Meme-tracking & Predicting Information Flow in Networks

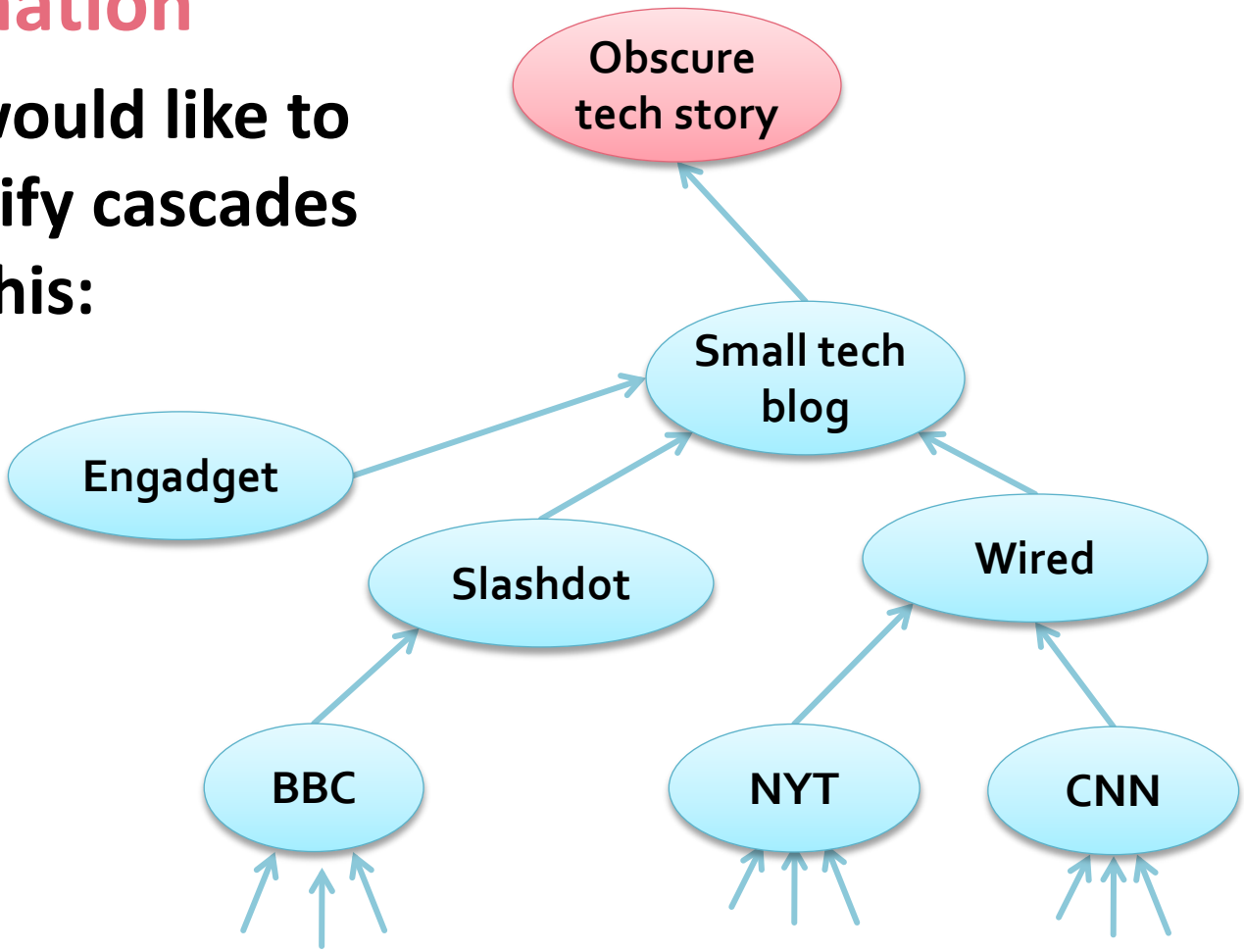
CS224W: Social and Information Network Analysis

Jure Leskovec, Stanford University

http://cs224w.stanford.edu

Tracking the Information Flow

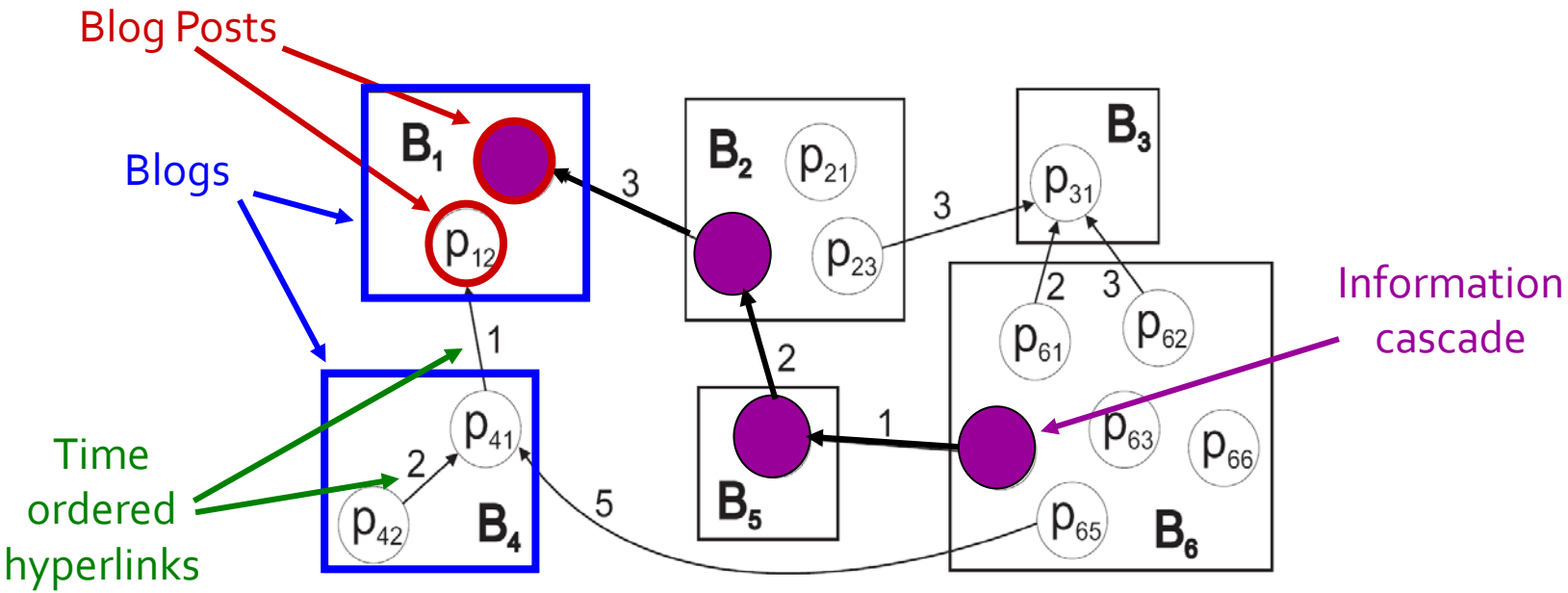
- Imagine you want to track the flow of information
  - We would like to identify cascades like this:



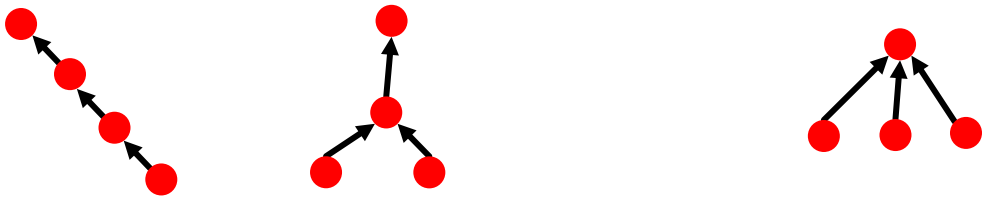
[SDM '07]

Approach 1: Trace Hyperlinks

- Tracking Hyperlinks on the Blogosphere

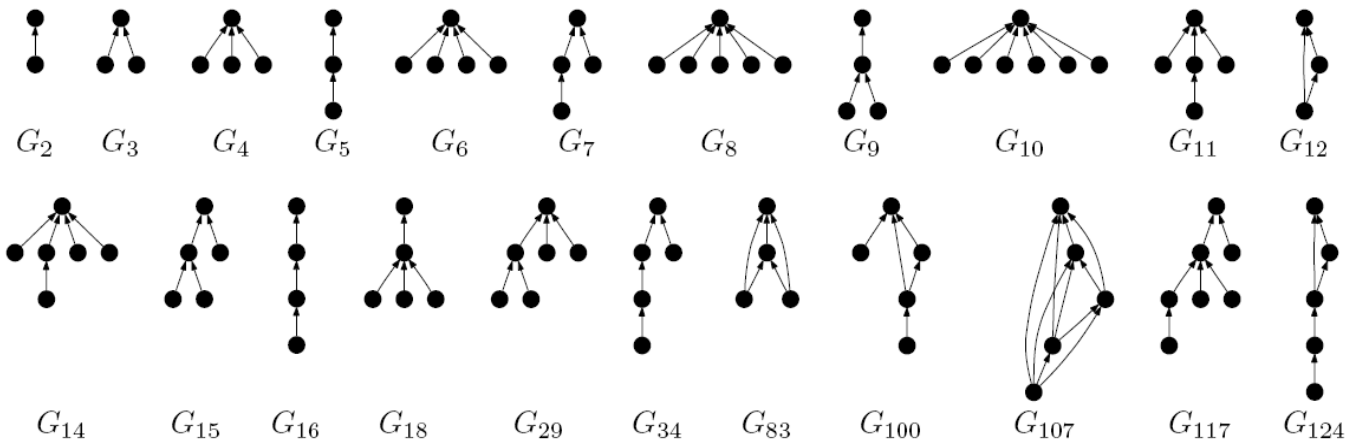


- **Identify cascades** – graphs induced by a time ordered propagation of information



