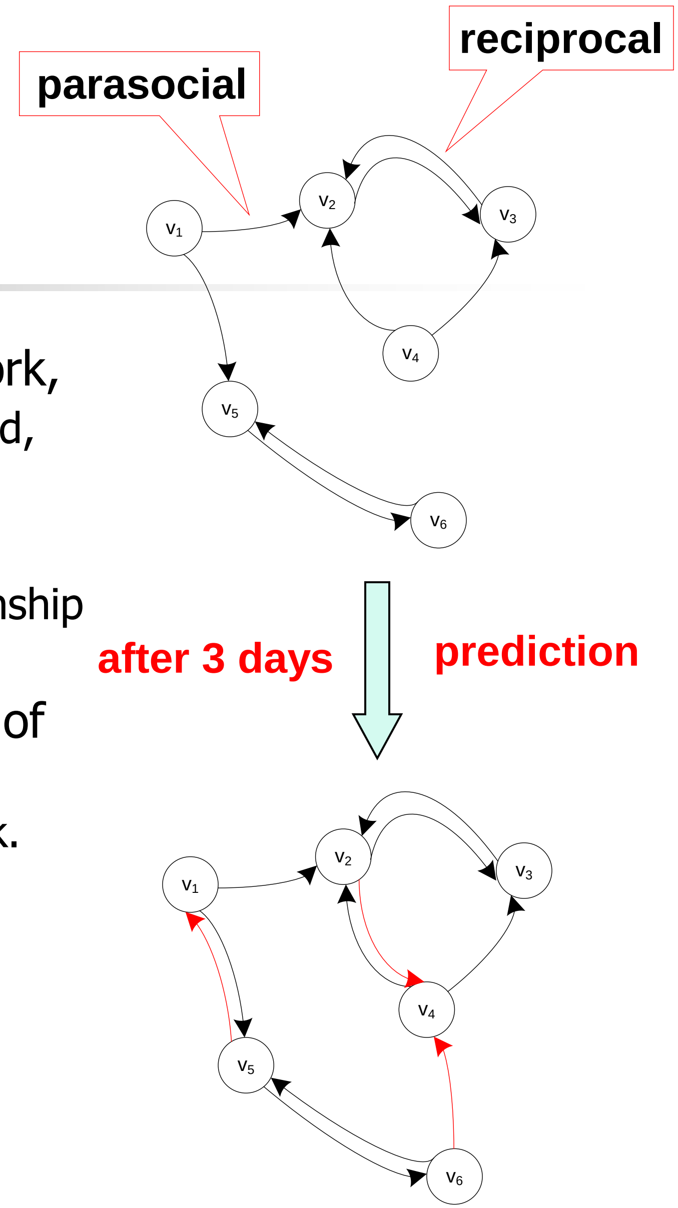
Who Will Follow You Back? Reciprocal Relationship Prediction*

