http://memetracker.org

Meme-tracking and the Dynamics of the News Cycle

Jure Leskovec
Computer Science Department
Stanford University

Includes joint work with Jawon Yang, Manuel Gomez-Rodriguez, Jon Kleinberg, Lars Backstrom, Andreas Krause, Christos Faloutsos, Carlos Guestrin

Information and Media

- Intersection of news media, technology, and the political process
- From its early stages, a tension between global effects from the mass media and local effects carried by social structure

How does information transmitted by the media interact with the personal influence arising from social networks?

Fragmentation and Acceleration

- Internet, blogging, and 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, with no pauses
- The ”24-hour news cycle”:
  - Difficult to define, but associated with technological acceleration and a challenge to healthy civic discourse [Kovach-Rosenstiel ‘99]

Defining News Cycle (1)

- Sept. 11, 2008 (New York Times): *"Mr. McCain's increasingly aggressive campaign has sought to put Mr. Obama on the defensive in each news cycle, using any development at hand, like Mr. Obama's colloquial comment this week about 'putting lipstick on a pig.'"*

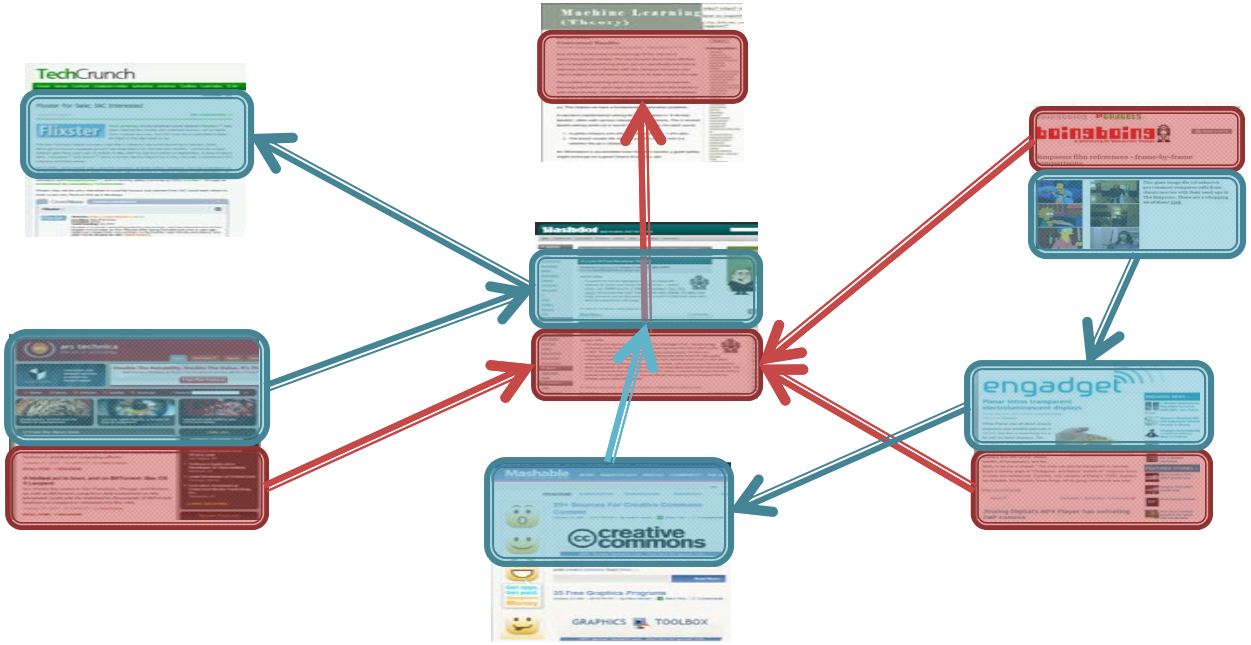
- Oct. 10, 2008 (New York Times): *"Mr. McCain's traveling road show has veered from message to message ... . Each news cycle seems to bring another tactic as the campaign appears to be trying anything and everything to see what might work."*

Defining News Cycle (2)

- **Question:** Is the "news cycle" simply a metaphorical construct describing our perception of the news, or is it visible in data?

- And if it's visible, can we measure some of its basic properties?

Information propagation




CULTURAL INFORMATION
PRACTICE OR IDEA OR CONCEPT
SONGS
NATURAL SELECTION
IDEAS
SELF-PROPAGATING
THEM
MEME

Units of analysis?

- What basic "units" make up the news cycle?
  - Cascasding hyper-links to articles: too fine-grained
    [Adar et al. 04, Gruhl et al. 04, Kumar et al. 03, Leskovec et al.]
  - Topics as probabilistic term mixtures: too coarse-grained [Blei-Lafferty 06, Wang-McCallum 06, Wang et al. 07]
  - Named entities: too coarse-grained

    Obama, McCain, Microsoft, Paris, Apple
  - Common sequence of words: too noisy

    "I love you", "web 2.0", "Oh my God", "Made in China"

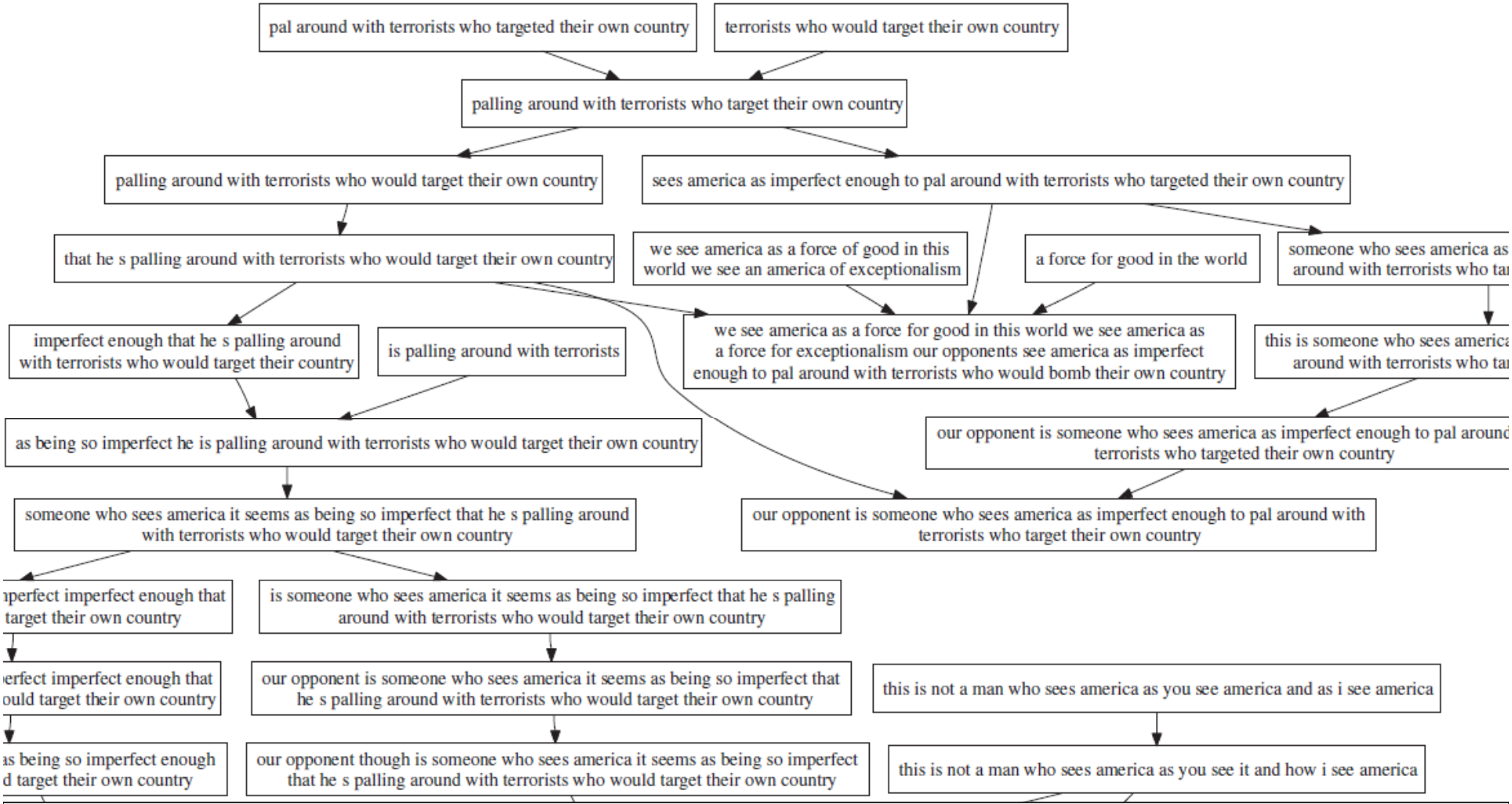
How to detect memes?

- Need units that:
  - correspond to aggregates of articles,
  - vary over the order of days,
  - and can be handled at terabyte scale
- **Plan:** identify text fragments, phrases, memes that travel relatively unchanged through many articles.
- **Idea: quoted phrases: ".*"**
  - 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

Online media

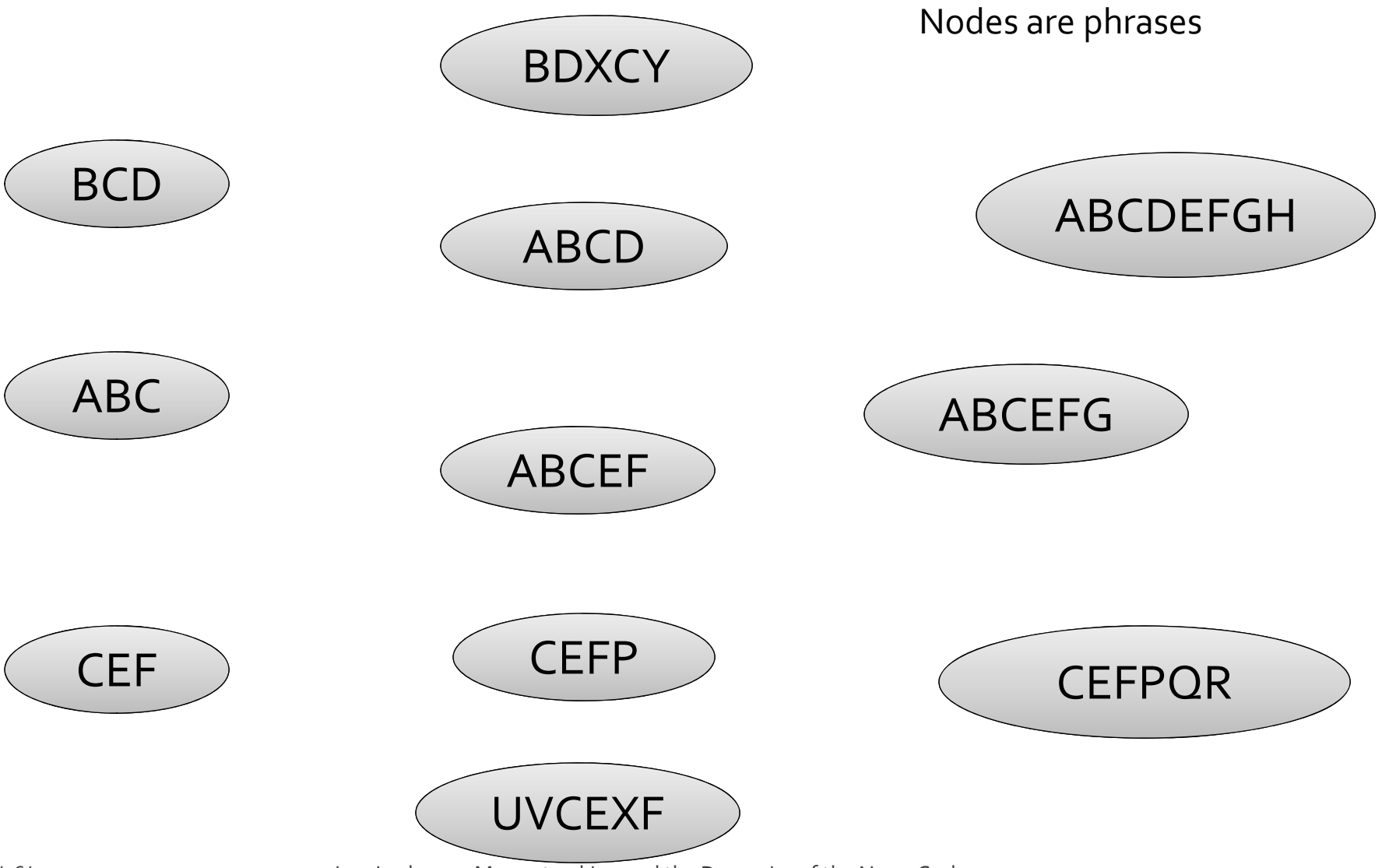
- Data from Spinn3r on the 3 months leading up to the 2008 U.S. Presidential Election:
  - 1 million news articles and blog posts per day
  - Essentially a complete online media coverage:
    - 20,000 sites that are part of Google News
    - 1.6 million blogs
  - From August 1 to October 31 2008
    - 90 million documents from 1.65 million sites, 390GB
  - We extract 112 million quotes (phrases)

Challenge: Phrases Mutate… A lot!