[SDM '07]

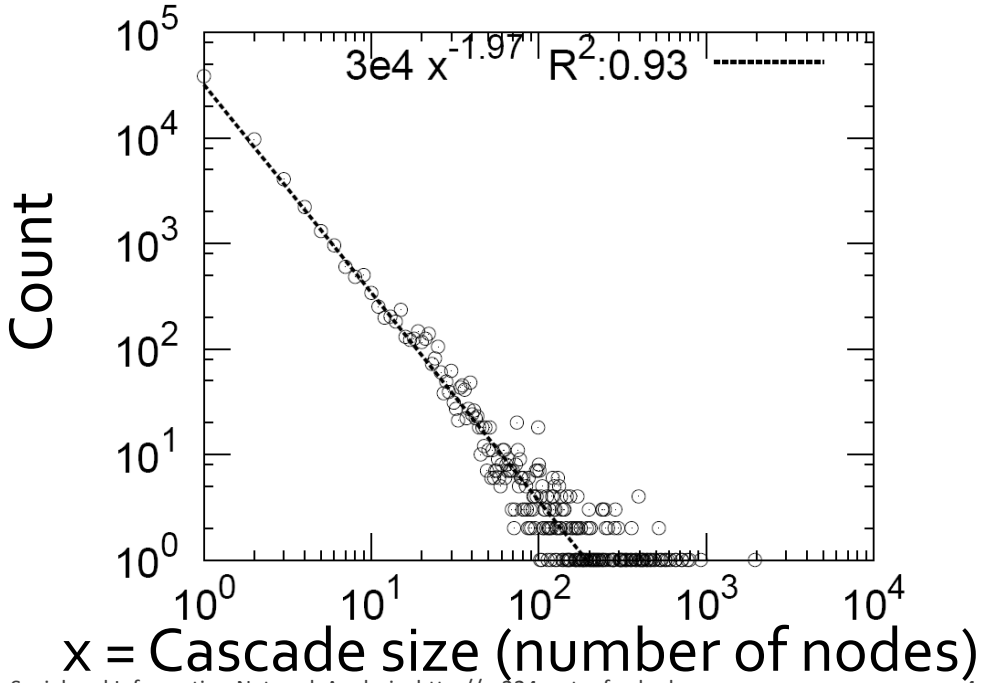
Cascade Shapes



Cascade shapes (ranked by frequency)

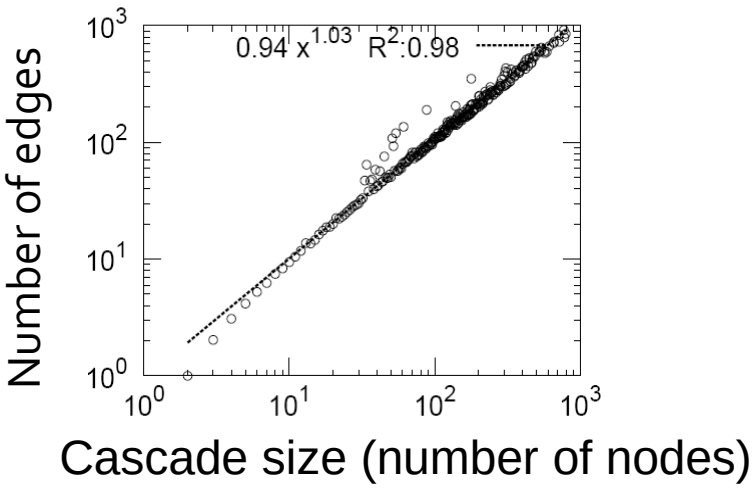
The probability of observing a cascade on n nodes follows:

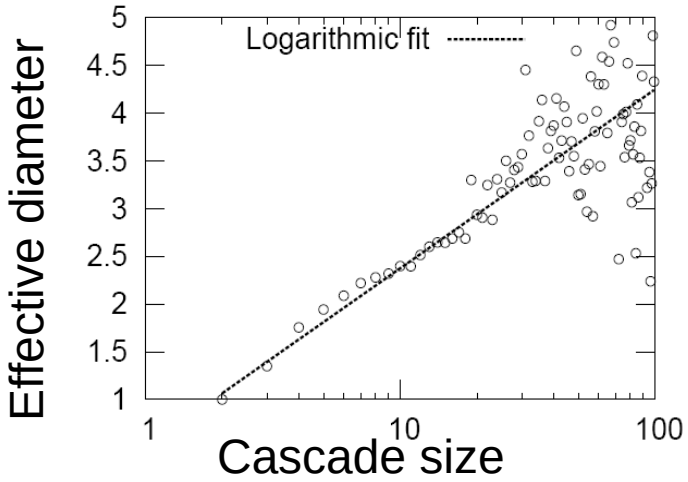
$$p(n) \sim n^{-2}$$



[SDM '07]

Properties of Blog Cascades

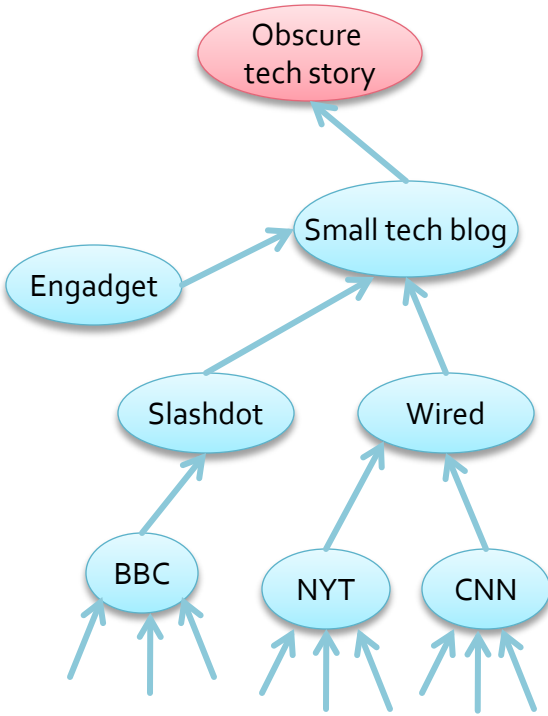




- **Most of cascades are trees:**
  - Number of edges is smaller than the number of nodes in a cascade
  - Diameter increases logarithmically

Tracing hyperlinks: Pros/Cons

- **Advantages:**
  - Unambiguous, precise and explicit way to trace information flow
  - We obtain both the times as well as the trace (graph) of information flow
- **Caveats:**
  - Not all links transmit information:
    - Navigational links, templates, adds
  - Many links are missing:
    - Mainstream media sites do not create links
    - Bloggers "forget" to link the source
      - (We will later see how to identify networks/cascades just based on what times sites mentioned information)



[KDD '09]

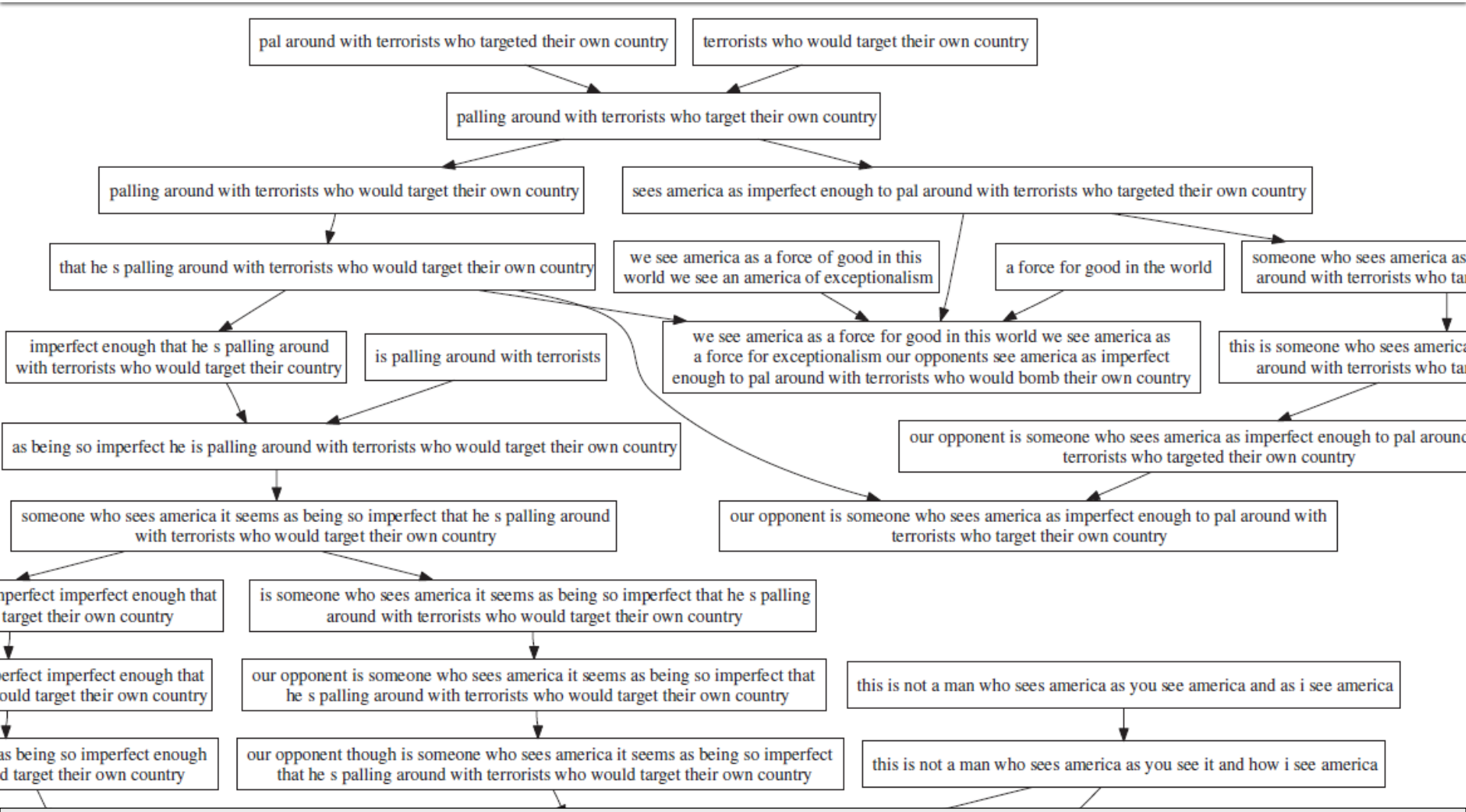
Meme-Tracking

- **Extract textual fragments that travel relatively unchanged, through many articles:**
  - **Look for phrases inside quotes: "…"**
    - About 1.25 quotes per document in our data
  - **Why it works?**

