The Flow of Information in Online Networks

Jure Leskovec

Includes joint work: Lars Backstrom, Manuel Gomez-Rodriguez, Jon Kleinberg, Andreas Krause, Seth Myers, Rok Sosic, Caroline Sue, Chenguan Zhu
Diffusion in Networks

- Information flows from a node to node like an epidemic

- How does information transmitted by mainstream media interact with social networks?

Diagram:
- Obscure tech story
  - Small tech blog
    - Engadget
    - Slashdot
    - Wired
    - BBC
    - NYT
    - CNN
Since August 2008 we have been collecting 30M articles/day: 6B articles, 20TB of data

Challenge:

How to track information as it spreads?
Goal: Trace textual phrases that spread through many news articles

Challenge 1: Phrases mutate!

Mutations of a meme about the Higgs boson particle.

2/13/2013

Jure Leskovec, Stanford University
Goal: Find mutational variants of a phrase

Objective:

- In a DAG of approx. phrase inclusion, delete min total edge weight such that each component has a single “sink”
The Algorithm

Challenge 2: 20TB of data!

Solution: Incremental phrase clustering

- Phrases arrive in a stream
- Simultaneously cluster the graph and attach new phrases to the graph
- Dynamically remove completed clusters

Overall, it takes 1 server, 60GB memory and 4 days to process 6B documents

- + 1 summer and 3 Stanford undergrads
Memes over Time

Visualization of 1 month of data from Aug 2012

- Browse all 4 years of data at http://snap.stanford.edu/nifty
Challenge 3: Information network is hidden

Goal: Infer the information diffusion network

- There is a hidden network, and
- We only see times when nodes get “infected”

Yellow info: (a,1), (c,2), (b,3), (e,4)
Blue info: (c,1), (a,4), (b,5), (d,6)
Inferring Networks

Virus propagation
- Viruses propagate through the network
- We only observe when people get sick
- But NOT who infected them

Word of mouth & Viral marketing
- Recommendations and influence propagate
- We only observe when people buy products
- But NOT who influenced them

Can we infer the underlying network?
Yes, a convex optimization problem!

Blogs

Mainstream media

5,000 news sites:

News Diffusion Network
Observe times when nodes adopt the information

But where did the first node find the information?

Potential node-to-node spread

External Influence

How did the information “jump”? 
Towards the Model

External source

Model the arrival of external exposures using **event profile**

Neighbors

Model the prob. of adoption using the **adoption curve**

Adopt!

[2/13/2013]

Jure Leskovec, Stanford University
Why is modeling external influence hard?

- External sources are unobservable
- Amount of external influence varies over time
- External influence can be confused with network influence

The post occurred both due to external and network effects!

Posts the rumor
Given:
- Network $G$
- Node adoption times $(i, t)$ of a contagion

Goal: Infer
- (1) External event profile
- (2) Adoption curve such that observed adoption times fit best
## Results: Different Topics

<table>
<thead>
<tr>
<th>Category</th>
<th>max P(k)</th>
<th>k at max P(k)</th>
<th>Duration (hours)</th>
<th>% Ext. Exposures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics (25)</td>
<td>0.0007 +/- 0.0001</td>
<td>4.59 +/- 0.76</td>
<td>51.24 +/- 16.66</td>
<td>47.38 +/- 6.12</td>
</tr>
<tr>
<td>World (824)</td>
<td>0.0013 +/- 0.0000</td>
<td>2.97 +/- 0.10</td>
<td>43.54 +/- 2.94</td>
<td>26.07 +/- 1.19</td>
</tr>
<tr>
<td>Entertain. (117)</td>
<td>0.0015 +/- 0.0002</td>
<td>3.52 +/- 0.28</td>
<td>89.89 +/- 16.13</td>
<td>17.87 +/- 2.51</td>
</tr>
<tr>
<td>Sports (24)</td>
<td>0.0010 +/- 0.0003</td>
<td>4.76 +/- 0.83</td>
<td>87.85 +/- 38.03</td>
<td>43.88 +/- 6.97</td>
</tr>
<tr>
<td>Health (81)</td>
<td>0.0016 +/- 0.0002</td>
<td>3.25 +/- 0.30</td>
<td>100.09 +/- 17.57</td>
<td>18.81 +/- 3.33</td>
</tr>
<tr>
<td>Tech. (226)</td>
<td>0.0013 +/- 0.0001</td>
<td>3.00 +/- 0.16</td>
<td>83.05 +/- 8.73</td>
<td>18.36 +/- 1.80</td>
</tr>
<tr>
<td>Business (298)</td>
<td>0.0015 +/- 0.0001</td>
<td>3.18 +/- 0.16</td>
<td>49.61 +/- 5.14</td>
<td>22.27 +/- 1.79</td>
</tr>
<tr>
<td>Science (106)</td>
<td>0.0012 +/- 0.0002</td>
<td>4.06 +/- 0.30</td>
<td>135.28 +/- 16.19</td>
<td>20.53 +/- 2.78</td>
</tr>
<tr>
<td>Travel (16)</td>
<td>0.0005 +/- 0.0001</td>
<td>2.33 +/- 0.29</td>
<td>151.73 +/- 39.70</td>
<td>39.99 +/- 6.60</td>
</tr>
<tr>
<td>Art (32)</td>
<td>0.0006 +/- 0.0001</td>
<td>5.26 +/- 0.66</td>
<td>188.55 +/- 48.17</td>
<td>27.54 +/- 5.30</td>
</tr>
<tr>
<td>Edu. (31)</td>
<td>0.0009 +/- 0.0001</td>
<td>3.77 +/- 0.51</td>
<td>130.53 +/- 38.63</td>
<td>21.45 +/- 6.40</td>
</tr>
</tbody>
</table>

So far we considered contagions as independently propagating.

