Social Influence Analysis and Measurement

Jing Zhang
Information School
Renmin University

Collaborate with
Wei Chen (MSRA), Cane Leung (Huawei Noah’s Ark), Hanghang Tong (ASU), Jimeng Sun (GIT), Jie Tang (THU), Juanzi Li (THU)
What is Social Influence?

• Social influence occurs when one's opinions, emotions, or behaviors are affected by others, intentionally or unintentionally.\(^1\)
  – Peer Pressure
  – Opinion leadership
  – Conformity
  – …

Two-step Flow Theory

Mass Media

Opinion leader

Individuals in social contact with an opinion leader
The theory of “Three Degree of Influence”

You are able to influence up to >1,000,000 persons in the world, according to the Dunbar’s number[3].

Asch’s Experiment

Which line matches the first line, A, B, or C?

74% of the participants followed the majority judgment on at least one trial, even when the majority was wrong.
Experiment on Voting

  – Will online political mobilization really work?

A controlled trial (with 61M users on FB)

- Social msg group: was shown with msg that indicates one’s friends who have made the votes.
- Informational msg group: was shown with msg that indicates how many other.

Social msg group were \textbf{2.08\%} more likely to click on the “I Voted” button.
Virtual Marketing

- Influence maximization
  - Initially targeting a few “influential” seeds, to trigger a maximal number of individuals to adopt the opinions/products through friend recommendation.
Existing Research

- **Influence Test**
  - Statistical causal inference
    [Arala et al. 2009] [La Fond and Neville 2010] [Anagnostopoulos et al. 2008]
  - Real controlled trials [Bakshy et al. 2012] [Bond et al. 2012]

- **Influence Learning**
  - Node influence [Weng et al. 2010]
  - Pairwise influence [Saito et al. 2008]
  - Group influence [Tang et al. 2013]

- **Influence Model**
  - Independent cascade model [Kemp et al, 2003]
  - Linear threshold model [Kemp et al, 2003]
Outline

• Node influence
  – Conformity influence

• Pairwise influence
  – Link influence

• Group influence
  – Structural influence

• An important assumption
  – A is more likely to be influence by B if A’s behaviors frequently follow B’s.
Conformity Influence

Who is more likely to conform to others, $v_1$ or $v_2$?

- Conformity is the *inclination* of a person to be influenced by others by yielding to perceived group pressure and copying the behavior and beliefs of others [Jenness 1932; Sherif 1935].

Asch’s Experiment
Formalize Conformity Influence

- Conformity theory [Bernheim 1994]
  - Everyone in a group expresses her own individuality.
  - Yet, even individualists pursue somewhat for status (esteem or popularity) and change their choices toward the social norm.

- Formalize conformity theory by a utility function:

\[
 f(y_i) = (1 - \lambda_i) d(y_i, \hat{y}_i) + \lambda_i \sum_{j \in N(i)} d(y_i, y_j) 
\]

- There exists Nash equilibria if all users in a network make the decisions for a given action according to the utility function.

\[ y_i = 1: \text{adopt an action} \]
\[ y_i = 0: \text{do not adopt an action} \]

\( \lambda_i \) represents the conformity tendency of \( v_i \)
To solve the **data sparsity** problem, we extend the utility function by incorporating role and topic.

- Conformity tendency is different for persons with different roles.
- Conformity tendency is different on actions with different topics.

Binary action $y_i$ vs a set of actions $W=\{w\}$

$$
f(y_i) = (1 - \lambda_i) d(y_i, \hat{y}_i) + \lambda_i \sum_{j \in N(i)} d(y_i, y_j).
$$

Conformity tendency of role $r$

$$
\gamma_{i,r}^w = \left(1 - \lambda_r\right) \sum_{z=1}^{K} \theta_i^z \phi_z^w + \lambda_r \frac{1}{|N_i|} \sum_{j \in N_i} \sum_{z=1}^{K} \theta_j^z \phi_z^w.
$$

Topic tendency of user $v_j$ on topic $z$

A score of taking action $w$ under topic $z$
Model Details

Basic Idea:
Users’ role distribution is determined by not only attributes but also actions.