    Quotes…
    - are integral parts of journalistic practices
    - tend to follow iterations of a story as it evolves
    - are attributed to individuals and have time and location

[KDD '09]

Challenge: Quotes Mutate

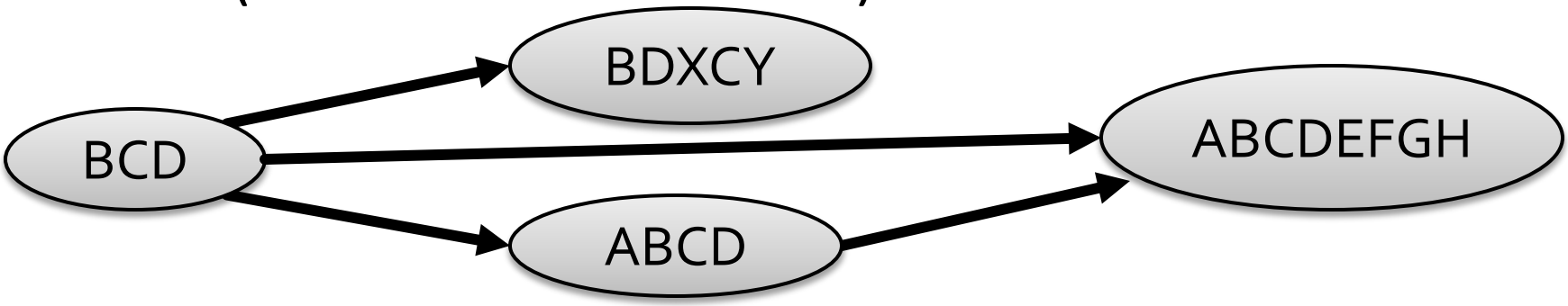


Quote: Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that he's palling around with terrorists who would target their own country.

[KDD '09]

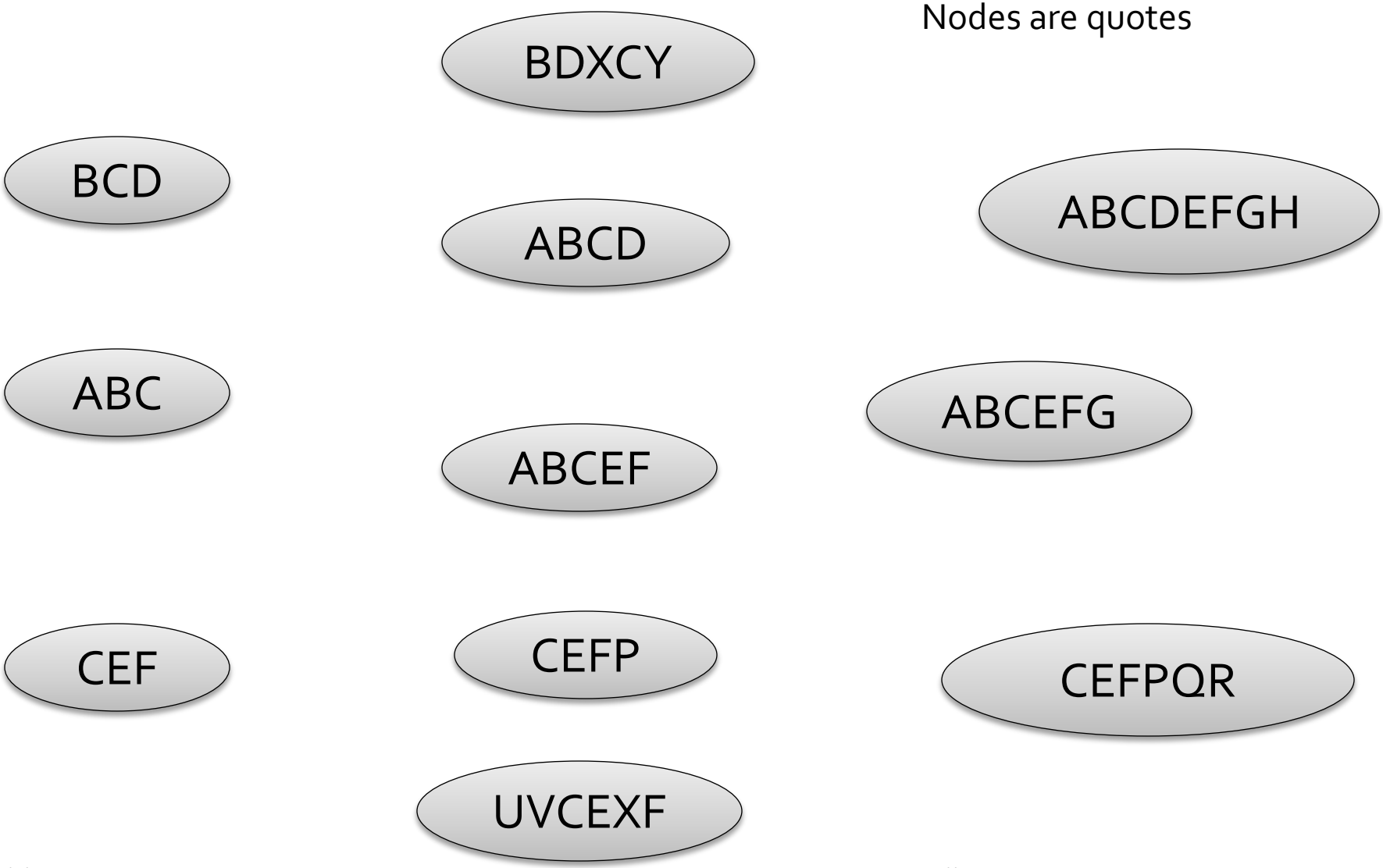
Finding Mutational Variants

- **Goal:** Find mutational variants of a quote
- Form **approximate quote inclusion graph**
  - Shorter quote is approximate substring of a longer one (word edit distance = 1)

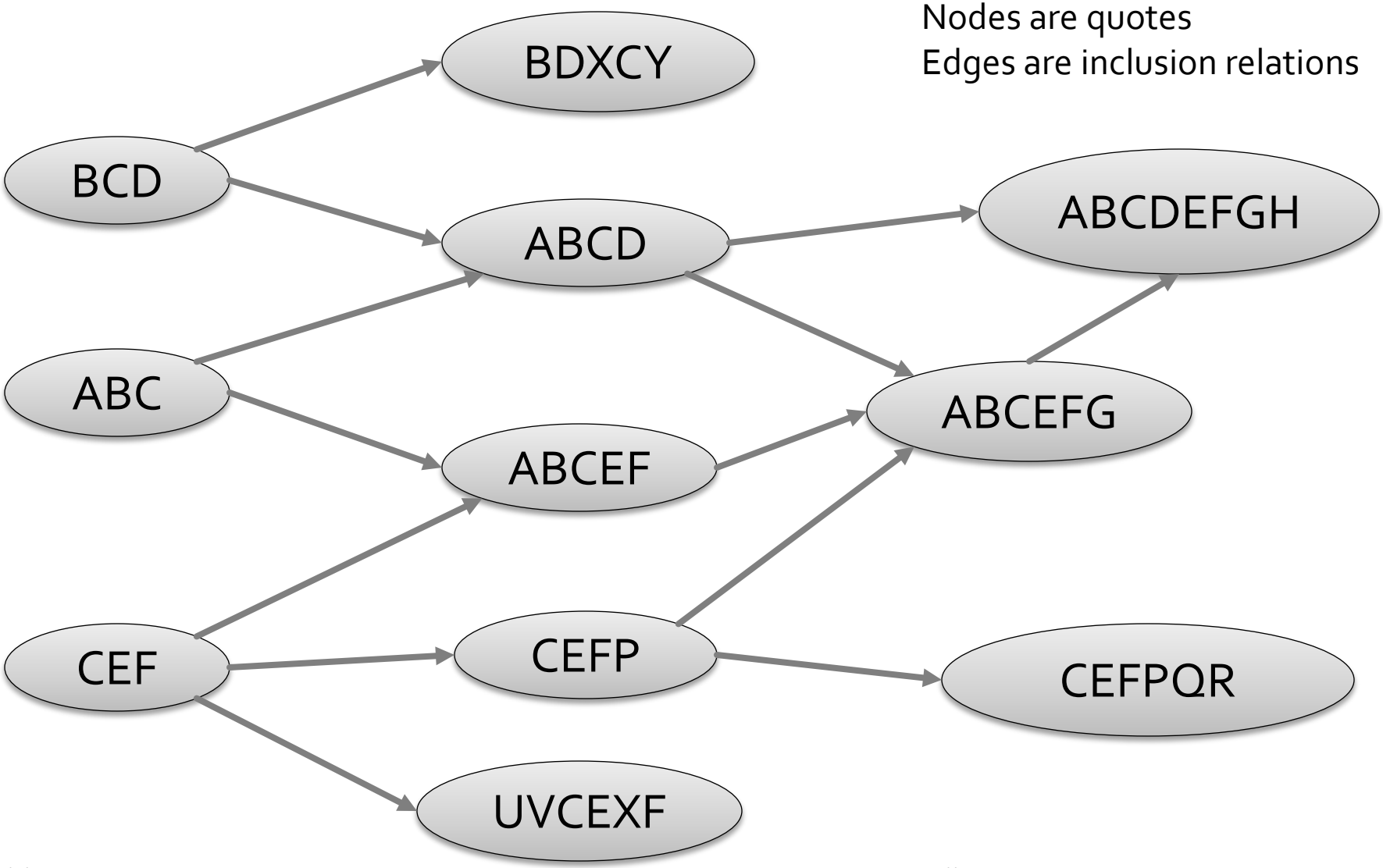


- **Objective:** In DAG of approx. quote inclusion, delete min total edge weight s.t. each connected component has a single "**sink**"

