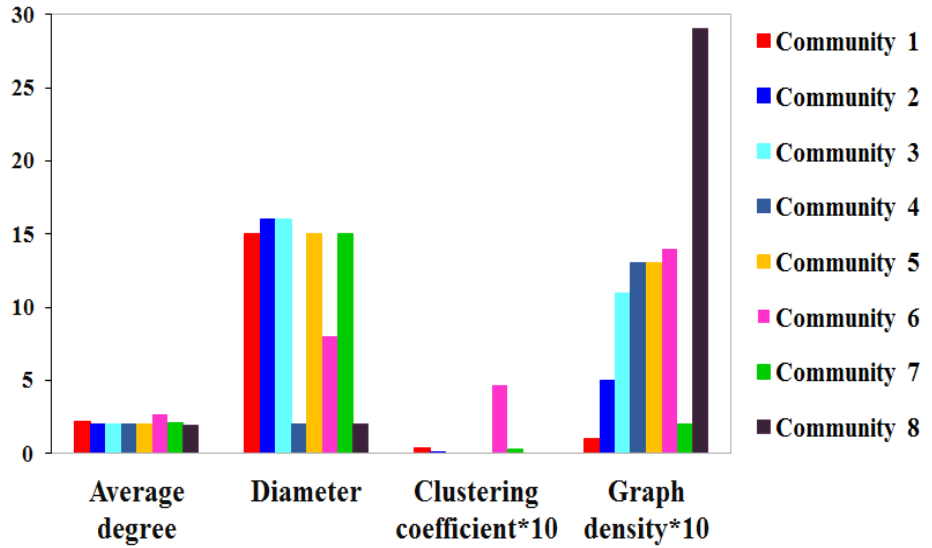
Poisson distribution fit (λ = 4.18)

Community level propagation

Assumptions:

- Each community has its own parameters
- Propagation among communities using source community's parameters





Protest forecasting

Twitter – data source

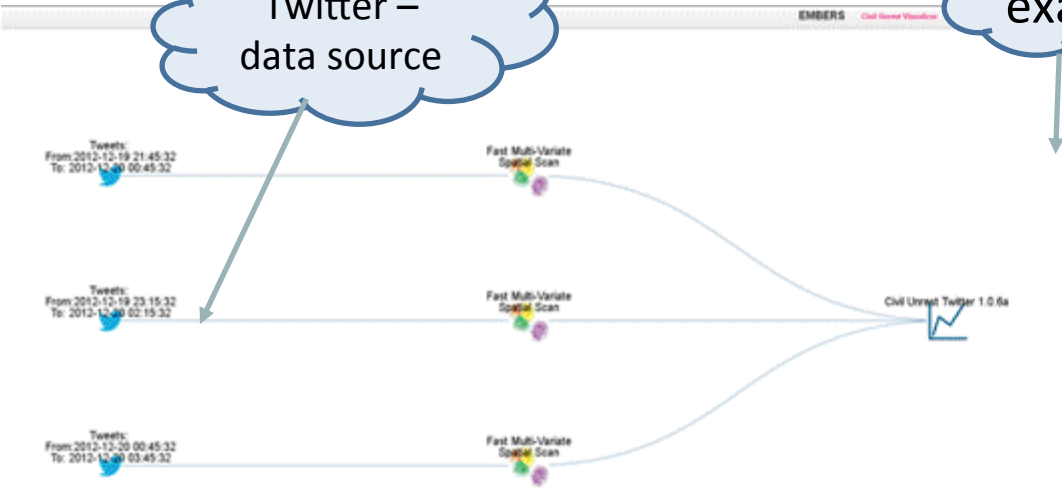
Protest example

Top Keywords for all three clusters

Geographical Map

Word Cloud

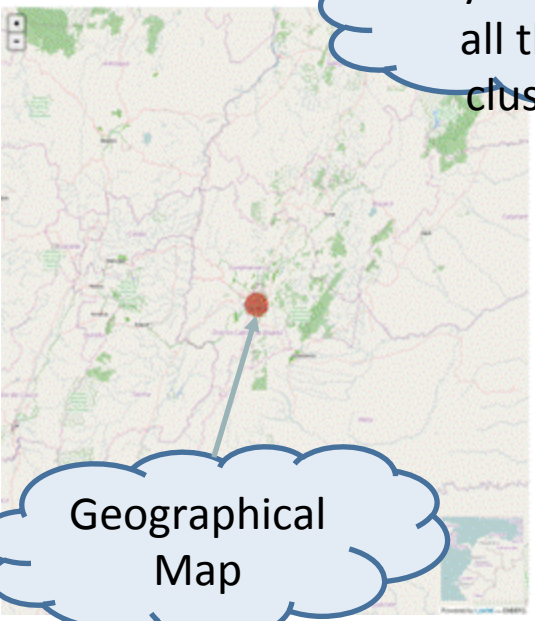
Relevant Tweets



Prediction Date	Event Date	Location	Population	Event Type
2012-12-20	2012-12-21	Argentina, Río Negro, Bariloche	General Population	0141
2012-12-20	2012-12-21	Ecuador, Pichincha, Quito	General Population	0142
2012-12-20	2012-12-21	Colombia, Bogotá, Bogotá	Agricultural	0142
2012-12-19	2012-12-20	Brazil, Distrito Federal, Brasília	General Population	0142
2012-12-18	2012-12-19	Ecuador, Carchi, Tulcán	Education	0122
2012-12-18	2012-12-19	Ecuador, Tungurahua, Ambato	Education	0122
2012-12-18	2012-12-19	Ecuador, Napo, Tena	Education	0122
2012-12-18	2012-12-19	Ecuador, Pichincha, Quito	Education	0122

