ANNOUNCEMENT:
Please fill out the course evaluation forms

Meme-tracking & Network Inference

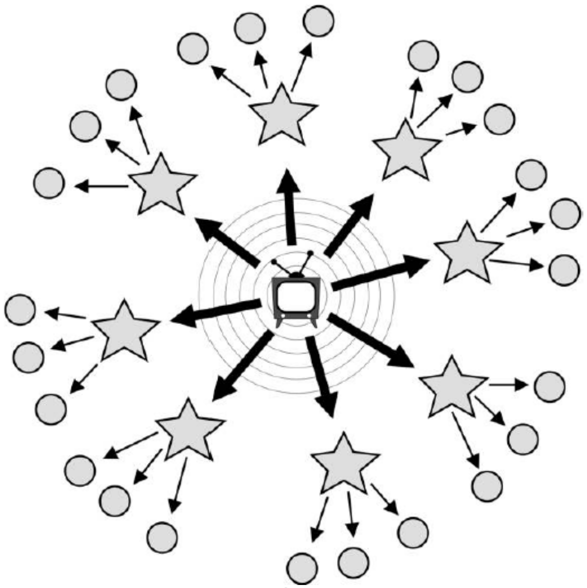
CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

Web as a “macroscope”

- **Explosion of online (social) media:**
  - Blogs (personal/professional)
  - Traditional (TV, Newspapers, Agencies)
  - Microblogging (Twitter)
- **How does information transmitted by the media interact with social networks?**
- **How does this feed back to the creators of news?**

Global vs. Local effects

- Interaction of global effects from mass media and local effects carried by the social structure (e.g., blogs, Twitter)



[Lazarsfeld-Katz '55]
[Watts-Dodds '07]

- **Internet, blogs, social media:**
  - 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

Challenges

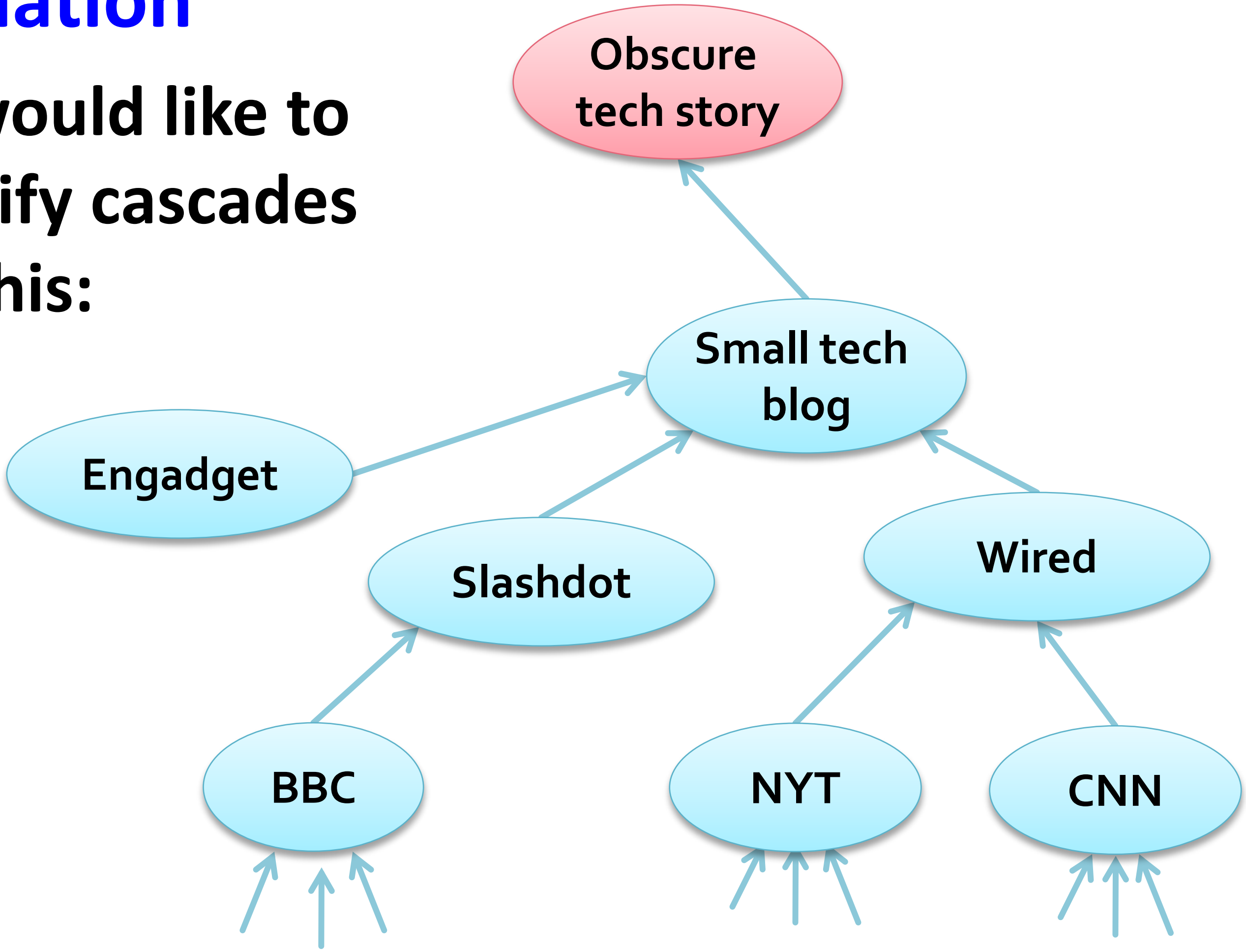
- **Media is about dynamics and information**

- **What are basic “units” of information that spread?**
  - Depends on the question we are asking
  - Depends on the “resolution” at which we want to capture/model news

- **How to automatically identify them?**
- **How to automatically track these units?**

Tracking the Information Flow

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



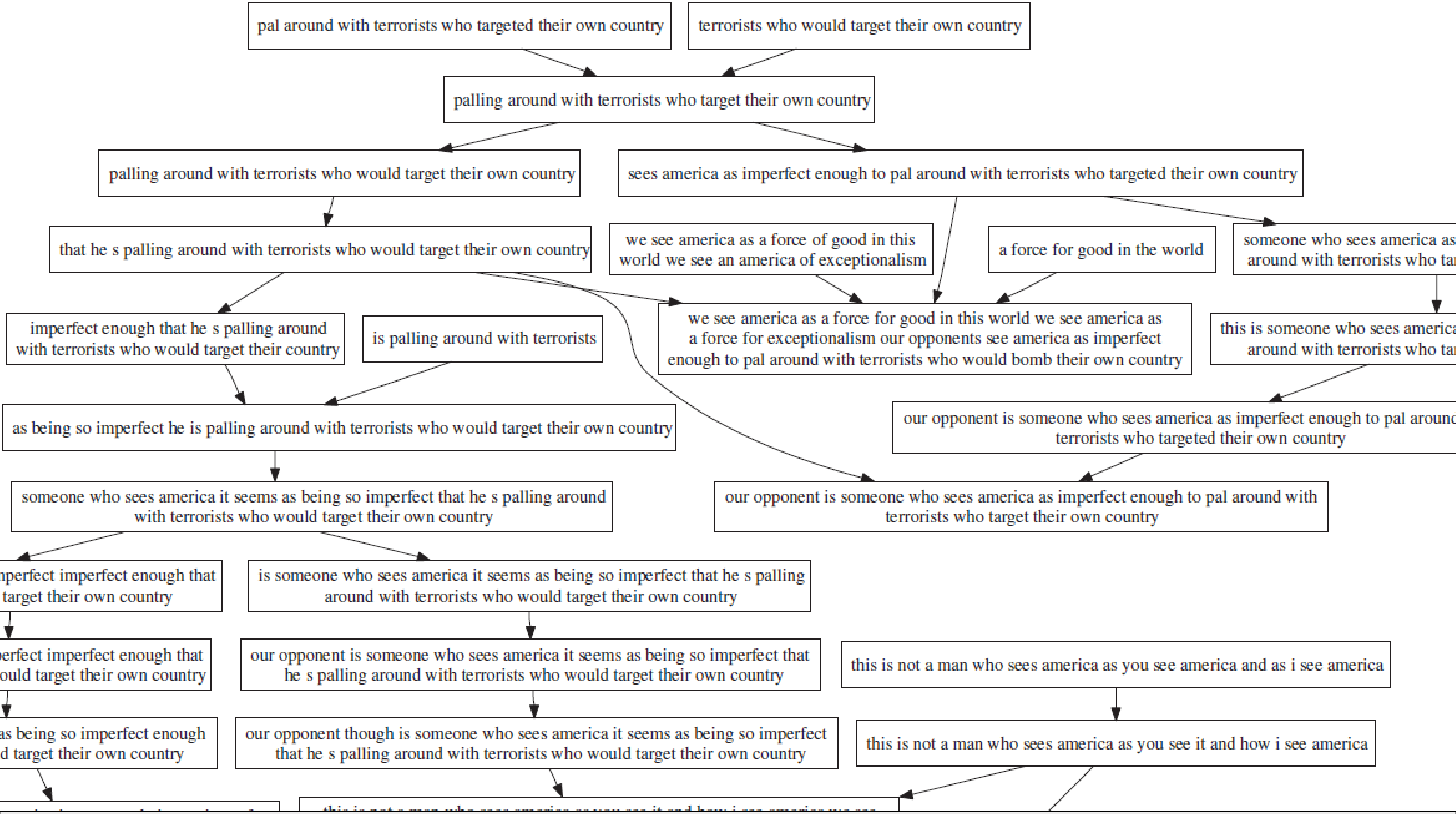
[KDD ‘09]

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
      - 6B news articles and blog posts
  - **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: Phrases Mutate



pal around with terrorists who targeted their own country

terrorists who would target their own country

palling around with terrorists who target their own country

palling around with terrorists who would target their own country

sees america as imperfect enough to pal around with terrorists who targeted their own country

that he s palling around with terrorists who would target their own country

we see america as a force of good in this world we see an america of exceptionalism

a force for good in the world

someone who sees america as around with terrorists who ta

imperfect enough that he s palling around with terrorists who would target their country

is palling around with terrorists

we see america as a force for good in this world we see america as a force for exceptionalism our opponents see america as imperfect enough to pal around with terrorists who would bomb their own country

this is someone who sees america around with terrorists who ta

as being so imperfect he is palling around with terrorists who would target their own country

our opponent is someone who sees america as imperfect enough to pal around terrorists who targeted their own country

someone who sees america it seems as being so imperfect that he s palling around with terrorists who would target their own country

our opponent is someone who sees america as imperfect enough to pal around with terrorists who target their own country

nperfect imperfect enough that target their own country

is someone who sees america it seems as being so imperfect that he s palling around with terrorists who would target their own country

erfect imperfect enough that ould target their own country

our opponent is someone who sees america it seems as being so imperfect that he s palling around with terrorists who would target their own country

this is not a man who sees america as you see america and as i see america

as being so imperfect enough d target their own country

our opponent though is someone who sees america it seems as being so imperfect that he s palling around with terrorists who would target their own country

this is not a man who sees america as you see it and how i see america

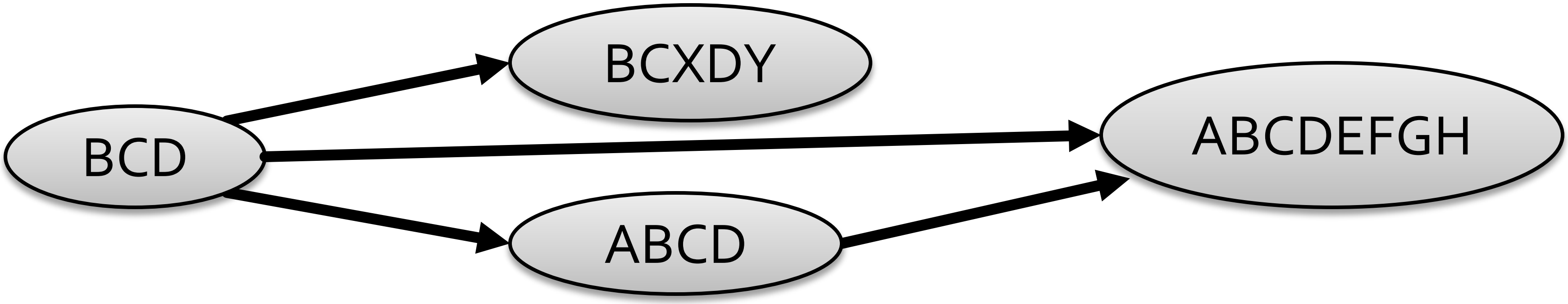
this is not a man who sees america as you see it and how i see america we see