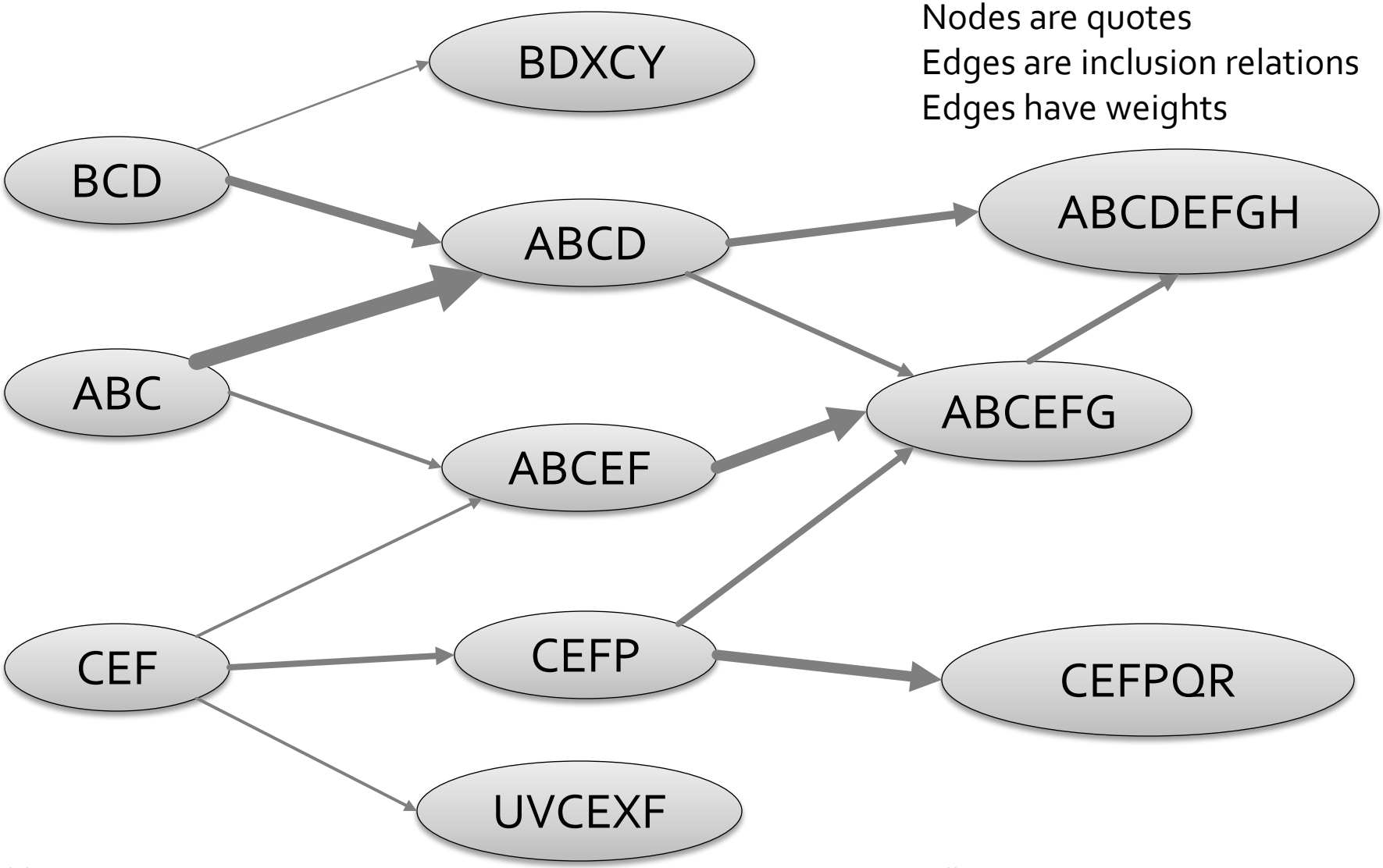
Creating Clusters of Mutations



Creating clusters of Mutations

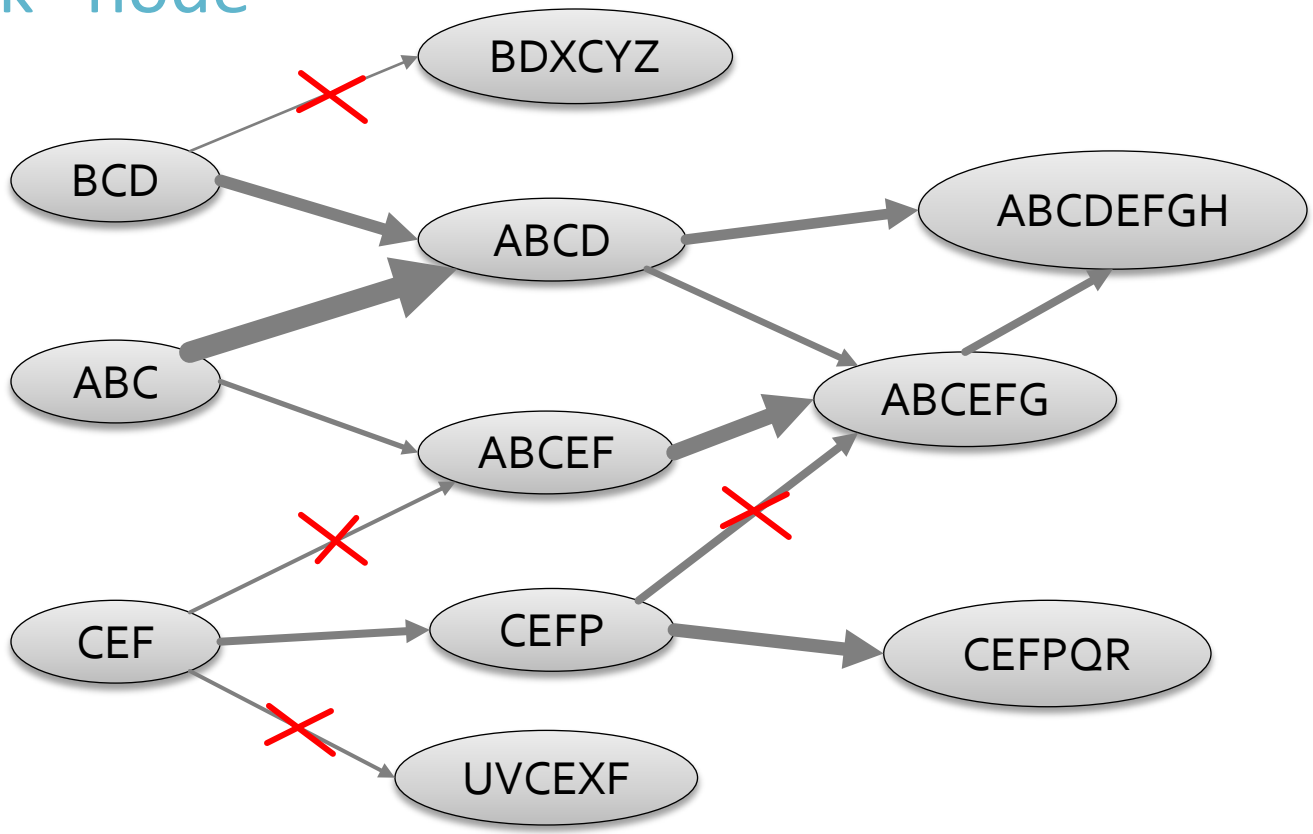


Creating clusters of Mutations



Quote Clustering: DAG Partitioning

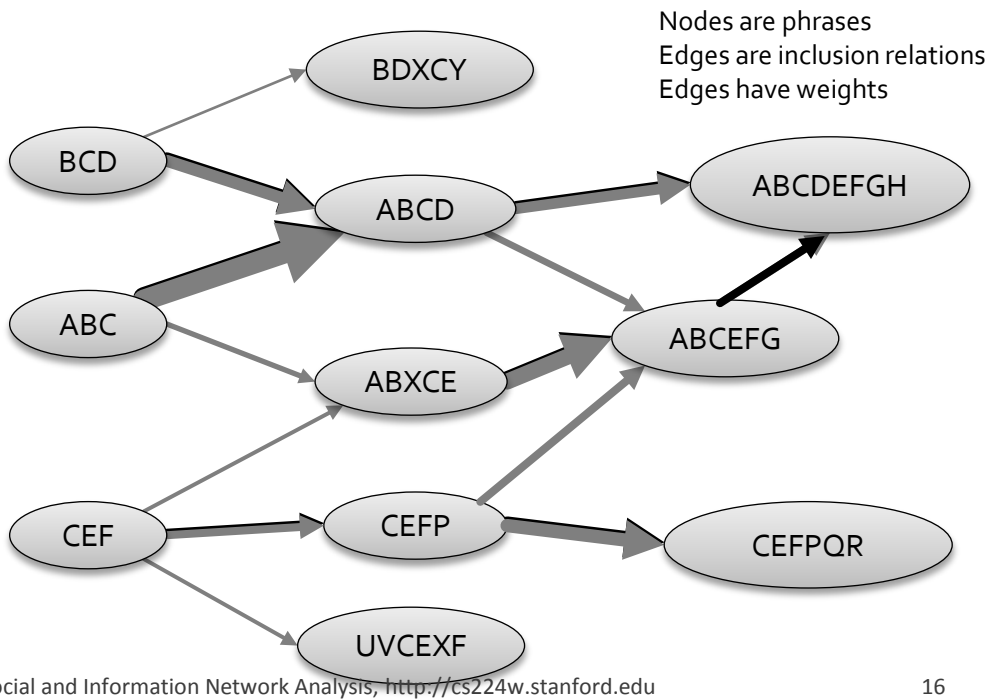
- **Objective:** In a directed acyclic graph (approx. quote inclusion), delete min total edge weight s.t. each connected component has a single "sink" node



