Meme-tracking and the Dynamics of the News Cycle

Leskovec, Backstrom, and Kleinberg

The Dichotomy of Influence

- Tension exists between the global influence of mass media and the local influence of social structures.
- How does information transmitted by the mass media interact with the inherent personal influence of social networks?

The rise of the 24-hour News Cycle

- Social media means the dichotomy between global and local influence is evaporating.
- The speed of media reporting and discussion has intensified, leading to a rapid progression of stories with no pauses.
- These factors have led to the creation of the 24-hour News Cycle.

24-hour News Cycle

- A succession of story lines that evolve and compete for attention within a relatively stable set of broader topics produce a dynamic effect called the news cycle.
- What is another word for these competing story lines?

Memes

- A meme is a unit of cultural ideas, symbols or practices, which is transmitted from one mind to another.
- Memes are regarded as cultural analogues to genes, in that they self-replicate, compete, and respond to selective pressures.

Important Questions...

- How do memes travel between blogs and mass media outlets?
- How do memes grow and decay?
- Can we model the dynamics of the 24-hour news cycle?

What this paper offers

- Methods for identification of unique memes in on-line text (meme clustering)
- Analysis of memetic interplay between blogs and mass media outlets (global meme analysis)
- Analysis of meme growth and decay over some finite lifespan (local meme analysis)
- Introduction of a model for representing the dynamics of a news cycle (global meme modeling)

Data set

- 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)

Terminology

item - a blog post or news article

phrase - a quoted string that occurs in one or more items

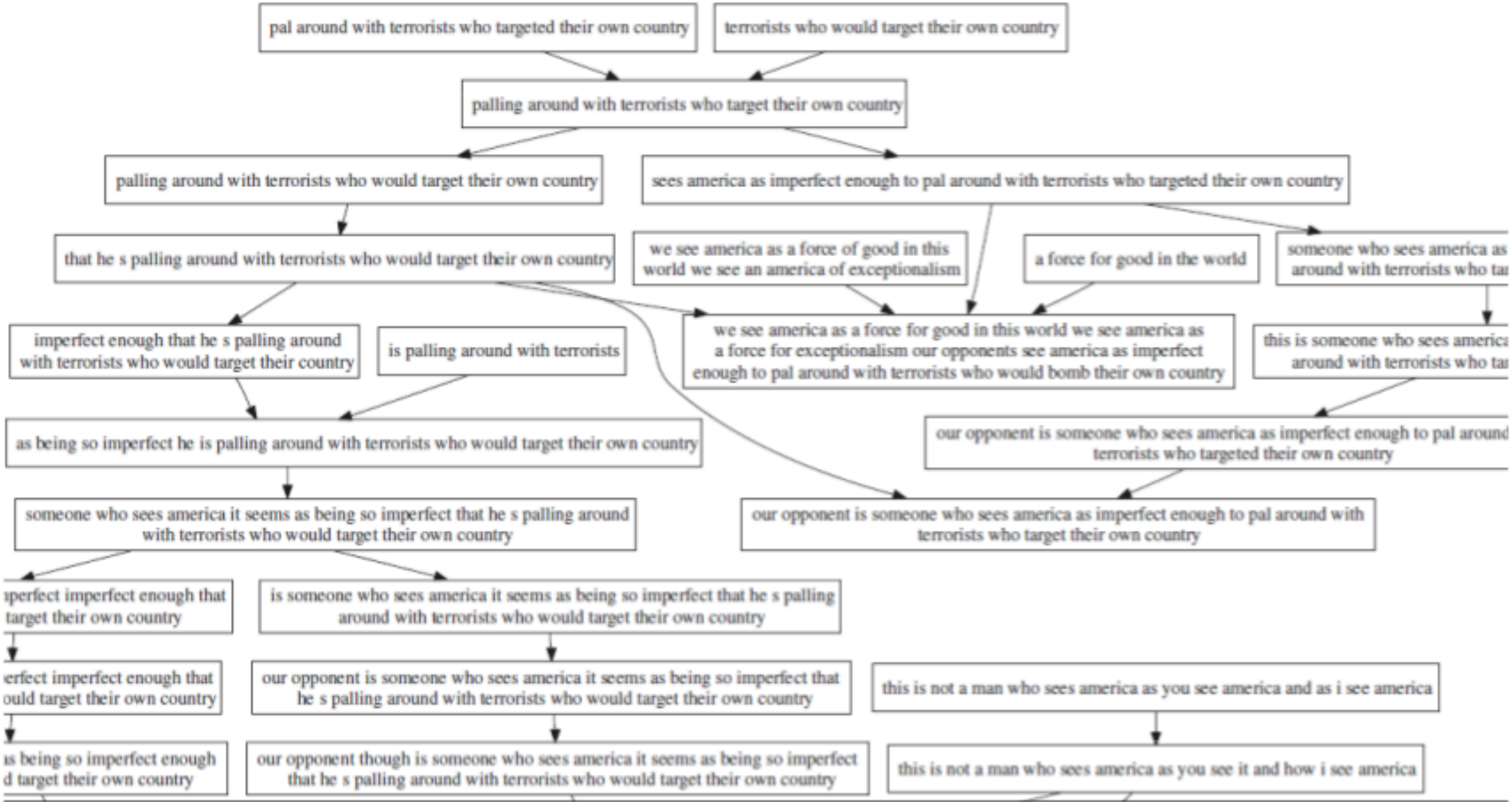
phrase graph - a graph where each node is a phrase variant, and directed edges connect related phrases

phrase cluster - collections of phrases deemed to be close textual variants of one another

Memetic Mutations

- Memes may undergo slight mutations over time; to recognize a meme is to recognize a cluster of all of its mutational variants.

Variations in Phrases



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

perfect 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

s 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

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.

Meme Clustering

Purpose: To produce phrase clusters which represent a discrete meme and all of its mutational variants so the meme can be tracked in on-line text.

This is a two-step process:

(1) Given a meme, build a phrase graph.

(2) Partition this graph into subgraphs of phrase clusters.

Phrase Graph

Pre-processing

lower bound L on word-length of phrases

lower bound M on the frequency of a phrase within the corpus

upper bound ϵ on the fraction of all occurrences that some phrase P occupies within any single item (to eliminate spam)

In this application:

$\epsilon = 0.25$, $L = 4$, $M = 10$

Phrase Graph

Building the graph:

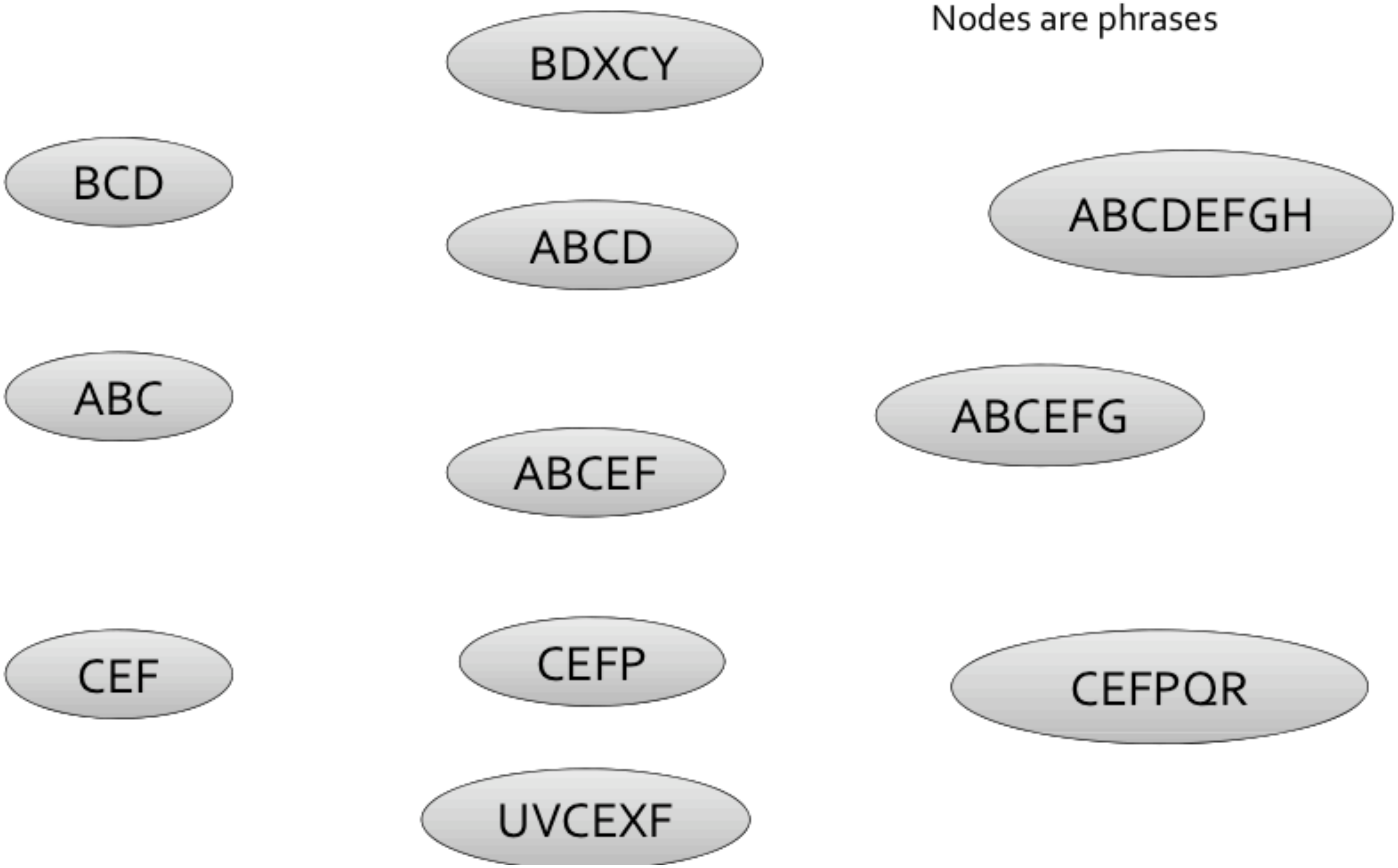
G is constructed on the set of quoted phrases

Vertices represent each variation of a phrase

Edges (p,q) are included for every pair of nodes p, q such that p is strictly shorter than q, or there is at least a k-word consecutive overlap between p and q

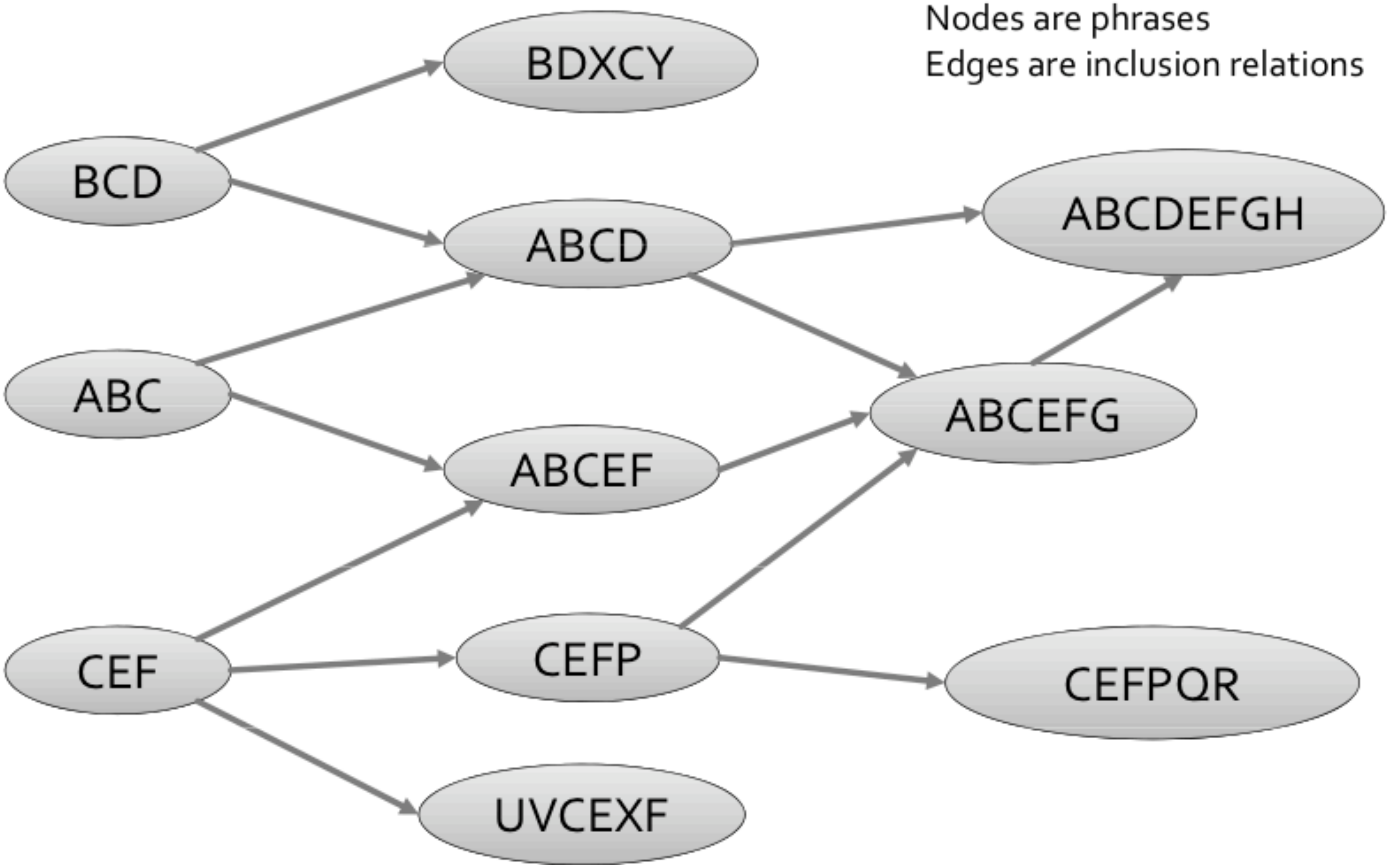
For each edge, a weight w is specified which decreases in the distance from p to q, and increases in the frequency of q in the corpus.