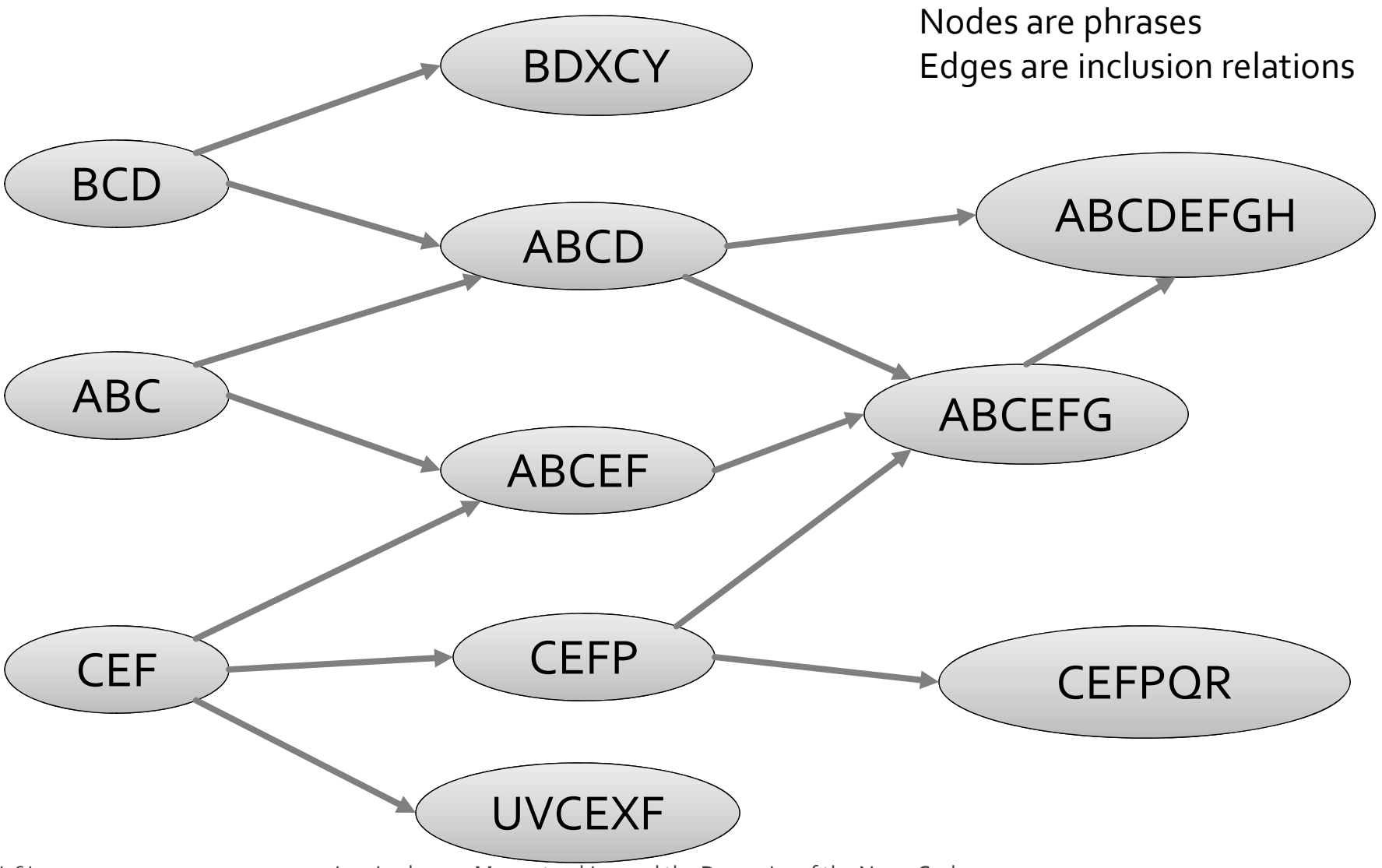


Phrase: Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that he's palling around with terrorists who would target their own country.