Probabilistic explanation

- Neighbors
- Role
- Individual attribute, e.g., degree, clustering coefficient, etc.
- Generation of all the actions
- Action
- Conformity tendency over role
- Generation of individual attributes
Parameter Estimation

- The objective is to estimate $\lambda_r$, i.e., the conformity over role.
- The method is to maximize the likelihood of generating both the individual attributes and the actions.

$$L_1 = \prod_{i=1}^{A} \prod_{t=1}^{T} \prod_{h=1}^{H} \prod_{r=1}^{R} \sum \frac{\rho_{i,t}^{r}}{\sqrt{2\pi\sigma_{r,h}^2}} \exp\left[-\frac{(x_{i,t,h} - \mu_{r,h})^2}{2\sigma_{r,h}^2}\right]$$

$$L_2 = \prod_{d,w} \sum_{i \in A_d} \sum_{r=1}^{R} \frac{\rho_{i,t}^{r} \gamma_{r,i}^{w}}{|A_d|}$$

- We iteratively optimize $L1$ and $L2$ by using EM algorithms and solve the parameters.
Evaluate Conformity through “Wording” Behavior Prediction

- **PLSA** only consider the intrinsic preference, and ignores the situation where a user’s topic distribution may change and become closer to her neighbors’ topic distribution over time.

- **CTM** (Citation influence model) directly learns the conformity tendency of each user, which becomes very difficult to be estimated accurately when very few historical actions of the user and/or her neighbors are available.

- Our model **RCM** clusters similar users into roles, and then learn the conformity tendency of each role.

<table>
<thead>
<tr>
<th>Query</th>
<th>Method</th>
<th>P@5</th>
<th>P@10</th>
<th>MAP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB&amp;DM</td>
<td>PLSA</td>
<td>20.10</td>
<td>15.49</td>
<td>09.26</td>
<td>77.61</td>
</tr>
<tr>
<td></td>
<td>CIM</td>
<td>22.26</td>
<td>17.98</td>
<td>11.59</td>
<td>85.50</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>30.40</td>
<td>24.94</td>
<td>14.16</td>
<td>86.90</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>14.25</td>
<td>11.65</td>
<td>5.71</td>
<td>67.37</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>18.67</td>
<td>15.34</td>
<td>8.12</td>
<td>73.39</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>19.16</td>
<td>15.40</td>
<td>8.92</td>
<td>75.32</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>17.33</td>
<td>14.39</td>
<td>8.47</td>
<td>79.96</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>19.62</td>
<td>16.25</td>
<td>10.83</td>
<td>88.67</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>20.57</td>
<td>17.12</td>
<td>11.37</td>
<td>89.21</td>
</tr>
<tr>
<td></td>
<td>PLSA</td>
<td>4.20</td>
<td>2.60</td>
<td>2.60</td>
<td>81.15</td>
</tr>
<tr>
<td></td>
<td>CIM</td>
<td>21.43</td>
<td>16.42</td>
<td>12.16</td>
<td>85.55</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>25.31</td>
<td>19.98</td>
<td>12.54</td>
<td>85.27</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>17.52</td>
<td>13.37</td>
<td>9.88</td>
<td>81.09</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>19.36</td>
<td>14.50</td>
<td>11.04</td>
<td>85.31</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>20.13</td>
<td>15.20</td>
<td>11.46</td>
<td>88.93</td>
</tr>
<tr>
<td></td>
<td>PLSA</td>
<td>19.92</td>
<td>15.50</td>
<td>9.40</td>
<td>84.10</td>
</tr>
<tr>
<td></td>
<td>CIM</td>
<td>21.24</td>
<td>16.41</td>
<td>10.85</td>
<td>90.70</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>23.60</td>
<td>18.02</td>
<td>11.41</td>
<td>90.92</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>26.68</td>
<td>20.33</td>
<td>12.99</td>
<td>80.63</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>29.36</td>
<td>21.62</td>
<td>14.75</td>
<td>86.92</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>31.20</td>
<td>23.35</td>
<td>15.22</td>
<td>88.41</td>
</tr>
<tr>
<td></td>
<td>PLSA</td>
<td>19.88</td>
<td>14.78</td>
<td>09.64</td>
<td>78.85</td>
</tr>
<tr>
<td></td>
<td>CTM</td>
<td>22.09</td>
<td>16.12</td>
<td>11.10</td>
<td>85.02</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>24.99</td>
<td>18.37</td>
<td>11.50</td>
<td>85.63</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>24.99</td>
<td>18.37</td>
<td>11.50</td>
<td>85.63</td>
</tr>
<tr>
<td></td>
<td>PLSA</td>
<td>17.49</td>
<td>13.51</td>
<td>8.49</td>
<td>78.85</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>21.75</td>
<td>16.83</td>
<td>11.31</td>
<td>85.13</td>
</tr>
<tr>
<td></td>
<td>RCM</td>
<td>24.36</td>
<td>19.05</td>
<td>12.07</td>
<td>85.95</td>
</tr>
</tbody>
</table>
The Correlation between Role and Conformity