Phrase Graph



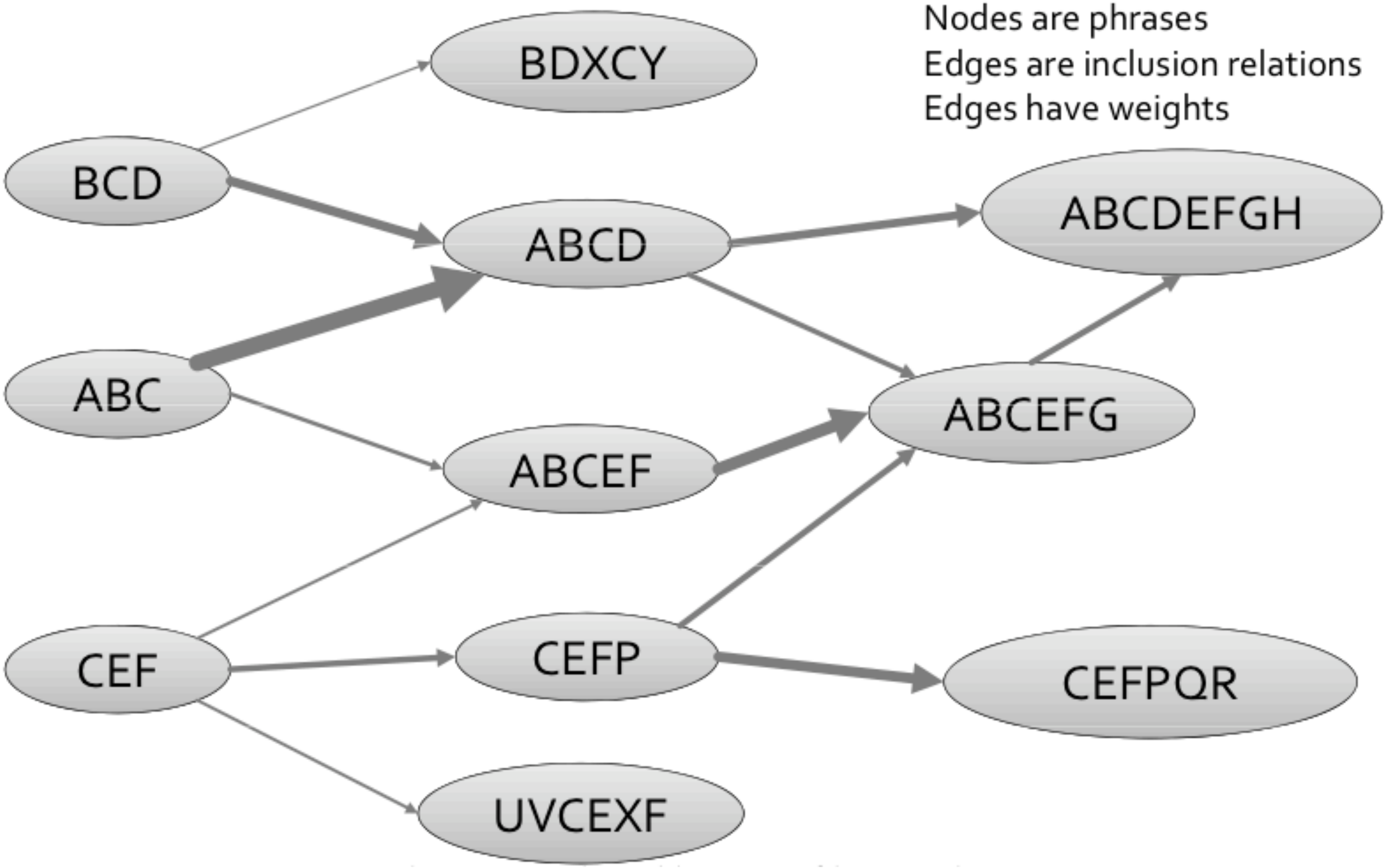
Nodes are phrases

Phrase Graph



Nodes are phrases
Edges are inclusion relations

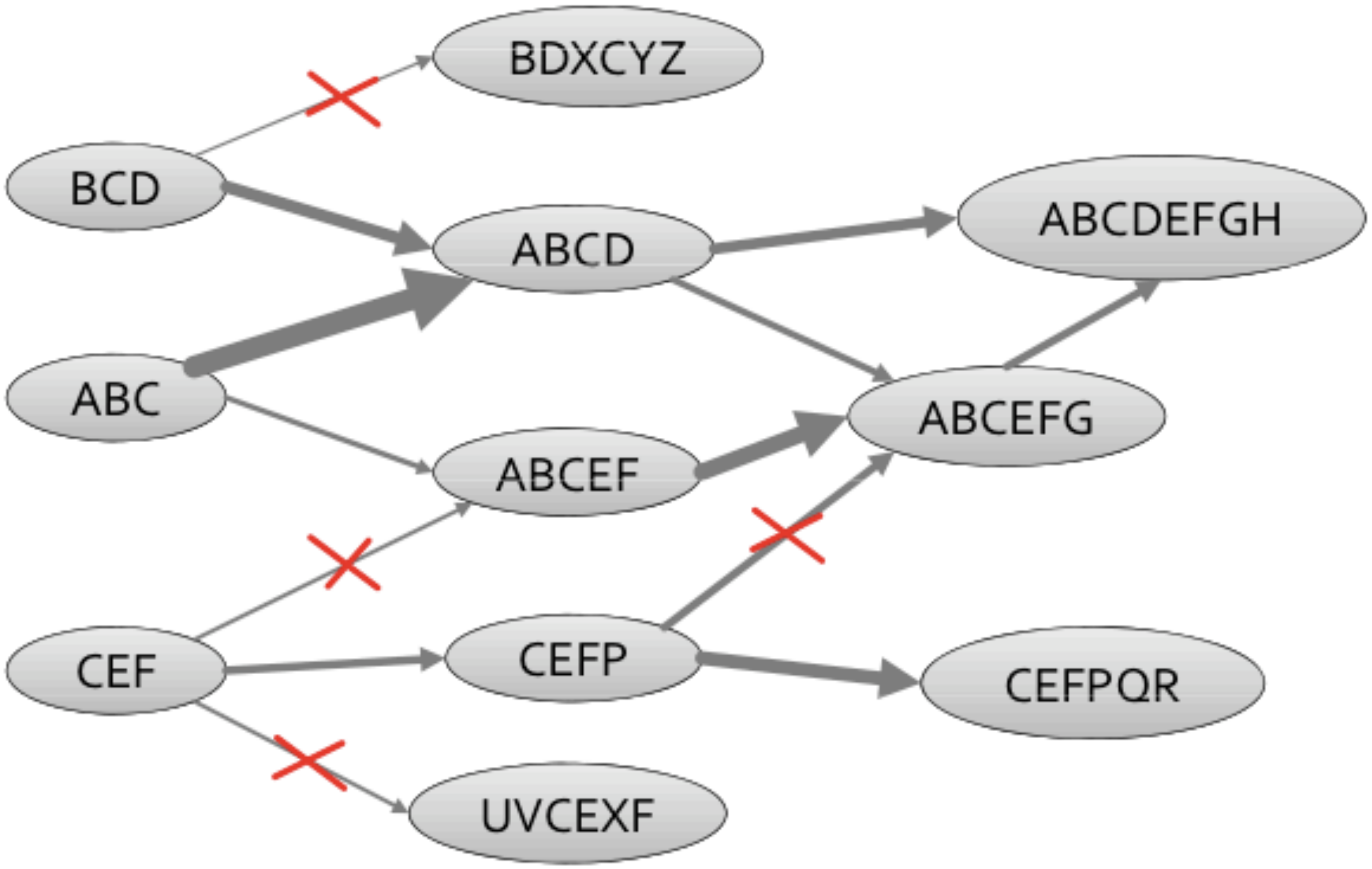
Phrase Graph



Nodes are phrases
Edges are inclusion relations
Edges have weights

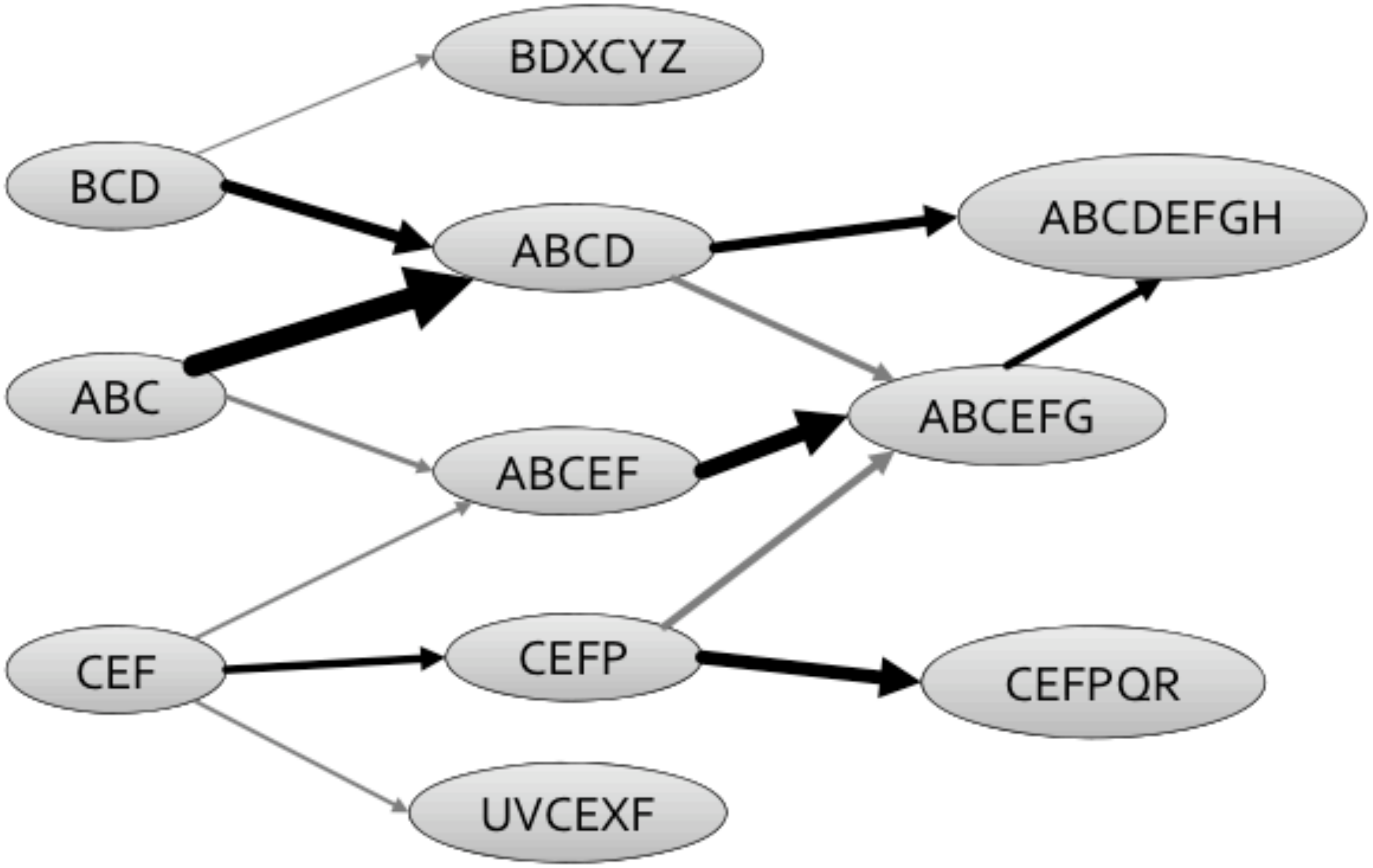
