Carnegie Mellon

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

Feng Chen and Daniel B. Neill

Carnegie Mellon University

10-7-2013

Outline

- Background
  - Research Goal
  - Why Can We Detect Events From Social Media?
  - Technical Challenges
- Our Approach
- Empirical Evaluations
- Conclusion

Research Goal

Develop methods for continuous and automated analysis of public available data in order to detect and interpret significant societal events

Model-free warning signals

Domain-specific behavior models

Civil Unrest

Disease Outbreak

Transportation Safety

Human Rights

Event Detection and Forecasting

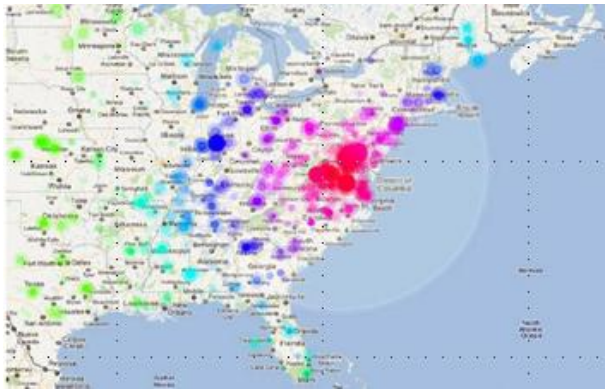
Event Casual Effects and Storytelling

Why Can We Detect Events from Social Media?

