Link Prediction and Recommendation across Heterogeneous Social Networks

Yuxiao Dong*, Jie Tang$, Sen Wu$, Jilei Tian#
Nitesh V. Chawla*, Jinghai Rao#, Huanhuan Cao#

*University of Notre Dame
$Tsinghua University
#Nokia Research China
Introduction

- Link Recommendation is ubiquitous ...
Transfer Link Prediction (Recommendation)

1. General Factors

2. Transfer Model

3. Source network → Target network
Link Recommendation

$G = (V, E)$: social network

$\nu_s$: a particular user

$C$: candidates for $\nu_s$

$Y$: candidates’ rank

Input: $G, \nu_s, C$

Output: $f$: $(G, \nu_s, C) \rightarrow Y$
Basic Idea
Data Sets and Methodologies
Data Sets

- **Networks**

<table>
<thead>
<tr>
<th></th>
<th>#nodes</th>
<th>#edges</th>
<th>+edges</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epinions</td>
<td>131,828</td>
<td>841,372</td>
<td>85%</td>
<td>Who-trust-whom online social website</td>
</tr>
<tr>
<td>Slashdot</td>
<td>82,144</td>
<td>549,202</td>
<td>78%</td>
<td>User community based technology news website</td>
</tr>
<tr>
<td>Twitter</td>
<td>63,803</td>
<td>153,098</td>
<td>38%</td>
<td>Who-follow-whom micro-blogging networks</td>
</tr>
</tbody>
</table>

- **Candidate Generation**
  - Randomly select 2000 nodes as the source users\[^2\] from the network.
  - For each source user, we generate the candidate list for her/him.

1. Epinions, Slashdot, Wikivote are available at [http://snap.stanford.edu](http://snap.stanford.edu) and Twitter available at [http://arnetminer.org/reciprocal](http://arnetminer.org/reciprocal)
2. Lars Backstrom, Jure Leskovec. Supervised Random Walks: Predicting and Recommending Links in social Networks. In *WSDM’11*
Baseline Predictors

- **Unsupervised methods**
  - Common neighbors \( |\psi(v_i) \cap \psi(v_j)| \)
  - Adamic/Adar \( \sum_{v_k \in \psi(v_i) \cap \psi(v_j)} \frac{1}{\log d(v_k)} \)
  - Jaccard Index \( \frac{|\psi(v_i) \cap \psi(v_j)|}{|\psi(v_i) \cup \psi(v_j)|} \)
  - Preferential Attachment \( d(v_i) \times d(v_j) \)

- **Supervised methods**
  - SVMRank (SVM-light)
  - Logistic Regression (Weka)
  - **Ranking Factor Graph Model**
Ranking Factor Graph Model

Joint distribution:

\[ p(Y|G) = \prod f(v_s, v_i, y_{si}) \cdot g(X_c, Y_c) \]

Attribute factors

Social factors
Ranking Factor Graph Model (Cont.)

- Joint distribution:
  \[ p(Y|G) = \prod f(v_s, v_i, y_{si}) g(X_c, Y_c) \]

- Exponential-linear functions to initialize factors
  - Attribute factor:
    \[ f(v_s, v_i, y_{si}) = \frac{1}{Z_\alpha} \exp\{\sum_{j=1}^{d} \alpha_j f_j(x_{si_j}, y_{si})\} \]
  - Social factor:
    \[ g_k(X_c, Y_c) = \frac{1}{Z_\beta} \exp\{\sum_c \sum_k \beta_k g_k(X_c, Y_c)\} \]
Ranking Factor Graph Model (Cont.)

- RFG objective function:

\[
\mathcal{O}(\theta) = \sum_{s=1}^{V} \sum_{i=1}^{C} \sum_{j=1}^{d} \alpha_j f_j(x_{si_j}, y_{si}) + \sum_{c} \sum_{k} \beta_k g_k(x_c, y_c) - \log Z
\]

- Learning\(^1\): Parameters: \(\theta = (\{\alpha\}, \{\beta\})\)

---

<table>
<thead>
<tr>
<th>Feature</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>in-degree</td>
<td>$d_{in}(v_i), d_{in}(v_j)$</td>
</tr>
<tr>
<td>out-degree</td>
<td>$d_{out}(v_i), d_{out}(v_j)$</td>
</tr>
<tr>
<td>all-degree</td>
<td>$d_{all}(v_i), d_{all}(v_j)$</td>
</tr>
<tr>
<td>common neighbors</td>
<td>$</td>
</tr>
<tr>
<td>Adamic/Adar Index</td>
<td>$\sum_{v_k \in \psi(v_i) \cap \psi(v_j)} \frac{1}{\log d(v_k)}$</td>
</tr>
<tr>
<td>Jaccard Index</td>
<td>$\frac{</td>
</tr>
<tr>
<td>Preferential Index</td>
<td>$d(v_i) \times d(v_j)$</td>
</tr>
</tbody>
</table>
Still Problems?

- **Unbalanced Data**: the number of potential candidates grows exponentially \((d(v_s)^{n-1})\) as the number of hops \(n\) increases.