Phrase Partitioning

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

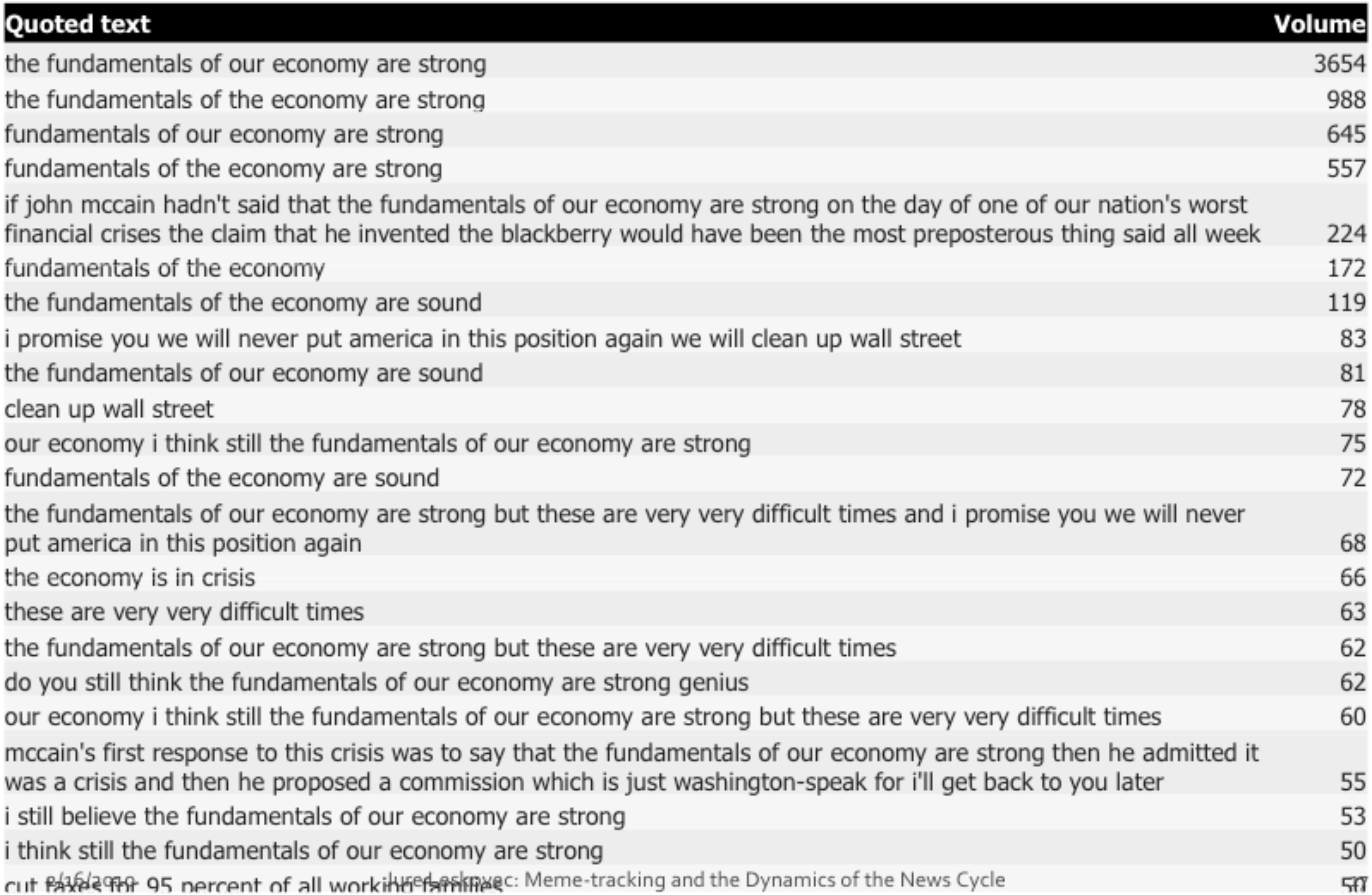


Phrase Partitioning

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



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

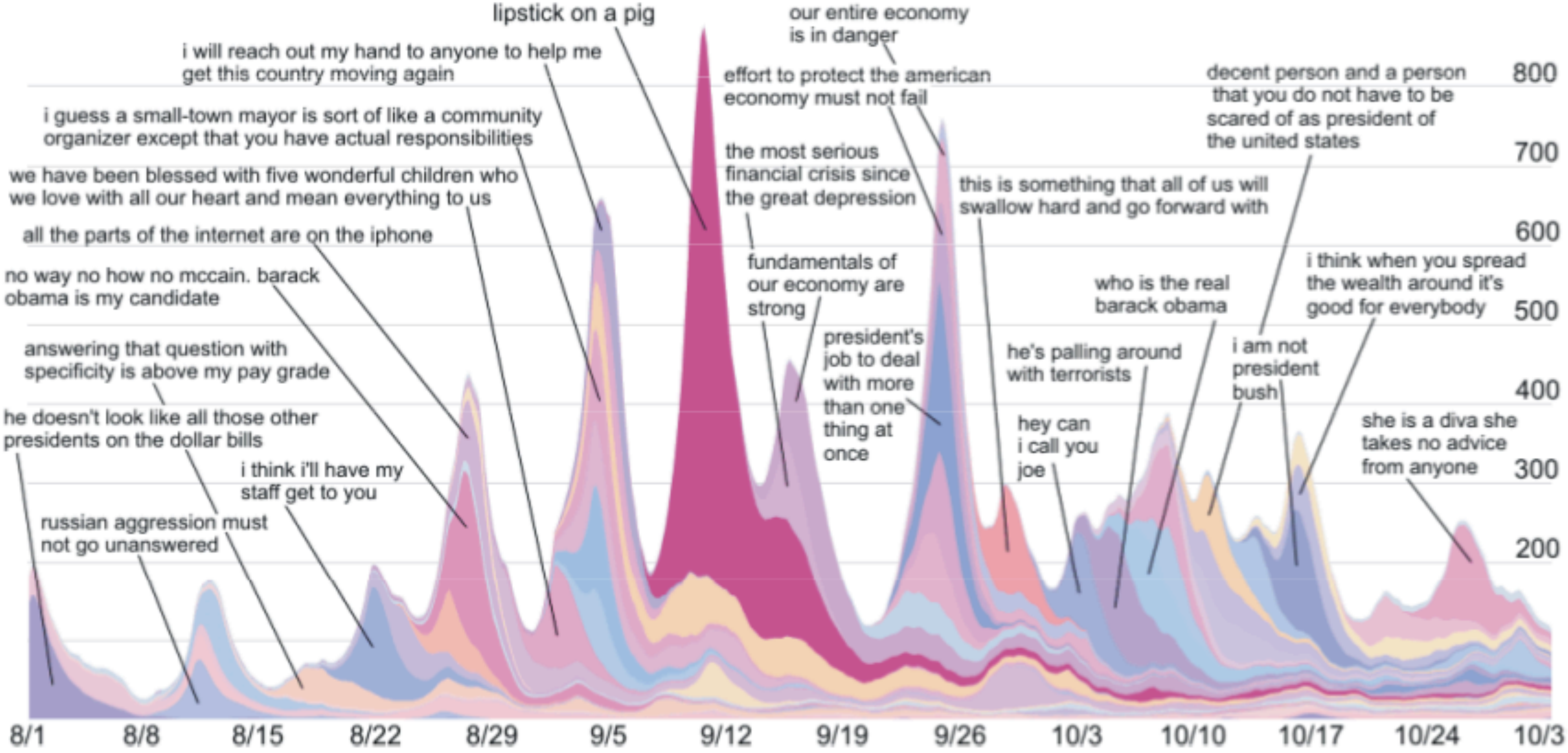
Global Analysis