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

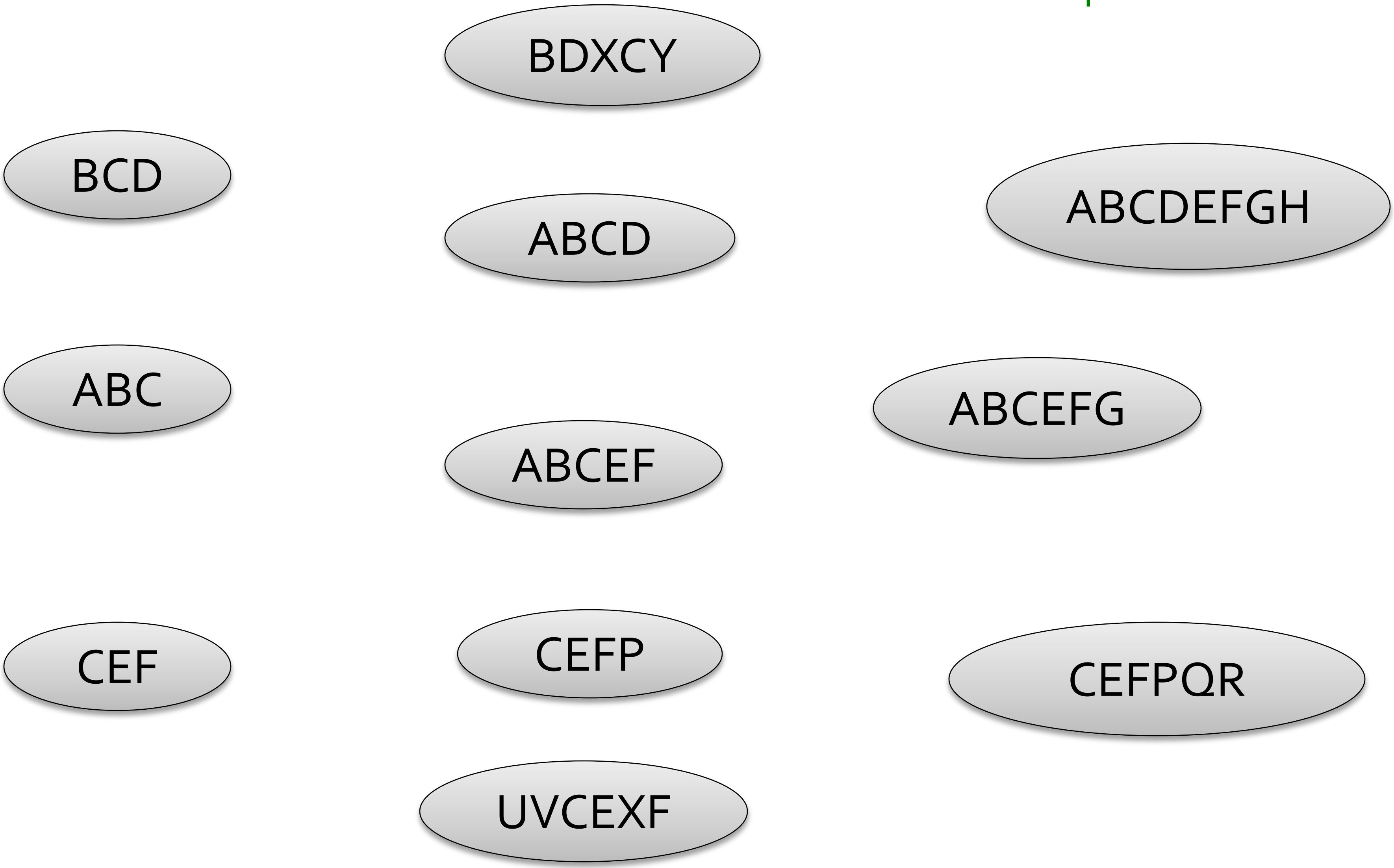
- **Goal:** Find mutational variants of a phrase
- Form **approximate phrase inclusion graph**
  - Shorter phrase is approximately included in a longer one (swap/add/delete a word, d(BCD,BCXDY)=2)