[KDD '09]

DAG Partitioning Heuristic

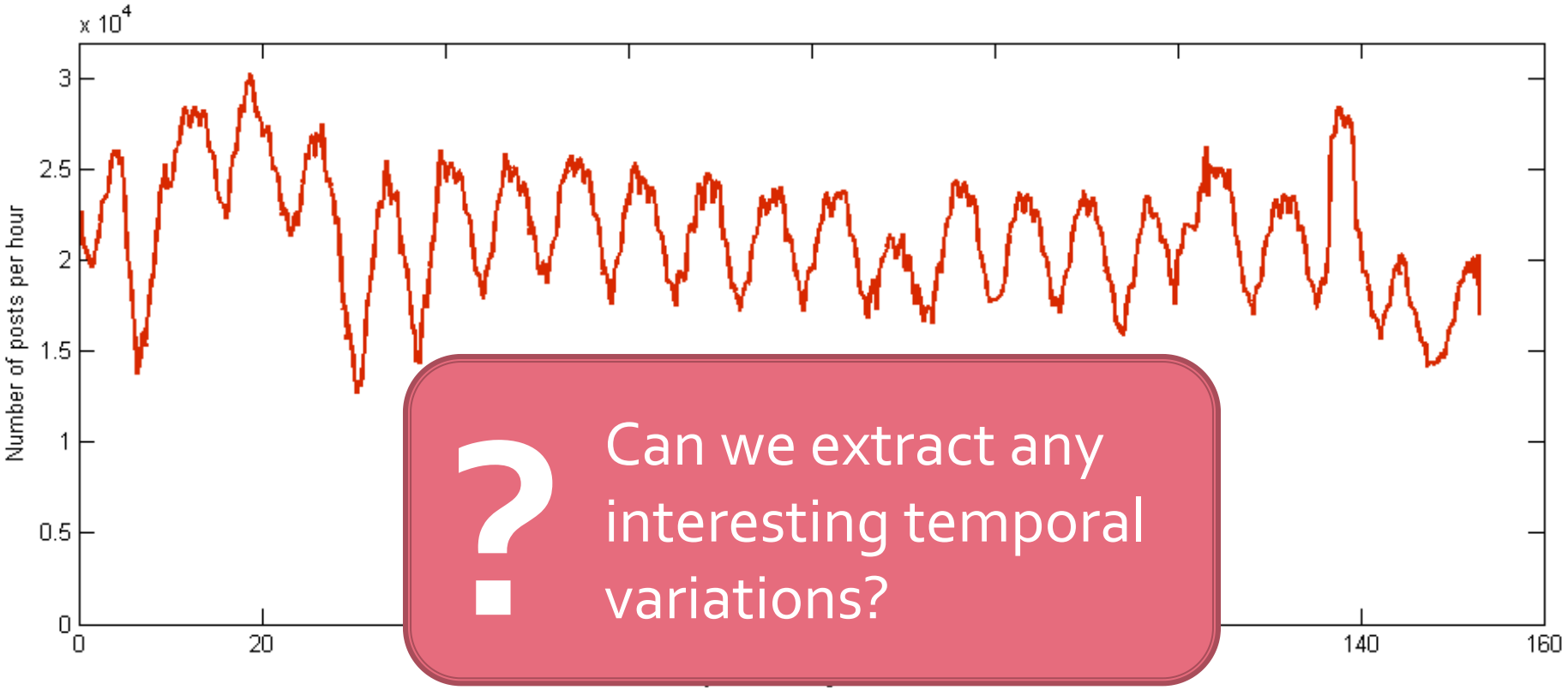
- **DAG-partitioning is NP-hard but heuristics are effective:**
  - **Observation:** Enough to know node's parent to reconstruct optimal solution
  - **Heuristic:**
    Proceed right-to-left and assign a node (keep a single edge) to the strongest cluster



A Quote Cluster

Quoted text	Volume
the fundamentals of our economy are strong	3654
the fundamentals of the economy are strong	988
fundamentals of our economy are strong	645
fundamentals of the economy are strong	557
if john mccain hadn't said that the fundamentals of our economy are strong on the day of one of our nation's worst financial crises the claim that he invented the blackberry would have been the most preposterous thing said all week	224
fundamentals of the economy	172
the fundamentals of the economy are sound	119
i promise you we will never put america in this position again we will clean up wall street	83
the fundamentals of our economy are sound	81
clean up wall street	78
our economy i think still the fundamentals of our economy are strong	75
fundamentals of the economy are sound	72
the fundamentals of our economy are strong but these are very very difficult times and i promise you we will never put america in this position again	68
the economy is in crisis	66
these are very very difficult times	63
the fundamentals of our economy are strong but these are very very difficult times	62
do you still think the fundamentals of our economy are strong genius	62
our economy i think still the fundamentals of our economy are strong but these are very very difficult times	60
mccain's first response to this crisis was to say that the fundamentals of our economy are strong then he admitted it was a crisis and then he proposed a commission which is just washington-speak for i'll get back to you later	55
i still believe the fundamentals of our economy are strong	53
i think still the fundamentals of our economy are strong	50
cut taxes for 95 percent of all working families	50

Quotes Over Time

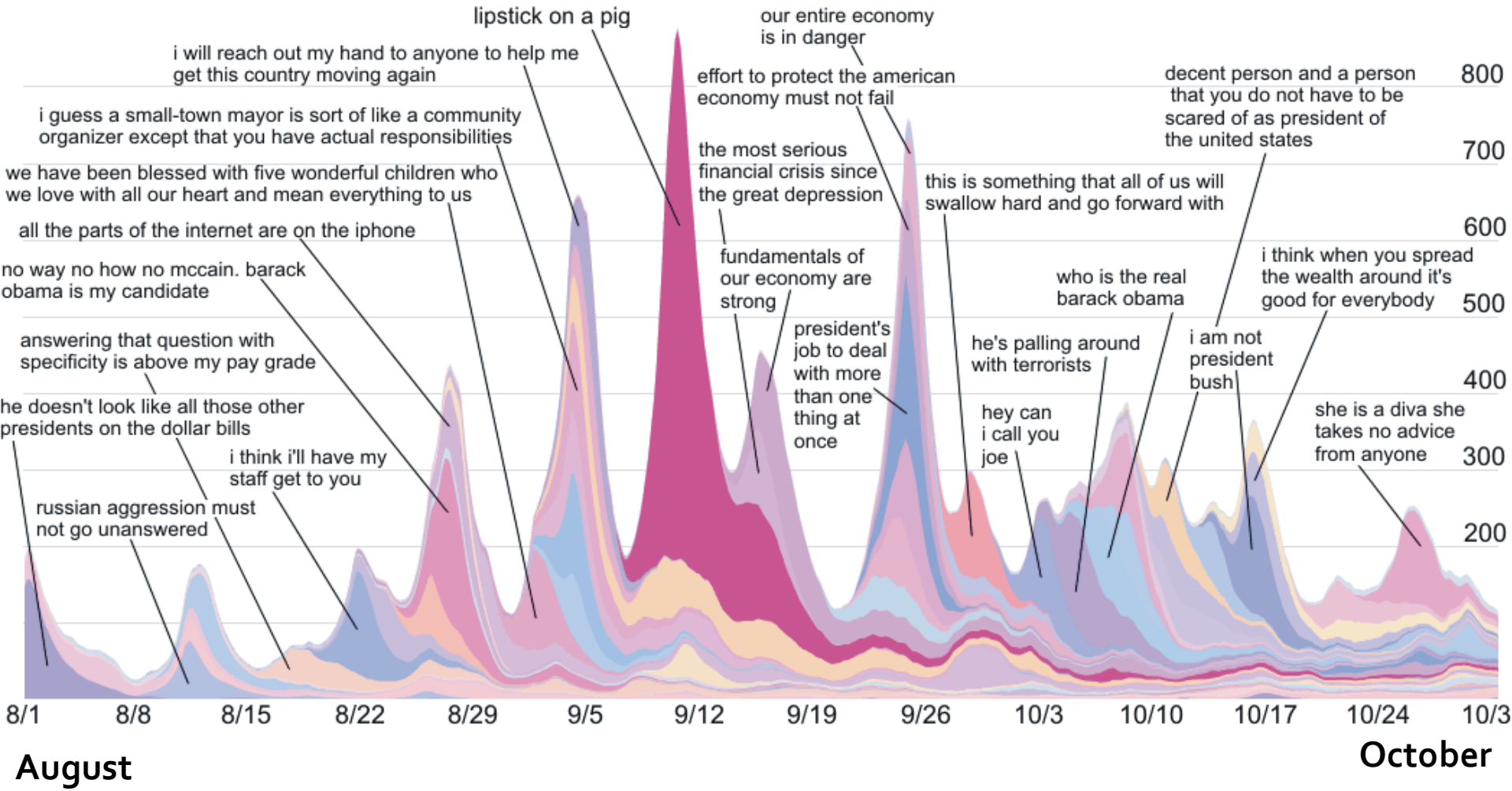


? Can we extract any interesting temporal variations?