thread - for a given phrase cluster, a thread is the set of all items (news articles or blog posts) containing some phrase from the cluster

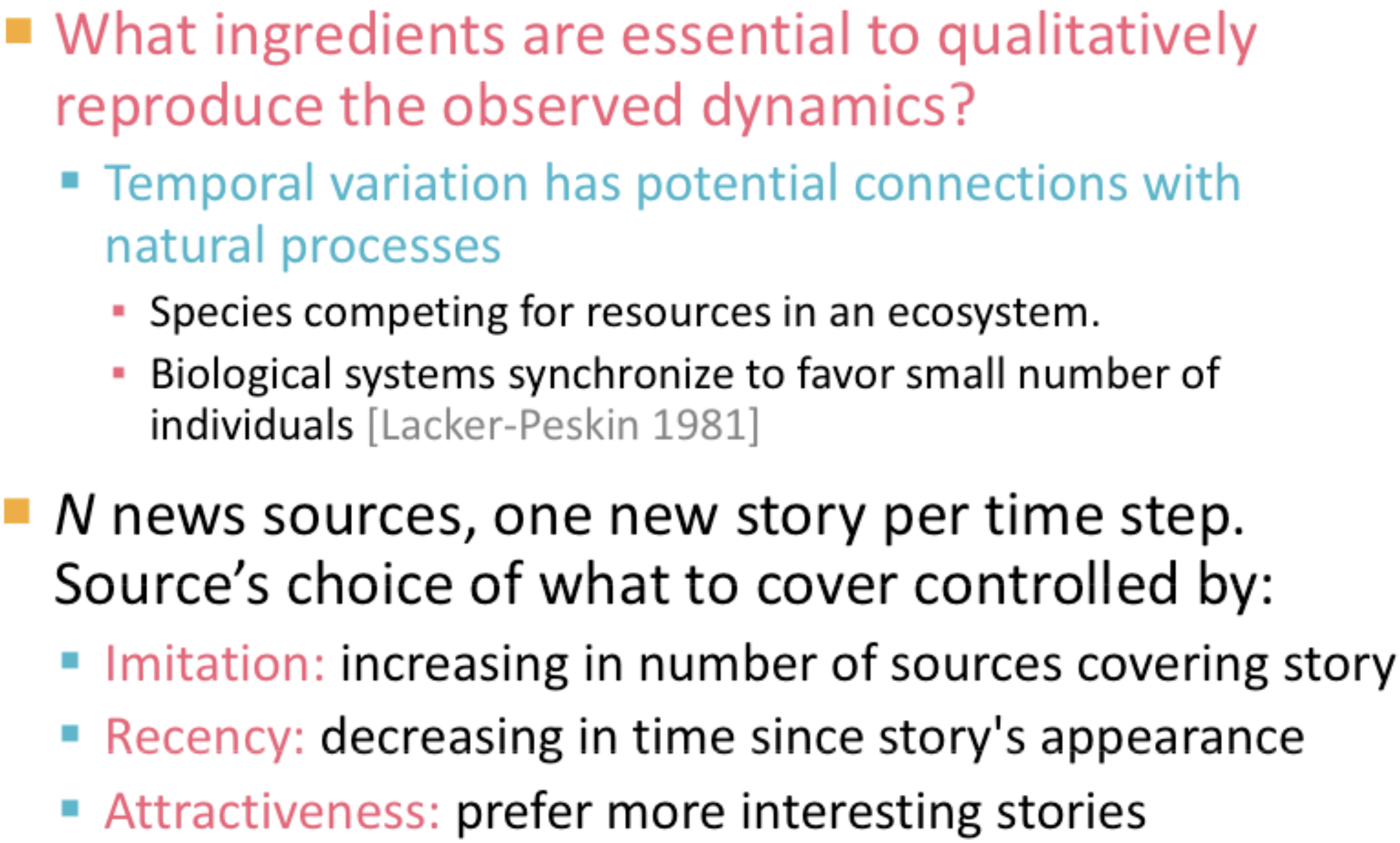
The top 50 threads for the period Aug. 1 2009 - Oct. 31 2009 are plotted in order to measure temporal variation.