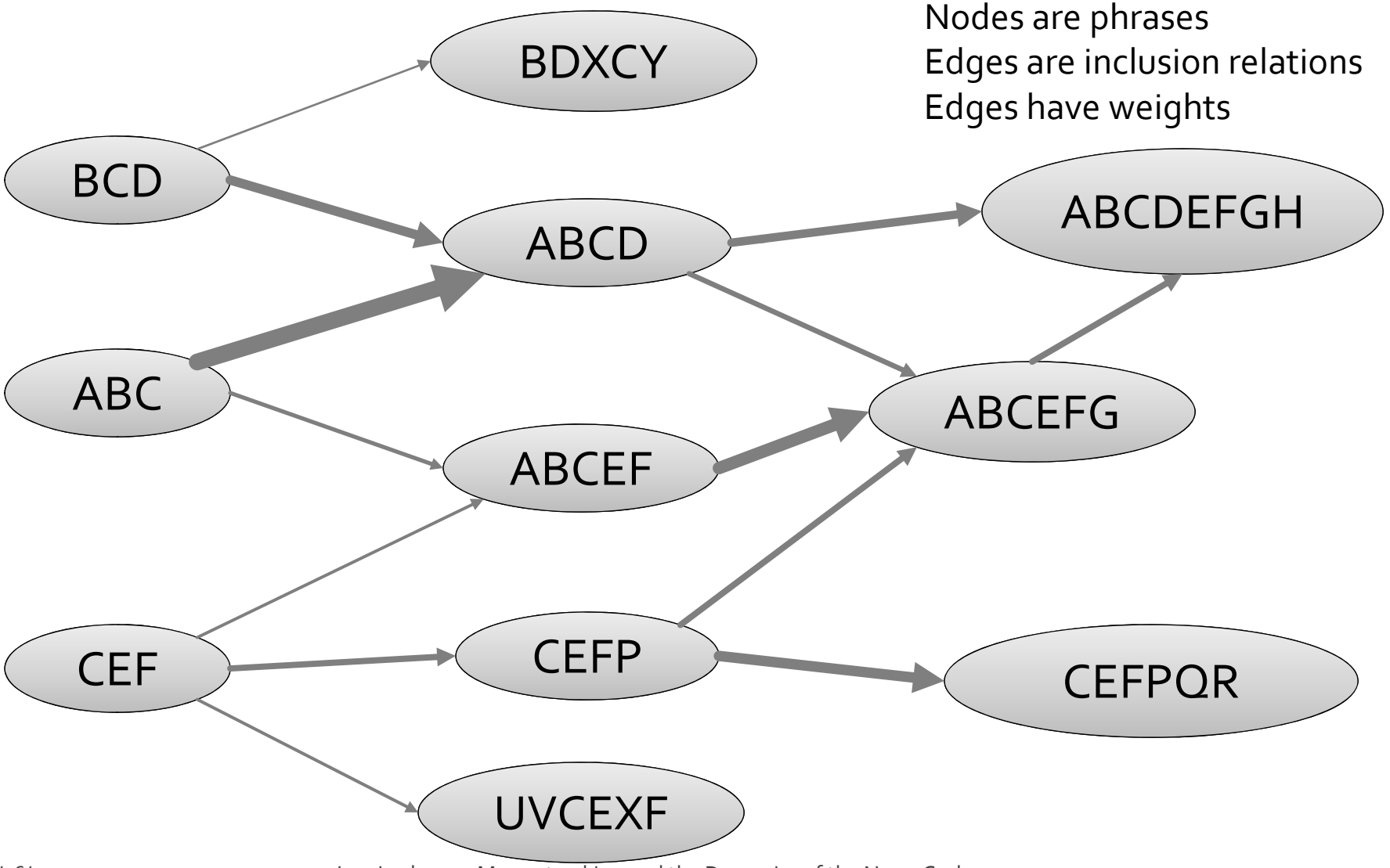
Creating clusters of Mutations



Creating clusters of Mutations

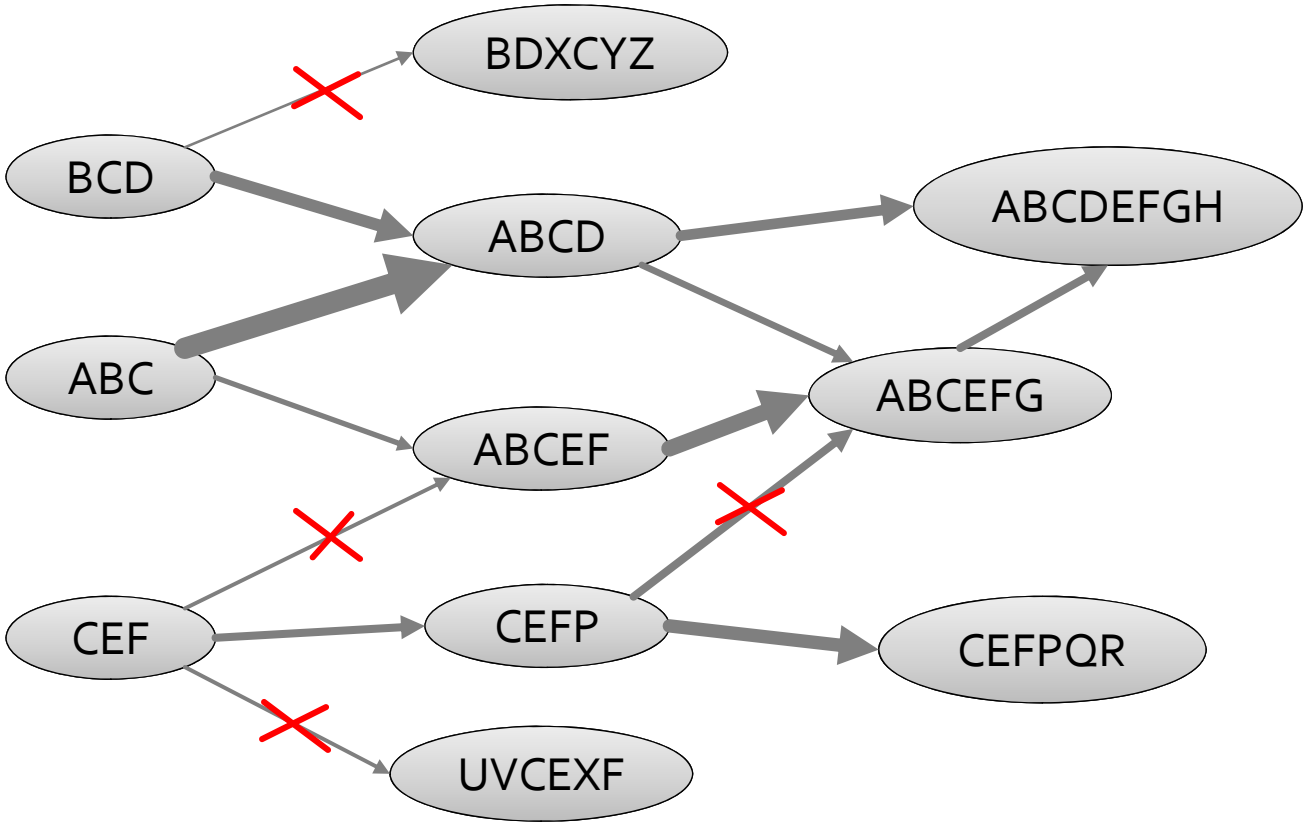


Creating clusters of Mutations



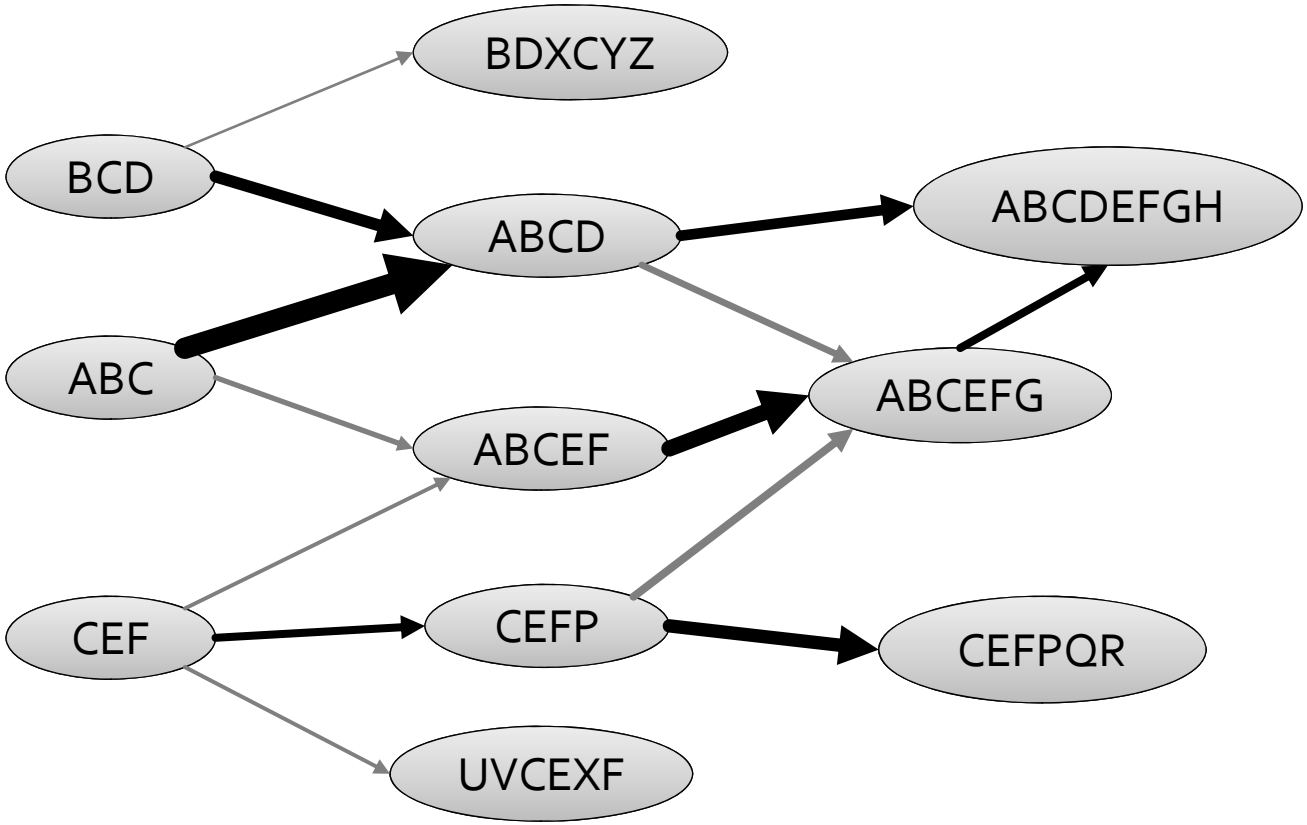
Phrase clustering: DAG partitioning

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



Phrase clustering: DAG partitioning

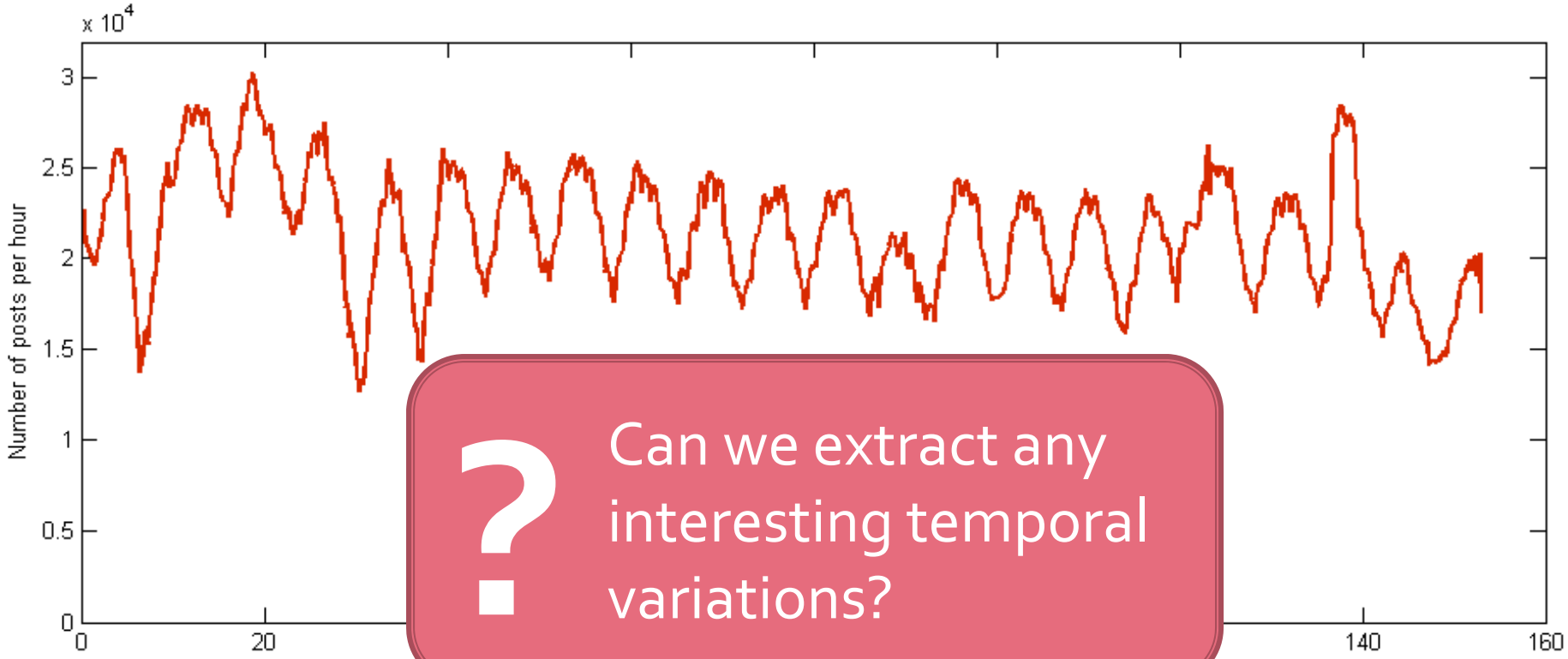
- **Observation:** enough to know node's parent
- **Heuristic:** proceed top down and assign node to strongest cluster



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 workin' families	50

Articles/phrases over time

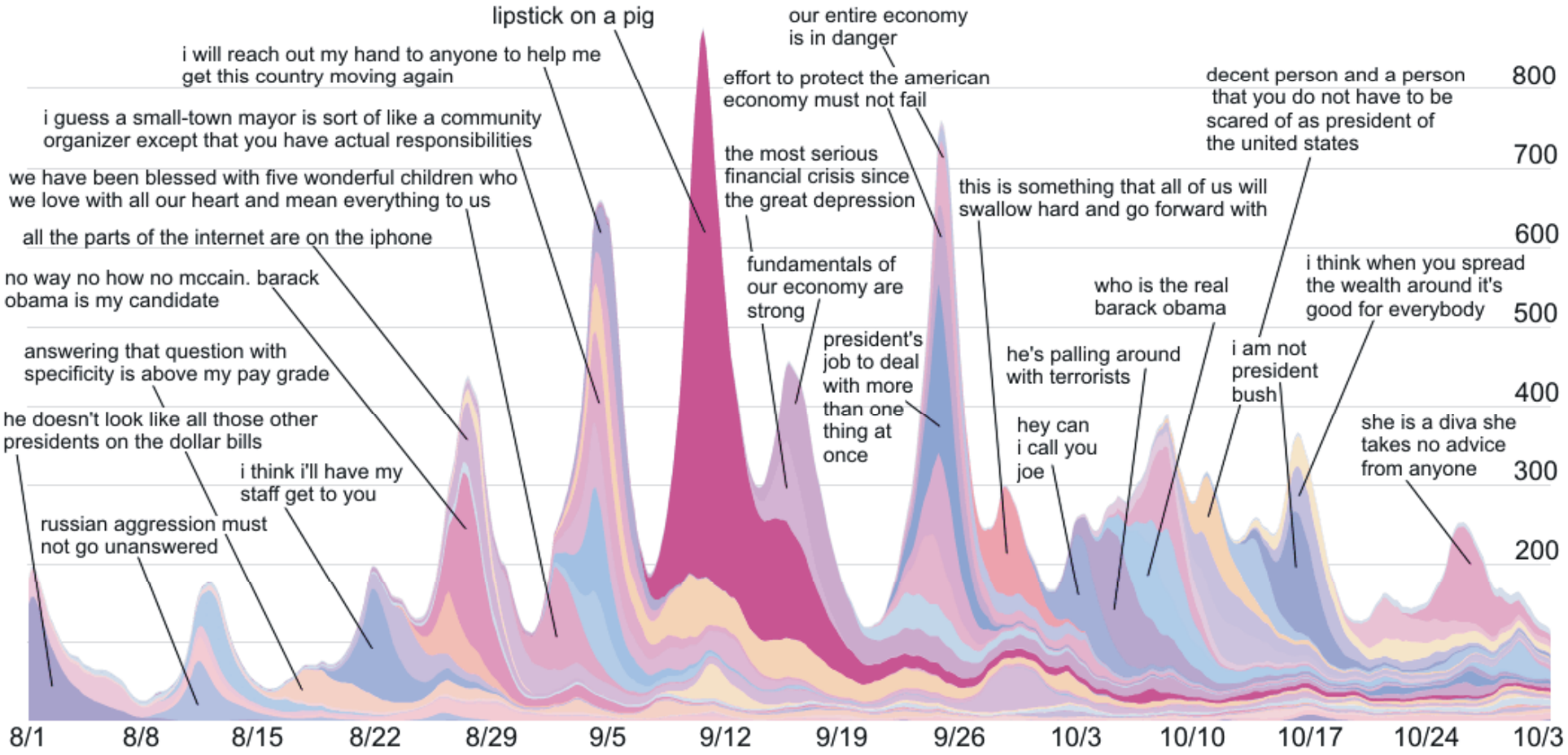


? Can we extract any interesting temporal variations?

… is periodic (weekly), has no trends.