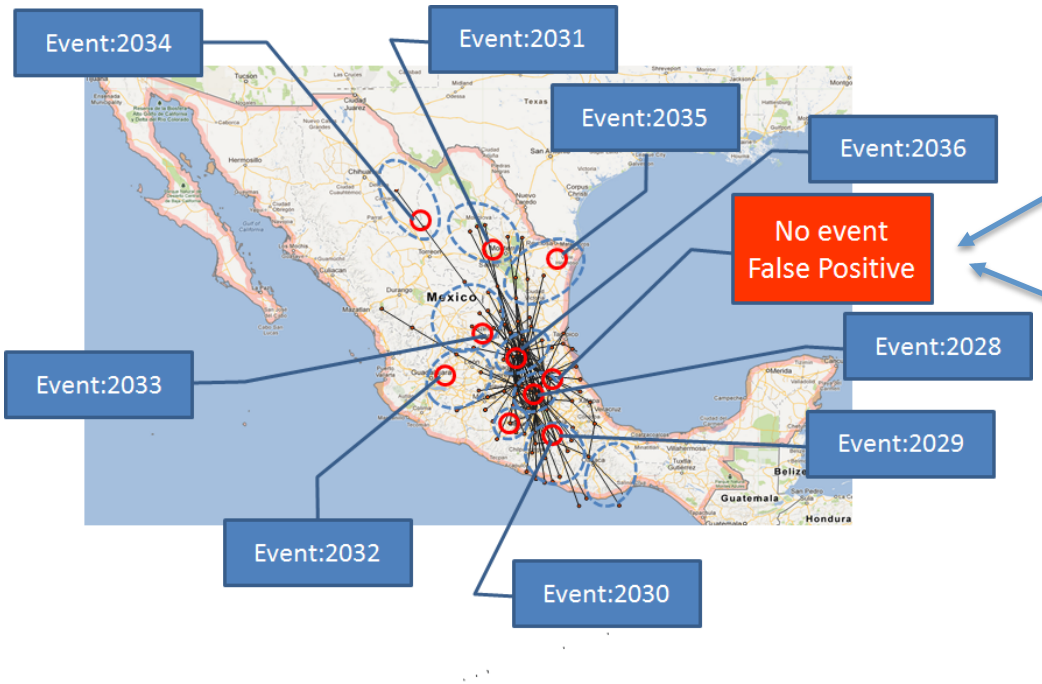
754 Clusters

Creation Date	Population	Country	Event Type
2012-12-20 20:17:58	General Population	Ecuador	0142
2012-12-20 20:17:50	General Population	Venezuela	0142
2012-12-20 20:17:50	General Population	Venezuela	0142
2012-12-20 20:17:41	General Population	Argentina	0142
2012-12-20 20:17:41	General Population	Argentina	0142
2012-12-20 20:17:18	General Population	Paraguay	0142
2012-12-20 20:17:14	Agricultural	Colombia	0142
2012-12-20 20:17:14	Agricultural	Colombia	0122



Case study: misinformation campaigns

Protest detection



Sept 5, 2012@ Mexico

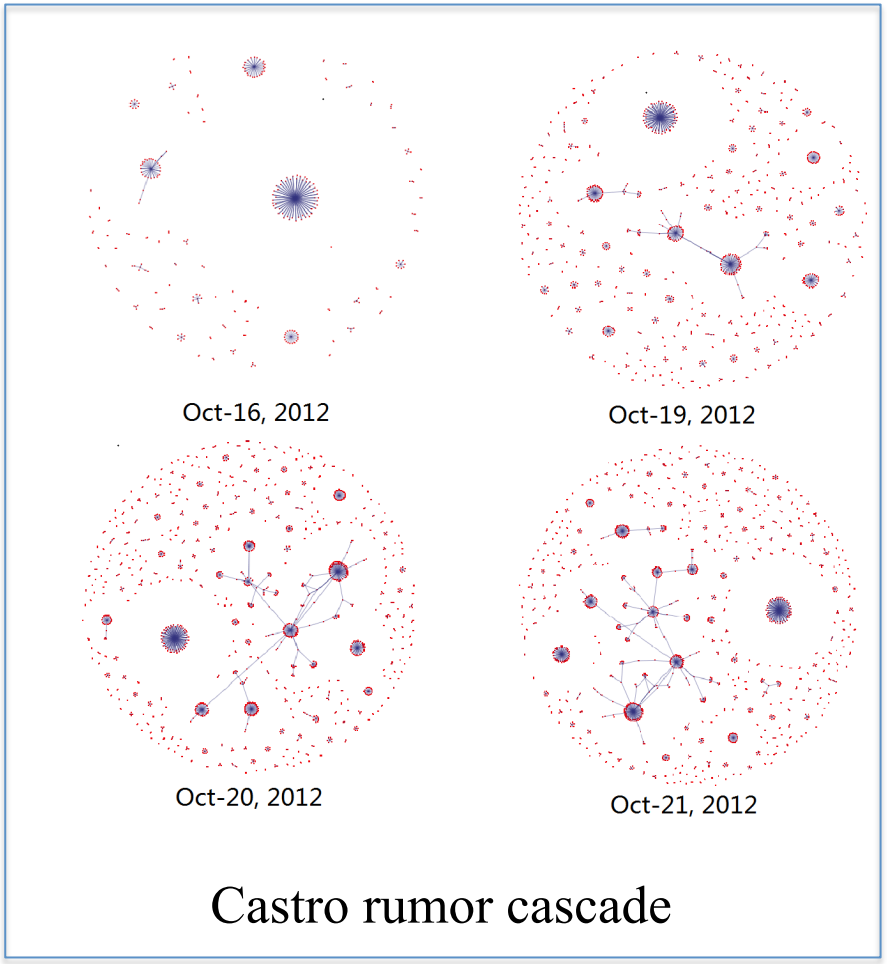
False rumors

How can we distinguish real movements from rumors?

Distinguish rumors from real news

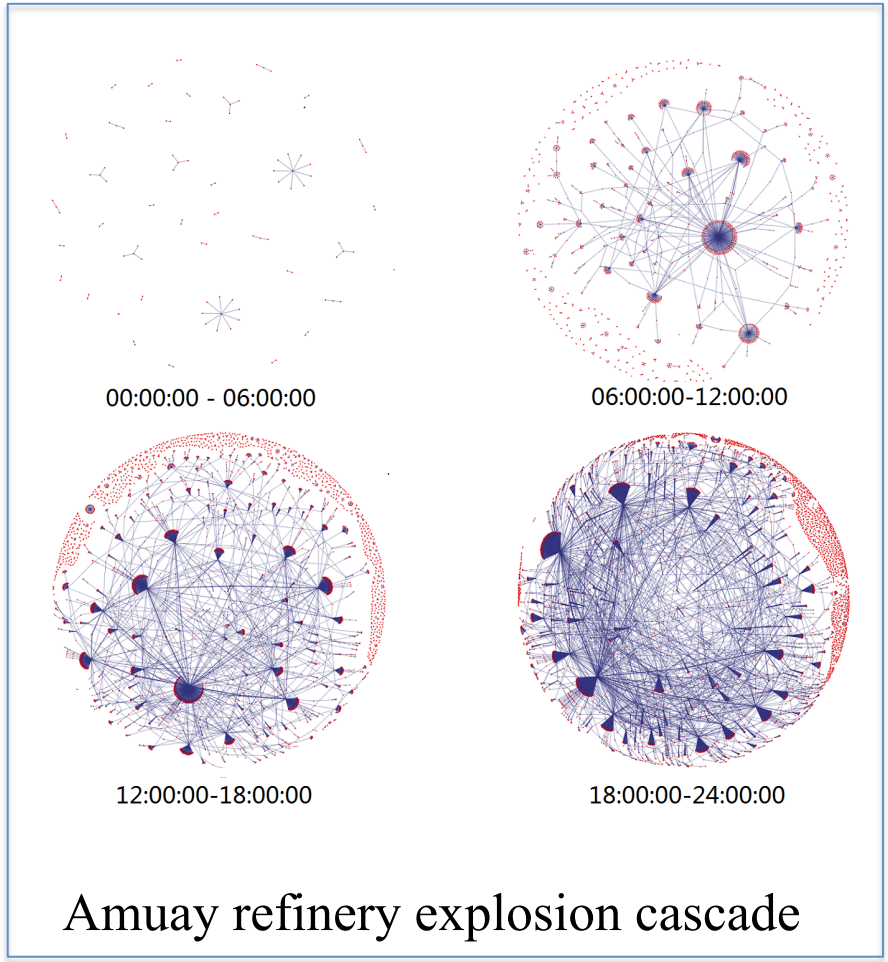
Difference between rumor and news propagation

