Information Flow and Graph Structure in On-Line Social Networks

Jon Kleinberg

Cornell University

Two Metaphors for the Web

On-line networks are balanced between two metaphors.

- **The library:** pages, hyperlinks, associations.
- **The crowd:** real-time awareness, memes, contagion.
Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them ... There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record.

— Vannevar Bush, *As We May Think*, 1945
Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them ... There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record.
— Vannevar Bush, As We May Think, 1945

.... radio and the printed page seemed to have only negligible effects on actual vote decisions .... When [people] were asked what had contributed to their decision, their answer was: other people.
— Elihu Katz and Paul Lazarsfeld, Personal Influence, 1955

Diffusion of innovations:
Ryan-Gross 43, Lazarsfeld et al 44, Coleman et al 66, Friedkin-Johnsen 90, Blume 93, Ellison 93, Domingos-Richardson 01, Kempe et al 03
What are on-line social networks accomplishing for their users?

1. Transport mechanism for information, opinions, behaviors.
2. Assistance for maintaining social ties over time.
Dear All, The US Congress has authorised the President of the US to go to war against Iraq. Please consider this an urgent request.

UN Petition for Peace:

[...]

Please COPY (rather than Forward) this e-mail in a new message, sign at the end of the list, and send it to all the people whom you know. If you receive this list with more than 500 names signed, please send a copy of the message to:

usa@un.int
president@whitehouse.gov

Chain-letter petitions as “tracers” through global social network
[LibenNowell-Kleinberg]
Analyses of information propagation in many domains.
Natural tree structure: $w$ acquires from $v \implies v$ is parent of $w$.

- Product recommendations [Leskovec et al 2006]
- Quoted phrases through news, blogs [Leskovec-Backstrom-Kleinberg 09]
- Facebook copy-paste and photo memes [Adamic et al 12, Cheng et al 14]
- Cascading invitations to join new platforms [Anderson et al 2015]
A first issue: networks have very low diameter, but trees are deep.
[Kleinberg-LibenNowell, Iribarren-Moro, Golub-Jackson]

- Selective sharing producing a sparse subgraph.
- Particular role for strong ties.
- Large heterogeneity in rate of node response.
  cf. literature on shortest paths with random edge lengths

Open: A reasonable model of tree depth with provable guarantees.
Roughly 30-40% of Facebook image/video memes recur. [Cheng-Adamic-Kleinberg-Leskovec 2016]

- Can define recurrence by inferring an on/off state transition in an underlying generative model for volume, or by simple thresholding.
Outbreaks of Moderate Size

Conditioned on size of first outbreak, expected number of future outbreaks maximized for moderate-sized cascades.

Consistent with a simple probabilistic contagion model:
- Meme appears spontaneously at random nodes at randomly spaced times.
- Nodes accept meme from neighbors with fixed probability $p$.
- Nodes have reduced probability $p' < p$ after first exposure (a variant of SIR epidemic model).

Giant cascades exhaust population in first outbreak; small cascades never get going.
Dear All, The US Congress has authorised the President of the US to go to war against Iraq. Please consider this an urgent request.

UN Petition for Peace:

[...]

Please COPY (rather than Forward) this e-mail in a new message, sign at the end of the list, and send it to all the people whom you know. If you receive this list with more than 500 names signed, please send a copy of the message to:

usa@un.int
president@whitehouse.gov

Does the language used help us predict the success of the meme?

e.g. [Hovland et al 1953; Nickerson-Rogers 2010; Milkman-Berger 2012]
Same user tweeting same URL with different text, within 12 hours. [Tan-Lee-Pang 2014]

- First post vs. second post gives essentially no predictive value.
- Are there useful features in the language of the tweet?

Human performance 61.3%; algorithmic performance 65.6%

- Key features included probability of tweet under language models built from: universe of Tweets; user’s past tweets; successful tweets.
Similar algorithmic performance on a corpus of movie quotes. [DanescuNiculescuMizil-Cheng-Kleinberg-Lee 2012]

- Compare pairs of movie lines of approx. same length, spoken by same character in same scene of same movie.
- Algorithmic performance 64%; now human performance 75%.