The ”bandwidth” of the online media is constant

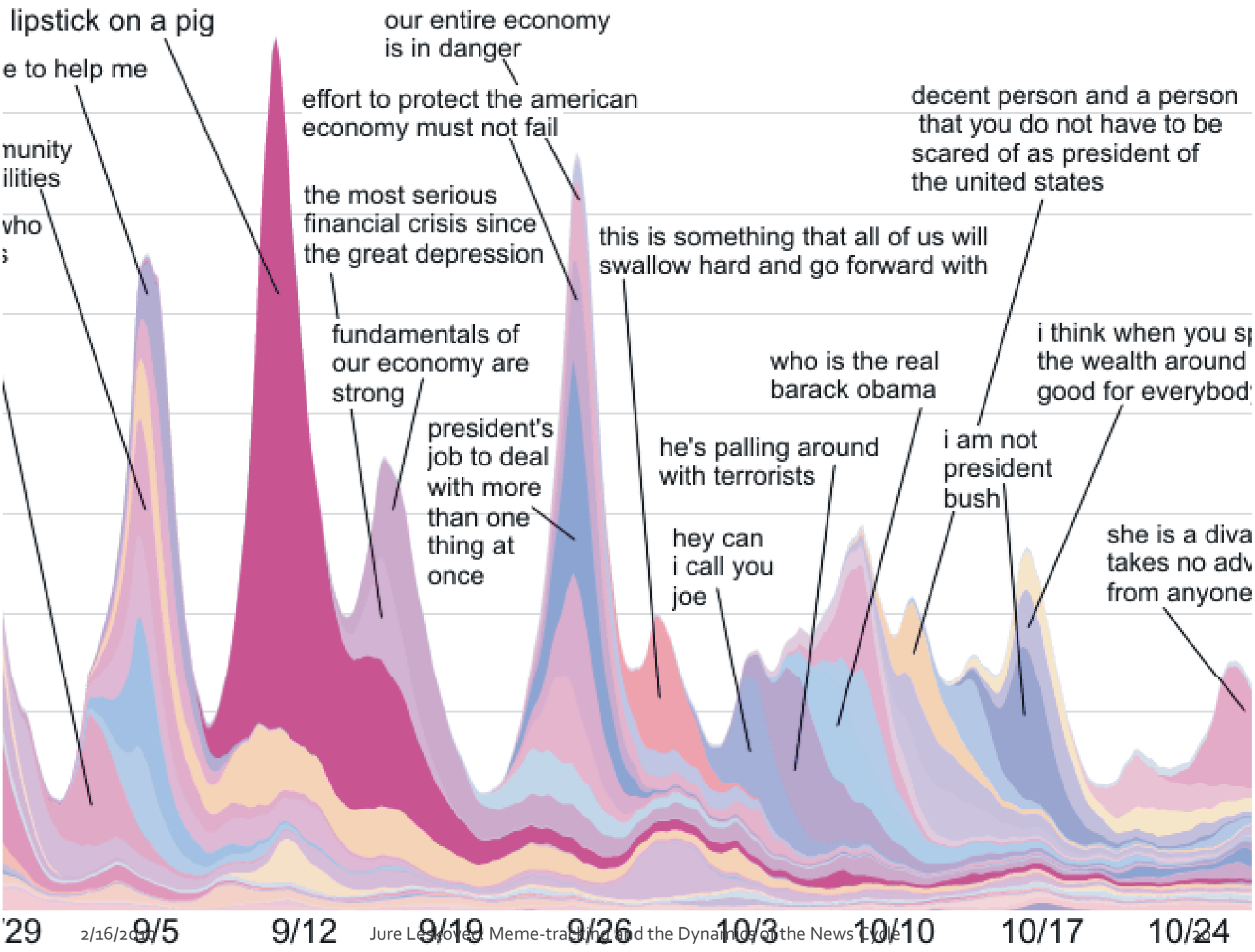
Cluster volume over time



August **October**

Volume over time of top 50 largest total volume phrase clusters

http://memetracker.org

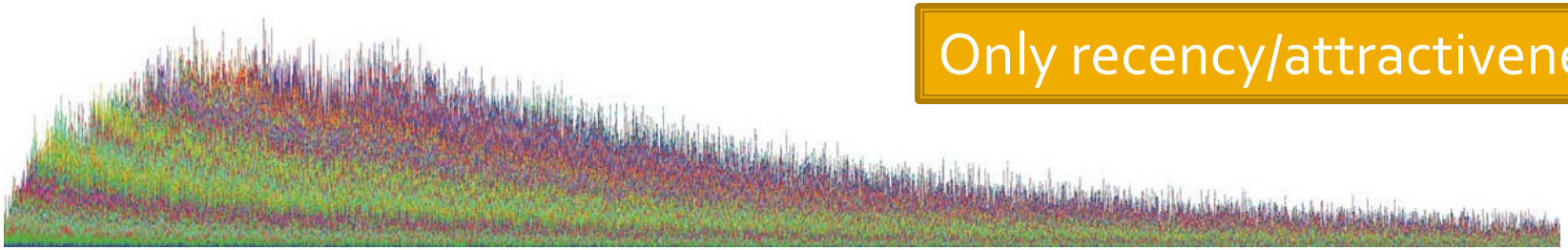

lipstick on a pig
e to help me
nunity
ilities
who
our entire economy
is in danger
effort to protect the american
economy must not fail
the most serious
financial crisis since
the great depression
fundamentals of
our economy are
strong
president's
job to deal
with more
than one
thing at
once
this is something that all of us will
swallow hard and go forward with
hey can
i call you
joe
he's palling around
with terrorists
who is the real
barack obama
decent person and a person
that you do not have to be
scared of as president of
the united states
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

Modeling the temporal variation

- What ingredients are essential to qualitatively reproduce the observed dynamics?
  - Temporal variation has potential connections with natural processes
    - Species competing for resources in an ecosystem.
    - Biological systems synchronize to favor small number of individuals [Lacker-Peskin 1981]
- *N* news sources, one new story per time step. Source's choice of what to cover controlled by:
  - Imitation: increasing in number of sources covering story
  - Recency: decreasing in time since story's appearance
  - Attractiveness: prefer more interesting stories

Modeling the temporal variation

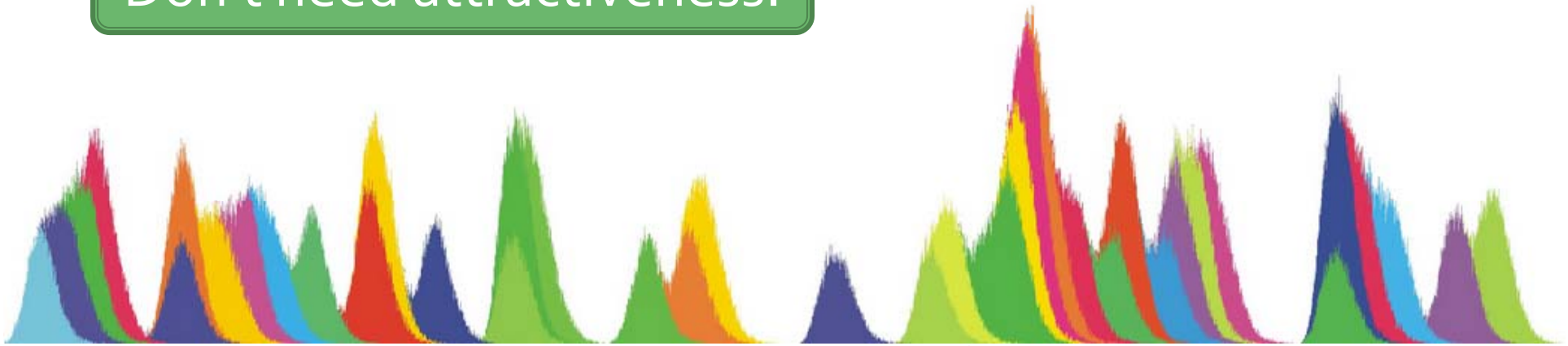
Only imitation



Only recency/attractiveness

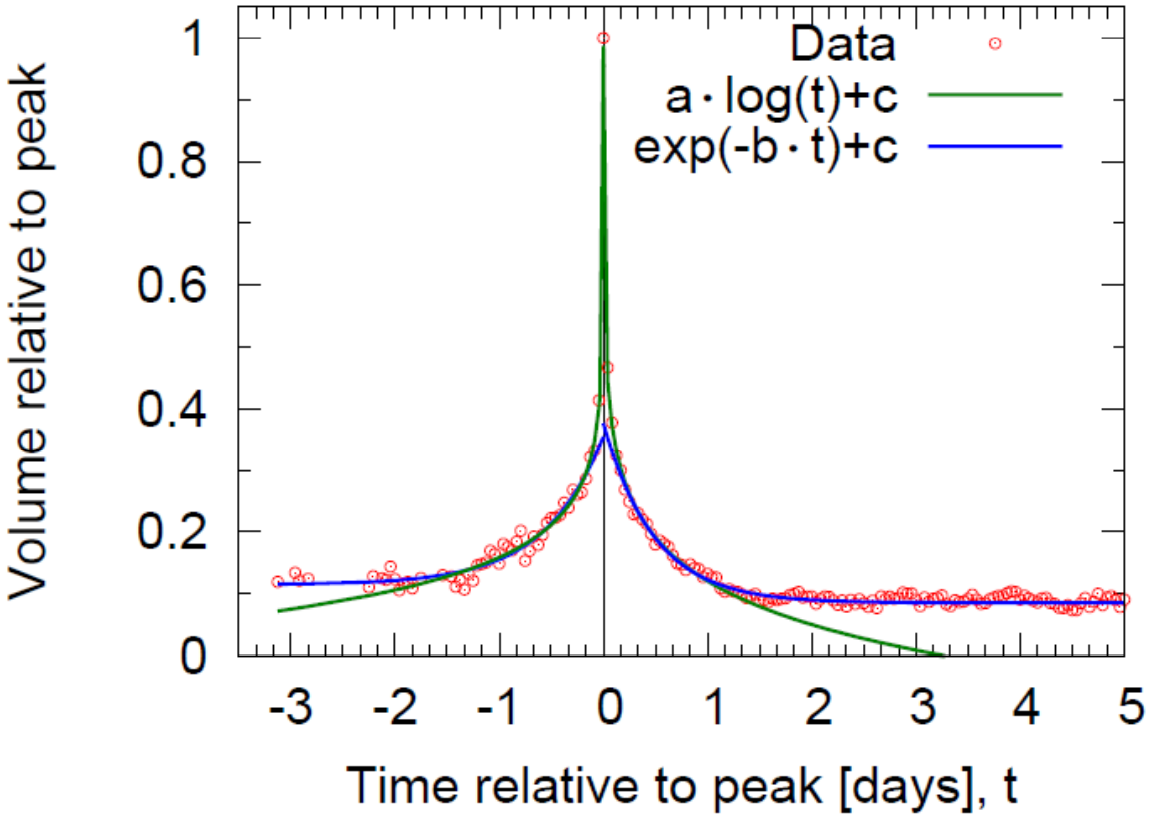
Don't need attractiveness!

Imitation & Recency

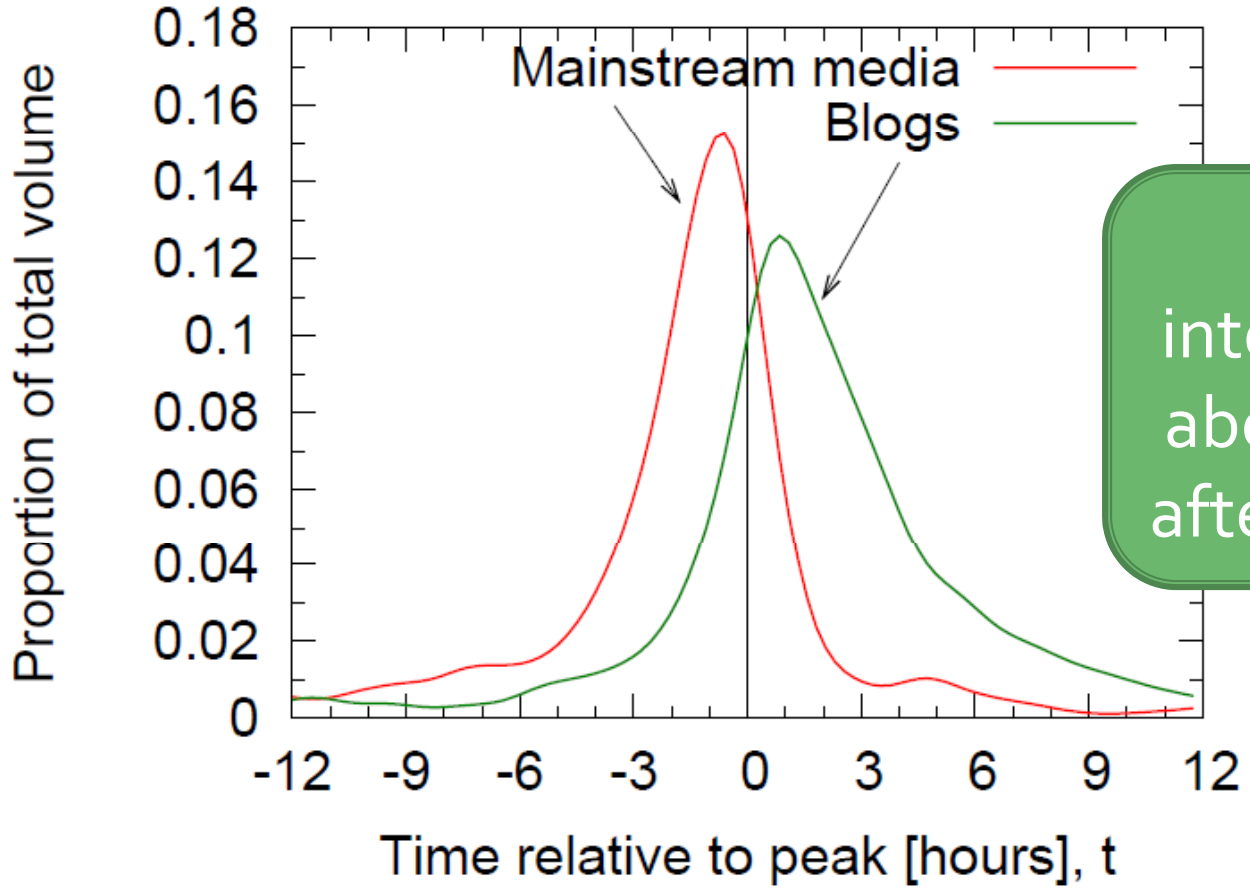


Interaction of News and Blogs

- Can study typical phrase cluster volume curve
- Peak behaves like a delta function (infinity at t=0)
- 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.)

