# Rhythms of Information Flow through Networks

Jure Leskovec,
Stanford University

Includes joint work with Jaewon Yang, Manuel Gomez-Rodriguez, Andreas Krause, Lars Backstrom and Jon Kleinberg.

# Information and Networks

- Information reaches us…
  - …by personal influence in our social networks
  - …through influence by mainstream media
- How does information transmitted by the mass media interact with the personal influence arising from social networks?
- From its early stages, a tension between global effects from the mass media and local effects carried by social structure

# Online (Social) Media

- Traditionally it was hard to capture and quantify the effects of media and social networks
- Explosion of online (social) media:
  - Traditional (TV, Newspapers, Agencies)
  - Blogs (personal/professional)
  - Microblogging (Twitter)

# Social Media: The New Picture

- Internet, blogs and social media changed the traditional picture:
  - Social media means the dichotomy between global and local influence is evaporating
  - Speed of media reporting and discussion has intensified: very rapid progression of stories
  - Information reaches us in small increments from real-time sources and via social networks
- How should this change our understanding of information consumption and of the role of social networks?

# Plan for the Talk

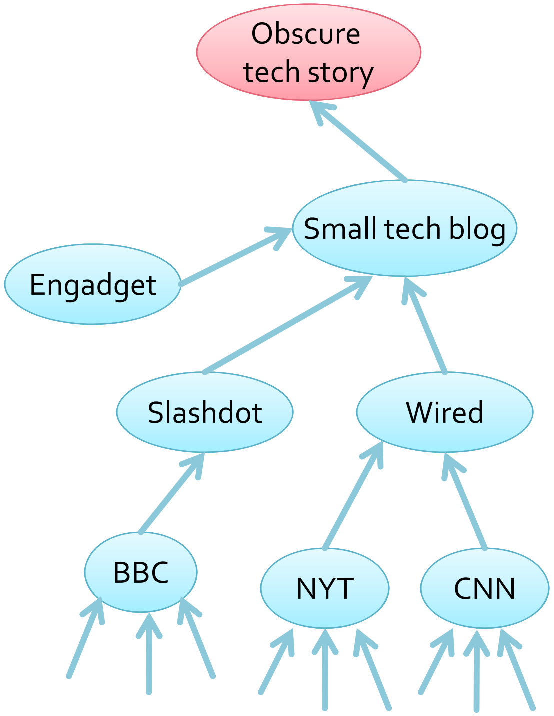
- Plan for the talk:
  - Analyze underlying mechanisms for the real-time spread of information through on-line networks
- Motivating questions:
  - How to track messages as they spread?
  - How to model/predict the spread of information?
  - How to identify networks over which the information spreads?

# Challenges and Opportunities

- In principle, we can collect nearly all (online) social media content:
  - 20 million articles/**day** (50GB of data/day)
    - 20,000 news sources + millions blogs and forums
- What are basic "units" of information?
  - Pieces of information that propagate between the nodes (users, media sites, …)
    - phrases, quotes, messages, links, tags

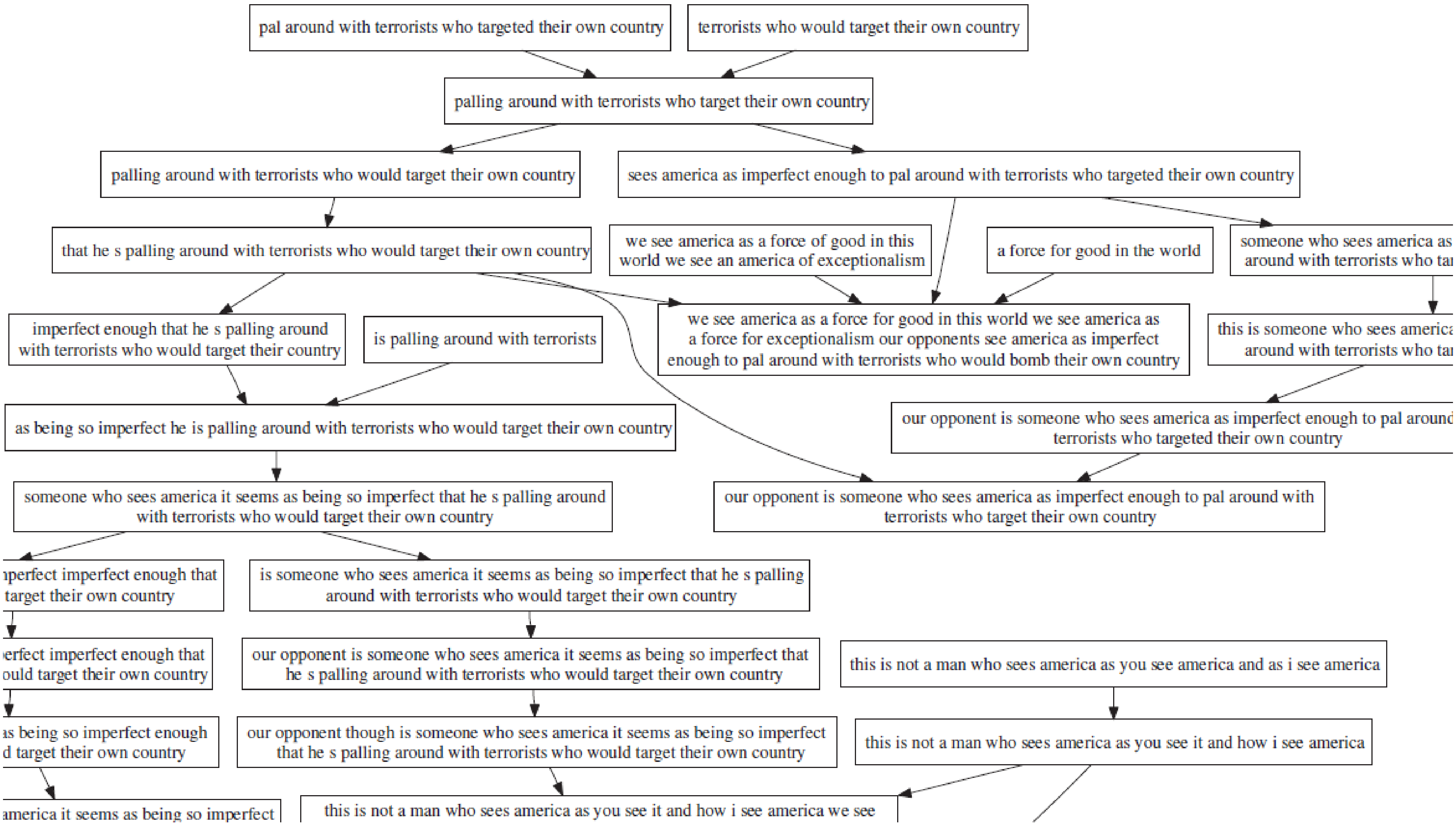
# Extracting Units of Information

- Would like units that:
  - correspond to pieces of information
  - vary over the order of days,
  - and can be handled at terabyte scale
- Approaches:
  - Cascading links to individual articles
    [Adamic-Adar ‘05] [SDM ‘07]
  - Textual fragments that travel relatively unchanged, in this case through many news articles:
    - Look for phrases inside quotes: “…”
    - About 1.25 quotes per document in our data