Rumor



Castro rumor cascade

Real News



Amuay refinery explosion cascade

Retweet cascade

Model intuition (comparing disease vs rumor propagation)

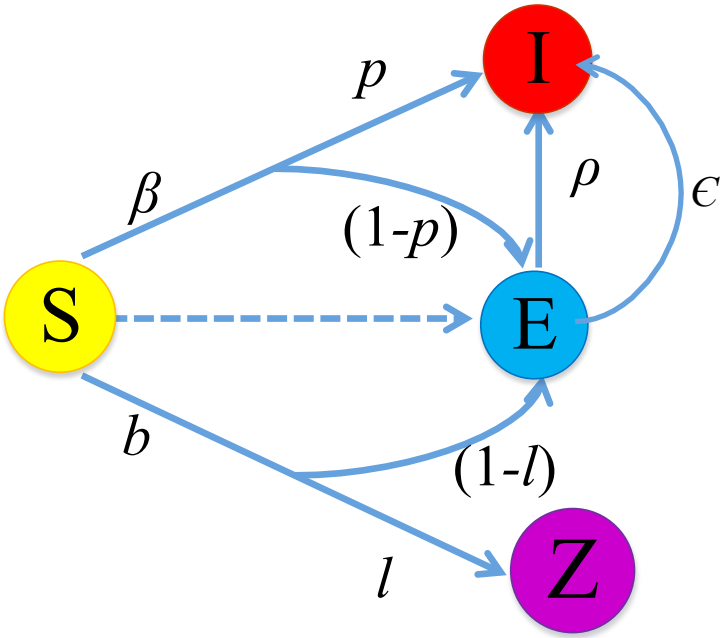
Similarities:

- susceptible, using status S
- infected, using status I
- may take time to accept, exposed status E
- with transmission route

Differences:

- Idea: can be skeptics, introduce skeptics Z
- Idea: no immune system, no recover "R"

SEIZ Model



$$\frac{d[S]}{dt} = -\beta S\frac{I}{N} - bS\frac{Z}{N}$$

$$\frac{d[E]}{dt} = (1-p)\beta S\frac{I}{N} + (1-l)bS\frac{Z}{N} - \rho E\frac{I}{N} - \epsilon E$$

$$\frac{d[I]}{dt} = p\beta S\frac{I}{N} + \rho E\frac{I}{N} + \epsilon E$$

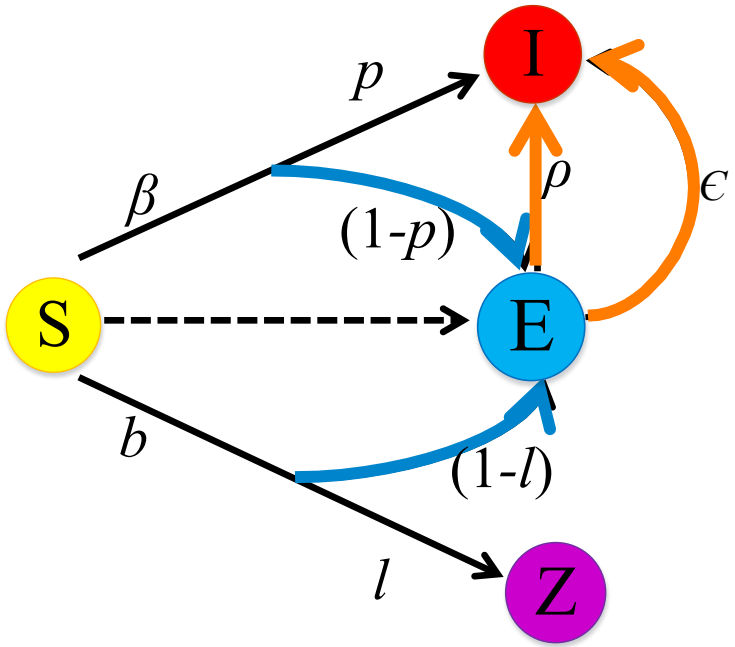
$$\frac{d[Z]}{dt} = lbS\frac{Z}{N}$$

Susceptible	S	Twitter accounts
Infected	I	Believe news / rumor, (I) post a tweet
Exposed	E	Be exposed but not yet believe
Skeptics	Z	Skeptics, do not tweet

Disease **Ideas**

Capturing people's acceptance of ideas

Response ratio: Compare the speed of **adding to the Exposed** compartment with removing from the Exposed compartment.



$$R_{SI} = \frac{\text{Inflow to Exposed}}{\text{Outflow from Exposed}}$$

$$R_{SI} = \frac{(1-p)\beta + (1-l)b}{\rho + \epsilon}$$

R_{SI}, a kind of flux ratio, the ratio of effects entering E to those leaving E.

Dataset: Ebola related rumors

Table 1. Top 10 Ebola-related rumors by Tweet volume from 28 September to 18 October 2014.

Rumor no.	Content	Label
1	Ebola vaccine only works on white people	White
2	Ebola patients have risen from the dead	Zombie
3	Ebola could be airborne in some cases	Airborne
4	Health officials might inject Ebola patients with lethal substances	Inject
5	There will be no 2016 election and complete anarchy	Vote
6	The US government owns a patent on the Ebola virus	Patent
7	Terrorists will purposely contract Ebola and spread it around	Terrorist
8	The new iPhone 6 is infecting people with Ebola	iPhone
9	There is a suspected Ebola case in Kansas City	Kansas
10	Ebola has been detected in hair extensions	Hair

Can you believe?

