Social Tie Analysis
—Computational aspect

Jie Tang
Tsinghua University, China

Collaborate with
Jon Kleinberg and John Hopcroft (Cornell)
Jiawei Han and Chi Wang (UIUC)
Tiancheng Lou, Wenbin Tang, Honglei Zhuang, and Jing Zhang (THU)
Iceberg Model for Social Network
Iceberg Model for Social Network

- Information Diffusion
- Collective Intelligence
- Tie
- Influence
- Traits and Motivates
Inferring Social Ties

KDD 2010, PKDD 2011 (*Best Paper Runnerup*), WSDM 2012, DMKD
Real social networks are complex...

• Nobody exists only in one social network.
  – Public network vs. private network
  – Business network vs. family network

• However, existing networks (e.g., Facebook and Twitter) are trying to lump everyone into one big network
  – FB tries to solve this problem via lists/groups
  – However...

• Google+

which circle? Users do not take time to create it.
Even complex than we imaged!

- Only 16% of mobile phone users in Europe have created custom contact groups
  - users do not take the time to create it
  - users do not know how to circle their friends

- The fact is that our social network is black-white...
Example 1: finding **boss** in email networks
(PKDD’11, Best Paper Runnerup)

**Enterprise email network**

**How to infer**
- CEO
- Manager
- Employee

User interactions may form *implicit groups*
Example 2: finding friends in mobile networks
Challenges

- What are the fundamental forces behind?
- Can we automatically infer the type of social ties?
Networks

- **Epinions**: a network of product reviewers: 131,828 nodes (users) and 841,372 edges
  - trust relationships between users

- **Slashdot**: 82,144 users and 59,202 edges
  - “friend” relationships between users

- **Mobile**: 107 mobile users and 5,436 edges
  - to infer friendships between users

- **Coauthor**: 815,946 authors and 2,792,833 coauthor relationships
  - to infer advisor-advisee relationships between coauthors

- **Enron**: 151 Enron employees and 3572 edges
  - to infer manager-subordinate relationships between users
Problem Formulation

Input: \( G = (V, E^L, E^U, R^L, W) \)

- \( V \): Set of Users
- \( E^L, R^L \): Labeled relationships
- \( E^U \): Unlabeled relationships

Output: \( f: G \rightarrow R \)

Partially Labeled Network
Basic Idea

Friend

User → Node

Relationship → Node
Partially Labeled Pairwise Factor Graph Model (PLP-FGM)

Problem:
For each relationship, identify which type has the highest probability?

Example:
A makes call to B immediately after the call to C.

Solutions (con’t)

• Different ways to instantiate factors
  – We use exponential-linear functions
    • Attribute Factor:
      \[ f(y_i, x_i) = \frac{1}{Z_\lambda} \exp\{\lambda^T \Phi(y_i, x_i)\} \]
    
    • Correlation / Constraint Factor:
      \[ g(y_i, G(y_i)) = \frac{1}{Z_\alpha} \exp\{\sum_{y_j \in G(y_i)} \alpha^T g(y_i, y_j)\} \]
      \[ h(y_i, H(y_i)) = \frac{1}{Z_\beta} \exp\{\sum_{y_j \in H(y_i)} \beta^T h(y_i, y_j)\} \]
      
      \[ \theta = [\lambda, \alpha, \beta], s = [\Phi^T, g^T, h^T]^T \]
      
      – Log-Likelihood of labeled Data:
      \[ \mathcal{O}(\theta) = \log \sum_{Y|Y^L} \exp\{\theta^T S\} - \log \sum_Y \exp\{\theta^T S\} \]
Learning Algorithm

• Maximize the log-likelihood of labeled relationships

```
Input: learning rate $\eta$
Output: learned parameters $\theta$
Initialize $\theta$;
repeat
    Calculate $\mathbb{E}_{p(\theta|Y,L,G)}S$ using LBP;
    Calculate $\mathbb{E}_{p(\theta|G)}S$ using LBP;
    Calculate the gradient of $\theta$ according to Eq. 7:
    \[
    \nabla_\theta = \mathbb{E}_{p(\theta|Y,L,G)}S - \mathbb{E}_{p(\theta|G)}S
    \]
    Update parameter $\theta$ with the learning rate $\eta$:
    \[
    \theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \nabla_\theta
    \]
until Convergence;
```

Algorithm 1: Learning PLP-FGM.

Gradient Ascent Method
Still Challenges?