# Challenge: Quotes Mutate… A lot!

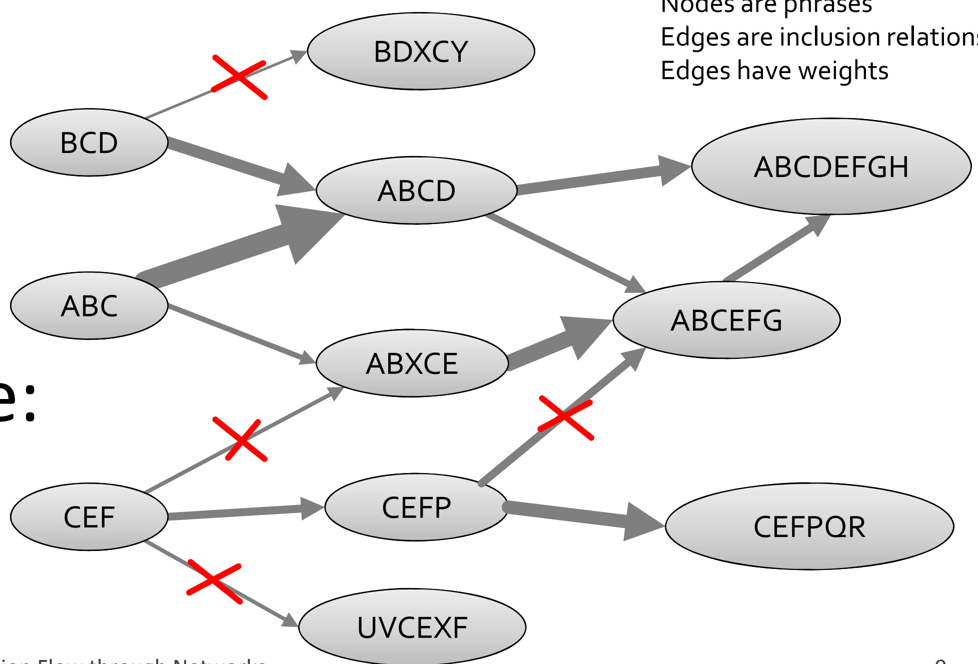
[KDD '09]



# Finding Mutational Variants

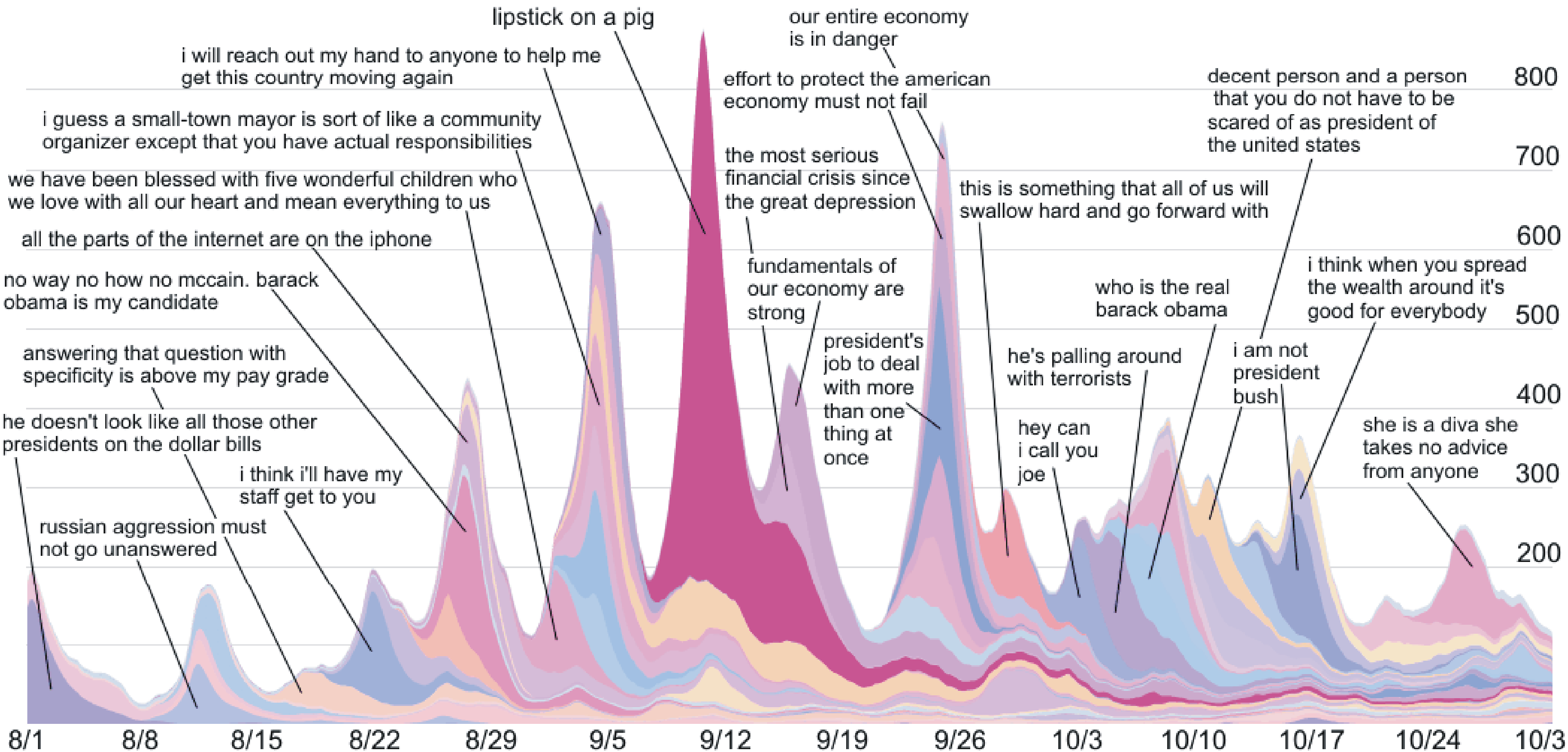
- Goal: Find mutational variants of a quote
- **Objective:**
  - In a DAG of approx. quote inclusion, delete min total edge weight s.t. each component has a single "sink"
- Our basic units are quotes
  - Length ≥ 4, freq. ≥ 10
    Gives 22M quotes
- DAG-partitioning is NP-hard but heuristics are effective:
  - Gives ~35,000 non-trivial clusters