The area of each thread corresponds to the number of unique occurrences of that thread at a specific point in time.

Top 50 Threads, Aug - Oct 2009



Modeling the News Cycle



- 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 News Cycle

t = 1, 2, 3, ... , T

N media sources; each reports on a single thread in a given time-step

each given source N chooses a thread j with probability proportional to:

$$f(n_j)\delta(t - t_j)$$

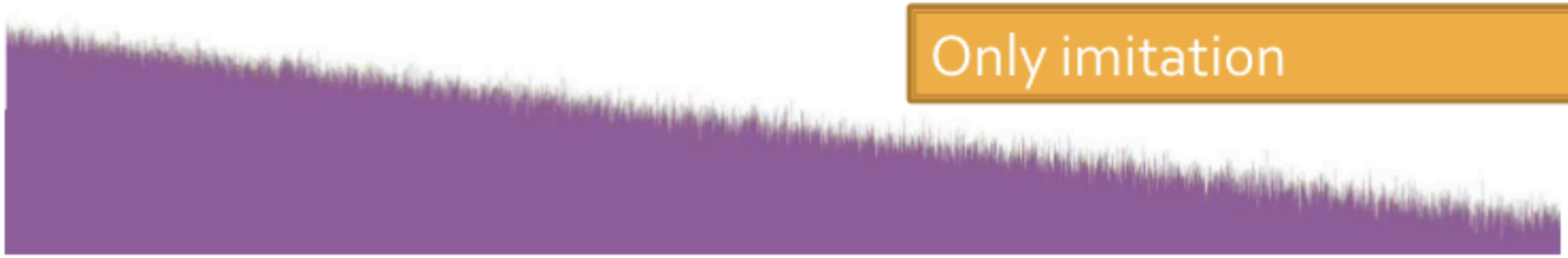
n_j is the number of stories previously written about thread j

t_j is the time when j was first produced

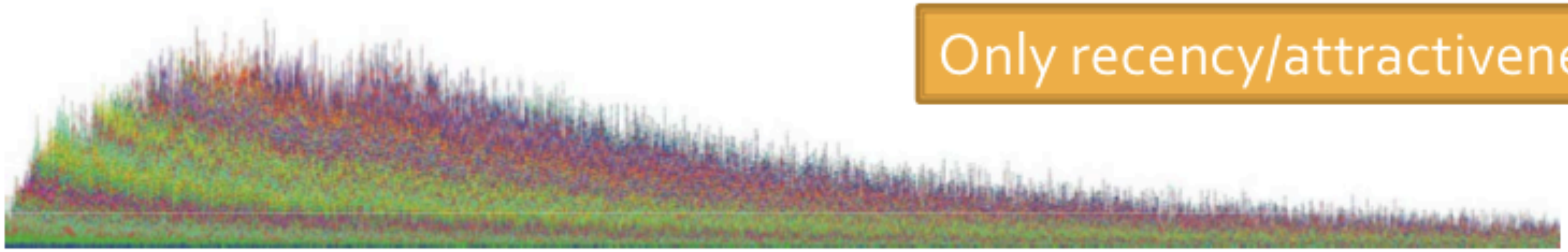
$\delta()$ is an exponentially decaying recency function

$f()$ is a monotonically increasing imitation function

Modeling the News Cycle



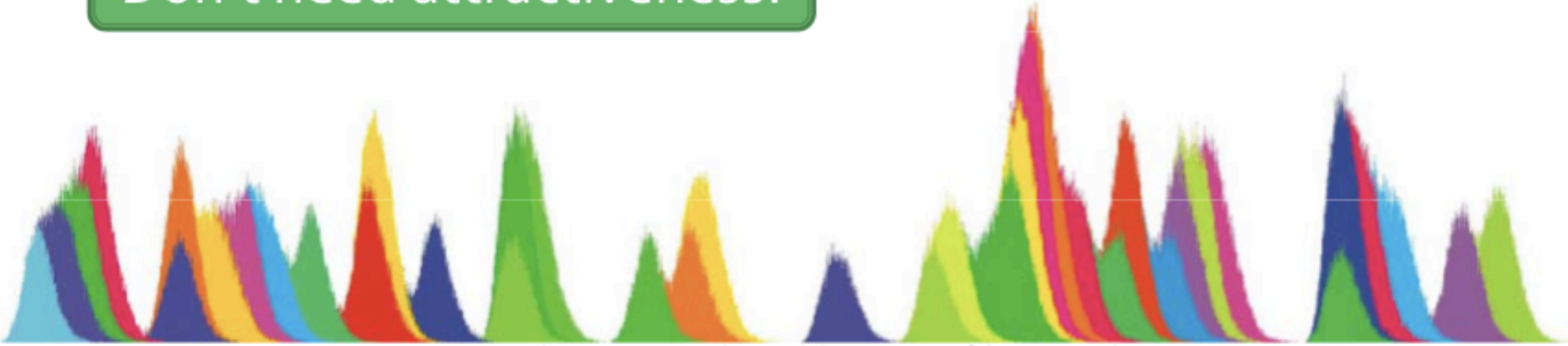
Only imitation



Only recency/attractiveness

Don’t need attractiveness!