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

Creating Clusters of Mutations

Nodes are phrases



BDXCY

BCD

ABCD

ABCDEFGH

ABC

ABCEFG

ABCEF

CEF

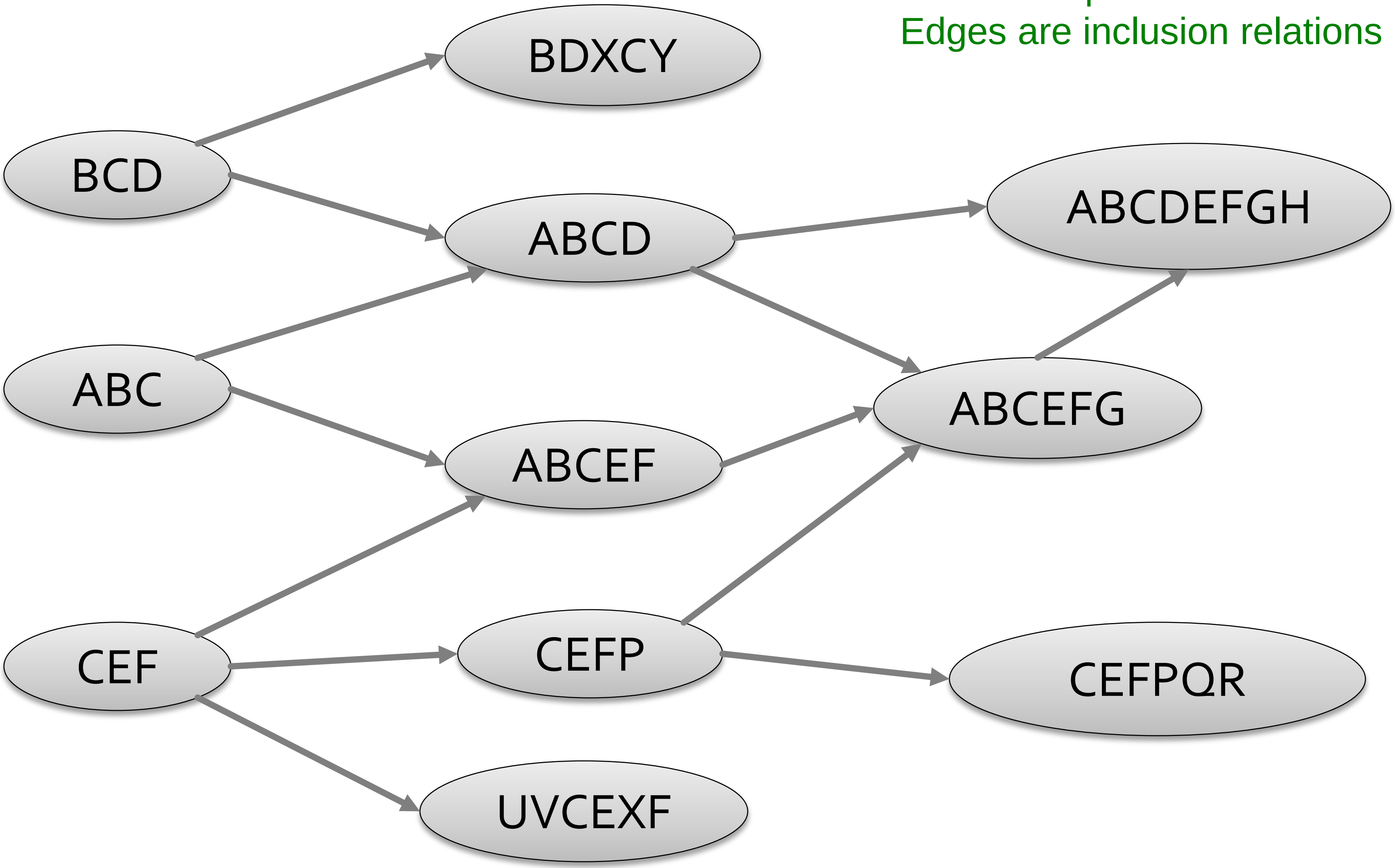
CEFP

CEFPQR

UVCEXF

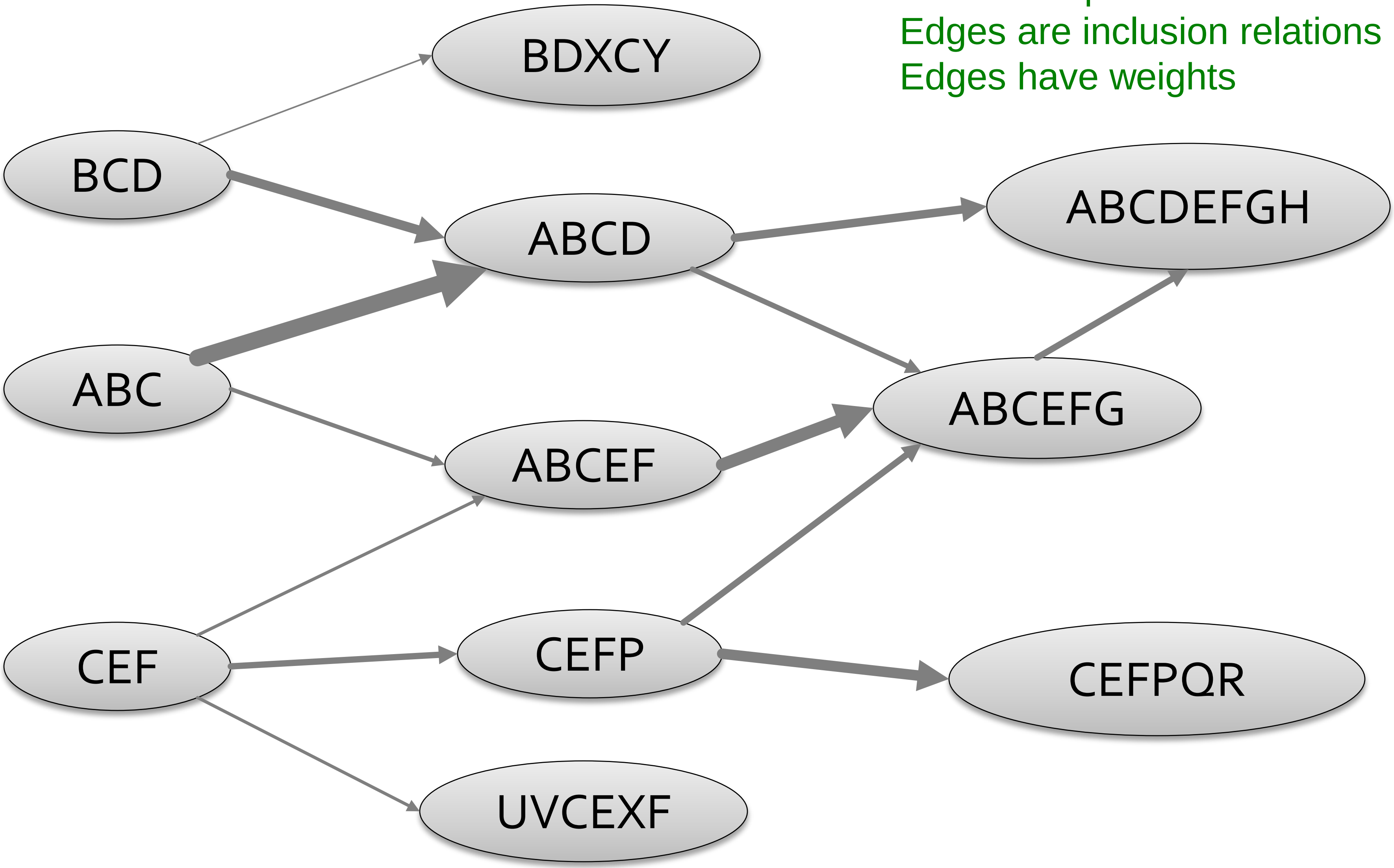
Creating Clusters of Mutations

Nodes are phrases
Edges are inclusion relations



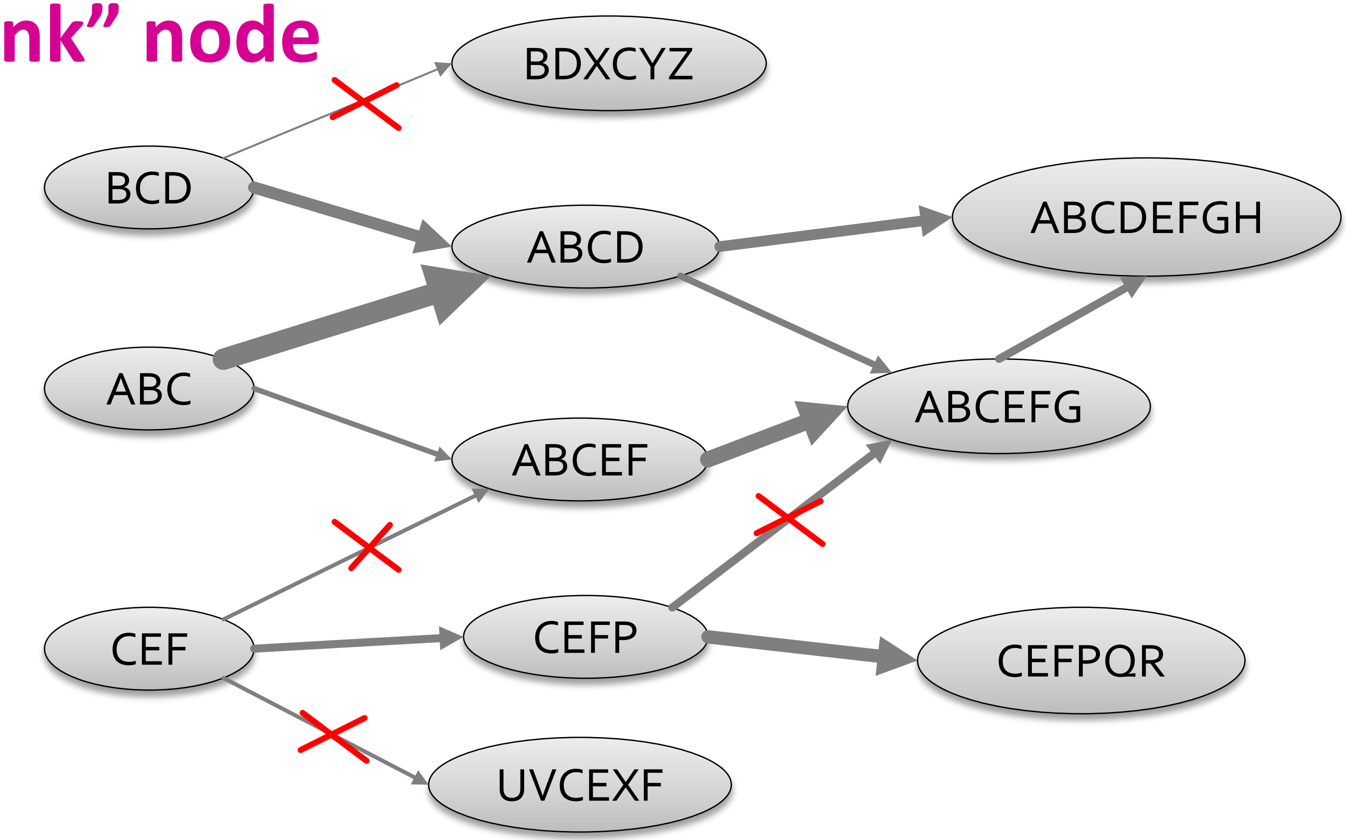
Creating clusters of Mutations

Nodes are phrases
Edges are inclusion relations
Edges have weights



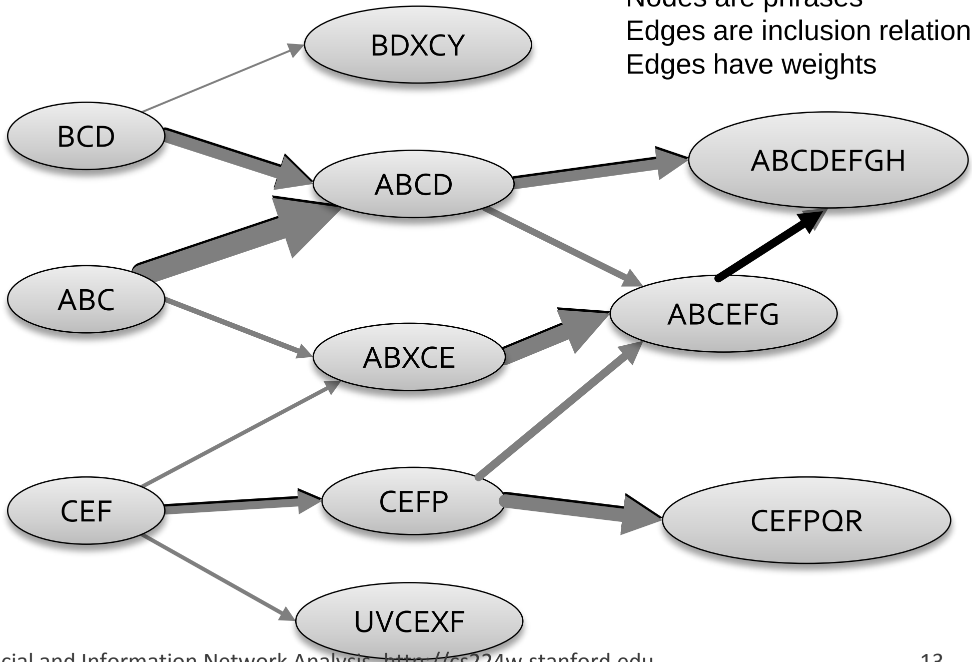
Quote Clustering: DAG Partitioning

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



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

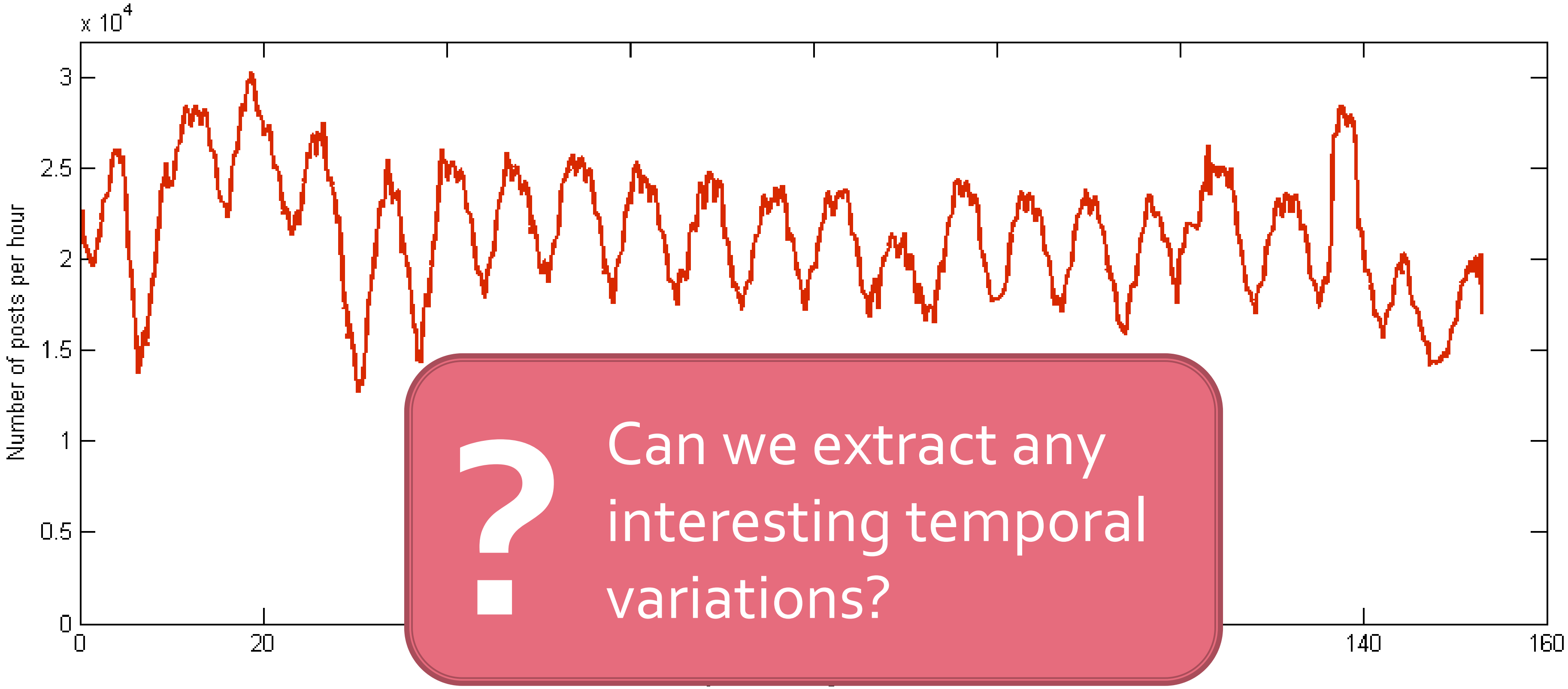


Nodes are phrases
Edges are inclusion relations
Edges have weights

Meme: A Phrase 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
today of all days john mccain's stubborn insistence that the fundamentals of the economy are strong shows that he is	

Memes Over Time

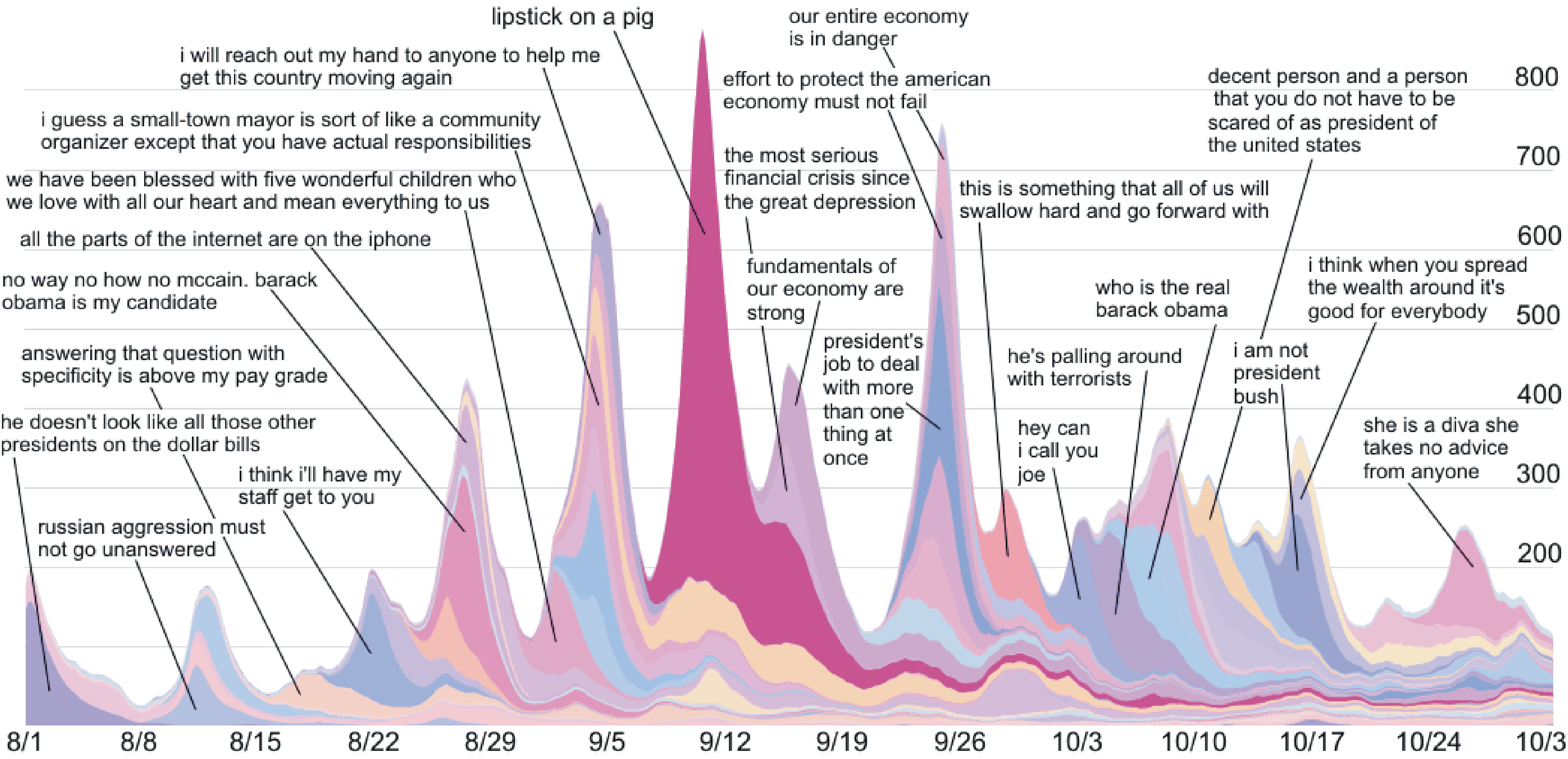


? Can we extract any interesting temporal variations?