2012 July-14, Mexico Protest

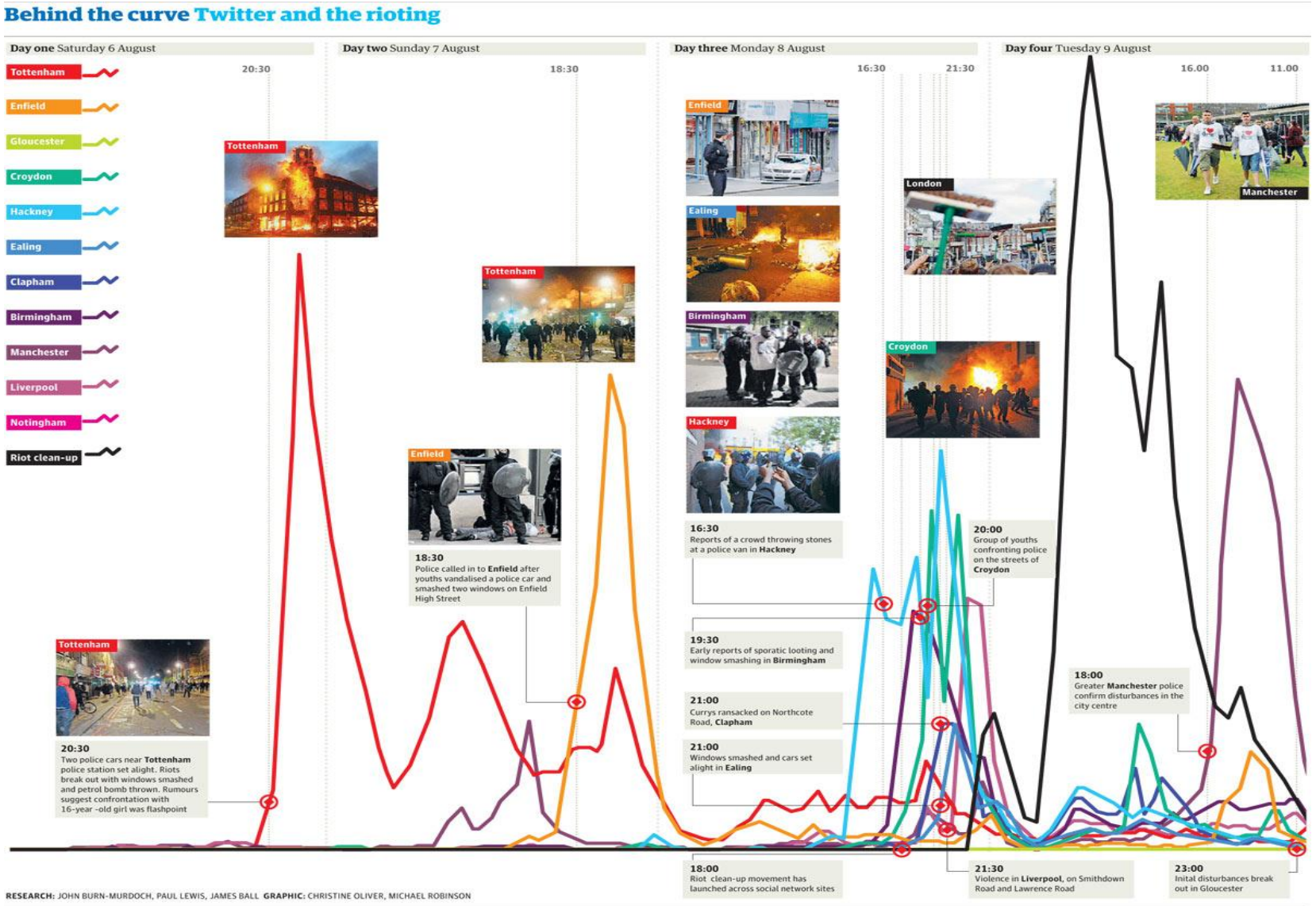
2012 Washington D.C. Traffic

Tweet Map for 2011 VA Earthquake



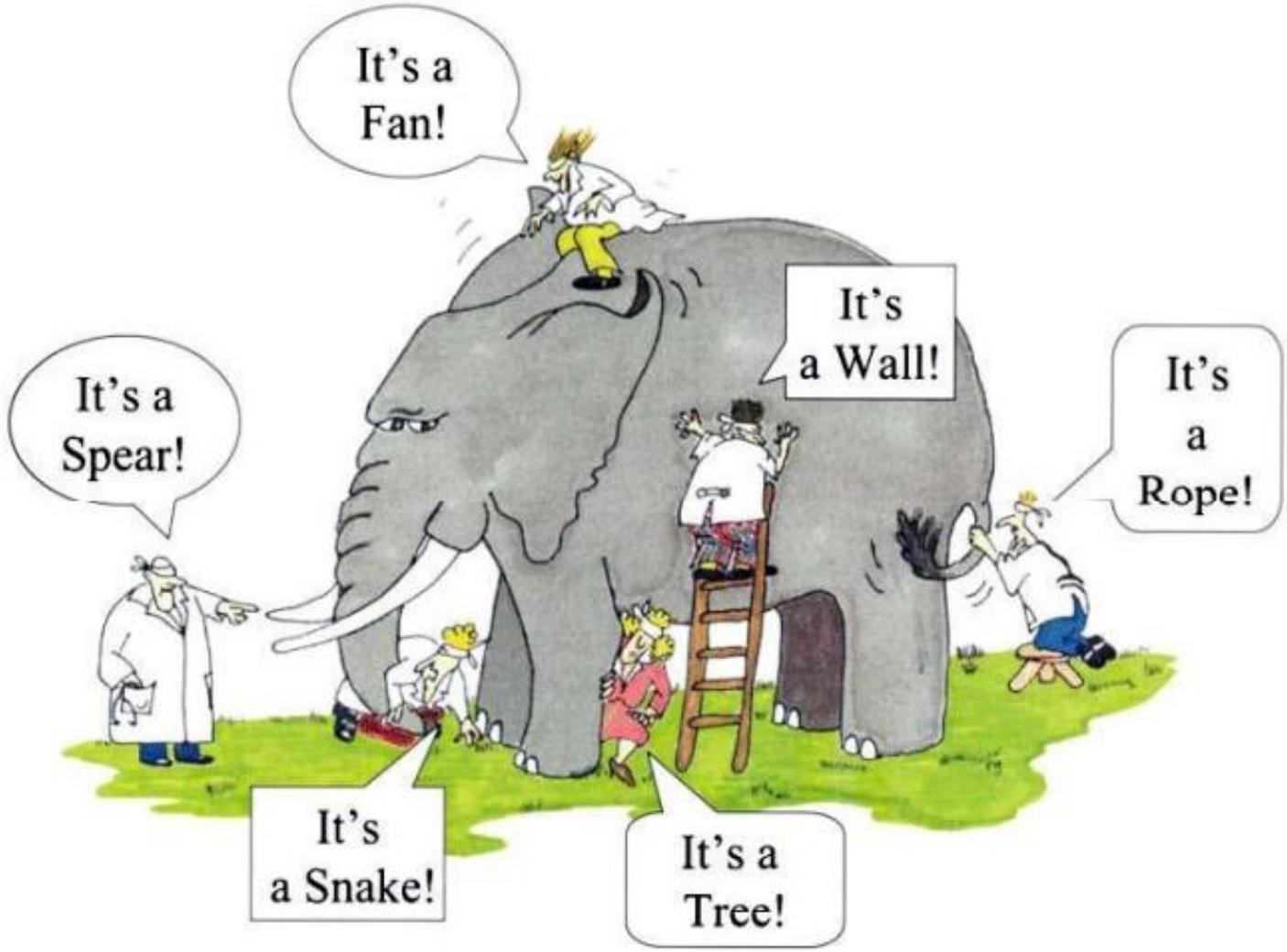
- **Event = Large-scale population behavior**
- **Social media is a real-time "sensor" of large population behavior**
- **Event Detection vs. Forecasting**
  - **Sense of public discussions about ongoing events vs. trigger events using social media**