Imitation & Recency



Local Analysis

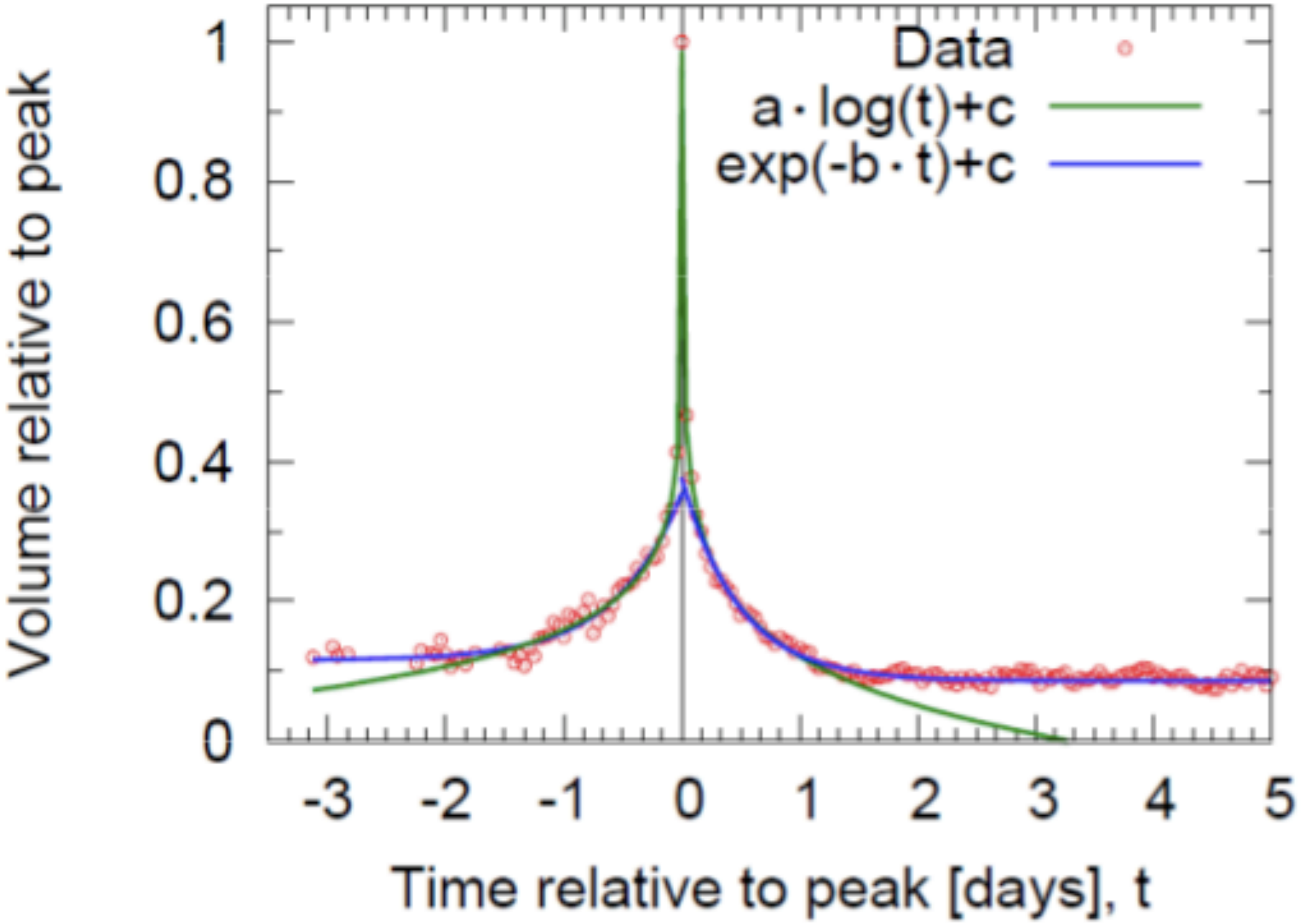
Given some thread p, its volume at a time t is simply the number of items it contains with time-stamp t.

Given some thread p, its peak time is defined as the median of the times at which p occurred in the data set.

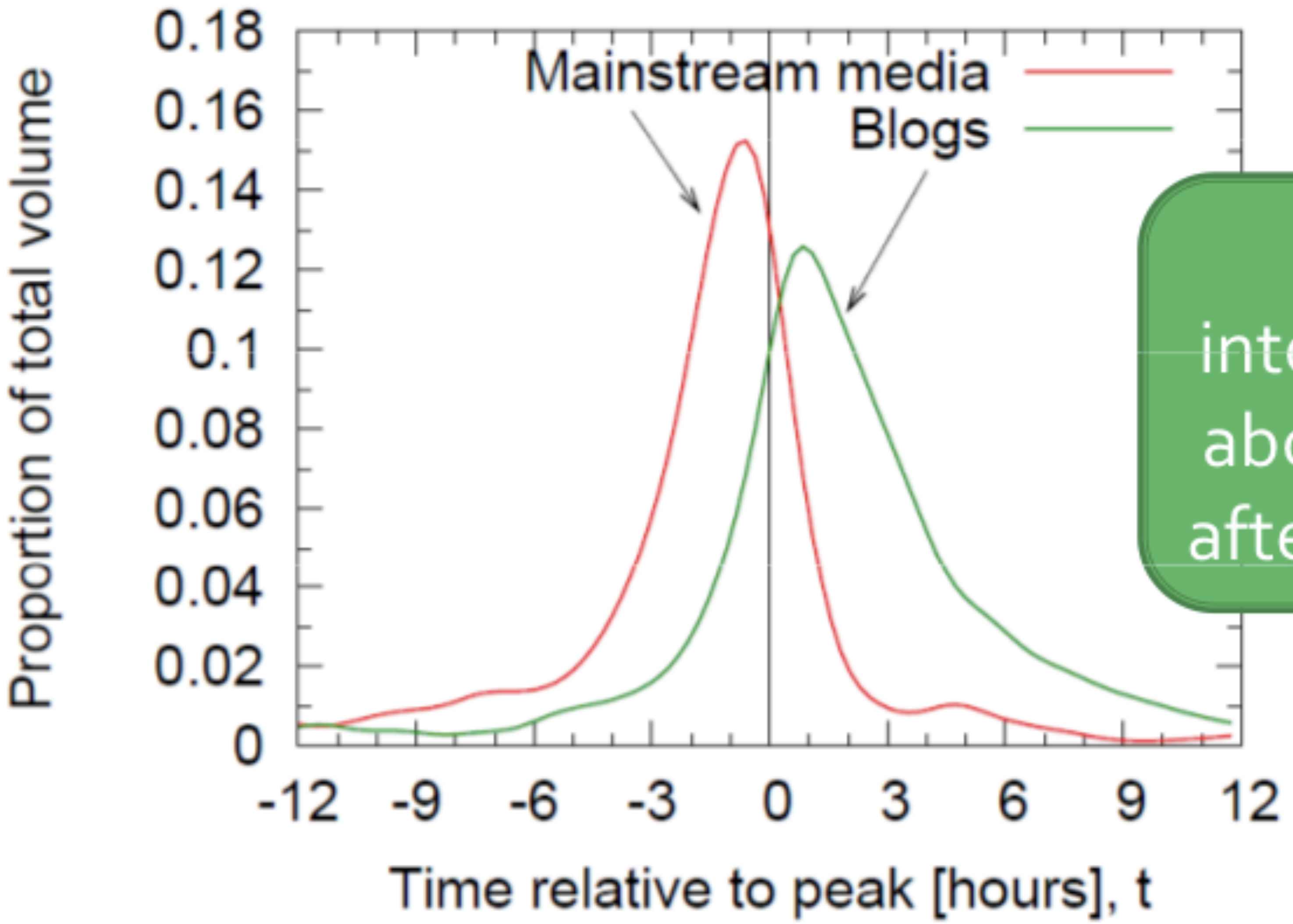
To study birth and decay of memes, the peak times of 1,000 phrase-clusters were calculated. The volume curves were plotted and normalized to t = 0.