… is periodic, has no “real” trends.

”Bandwidth” of the online media is constant

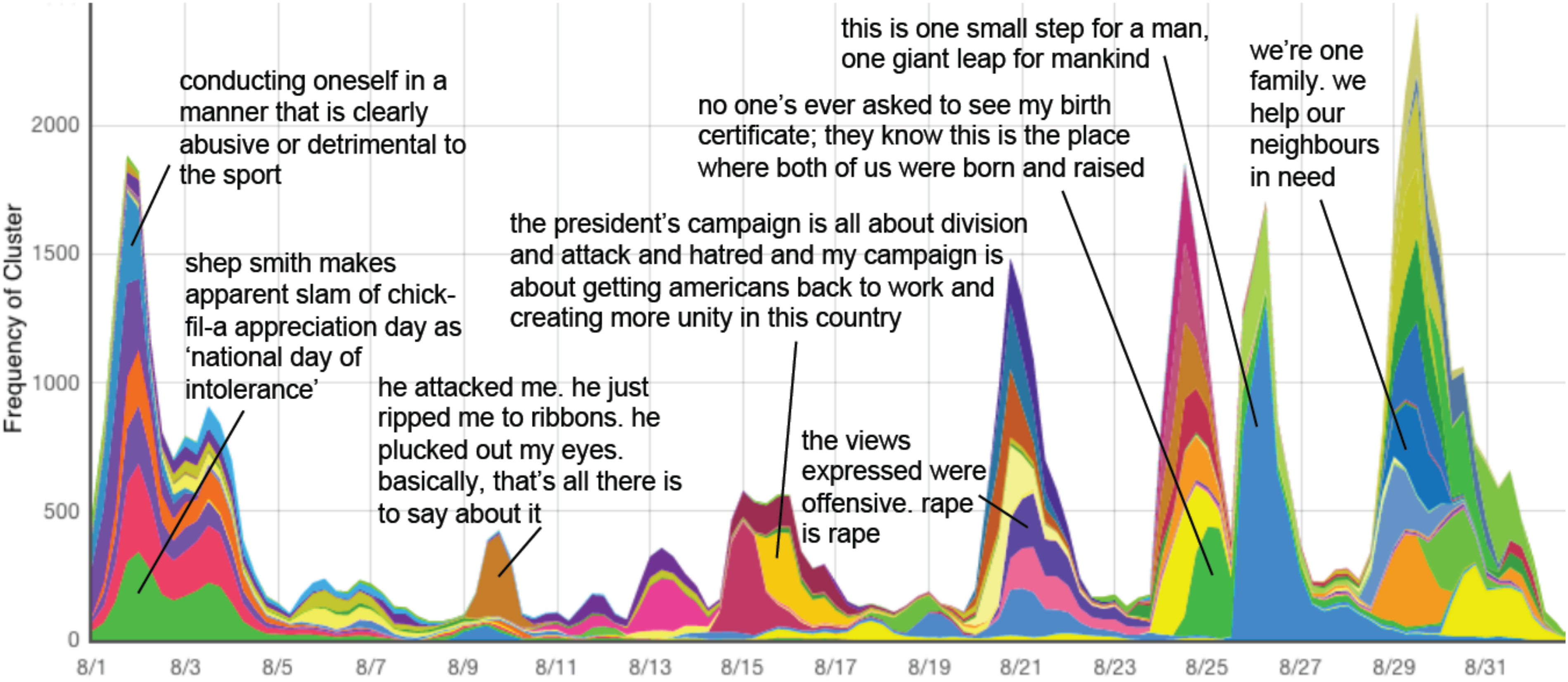
Memes Over Time



August

October

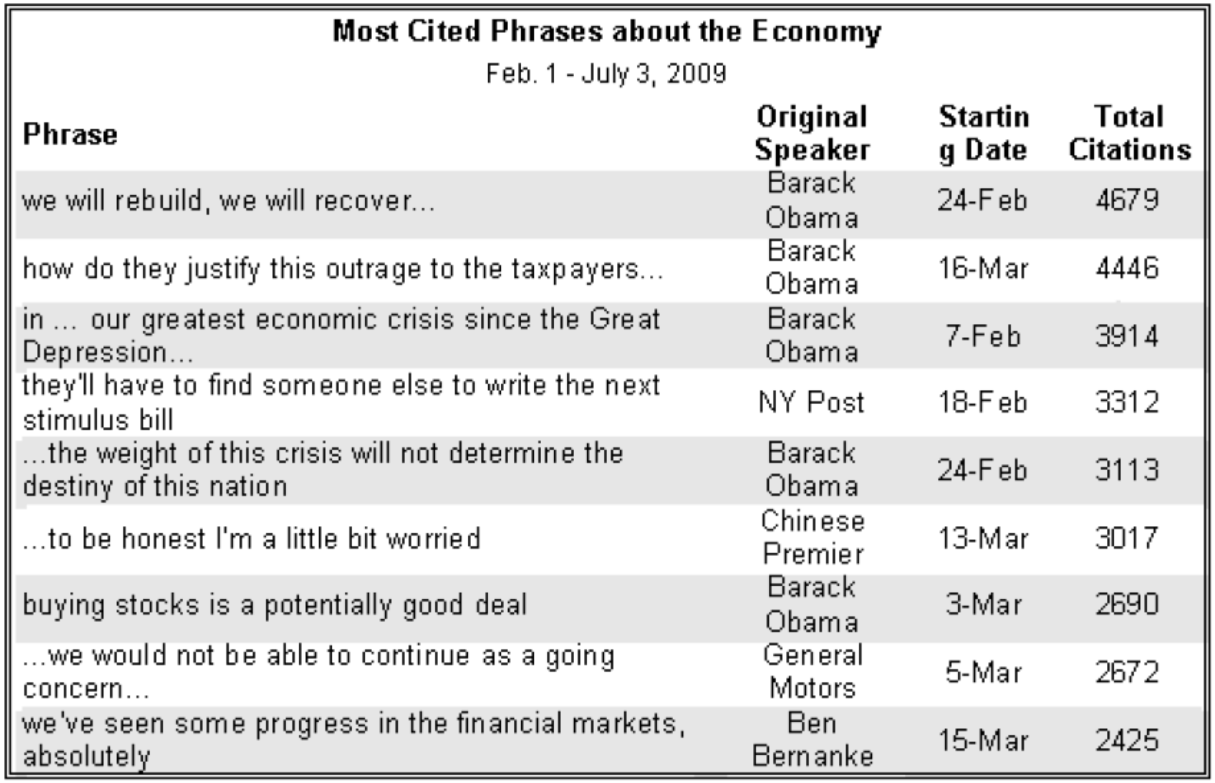
Meme Volume Over Time



- Volume over time of top 50 largest total volume memes (phrase clusters)
- More at: http://snap.stanford.edu/nifty