- **Few Training Data**: obtaining sufficient training data is difficult.
Transfer Link Recommendation
Transfer Link Recommendation Framework

1. General Factors
2. Transfer Model
3. Candidate Generation
Social Factors for Transfer Learning

- What are the **General social factors** to form links?
  - Homophily
  - Social Balance
  - Microscopic Mechanism
General social factors

- Homophily / Social Balance / Microscopic

The principle of homophily suggests that users with similar characteristics tend to associate with each other.

The likelihood of two users creating a link increases when the number of their common neighbors increases in the four networks.

This effect of homophily is more pronounced when the number reaches 100, where the probabilities are all higher than 50% in the four networks.
Social balance theory is based on the principles that “the friend of my friend is my friend” and “the enemy of my enemy is my friend”.

It is more likely (more than 80% likelihood) for users to establish balanced triangle of friendships in all four online networks.
Microsociology is one of the branches of sociology, concerning the nature of human social interactions and agency on a small scale.

Why?

student

student

Lady Gaga

Barack Obama

twitter
General social factors

- Homophily / Social Balance / Microscopic

The enumeration is conditioned on whether X, Y, Z are opinion leaders (green means it is an opinion leader).

Four networks share a very similar distribution on probabilities of close triad formation in all six cases, though the four networks are totally different.
Transfer based factor graph model was first proposed as a classification model in [1].

Transfer RFG (Cont.)

- **RFG**
  - Objective function:
    \[ O(\theta) = \sum_{s=1}^{V} \sum_{i=1}^{C} \sum_{j=1}^{d} \alpha_j f_j(x_{si}, y_{si}) + \sum_{c} \sum_{k} \beta_k g_k(X_c, Y_c) - \log Z \]

- **TRFG**
  - Objective function:
    \[ O(\alpha, \beta, \mu) = O_S(\alpha, \beta) + O_T(\mu, \beta) \]
    \[ = \sum_{s=1}^{|V_S|} \sum_{i=1}^{|C_S|} \sum_{j=1}^{d} \alpha_j f_j(x_{si}, y_{si}) + \sum_{s=1}^{|V_T|} \sum_{i=1}^{|C_T|} \sum_{j=1}^{d'} \mu_j f_j'(x_{si}, y_{si}) + \sum_{k} \beta_k \left( \sum_{c \in G_S} g_k(X_c^S, Y_c^S) + \sum_{c \in G_T} g_k(X_c^T, Y_c^T) \right) - \log Z \]

- General social factors across source and target networks
- Attributes factor in source network
- Attributes factor in target network
- Bridge source & target networks
Learning Algorithm

Input: a source network $G_S$, a target network $G_T$, and the learning rate $\eta$

Output: estimated parameters $\theta = (\{\alpha\}, \{\beta\}, \{\mu\})$

Initialize $\theta \leftarrow 0$;
Perform statistics according to social theories;
Construct social theories based features $h_k(Y_c)$;
repeat
   Step 1: Perform LBP to calculate marginal distribution of unknown variables in the source network $P(y_i|x_i, G_S)$;
   Step 2: Perform LBP to calculate marginal distribution of unknown variables in the target network $P(y_i|x_i, G_T)$;
   Step 3: Perform LBP to calculate the marginal distribution of clique $c$, i.e., $P(y_c|X_c^S, X_c^T, G_S, G_T)$;
   Step 4: Calculate the gradient of $\mu_k$ according to Eq. 8 (for $\alpha_j$ and $\beta_j$ with a similar formula);
   Step 5: Update parameter $\theta$ with the learning rate $\eta$:

   $$\theta_{\text{new}} = \theta_{\text{old}} + \eta \cdot \frac{\partial \ell(\theta)}{\partial \theta}$$

until Convergence;

Gradient Decent method

### Results

**Prediction without transfer, evaluated by AUC**

<table>
<thead>
<tr>
<th>Method</th>
<th>Epinions</th>
<th>Slashdot</th>
<th>Wikivote</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN</td>
<td>0.8728</td>
<td>0.5048</td>
<td>0.7842</td>
<td>0.5920</td>
</tr>
<tr>
<td>AA</td>
<td>0.8736</td>
<td>0.5362</td>
<td>0.7924</td>
<td>0.6198</td>
</tr>
<tr>
<td>JA</td>
<td>0.6850</td>
<td>0.3277</td>
<td>0.7241</td>
<td>0.5718</td>
</tr>
<tr>
<td>PA</td>
<td>0.8300</td>
<td>0.7108</td>
<td>0.7433</td>
<td>0.5725</td>
</tr>
<tr>
<td>SVMRank</td>
<td>0.8943</td>
<td>0.7880</td>
<td>0.7907</td>
<td>0.6834</td>
</tr>
<tr>
<td>LRC</td>
<td>0.9405</td>
<td>0.9200</td>
<td>0.8905</td>
<td>0.8044</td>
</tr>
<tr>
<td>RFG</td>
<td><strong>0.9821</strong></td>
<td><strong>0.9866</strong></td>
<td><strong>0.9298</strong></td>
<td><strong>0.8905</strong></td>
</tr>
</tbody>
</table>
Transfer Recommendation Results

- 50% labeled data and 50% unlabeled data in target network;
- #source data is 50% of #target data.