People with higher degree and lower clustering coefficient are more likely to conform to others.

When a person collaborates with more authors and the coauthors are more structurally diverse, she may become more open-minded and tend to accept new ideas from others.

When the social circle of the user is restricted to a few coauthors forming a dense collaboration network, the person will be more conservative and tend not to accept other ideas.
Pairwise Influence between Links

Will the formation of AC influences AE, BC, and DC to be formed?

- **Active link**
- **Link to be influenced**
Two Categories of Link Influence

(a) Follower diffusion
(b) Followee diffusion

-->: pre-existing relationships
-->: a new link added at time $t'$
-->: a possible link added at time $t$

$$0 \leq t - t' \leq \delta$$
Randomization Test

- Randomization test is a model-free, computationally intensive statistical technique for hypothesis testing, the main steps are
  1. Compute some test statistic using the set of original observations;
  2. Carry out the random shuffle according to the null hypothesis a large number of times, and compute the test statistic for each random data;
  3. By the law of large numbers, the permutation p-value is approximated by the proportion of randomly generated values that more or less than the observed value of the test statistic.
- **Null hypothesis**: the formation of neighboring links is temporally independent of one another.
- **Test statistic**: 
  \[
  \text{rate} = \frac{|\{\text{triad}(A,B,C) \mid 0 \leq t_{BC} - t_{AC} \leq \delta\}|}{|\{\text{triad}(A,B,C)\}|}
  \]
The link $e_{AC}$ is formed most probably due to the “following” behavior from ordinary user to celebrity user.

The most probable reason of B “following” C is C “following” B before and B “following” back, rather than the influence from A “following” C.

The most probable reason why A follows C is “following” back, and thus C is more likely to be an ordinary user.

There are more two-way links in a triadic closure, which can strengthen the diffusion effect from $e_{AC}$. 

P-values on 24 Triads
Diffusion Decay

- The increasing rate becomes slower over time.
- When $\delta$ is larger than 7 days, the rate almost stops increasing.
- The formation of B following C in followee diffusion is easier than that in follower diffusion.
Follower Diffusion: Power of Reciprocity

Observation: Reciprocal relationships are much more likely to be actual “social” relationships, rather than “celebrity following”, and thus have stronger social influence.
Followee Diffusion: Easy Discovery

Observation: When a user B follows another user A, who already follows user C, B is likely to discover C through browsing A’s retweets of C’s messages or directly checking A’s followee list, and A’s interest in C may indicates that B would also be interested in C.
“Following” Link Cascade Model

- When a link $e'$ is added at time $t'$, at each time slot from time $t'$ to $t' + \delta$:
  - The follower end point $B$ of link $e$ may discover the link $e'$ with discovery probability $g_{e'e}$.
  - Once discovered, $e'$ may trigger $e$ to be formed with influence probability $h_{e'e}$.
  - If failed, $e'$ will have no chance to activate $e$ again.
  - When multiple links activate $e$, $e$ is activated at the time of the first successful attempt.
- The time delay $\lambda$ for discovery follows a geometric distribution with parameter $g_{e'e}$ and after discovery there is one chance at time $t' + \lambda$ that $e'$ could activate $e$. 
Influence Estimation