Memes 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

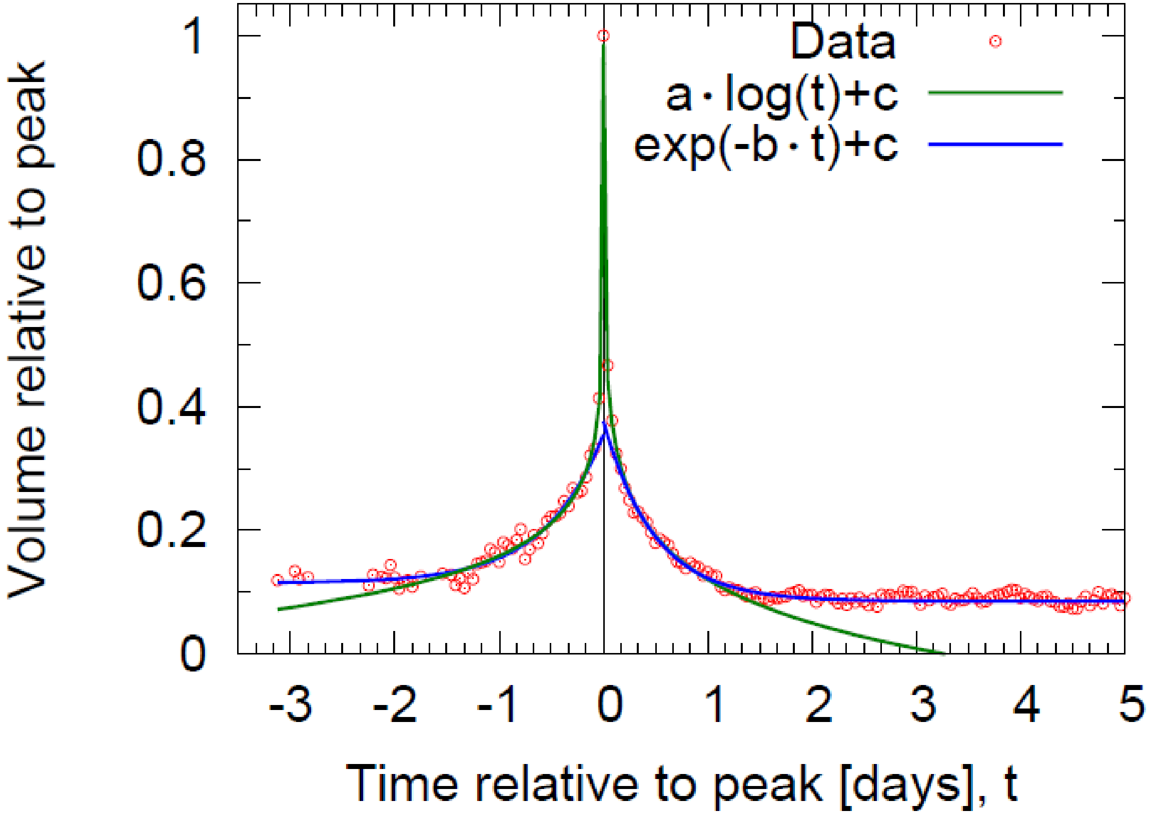
THEY'LL HAVE TO FIND SOMEONE ELSE TO WRITE THE NEXT STIMULUS BILL

60-minutes interview

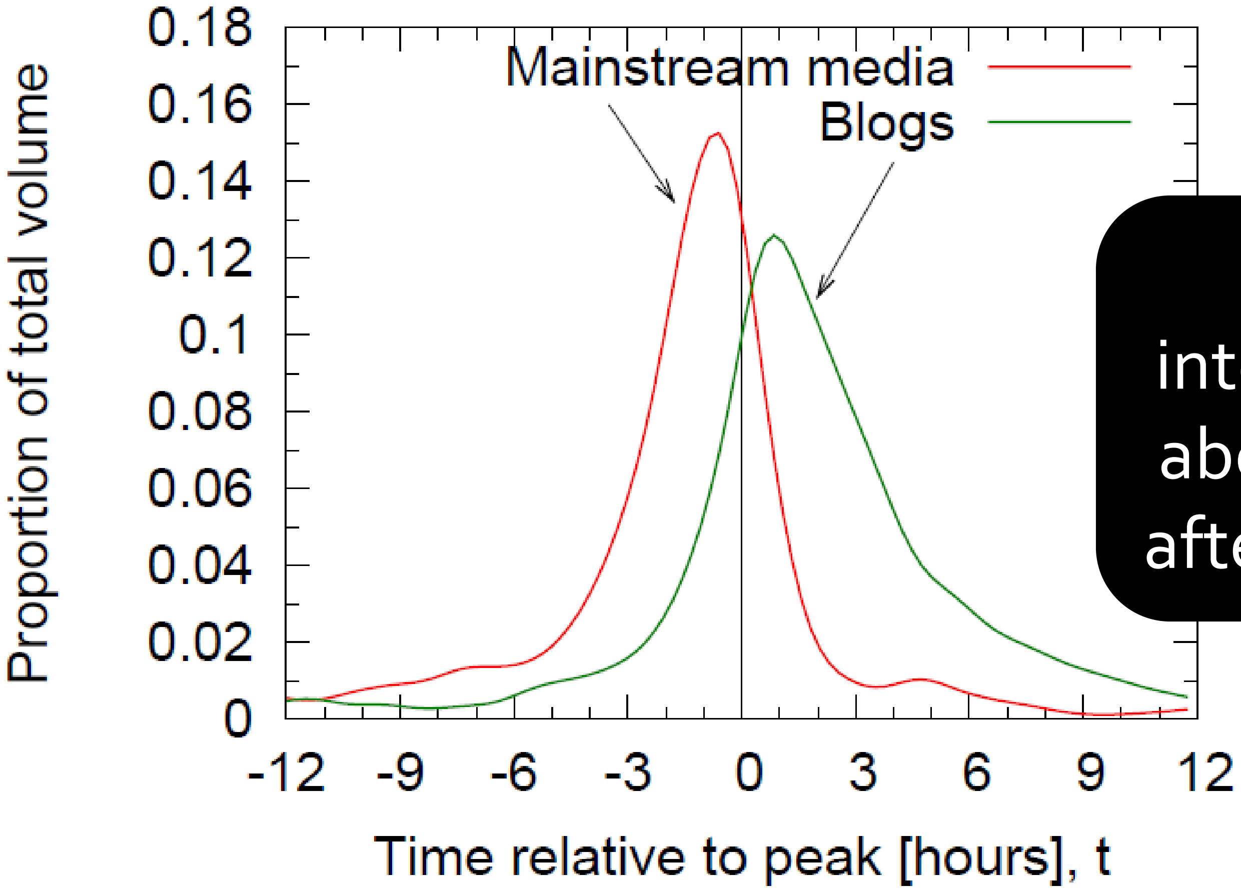
Top republican voice ranks only 14th

Interaction of News and Blogs

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



Interaction of News and Blogs

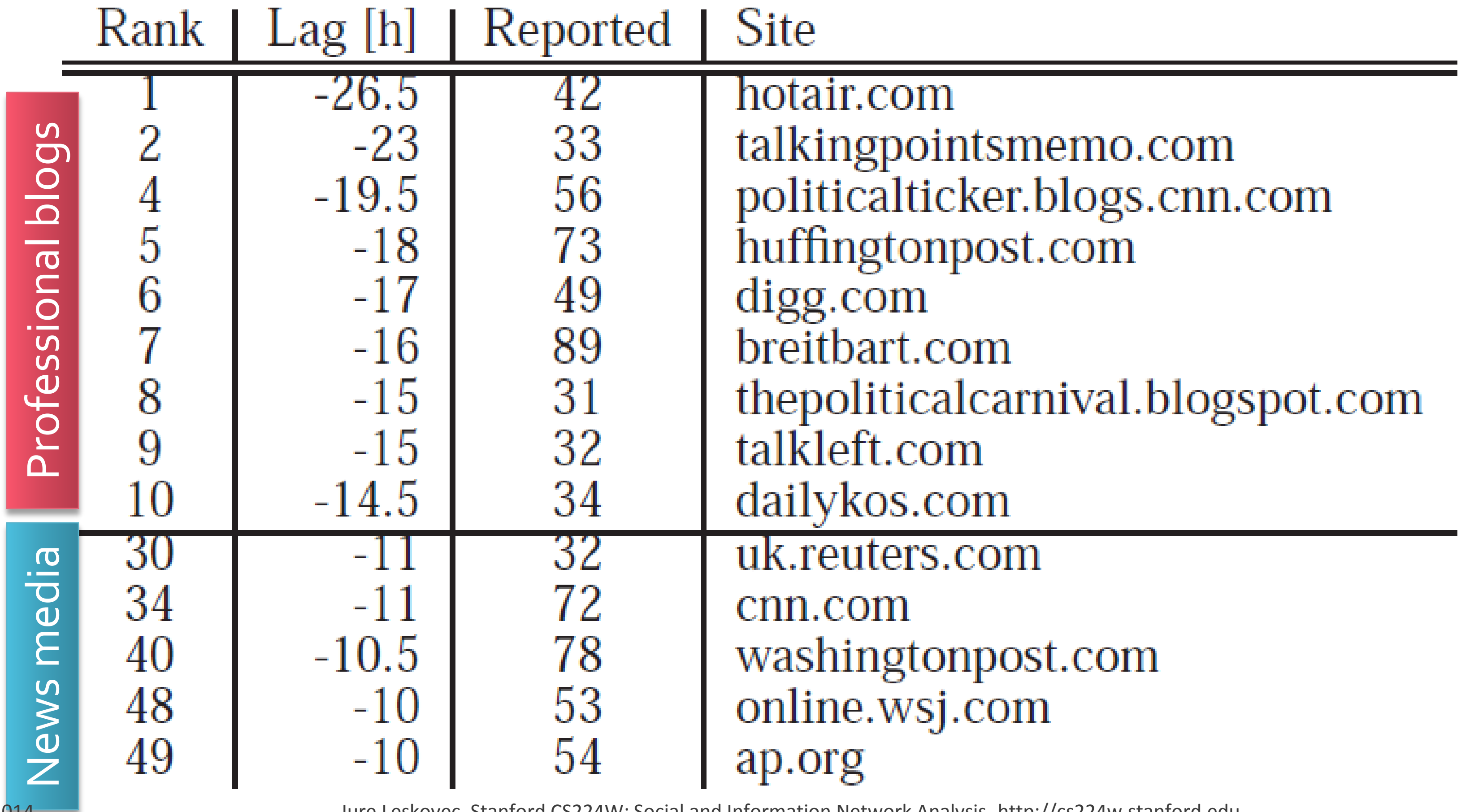


Peak blog intensity comes about 2.5 hours after news peak.

- **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?

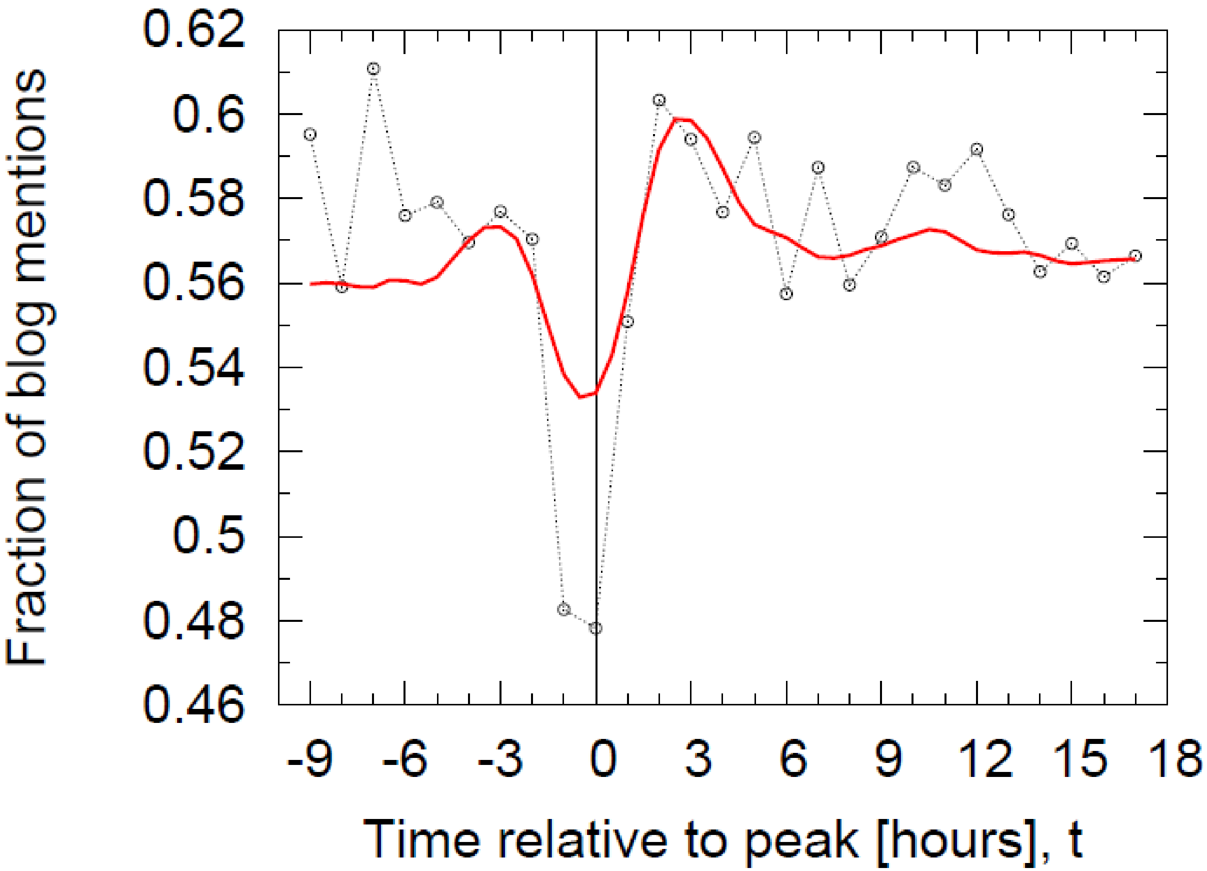
- **Classify individual sources by their typical timing relative to the peak aggregate intensity**



	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

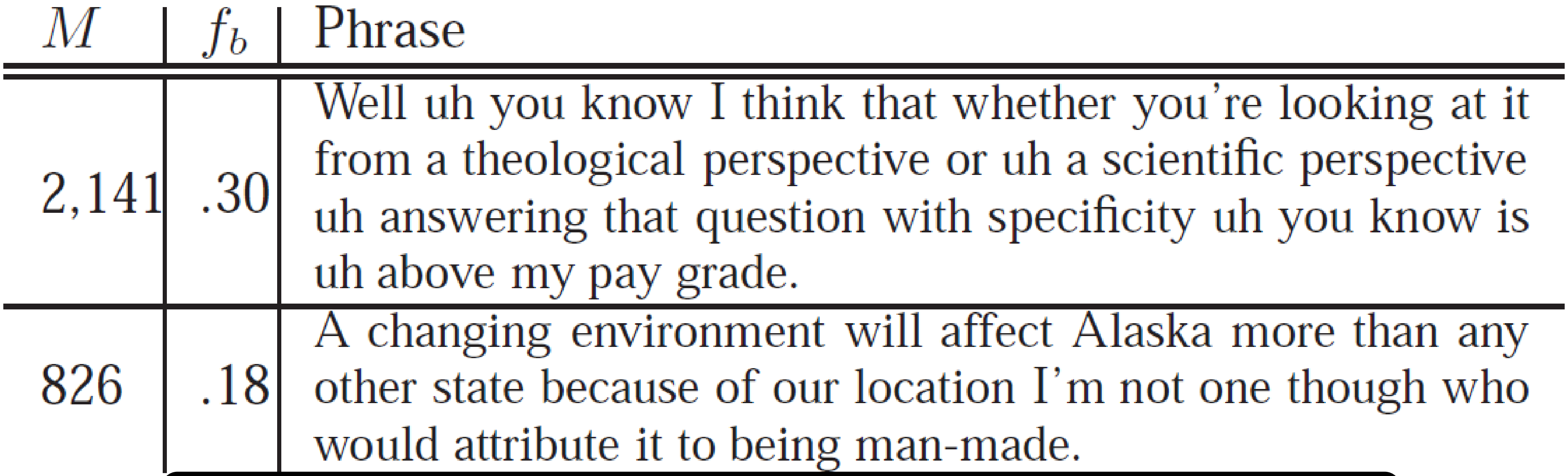
- The “oscillation” of attention between mainstream media and blogs



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

Network Inference

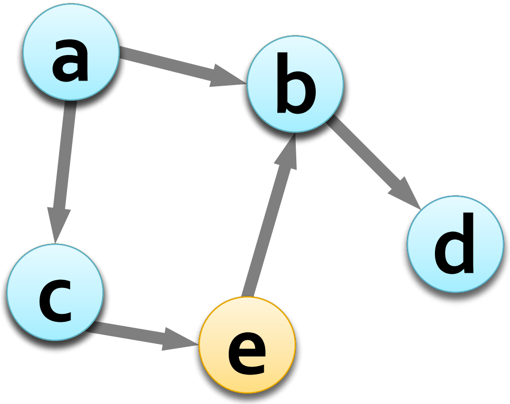
Hidden and Implicit Networks

- **Many networks are implicit or hard to observe:**
  - Hidden/hard-to-reach populations:
    - Network of needle sharing between drug injection users
  - Implicit connections:
    - Network of information propagation in online news media
- **But we can observe results of the processes taking place on such (invisible) networks:**
  - **Virus propagation:**
    - Drug users get sick, and we observe when they see the doctor
  - **Information networks:**
    - We observe when media sites mention information
- **Question: Can we infer the hidden networks?**

[KDD '10, NIPS '10]