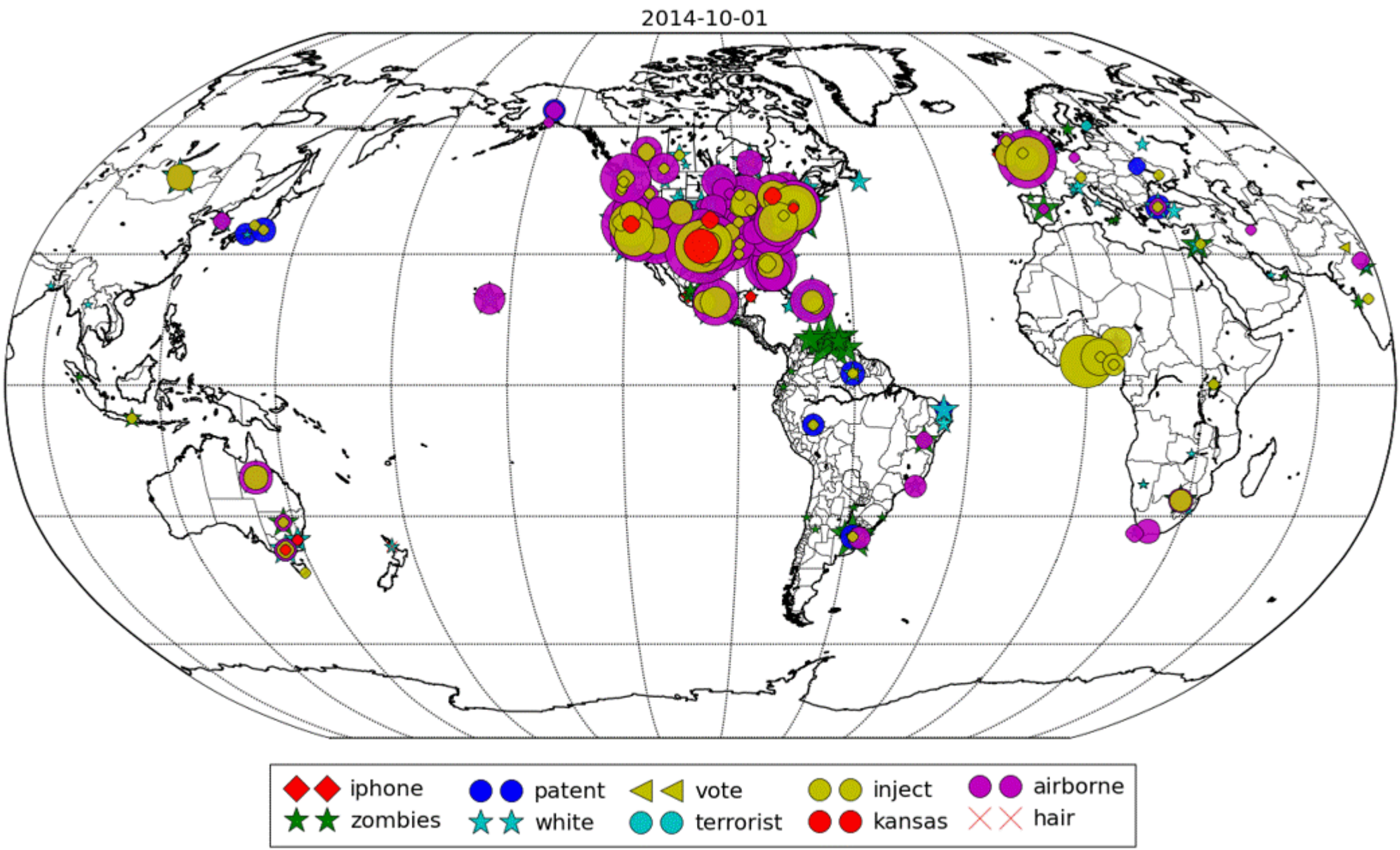
Can you believe?

Can you believe?

Table 2: Ebola related news stories

1	The first Ebola patient (Duncan) identified in US (Dallas).	Dallas
2	The specific symptoms and travel activities of Spencer in the days before he was diagnosed.	Spencer
3	The first confirmation of an Ebola patient in New York City	NYC