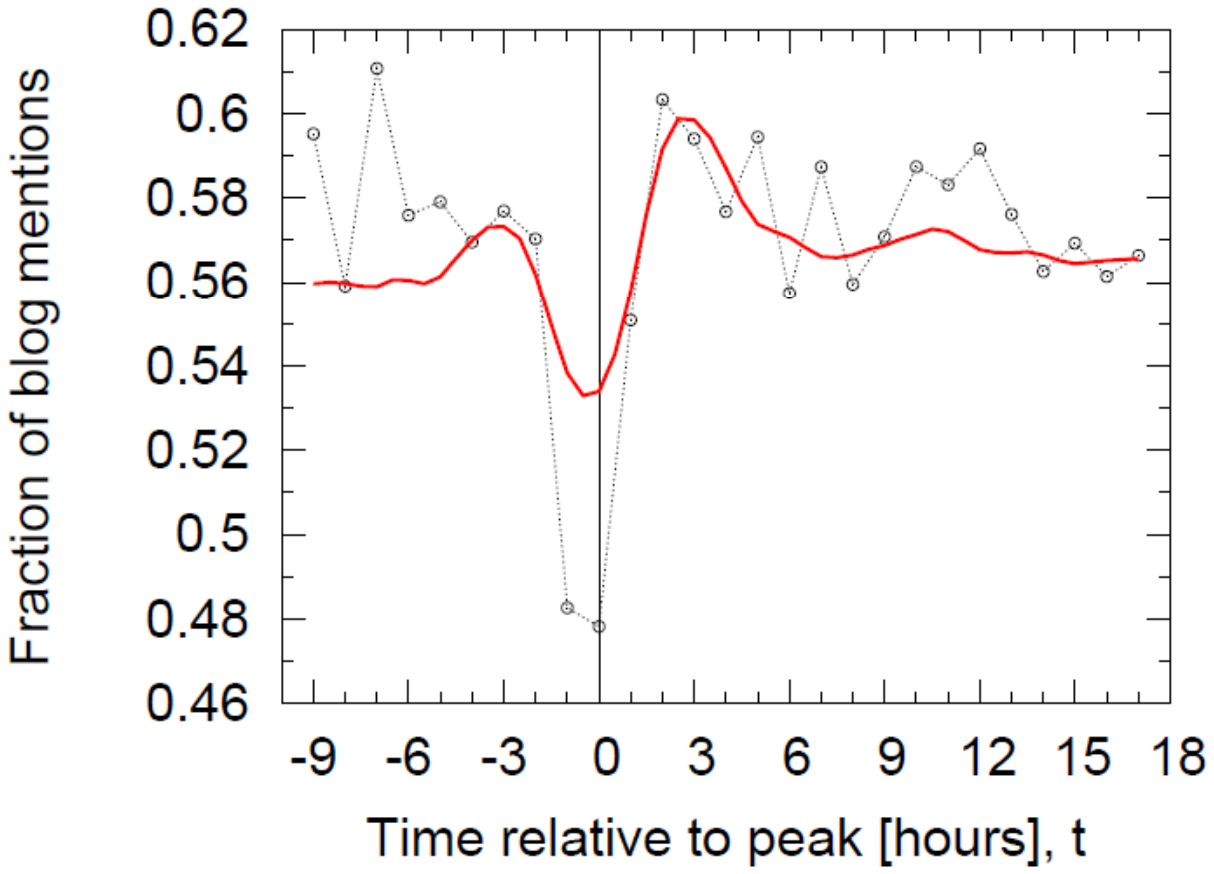
How quickly sites mention quotes?

- Can 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

- Can study "oscillation" of attention between news and media



Stories catalyzed by blogs

- Can formulate queries for different temporal "signatures": e.g., stories catalyzed by blogs:

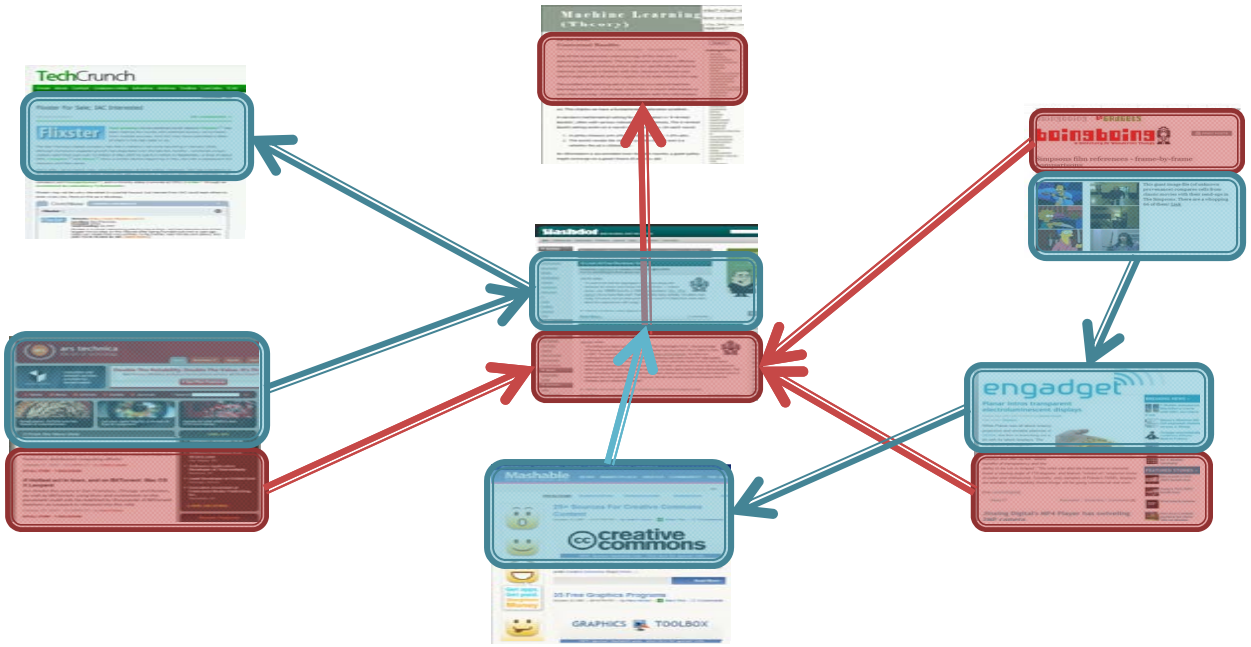
  [*x*; *y*; *t*]-query: between *x* and *y* frac. of total phrase volume (f_b) occurred on blogs at least *t* days before overall the peak

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 about 3.5% of phrases migrate from blogs to media

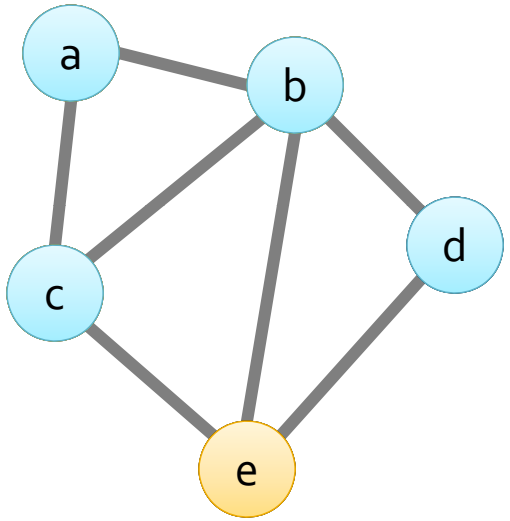
Information Propagation

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



Inferring the Diffusion Network

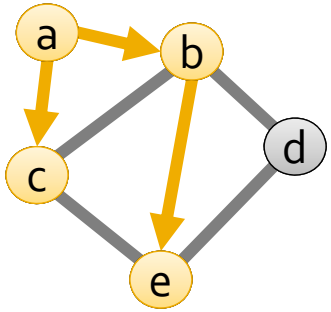
- There is a hidden diffusion network:



- 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-influences-whom network

Inferring the Diffusion Network

Given a cascade *c*:



- Probability of propagation

$$P_c(i,j) \propto e^{-\Delta t}$$

- Prob. cascade c that propagates in pattern T

$$P(c|T) \propto \prod_{(i,j)\in T} P_c(i,j)$$

- But we do not know the propagation tree T thus need to consider all trees

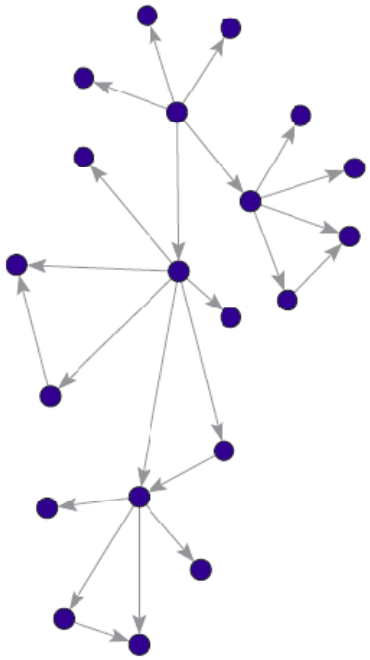
$$P(c|G) = \sum_{T\in\mathcal{T}(G)} P(c|T)P(T|G)$$

Finding the Influence Graph

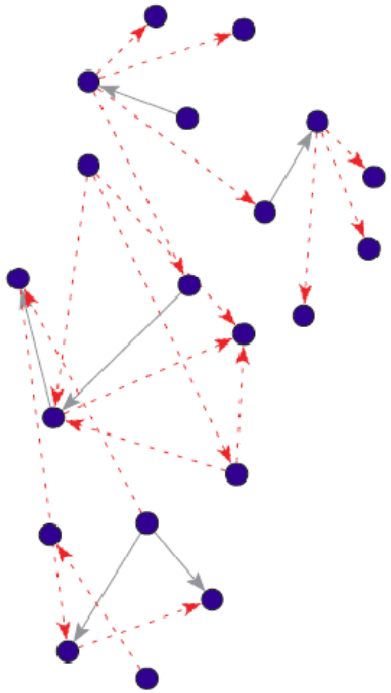
- We want: $\hat{G} = \underset{|G| \leq k}{\operatorname{argmax}} P(C|G)$
- Computing the P(C|G) is intractable
  - Need to consider all possible propagation patterns
  - Apply the Matrix tree theorem ($O(n^3)$)
- How to maximize over P(C|G)?
  - Theorem: P(C|G) is submodular
    - Diminishing returns
  - We can find near optimal G

Synthetic example

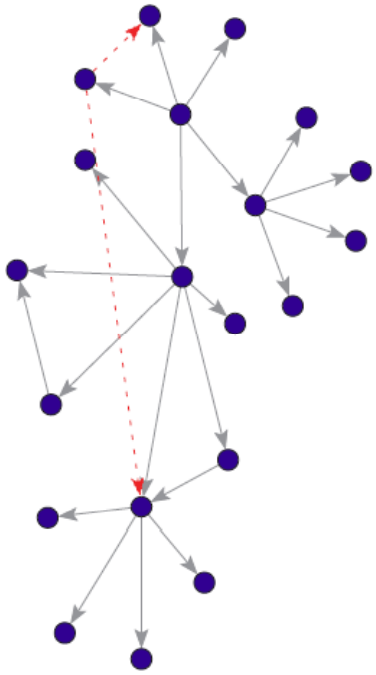
- Small synthetic network:



True network



Baseline network

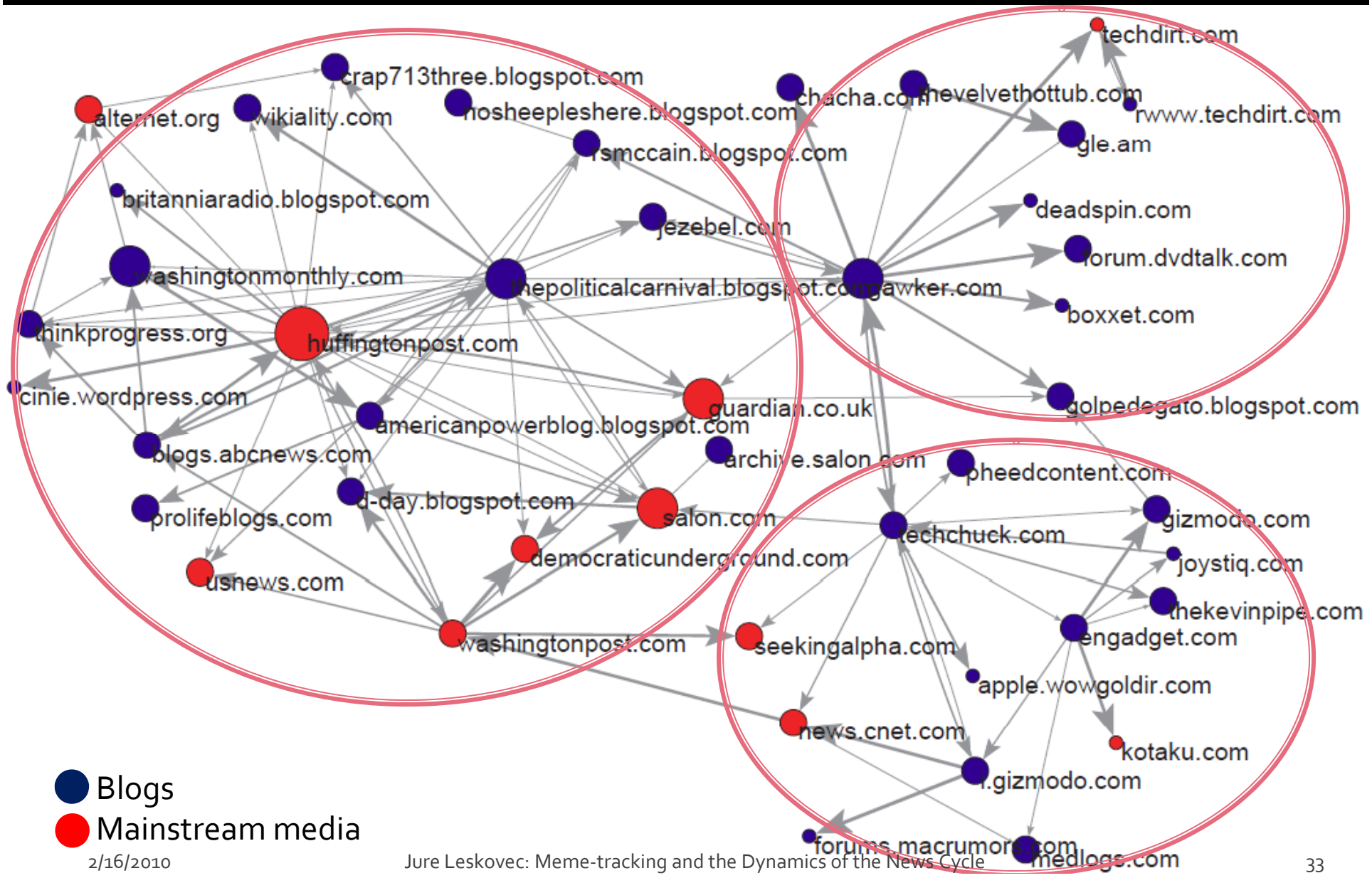


Our method

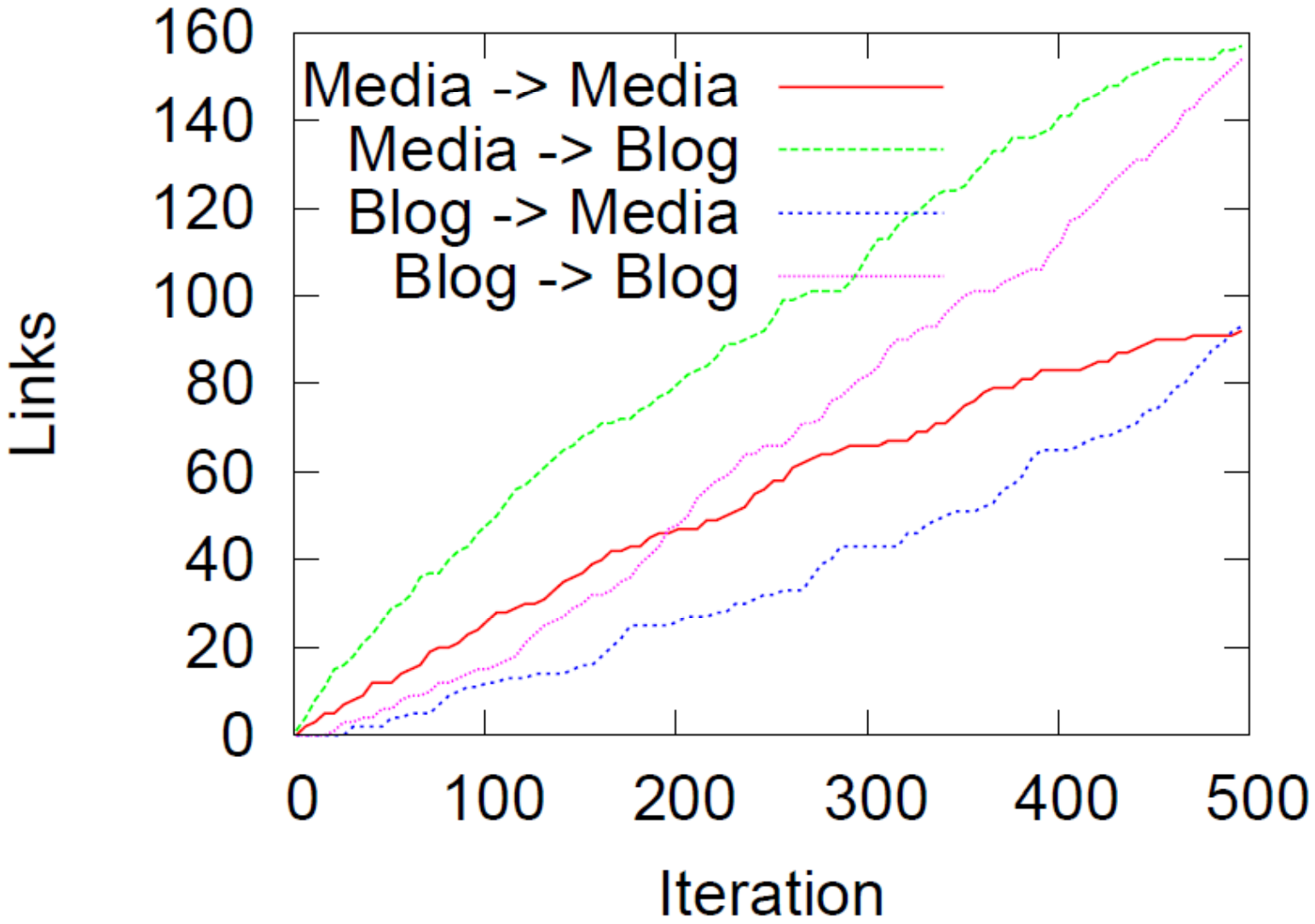
Pick strongest edges

$$w(u,v) = \sum_{c \in C} P_c(u,v)$$

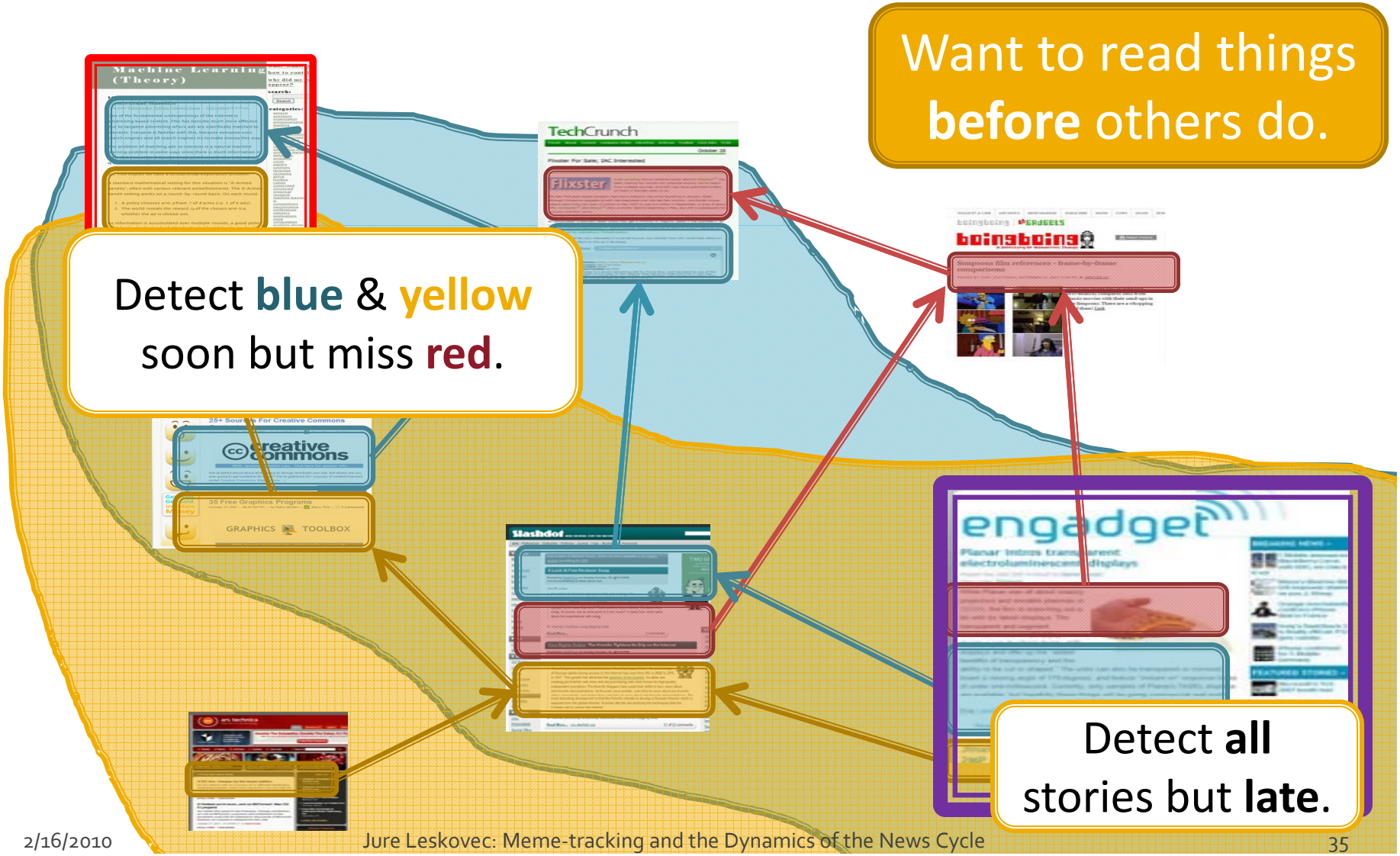
Diffusion network (small part)



