Cascading Processes and Network Structure

Jon Kleinberg

Cornell University

Including joint work with Lars Backstrom, Larry Blume, David Easley, Bobby Kleinberg, Cameron Marlow, Éva Tardos, and Johan Ugander
Two Metaphors for the Web

The Web is balanced between two metaphors.

- The library:
  knowledge, pages, hyperlinks, associations.
- The crowd:
  real-time awareness, memes, contagion.
Graph Structure in On-Line Social Systems

The graph-theoretic terrain of the on-line world.

This talk:

- Processes operating on the network: the flow of information
- Analysis of the structure: node neighborhoods
  [Backstrom-Kleinberg 2014]
Cascading behavior has become a basic “transport mechanism” in many kinds of networks.

- Agricultural, medical innovations [Ryan-Gross 1943, Coleman et al 1966]
- Media influence and two-stage flow [Lazarsfeld et al 1944]
- Collective action, social movements [McAdam 1986, Chwe 1999]
- Viral marketing [Jurvetson 2000, Domingos-Richardson 2001]
- Current social media systems ...
Cascading behavior has become a basic “transport mechanism” in many kinds of networks.

- Agricultural, medical innovations [Ryan-Gross 1943, Coleman et al 1966]
- Media influence and two-stage flow [Lazarsfeld et al 1944]
- Collective action, social movements [McAdam 1986, Chwe 1999]
- Viral marketing [Jurvetson 2000, Domingos-Richardson 2001]
- Current social media systems ...
Cascading behavior has become a basic “transport mechanism” in many kinds of networks.

- Agricultural, medical innovations [Ryan-Gross 1943, Coleman et al 1966]
- Media influence and two-stage flow [Lazarsfeld et al 1944]
- Collective action, social movements [McAdam 1986, Chwe 1999]
- Viral marketing [Jurvetson 2000, Domingos-Richardson 2001]
- Current social media systems ...

Diffusion and Contagion
Contagion in Networks

Book recommendations (Leskovec et al 2006)

On-line group formation (Backstrom-Huttenlocher-Kleinberg-Lan 2006)

Networks of sexual contacts (Potterat et al 02)

Overnight loans among financial institutions (Bech and Atalay, 2008)
Models of Contagion

Independent contagion: each exposure to an affected node carries a prob. $p > 0$ of becoming affected.

Threshold contagion: the $j^{\text{th}}$ exposure to an affected node carries a prob. $p_j$ of becoming affected [Granovetter 1978, Schelling 1978].

- $p_1 \geq p_2 \geq p_3 \geq \cdots$: E.g. social contagion with attenuating influence.
- $p_1 \leq p_2 \leq p_3 \leq \cdots$: E.g. financial contagion where earlier shocks weaken you.
An equivalent view: Threshold distributions

Instead of applying probs. $p_j$ throughout the process, each node $v$ chooses a threshold $\ell(v)$ at the start.

- $v$ will be affected as soon as it has $\ell(v)$ affected neighbors.
- From $p_1, p_2, p_3, \ldots$, obtain distribution $\mu$ on thresholds:
  $v$ chooses threshold $j$ with prob. $\mu(j)$.

Despite simple formulation, a challenging model to work with.
- Special-case results for diminishing thresholds
  ($\mu(1) \geq \mu(2) \geq \cdots$) [Kempe-Kleinberg-Tardos 03, Mossel-Roch 07].
- Special-case results when graph $G$ is a tree [Dodds-Watts 04], lattice [Cox-Durrett 91], or clique [Granovetter 78, Schelling 78].
Which networks are least susceptible to cascading failures?

- Take edge density out of consideration: set of all graphs where each node has \( d \) neighbors.
- Choose \( \mu \) from set of all distributions on \( \{0, 1, 2, \ldots, d + 1\} \).
- Risk of \( G = \) maximum failure probability of any node in \( G \) when thresholds are drawn from \( \mu \).

Given \( \mu \), which graphs have the lowest risk?

[Blume-Easley-Kleinberg-Kleinberg-Tardos 2011]
Intuition from epidemiology:
- Dangerous to belong to a large connected component: the clique $K_{d+1}$ is a resilient graph.

Intuition from financial markets:
- Want diversity among neighbors, uncorrelated shocks: the tree $T_d$ is a resilient graph.

These two forms of intuition are in direct opposition to each other.
To get further insight into the model:
Let’s test these intuitions on distributions of the form

\[(\mu(0), \mu(1), \mu(2)) = (s, t, 1 - s - t)\]

where \(s\) is fixed and small, and \(t\) varies. (All thresholds above 2 have prob. 0.)