Nodes are phrases
Edges are inclusion relations
Edges have weights



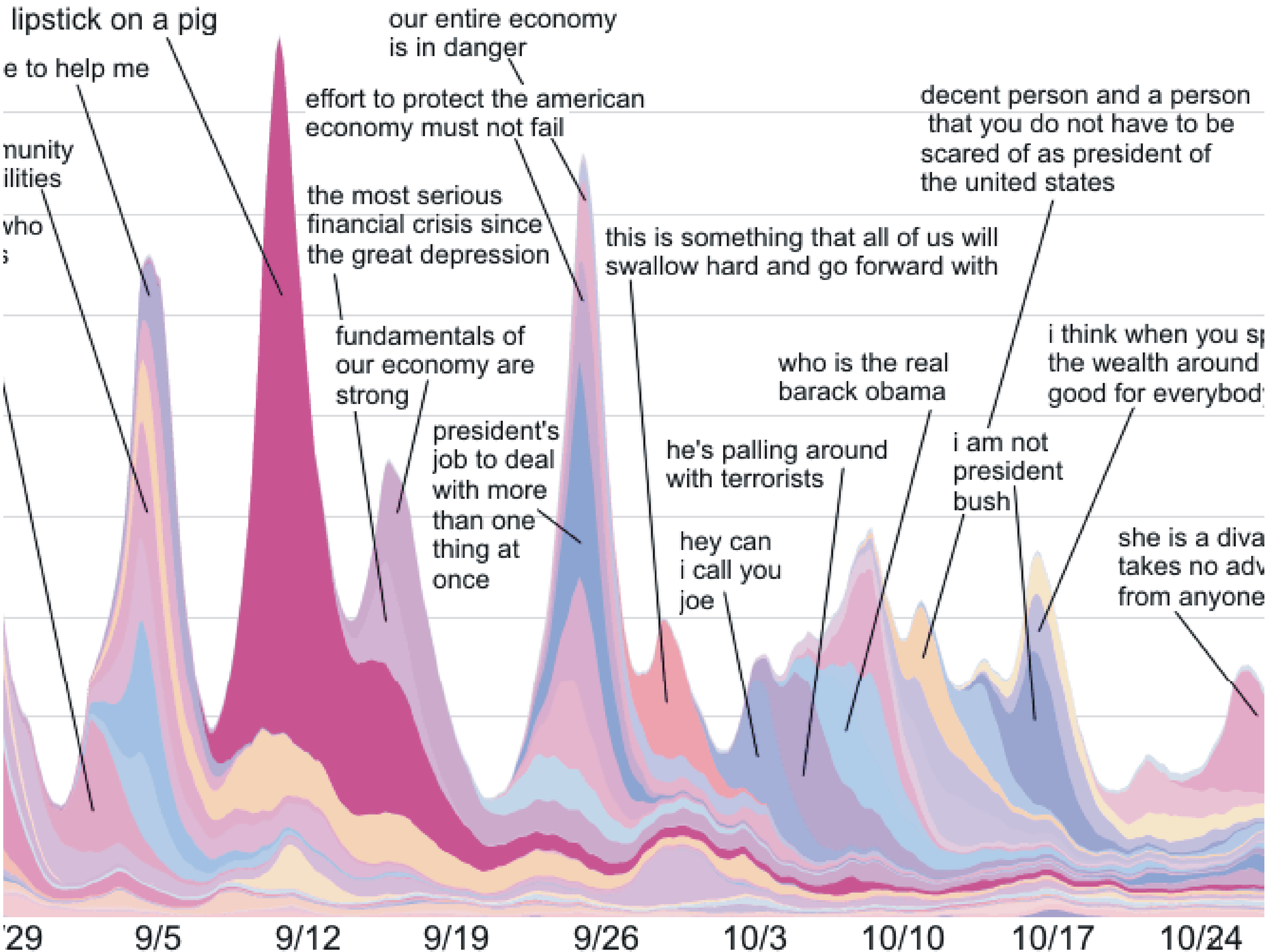
http://memetracker.org

# Cluster Volume over Time




**August** **October**

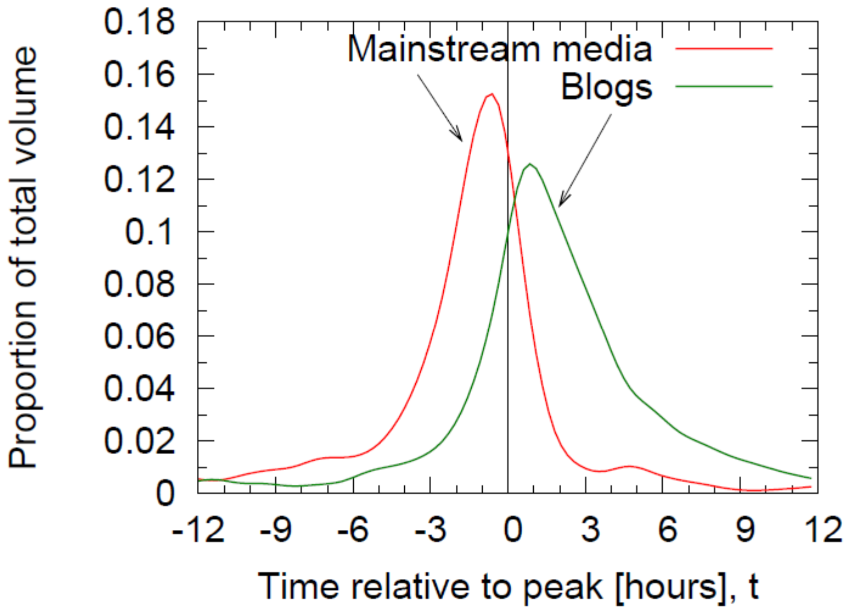
Volume over time of top 50 largest total volume clusters



lipstick on a pig
e to help me
nunity
ilities
vho
our entire economy
is in danger
effort to protect the american
economy must not fail
the most serious
financial crisis since
the great depression
fundamentals of
our economy are
strong
president's
job to deal
with more
than one
thing at
once
this is something that all of us will
swallow hard and go forward with
decent person and a person
that you do not have to be
scared of as president of
the united states
who is the real
barack obama
he's palling around
with terrorists
hey can
i call you
joe
i am not
president
bush
i think when you sp
the wealth around
good for everybod
she is a diva
takes no adv
from anyone
'29
9/5
9/12
9/19
9/26
10/3
10/10
10/17
10/24

[KDD '09]

# Interaction of News and Blogs

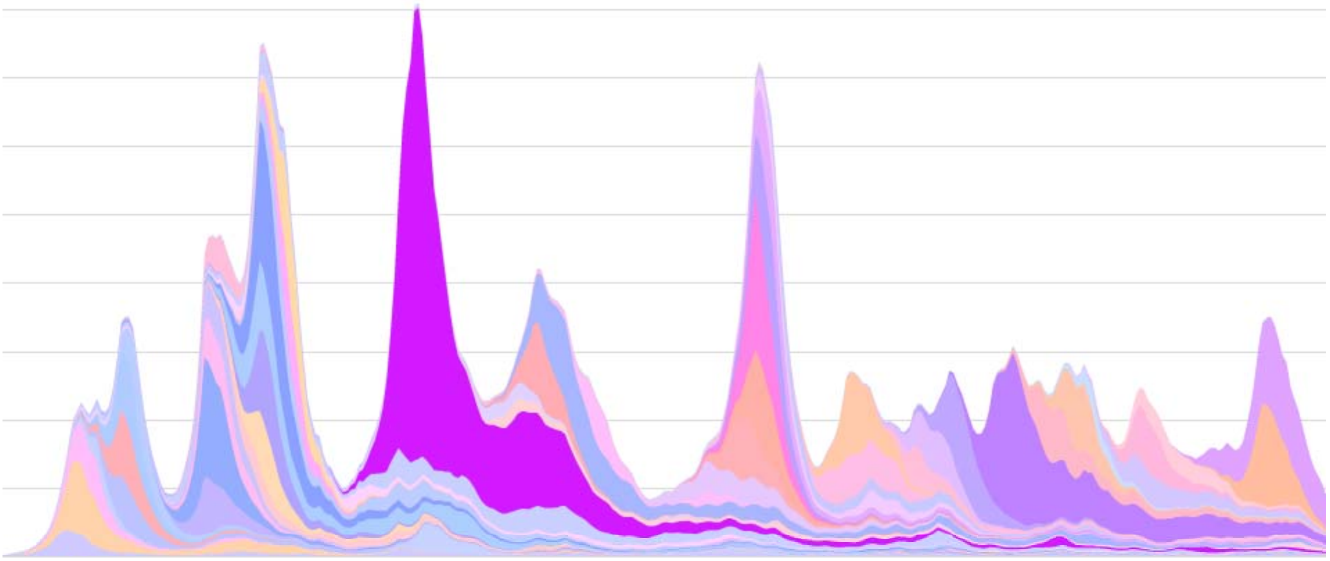



| Rank | Lag [h] | Reported | Site |
|---|---|---|---|
| 1 | -26.5 | 42 | hotair.com |
| 2 | -23 | 33 | talkingpointsmemo.com |
| 4 | -19.5 | 56 | politicalticker.blogs.cnn.com |
| 5 | -18 | 73 | huffingtonpost.com |
| 6 | -17 | 49 | digg.com |
| 7 | -16 | 89 | breitbart.com |
| 8 | -15 | 31 | thepoliticalcarnival.blogspot.com |
| 9 | -15 | 32 | talkleft.com |
| 10 | -14.5 | 34 | dailykos.com |
| 30 | -11 | 32 | uk.reuters.com |
| 34 | -11 | 72 | cnn.com |
| 40 | -10.5 | 78 | washingtonpost.com |
| 48 | -10 | 53 | online.wsj.com |
| 49 | -10 | 54 | ap.org |

- Can study a division of sources into news and blogs
  - Peak intensity from blogs typically comes about 2.5 hours after peak intensity from news
  - Can classify individual sources by their typical timing relative to the peak aggregate intensity.