Event Signals from Twitter Data



Credits: http://www.guardian.co.uk/uk/2011/aug/24/twitter-study-post-riot-plans

Technical Challenges



Technical Challenges

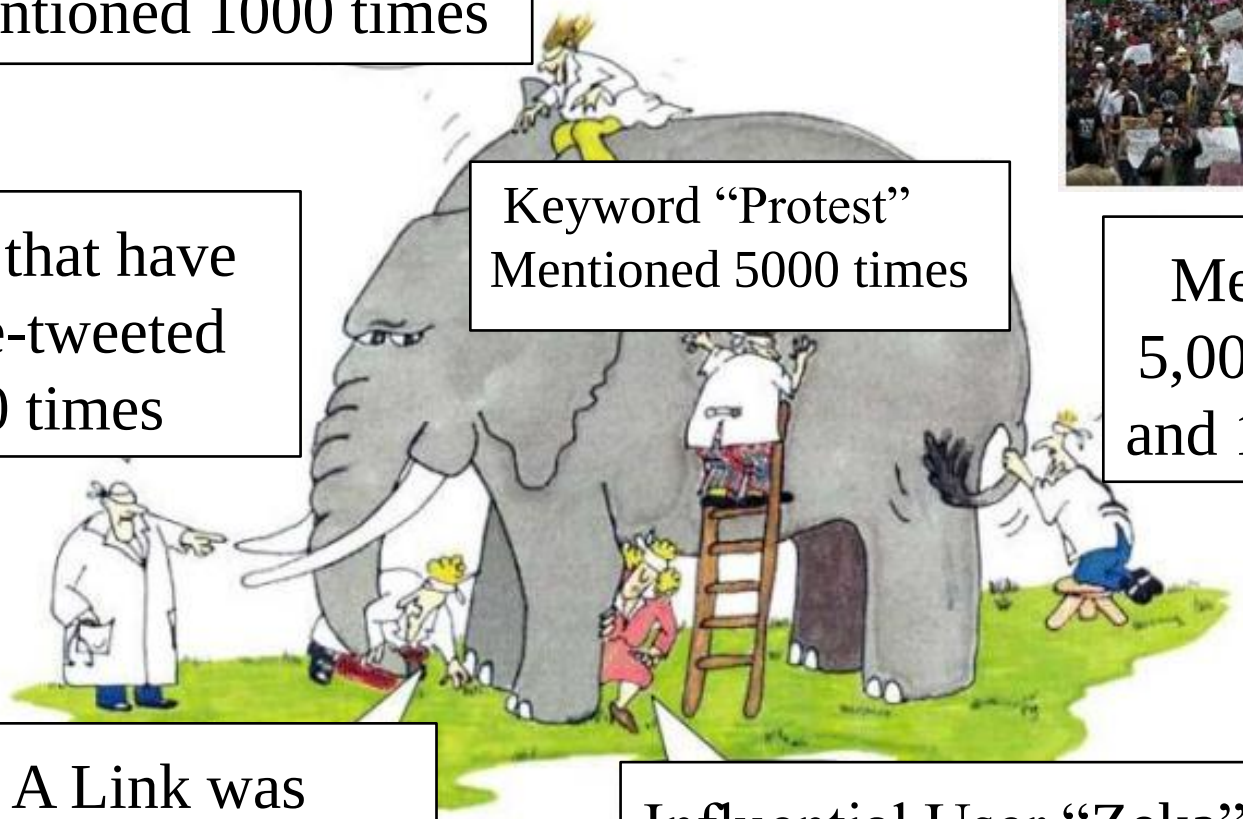
One week before 2012 Mexico President Election

"#Megamarch" mentioned 1000 times

Keyword "Protest" Mentioned 5000 times

Tweets that have been re-tweeted 1000 times

Mexico City has 5,000 Active Users and 100,000 tweets



A Link was mentioned 866 times

Influential User "Zeka" posted 10 tweets

Technical Challenges

One week before 2012 Mexico President Election

Hashtag "#Megamarch" mentioned 1000 times

Tweets that have been re-tweeted 1000 times

Keyword "Protest" Mentioned 5000 times

Mexico City: 5,000 Active Users and 100,000 tweets

A Link was mentioned 866 times

Influential User "Zeka" posted 10 tweets

?