How do contagions interact?

- Does being exposed to blue change the probability of talking about red contagion?
Goal: Model interaction between many contagions spreading over the network simultaneously

- Some contagions may help each other in adoption
- Others may compete for attention
User is reading posts on Twitter:

- User examines posts one by one
- Currently she is examining $X$
- How does the probability of reposting $X$ depend on what she has seen in the past?

$$P(\text{post } X \mid \text{exposed to } X, Y_1, Y_2, Y_3) = ?$$
What’s the goal?

Given:
- For a single user: Exposure and infection events

Goal: Infer tweet topic memberships and topic interactions
- Reinforces
- But suppresses
The Model

- **Goal**: Model $P(\text{post } X | \text{ exp. } X, Y_1, Y_2, Y_3)$

- **Assume exposures are not correlated**:

  \[
P \left( X | \{Y_k\}_{K=1}^K \right) = \frac{1}{P(X)^{K-1}} \prod_{k=1}^{K} P(X|Y_k)
  \]

- Where each $P(X|Y_k)$:

  \[
P(X = u_j | Y_k = u_i) = P(X = u_j) + \sum_t \sum_s M_{i,t} \cdot \Delta_{t,s}^{(k)} \cdot M_{j,s}
  \]
Dataset: Twitter

- **Data from Twitter**
  - *Complete* data from Jan 2011: 3 billion tweets
  - All URLs tweeted by at least 50 users: 191k
- **Task:**
  Predict whether a user will post URL $X$
  - Train on 90% of the data, test on 10%
- **Baselines:**
  - Infection Probability (IP): $P(X = u_i | Y_k = u_j) = P(X = u_i)$
  - IP + Node bias (NB): $P(X = u_i) + \gamma_n$
  - Exposure curve (EC): $P(X | \# \text{ times exposed to } X)$
Predicting Retweets

- Task: Predict a retweet given the context

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Log-Like.</th>
<th>max $F_1$</th>
<th>Area under PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>-335,550.39</td>
<td>0.0150</td>
<td>0.0157</td>
</tr>
<tr>
<td>UB</td>
<td>-338,821.54</td>
<td>0.0112</td>
<td>0.0123</td>
</tr>
<tr>
<td>EC</td>
<td>-338,367.86</td>
<td>0.0181</td>
<td>0.0250</td>
</tr>
<tr>
<td><strong>Our Model - With Prior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMM K=1</td>
<td>-313,843.93</td>
<td>0.0412</td>
<td>0.0515</td>
</tr>
<tr>
<td>IMM K=2</td>
<td>-299,884.86</td>
<td><strong>0.0465</strong></td>
<td><strong>0.1238</strong></td>
</tr>
<tr>
<td>IMM K=3</td>
<td><strong>-299,352.32</strong></td>
<td>0.0380</td>
<td>0.0926</td>
</tr>
<tr>
<td>IMM K=4</td>
<td>-315,319.54</td>
<td>0.0321</td>
<td>0.0804</td>
</tr>
<tr>
<td>IMM K=5</td>
<td>-352,687.54</td>
<td>0.0386</td>
<td>0.0924</td>
</tr>
</tbody>
</table>
How do Tweets Interact?

- How $P(\text{post } u_2| \text{ exp. } u_1)$ changes if ...
  - $u_2$ and $u_1$ are similar/different in content?
  - $u_1$ is highly viral?

**Observations:**
- If $u_1$ is not viral, this boost $u_2$.
- If $u_1$ is highly viral, this kills $u_2$.

**BUT:**
Only if $u_1$ and $u_2$ are of low content similarity (LCS) else, $u_1$ helps $u_2$.

Relative change in infection prob.
Final Remarks

- **Modeling contagion interactions**
  - 71% of the adoption probability comes from the topic interactions!
  - Modeling user bias does not matter

- **Detecting external events**
  - Overall, 69% exposures on Twitter come from the network and 29% from external sources
    - About the same for URLs as well as hashtags!
Can such analysis help identify dynamics of polarization [Adamic-Glance ‘05]?

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?
Methodology:

- Each node of the cascade is a blog post that belongs to a blog
- For each blog compute the baseline sentiment (over all its posts)
  - Subjectivity: deviation in sentiment from the baseline (in positive or negative direction)

Question:

- Does sentiment flow in cascade?
Cascades “heats” up early, then cool off

Subjectivity of the child and the parent are correlated. 

Sentiment flows!
Chirp! Chirp!

RT @birdnexttome
chirp! chirp! << LoL
#funnybirdnoises

wtf!??
Thank you!
http://snap.stanford.edu
References