Inferring the Diffusion Networks

- **There is a hidden diffusion network:**

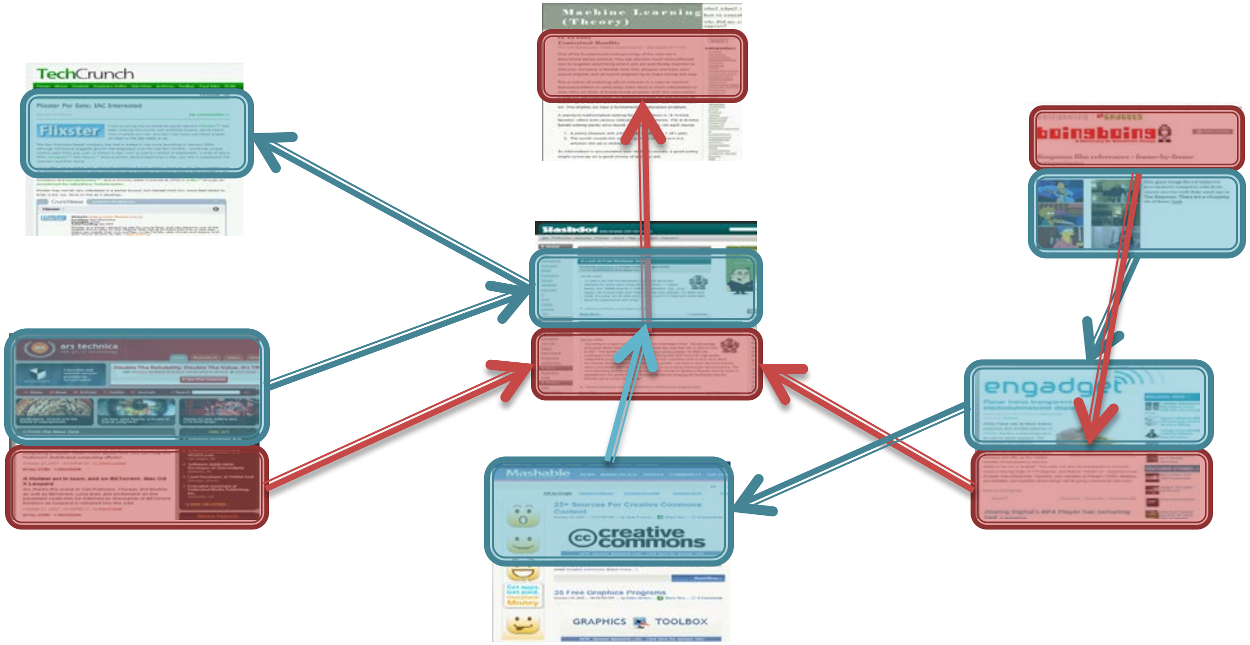


- We only see times when nodes get "infected":
  - Cascade c_1: (a,1), (c,2), (b,3), (e,4)
  - Cascade c_2: (c,1), (a,4), (b,5), (d,6)
- **Want to infer who-infects-whom network!**

[KDD, '09]

Examples and Applications

- **Information diffuses through the blogosphere**



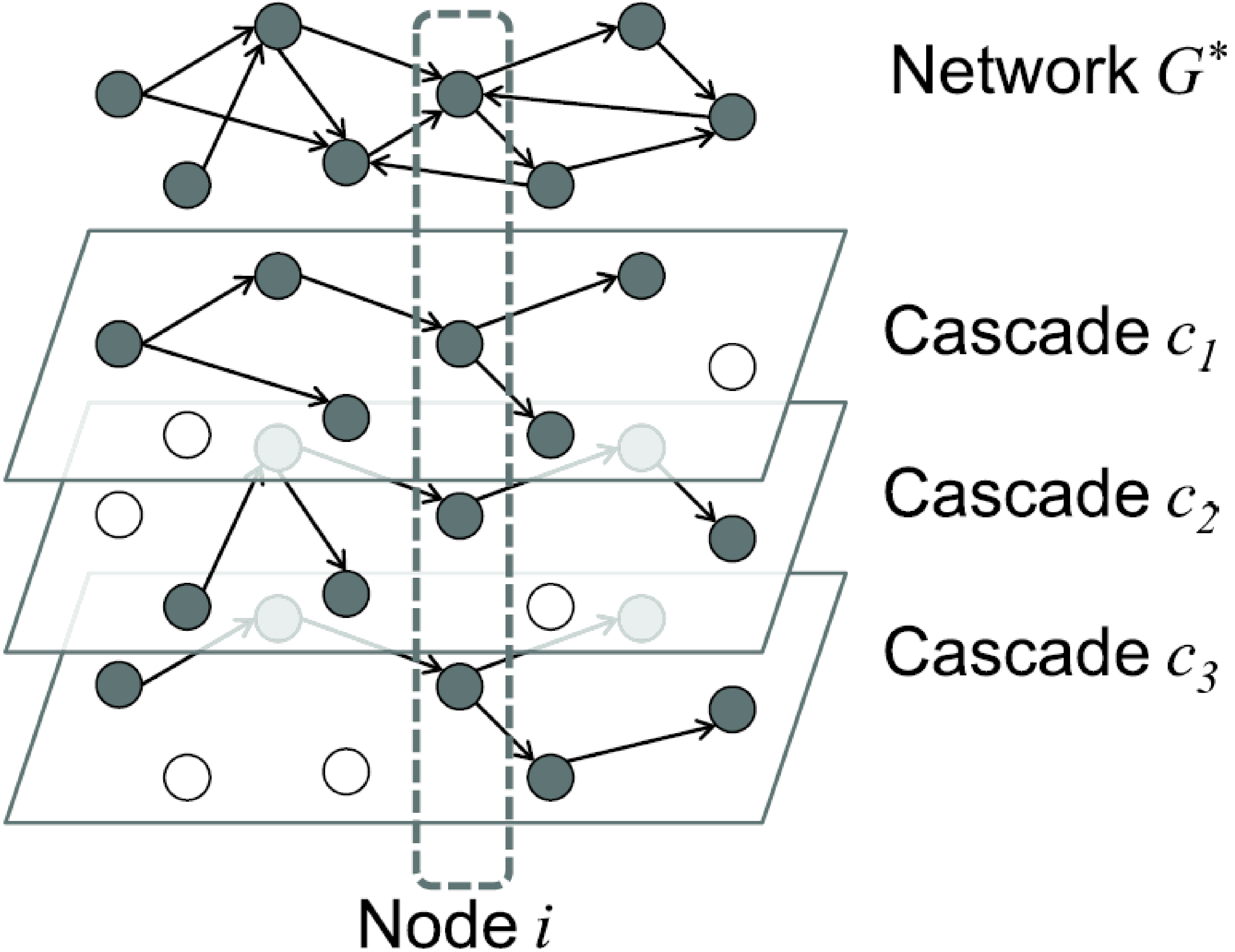
- We only see the mention but not the source
- Can we reconstruct (hidden) **diffusion network**?

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

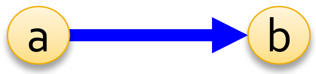
Can we infer the underlying network?

Inferring the Diffusion Network

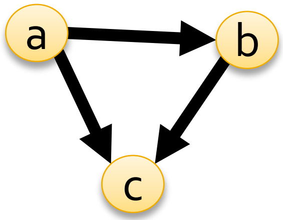


Network Inference: The Task

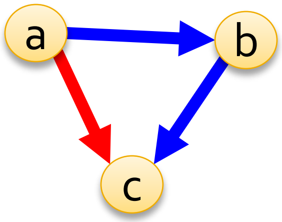
- **Goal: Find a graph G that best explains the observed infection times**
  - **Given a graph G, define the likelihood P(C|G):**



P_c(a,b): How likely is *a* to infect *b*

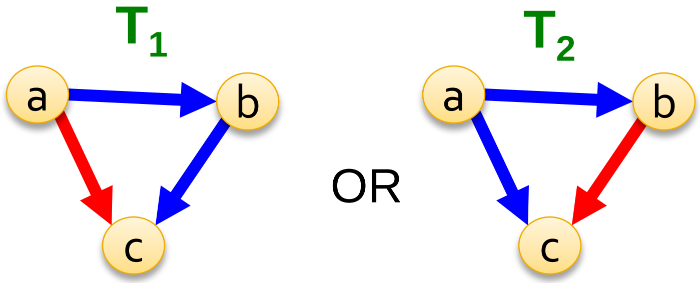


Graph G



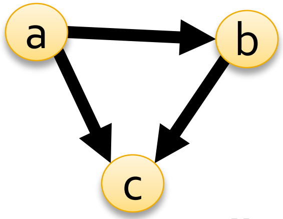
P(c|T): How likely is *c* to propagate via cascade-tree T

Here: T={a → b → c}



P(c|G): How likely is *c* to propagate in graph G

In both T_1,T_2 the order of infections is the same: a,b,c



P(C|G): How likely is a set of $c \in C$ to propagate in G

Network Inference: The Task

- **Goal: Find a graph G that best explains the observed infection times**
  - **Given a graph G, define the likelihood P(C|G):**

    Define a model of information diffusion over a graph

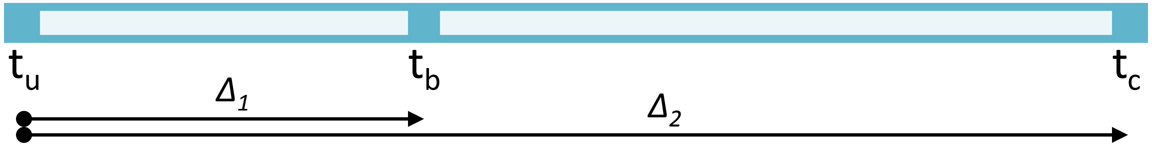
    - **P_c(a,b)** … prob. that *a* infects *b* in contagion *c*
    - **P(c|T)** … prob. that *c* spread in particular cascade-tree *T*
    - **P(c|G)** … prob. that cascade *c* occurred in *G*
    - **P(C|G)** … prob. that a set of cascades *C* occurred in *G*
- **Questions:**
  - How to efficiently compute P(G|C)? (given a single G)
  - How to efficiently find G^* that maximizes P(G|C)? (over $O(2^{N*N})$ graphs)

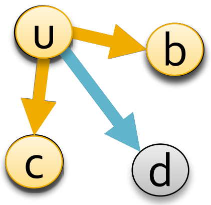
Cascade Diffusion Model

- **Continuous time cascade diffusion model:**
  - Cascade *c* reaches node *u* at t_u and spreads to *u*'s neighbors:
    - With probability *β* cascade propagates along edge *(u, v)* and we determine the infection time of node v

      $t_v = t_u + \Delta$

      e.g.: *Δ ~ Exponential*



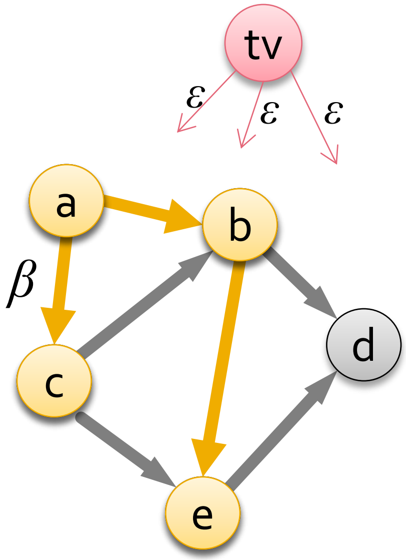


We assume each node *v* has only one parent!

