Exploring the structure of online social networks: Role of positive and negative links

Jure Leskovec
Computer Science, Stanford University
Joint work with Daniel Huttenlocher and Jon Kleinberg
Web – Diverse applications

Diverse on-line computing applications
Rich and highly dynamic content and interactions
Web – Billions of users

Billions of users,
Massive traces of human activity
Web – Immense possibilities

Web is like a “laboratory” for studying billions of humans
Web – Interaction networks

Model as an interaction network
Many data is a Network!

- Internet
- Citations
- Web
- Brain/neurons
- Text
- Genes/proteins
- Software
- Friendship
Traditional obstacle:
Can only choose 2 of 3:
- Large-scale
- Realistic
- Completely mapped

Now: Large on-line systems with detailed records of social activity
- Communities: Facebook, MySpace, LiveJournal
- Communication: Twitter, blogs, markets, IM
- On-line publication repositories: arXiv, MedLine
Networks: Size *matters*

- Network data spans many orders of magnitude:
  - 436-node network of email exchange over 3-months at corporate research lab [Adamic-Adar, SocNets ‘03]
  - 43,553-node network of email exchange over 2 years at a large university [Kossinets-Watts, Science ‘06]
  - 4.4-million-node network of declared friendships on a blogging community [Liben-Nowell et al., PNAS ‘05]
  - 240-million-node network of communication on Microsoft Messenger [Leskovec-Horvitz, WWW‘08]

- Massive data – Benefits: Patterns become “visible”
Rich social structure in online computing applications

Such structures are modeled by networks

Most social network analyses view links as positive

- Friends
- Fans
- Followers

But generally links can convey either friendship or antagonism
Study online social structure using large scale data
- Networks indicating positive and negative relations

Question:
- How do edge signs and network structure interact?
- How to model and predict edge signs?

Goal:
- Offer insights into how online computing systems are being used

Applications:
- Friend recommendation:
  - Easy to predict whether you know someone vs. to predict what you think of them
Our work: Nets with explicit signs

- Each link $A \rightarrow B$ is explicitly tagged with a sign:
  - **Epinions**: Trust/Distrust
    - Does A trust B’s product reviews? (only positive links are visible)
  - **Wikipedia**: Support/Oppose
    - Does A support B to become Wikipedia administrator?
  - **Slashdot**: Friend/Foe
    - Does A like B’s comments?
How to reason about links and signs?

Start with intuition:

- Friend of my friend is my friend
- Enemy of friend is my enemy
- Friend of enemy is my enemy
- Enemy of enemy is my friend

Look at connected triples of nodes that are consistent with this logic.
Structural Balance

- Three-Node Signed Triads [Heider ’46]

Balanced

Consistent with “friend of a friend” or “enemy of the enemy” intuition

Unbalanced

Inconsistent with the “friend of a friend” or “enemy of the enemy” intuition
Links are directed

Balance theory:
- Traditionally: disregard direction, apply balance
- Status theory [Davis-Leinhardt ‘68, Guha et al. ’04, Leskovec et al. ‘10]
  - Link $A \rightarrow B$ means: B has higher status than A
  - Link $A \leftarrow B$ means: B has lower status than A

Status and balance can give different predictions:
Plan for the talk

- How do these two theories align with ways users create links:
  - Not just “which is right” but how are aspects of each reflected in the data
  - Provide insights into how these linking systems are being used

- Outline:
  - Study links as undirected: Balance theory
  - Study links as directed and evolving: Status theory
- Consider networks as undirected
- Compare frequencies of signed triads in real and shuffled data
  - 4 triad types $t$:
    - ![Triad Type Diagram]
  - **Surprise** ($z$-score) for triad type $t$:
    - Number of std. deviations by which real number of $t$ differs from the expected number in shuffled data
**Undirected links: Balance**

- **Surprise values:**

- **Observations:**
  - Strong signal for balance
  - Consistency with Davis’s [1967] weak balance
  - Epinions and Wikipedia agree on all types
  - Question for later: How can all triads with negative edges be underrepresented in Slashdot?

