Jure Leskovec
Machine Learning Department
Carnegie Mellon University

How to detect epidemics and influential blogs?
Networks: Rich data

- **Today**: Large on-line systems have detailed records of human activity
  - **On-line communities**:
    - Facebook (64 million users, billion dollar business)
    - MySpace (300 million users)
  - **Communication**:
    - Instant Messenger (~1 billion users)
  - **News and Social media**:
    - Blogging (250 million blogs world-wide, presidential candidates run blogs)
  - **On-line worlds**:
    - World of Warcraft (internal economy 1 billion USD)
    - Second Life (GDP of 700 million USD in ‘07)

Opportunities for impact in science and industry
We need **massive network data for the patterns to emerge:**

- **MSN Messenger network** [WWW ’08]
  - 240M people, 255B messages, 4.5 TB data
- **Blogosphere**
  - 60M posts, 120M links
**Diffusion and Cascades**

- Behavior that cascades from node to node like an epidemic
  - News, opinions, rumors
  - Word-of-mouth in marketing
  - Infectious diseases
- As activations spread through the network they leave a trace – a cascade

![Network](image1.png)

![Cascade](image2.png)

**Network**

**Cascade** (propagation graph)
Where do cascades occur?
- On the Web we can actually observe and measure a number of cascades

What do cascades look like?
- How do information and influence spread?

How to detect who is influential?
- Effective and efficient algorithms
- Saving lives
Setting 1: Viral marketing

- People send and receive product recommendations, purchase products

  - 10% credit
  - 10% off

Data: Large online retailer: 4 million people, 16 million recommendations, 500k products
Bloggers write posts and refer (link) to other posts and the information propagates.

- **Data**: 10.5 million posts, 16 million links
What do cascades look like?

- Are they stars? Chains? Trees?

- Information cascades (blogosphere):

- Viral marketing (DVD recommendations):

- Viral marketing cascades are more social:
  - Collisions (no summarizers)
  - Richer non-tree structures
Prob. of adoption depends on the number of friends who have adopted [Bass ‘69, Granovetter ’78]

What is the shape?
- Distinction has consequences for models and algorithms

Diminishing returns? Critical mass?

To find the answer we need lots of data
Later similar findings were made for group membership [Backstrom-Huttenlocher-Kleinberg ‘06], and probability of communication [Kossinets-Watts ‘06].

Adoption curve follows the **diminishing returns**. Can we exploit this?
Cascade & outbreak detection

- Blogs – information epidemics
  - Which are the influential/infectious blogs?

- Viral marketing
  - Who are the trendsetters?
  - Influential people?

- Disease spreading
  - Where to place monitoring stations to detect epidemics?
The problem: Detecting cascades

How to quickly detect cascades as they spread?

[w/ Krause-Guestrin et al., KDD ’07]
(best student paper)
Two parts to the problem

- **Cost:**
  - Cost of monitoring is blog dependent (big blogs cost more time to read)

- **Reward:**
  - Minimize the number of people that know the story before we do
The solution: Covering blogs

= Given a budget (e.g., of 3 blogs)
= Select blogs to cover the most of the blogosphere?
= Bad news: Solving this exactly is NP-hard
= Good news: **Theorem**: Our algorithm CELF can do it in linear time and with factor 3 approximation

[w/ Krause-Guestrin et al., KDD ’07] (best student paper)
Gain of adding a node to small set is larger than gain of adding a node to large set

Submodularity: diminishing returns, think of it as “concavity”
Back to the Question...

= I have 10 minutes. Which blogs should I read to be most up to date?

= Who are the most influential bloggers?
A single story propagates...

Sooner we read the story, more of its influence area we cover.
Blogs: Information epidemics

- Which blogs should one read?

"Covered" blogosphere (higher is better)

For more info see our website: www.blogcascades.org
**CELF: Scalability**

CELF runs 700x faster than simple greedy algorithm.
## So, who is influential? What should I read?

<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://themo">http://themo</a> modulator.org</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>
Given:
- a real city water distribution network
- data on how contaminants spread over time

Place sensors (to save lives)

Problem posed by the *US Environmental Protection Agency*
Our approach performed best at the Battle of Water Sensor Networks competition.
How do news and information spread
  - New ranking and influence measures for blogs
  - Recommendations and incentives
  - Diffusion of topics (news, media)

Predictive models of information diffusion
  - Social Media Marketing

How to design better systems incorporating diffusion and incentives
References

- Jure Leskovec, jure@cs.cmu.edu
- http://www.cs.cmu.edu/~jure/