… is periodic, has no trends.

"Bandwidth" of the online media is constant

Cluster Volume Over Time



Volume over time of top 50 largest total volume quote clusters

Quotes on the "Great Depression"

- **Media coverage of the current economic crisis**
- **Main proponents of the debate:**

Most Cited Phrases about the Economy

Feb. 1 - July 3, 2009

Phrase	Original Speaker	Starting Date	Total Citations
we will rebuild, we will recover...	Barack Obama	24-Feb	4679
how do they justify this outrage to the taxpayers...	Barack Obama	16-Mar	4446
in ... our greatest economic crisis since the Great Depression...	Barack Obama	7-Feb	3914
they'll have to find someone else to write the next stimulus bill	NY Post	18-Feb	3312
...the weight of this crisis will not determine the destiny of this nation	Barack Obama	24-Feb	3113
...to be honest I'm a little bit worried	Chinese Premier	13-Mar	3017
buying stocks is a potentially good deal	Barack Obama	3-Mar	2690
...we would not be able to continue as a going concern...	General Motors	5-Mar	2672
we've seen some progress in the financial markets, absolutely	Ben Bernanke	15-Mar	2425

Speech in congress

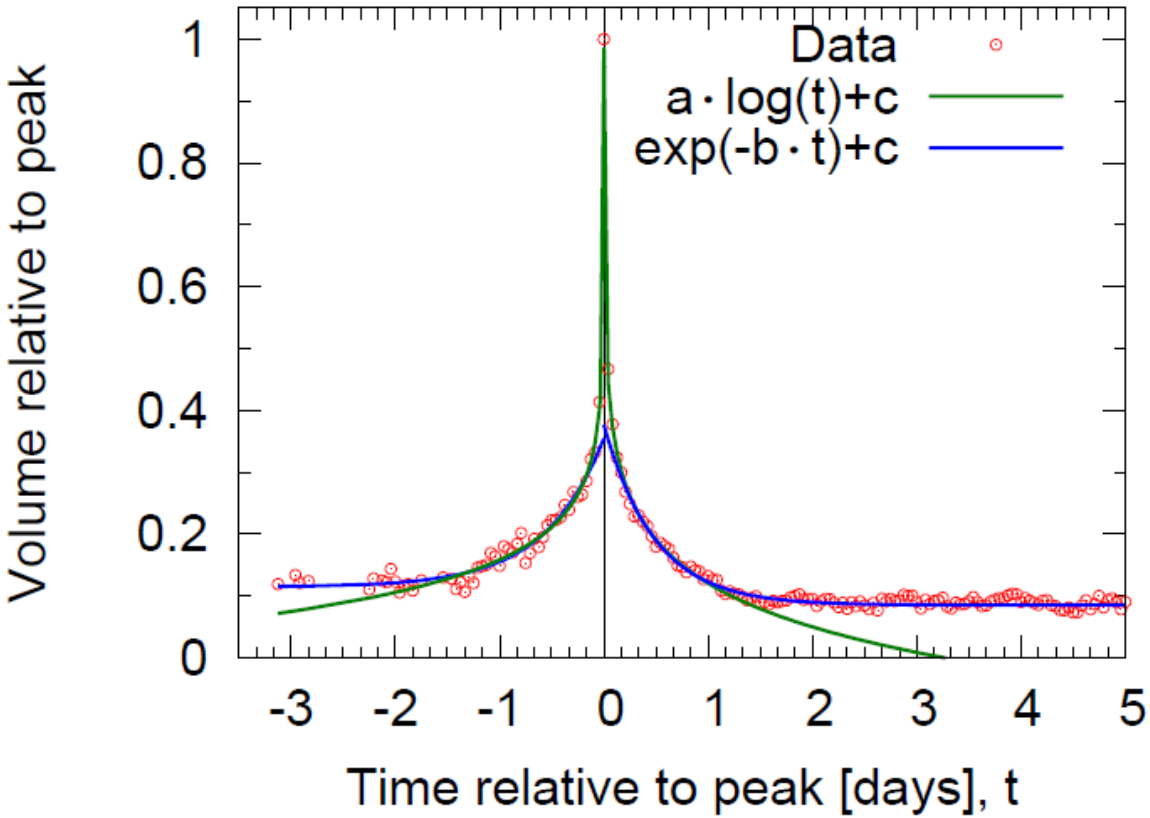
Dept. of Labor release

60-minutes interview

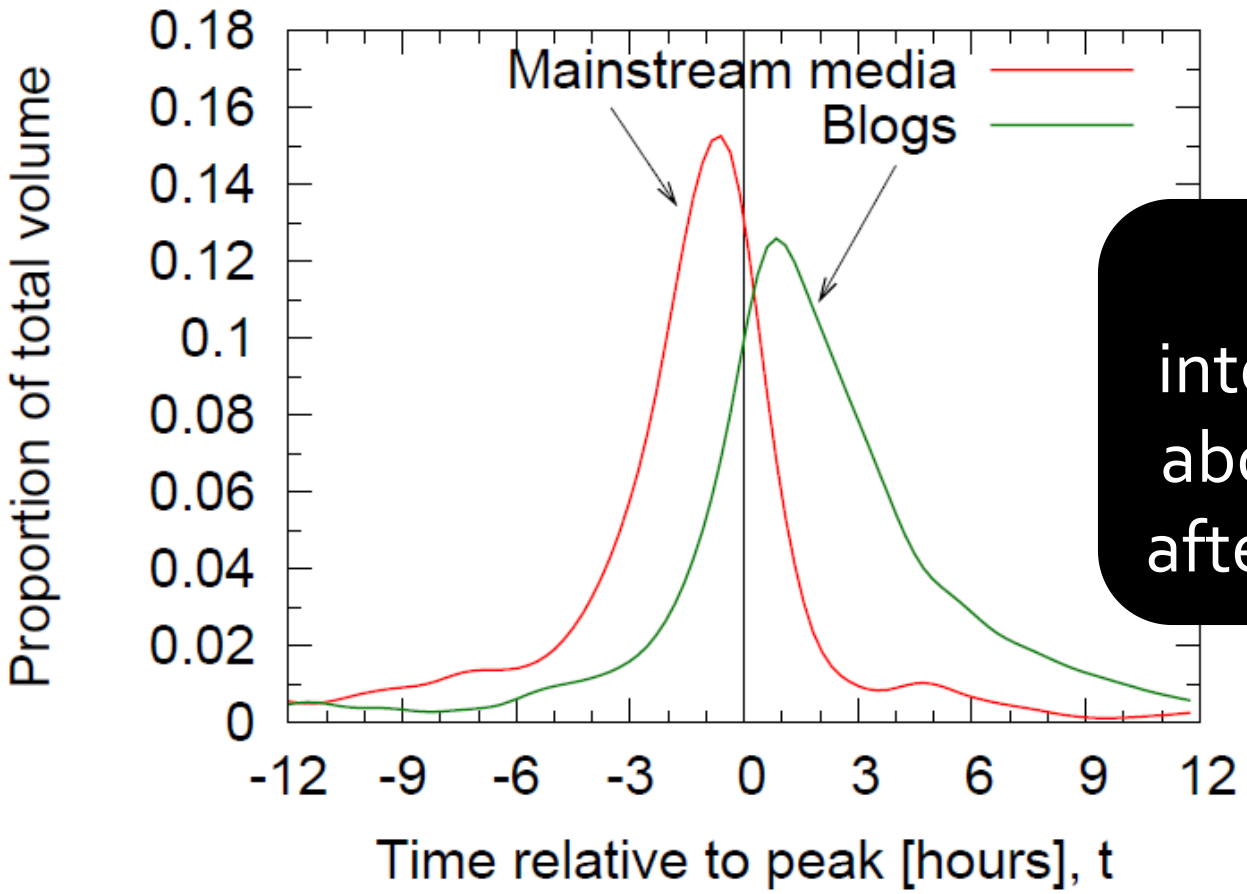
Top republican voice ranks only 14th

Interaction of News and Blogs

- Can study typical quote cluster volume curve
- Phrases are very short lived:



Interaction of News and Blogs



Peak blog intensity comes about 2.5 hours after news peak.

- **Using Google News we label:**
  - Mainstream media: 20,000 sites (44% vol.)
  - Blog (everything else): 1.6 million sites (56% vol.)