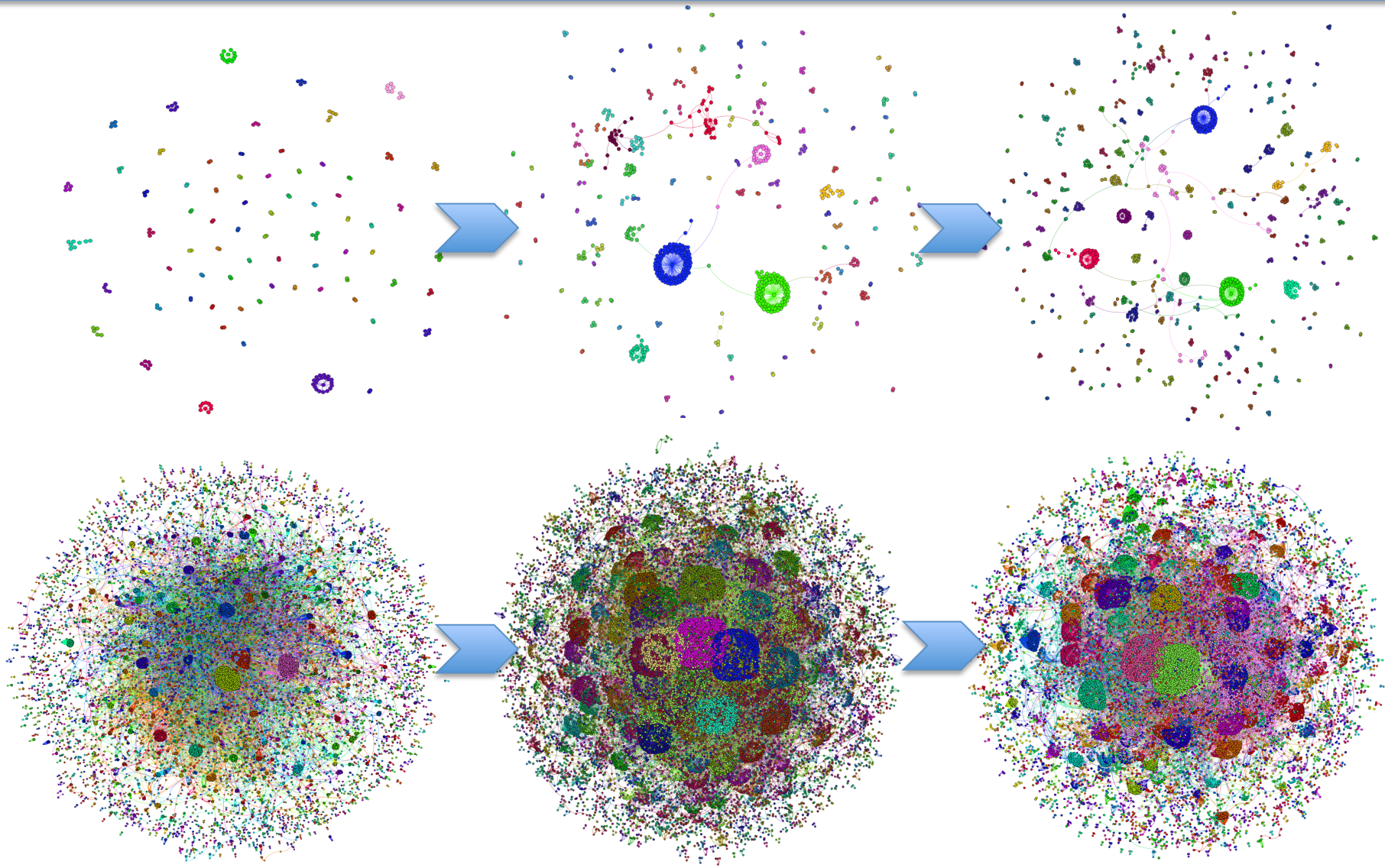
Ebola related rumor distribution



Difference between rumor and news propagation

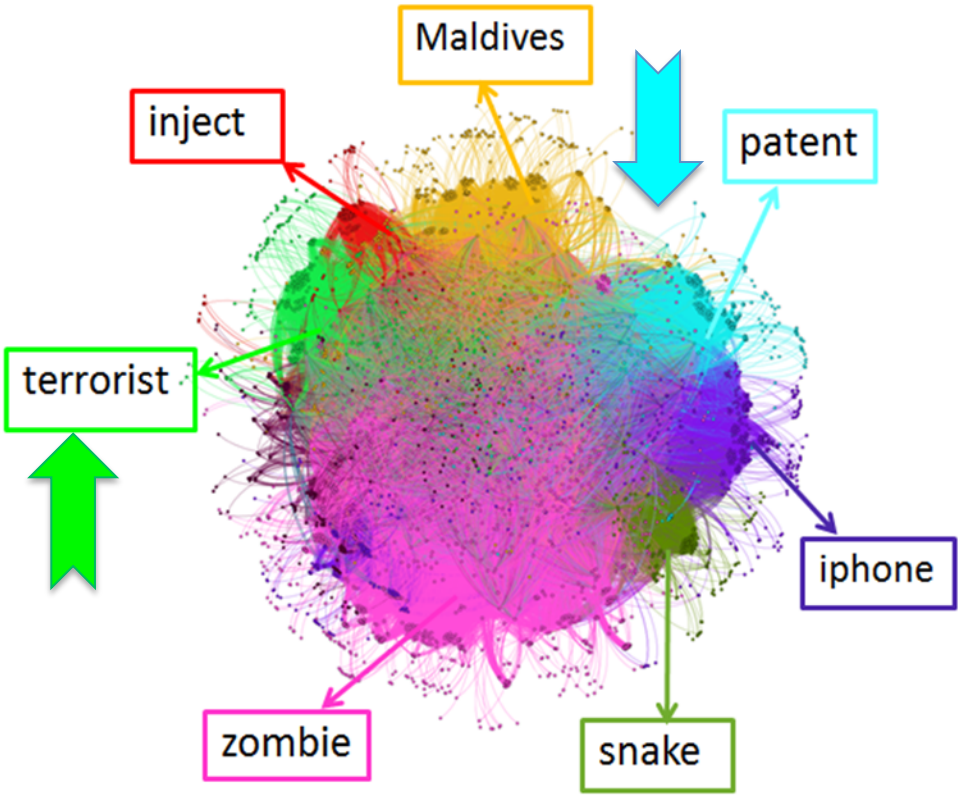
Patent rumor

First US patient news

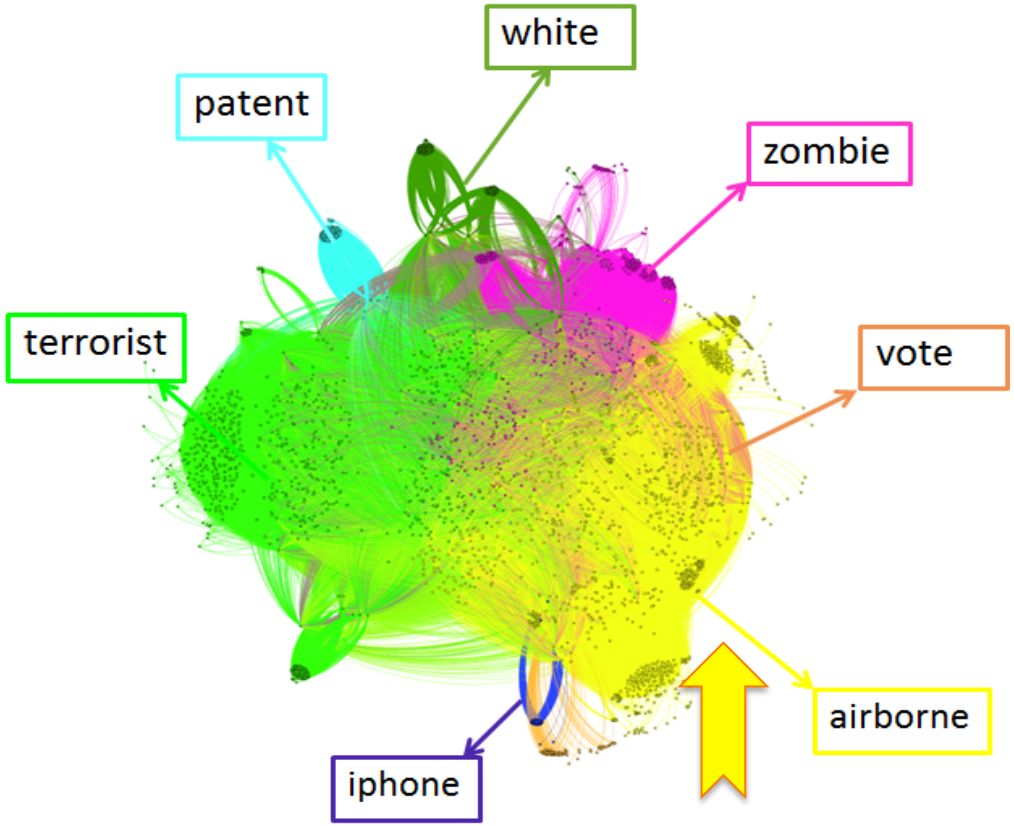