<table>
<thead>
<tr>
<th>Triad</th>
<th>Epin</th>
<th>Wiki</th>
<th>Slashdot</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ + +</td>
<td>1,881</td>
<td>380</td>
<td>927</td>
</tr>
<tr>
<td>- + -</td>
<td>-2,105</td>
<td>-573</td>
<td>-824</td>
</tr>
<tr>
<td>+ - +</td>
<td>249</td>
<td>289</td>
<td>-175</td>
</tr>
<tr>
<td>- - +</td>
<td>288</td>
<td>11</td>
<td>-9</td>
</tr>
</tbody>
</table>
Our networks are really directed
  - trust, opinion (friendship)

How many $\triangle$ are now explained by balance?
  - Half (8 out of 16)

Is there a better explanation?
  - Yes. **Theory of Status.**
Links are directed and created over time

**Status theory** [Davis-Leinhardt ‘68, Guha et al. ’04, Leskovec et al. ‘10]

- Link $A \rightarrow B$ means: $B$ has higher status than $A$
- Link $A \leftarrow B$ means: $B$ has lower status than $A$

Status and balance can give different predictions:
Edges are directed
Edges are created over time
- X has links to A and B
- Now, A links to B (triad A-B-X)
- How does sign of A-B depend on signs of X?

We need to formalize:
- Links are embedded in triads – provides context for signs
- Users are heterogeneous in their linking behavior
A contextualized link is a triple \((A,B;X)\) such that a directed \(A-B\) link forms after there is a two-step semi-path \(A-X-B\). A-X and B-X links can have either direction and either sign: 16 possible types.
Users differ in frac. of + links they give/receive

For a user U:
- **Generative baseline** $p_g(U)$: frac. of + given by U
- **Receptive baseline** $p_r(U)$: frac. of + received by U

Basic kinds of questions:
- How do different link contexts cause users to deviate from baselines?
- Link contexts as modifiers on a person’s predicted behavior
For a triad of type \( t \) let \((A_1, B_1; X_1) \ldots (A_k, B_k; X_k)\) be all instances of triad \( t \):

- **Generative baseline of \( t \):**
  - sum of gen. baselines \( p_g(A_i) \) of \( A_i \)

- **Receptive baseline of \( t \):**
  - sum of rec. baselines \( p_r(B_i) \) of \( B_i \)

**Surprise:** How much behavior of users \( A_i/B_i \) deviates from **baseline** when they are in context \( t \):

- **Generative surprise of \( t \):** number of std. deviations by which num. of positive \( A_i-B_i \) links differs from generative baseline of \( t \)

- **Receptive surprise for \( t \):** number of std. deviations by which num. of positive \( A_i-B_i \) links differs from receptive baseline of \( t \)
Two basic examples:

More negative than gen. baseline of A
More negative than rec. baseline of B
X positively endorses A and B
Now A links to B

A puzzle:
In our data we observe:
Fraction of positive links deviates
- Above generative baseline of A
- Below receptive baseline of B

Why?
A story: Soccer team

- Ask every node: How does skill of B compare to yours?
  - Build a signed directed network

- We haven’t asked A about B
- But we know that X thinks A and B are both better than him

- What can we infer about A’s answer?
A story: Soccer team

- A’s viewpoint:
  - Since B has positive evaluation, B is high status
  - Thus, evaluation A gives is more likely to be positive than the baseline

How does A evaluate B?
A is evaluating someone who is better than avg.

\[ Y \quad \text{A} \quad \text{X} \quad \text{B} \]

Y... average node
In brief: Users on Epinions behave like members of our hypothetical soccer team.
Determine node status:
- Assign X status 0
- Based on signs and directions of edges set status of A and B

Surprise is **status**-consistent, if:
- **Gen.** surprise is status-consistent if it has *same* sign as status of B
- **Rec.** surprise is status-consistent if it has the *opposite* sign from the status of A

Surprise is **balance**-consistent, if:
- If it completes a balanced triad

Status-consistent if:
- Gen. surprise > 0
- Rec. surprise < 0
## Predictions:

<table>
<thead>
<tr>
<th>$t_i$</th>
<th>count</th>
<th>$P(\cdot)$</th>
<th>$s_{out}$</th>
<th>$s_{in}$</th>
<th>$B_{out}$</th>
<th>$B_{in}$</th>
<th>$S_{out}$</th>
<th>$S_{in}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>178,051</td>
<td>0.97</td>
<td>95.9</td>
<td>197.8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_2$</td>
<td>45,797</td>
<td>0.54</td>
<td>-151.3</td>
<td>-229.9</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>○</td>
</tr>
<tr>
<td>$t_3$</td>
<td>246,371</td>
<td>0.94</td>
<td>89.9</td>
<td>195.9</td>
<td>✓</td>
<td>✓</td>
<td>○</td>
<td>✓</td>
</tr>
<tr>
<td>$t_4$</td>
<td>25,384</td>
<td>0.89</td>
<td>1.8</td>
<td>44.9</td>
<td>○</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_5$</td>
<td>45,925</td>
<td>0.30</td>
<td>18.1</td>
<td>-333.7</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_6$</td>
<td>11,215</td>
<td>0.23</td>
<td>-15.5</td>
<td>-193.6</td>
<td>○</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_7$</td>
<td>36,184</td>
<td>0.14</td>
<td>-53.1</td>
<td>-357.3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_8$</td>
<td>61,519</td>
<td>0.63</td>
<td>124.1</td>
<td>-225.6</td>
<td>✓</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_9$</td>
<td>338,238</td>
<td>0.82</td>
<td>207.0</td>
<td>-239.5</td>
<td>✓</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_{10}$</td>
<td>27,089</td>
<td>0.20</td>
<td>-110.7</td>
<td>-449.6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_{11}$</td>
<td>35,093</td>
<td>0.53</td>
<td>-7.4</td>
<td>-260.1</td>
<td>○</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_{12}$</td>
<td>20,933</td>
<td>0.71</td>
<td>17.2</td>
<td>-113.4</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_{13}$</td>
<td>14,305</td>
<td>0.79</td>
<td>23.5</td>
<td>24.0</td>
<td>○</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$t_{14}$</td>
<td>30,235</td>
<td>0.69</td>
<td>-12.8</td>
<td>-53.6</td>
<td>○</td>
<td>○</td>
<td>✓</td>
<td>○</td>
</tr>
<tr>
<td>$t_{15}$</td>
<td>17,189</td>
<td>0.76</td>
<td>6.4</td>
<td>24.0</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>✓</td>
</tr>
<tr>
<td>$t_{16}$</td>
<td>4,133</td>
<td>0.77</td>
<td>11.9</td>
<td>-2.6</td>
<td>✓</td>
<td>○</td>
<td>✓</td>
<td>○</td>
</tr>
</tbody>
</table>

Number of correct predictions: | 8 | 7 | 14 | 13 |
Status seems to fail on triad types where A and B have both low status relative to X

- Mistakes for gen. surprise (a) and (b) are “duals” of mistakes for receptive surprise (c) and (d)
- Type (a) is one of most basic cases of balance theory
Intuitive picture of social network in terms of densely linked clusters

How does structure interact with links?

Embeddedness of link (A,B): number of shared neighbors
Global structure: Embeddedness

- Embeddedness of ties:
  - Embedded ties tend to be more positive

- A natural connection to closure based social capital [Coleman ‘88]

- Public display of signs (votes) in Wikipedia further strengthens this
Both theories make predictions about the global structure of the network

- **Structural balance – Factions**
  - Put nodes into groups such that the number of in group “+” and between group “-” edges is maximized

- **Status theory – Global Status**
  - Flip direction and sign of negative edges
  - Assign each node a unique status value so that most edges point from low to high
Fraction of edges of the network that satisfy Balance and Status?