Questions:
- How to obtain sufficiently training data?
- Can we leverage knowledge from other network?
Distributed Learning

Compute Gradient via LBP

Optimize with Gradient Descent

Graph Partition by Metis
Master-Slave Computing
Inferring Social Ties Across Networks

Input: Heterogeneous Networks

Output: Inferred social ties in different networks

What is the knowledge to transfer?

Social Theories

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

Observations:
(1) The underlying networks are unbalanced;
(2) While the friendship networks are balanced.
Social Theories—Structural hole

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

**Observations:** Users are more likely (+25-150% higher than change) to have the same type of relationship with C if C spans structural holes.
Social Theories—Social status

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

**Observations:** 99% of triads in the networks satisfy the social status theory

**Note:** Given a triad (A,B,C), let us use 1 to denote the advisor-advisee relationship and 0 colleague relationship. Thus the number 011 to denote A and B are colleagues, B is C’s advisor and A is C’s advisor.
Social Theories—Two-step-flow

• Social balance theory
• Structural hole theory
• Social status theory
• Two-step-flow theory

Observations: Opinion leaders are more likely (+71%-84% higher than chance) to have a higher social-status than ordinary users.
Transfer Factor Graph Model

TrFG model

\[ y_1 = 1 \]

\[ v_1 \]

\[ v_2 \]

\[ v_3 \]

\[ v_4 \]

\[ v_5 \]

\[ v_6 \]

Input: social network

Bridge via social theories

Coauthor network

mobile
Mathematical Formulation

\[ O(\alpha, \beta, \mu) = O_S(\alpha, \mu) + O_T(\beta, \mu) \]

\[ = \sum_{i=1}^{V_S} \sum_{j=1}^{d} \alpha_{ij} g_{ij}(x_{ij}^S, y_i^S) + \sum_{i=1}^{V_T} \sum_{j=1}^{d'} \beta_{ij} g'_{ij}(x_{ij}^T, y_i^T) \]

\[ + \sum_{k} \mu_k (\sum_{c \in G_S} h_k(Y_c^S) + \sum_{c \in G_T} h_k(Y_c^T)) \]

\[ - \log Z \]

Features defined in source network

Features defined in target network

Triad-based features shared across networks

Experiments

• Data sets
  – Epinions: 131,828 nodes (users) and 841,372 edges
  – Slashdot: 82,144 users and 59,202 edges
  – Mobile: 107 mobile users and 5,436 edges
  – Coauthor: 815,946 authors and 2,792,833 coauthor relationships
  – Enron: 151 Enron employees and 3572 edges

• Comparison methods
  – SVM and CRF are two baseline methods
  – PFG is the partially-labeled factor graph model
  – TranFG is the transfer–based factor graph model
**Results – undirected networks**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epinions (S) to Slashdot (T)</td>
<td>SVM</td>
<td>0.7157</td>
<td>0.9733</td>
<td>0.8249</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.8919</td>
<td>0.6710</td>
<td>0.7658</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>0.9300</td>
<td>0.6436</td>
<td>0.7607</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.9414</td>
<td>0.9446</td>
<td>0.9430</td>
</tr>
<tr>
<td>Slashdot (S) to Epinions (T)</td>
<td>SVM</td>
<td>0.9132</td>
<td>0.9925</td>
<td>0.9512</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.8923</td>
<td>0.9911</td>
<td>0.9393</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>0.9954</td>
<td>0.9787</td>
<td>0.9870</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.9954</td>
<td>0.9787</td>
<td>0.9870</td>
</tr>
<tr>
<td>Epinions (S) to Mobile (T)</td>
<td>SVM</td>
<td>0.8983</td>
<td>0.5955</td>
<td>0.7162</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.9455</td>
<td>0.5417</td>
<td>0.6887</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>1.0000</td>
<td>0.5924</td>
<td>0.7440</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.8239</td>
<td>0.8344</td>
<td>0.8291</td>
</tr>
<tr>
<td>Slashdot (S) to Mobile (T)</td>
<td>SVM</td>
<td>0.8983</td>
<td>0.5955</td>
<td>0.7162</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.9455</td>
<td>0.5417</td>
<td>0.6887</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>1.0000</td>
<td>0.5924</td>
<td>0.7440</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.7258</td>
<td>0.8599</td>
<td>0.7872</td>
</tr>
</tbody>
</table>