Cascade Diffusion Model

- **The model for one cascade:**
  - Cascade reaches node *u* at time t_u, and spreads to *u*'s neighbors *v*:

    With prob. β cascade propagates along edge *(u,v)* and $t_v = t_u + \Delta$



- **Transmission probability:**

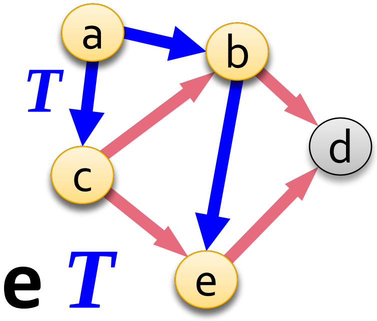
$$P_c(u,v) \propto P(t_v - t_u) \text{ if } t_v > t_u \text{ else } \varepsilon$$

$$e.g.: P_c(u,v) \propto e^{-\Delta t}$$

  - ε captures influence external to the network
    - At any time a node can get infected from outside with small probability ε, equal for all nodes

Cascade Probability

- **Given node infection times & cascade-tree *T*:**
  - *c* = { (a,1), (c,2), (b,3), (e,4) }
  - *T* = { a→b, a→c, b→e }



Graph G

- **Prob. that *c* propagates in cascade-tree *T***

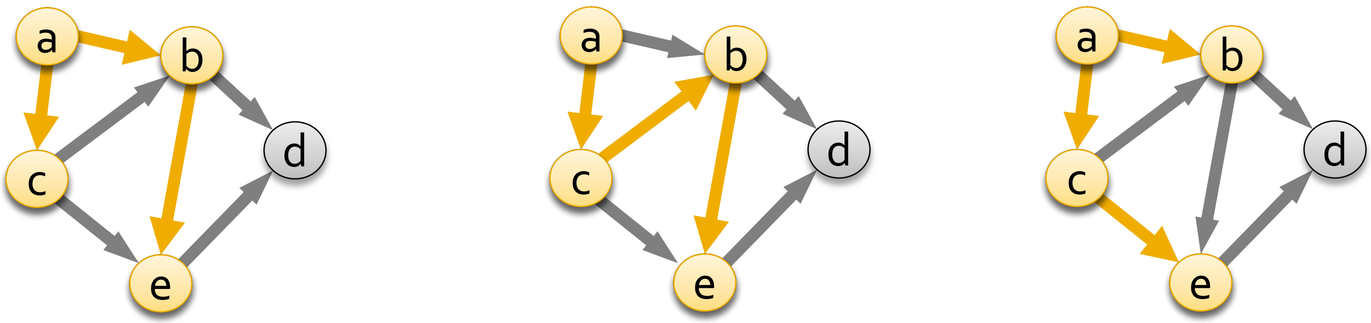
$$P(c|T) = \prod_{(u,v)\in E_T} \beta P_c(u,v) \prod_{u\in V_T,(u,x)\in E\backslash E_T} (1-\beta)$$

Edges that "propagated" Edges that failed to "propagate"

- **Approximate it as:** $P(c|T) \approx \prod_{(u,v)\in E_T} P_c(v,u)$

Complication: Too Many Trees

- **How likely is cascade *c* to spread in graph G?**
  - $c = \{(a,1), (c,2), (b,3), (e,4)\}$



- Need to consider **all possible ways for *c* to spread over *G*** (i.e., all spanning trees *T*):

$$P(c|G) = \sum_{T \in \mathcal{T}_c(G)} P(c|T) \approx \max_{T \in \mathcal{T}_c(G)} P(c|T)$$

Consider only the most likely propagation tree

The Optimization Problem

- **Score of a graph G for a set of cascades C:**

$$P(C|G) = \prod P(c|G)$$

$$F_C(G) = \sum_{c \in C} \log P(c|G)$$

- **Want to find the "best" graph:**

$$G^* = \underset{|G| \leq k}{\operatorname{argmax}} F_C(G)$$

The problem is NP-hard:
MAX-k-COVER [KDD '10]

How to Find the Best Tree?

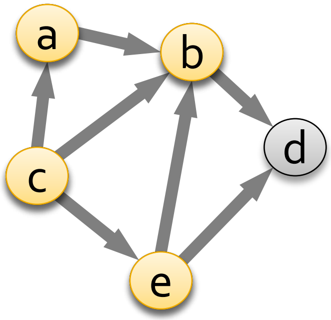
- **Given a cascade c, what is the most likely propagation tree?**

$$\max_{T\in\mathcal{T}_c(G)} P(c|T) = \max_{T\in\mathcal{T}(G)} \sum_{(i,j)\in T} w_c(i,j)$$

  - Maximum **directed** spanning tree
    - Edge *(i,j)* in G has weight $w_c(i,j) = \log P_c(i,j)$
    - The maximum weight spanning tree on infected nodes: Each node picks an in-edge of max weight:

$$= \sum_{i\in V} \max_{Par_T(i)} w(Par_T(i), i)$$

Parent of node *i* in tree *T*

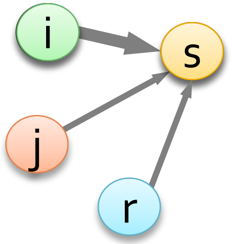


Local greedy selection gives optimal tree!

Great News: Submodularity!

- **Theorem:**
  $F_c(G)$ is monotonic, and submodular
- Proof:
  - Single cascade *c*, some edge *e=(r,s)* of weight. w_{rs}
  - Show $F_c(G \cup \{e\}) - F_c(G) \geq F_c(G' \cup \{e\}) - F_C(G')$
  - Let $w_{.s}$ be max weight in-edge of *s* in G
  - Let $w'_{.s}$ be max weight in-edge of *s* in G'
  - Since $\boldsymbol{G \subseteq G'}$: $w_{.s} \leq w'_{.s}$ and $w_{rs} = w'_{rs}$
  - $F_c(G \cup \{(r,s)\} - F_c(G)$

$$= \max(w_{.s}, w_{rs}) - w_{.s}$$
$$\geq \max(w'_{.s}, w_{rs}) - w'_{.s}$$
$$= F_c(G' \cup \{(r,s)\}) - F_c(G')$$



s picks in-edge of max weight

[KDD '10]

NetInf: The Algorithm

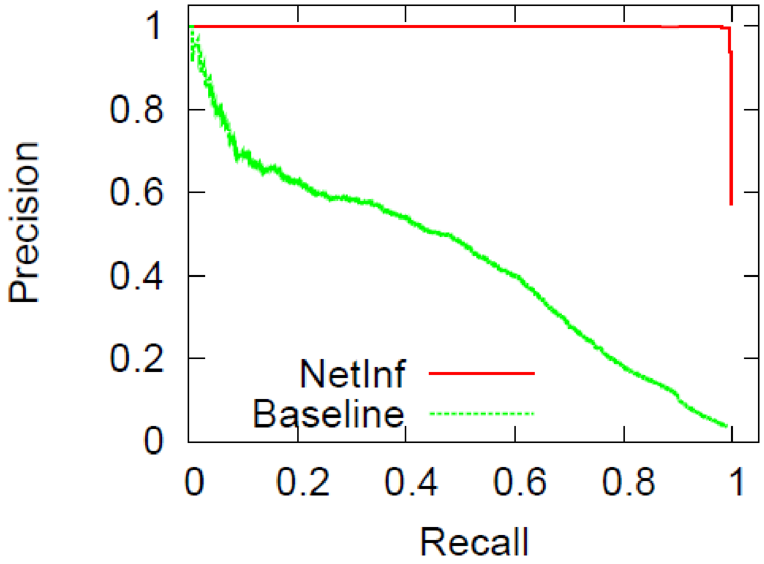
- **The NetInf algorithm:**
  Use **greedy hill-climbing** to maximize $F_C(G)$:
  - Start with empty G_0 (G with no edges)
  - Add k edges (k is parameter)
  - At every step i add an edge to the graph G_i that maximizes the marginal improvement

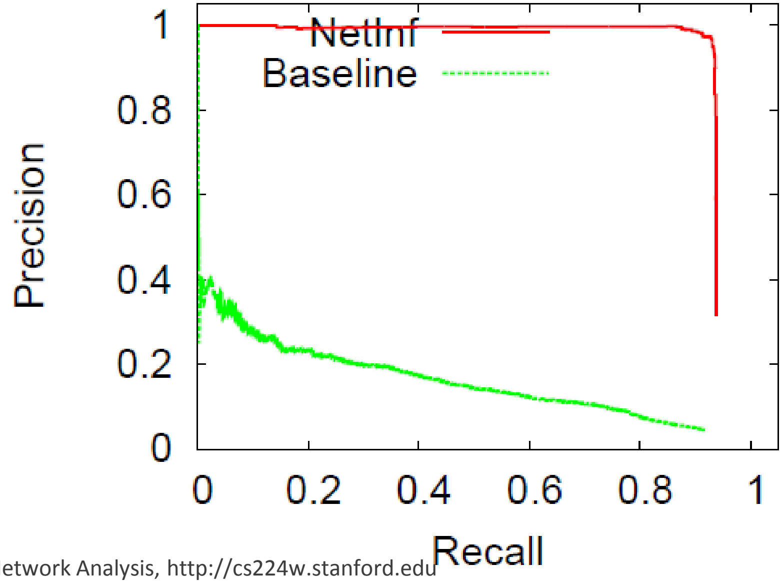
$$e_i = \underset{e \in G \setminus G_{i-1}}{\operatorname{argmax}} \; F_C(G_{i-1} \cup \{e\}) - F_C(G_{i-1})$$

Note: This is the same algorithm we used for influence maximization

Experiments: Synthetic data

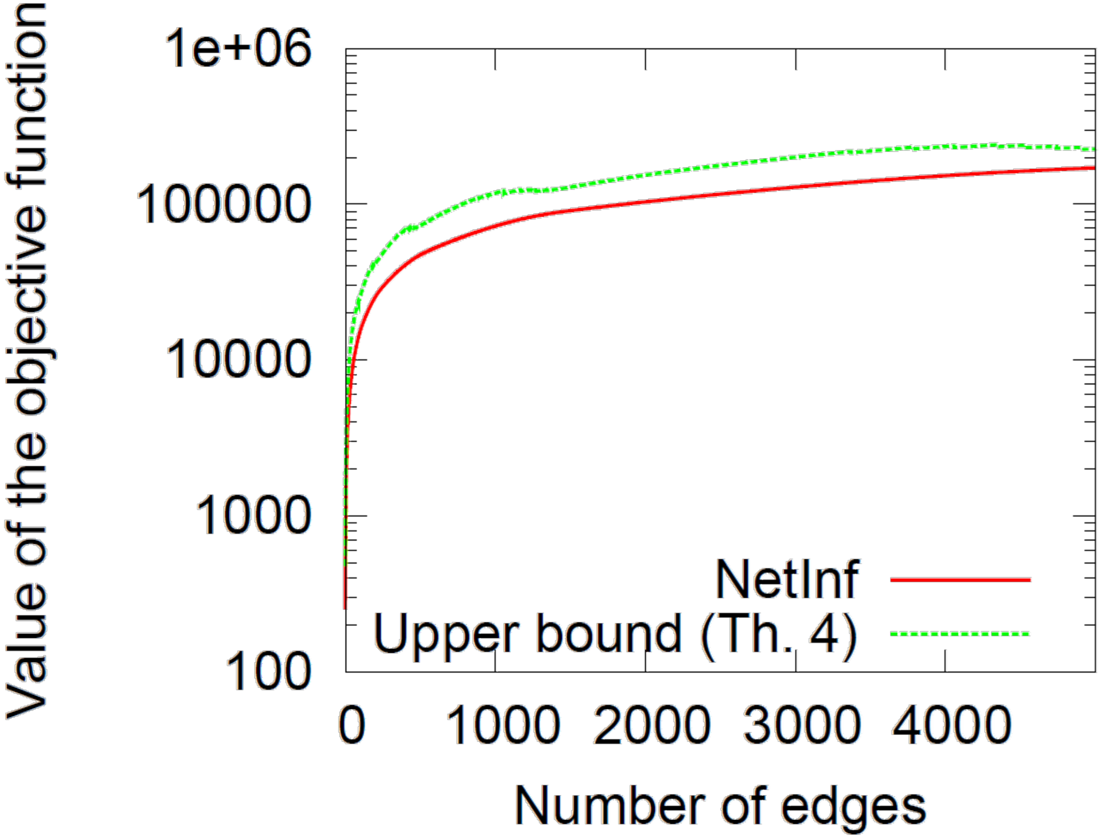
- **Synthetic data:**
  - Take a graph G on *k* edges
  - Simulate info. diffusion
  - Record node infection times
  - Reconstruct G
- **Evaluation:**
  - How many edges of G can NetInf find?
    - Break-even point (precision=recall): 0.95
    - Performance is independent of the structure of G!





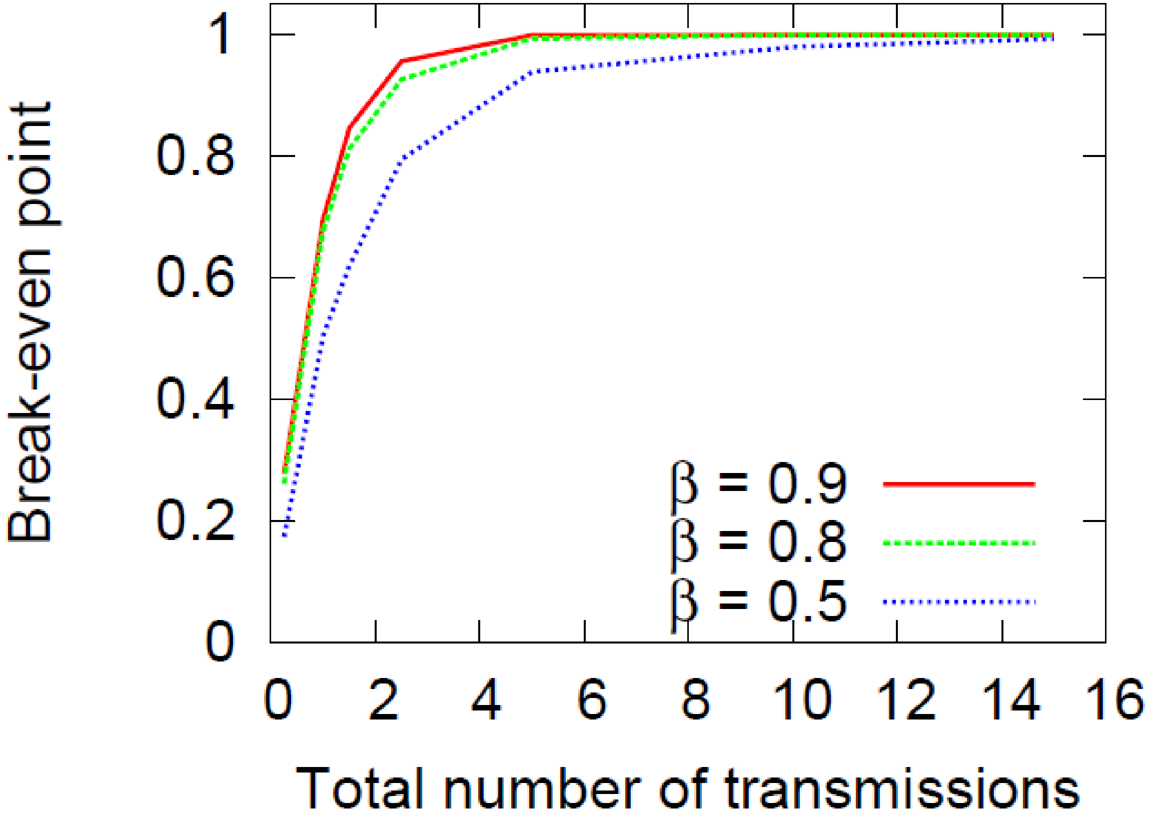
How Good is Our Graph?