[WSDM '11]

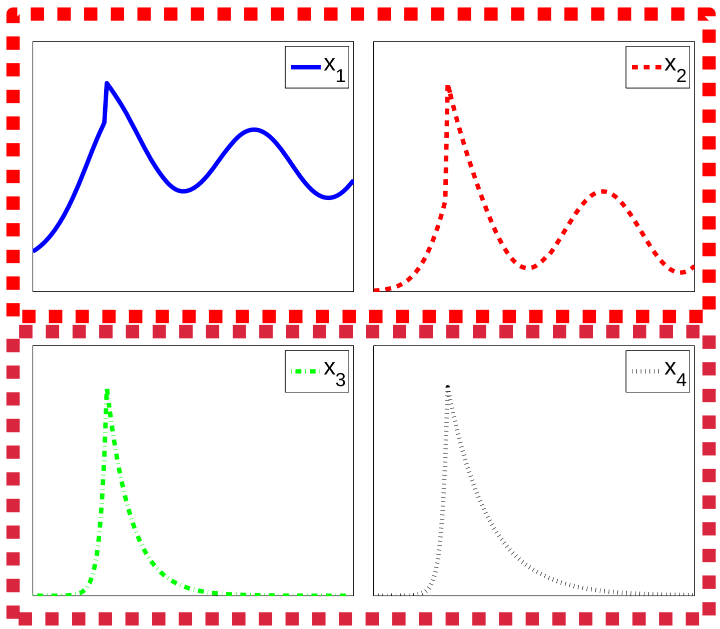
# Patterns of Information Attention



- **Q: What are temporal patterns of information attention?**
  - Item *i*: Piece of information (e.g., quote, url, hashtag)
  - Volume $x_i(t)$: # of times *i* was mentioned at time *t*
    - Volume = number of mentions = attention = popularity
  - Q: Typical classes of shapes of $x_i(t)$

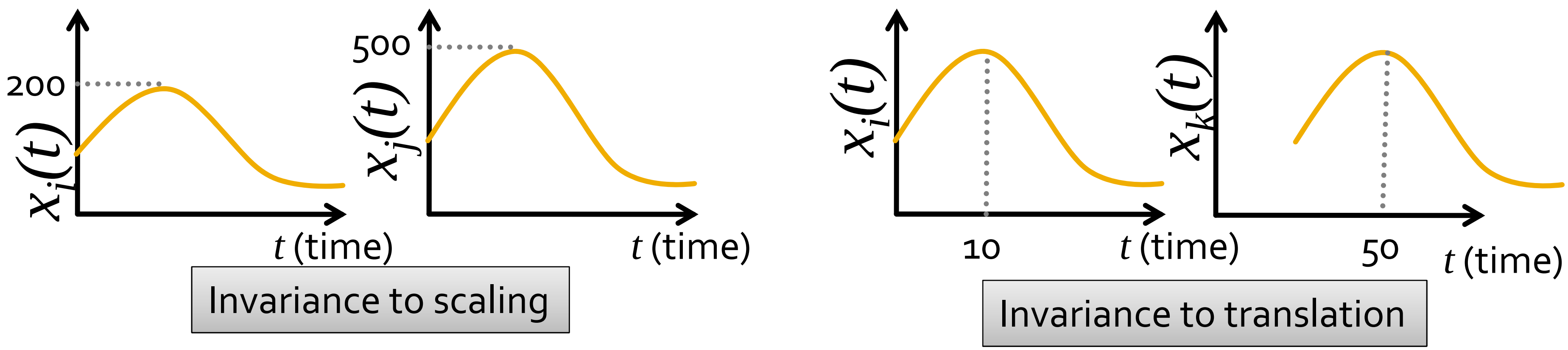
# Discovering Attention Patterns

- Given: Volume of an item over time
- Goal: Want to discover types of shapes of volume time series



# Clustering Temporal Signatures

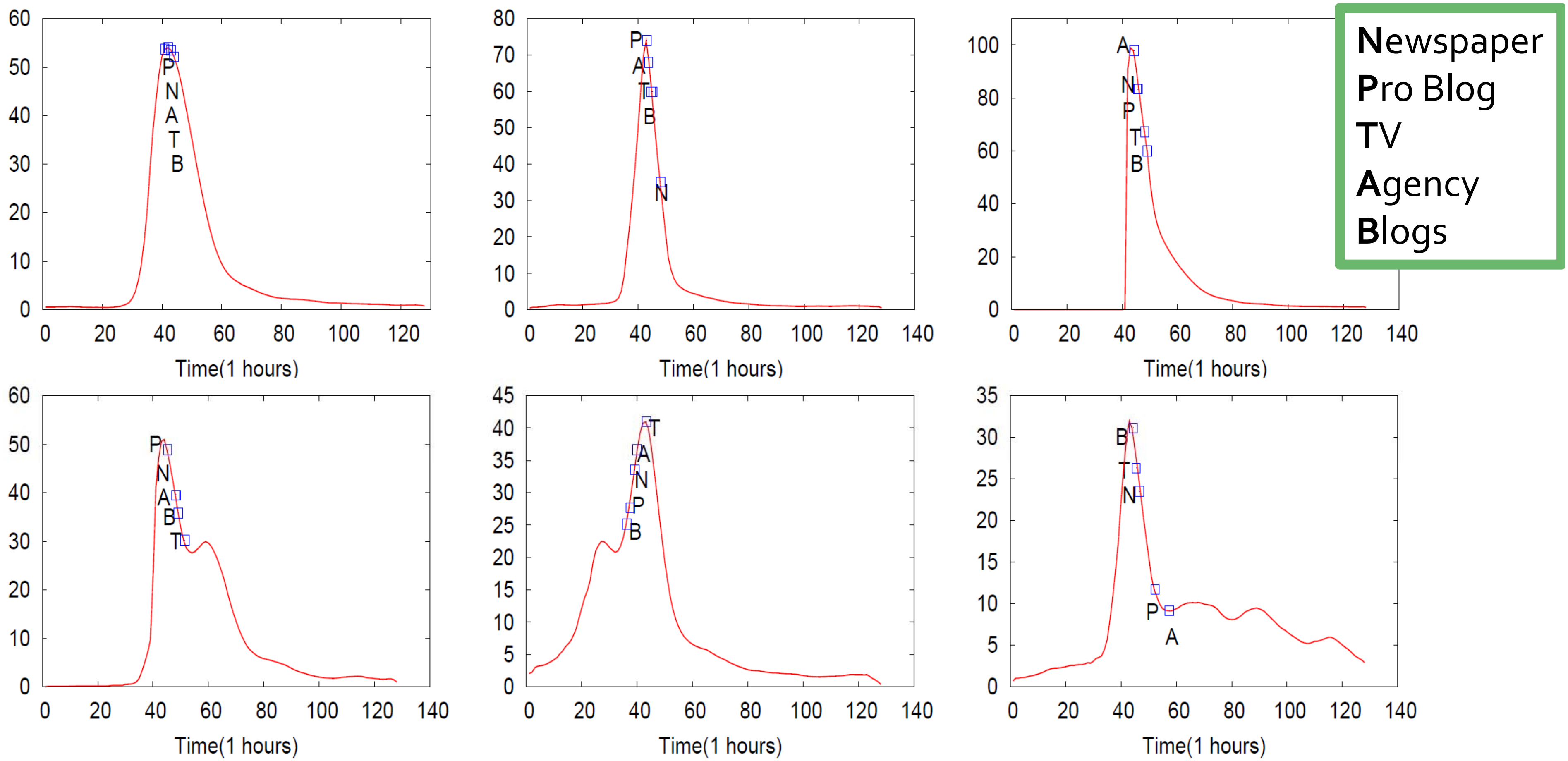
- Goal: Cluster time series & find cluster centers
- Time series distance function needs to be:




$$d(x,y) = \min_{a,q} \sum_t \left(x(t) - a \cdot y(t-q)\right)^2$$

- K-Spectral Centroid clustering [WSDM '11]