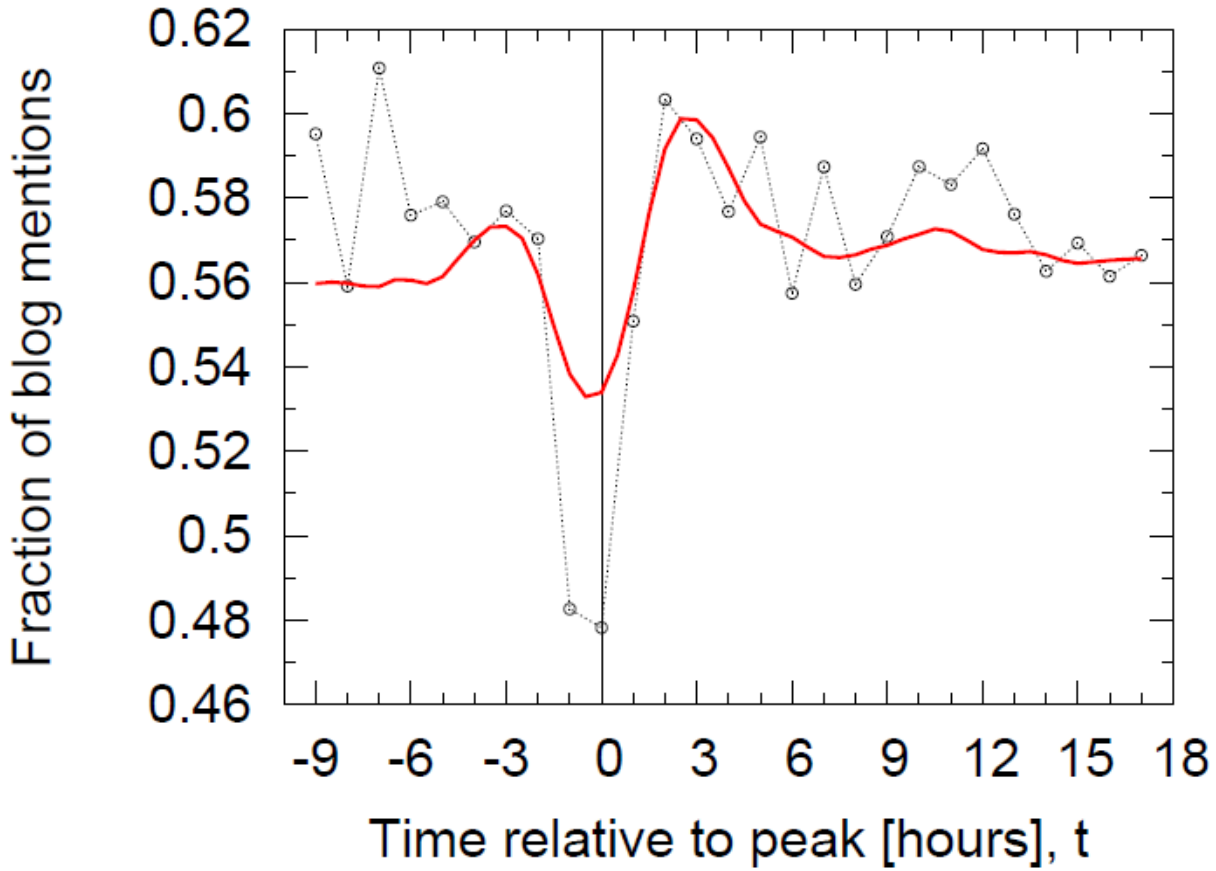
How quickly sites mention quotes?

- **Classify individual sources by their typical timing relative to the peak aggregate intensity**

	Rank	Lag [h]	Reported	Site
Professional blogs	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
News media	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

Interaction of News and Blogs

- The "oscillation" of attention between mainstream media and blogs



Stories catalyzed by blogs

- **Queries for different temporal "signatures":**

  e.g., stories catalyzed by blogs:

  [*x*; *y*; *t*]-query: between *x* and *y* frac. of total quote volume (f_b) occurred on blogs at least *t* days before overall the peak

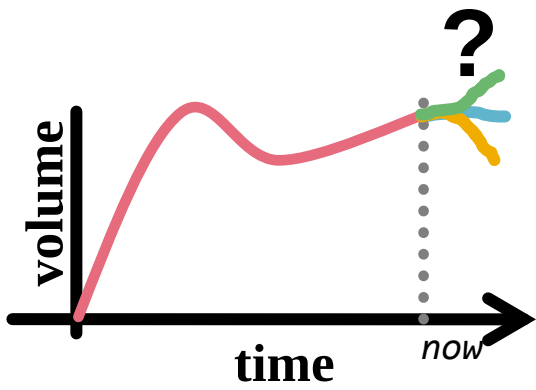
M	f_b	Phrase
2,141	.30	Well uh you know I think that whether you're looking at it from a theological perspective or uh a scientific perspective uh answering that question with specificity uh you know is uh above my pay grade.
826	.18	A changing environment will affect Alaska more than any other state because of our location I'm not one though who would attribute it to being man-made.

In total 3.5% of phrases migrate from blogs to media

Predicting Information Volume

Predicting Information Attention

- **How much attention will information get?**
  - How many sites mention information at particular time?
- **Idea:** Predict the future number of mentions based on who got "infected" in the past
- **Linear Influence Model (LIM)**
  - Assume no network
  - Model the global influence of each node
  - Predict future volume from node influences



Predicting Information 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 sites will mention the info at time 4, 5,…?
- **Motivating question:**
  - If NYT mentions info at time *t*
  - How many additional mentions does this "generate" (on other sites) at time *t+1*, *t+2*, …?

LIM: Strategy

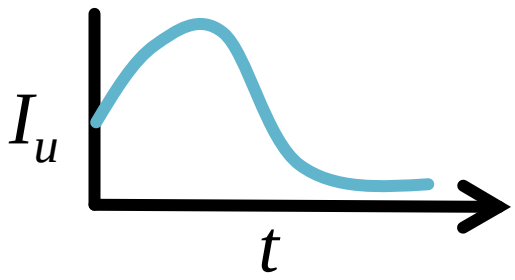
t	A(t)	V(t)
1	u, w	2
2	u, w, v, x, y	3
3		?

- **K=1 piece of information:**
  - $V(t)$…volume (number of new infections at time t)
  - $A(t)$…set of already infected nodes by time t
- How does **LIM** predict the future number of infections $V(t+1)$?
  - Each node u has an **influence function:**
    - After node u gets infected, how many other nodes tend to get infected
    - 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

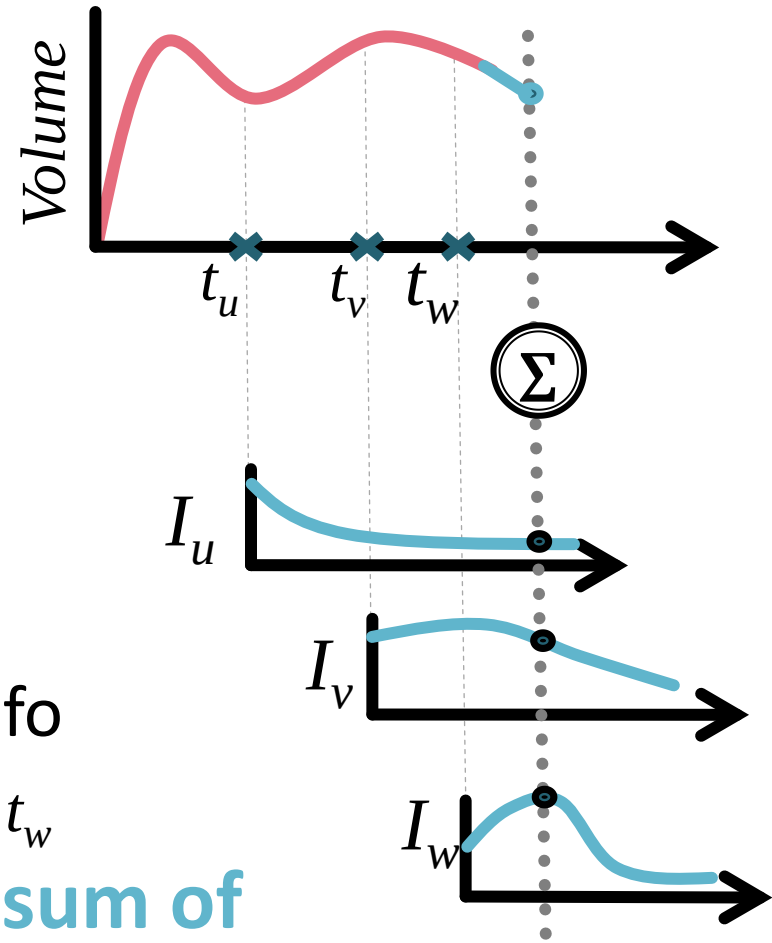
- Each node *u* has an "**influence" function** $I_u(t)$:
  - $I_u(t)$: After node *u* gets mentions, how many other nodes tend to mention *t* hours later
    - e.g.: Influence function of NYT: How many sites say the info after NYT says it?



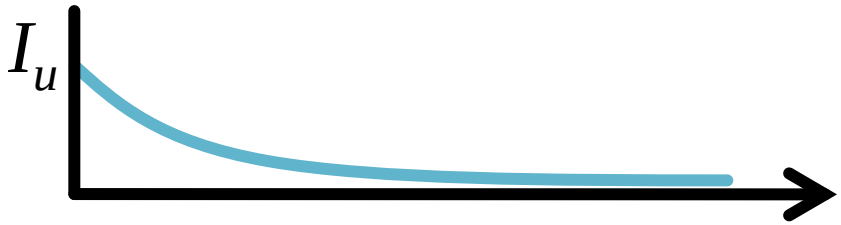
- **How to predict future volume *V(t+1)*?**
  - Predict future volume using the influence functions of nodes infected in the past

The Linear Influence Model

- **LIM model:**
  - Volume V*(t)* at time *t*
  - *A(t)* … a set of nodes that mentioned info before time *t*
- **And let:**
  - $I_u(t)$: influence function of *u*
  - t_u: time when *u* mentioned info
    - *u*, *v*, *w* mentioned at times t_u, t_v, t_w
- **Predict future volume as a sum of influences:** $V(t+1) = \sum_{u \in A(t)} I_u(t - t_u)$



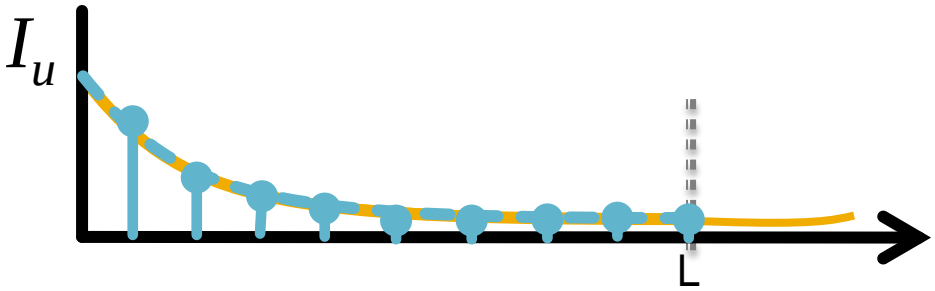
Node Influence Function



- After node *u* is infected, it will infect $I_u(t)$ other nodes over time
- **Influence function $I_u(t)$ of node *u*:**
  - Number of infections caused by *u* *t*-time steps after it gets infected
  - $I_u(t)$ is unobserved, need to estimate it
- **Influence function $I_{CNN}(t)$ of CNN**
  - How many people mention the information over time after they see it on CNN?