Link types by strength

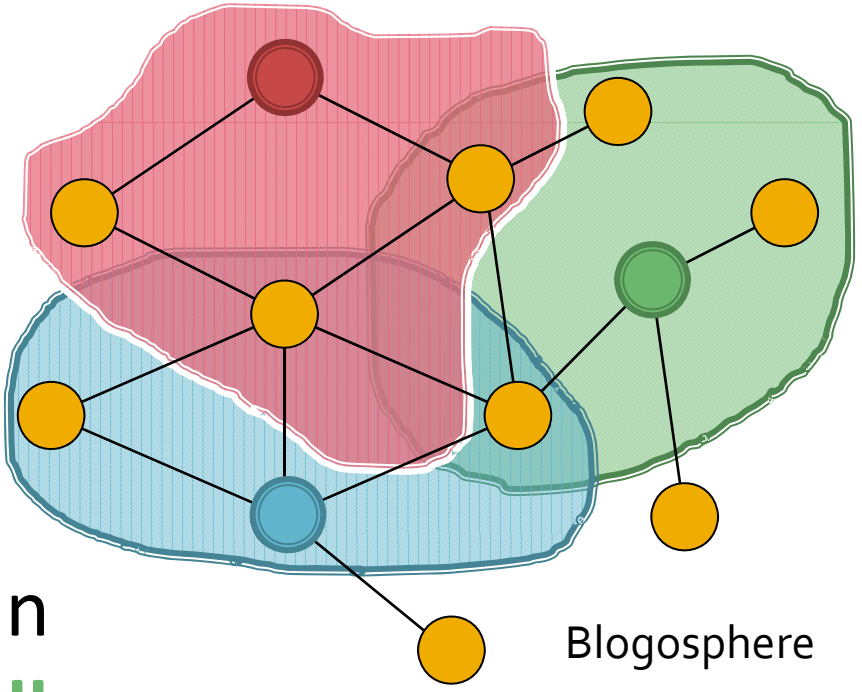


Detecting information outbreaks



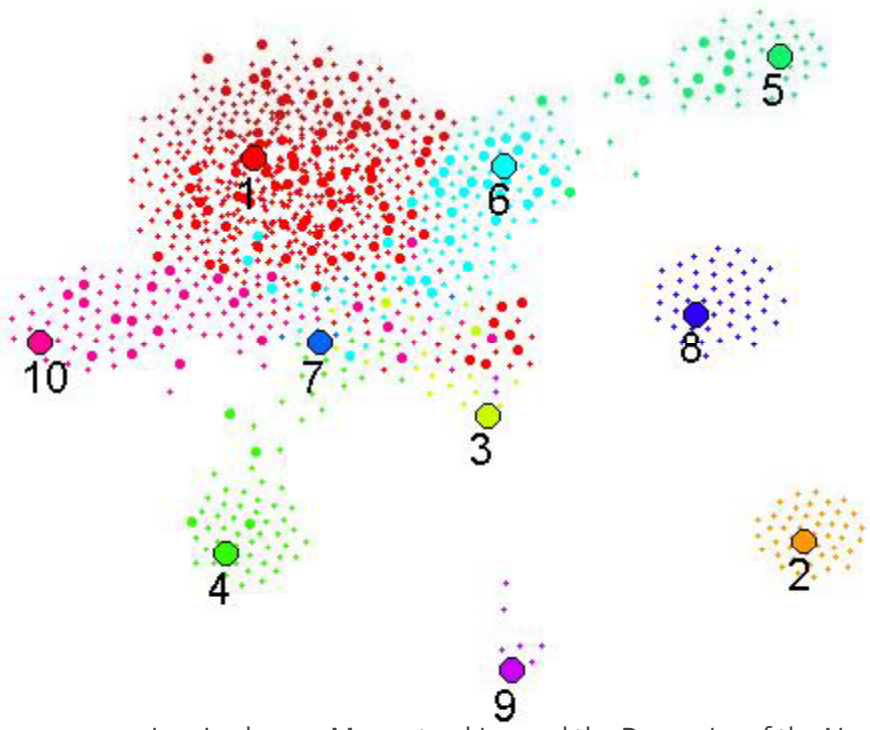
Problem: Covering stories

- Given a budget (e.g., of 3 blogs)
- Select sites to cover the most of the Web
- Bad news: Solving this exactly is **NP-hard**
- Good news: **Theorem**: Our algorithm can do it in **linear time near-optimally**



Blogs: Information epidemics

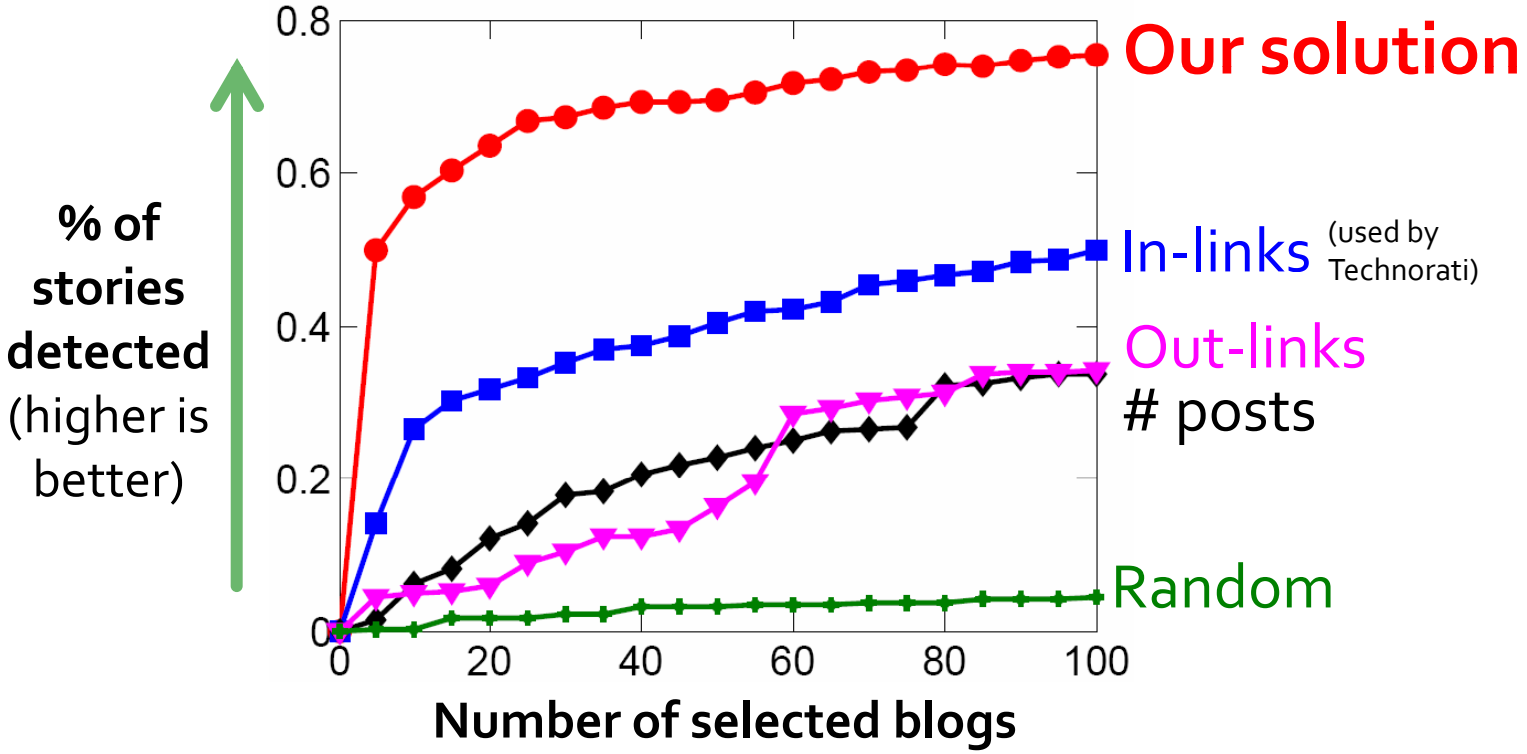
- **Question:** Which websites should one read to catch big stories?
- Idea: Each blog covers part of the Web



• Each dot is a blog
• Proximity is based on the number of common cascades

Experimental results

Which blogs to read to be most up to date?



www.blogcascades.org

So, who was influential (in 2006)?

k	Score	Blog	Posts	InLinks	OutLinks
1	0.13	http://instapundit.com	4593	4636	5255
2	0.18	http://donsurber.blogspot.com	1534	1206	3495
3	0.22	http://sciencepolitics.blogspot.com	924	576	2701
4	0.26	http://www.watcherofweasels.com	261	941	3630
5	0.29	http://michellemalkin.com	1839	12642	6323
6	0.32	http://blogometer.nationaljournal.com	189	2313	9272
7	0.34	http://themodulator.org	475	717	4944
8	0.35	http://www.bloggersblog.com	895	247	10201
9	0.37	http://www.boingboing.net	5776	6337	6183
10	0.38	http://atrios.blogspot.com	4682	3205	3102
11	0.39	http://lawhawk.blogspot.com	1862	463	6597
12	0.40	http://www.gothamist.com	6223	3324	17172
13	0.41	http://mparent7777.livejournal.com	25925	199	47933
14	0.42	http://wheelgun.blogspot.com	1174	128	939
15	0.43	http://gevkaffeegal.typepad.com/the_alliance	302	428	2481

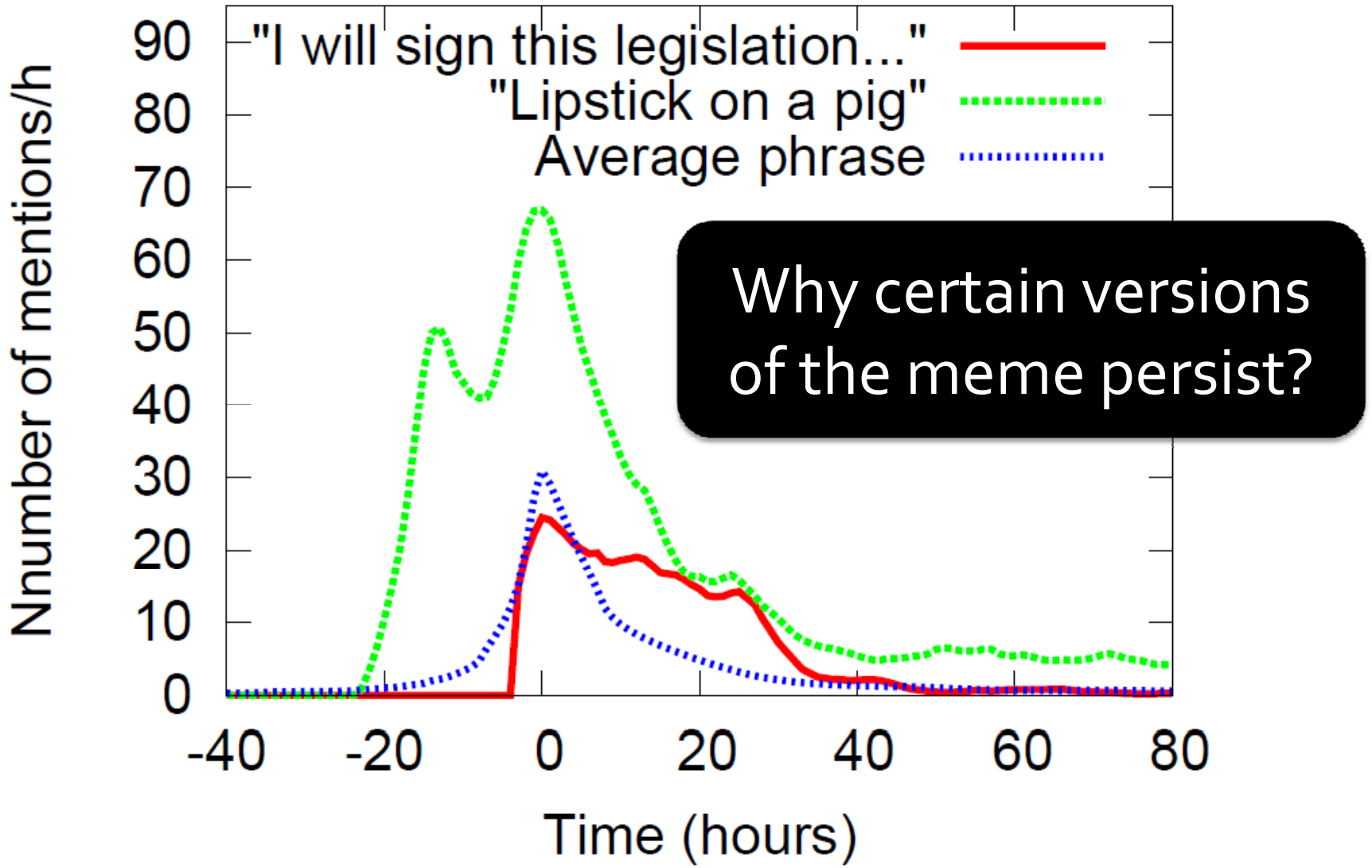
Conclusion & Further questions

- A framework for tracking memes through the news, to quantify the dynamics of the news cycle.
- Demo + Data:

  http://memetracker.org

- Many further questions:
  - Which elements of the news cycle do we miss?
  - Can this analysis of memes help identify dynamics of polarization? (cf. [Adamic-Glance, 2005])
  - How are these memes actually spreading among people?

