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)

- 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

[Lazarsfeld-Katz ‘55]
[Watts-Dodds ‘07]
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?
Imagine you want to track the flow of information

- We would like to identify cascades like this:
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
Challenge: Phrases Mutate

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.
**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
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
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 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

[KDD '09]
### Meme: 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>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</td>
<td>988</td>
</tr>
<tr>
<td>fundamentals of the economy are strong</td>
<td>645</td>
</tr>
<tr>
<td>fundamentals of the economy are sound</td>
<td>557</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>our economy i think still the fundamentals of our economy are strong</td>
<td>172</td>
</tr>
<tr>
<td>fundamentals of the economy are sound</td>
<td>119</td>
</tr>
<tr>
<td>the fundamentals of the economy are sound</td>
<td>81</td>
</tr>
<tr>
<td>clean up wall street</td>
<td>78</td>
</tr>
<tr>
<td>the fundamentals of our economy are sound</td>
<td>75</td>
</tr>
<tr>
<td>our economy i think still the fundamentals of our economy are strong</td>
<td>72</td>
</tr>
<tr>
<td>these are very very difficult times</td>
<td>68</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>66</td>
</tr>
<tr>
<td>the economy is in crisis</td>
<td>63</td>
</tr>
<tr>
<td>these are very very difficult times</td>
<td>63</td>
</tr>
<tr>
<td>do you still think the fundamentals of our economy are strong</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>62</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>60</td>
</tr>
<tr>
<td>i still believe the fundamentals of our economy are strong</td>
<td>55</td>
</tr>
<tr>
<td>i think still the fundamentals of our economy are strong</td>
<td>53</td>
</tr>
<tr>
<td>cut taxes for 95 percent of all working families</td>
<td>50</td>
</tr>
<tr>
<td>today of all days john mccain's stubborn insistence that the fundamentals of the economy are strong shows that he is</td>
<td>50</td>
</tr>
</tbody>
</table>
... is periodic, has no “real” trends.

“Bandwidth” of the online media is constant.
Memes Over Time

August

12/2/2014

- Volume over time of top 50 largest total volume memes (phrase clusters)
- More at: [http://snap.stanford.edu/nifty](http://snap.stanford.edu/nifty)
Memes on the “Great Depression”

- Media coverage of the current economic crisis
- Main proponents of the debate:

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Original Speaker</th>
<th>Starting Date</th>
<th>Total Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>we will rebuild, we will recover...</td>
<td>Barack Obama</td>
<td>24-Feb</td>
<td>4679</td>
</tr>
<tr>
<td>how do they justify this outrage to the taxpayers...</td>
<td>Barack Obama</td>
<td>16-Mar</td>
<td>4446</td>
</tr>
<tr>
<td>in ... our greatest economic crisis since the Great Depression...</td>
<td>Barack Obama</td>
<td>7-Feb</td>
<td>3914</td>
</tr>
<tr>
<td>they'll have to find someone else to write the next stimulus bill</td>
<td>NY Post</td>
<td>16-Feb</td>
<td>3312</td>
</tr>
<tr>
<td>...the weight of this crisis will not determine the destiny of this nation</td>
<td>Barack Obama</td>
<td>24-Feb</td>
<td>3113</td>
</tr>
<tr>
<td>...to be honest I'm a little bit worried</td>
<td>Chinese Premier</td>
<td>13-Mar</td>
<td>3017</td>
</tr>
<tr>
<td>buying stocks is a potentially good deal</td>
<td>Barack Obama</td>
<td>3-Mar</td>
<td>2690</td>
</tr>
<tr>
<td>...we would not be able to continue as a going concern...</td>
<td>General Motors</td>
<td>5-Mar</td>
<td>2672</td>
</tr>
<tr>
<td>we’ve seen some progress in the financial markets, absolutely</td>
<td>Ben Bernanke</td>
<td>15-Mar</td>
<td>2425</td>
</tr>
</tbody>
</table>

Top republican voice ranks only 14th
Interaction of News and Blogs

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

![Graph showing volume relative to peak over time.](chart.png)

Data

\[ a \cdot \log(t) + c \]

\[ \exp(-b \cdot t) + c \]
Interaction of News and Blogs

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

Peak blog intensity comes about 2.5 hours after news peak.
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>politico.ticker.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>
Interaction of News and Blogs

- The “oscillation” of attention between mainstream media and 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

<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 3.5% of phrases migrate from blogs to media
Network Inference
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?
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!
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**
- 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 the Diffusion Network

Network $G^*$

Cascade $c_1$

Cascade $c_2$

Cascade $c_3$

Node $i$
**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 $b$

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

Here: $T = \{a \rightarrow b \rightarrow c\}$

- OR

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

In both $T_1, T_2$ the order of infections is the same: $a,b,c$
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 $O(2^{N^2})$ graphs)
**Continuous time cascade diffusion model:**

- Cascade $c$ reaches node $u$ at $t_u$
  - and spreads to $u$’s neighbors:
    - With probability $\beta$ cascade propagates along edge $(u, v)$
    - and we determine the infection time of node $v$

  $t_v = t_u + \Delta$

  e.g.: $\Delta \sim \text{Exponential}$

We assume each node $v$ has only one parent!
The model for one cascade:

- Cascade reaches node $u$ at time $t_u$, and spreads to $u$’s neighbors $v$: With prob. $\beta$ cascade propagates along edge $(u,v)$ and $t_v = t_u + \Delta$

Transmission probability:

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

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

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

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

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

  - 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 $C$:

$$P(C|G) = \prod_{c \in C} P(c|G)$$

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

- Want to find the “best” graph:

$$G^* = \arg\max_{|G| \leq k} F_C(G)$$

The problem is **NP-hard**: MAX-k-COVER [KDD ’10]
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:
    $$= \sum_{i \in V} \max_{\text{Par}_T(i)} w(\text{Par}_T(i), i)$$

Local greedy selection gives optimal tree!
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 \( G \subseteq G' \): \( w_s \leq w'_s \) and \( w_{rs} = w'_{rs} \)
- \( F_c(G \cup \{(r,s)\}) - F_c(G) \)
  \[
  = \max(w_s, w_{rs}) - w_s \\
  \geq \max(w'_s, w_{rs}) - w'_s \\
  = F_c(G' \cup \{(r,s)\}) - F_c(G')
  \]
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

$$e_i = \arg\max_{e \in G \setminus G_{i-1}} F_C(G_{i-1} \cup \{e\}) - F_C(G_{i-1})$$

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 $G$!
NetInf achieves ≈ 90 % of the best possible network!
With 2x as many infections as edges, the break-even point is already 0.8 - 0.9!
Memetracker dataset:
- 172m news articles
- Aug ‘08 – Sept ‘09
- 343m textual phrases
- Times $t_c(w)$ when site $w$ mentions phrase $c$

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:
Diffusion Network (small part)