Rhythms of Information Flow through Networks

Jure Leskovec,
Stanford University

Includes joint work with Jaewon Yang, Manuel Gomez-Rodriguez, Andreas Krause, Lars Backstrom and Jon Kleinberg.
Information and Networks

- Information reaches us...
  - ...by personal influence in our social networks
  - ...through influence by mainstream media

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

- From its early stages, a tension between global effects from the mass media and local effects carried by social structure
Traditionally it was hard to capture and quantify the effects of media and social networks.

Explosion of online (social) media:
- Traditional (TV, Newspapers, Agencies)
- Blogs (personal/professional)
- Microblogging (Twitter)
Internet, blogs and social media changed the traditional picture:

- **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.
- **Information** reaches us in small increments from real-time sources and via social networks.

How should this change our understanding of information consumption and of the role of social networks?
Plan for the Talk

- Plan for the talk:
  - Analyze underlying mechanisms for the real-time spread of information through on-line networks

- Motivating questions:
  - How to track messages as they spread?
  - How to model/predict the spread of information?
  - How to identify networks over which the information spreads?
In principle, we can collect nearly all (online) social media content:
- 20 million articles/day (50GB of data/day)
  - 20,000 news sources + millions blogs and forums

What are basic “units” of information?
- Pieces of information that propagate between the nodes (users, media sites, ...)
  - phrases, quotes, messages, links, tags
Would like **units** that:
- correspond to pieces of information
- vary over the order of days,
- and can be handled at terabyte scale

**Approaches:**
- Cascading links to individual articles
  [Adamic-Adar ‘05] [SDM ‘07]
- Textual fragments that travel relatively unchanged, in this case through many news articles:
  - Look for phrases inside quotes: “…”
  - About 1.25 quotes per document in our data
Challenge: Quotes Mutate... A lot!

- 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
- Sees America as imperfect enough to pal around with terrorists who targeted their own country
- Someone who sees America as around with terrorists who target their own country
- Imperfect enough that he's palling around with terrorists who would target their own country
- Is palling around with terrorists
- We see America as a force for good in this world we see an America of exceptionalism
- A force for good in the world
- This is someone who sees America around with terrorists who target their own country
- As being so imperfect he is palling around with terrorists who would target their own country
- Someone who sees America it seems as being so imperfect that 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 with terrorists who targeted their own country
- This is not a man who sees America as you see America and as I see America
- This is not a man who sees America as you see it and how I see America
- America it seems as being so imperfect
- 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 we see
Goal: Find mutational variants of a quote

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

Our basic units are quotes
  - Length $\geq 4$, freq. $\geq 10$
    - Gives 22M quotes

DAG-partitioning is NP-hard but heuristics are effective:
  - Gives $\sim 35,000$ non-trivial clusters
Cluster Volume over Time

August
Volume over time of top 50 largest total volume clusters

Jure Leskovec: Rhythms of Information Flow through Networks
Interaction of News and Blogs

- Can study a division of sources into news and blogs
  - Peak intensity from blogs typically comes about 2.5 hours after peak intensity from news
  - 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>
Q: What are temporal patterns of information attention?

- Item $i$: Piece of information (e.g., quote, url, hashtag)
- Volume $x_i(t)$: # of times $i$ was mentioned at time $t$
  - Volume = number of mentions = attention = popularity
- Q: Typical classes of shapes of $x_i(t)$
- **Given:** Volume of an item over time
- **Goal:** Want to discover types of shapes of volume time series
Cluster time series & find cluster centers

Time series distance function needs to be:

\[ d(x, y) = \min_{a, q} \sum_{t} (x(t) - a \cdot y(t - q))^2 \]

K-Spectral Centroid clustering [WSDM '11]
Patterns of Attention

- **Quotes**: 1 year, 172M docs, 343M quotes
- **Same 6 shapes for Twitter**: 580M tweets, 8M #tags

Jure Leskovec: Rhythms of Information Flow through Networks
Analysis of Attention Patterns

Different media give raise to different patterns

- Spike created by News Agencies (AP Reuters)
- The only cluster that is dominated by loggers both in time and volume
- Slow & small response of blogs
  - Blogs mention 1.3 hours after the mainstream media
  - Blogs mention 20 min before mainstream media
- Blog volume = 29.1% in the mainstream media
- Blog volume = 53.1% in mainstream media
Predicting 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 will mention the info at time 4, 5,...?