Stormtrooper: Let me see your identification.
Obi-Wan: You don’t need to see his identification.
Stormtrooper: We don’t need to see his identification.
Obi-Wan: These aren’t the droids you’re looking for.
Stormtrooper: These aren’t the droids we’re looking for.
Obi-Wan: He can go about his business.
Stormtrooper: You can go about your business.
These aren’t the droids you’re looking for.

These aren’t the droids you’re looking for.

These aren’t the droids you’re looking for.

These aren’t the droids you’re looking for.

These aren’t the droids you’re looking for.

These aren’t the droids you’re looking for.

Google

We do other things besides look for droids, but that’s all anyone ever remembers.

Bye, daddy. I hope you find the droids you’re looking for.
Mutation of textual memes as they travel from source to source.

- Used for phrase clustering in Leskovec-Backstrom-Kleinberg 2009
- More extensive analysis by Simmons-Adamic-Adar 2011

Genetic analogy for memes: beginning of a formalization?

- Fitness functions
- Mutation mechanisms
- Preservation of “functional” elements
What are on-line social networks accomplishing for their users?

1. Transport mechanism for information, opinions, behaviors.

2. Assistance for maintaining social ties over time.
One person’s network neighborhood.

Think of Facebook not as a billion-node network, but instead as a collection of a billion (relatively dense) small networks. [Ugander-Backstrom-Kleinberg 2013]
A Baseline Model

\( G_{n,p} \): place \( n \) nodes; connect each pair independently with probability \( p \).

- Erdős-Rényi 1960, Gilbert 1959

Deficiencies with the \( G_{n,p} \) model:

- Doesn’t produce nodes with enormous numbers of neighbors. (More a problem for Web graphs than social networks.)
- Real social networks are rich in triangles: triadic closure.
A Baseline Model

$G_{n,p}$: place $n$ nodes; connect each pair independently with probability $p$.

- Erdös-Rényi 1960, Gilbert 1959

Deficiencies with the $G_{n,p}$ model:

- Doesn’t produce nodes with enormous numbers of neighbors. (More a problem for Web graphs than social networks.)
- Real social networks are rich in triangles: triadic closure.
A Baseline Model

$G_{n,p}$: place $n$ nodes; connect each pair independently with probability $p$.

- Erdős-Rényi 1960, Gilbert 1959

Deficiencies with the $G_{n,p}$ model:

- Doesn’t produce nodes with enormous numbers of neighbors. (More a problem for Web graphs than social networks.)
- Real social networks are rich in triangles: triadic closure.
Describe neighborhood $G$ by vector of subgraph frequencies: For small $k$, and each $k$-node graph $H$, let $f_G(H) = \text{frac. of } k\text{-node sets inducing } H$.

- Triad census: Davis-Leinhardt 71
- Network motifs: Milo et al 02
- Frequent subgraph mining: Yan-Han 02, Kuramochi-Karypis 04
- Subgraph density, homomorphisms: Borgs-Chayes-Lovász-Sós-Vesztergombi, Razborov 07
- Characterizing neighborhoods: Ugander-Backstrom-Kleinberg 13

\[
\begin{align*}
(x_1, x_2, x_3, x_4) &= (0.18, 0.37, 0.14, 0.31) \\
(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11})
\end{align*}
\]
Describe neighborhood $G$ by vector of subgraph frequencies: For small $k$, and each $k$-node graph $H$, let $f_G(H) = \text{frac. of } k\text{-node sets inducing } H$.