• The objective is to estimate $h_{e'e}$ and $g_{e'e}$.
• The method is to maximize the likelihood of generating all the links and solve the parameters in the likelihood function.

\[
\mathcal{L} = \prod_{e \in \mathcal{E}} \left\{ p(e | S_e) \prod_{e' \in \mathcal{R}_e} y_{ee'} \right\}.
\]

1. We formalize the formation of each newly added link.
2. For each newly added link, we also formalize its effect on its unformed neighboring links.
Log-likelihood

- A link $e$ is successfully added if at least one of its recently added neighboring links $e' \in S_e$ successfully activated it.
- Use a latent binary vector $\alpha_{Se} = \{\alpha_e\}_{e' \in S_e}$ to represent the statuses of $S_e$.
  - $\alpha_{e'} = 1$: $e'$ tried to activate $e$ and succeeded.
  - $\alpha_{e'} = 0$: $e'$ failed to activate $e$ within $[t_{e'}, t_e]$.

Assume $e'$ activates $e$ independently

$$p(e|S_e) = \sum_{\tilde{\alpha}_{Se}} p(e|\tilde{\alpha}_{Se})p(\tilde{\alpha}_{Se})$$

Assume $p(\alpha_{Se})$ is uniformly distributed.

$$p(e|\tilde{\alpha}_{Se}) = \prod_{e' \in S_e} x_{e'e}^{\alpha_{e'}} y_{e'e}^{1-\alpha_{e'}}$$

The probability of $e'$ not activating $e$ within $[t_{e'}, t_e]$

$$y_{e'e} = 1 - h_\Delta g_\Delta \sum_{t=t_{e'}}^{t_e} (1 - g_\Delta)^{t-t_{e'}}$$

$$= h_\Delta (1 - g_\Delta)^{t_e-t_{e'}} + 1 + (1 - h_\Delta)$$

The final log-likelihood:

$$\log L = \sum_{e \in E} \left\{ \log \sum_{\tilde{\alpha}_{Se}} \prod_{e' \in S_e} x_{e'e}^{\alpha_{e'}} y_{e'e}^{1-\alpha_{e'}} + \sum_{e' \in R_e} \log y_{ee'} \right\}$$
EM Algorithm

• Estimate the influence probabilities associated to 18 triads instead of link pairs.
  – Associate each link pair \((e,e')\) to a triad structure.
  – Aggregate different pairs with the same structure together.

\[ \theta = \{ h_{e'e}, g_{e'e} \} \rightarrow \theta = \{ h_\Delta, g_\Delta \} \]

• Introduce a posterior distribution \(q(e|\alpha_{Se})\) of \(p(e|\alpha_{Se})\), and get a lower bound of the original log-likelihood function.

• Differentiate the lower bound with respect to each parameter and set the partial differential to zero.
Ranking-based Link Prediction

<table>
<thead>
<tr>
<th>Model</th>
<th>P@1</th>
<th>P@2</th>
<th>P@5</th>
<th>P@10</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>47.69</td>
<td>44.24</td>
<td>35.78</td>
<td>30.26</td>
<td>61.55</td>
</tr>
<tr>
<td>SimRank</td>
<td>27.44</td>
<td>30.11</td>
<td>28.90</td>
<td>27.53</td>
<td>46.11</td>
</tr>
<tr>
<td>Katz</td>
<td>50.46</td>
<td>45.38</td>
<td>36.22</td>
<td>30.16</td>
<td>62.54</td>
</tr>
<tr>
<td>RR</td>
<td>54.57</td>
<td>46.87</td>
<td>36.11</td>
<td>29.99</td>
<td>64.53</td>
</tr>
<tr>
<td>PAC</td>
<td>47.69</td>
<td>40.85</td>
<td>33.36</td>
<td>28.99</td>
<td>59.68</td>
</tr>
<tr>
<td>FCM</td>
<td>75.54</td>
<td>60.43</td>
<td>40.37</td>
<td>31.17</td>
<td>79.66</td>
</tr>
</tbody>
</table>

- **CF, SimRank, and Katz**
  - They only consider the static structure information and ignore the dynamic evolution of the network structure.