- With threshold distribution \((s, 1 - s, 0)\), \(r\)’s failure prob. is monontonic in the size of its component.
- The clique is uniquely optimal.
Recall $(\mu(0), \mu(1), \mu(2)) = (s, t, 1 - s - t)$.

A first result:

- There exist $s, t$ (both small, with $t$ larger than $s$) so that the tree $T_d$ has lower risk than the clique $K_{d+1}$.
- Qualitative point: very different kinds of graphs are safer against different kinds of threshold contagion processes.
Cliques vs. Trees

Tree:

\[ s + \binom{d}{2}s^2 + dst + \text{terms of degree } \geq 3 \cdots \]

Clique:

\[ s + \binom{d}{2}s^2 + dst + d(d - 1)st + \text{terms of degree } \geq 3 \cdots \]

Now let \( s, t \to 0 \) so higher-order terms are negligible.
(Justification becomes subtle.)
Call a set of graphs $\mathcal{H}$ “sufficient” if:

- For every distribution $\mu$, some graph in $\mathcal{H}$ achieves minimum risk over all $d$-regular graphs.

Question:

- Do $K_{d+1}$ and $T_d$ form a two-element sufficient set?
- Is there a finite sufficient set for family of all $d$-regular graphs?
Sufficient Sets

- $K_{d+1}$ and $T_d$ form a sufficient set for $d = 2$.
- They don’t for $d \geq 3$.

To construct a graph with lower risk than both $K_{d+1}$ and $T_d$

- Study threshold distributions with very small $\mu(j)$ for $j = 0, 1, \ldots, h$, and remaining probability on $\mu(h+1)$.
- Write failure probability as a power series in $\mu(j)$ for $j \leq h$.
- Find graph-theoretic conditions for positive radius of convergence.
Social networks where people make decisions about new behaviors.

- User-defined groups in on-line communities; participation in on-line collaborative projects; decision to use a hashtag on Twitter; ...

- Many instances in Facebook data: accepting an invitation to join the site; clicking on an ad; liking a page; commenting on a post.

Does set/structure of adopting neighbors help predict tendency to adopt?

Network Neighborhoods

One person's network neighborhood:

- Their interface to the rest of the social network.
You’re more likely to do something when more friends are doing it. Why is that?

The issue of homophily/selection vs. influence

[Cohen 77, Kandel 78, Manski 93, Aral et al. 09, Shalizi-Thomas 11]

An experiment to sort out these effects

[Bakshy-Eckles-Yan-Rosenn 2012]
Structural Diversity

Dependence on number of friends: a first step toward general prediction.

- Given the full pattern of connections among your friends, estimate probability of adopting a new behavior.

Structural diversity

[Ugander-Backstrom-Marlow-Kleinberg]

- Data from invitations to join Facebook.
Structural Diversity

With four neighbors:
Your interface to the information is non-uniform.

- Algorithmically managed by systems like News Feed.
- Typical Facebook user writes 60-70% of comments to \( \approx 15 \) people.
  
  [Backstrom-Bakshy-Kleinberg-Lento-Rosenn 2011]
Finding Significant People

Given a person’s network neighborhood, can we identify their most significant social ties?

Theories of strong and weak ties [Granovetter 1973, 1985].
- **Embeddedness**: # of mutual friends shared by e’s endpoints.

If an edge is highly embedded, it is likely to be a stronger tie.
- **Rank neighbors by embeddedness?**
Network structure via neighborhoods

In practice: embeddedness finds many nodes from the largest cluster.
Network structure via neighborhoods

In practice: embeddedness finds many nodes from the largest cluster.

- Often this is a large collection of co-workers or college alumni friends. Compare: node in lower left — the spouse.
In practice: embeddedness finds many nodes from the largest cluster.

- Motivating question: Given a Facebook user in a relationship, find their partner just from network structure [Backstrom-Kleinberg 2014]
Instead of just counting mutual friends, look at their structure.

- How well connected are the common endpoints of edge $e$?
- If not well connected, suggests something about $v$-$w$ relationship.
- $v$-$w$ cannot be easily “explained” by any one social focus.

Type of bridging/brokerage role [Granovetter 73, Burt 92, Watts 99] but played jointly by $v$ and $w$, and implying a form of tie strength.
Dispersion

$$C_{vw} = \text{common neighbors of } v \text{ and } w.$$  

Sum of distances between pairs in $C_{vw}$, after deleting $v$ and $w$:

$$\sum_{s,t \in C_{vw}} d_{G \setminus \{v,w\}}(s,t).$$

The dispersion of edge $(v, w)$ with respect to distance function $d$.

- Based on a 0-1-valued metric, normalized by $|C_{vw}|$. 

\[ G \]

\[ w \]

\[ v \]
Can use many possible distance functions $d$ when summing over pairs of mutual neighbors.