- **NetInf achieves ≈ 90 % of the best possible network!**



How Many Cascades Do We Need?

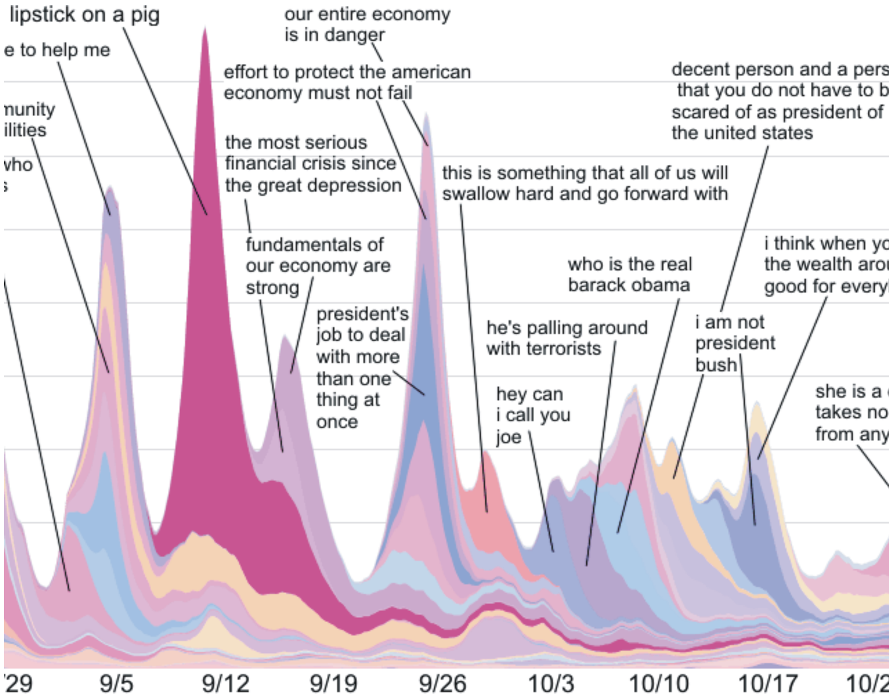
- **With 2x as many infections as edges, the break-even point is already 0.8 - 0.9!**



[KDD, '09]

Experiments: Real data

- **Memetracker dataset:**
  - 172m news articles
  - Aug '08 – Sept '09
  - 343m textual phrases
  - Times $t_c(w)$ when site w mentions phrase c

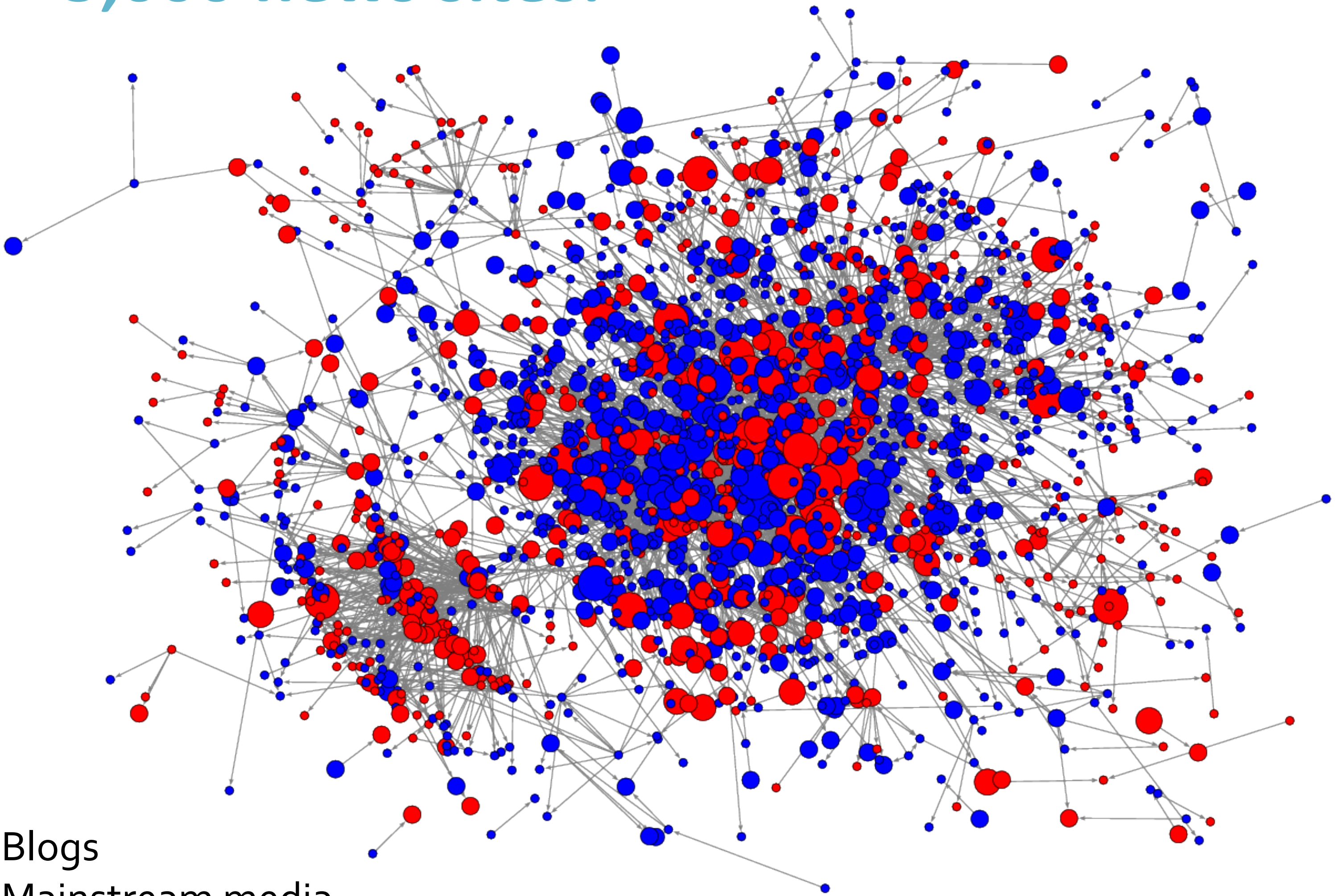


http://memetracker.org

- Given times when sites mention phrases
- Infer the network of information diffusion:
  - Who tends to copy (repeat after) whom

Example: Diffusion Network

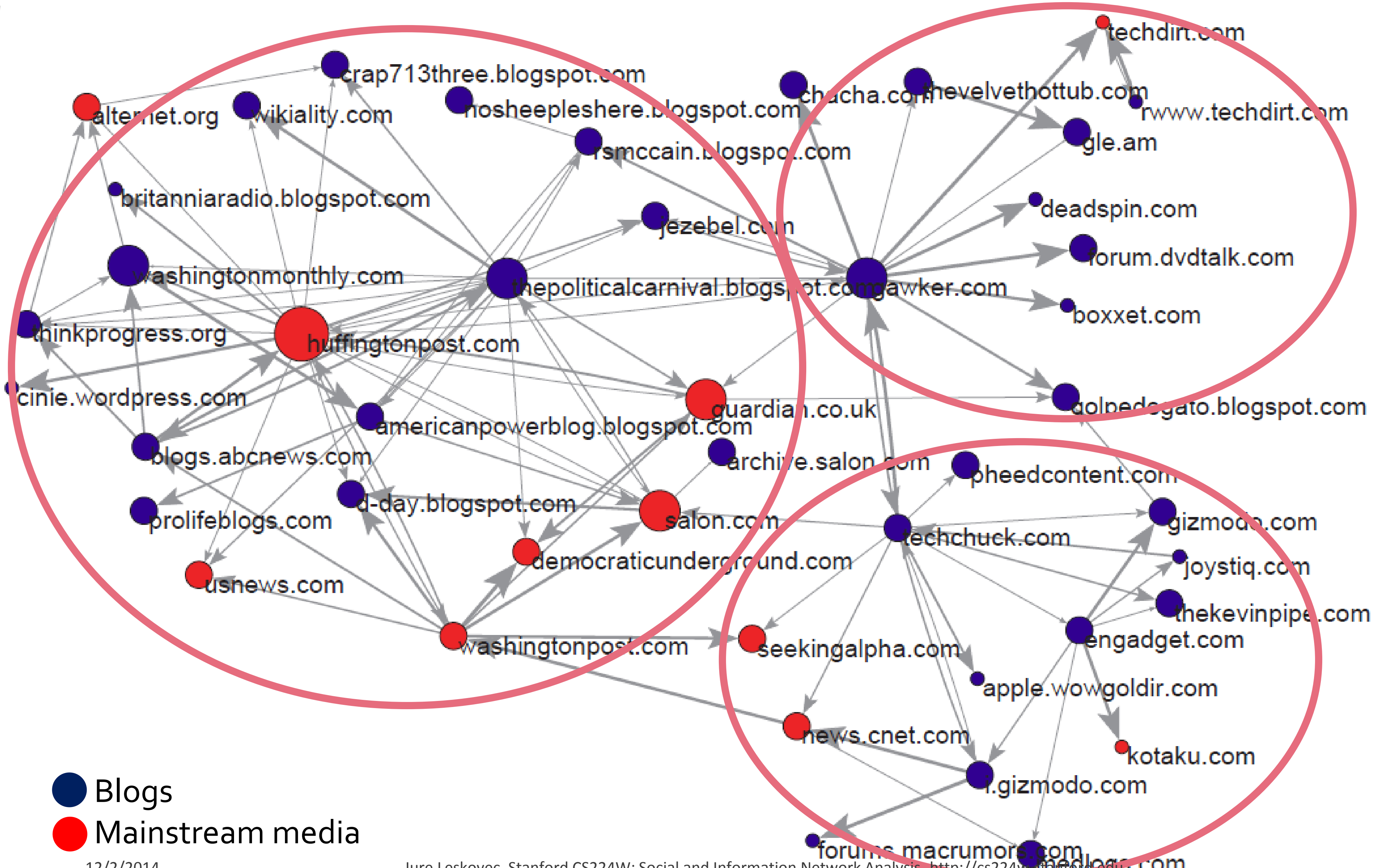
- 5,000 news sites:



● Blogs

● Mainstream media

Diffusion Network (small part)


techdirt.com
crap713three.blogspot.com
alternet.org
wikiality.com
nosheepleshere.blogspot.com
chacha.com
thevelvethottub.com
www.techdirt.com
rsmccain.blogspot.com
gle.am
britanniaradio.blogspot.com
deadspin.com
jezebel.com
forum.dvdtalk.com
washingtonmonthly.com
thepoliticalcarnival.blogspot.com
gawker.com
boxxet.com
thinkprogress.org
huffingtonpost.com
cinie.wordpress.com
guardian.co.uk
golpedegato.blogspot.com
americanpowerblog.blogspot.com
blogs.abcnews.com
archive.salon.com
pheedcontent.com
d-day.blogspot.com
prolifeblogs.com
salon.com
gizmodo.com
techchuck.com
democraticunderground.com
joystiq.com
usnews.com
thekevinpipe.com
engadget.com
washingtonpost.com
seekingalpha.com
apple.wowgoldir.com
news.cnet.com
kotaku.com
i.gizmodo.com
forums.macrumors.com
medlogs.com
● Blogs
● Mainstream media