Technical Challenges

One week before 2012 Mexico President Election

"#Megamarch" mentioned 1000 times

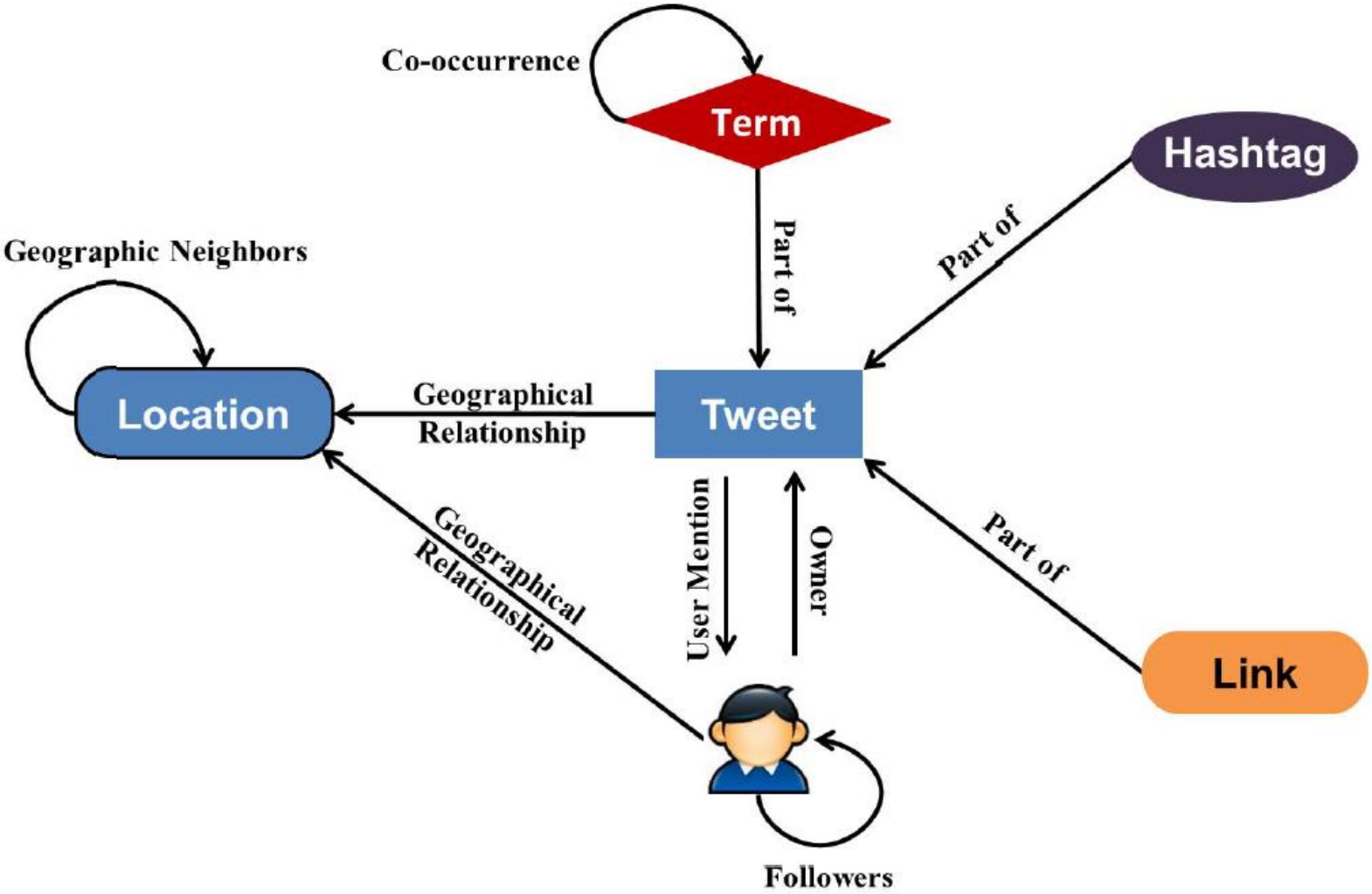
Our Solution

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

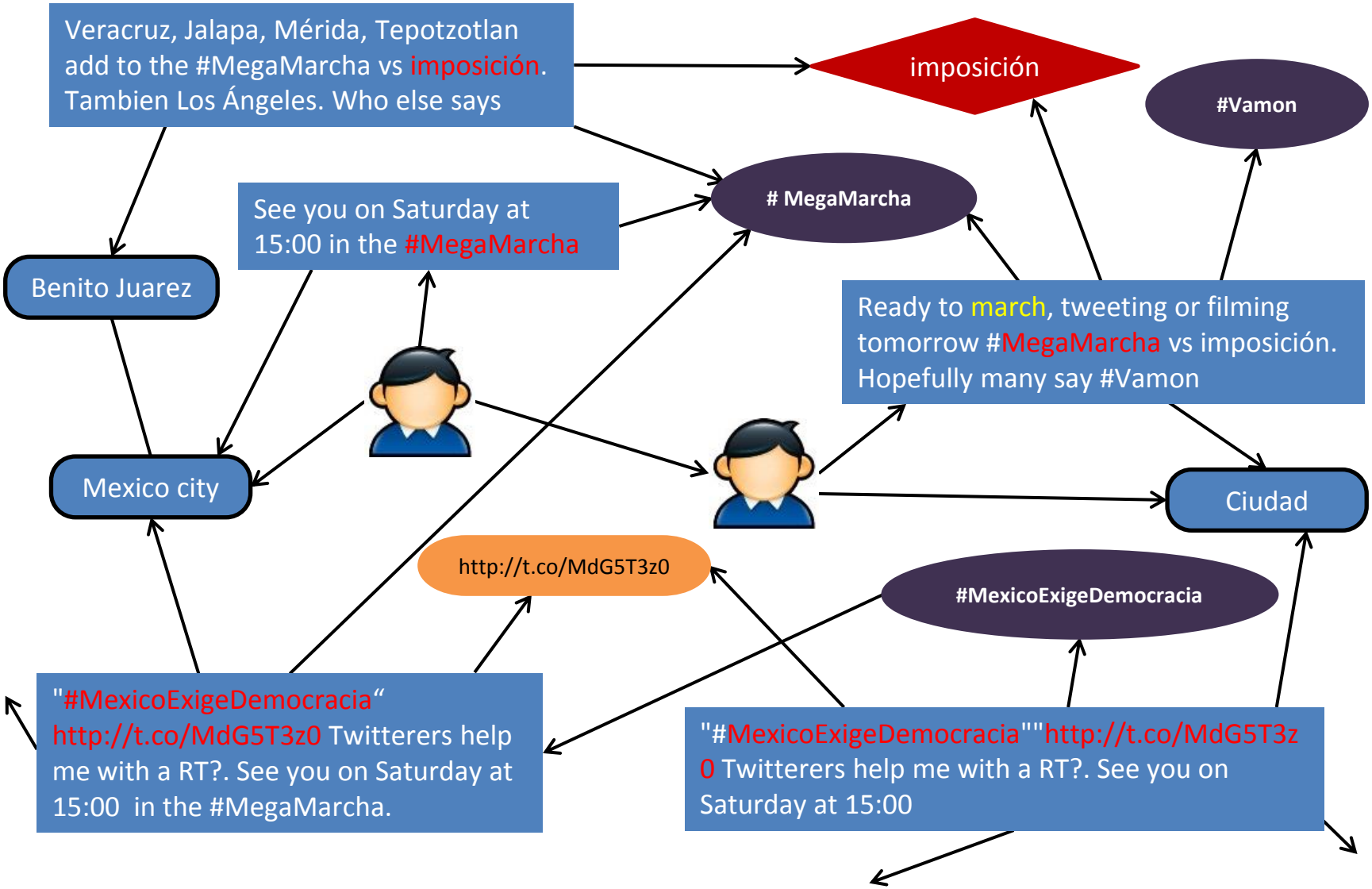
A Link was mentioned 866 times

Influential User "Zeka" posted 10 tweets

Twitter Heterogeneous Network



Twitter Heterogeneous Network (Example)