- $d(s, t) = \begin{cases} 
0 & \text{if } (s, t) \text{ is an edge} \\
1 & \text{otherwise}
\end{cases}$

- $d(s, t) = \begin{cases} 
0 & \text{if shortest } s-t \text{ path avoiding } v, w \text{ has } \leq k \text{ edges} \\
1 & \text{otherwise}
\end{cases}$

- Many other choices for $d$ based on community detection, brokerage measures, spring embedding, ...

Can also normalize the dispersion:

$$\frac{\text{dispersion}(v, w)}{|C_{vw}|^\alpha}.$$ 

- Searching over choices of $k, \alpha$ shows $k = 2$ and $\alpha = 1$ nearly optimal.

- A slight improvement if we apply this recursively (details omitted here ... )
Evaluating the Methods

For evaluation, use 1.3 million Facebook users who:

- Declare a relationship partner in their profile (symmetric).
- Have between 50 and 2000 friends.
- Are at least 20 years old.

For each user $v$, rank all friends $w$ by competing metrics:

- Embeddedness of $v$-$w$ edge.
- Dispersion of $v$-$w$ edge.
- Number of photos in which $v$ and $w$ are both tagged.
- Number of times $v$ viewed $w$'s profile in last 90 days.

For what fraction of all users $v$ is the top-ranked $w$ the relationship partner?
<table>
<thead>
<tr>
<th>type</th>
<th>embed</th>
<th>dispersion</th>
<th>photo</th>
<th>profile view</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.247</td>
<td>0.506</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:

Embeddedness vs. dispersion
<table>
<thead>
<tr>
<th>type</th>
<th>embed</th>
<th>dispersion</th>
<th>photo</th>
<th>profile view</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.247</td>
<td>0.506</td>
<td>0.415</td>
<td>0.301</td>
</tr>
</tbody>
</table>

Notes:

Embeddedness vs. dispersion

Structural vs. activity-based
<table>
<thead>
<tr>
<th>type</th>
<th>embed</th>
<th>dispersion</th>
<th>photo</th>
<th>profile view</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.247</td>
<td>0.506</td>
<td>0.415</td>
<td>0.301</td>
</tr>
<tr>
<td>married</td>
<td>0.321</td>
<td>0.607</td>
<td>0.449</td>
<td>0.210</td>
</tr>
<tr>
<td>relationship</td>
<td>0.132</td>
<td>0.344</td>
<td>0.347</td>
<td>0.441</td>
</tr>
</tbody>
</table>

Notes:

- Embeddedness vs. dispersion
- Structural vs. activity-based
- Married vs. in a relationship
<table>
<thead>
<tr>
<th>type</th>
<th>embed</th>
<th>dispersion</th>
<th>photo</th>
<th>profile view</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.247</td>
<td>0.506</td>
<td>0.415</td>
<td>0.301</td>
</tr>
<tr>
<td>married</td>
<td>0.321</td>
<td>0.607</td>
<td>0.449</td>
<td>0.210</td>
</tr>
<tr>
<td>married (female)</td>
<td>0.296</td>
<td>0.551</td>
<td>0.391</td>
<td>0.202</td>
</tr>
<tr>
<td>married (male)</td>
<td>0.347</td>
<td>0.667</td>
<td>0.511</td>
<td>0.220</td>
</tr>
<tr>
<td>relationship</td>
<td>0.132</td>
<td>0.344</td>
<td>0.347</td>
<td>0.441</td>
</tr>
<tr>
<td>relationship (female)</td>
<td>0.139</td>
<td>0.316</td>
<td>0.290</td>
<td>0.467</td>
</tr>
<tr>
<td>relationship (male)</td>
<td>0.125</td>
<td>0.369</td>
<td>0.399</td>
<td>0.418</td>
</tr>
</tbody>
</table>

Notes:

- Embeddedness vs. dispersion
- Structural vs. activity-based
- Married vs. in a relationship
- Female vs. male
<table>
<thead>
<tr>
<th>type</th>
<th>embed</th>
<th>dispersion</th>
<th>photo</th>
<th>profile view</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.247</td>
<td>0.506</td>
<td>0.415</td>
<td>0.301</td>
</tr>
<tr>
<td>married</td>
<td>0.321</td>
<td>0.607</td>
<td>0.449</td>
<td>0.210</td>
</tr>
<tr>
<td>married (female)</td>
<td>0.296</td>
<td>0.551</td>
<td>0.391</td>
<td>0.202</td>
</tr>
<tr>
<td>married (male)</td>
<td>0.347</td>
<td>0.667</td>
<td>0.511</td>
<td>0.220</td>
</tr>
<tr>
<td>relationship</td>
<td>0.132</td>
<td>0.344</td>
<td>0.347</td>
<td>0.441</td>
</tr>
<tr>
<td>relationship (female)</td>
<td>0.139</td>
<td>0.316</td>
<td>0.290</td>
<td>0.467</td>
</tr>
<tr>
<td>relationship (male)</td>
<td>0.125</td>
<td>0.369</td>
<td>0.399</td>
<td>0.418</td>
</tr>
</tbody>
</table>