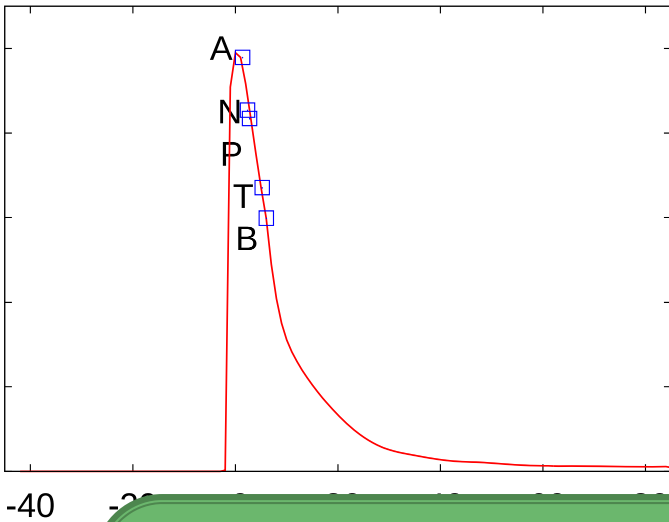
# Patterns of Attention

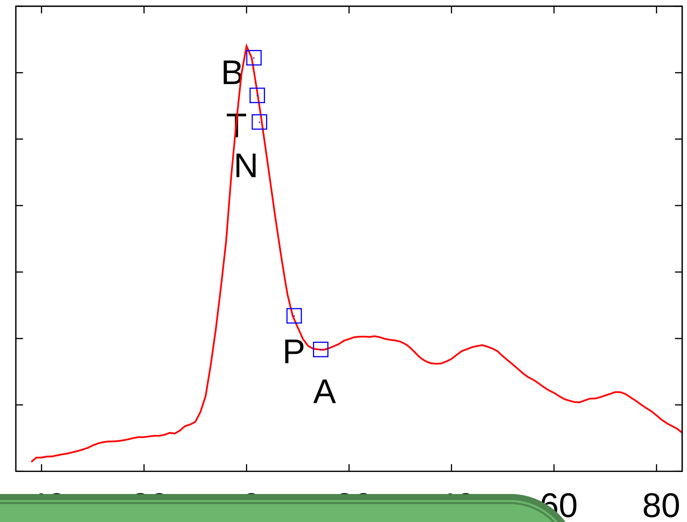



- Quotes: 1 year, 172M docs, 343M quotes
- **Same 6 shapes** for Twitter: 580M tweets, 8M #tags

[WSDM '11]

# Analysis of Attention Patterns




Different media give raise to different patterns

•Spike

Agenci

•Slow &

• Blogs

the mainstream media

• Blog volume = 29.1%

h in

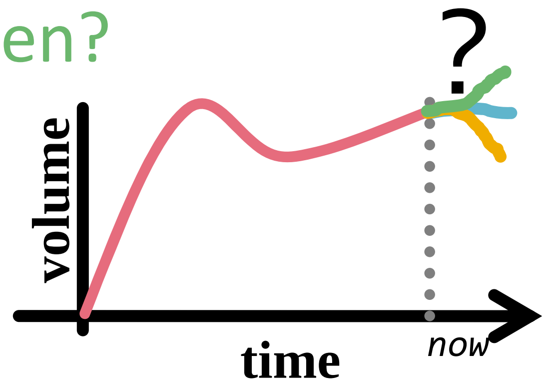
fore

mainstream media

• Blog volume = 53.1%

# Predicting Attention

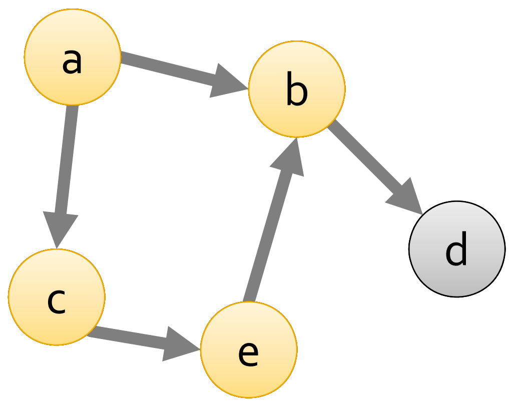
- **How much attention will information get?**
  - Who reports the information and when?
    - 1h: Gizmodo, Engadget, Wired
    - 2h: Reuters, Associated Press
    - 3h: New York Times, CNN
    - How many will mention the info at time 4, 5,…?




- Motivating question:
  - If NYT mentions info at time *t*
  - How many subsequent mentions of the info will this generate at time *t+1*, *t+2*, …?

# Predicting Information Diffusion

- Goal:
  - Predict future attention (number of mentions)
- Traditional view:
  - In a network "infected" nodes spread info to their neighbors




- Problem:
  - The network may be unknown
- **Idea:** Predict the future attention based on which nodes got "infected" in the past

[ICDM '10]

# The Linear Influence Model

How to predict future volume $x_i(t+1)$ of info $i$ ?

- Node $u$ has an **influence function** $I_u(q)$:
  - $I_u(q)$: After node $u$ gets "infected", how many other nodes tend to get infected $q$ hours later
    - E.g.: Influence function of CNN:
      How many sites say the info after they see it on CNN?
  - Estimate the influence function from past data
- Predict future volume using the influence functions of nodes infected in the past

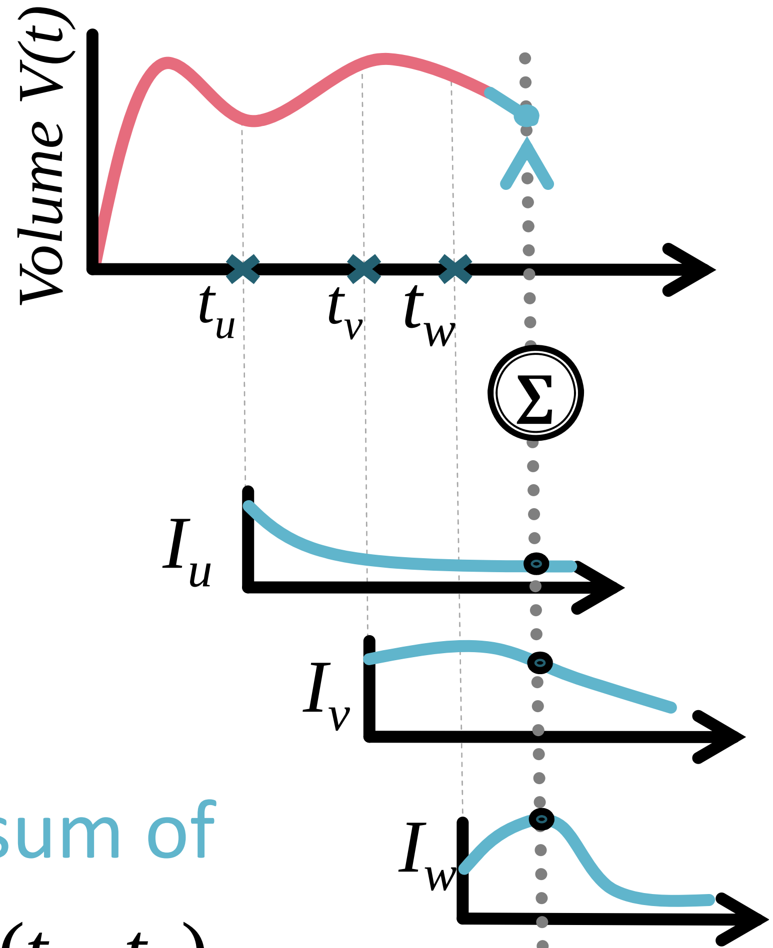
[ICDM '10]

# The Linear Influence Model

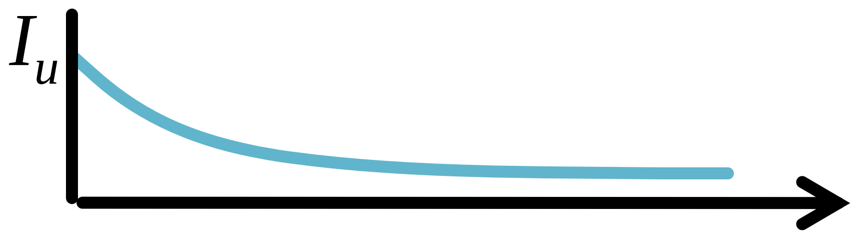
LIM model:

- Volume $x_i(t)$ of $i$ at time $t$
- $A_i(t)$ … a set of nodes that mentioned $i$ before time $t$

- And let:
  - $I_u(q)$: influence function of $u$
  - $t_u$: time when $u$ mentioned $i$
- Predict future volume as a sum of influences: $x_i(t+1) = \sum_{u \in A_i(t)} I_u(t - t_u)$



# Estimating Influence Functions

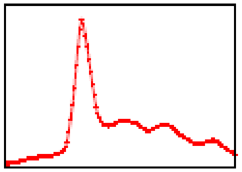


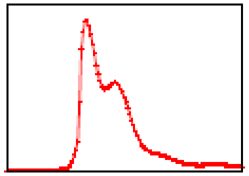

- After node *u* mentions the info, $I_u(q)$ other mentions tend to occur *q* hours later
  - $I_u(q)$ is not observable, need to estimate it
  - Make no assumption about its shape
    - Model $I_u(q)$ as a vector: $I_u(q) = [I_u(1), I_u(2), I_u(3),\ldots, I_u(L)]$
  - Find $I_u(q)$ by solving a least-squares-like problem:

$$\min_{I_u, \forall u} \sum_i \sum_t \left( x_i(t+1) - \sum_{u \in A_i(t)} I_u(t - t_u) \right)^2$$

# The model: Performance

- Take top 1,000 quotes by the total volume:
  - Total 372,000 mentions on 16,000 websites
- Build LIM on 100 highest-volume websites
  - $x_i(t)$ … number of mentions across 16,000 websites
  - $A_i(t)$ … which of 100 sites mentioned quote *i* and when
- Improvement in L2-norm over 1-time lag predictor





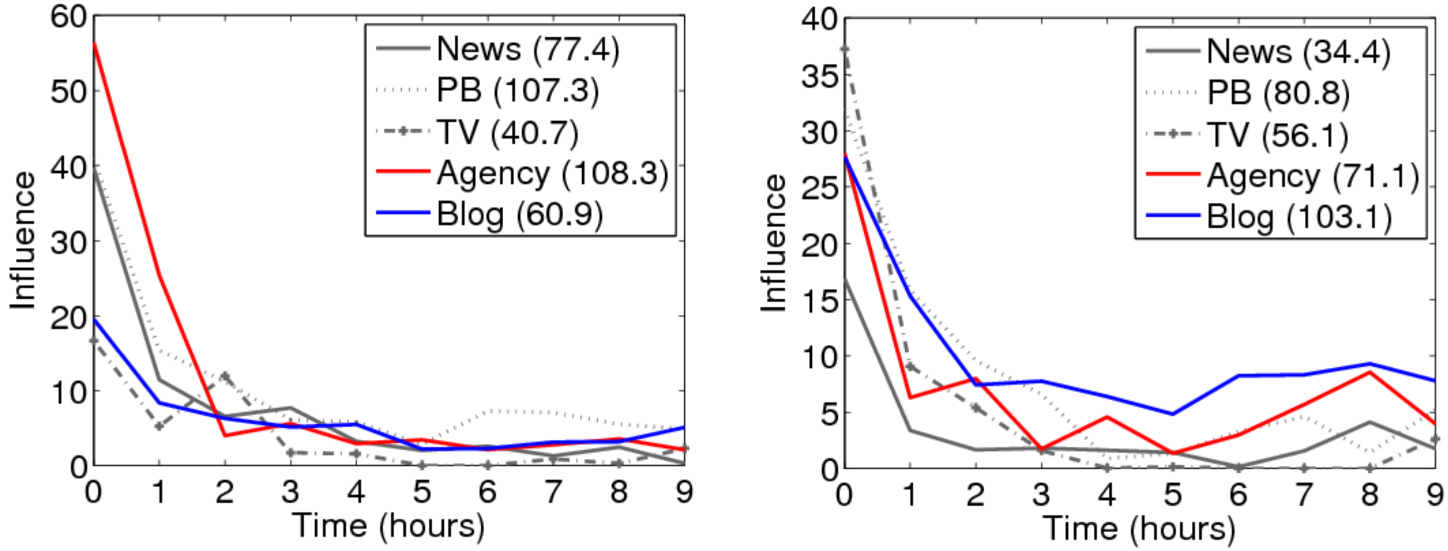
| | Bursty phrases | Steady phrases | Overall |
|---|---|---|---|
| AR | 7.21% | 8.30% | 7.41% |
| ARMA | 6.85% | 8.71% | 7.75% |
| **LIM (N=100)** | **20.06%** | **6.24%** | **14.31%** |

[ICDM '10]

# Analysis of Influence Functions

- Influence functions give insights:
  - **Q:** NYT writes a post on politics, how many people tend to mention it next day?
  - **A:** Influence function of NYT for political phrases!
- Experimental setup:
  - 5 media types:
    - Newspapers, Pro Blogs, TVs, News agencies, Blogs
  - 6 topics:
    - Politics, nation, entertainment, business, technology, sports
  - For all phrases in the topic, estimate average influence function by media type

# Analysis of Influence




Politics Entertainment

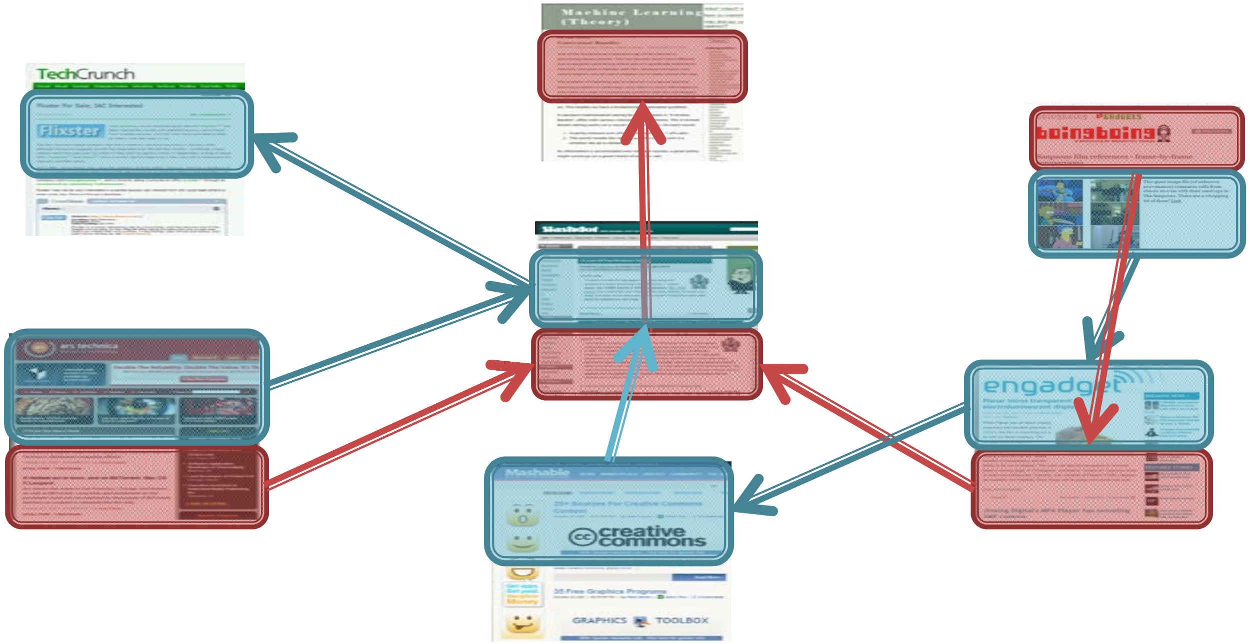
News Agencies, Personal Blogs (Blog), Newspapers, Professional Blogs, TV

- Politics is dominated by traditional media
- Blogs:
  - Influential for Entertainment phrases
  - Influence lasts longer than for other media types

[KDD '10, NIPS '10]

# Inferring the Diffusion Network

- But how does information **really** spread?