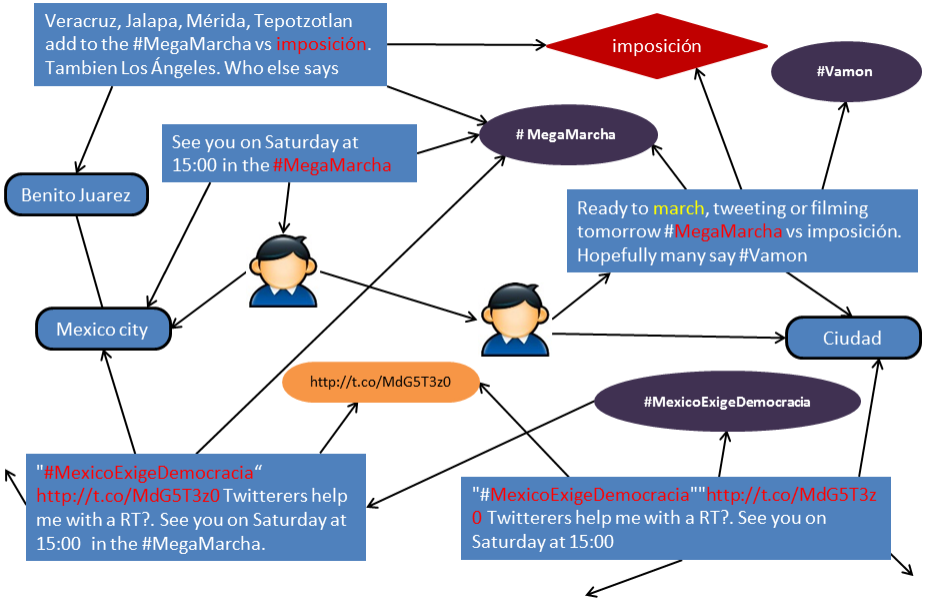
Step 1: “Sensor Network” Modeling

- **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).**

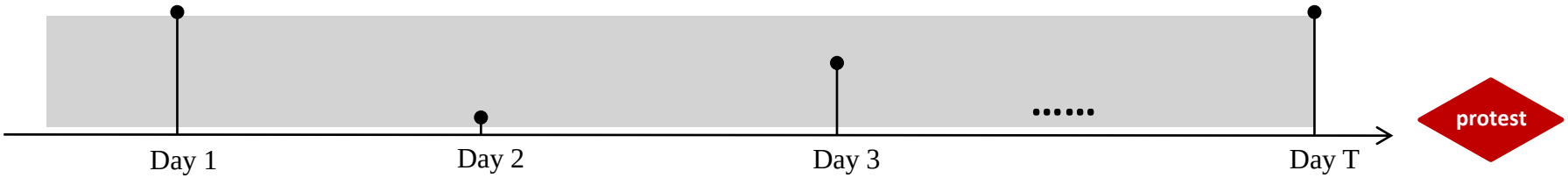
Object Type	Features
User	# tweets, # retweets, # followers, #followees, #mentioned_by, #replied_by, diffusion graph depth, diffusion graph size
Tweet	Klout, sentiment, replied_by_graph_size, reply_graph_size, retweet_graph_size, retweet_graph_depth
City, State, Country	# tweets, # active users
Term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