**SVM** and **CRF** are two baseline methods.

**PFG** is the proposed partially-labeled factor graph model.

**TranFG** is the proposed transfer–based factor graph model.
Results – directed networks

SVM and CRF are two baseline methods.
PFG is the proposed partially-labeled factor graph model.
TranFG is the proposed transfer–based factor graph model.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coauthor (S) to</td>
<td>SVM</td>
<td>0.9524</td>
<td>0.5556</td>
<td>0.7018</td>
</tr>
<tr>
<td>Enron (T) (40%)</td>
<td>CRF</td>
<td>0.9565</td>
<td>0.5366</td>
<td>0.6875</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>0.9730</td>
<td>0.6545</td>
<td>0.7826</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.9556</td>
<td>0.7818</td>
<td>0.8600</td>
</tr>
<tr>
<td>Enron (S) to</td>
<td>SVM</td>
<td>0.6910</td>
<td>0.3727</td>
<td>0.4842</td>
</tr>
<tr>
<td>Coauthor (T) (40%)</td>
<td>CRF</td>
<td>1.0000</td>
<td>0.3043</td>
<td>0.4666</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>0.9916</td>
<td>0.4591</td>
<td>0.6277</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.5936</td>
<td>0.7611</td>
<td>0.6669</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.9793</td>
<td>0.5525</td>
<td>0.7065</td>
</tr>
</tbody>
</table>
Factor Contribution Analysis

**Undirected Network**

- **SH** - Structural hole;
- **SB** - Social balance.

**Directed Network**

- **OL** - Opinion leader;
- **SS** - Social status.

![Graphs showing F1-Measure for different networks and factor contributions.](image)
Parasocial vs. Reciprocal
Who will follow you back?

On Twitter…

Ladygaga

Obama

Shiteng

Huwei

JimmyQiao

100%

30%

1%

60%
Homophily

**Link homophily:** users who share common links will have a tendency to follow each other.

**Status homophily:** Elite users have a much stronger tendency to follow each other.
Interaction

Retweet vs. reply

*Retweeting seems to be more helpful
Structural Balance

- Reciprocal relationships are balanced (88%);
- Parasocial relationships are not (only 29%).
Triad Factor Graph (TriFG)

TriFG model

Observations

Input: Mobile Network

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_4 = \text{?}
\]

\[
y_5 = \text{non-friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_4 = \text{?}
\]

\[
y_5 = \text{non-friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]

\[
y_1 = \text{friend}
\]

\[
y_2 = \text{friend}
\]

\[
y_6 = \text{non-friend}
\]

\[
y_3 = \text{?}
\]
Experiments

- Huge sub-network of twitter
  - 13,442,659 users and 56,893,234 following links.
  - Extracted 35,746,366 tweets.

- Dynamic networks
  - With an average of 728,509 new links per day.
  - Averagely 3,337 new follow-back links per day.
  - 13 time stamps by viewing every four days as a time stamp

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1Measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Case 1</strong></td>
<td>SVM</td>
<td>0.6908</td>
<td>0.6129</td>
<td>0.6495</td>
<td>0.9590</td>
</tr>
<tr>
<td></td>
<td>LRC</td>
<td>0.6957</td>
<td>0.2581</td>
<td>0.3765</td>
<td>0.9510</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>1.0000</td>
<td>0.6290</td>
<td>0.7723</td>
<td>0.9770</td>
</tr>
<tr>
<td></td>
<td>TriFG</td>
<td>1.0000</td>
<td>0.8548</td>
<td><strong>0.9217</strong></td>
<td><strong>0.9910</strong></td>
</tr>
<tr>
<td><strong>Test Case 2</strong></td>
<td>SVM</td>
<td>0.7323</td>
<td>0.6212</td>
<td>0.6722</td>
<td>0.9534</td>
</tr>
<tr>
<td></td>
<td>LRC</td>
<td>0.8333</td>
<td>0.3030</td>
<td>0.4444</td>
<td>0.9417</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>1.0000</td>
<td>0.6333</td>
<td>0.7755</td>
<td>0.9717</td>
</tr>
<tr>
<td></td>
<td>TriFG</td>
<td>1.0000</td>
<td>0.8788</td>
<td><strong>0.9355</strong></td>
<td><strong>0.9907</strong></td>
</tr>
</tbody>
</table>
Effect of Time Span

- Distribution of follow back time
  - 60% for next-time stamp;
  - 37% for following 3 time stamps.
- Different settings of the time span
  - Performance drops sharply when two or less;
  - Acceptable for three time stamps.
Case Study

(a) Ground Truth

(b) SVM

(c) Our approach (TriFG)
Triadic Closure
Triadic Closure

Ladygaga → 0.5% → Obama → 1% → Shiteng → 60% → Huwei → 50% → JimmyQiao → 90% → Ladygaga
**Triad Status**