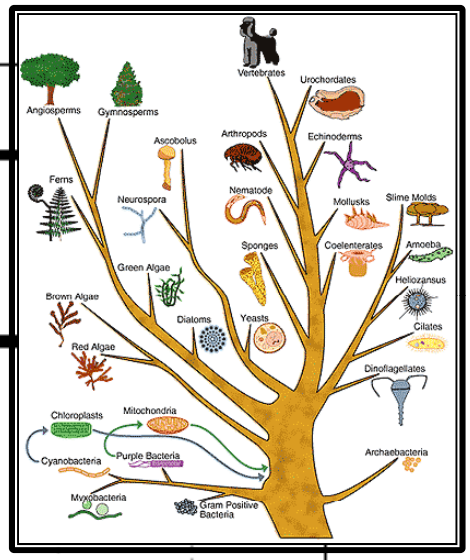
Question 1: Persistence

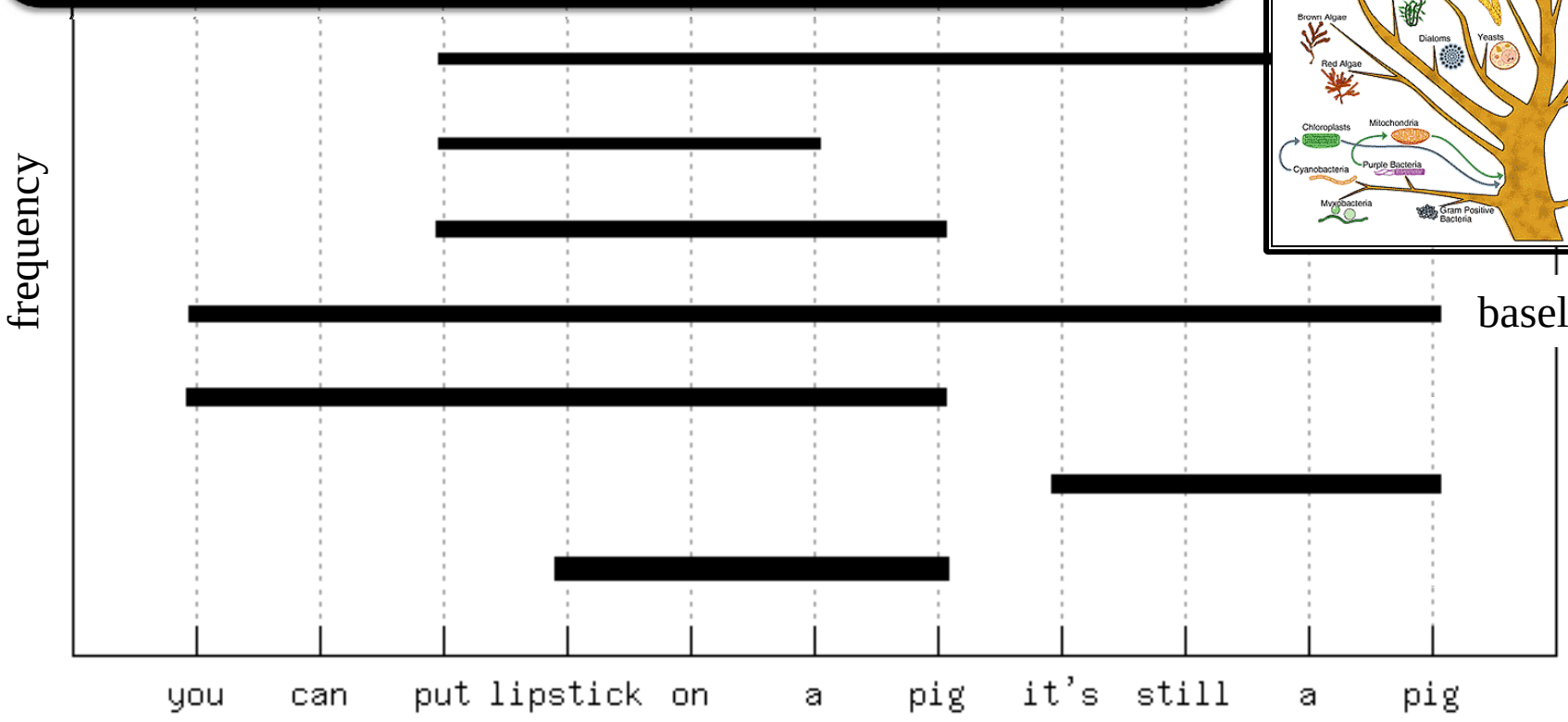


Why certain versions of the meme persist?

Questions 2: Mutation

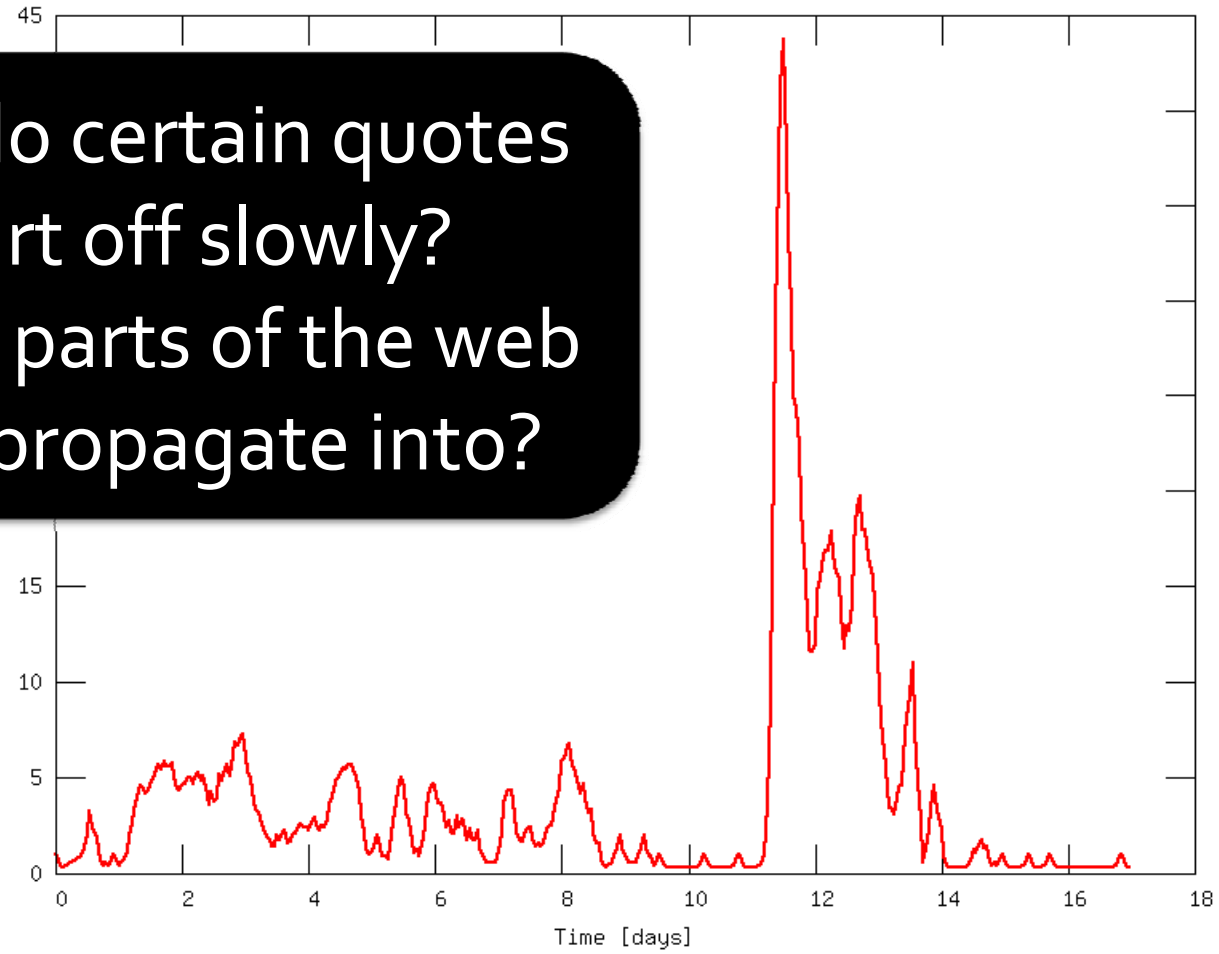
What is a **phylogenetic tree** of information?





Question 3: Trend prediction

Why do certain quotes start off slowly?
Which parts of the web they propagate into?



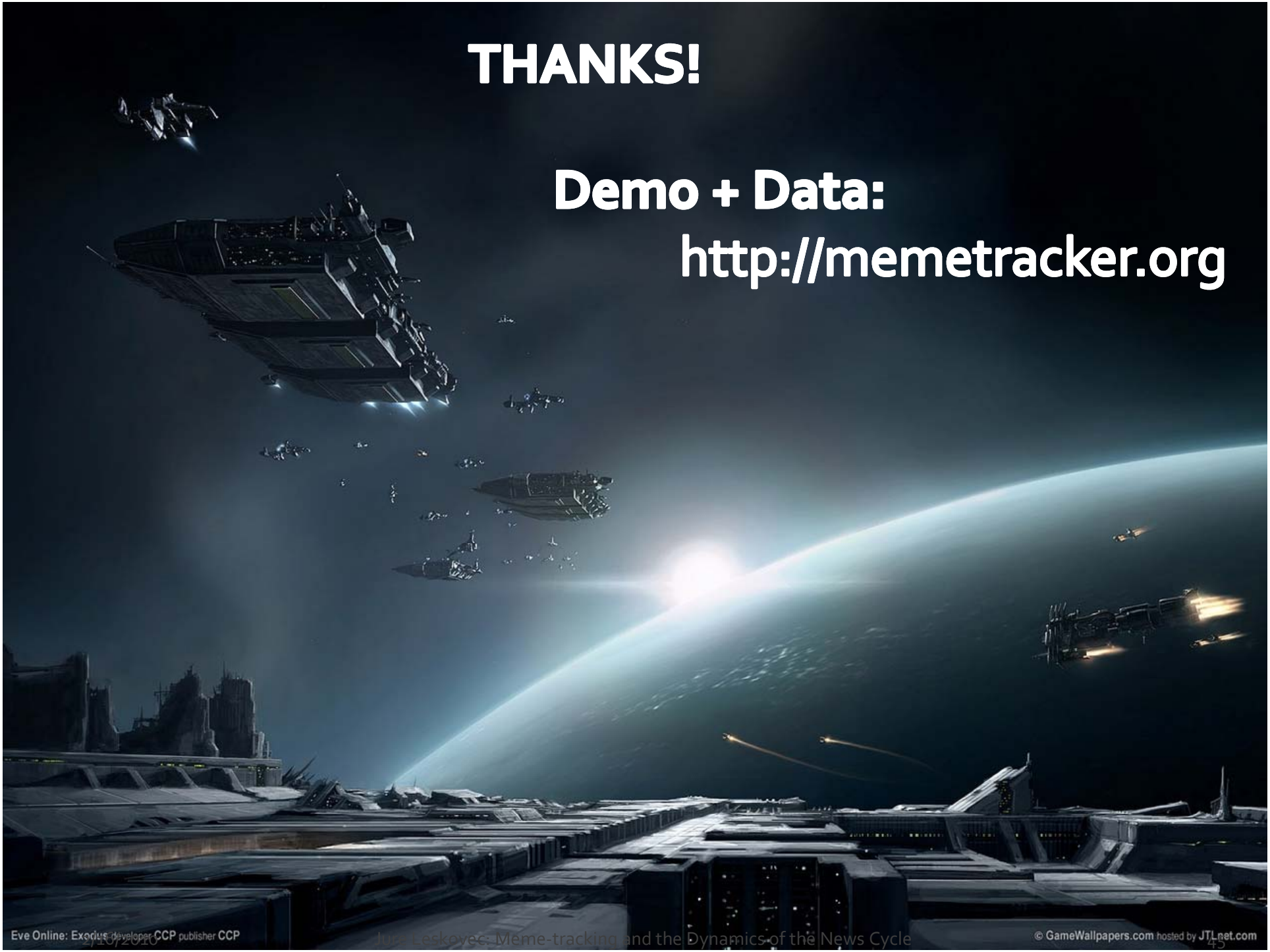
"A changing environment will affect Alaska more than any other state because of our location I'm not one though who would attribute it to being man-made"

References

- *Meme-tracking and the Dynamics of the News Cycle*, by J. Leskovec, L. Backstrom, J. Kleinberg. KDD, 2009
  http://cs.stanford.edu/people/jure/pubs/quotes-kdd09.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/people/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


THANKS!
Demo + Data:
http://memetracker.org
Eve Online: Exodus developer CCP publisher CCP
© GameWallpapers.com hosted by JTLnet.com