- **RR and PAC**
  - They fit the distributions of some macroscopic properties such as clustering coefficient and closure ratio.
  - They also do not consider the temporal dependence between two links.
Classification-based Link Prediction

- SVM and LRC perform poorer than FCM on the triads presenting relatively weak diffusion effects, especially on triads 1, 2, 3, and 6.
- The performance of SVM and LRC may be dominated by the effects from the statistically significant triads.
- FCM smooths the effects from different factors using a generative process.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>74.09</td>
<td>54.66</td>
<td>62.90</td>
<td>77.00</td>
</tr>
<tr>
<td>SVM</td>
<td>73.54</td>
<td>56.18</td>
<td>63.69</td>
<td>75.28</td>
</tr>
<tr>
<td>LRC</td>
<td>63.37</td>
<td>63.51</td>
<td>63.43</td>
<td>88.67</td>
</tr>
<tr>
<td>FCM</td>
<td>70.58</td>
<td>60.04</td>
<td>64.88</td>
<td>91.95</td>
</tr>
</tbody>
</table>

Fig. 8. Performance analysis in different triadic structures on Twitter. X-axis: triadic structure index. Y-axis: F1-measure
Learned Model Parameters

Fig. 9. Learned model parameters on Twitter. X-axis: triadic structure index. Y-axis: Discovery/Diffusion probability.

• The discoveries in followee diffusion are easier than those in follower diffusion.
• The diffusion effects in followee diffusion are stronger than those in follower diffusion.
Application: Follower Maximization

Find a set $S$ of $k$ initial followers to follow user $v$ such that the number of subsequent new followers to follow $v$ is maximized.
Application: Friend Recommendation

Find a set $S$ of $k$ initial followees for user $v$ such that the total number of subsequent new followees accepted by $v$ is maximized.
Structural Influence

Whose ego network has more influence, $v_1$ or $v_2$?

- Active neighbor
- Inactive neighbor
- User to be influenced
Test Influence Locality

Goal: evaluate the correlation between active probability and the active neighbors.

Randomized experiment

<table>
<thead>
<tr>
<th>Treatment group</th>
<th>Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users have &gt; 1 active neighbors</td>
<td>Users have =1 active neighbor</td>
</tr>
</tbody>
</table>

Selection bias: users assigned in the treatment group are more likely to retweet than those in the control group even though they do not have >1 active neighbors, because of homophily.

Matched sampling: Match the users in treatment group to those in control group with similar probability to be treated.

\[ p_{it} = P(T_{it} = 1 | X_{it}) \]

A binary variable indicating whether user \( i \) will be treated at time \( t \)

All attributes associated with user \( i \) at time \( t \)
The fraction of active users with 2 active neighbors is about **2 times** the fraction of active users with only 1 active neighbor.

The ratio increases with the number of active neighbors.

After 48 hours when the original tweet has been published, the increasing rate slows down.

\[ N_{T=1} / N_{T=0} > 1 \] indicates the influence locality exerts positive effect on users’ retweet behaviors.

**NT=1**: the average number of active users in the treatment group.

**NT=0**: the average number of active users in the control group.
The probability of a user retweeting a microblog is \textbf{negatively} correlated with the structure diversity of the active neighbors.
Evaluate through Retweet Prediction

Ego network Influence Q = #active neighbors + #circles formed by active neighbors

With only ego network influence factor, we can obtain a F1-score of 71.65%.

Table 2: Performance of retweet behavior prediction. (%)

<table>
<thead>
<tr>
<th>Model</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRC-B</td>
<td>68.11</td>
<td>74.26</td>
<td>71.05</td>
<td>69.74</td>
</tr>
<tr>
<td>LRC-Q</td>
<td>66.82</td>
<td>77.22</td>
<td>71.65</td>
<td>69.44</td>
</tr>
<tr>
<td>LRC-BQ</td>
<td>69.89</td>
<td>77.06</td>
<td>73.30</td>
<td>71.93</td>
</tr>
</tbody>
</table>

LRC-B: logistic regression classifier with only basic features
LRC-Q: logistic regression classifier with only the feature of ego network influence.
LRC-BQ: Combine basic features and influence locality function together.