- **Triad census**: Davis-Leinhardt 71
- **Network motifs**: Milo et al 02
- **Frequent subgraph mining**:
  - Yan-Han 02, Kuramochi-Karypis 04
- **Subgraph density, homomorphisms**:
  - Borgs-Chayes-Lovász-Sós-Vesztergombi, Razborov 07
- **Characterizing neighborhoods**:
  - Ugander-Backstrom-Kleinberg 13

\[
\begin{align*}
(x_1, x_2, x_3, x_4) &= (0.18, 0.37, 0.14, 0.31) \\
(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11})
\end{align*}
\]
The geography of Facebook neighborhoods

Axes: triad frequencies

Axes: triad frequencies
The geography of Facebook neighborhoods

Axes: triad frequencies

- “Coastlines:” freq of 1-edge triad is $\leq \frac{3}{4}$.
- Unpopulated areas: freq of 2-edge triad never close to $\frac{3}{4}$ in real life.
- Full feasible region contains hard extremal graph theory questions [Razborov 2007].
The geography of Facebook neighborhoods

Axes: triad frequencies

- “Coastlines:” freq of 1-edge triad is $\leq \frac{3}{4}$.
- Unpopulated areas: freq of 2-edge triad never close to $\frac{3}{4}$ in real life.
- Full feasible region contains hard extremal graph theory questions [Razborov 2007].

$G_{n,p}$ is the “backbone” that runs through the points.
- With deviations based on triadic closure.
The Role of Triadic Closure

Continuous-time Markov chain on unlabeled, undirected graphs. For each pair of nodes $v, w$:
- If $v, w$ not linked, they form an edge at rate $\gamma$.
- If $v, w$ linked, they drop their edge at rate $\delta$.

For each three-node path on $v, u, w$:
- $v$ and $w$ form an edge at rate $\lambda$. 
The Model on Network Data

Define $\nu = \gamma/\delta =$ formation / destruction.

- If $\lambda = 0$, the Markov chain produces the $G_{n,p}$ distribution.
- With just $\lambda$ and $\nu$, a much richer family of structures.
Finding Significant People

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

- Central issue in ranking shared content.

Triangles play a central role here. Theories of strong and weak ties [Granovetter 1973, 1985].

Embeddedness of edge $e$: number of triangles containing $e$. Equivalently: # of mutual friends shared by $e$’s endpoints. If an edge is highly embedded, it is likely to be a stronger tie.
Finding Significant People

Given a person’s network neighborhood, can we identify their most significant social ties?
- Central issue in ranking shared content.

Triangles play a central role here.
- Theories of strong and weak ties [Granovetter 1973, 1985].
- Embeddedness of edge $e$: number of triangles containing $e$.
- Equivalently: # 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.
Network structure via neighborhoods

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 [Burt 1992] 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 $w$:

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

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

- Should use 0-1-valued metric; normalize by $|C_{vw}|$. 

Can use many possible functions $d$.

\[ disp(v, w) = \sum_{s,t \in C_{vw}} d_G - \{v,w\}(s, t). \]

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

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

Can also normalize the dispersion:

\[ \frac{disp(v, w)}{|C_{vw}|^\alpha}. \]

- Analogue of clustering coefficient [Watts-Strogatz 98] is $k = 1$ and $\alpha = 2$.

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

time since relationship reported (months)

relationship, individual features

view

photo

rec disp

emb
A General Structure for Network Neighborhoods

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.
Further Directions

Other information latent in network neighborhoods?

- Structural signatures for other types of roles in social circles?
- Relation of neighborhood structure to the spread of information [Blume-Easley-Kleinberg-Kleinberg-Tardos 2011].
- Implications for designing systems to share information [Backstrom-Kleinberg-Lee-DanescuNiculescuMizil 2013].
Final Reflections

MySpace is doubly awkward because it makes public what should be private. It doesn’t just create social networks, it anatomizes them. It spreads them out like a digestive tract on the autopsy table. You can see what’s connected to what, who’s connected to whom.


- Social networks — implicit for millenia — are being recorded at high resolution.

- What is the right framework for capturing the structures and phenomena that we see?

- What are the dangers of stockpiling this much personal data?

- Opportunity for fundamental computational and mathematical models to inform the next steps on all these questions.