Step 2: Sensor Signals → Empirical P-values

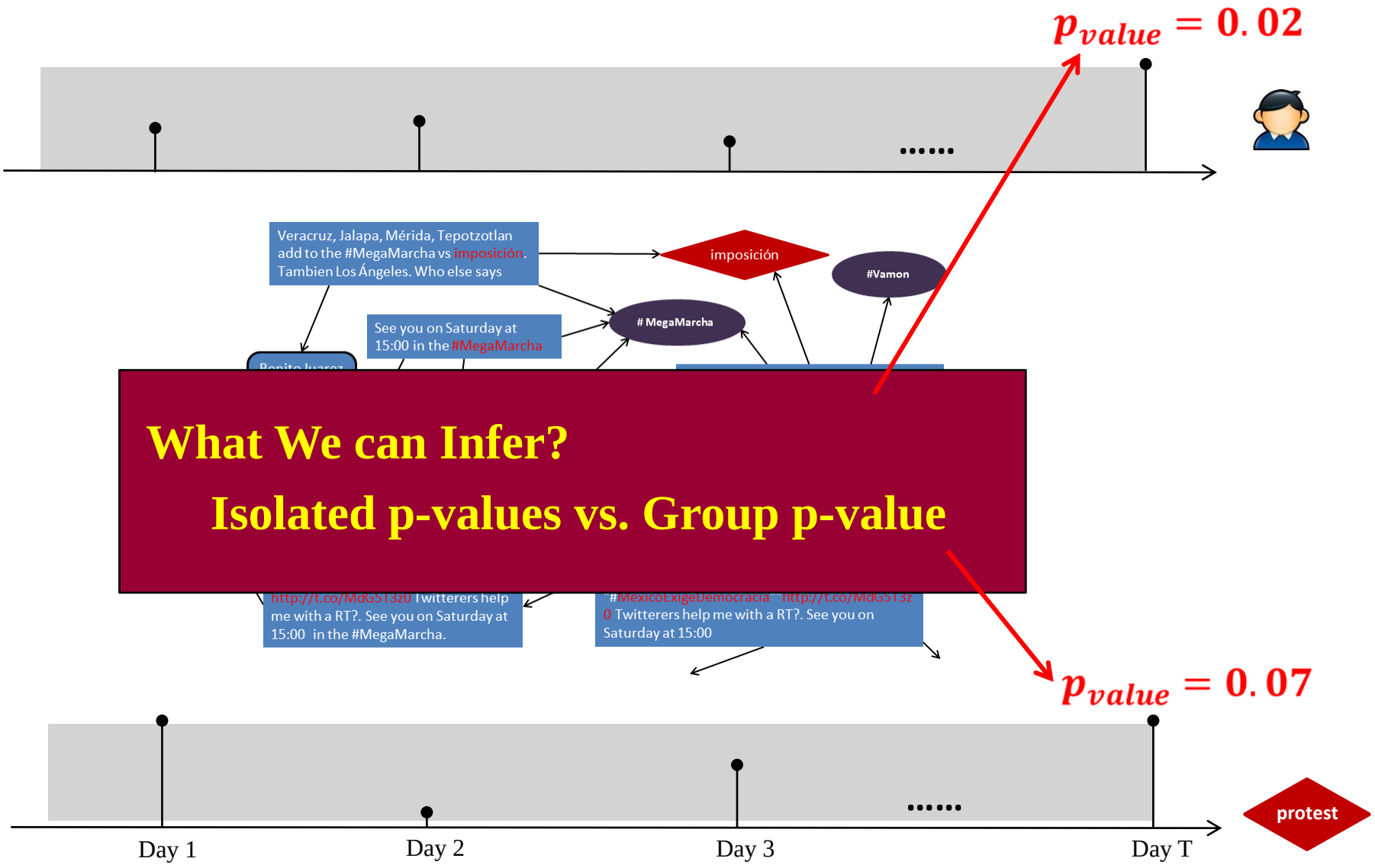
$p_{value} = 0.02$



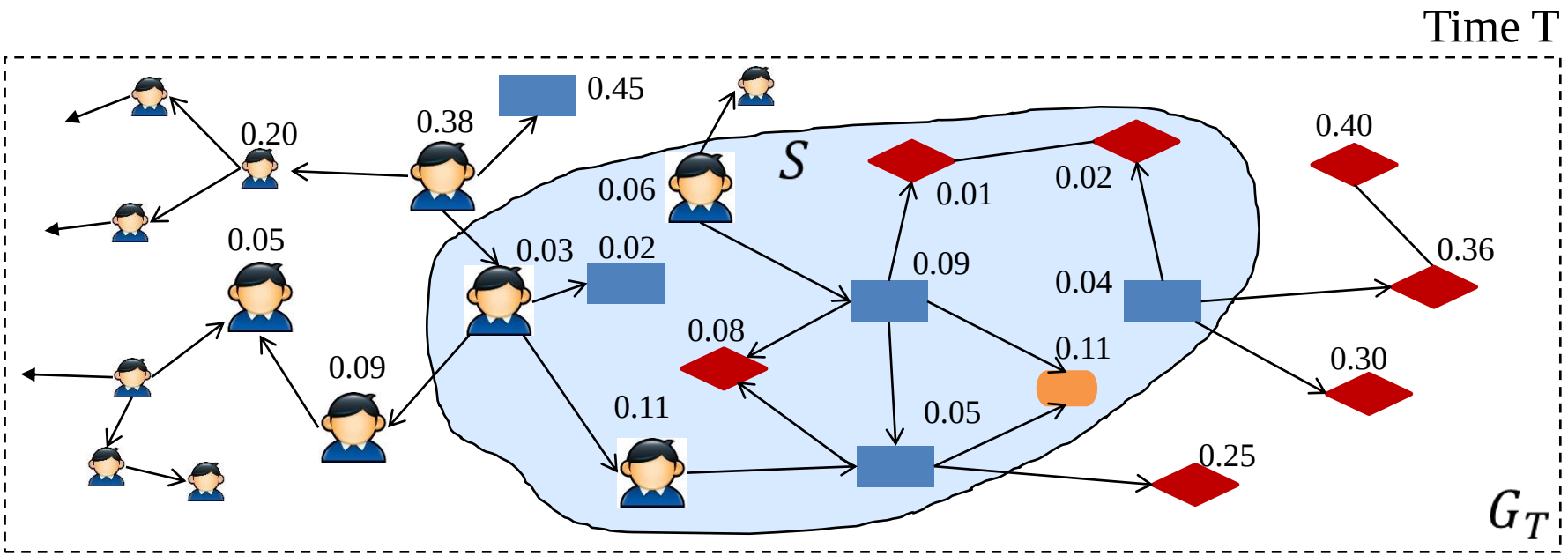
$p_{value} = 0.07$



Step 2: Sensor Signals → Empirical P-values



Step 3: Event Detection on "Sensor Network"



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

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