- Motivating question:
  - If NYT mentions info at time $t$
  - How many subsequent mentions of the info will this generate at time $t+1$, $t+2$, ...?
Goal:
- Predict future attention (number of mentions)

Traditional view:
- In a network “infected” nodes spread info to their neighbors

Problem:
- The network may be unknown

Idea: Predict the future attention based on which nodes got “infected” in the past
How to predict future volume $x_i(t+1)$ of info $i$?

- Node $u$ has an **influence function** $I_u(q)$:
  - $I_u(q)$: After node $u$ gets “infected”, how many other nodes tend to get infected $q$ hours later
  - E.g.: Influence function of CNN:
    How many sites say the info after they see it on CNN?
- Estimate the influence function from past data
- Predict future volume using the influence functions of nodes infected in the past
LIM model:
- Volume $x_i(t)$ of $i$ at time $t$
- $A_i(t)$ ... a set of nodes that mentioned $i$ before time $t$

And let:
- $I_u(q)$: influence function of $u$
- $t_u$: time when $u$ mentioned $i$

Predict future volume as a sum of influences:
$$x_i(t+1) = \sum_{u \in A_i(t)} I_u(t - t_u)$$
After node $u$ mentions the info, $I_u(q)$ other mentions tend to occur $q$ hours later

- $I_u(q)$ is not observable, need to estimate it
- Make no assumption about its shape
  - Model $I_u(q)$ as a vector: $I_u(q) = [I_u(1), I_u(2), I_u(3), \ldots, I_u(L)]$
- Find $I_u(q)$ by solving a least-squares-like problem:

$$
\min_{I_u, \forall u} \sum_i \sum_t \left( x_i(t+1) - \sum_{u \in A_i(t)} I_u(t-t_u) \right)^2
$$
The model: Performance

- Take top 1,000 quotes by the total volume:
  - Total 372,000 mentions on 16,000 websites
- Build LIM on 100 highest-volume websites
  - \( x_i(t) \) ... number of mentions across 16,000 websites
  - \( A_{ij}(t) \) ... which of 100 sites mentioned quote \( i \) and when
- Improvement in L2-norm over 1-time lag predictor

<table>
<thead>
<tr>
<th></th>
<th>Bursty phrases</th>
<th>Steady phrases</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>7.21%</td>
<td>8.30%</td>
<td>7.41%</td>
</tr>
<tr>
<td>ARMA</td>
<td>6.85%</td>
<td>8.71%</td>
<td>7.75%</td>
</tr>
<tr>
<td>LIM (N=100)</td>
<td>20.06%</td>
<td>6.24%</td>
<td>14.31%</td>
</tr>
</tbody>
</table>

Jure Leskovec: Rhythms of Information Flow through Networks
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
Politics is dominated by traditional media

- **Blogs:**
  - Influential for Entertainment phrases
  - Influence lasts longer than for other media types
Inferring the Diffusion Network

- But how does information really spread?

- We only see time of mention but not the edges
- Can we reconstruct (hidden) diffusion network?
### Virus propagation
- **Process**: Viruses propagate through the network
- **We observe**: We only observe when people get sick
- **It’s hidden**: But NOT who infected whom

### Word of mouth & Viral marketing
- **Process**: Recommendations and influence propagate
- **We observe**: We only observe when people buy products
- **It’s hidden**: But NOT who influenced whom

---

**Can we infer the underlying network?**
Inferring Diffusion Networks

- There is a hidden diffusion network:

```
  a -- b -- d
  |   |   |
  c   e   c
```

- 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-infests-whom network!**
  - The problem is NP-hard (there are $O(2^{N^2})$ graphs!)
  - Our algorithm can do it near-optimally in $O(N^2)$
Experiments

- Propagation of quotes:
  - 172 million news and blog articles over 1 year
  - Extract 343 million different quotes
  - Record times $t_i(w)$ when site $w$ mentioned quote $i$

- Given the times when sites mention quotes, infer the network of information diffusion:
  - Who tends to copy (repeat after) whom
5,000 news sites:

- Blogs
- Mainstream media
Diffusion Network (zoom-in)

Blogs
Mainstream media

Jure Leskovec: Rhythms of Information Flow through Networks
Messages arriving through networks from real-time sources requires new ways of thinking about information dynamics and consumption:

- Tracking information through (implicit) networks
- Quantify the dynamics of online media
- Predict the diffusion of information
- And infer networks of information diffusion
**Further Qs: Opinion dynamics**

- Can this analysis help identify dynamics of polarization [Adamic-Glance ‘04]?  

- Connections to mutation of information:  
  - How does *attitude* and *sentiment* change in different parts of the network?  
  - How does *information* change in different parts of the network?
THANKS!

http://snap.stanford.edu
http://memetracker.org
References