[ICDM '10]

Estimating Influence Functions



- $I_u(t)$ is **not observable**, need to estimate it
- **Discrete non-parametric influence functions:**
  - Discrete time units
  - $I_u(t)$ … non-negative vector of length L

$$I_u(t) = [I_u(1), I_u(2), I_u(3), \ldots, I_u(L)]$$

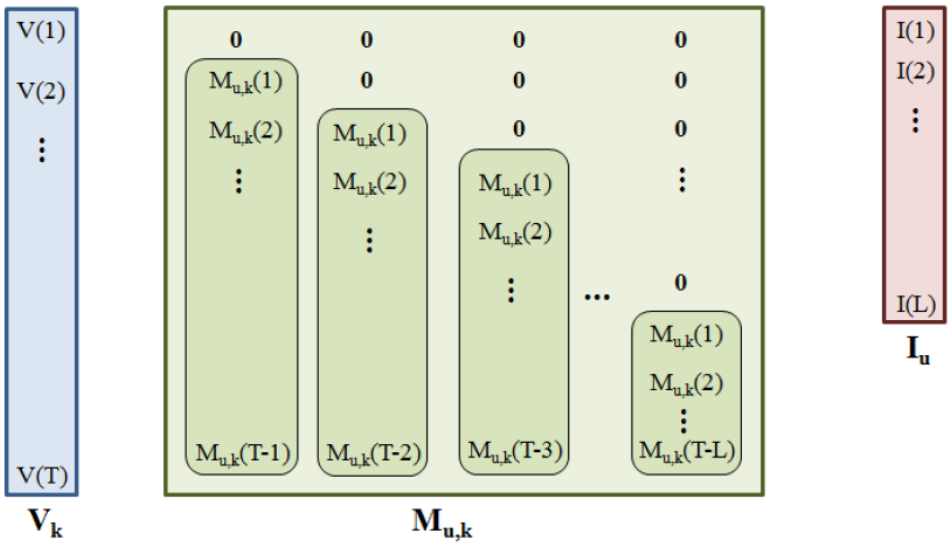
- **Find $I_u(t)$ by solving a optimization problem:**

$$\min_{I_u, \forall u} \sum_k \sum_t \left(V_k(t+1) - \sum_{u \in A_k(t)} I_u(t - t_u) \right)^2$$

$V_k(t)$… volume of k-th info
$A_k(t)$… infected set with k-th info

LIM as Matrix Equation

- **Input data:** 1 contagion, 1 node
- Write **LIM** as a matrix equation:



- **Volume vector:**

  $V_k(t)$ … volume of contagion *k* at time *t*
- **Infection indicator matrix:**

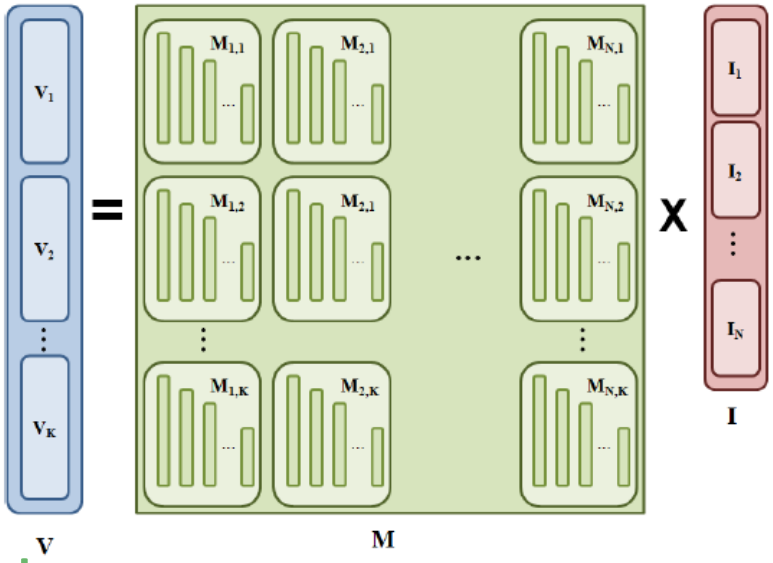
  $M_{u,k}(t) = 1$ if node *u* gets infected by contagion *k* at time *t*
- **Influence functions:**

  $I_u(t)$ … influence of node *u* on diffusion

[ICDM '10]

LIM as Matrix Equation

- **Input data:** K contagions, N nodes
- Write **LIM** as a matrix equation:



- **Volume vector:**
  $V_k(t)$ … volume of contagion *k* at time *t*
- **Infection indicator matrix:**
  $M_{u,k}(t) = 1$ if node *u* gets infected by contagion *k* at time *t*
- **Influence functions:**
  $I_u(t)$ … influence of node *u* on diffusion

[ICDM '10]

Estimating Influence Functions

- **LIM** as a matrix equation: **V = M * I**
- **Estimate influence functions:**

$$\hat{\mathbf{I}} = \arg\min_{\mathbf{I} \geq 0} ||\mathbf{V} - \mathbf{M} \cdot \mathbf{I}||_2^2$$

  - Solve using (Non-Negative) **Least Squares**
    - Well known, can use gradient descent
    - Time ~1 sec when *M* is 200,000 x 4,000 matrix
- **Predicting future volume: Simple!**
  - Given M and I, then

    V = M * I

[ICDM '10]

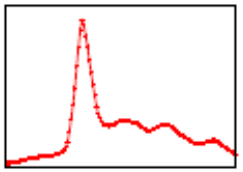
LIM: Performance

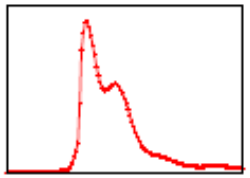
- **Memetracker data**
  - **Node:** website,
  - **Contagion:** textual phrase
- **Take top 1,000 quotes by the total volume:**
  - Total 372,000 mentions on 16,000 websites
- **Build LIM on 100 highest-volume websites**
  - $V_i(t)$ … number of mentions across 16,000 websites
  - $A_i(t)$ … which of 100 sites mentioned quote *i* and when
- Improvement in L2-norm over 1-time lag predictor: $\hat{V}_k(t+1) = V_k(t)$

[ICDM '10]

LIM: Performance

- **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%**

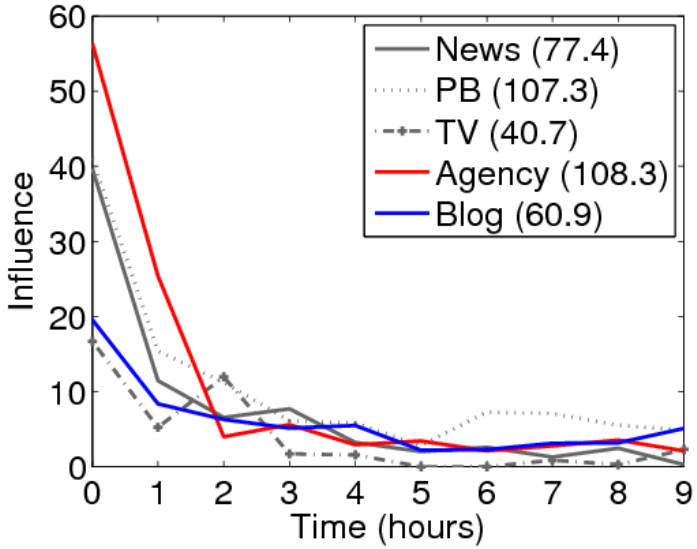
[ICDM '10]

Analysis of Influence Functions

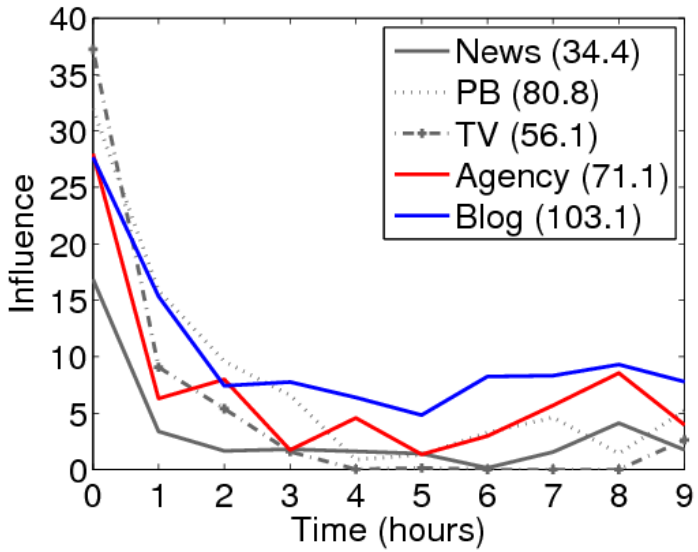
- **Influence functions give insights:**
  - **Q:** NYT writes a post on politics, how many people tend to mention it next day?
  - **A:** Influence function of NYT for political phrases!
- **Experimental setup:**
  - 5 media types:
    - Newspapers, Pro Blogs, TVs, News agencies, Blogs
  - 6 topics: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

[ICDM '10]

Analysis of Influence Functions



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