Basic features: Gender, verification status, #followers, #two-way following relationships, #one-way following relationships, #historical microblogs, topic propensity, the elapsed time
## Structural Influence

<table>
<thead>
<tr>
<th>$k$</th>
<th>$C_k$</th>
<th>$IP_k$</th>
<th>$\widetilde{IP}_k$</th>
<th>$U_{IP_k}$</th>
<th>$k$</th>
<th>$C_k$</th>
<th>$IP_k$</th>
<th>$\widetilde{IP}_k$</th>
<th>$U_{IP_k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.066</td>
<td>0.066</td>
<td>0.020</td>
<td>11</td>
<td></td>
<td>0.038</td>
<td>0.038</td>
<td>0.720</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.074</td>
<td>0.074</td>
<td>0.085</td>
<td>12</td>
<td></td>
<td>0.186</td>
<td>0.186</td>
<td>0.088</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.111</td>
<td>0.110</td>
<td>0.425</td>
<td>13</td>
<td></td>
<td>0.399</td>
<td>0.392</td>
<td>1.785</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.307</td>
<td>0.304</td>
<td>0.928</td>
<td>14</td>
<td></td>
<td>0.063</td>
<td>0.062</td>
<td>0.616</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.069</td>
<td>0.069</td>
<td>0.530</td>
<td>15</td>
<td></td>
<td>0.619</td>
<td><strong>0.615</strong></td>
<td>0.548</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.091</td>
<td>0.090</td>
<td>0.358</td>
<td>16</td>
<td></td>
<td>0.444</td>
<td><strong>0.439</strong></td>
<td>1.378</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.067</td>
<td>0.067</td>
<td>0.236</td>
<td>17</td>
<td></td>
<td>0.070</td>
<td>0.070</td>
<td>0.074</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.106</td>
<td>0.099</td>
<td>5.852</td>
<td>18</td>
<td></td>
<td>0.420</td>
<td><strong>0.416</strong></td>
<td>0.890</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.381</td>
<td>0.388</td>
<td>1.666</td>
<td>19</td>
<td></td>
<td>0.662</td>
<td><strong>0.645</strong></td>
<td>2.696</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.165</td>
<td>0.162</td>
<td>1.128</td>
<td>20</td>
<td></td>
<td>0.485</td>
<td><strong>0.479</strong></td>
<td>1.239</td>
</tr>
</tbody>
</table>
Problem of Structural Influence Measurement

\[ \text{Influence Probability} = \frac{x_k}{x_k + y_k} \]

<table>
<thead>
<tr>
<th>Influence Probability</th>
<th>( l_t \in L )</th>
<th>( l_t \notin L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP(( C_k )) = \frac{C_k}{x_k} \frac{z_k}{y_k} \frac{w_k}{C_k}</td>
<td>( C_k )</td>
<td>( x_k )</td>
</tr>
</tbody>
</table>

Network + Action logs → Action diffusion graphs

Active target action

Inactive target action

v0, a1, t2 → v1, a1, t1
v2, a2, t4 → v3, a2, t3
v3, a1, t2
v2, a1, t3
v3, a2, t3
v2, a2, t4
v1, a1, t1
v3, a1, t2
v3, a2, t3
v2, a2, t4
v5, a2, t5
Approach: StructInf-Basic

- Identify active and inactive target actions
  - Count active actions when an action newly arrives
  - Count inactive actions when an action is outdated
Approach: StructInf-Basic

- Enumerate influence patterns
  - Extend nodes instead of edges
  - Dynamic labeling to avoid duplication
Fast Sampling: StructInf-S1

- Randomly sample nodes when enumerating influence patterns using **sampling probability** $p$. 