- **P(1XX) > P(0XX)**. Elites users play a more important role to form the triadic closure. The average probability of 1XX is three times higher than that of 0XX.
- **P(X0X) > P(X1X)**. Low-status users act as a bridge to connect users so as to form a closure triad. The likelihood of X0X is 2.8 times higher than X1X.
- **P(XX1) > P(XX0)**. The rich gets richer. This phenomenon validates the mechanism of preferential attachment [Newman 2001].
## Triad Closure Prediction Result

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Test Case 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0870</td>
<td>0.1429</td>
<td>0.1081</td>
</tr>
<tr>
<td></td>
<td>LRC</td>
<td>0.0536</td>
<td>0.1404</td>
<td>0.0759</td>
</tr>
<tr>
<td></td>
<td>CRF-balance</td>
<td>0.0208</td>
<td>0.0436</td>
<td>0.0282</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.1111</td>
<td>0.0870</td>
<td>0.0976</td>
</tr>
<tr>
<td></td>
<td>wTriFG</td>
<td>0.3333</td>
<td>0.0373</td>
<td>0.0671</td>
</tr>
<tr>
<td></td>
<td>TriFG</td>
<td>0.4545</td>
<td>0.2174</td>
<td>0.2941</td>
</tr>
<tr>
<td></td>
<td><strong>Test Case 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.2000</td>
<td>0.2222</td>
<td>0.2105</td>
</tr>
<tr>
<td></td>
<td>LRC</td>
<td>0.1071</td>
<td>0.1667</td>
<td>0.1304</td>
</tr>
<tr>
<td></td>
<td>CRF-balance</td>
<td>0.0909</td>
<td>0.0556</td>
<td>0.0690</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.2222</td>
<td>0.2222</td>
<td>0.2222</td>
</tr>
<tr>
<td></td>
<td>wTriFG</td>
<td>0.5000</td>
<td>0.0556</td>
<td>0.1000</td>
</tr>
<tr>
<td></td>
<td>TriFG</td>
<td>0.8571</td>
<td>0.3333</td>
<td>0.4800</td>
</tr>
</tbody>
</table>
Follow Influence

Lady Gaga → You → Lady Gaga

Obama → Shiteng
Will the “following” be Influenced?

Large neighbors, but may not be influenced

Ladygaga

50%

JimmyQiao

90%

Few neighbors, but may be significantly influenced

Obama

40%

Huwei

5%

Shiteng

60%

2%

1%?

Large neighbors, but may not be influenced

2%
Influence Test

Question:
Whether there exist follow influence?
In which kind of triad the influence is significant?

Method:
Compare the same kind of triad with different timestamp.

Assumption:
If $P_1(B\rightarrow C)$ is much larger than $P_2(B\rightarrow C)$, then influence exists.
Test Result

Two categories of triads have significant influence, compared with two other categories.

- **Attract more followers**
  - P1(B→C) = 0.5%
  - P2(B→C) = 0.1%
  - [Diagram]

- **Follow More**
  - P1(B→C) = 14.4%
  - P2(B→C) = 0.1%
  - [Diagram]

- **No influence**
  - P1(B→C) = 0.02%
  - P2(B→C) = 0.02%
  - [Diagram]

- **No influence**
  - P1(B→C) = 0.02%
  - P2(B→C) = 0.02%
  - [Diagram]
More…

P(B->C) is significantly boosted when the reversed follow link is pre-formed.

Question: Are there any other factors that can boost P(B->C)?
Structural Balance

$P(B\rightarrow C)$ is significantly boosted when the resultant triad satisfies the balance theory.
Application: Follow Influence Maximization

- Influence: Select seeds which can influence most users
- Followback: Select seeds which can follow back with the highest probabilities
- Random: Select seeds randomly
Summary

• Computational models for social tie analysis
  – Inferring social tie
  – Parasocial -> Reciprocal
  – Tradic closure
  – Follow influence

• This is just a start for social tie analysis
  – How social tie influences user behaviors?
  – How social tie influences the network structure?
  – …
Related Publications

- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, (accepted).


- Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogenous Networks. *WSDM’12*.


- Chi Wang, Jiawei Han, Yuntao Jia, Duo Zhang, Yintao Yu, Jie Tang, Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. *KDD’10*. 
Thank you!

QA?

Data & Code:
http://arnetminer.org/socialtieacross
http://arnetminer.org/socialtie
http://arnetminer.org/reciprocity