- We only see time of mention but not the edges
- Can we reconstruct (hidden) **diffusion network**?

[KDD '10, NIPS '10]

# Examples and Applications

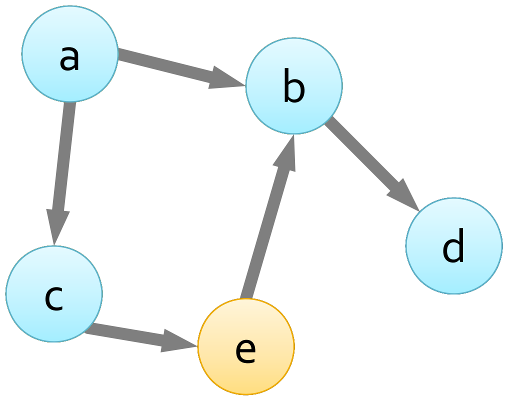
| | Virus propagation | Word of mouth & Viral marketing |
|---|---|---|
| **Process** | Viruses propagate through the network | Recommendations and influence propagate |
| **We observe** | We only observe when people get sick | We only observe when people buy products |
| **It's hidden** | But NOT who **infected** whom | But NOT who **influenced** whom |

Can we infer the underlying network?

[KDD '10, NIPS '10]

# Inferring Diffusion Networks

- There is a hidden diffusion network:



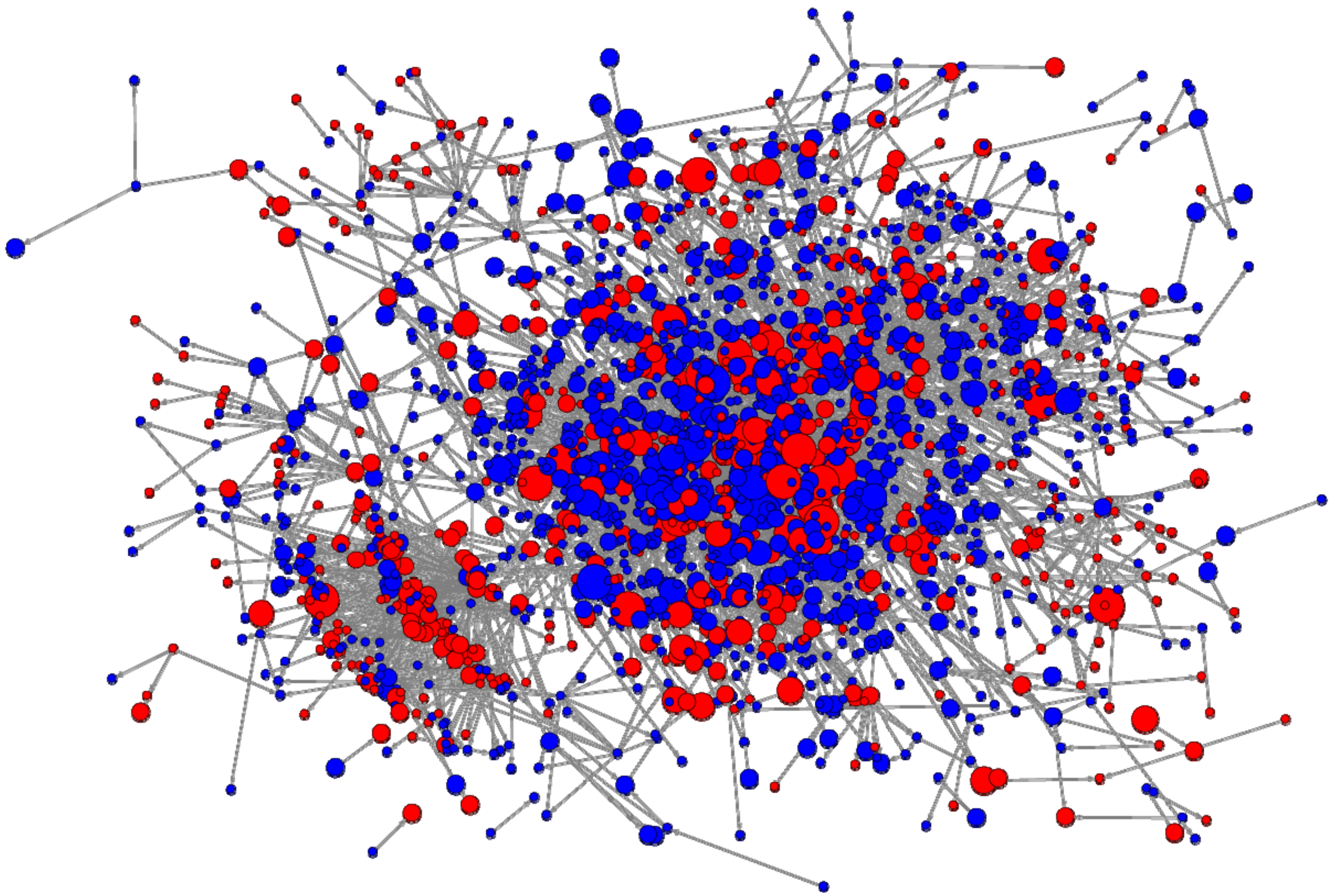

- We only see times when nodes get "infected":
  - $c_1$: (a,1), (c,2), (b,3), (e,4)
  - $c_2$: (c,1), (a,4), (b,5), (d,6)
- **Want to infer who-infects-whom network!**
  - The problem is NP-hard (there are $O(2^{N^2})$ graphs!)
  - Our algorithm can do it near-optimally in $O(N^2)$

# Experiments

- Propagation of quotes:
  - 172 million news and blog articles over 1 year
    Extract 343 million different quotes
  - Record times $t_i(w)$ when site $w$ mentioned quote $i$
- Given the times when sites mention quotes infer the network of information diffusion:
  - Who tends to copy (repeat after) whom