Local Analysis

- 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 & Blogs

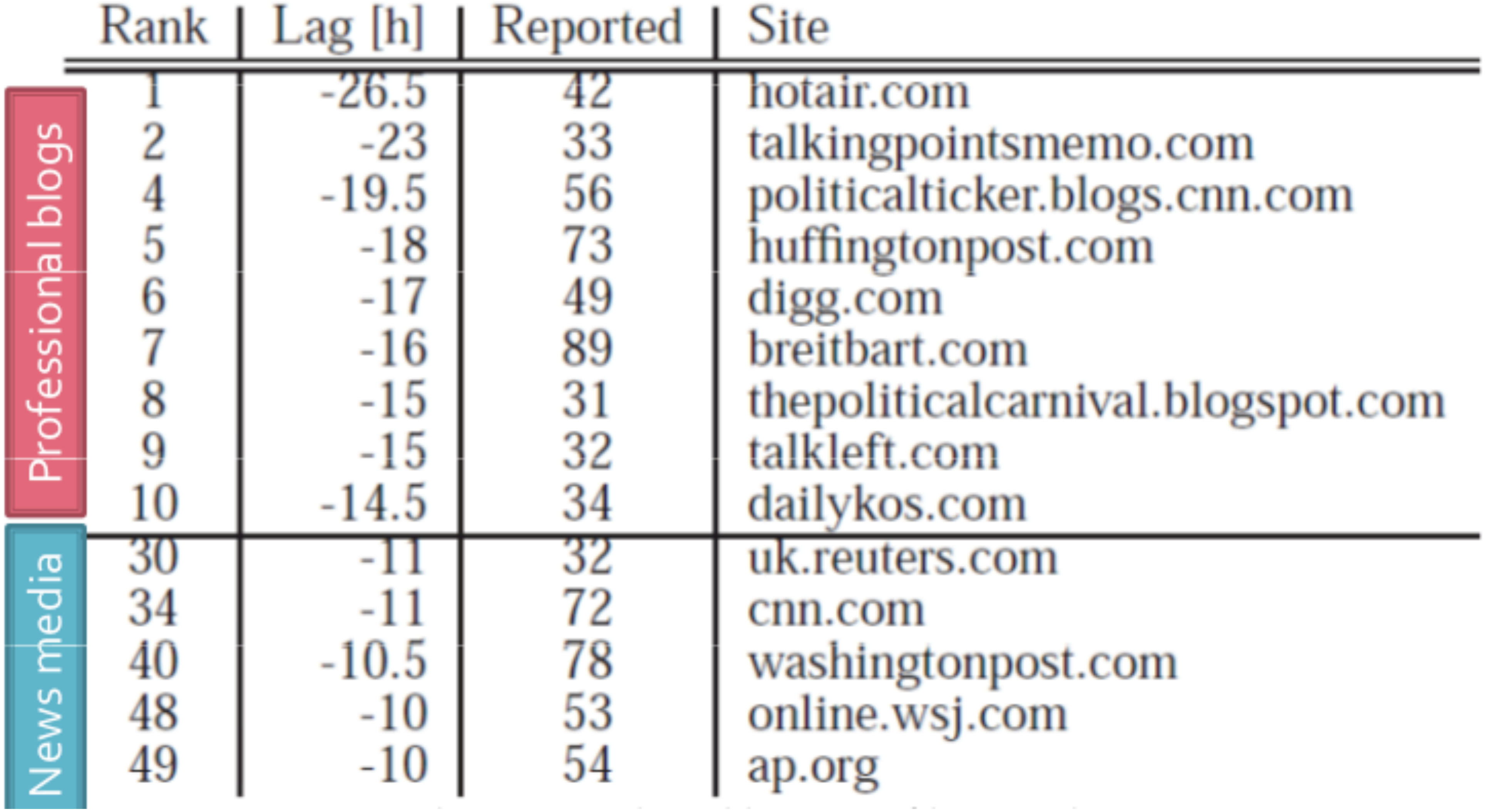


Peak blog intensity comes about 2.5 hours after news peak.

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

Relative Speed of Blogs

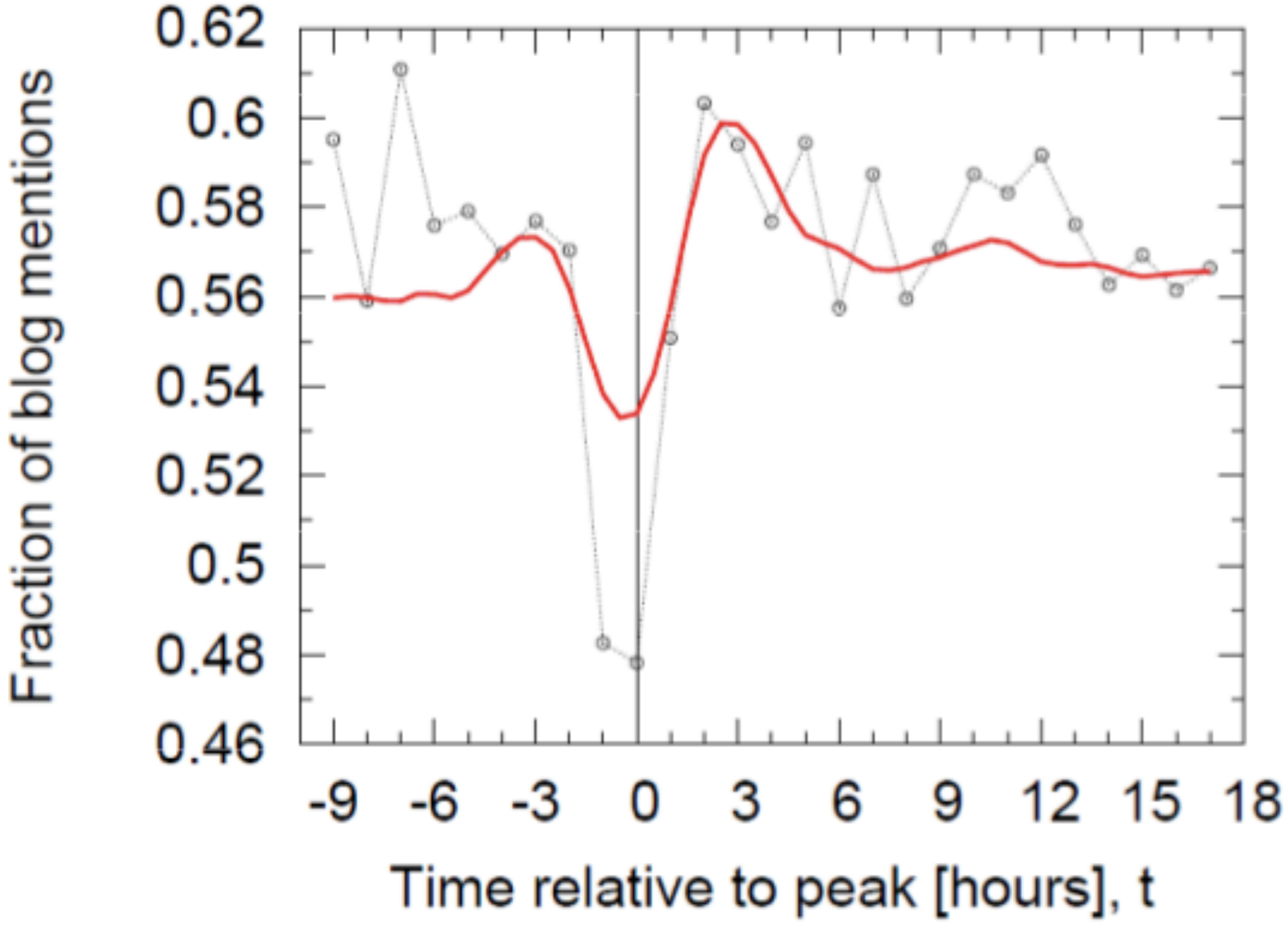
- Can classify individual sources by their typical timing relative to the peak aggregate intensity



	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

News “Pulse”

- Can study “oscillation” of attention between news and media



Questions?