Notes:

- Embeddedness vs. dispersion
- Structural vs. activity-based
- Married vs. in a relationship
- Female vs. male

Combining all via machine learning: 0.716 married, 0.682 relationship
<table>
<thead>
<tr>
<th>type</th>
<th>embed</th>
<th>dispersion</th>
<th>photo</th>
<th>profile view</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.247</td>
<td>0.506</td>
<td>0.415</td>
<td>0.301</td>
</tr>
<tr>
<td>married</td>
<td>0.321</td>
<td>0.607</td>
<td>0.449</td>
<td>0.210</td>
</tr>
<tr>
<td>married (female)</td>
<td>0.296</td>
<td>0.551</td>
<td>0.391</td>
<td>0.202</td>
</tr>
<tr>
<td>married (male)</td>
<td>0.347</td>
<td>0.667</td>
<td>0.511</td>
<td>0.220</td>
</tr>
<tr>
<td>relationship</td>
<td>0.132</td>
<td>0.344</td>
<td>0.347</td>
<td>0.441</td>
</tr>
<tr>
<td>relationship (female)</td>
<td>0.139</td>
<td>0.316</td>
<td>0.290</td>
<td>0.467</td>
</tr>
<tr>
<td>relationship (male)</td>
<td>0.125</td>
<td>0.369</td>
<td>0.399</td>
<td>0.418</td>
</tr>
</tbody>
</table>

Notes:

- Embeddedness vs. dispersion
- Structural vs. activity-based
- Married vs. in a relationship
- Female vs. male
- Combining all via machine learning: 0.716 married, 0.682 relationship
- Approx 34-38% of dispersion’s incorrect guesses are family members.
<table>
<thead>
<tr>
<th>type</th>
<th>embed</th>
<th>dispersion</th>
<th>photo</th>
<th>profile view</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.247</td>
<td>0.506</td>
<td>0.415</td>
<td>0.301</td>
</tr>
<tr>
<td>married</td>
<td>0.321</td>
<td>0.607</td>
<td>0.449</td>
<td>0.210</td>
</tr>
<tr>
<td>married (female)</td>
<td>0.296</td>
<td>0.551</td>
<td>0.391</td>
<td>0.202</td>
</tr>
<tr>
<td>married (male)</td>
<td>0.347</td>
<td>0.667</td>
<td>0.511</td>
<td>0.220</td>
</tr>
<tr>
<td>relationship</td>
<td>0.132</td>
<td>0.344</td>
<td>0.347</td>
<td>0.441</td>
</tr>
<tr>
<td>relationship (female)</td>
<td>0.139</td>
<td>0.316</td>
<td>0.290</td>
<td>0.467</td>
</tr>
<tr>
<td>relationship (male)</td>
<td>0.125</td>
<td>0.369</td>
<td>0.399</td>
<td>0.418</td>
</tr>
</tbody>
</table>

Notes:

- Embeddedness vs. dispersion
- Structural vs. activity-based
- Married vs. in a relationship
- Female vs. male
- Combining all via machine learning: 0.716 married, 0.682 relationship
- Approx 34-38% of dispersion’s incorrect guesses are family members.
married, individual features

performance (precision at first position) vs. time since relationship reported (months)

- rec disp
- photo
- emb
- view

relationship, individual features

- view
- photo
- rec disp
- emb
married, individual features

performance (precision at first position) vs. time since relationship reported (months)

relationship, individual features

performance (precision at first position) vs. time since relationship reported (months)
Probability a user transitions to ‘single’ status in next 60 days.

- Relationships where dispersion is correct vs. incorrect.
- Separately over relationships in 2-month age ranges.
A schematic picture for a node’s neighborhood:

A constant number of homogeneous clusters.
A General Structure for Network Neighborhoods

A schematic picture for a node’s neighborhood:

A constant number of homogeneous clusters.
Plus a constant number of nodes that defy classification.
Final Reflections

- Deeper understanding of how neighborhoods modulate information flow.
- Characterizing very large graphs through family of network neighborhoods [Ugander-Backstrom-Kleinberg 2013].
- As information spreads, what structures are traced out? What can we infer about the process from the structure? [Liben-Nowell-Kleinberg 2008, Golub-Jackson 2010, Chierichetti et al 11, Adamic-Lento-Fiore 12]