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?
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]
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
Meme-tracking and the Dynamics of the News Cycle
What basic “units” make up the news cycle?

- Cascading 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)
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

Nodes are phrases

BDXCY
ABCD
ABCEFG
ABCEF
CEFP
CEF
UVCEXF
ABCD
ABCDEFGH
ABCEFG
CEFPQR
UVCEXF
Creating clusters of Mutations

Nodes are phrases
Edges are inclusion relations

BDXCY

BCD

ABCD

ABC

CEF

CEF

UVCEXF

ABCEFG

ABCEFG

ABCD

ABCDEF

CEF

CEFP

CEFPQR

ABCDEFGH

2/16/2010 Jure Leskovec: Meme-tracking and the Dynamics of the News Cycle
Nodes are phrases
Edges are inclusion relations
Edges have weights

Creating clusters of Mutations
**Objective:** in directed acyclic graph (approx. quote inclusion), delete min total edge weight s.t. each connected component has a single “sink” node.
**Observation:** enough to know node’s parent

**Heuristic:** proceed top down and assign node to strongest cluster
### A phrase cluster

<table>
<thead>
<tr>
<th>Quoted text</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>the fundamentals of our economy are strong</td>
<td>3654</td>
</tr>
<tr>
<td>the fundamentals of the economy are strong</td>
<td>988</td>
</tr>
<tr>
<td>fundamentals of our economy are strong</td>
<td>645</td>
</tr>
<tr>
<td>fundamentals of the economy are strong</td>
<td>557</td>
</tr>
<tr>
<td>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</td>
<td>224</td>
</tr>
<tr>
<td>fundamentals of the economy are sound</td>
<td>172</td>
</tr>
<tr>
<td>the fundamentals of the economy are sound</td>
<td>119</td>
</tr>
<tr>
<td>i promise you we will never put america in this position again we will clean up wall street</td>
<td>83</td>
</tr>
<tr>
<td>the fundamentals of our economy are sound</td>
<td>81</td>
</tr>
<tr>
<td>clean up wall street</td>
<td>78</td>
</tr>
<tr>
<td>our economy i think still the fundamentals of our economy are strong</td>
<td>75</td>
</tr>
<tr>
<td>fundamentals of the economy are sound</td>
<td>72</td>
</tr>
<tr>
<td>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</td>
<td>68</td>
</tr>
<tr>
<td>the economy is in crisis</td>
<td>66</td>
</tr>
<tr>
<td>these are very very difficult times</td>
<td>63</td>
</tr>
<tr>
<td>the fundamentals of our economy are strong but these are very very difficult times</td>
<td>62</td>
</tr>
<tr>
<td>do you still think the fundamentals of our economy are strong genius</td>
<td>62</td>
</tr>
<tr>
<td>our economy i think still the fundamentals of our economy are strong but these are very very difficult times</td>
<td>60</td>
</tr>
<tr>
<td>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</td>
<td>55</td>
</tr>
<tr>
<td>i still believe the fundamentals of our economy are strong</td>
<td>53</td>
</tr>
<tr>
<td>i think still the fundamentals of our economy are strong</td>
<td>50</td>
</tr>
<tr>
<td>cut taxes for 95 percent of all working families</td>
<td>50</td>
</tr>
</tbody>
</table>
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

Volume over time of top 50 largest total volume phrase clusters

http://memetracker.org
lipstick on a pig
e to help me
nunity ilities

our entire economy is in danger
effort to protect the american economy must not fail
the most serious financial crisis since the great depression
this is something that all of us will swallow hard and go forward with
fundamentals of our economy are strong
president's job to deal with more than one thing at once

who is the real barack obama
he's palling around with terrorists
i am not president bush
hey can i call you joe?

i think when you si the wealth around good for everybody
she is a diva takes no adv from anyone
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

2/16/2010 Jure Leskovec: Meme-tracking and the Dynamics of the News Cycle
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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Lag [h]</th>
<th>Reported</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-26.5</td>
<td>42</td>
<td>hotair.com</td>
</tr>
<tr>
<td>2</td>
<td>-23</td>
<td>33</td>
<td>talkingpointsmemo.com</td>
</tr>
<tr>
<td>4</td>
<td>-19.5</td>
<td>56</td>
<td>politicalticker.blogs.cnn.com</td>
</tr>
<tr>
<td>5</td>
<td>-18</td>
<td>73</td>
<td>huffingtonpost.com</td>
</tr>
<tr>
<td>6</td>
<td>-17</td>
<td>49</td>
<td>digg.com</td>
</tr>
<tr>
<td>7</td>
<td>-16</td>
<td>89</td>
<td>breitbart.com</td>
</tr>
<tr>
<td>8</td>
<td>-15</td>
<td>31</td>
<td>thepoliticalcarnival.blogspot.com</td>
</tr>
<tr>
<td>9</td>
<td>-15</td>
<td>32</td>
<td>talkleft.com</td>
</tr>
<tr>
<td>10</td>
<td>-14.5</td>
<td>34</td>
<td>dailykos.com</td>
</tr>
<tr>
<td>30</td>
<td>-11</td>
<td>32</td>
<td>uk.reuters.com</td>
</tr>
<tr>
<td>34</td>
<td>-11</td>
<td>72</td>
<td>cnn.com</td>
</tr>
<tr>
<td>40</td>
<td>-10.5</td>
<td>78</td>
<td>washingtonpost.com</td>
</tr>
<tr>
<td>48</td>
<td>-10</td>
<td>53</td>
<td>online.wsj.com</td>
</tr>
<tr>
<td>49</td>
<td>-10</td>
<td>54</td>
<td>ap.org</td>
</tr>
</tbody>
</table>
Can study “oscillation” of attention between news and media

![Graph showing the oscillation of attention between news and media over time.]
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

<table>
<thead>
<tr>
<th>(M)</th>
<th>(f_b)</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,141</td>
<td>.30</td>
<td>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.</td>
</tr>
<tr>
<td>826</td>
<td>.18</td>
<td>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.</td>
</tr>
</tbody>
</table>

In total about 3.5% of phrases migrate from blogs to media
But how does information really spread?
Inferring the Diffusion Network

- There is a hidden diffusion network:

![Diagram of a network with nodes a, b, c, d, and e connected in a cycle]

- 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
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
  \[ P(c|G) = \sum_{T \in \mathcal{T}(G)} P(c|T)P(T|G) \]
We want: \( \hat{G} = \arg\max_{|G| \leq k} P(C|G) \)

- 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
\[ w(u, v) = \sum_{c \in C} P_c(u, v) \]
Link types by strength

- Media -> Media
- Media -> Blog
- Blog -> Media
- Blog -> Blog
Detecting information outbreaks

Want to read things before others do.

Detect blue & yellow soon but miss red.

Detect all stories but late.
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
- **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?

Our solution

% of stories detected (higher is better)

In-links (used by Technorati)

Out-links

# posts

Random

Number of selected blogs

www.blogcascades.org

Jure Leskovec: Meme-tracking and the Dynamics of the News Cycle
## So, who was influential (in 2006)?

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

- You can put lipstick on a pig, it's still a pig.

frequency

baseline
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


THANKS!

Demo + Data:
http://memetracker.org