09/30/2014 10/01/2014 10/02/2014

Ebola rumors cluster



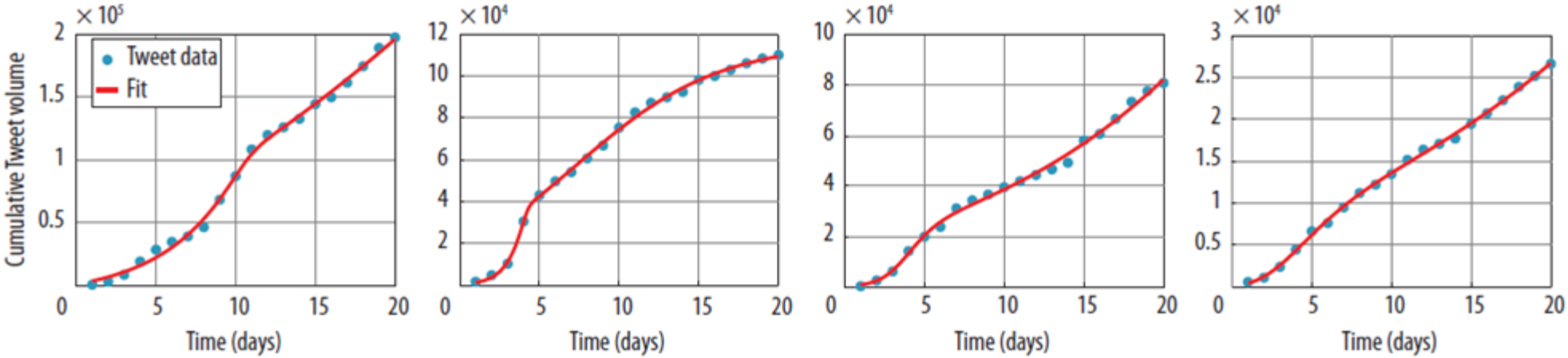
09/29/2014



10/06/2014

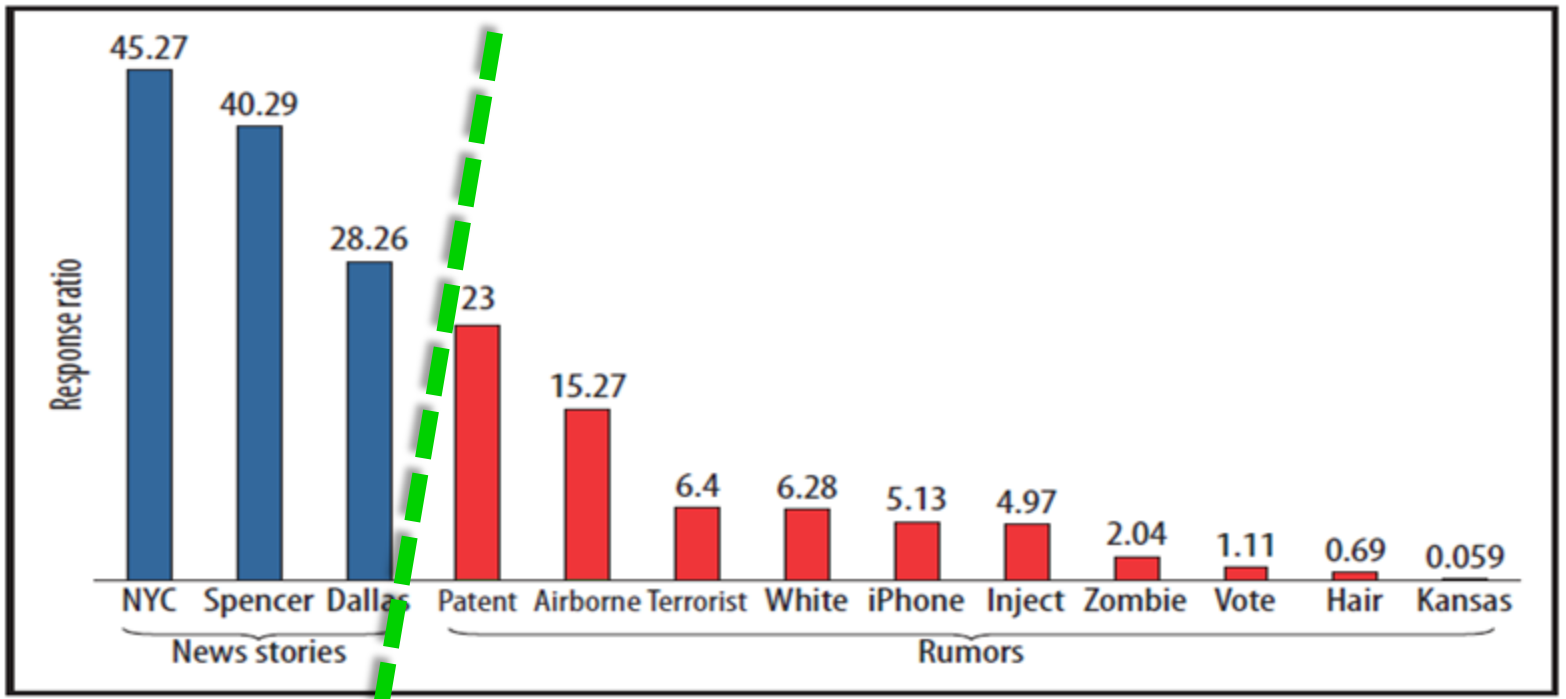
Rumors are color coded consistently across the two frames.