Empirical Evaluations

Country	# of tweets	News source*
Argentina	29 ,000,000	Clarín; La Nación; Infobae
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

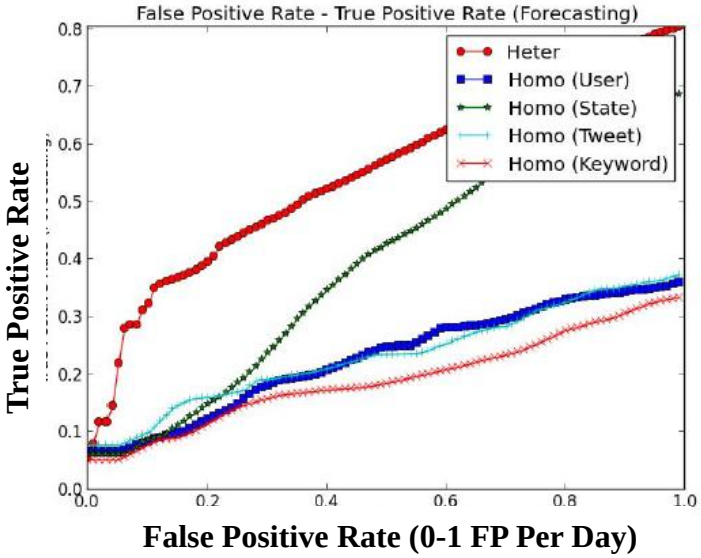
Time Period: From **2012 Jul.** to **2012 Dec.** Totally **918** civil unrest events

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/").