Observations:
- No evidence for global balance beyond the random baselines
  - Real data is 80% consistent vs. 80% consistency under random baseline
- Evidence for global status beyond the random baselines
  - Real data is 80% consistent, but 50% consistency under random baseline
Predicting edge signs

Edge sign prediction problem
- Given a network and signs on all but one edge, predict the missing sign

Machine Learning formulation:
- Predict sign of edge (u, v)
- Class label:
  - +1: positive edge
  - -1: negative edge
- Learning method:
  - Logistic regression

\[
P(+) = \frac{1}{1 + e^{-(b_0 + \sum_{i} b_i x_i)}}
\]

- Dataset:
  - Original: 80% +edges
  - Balanced: 50% +edges
- Evaluation:
  - Accuracy and ROC curves
- Features for learning:
  - Next slide
For each edge \((u,v)\) create features:

- **Triad counts (16):**
  - Counts of signed triads edge \(u \rightarrow v\) takes part in

- **Degree (7 features):**
  - Signed degree:
    - \(d_{out}^+(u), d_{out}^-(u), d_{in}^+(v), d_{in}^-(v)\)
  - Total degree:
    - \(d_{out}(u), d_{in}(v)\)
  - Embeddedness of edge \((u,v)\)
Error rates:
- Epinions: 6.5%
- Slashdot: 6.6%
- Wikipedia: 19%

Signs can be modeled from local network structure alone
- Trust propagation model of [Guha et al. ‘04] has 14% error on Epinions

Triad features perform less well for less embedded edges

Wikipedia is harder to model:
- Votes are publicly visible
Do people use these very different linking systems by obeying the same principles?

- How generalizable are the results across the datasets?
  - Train on row “dataset”, predict on “column”

Almost perfect generalization of the models even though networks come from very different applications

<table>
<thead>
<tr>
<th></th>
<th>All23</th>
<th>Epinions</th>
<th>Slashdot</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epinions</td>
<td>0.9342</td>
<td>0.9289</td>
<td>0.7722</td>
<td></td>
</tr>
<tr>
<td>Slashdot</td>
<td>0.9249</td>
<td>0.9351</td>
<td>0.7717</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.9272</td>
<td>0.9260</td>
<td>0.8021</td>
<td></td>
</tr>
</tbody>
</table>
Signed networks provide insight into how social computing systems are used:

- Status vs. Balance
- Different role of reciprocated links
  - Only 3-5% reciprocal, but those follow balance more closely
- Role of embeddedness and public display

Sign of relationship can be reliably predicted from the local network context
- ~90% accuracy sign of the edge
More evidence that networks are globally organized based on status

People use signed edges consistently regardless of particular application
  - Near perfect generalization of models across datasets

Many further directions:
  - Status difference [ICWSM ‘10]
Conclusions: Status

- Status difference on Wikipedia:
THANKS!

Data + Code:

http://snap.stanford.edu
Reciprocation of edges

- Between 3 to 5% of edges are reciprocated
- Different predictions:
  - **Balance**: reciprocation is *same* sign
  - **Status**: reciprocation is *different* sign
- Observations:
  - **Positive links**: strongly follow balance
  - **Negative links**: pos. links occur more than neg. links but lower than baseline
- **Role of triadic structure**:
  - Suppose A-B is part of a triad and B reciprocates
  - B is more likely to reverse the sign if the triad was *unbalanced*
### Undirected: Balance

- Consider networks as undirected
- Frequencies of signed triads:

<table>
<thead>
<tr>
<th>Triad</th>
<th>Epinions</th>
<th>Slashdot</th>
<th>Wikipedia</th>
<th>Conforms to balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Triad Diagram]</td>
<td>P(T)</td>
<td>P₀(T)</td>
<td>P(T)</td>
<td>P₀(T)</td>
</tr>
<tr>
<td>![Triad Diagram]</td>
<td>0.87</td>
<td>0.62</td>
<td>0.84</td>
<td>0.46</td>
</tr>
<tr>
<td>![Triad Diagram]</td>
<td>0.07</td>
<td>0.05</td>
<td>0.07</td>
<td>0.12</td>
</tr>
<tr>
<td>![Triad Diagram]</td>
<td>0.05</td>
<td>0.32</td>
<td>0.08</td>
<td>0.40</td>
</tr>
<tr>
<td>![Triad Diagram]</td>
<td>0.007</td>
<td>0.003</td>
<td>0.012</td>
<td>0.012</td>
</tr>
</tbody>
</table>

P(T) ... probability of a triad
P₀(T)... triad prob. if edge signs are shuffled