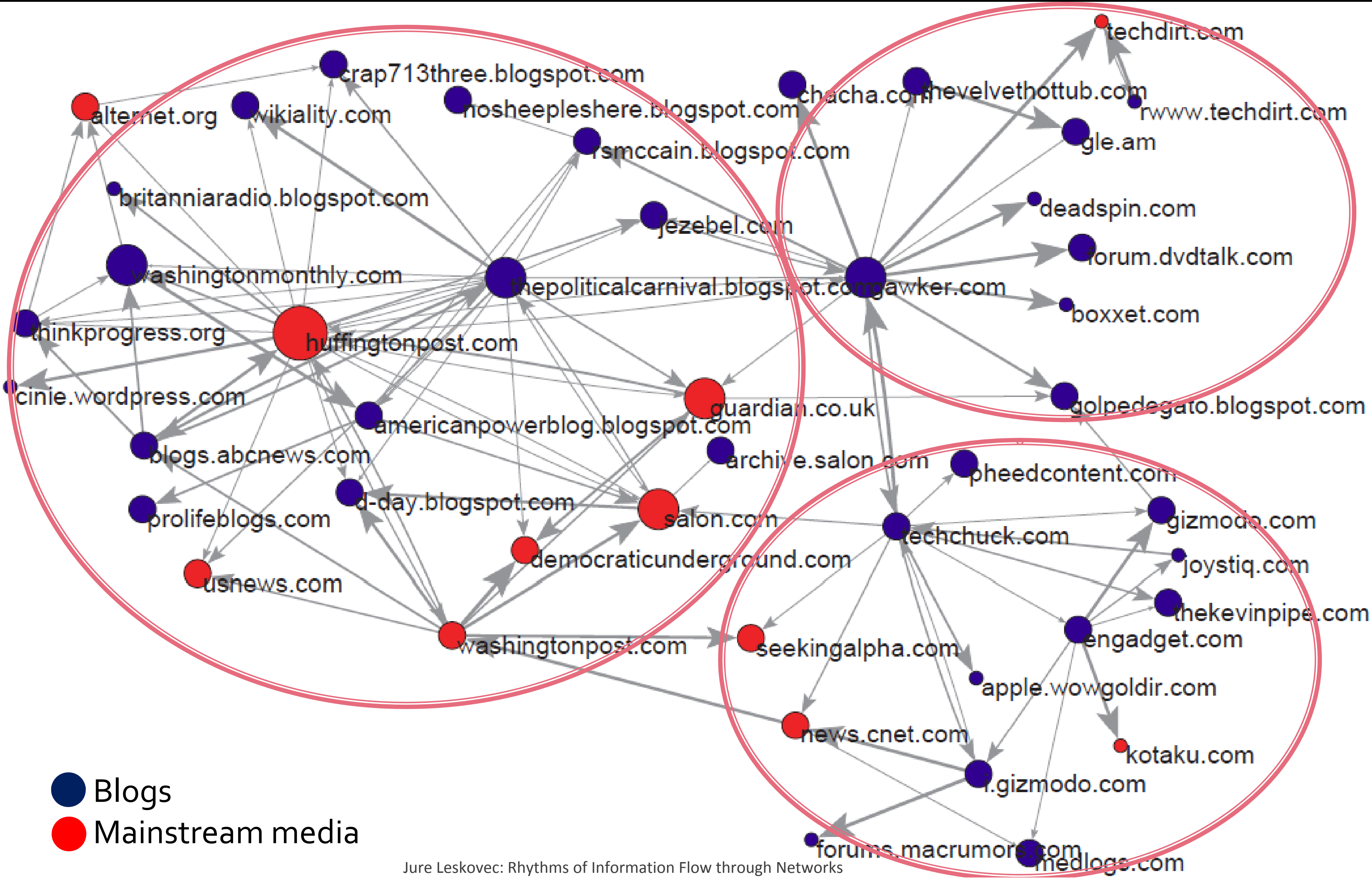
# Diffusion Network

- 5,000 news sites:



Blogs
Mainstream media

# Diffusion Network (zoom-in)



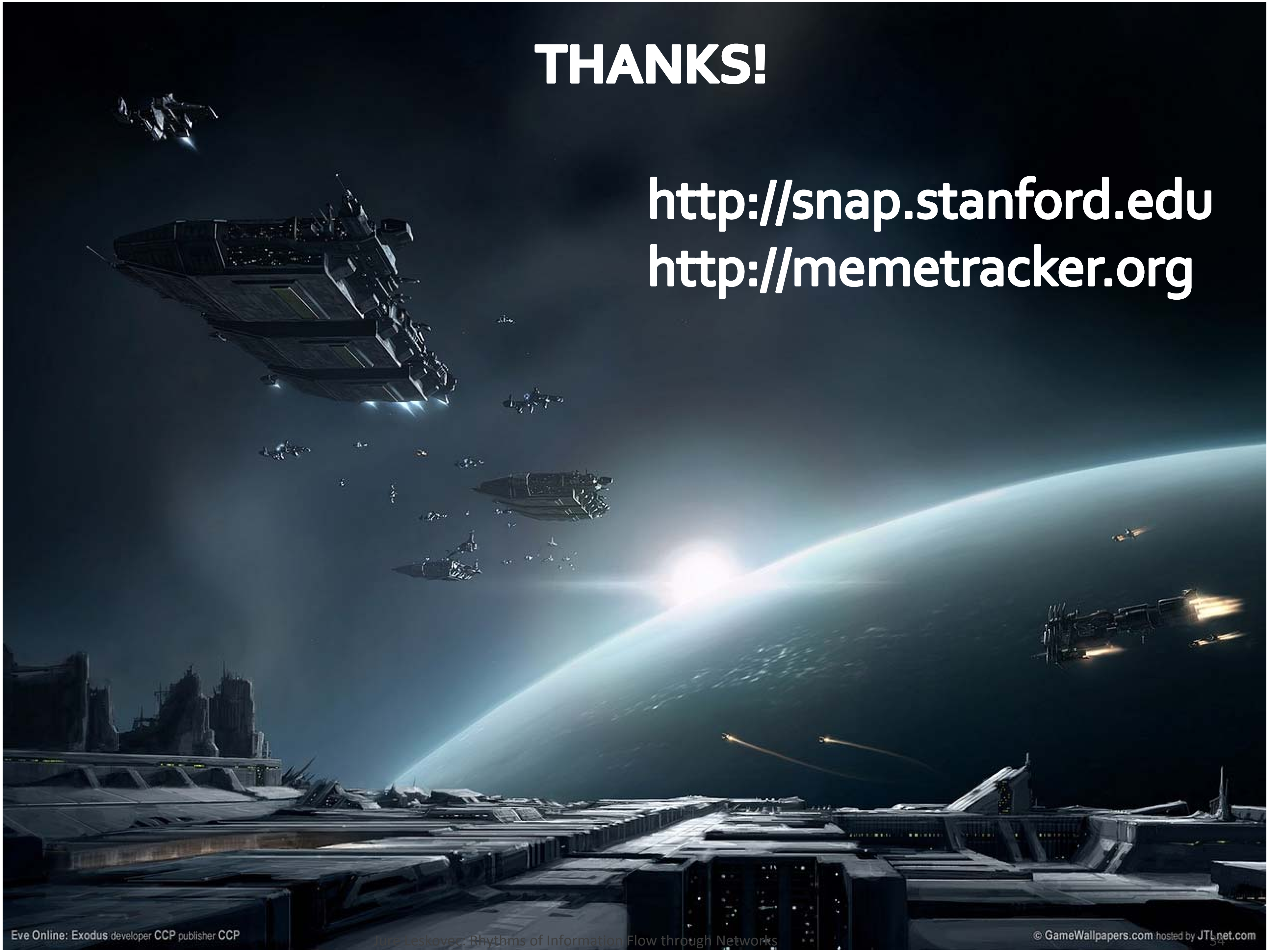
# Conclusions and Connections

- Messages arriving through networks from real-time sources requires new ways of thinking about information dynamics and consumption:
  - Tracking information through (implicit) networks
  - Quantify the dynamics of online media
  - Predict the diffusion of information
  - And infer networks of information diffusion

# Further Qs: Opinion dynamics

- Can this analysis help identify dynamics of polarization [Adamic-Glance '04]?
- Connections to mutation of information:
  - How does attitude and sentiment change in different parts of the network?
  - How does information change in different parts of the network?

# THANKS!


http://snap.stanford.edu
http://memetracker.org
Eve Online: Exodus developer CCP publisher CCP
© GameWallpapers.com hosted by JTLnet.com

# References


- *Meme-tracking and the Dynamics of the News Cycle*, by J. Leskovec, L. Backstrom, J. Kleinberg. KDD, 2009. http://cs.stanford.edu/~jure/pubs/quotes-kdd09.pdf
- *Patterns of Temporal Variation in Online Media* by J. Yang, J. Leskovec. ACM International Conference on Web Search and Data Minig (WSDM), 2011. http://cs.stanford.edu/people/jure/pubs/memeshapes-wsdm11.pdf
- *Modeling Information Diffusion in Implicit Networks* by J. Yang, J. Leskovec. IEEE International Conference On Data Mining (ICDM), 2010. http://cs.stanford.edu/people/jure/pubs/lim-icdm10.pdf
- *Inferring Networks of Diffusion and Influence* by M. Gomez-Rodriguez, J. Leskovec, A. Krause. KDD, 2010. http://cs.stanford.edu/~jure/pubs/netinf-kdd2010.pdf
- *On the Convexity of Latent Social Network Inference* by S. A. Myers, J. Leskovec. Neural Information Processing Systems (NIPS), 2010. http://cs.stanford.edu/people/jure/pubs/connie-nips10.pdf
- *Cost-effective Outbreak Detection in Networks* by J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance. KDD 2007. http://cs.stanford.edu/~jure/pubs/detect-kdd07.pdf
- *Cascading Behavior in Large Blog Graphs* by J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst. SDM, 2007. http://cs.stanford.edu/~jure/pubs/blogs-sdm07.pdf