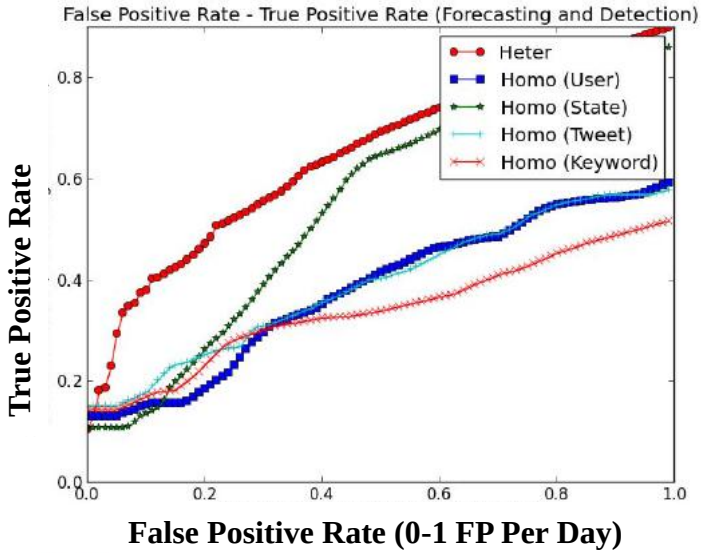
Our approach (NPHGS) vs. homogenous graph scan methods

Our approach (NPHGS) vs. competitive event detection methods

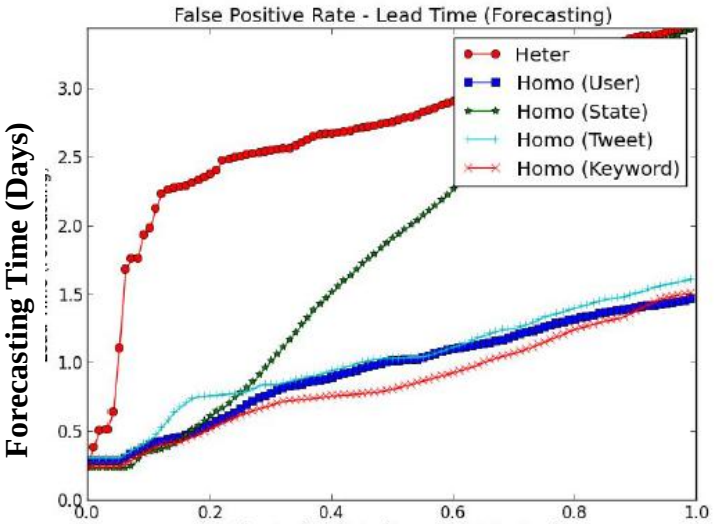
Our Approach vs. Homogenous Graph Scan Methods



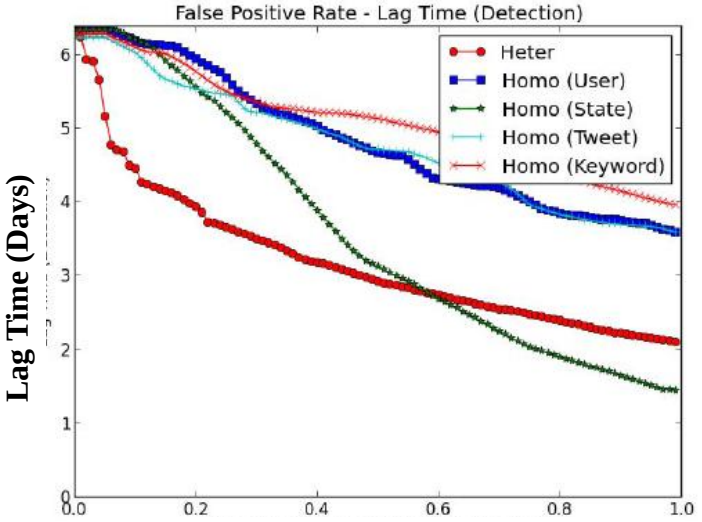
(a) FPR vs. TPR (Forecasting)



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



(c) FPR vs. Lead Time (Forecasting)



(d) FPR vs. Lag Time (Detection)

Our Approach vs. Competitive Event Detection 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

Conclusion

- Social media is a “sensor network” of people’s sentiments and opinions
- 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 problem of automatic event detection and storytelling using social media

Thank you!

Questions?