SEIZ results of Ebola rumors

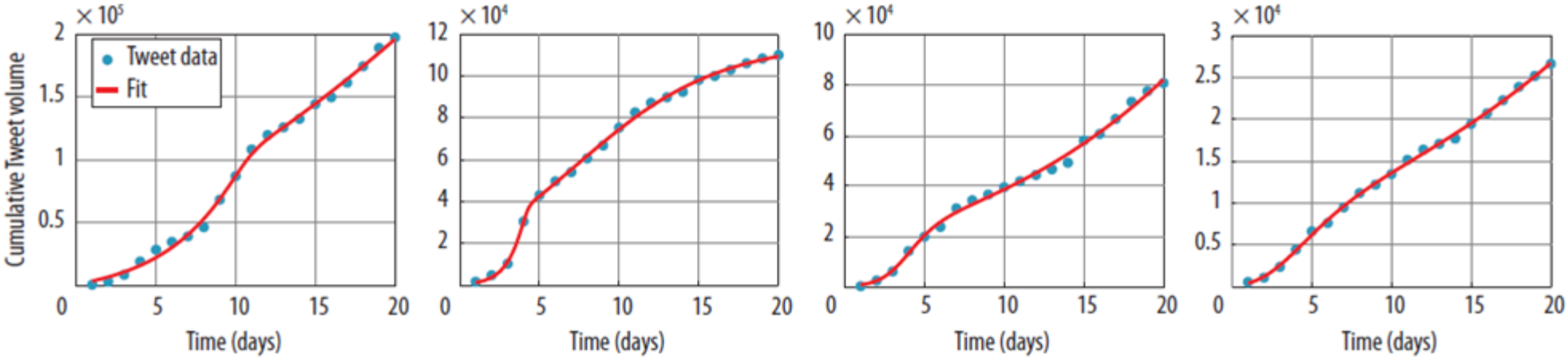


White Zombie Airborne Patent

Response ratio of 3 real news and 10 rumors

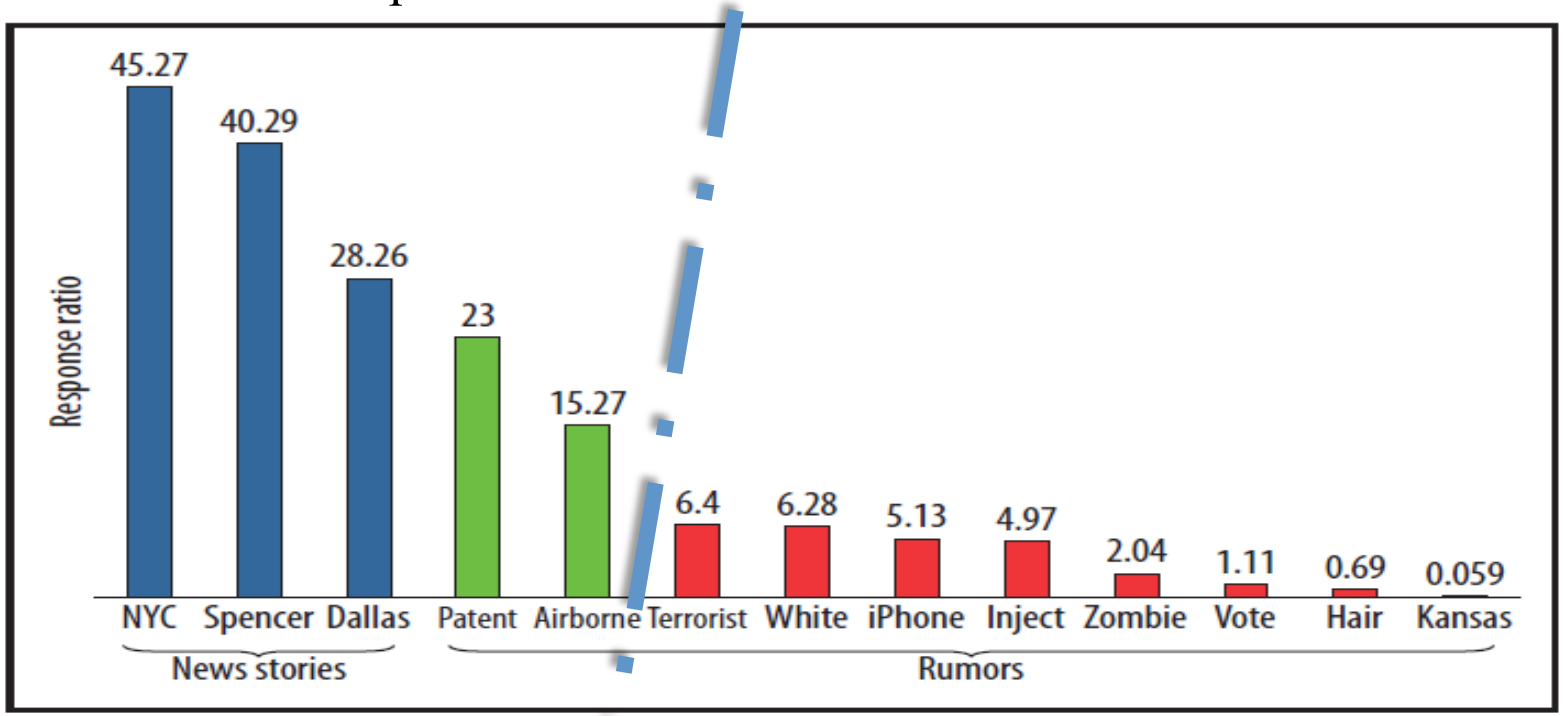


SEIZ results of Ebola rumors



White Zombie Airborne Patent

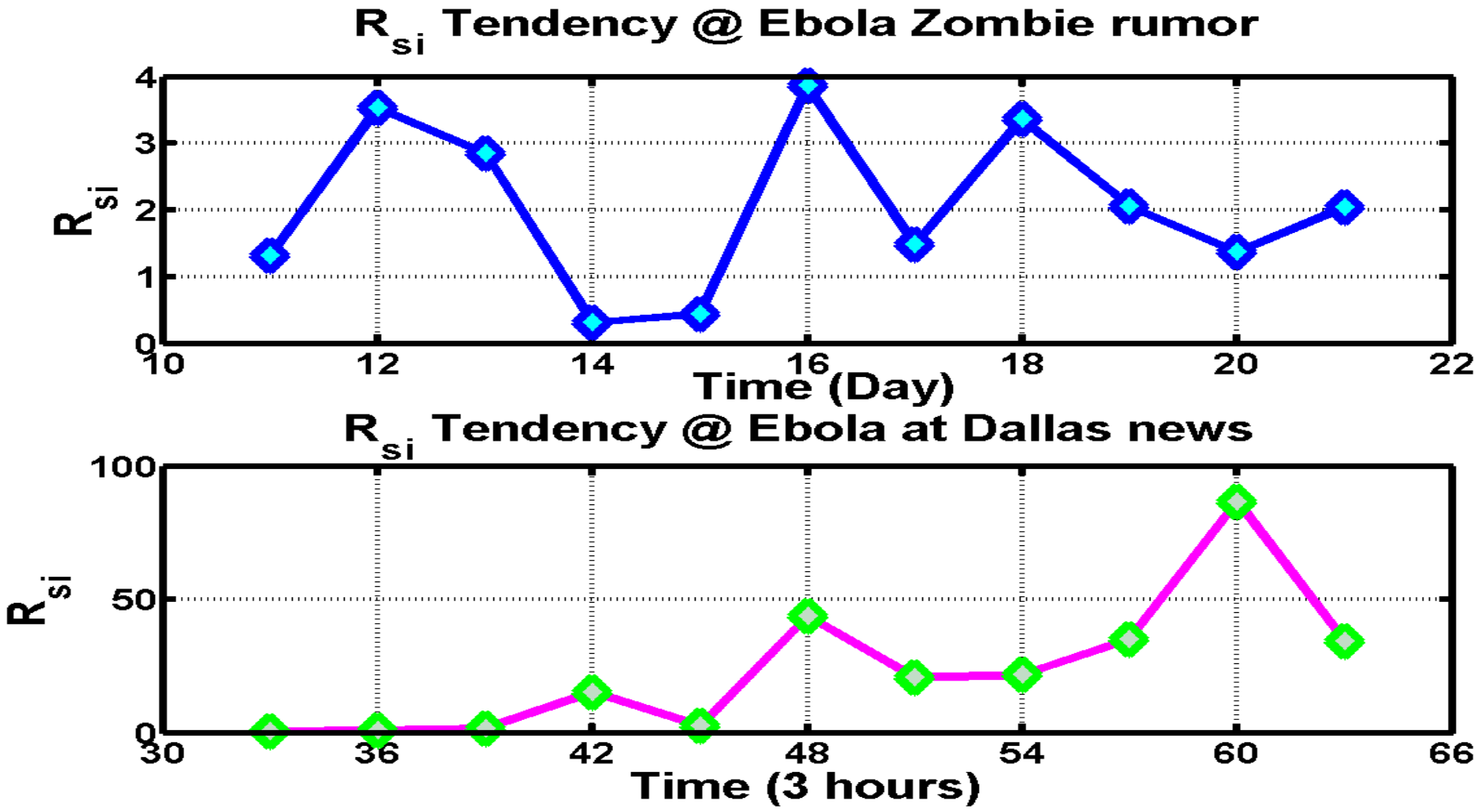
Response ratio of 3 real news and 10 rumors



