Who Will Follow You Back?
Reciprocal Relationship Prediction*

1John Hopcroft, 2Tiancheng Lou, 3Jie Tang
1Department of Computer Science, Cornell University,
2Institute for Interdisciplinary Information Sciences, Tsinghua University
3Department of Computer Science, Tsinghua University
Motivation

- Two kinds of relationships in social network,
  - one-way (called parasocial) relationship and,
  - two-way (called reciprocal) relationship
- Two-way (reciprocal) relationship
  - usually developed from a one-way relationship
  - more trustful.
- Try to understand (predict) the formation of two-way relationships
  - micro-level dynamics of the social network.
  - underlying community structure?
  - how users influence each other?
Example: real friend relationship

On Twitter: Who Will Follow You Back?

Ladygaga

Obama

Shiteng

Huwei

JimmyQiao
Several key challenges

- How to **model** the formation of two-way relationships?
  - SVM & CRF
- How to **combine** many social theories into the prediction model?
Outline

- Previous works
- Our approach
- Experimental results
- Conclusion & future works
Link prediction

- Unsupervised link prediction
  - Scores & intuition, such as preferential attachment [N01].

- Supervised link prediction
  - supervised random walks [BL11].
  - logistic regression model to predict positive and negative links [L10].

- Main differences:
  - We predict a directed link instead of only handles undirected social networks.
  - Our model is dynamic and learned from the evolution of the Twitter network.
Social behavior analysis

- Existing works on social behavior analysis:
  - The difference of the social influence on different topics and to model the topic-level social influence in social networks. [T09]
  - How social actions evolve in a dynamic social network? [T10]

- Main differences:
  - The proposed methods in previous work can be used here
  - but the problem is fundamentally different.
Twitter study

- The twitter network.
  - The topological and geographical properties. [J07]
  - Twittersphere and some notable properties, such as a non-power-law follower distribution, and low reciprocity. [K10]

- The twitter users.
  - Influential users.
  - Tweeting behaviors of users.

- The tweets.
  - Utilize the real-time nature to detect a target event. [S10]
  - TwitterMonitor, to detect emerging topics. [M10]
Outline

- Previous works
- Our approach
- Experimental results
- Conclusion & future works
Factor graph model

- Problem definition
  - Given a network at time $t$, i.e., $G^t = (V^t, E^t, X^t, Y^t)$
  - Variables $y$ are partially labeled.
  - Goal: infer unknown variables.

- Factor graph model
  - $P(Y \mid X, G) = P(X, G \mid Y) \frac{P(Y)}{P(X, G)} = C_0 P(X \mid Y) P(Y \mid G)$
  - In $P(X \mid Y)$, assuming that the generative probability is conditionally independent,
  - $P(Y \mid X, G) = C_0 P(Y \mid G) \prod P(x_i \mid y_i)$
  - Model them in a Markov random field, by the Hammersley-Clifford theorem,
    - $P(x_i \mid y_i) = \frac{1}{Z_1} \exp \{\sum \alpha_j f_j (x_{ij}, y_i)\}$
    - $P(Y \mid G) = \frac{1}{Z_2} \exp \{\sum \sum \mu_k h_k (Y_c)\}$
  - $Z_1$ and $Z_2$ are normalization factors.
Maximize likelihood

- Objective function
  - \( O(\theta) = \log P_{\theta}(Y | X, G) = \sum_i \sum_j \alpha_j f_j(x_{ij}, y_i) + \sum \mu_k h_k(Y_c) - \log Z \)

- Learning the model to
  - estimate a parameter configuration \( \theta = \{ \alpha, \mu \} \) to maximize the objective function:
  - that is, the goal is to compute \( \theta^* = \text{argmax } O(\theta) \)
Learning algorithm

- Goal : \( \theta^* = \arg\max O(\theta) \)

- The gradient of each \( \mu_k \) with regard to the objective function.
  - \( \frac{d\theta}{d\mu_k} = \mathbb{E}[h_k(Y_c)] - \mathbb{E}_{P_{\mu_k}(Y_c|X, G)}[h_k(Y_c)] \)
- A similar gradient can be derived for parameter \( \alpha_j \)

- One challenge : how to calculate the marginal distribution \( P_{\mu_k}(Y_c|X, G) \).
  - Approximate algorithms : Loopy Belief Propagation and Meanfield.
  - LBP : easy for implementation and effectiveness.
Learning algorithm (TriFG model)

Input: network $G^t$, learning rate $\eta$
Output: estimated parameters $\theta$

Initialize $\theta = 0$;
Repeat
  Perform LBP to calculate marginal distribution of unknown variables $P(y_i|x_i, G)$;
  Perform LBP to calculate marginal distribution of triad $c$, i.e. $P(y_c|x_c, G)$;
  Calculate the gradient of $\mu_k$ according to:
  \[ \frac{d\theta}{d\mu_k} = E[h_k(Y_C)] - E_{P_{\mu_k}(Y_c|X, G)}[h_k(Y_C)] \]
  Update parameter $\theta$ with the learning rate $\eta$:
  \[ \theta_{new} = \theta_{old} + \eta \frac{d\theta}{d\mu_k} \]
Until Convergence;
Prediction features

- Geographic distance
  - Global vs Local
- Homophily
  - Link homophily
  - Status homophily
- Implicit structure
  - Retweet or reply
  - Retweeting seems to be more helpful
- Structural balance
  - Two-way relationships are balanced (88%),
  - But, one-way relationships are not (only 29%).

(A) and (B) are balanced, but (C) and (D) are not.
Elite users have a stronger tendency to follow each other.
Our approach: TriFG

- TriFG model
  - Features based on observations
  - Partially labeled
  - Conditional random field
  - Triad correlation factors
Outline

- Previous works
- Our approach
- Experimental results
- Conclusion & future works
Data collection

- 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
Prediction performance

- Baseline algorithms
  - SVM & LRC & CRF
- Accurately infer 90% of reciprocal relationships in twitter.

<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>Test Case 1</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>Test Case 2</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.
Outline

- Previous works
- Our approach
- Experimental results
- Conclusion & future works
Conclusion

- Reciprocal relationship prediction in social network
- Incorporates social theories into prediction model.
- Several interesting phenomena.
  - Elite users tend to follow each other.
  - Two-way relationships on Twitter are balanced, but one-way relationships are not.
  - Social networks are going global, but also stay local.
Future works

- Other social theories for reciprocal relationship prediction.
- User feedback.
- Incorporating user interactions.
- Building a theory for different kinds of networks.
Thanks!
Q & A
Reference


Reference