<table>
<thead>
<tr>
<th>Transfer cases</th>
<th>AUC</th>
<th>Pre@30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epinions</td>
<td>0.9821</td>
<td>5.38</td>
</tr>
<tr>
<td>Slashdot (S) to Epinions (T)</td>
<td>0.9873</td>
<td>5.90</td>
</tr>
<tr>
<td>Wikivote (S) to Epinions (T)</td>
<td>0.9833</td>
<td>5.38</td>
</tr>
<tr>
<td>Twitter (S) to Epinions (T)</td>
<td>0.9825</td>
<td>5.39</td>
</tr>
<tr>
<td>Slashdot</td>
<td>0.9866</td>
<td>11.96</td>
</tr>
<tr>
<td>Epinions (S) to Slashdot (T)</td>
<td>0.9882</td>
<td>11.99</td>
</tr>
<tr>
<td>Wikivote (S) to Slashdot (T)</td>
<td>0.9833</td>
<td>11.21</td>
</tr>
<tr>
<td>Twitter (S) to Slashdot (T)</td>
<td>0.9762</td>
<td>10.38</td>
</tr>
<tr>
<td>Wikivote</td>
<td>0.9298</td>
<td>12.02</td>
</tr>
<tr>
<td>Epinions (S) to Wikivote (T)</td>
<td>0.9343</td>
<td>12.36</td>
</tr>
<tr>
<td>Slashdot (S) to Wikivote (T)</td>
<td>0.9313</td>
<td>12.41</td>
</tr>
<tr>
<td>Twitter (S) to Wikivote (T)</td>
<td>0.9048</td>
<td>10.74</td>
</tr>
<tr>
<td>Twitter</td>
<td>0.8905</td>
<td>5.07</td>
</tr>
<tr>
<td>Epinions (S) to Twitter (T)</td>
<td>0.8905</td>
<td>5.08</td>
</tr>
<tr>
<td>Slashdot (S) to Twitter (T)</td>
<td>0.8914</td>
<td>5.08</td>
</tr>
<tr>
<td>Wikivote (S) to Twitter (T)</td>
<td>0.8906</td>
<td>5.07</td>
</tr>
<tr>
<td>Transfer cases</td>
<td>AUC</td>
<td>Pre@30</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>Epinions</td>
<td>0.9821</td>
<td>5.38</td>
</tr>
<tr>
<td>Slashdot (S) to Epinions (T)</td>
<td>0.9873</td>
<td>5.90</td>
</tr>
<tr>
<td>Wikivote (S) to Epinions (T)</td>
<td>0.9833</td>
<td>5.38</td>
</tr>
<tr>
<td>Twitter (S) to Epinions (T)</td>
<td>0.9825</td>
<td>5.39</td>
</tr>
<tr>
<td>Slashdot</td>
<td>0.9866</td>
<td>11.96</td>
</tr>
<tr>
<td>Epinions (S) to Slashdot (T)</td>
<td>0.9882</td>
<td>11.99</td>
</tr>
<tr>
<td>Wikivote (S) to Slashdot (T)</td>
<td>0.9833</td>
<td>11.21</td>
</tr>
<tr>
<td>Twitter (S) to Slashdot (T)</td>
<td>0.9762</td>
<td>10.38</td>
</tr>
<tr>
<td>Wikivote</td>
<td>0.9298</td>
<td>12.02</td>
</tr>
<tr>
<td>Epinions (S) to Wikivote (T)</td>
<td>0.9343</td>
<td>12.36</td>
</tr>
<tr>
<td>Slashdot (S) to Wikivote (T)</td>
<td>0.9313</td>
<td>12.41</td>
</tr>
<tr>
<td>Twitter (S) to Wikivote (T)</td>
<td>0.9048</td>
<td>10.74</td>
</tr>
<tr>
<td>Twitter</td>
<td>0.8905</td>
<td>5.07</td>
</tr>
<tr>
<td>Epinions (S) to Twitter (T)</td>
<td>0.8905</td>
<td>5.08</td>
</tr>
<tr>
<td>Slashdot (S) to Twitter (T)</td>
<td>0.8914</td>
<td>5.08</td>
</tr>
<tr>
<td>Wikivote (S) to Twitter (T)</td>
<td>0.8906</td>
<td>5.07</td>
</tr>
</tbody>
</table>

The graph shows the distribution of positive degrees with probability on the y-axis and positive degree on the x-axis, with different colors representing different transfer cases.
Results (Cont.)

- Transfer efficiency

Performance of link recommendation with transfer by varying the percent of source instances to target instances.
Summary

Transfer Link recommendation across heterogeneous networks

- Social Theories on Link Formation
- Transfer Ranking Factor Graph Model
Thanks

Q & A

http://aminer.org/