![Diagram](image)
Fast Sampling: StructInf-S2

- Randomly sample edges when constructing action diffusion graphs using **sampling probability** $q$. 

```latex
Network

\begin{align*}
  v_0, a_1, t_1 \\
  v_0, a_1, t_2 \\
  v_3, a_1, t_2 \\
  v_3, a_1, t_3 \\
  v_2, a_1, t_3 \\
  v_3, a_2, t_3 \\
  v_2, a_2, t_4 \\
  v_5, a_2, t_5 \\
  \vdots
\end{align*}
```

Action logs

Action diffusion graphs
Fast Sampling: StructInf-S3

• Combine StructInf-S1 and StructInf-S2
• Randomly sample edges when constructing action diffusion graphs using *sampling probability* $q$ and sample nodes when enumerating influence patterns using *sampling probability* $p$ together.
UnBiasness Property

• StructInf-S1

\[ \tilde{x}_k = \hat{x}_k / p^{n_k} \]

• StructInf-S2
  – Complete subgraph
  \[ \hat{x}_k / q^{m_k} \]
  – Incomplete subgraph

• StructInf-S3
  – Complete subgraph
  \[ \hat{x}_k / (p^{n_k} q^{m_k}) \]
  – Incomplete subgraph

Complete subgraph

Incomplete subgraph
Results

(a) $C_1$

(b) $C_2$

(c) $C_3$

(d) $C_4$

(e) $C_{15}$

(f) $C_{20}$
Sampling Variance and Time

• Variance:

\[
\hat{V}(\hat{x}_k) = \sum_{i=1}^{\hat{x}_k} \frac{1 - p(c^k_i)}{p^2(c^k_i)} + \sum_{i \neq j} \frac{p(c^k_i c^k_j) - p(c^k_i)p(c^k_j)}{p(c^k_i)c^k_j)p(c^k_i)p(c^k_j)}
\]

• The higher the sampling probability \( p(c^k_i) \), the smaller the variance will be, while the sampling speed will be slower.

• Trade off error and time by \( p \) and \( q \)
Results

Varying the probabilities $p$ and $q$.

- StructInf-S1 performs better than StructInf-S2
- StructInf-S3 is less sensitive to the parameters

**Weibo dataset**
- 1,787,443 nodes
- 413,503,687 edges
- 20,134,307 actions
Application: Retweet Prediction

Basic: #friends, gender, status, etc.

$C_1$: the number of active neighbors

Weak: $\tilde{IP}_k < 0.1$

Moderate: $0.1 \leq \tilde{IP}_k < 0.3$

Strong: $\tilde{IP}_k > 0.3$

<table>
<thead>
<tr>
<th></th>
<th>F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>71.46</td>
</tr>
<tr>
<td>+$C_1$</td>
<td>71.74</td>
</tr>
<tr>
<td>+Weak</td>
<td>72.84</td>
</tr>
<tr>
<td>+Moderate</td>
<td>73.19</td>
</tr>
<tr>
<td>+Strong</td>
<td>74.92</td>
</tr>
</tbody>
</table>
Summary

• **Node conformity influence**
  _ People with higher degree and lower clustering coefficient are more likely to conform to others.

• **Pairwise link influence**
  _ A two-way relationship between two users can trigger more links than a one-way relationship.

• **Group influence**
  – **Structural diversity**
    • The probability of a user retweeting a tweet is negatively correlated with structural diversity of the active neighbors.
  – **Structural influence**
    • Sampling algorithms can achieve a 10 speedup compared to the exact influence pattern mining algorithm
Future Work

• How to design a diffusion model that considers different kinds of influence together?
• What’s the difference between influence in different kinds of social medias?
• How to leverage different kinds of influence to do social recommendation?
Code & Dataset

• Conformity Influence on “wording” behavior
  – http://arnetminer.org/roleconformity

• Link Influence
  – http://cs.aminer.org/followinf
  – Jing Zhang, Zhanpeng Fang, Wei Chen, and Jie Tang. Diffusion of “Following” Links in Microblogging Networks. IEEE Transaction on Knowledge and Data Engineering (TKDE)

• Structural influence
  – http://arnetminer.org/influencelocality
  – https://cn.aminer.org/structinf
  – Jing Zhang, Jie Tang, Yuanyi Zhong, Yuchen Mo, Jimeng Sun, and Juanzi Li. StructInf: Mining Structural Influence from Social Streams. In AAAI’17.
Thank You

https://scholar.google.com/citations?user=T7Wa3GQAAAAJ&hl=en