Reference

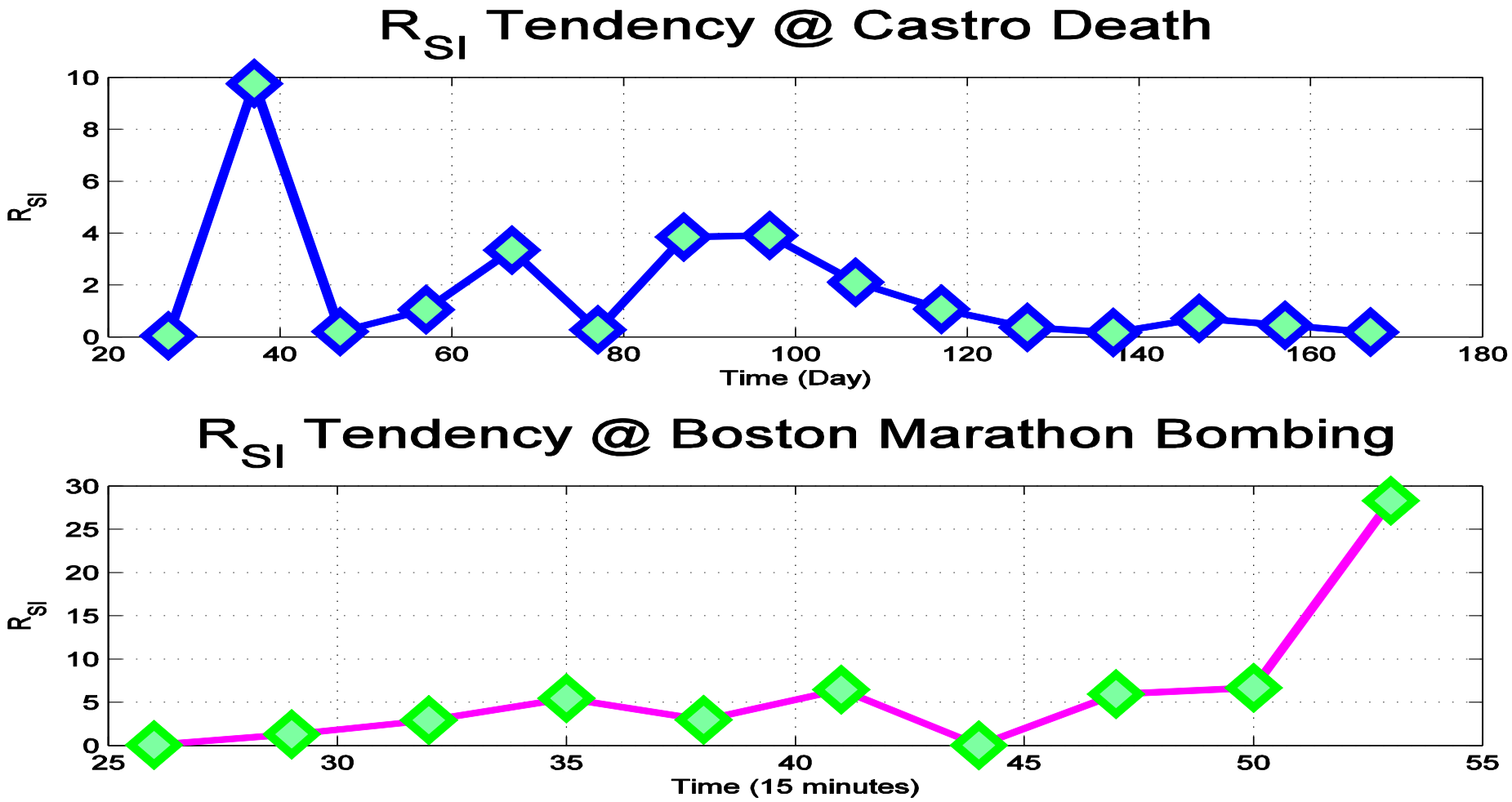
1. Liang Zhao, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. "Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling." PLOS ONE, vo. 9, no. 10 (2014): e110206.
2. **Fang Jin**, Feng Chen, Rupinder Paul Khandpur, Chang-Tien Lu, Naren Ramakrishnan. ***Absenteeism Detection in Social Media***, in Proceedings of the SIAM International Conference on Data Mining (SDM'17), Houston, TX, April 2017.
3. **Fang Jin**, Rupinder Paul Khandpur, Nathan Self, Edward Dougherty, Sheng Guo, Feng Chen, B. Aditya Prakash, Naren Ramakrishnan. ***Modeling Mass Protest Adoption in Social Network Communities using Geometric Brownian Motion***, in Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'14), Aug 2014.
4. **Fang Jin**, Edward Dougherty, Parang Saraf, Peng Mi, Yang Cao, and Naren Ramakrishnan. ***Epidemiological modeling of news and rumors on twitter***, in Proceedings of the 7th ACM SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD 2013), Chicago, IL, 2013, pages 8:1-8:9.
5. **Fang Jin**, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, Chang-Tien Lu, Naren Ramakrishnan. ***Misinformation Propagation in the age of Twitter***, IEEE Computer, Volume 47, Issue 12, pages 90-94, Dec 2014.
6. **Fang Jin**, Nathan Self, Parang Saraf, Patrick Butler, Wei Wang, Naren Ramakrishnan. ***Forex-Foreteller: Currency Trend Modeling using News Articles***, in Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining - Demo Track, pages 1470--1473, Aug 2013.

Response ratio time series



- Response ratio is dynamically changing
- Need to train classifier to dynamically classify two classes.

Response ratio time series



- Response ratio is dynamically changing
- Need to train classifier to dynamically classify two classes.

Thank you

Fang Jin: fang.jin@ttu.edu