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

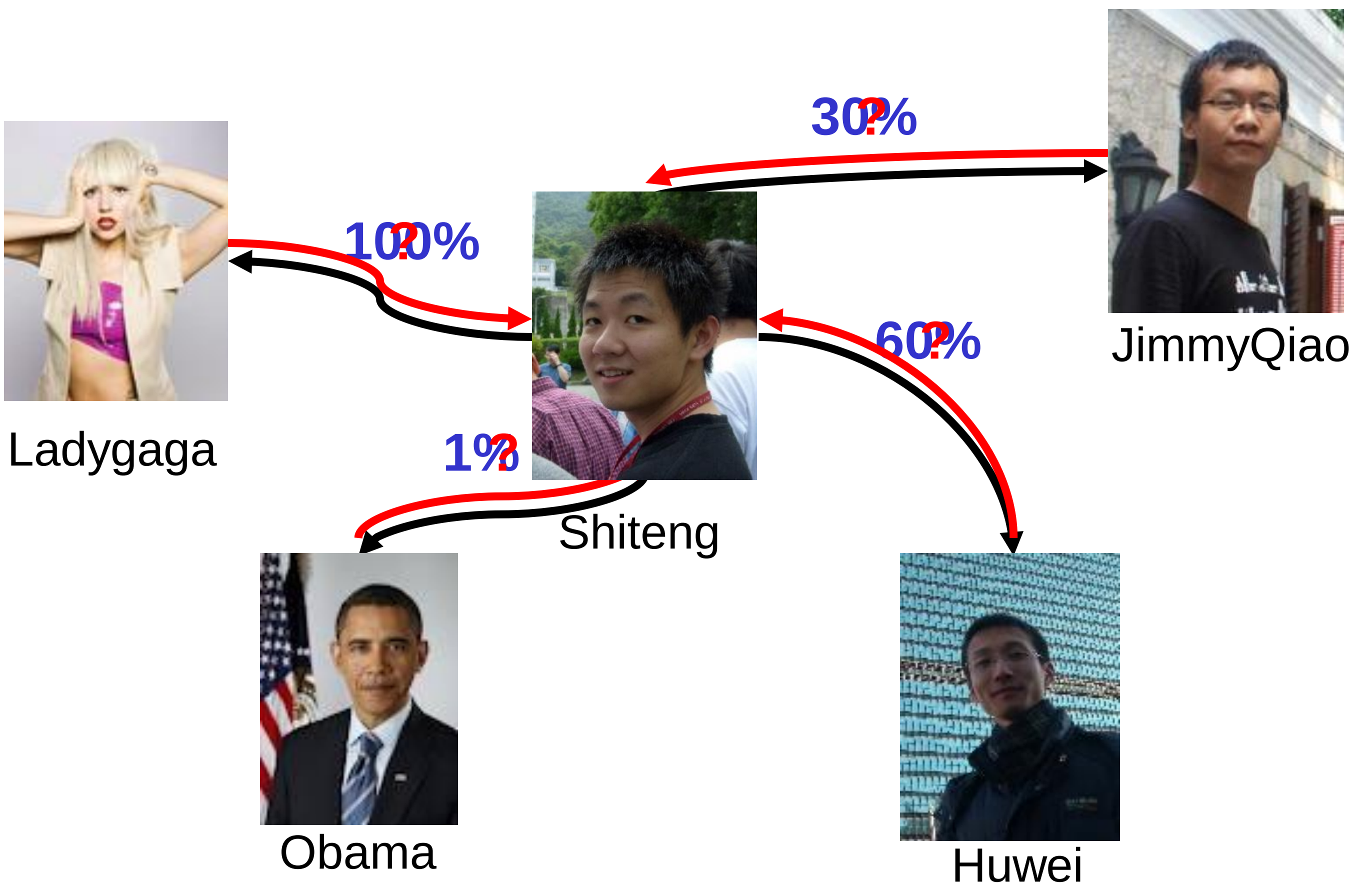
Motivation

- 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?



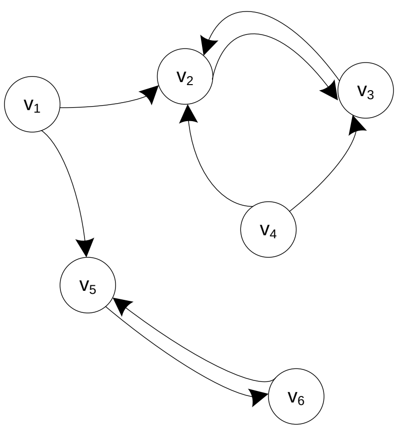
Example : real friend relationship

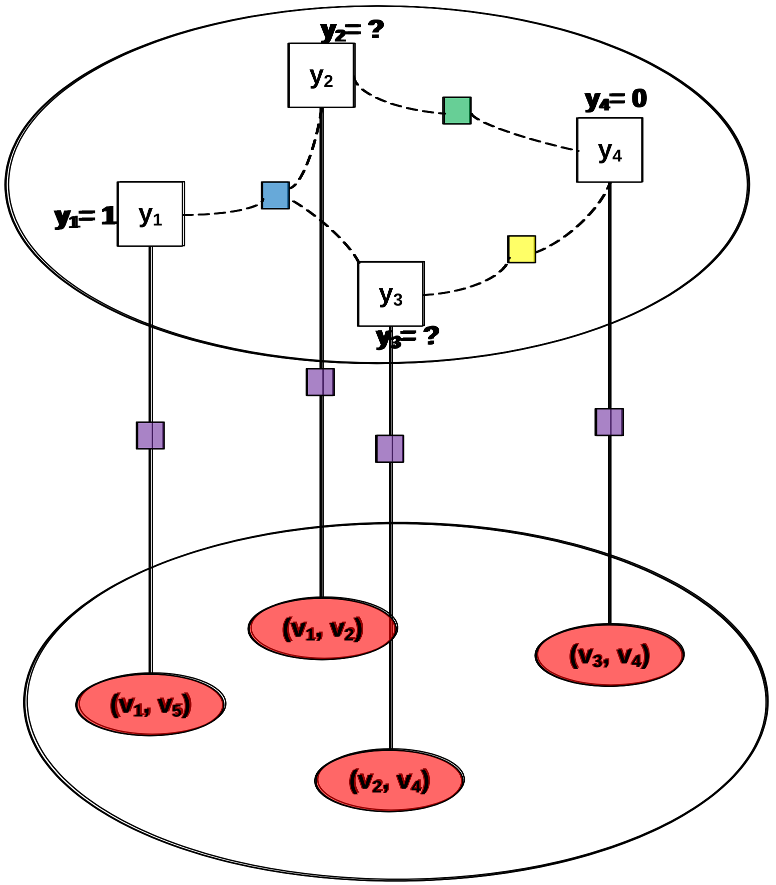
On Twitter : Who Will Follow You Back?



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 & intution, 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 difference 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
  - $P(Y \mid X, G) = P(X, G|Y)\, P(Y) / P(X, G) = C_0\, P(X \mid Y)\, P(Y \mid G)$
  - In $P(X \mid Y)$, assuming that the generative probability is conditionally independent,
  - $P(Y \mid X, G) = C_0 P(Y \mid G) \Pi P(x_i|y_i)$
  - Model them in a Markov random field, by the Hammersley-Clifford theorem,
    - $P(x_i|y_i) = 1/Z_1 * \exp\{\Sigma\alpha_j f_j (x_{ij}, y_i)\}$
    - $P(Y|G) = 1/Z_2 * \exp\{\Sigma_c \Sigma_k \mu_k h_k(Y_c)\}$
    - Z_1 and Z_2 are normalization factors.

Maximize likelihood

- Objective function
  - $O(\theta) = \log P_\theta(Y \mid X, G) = \Sigma_i \Sigma_j \alpha_j f_j (x_{ij}, y_i) + \Sigma\Sigma \mu_k h_k(Y_c) - \log Z$
- 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 : $\theta^* = \text{argmax } O(\theta)$

- The gradient of each μ_k with regard to the objective function.
  - $d\theta / d\mu_k = E[h_k(Y_c)] - E_{P_{\mu^k}(Yc|X, G)}[h_k(Y_c)]$
- A similar gradient can be derived for parameter α_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.

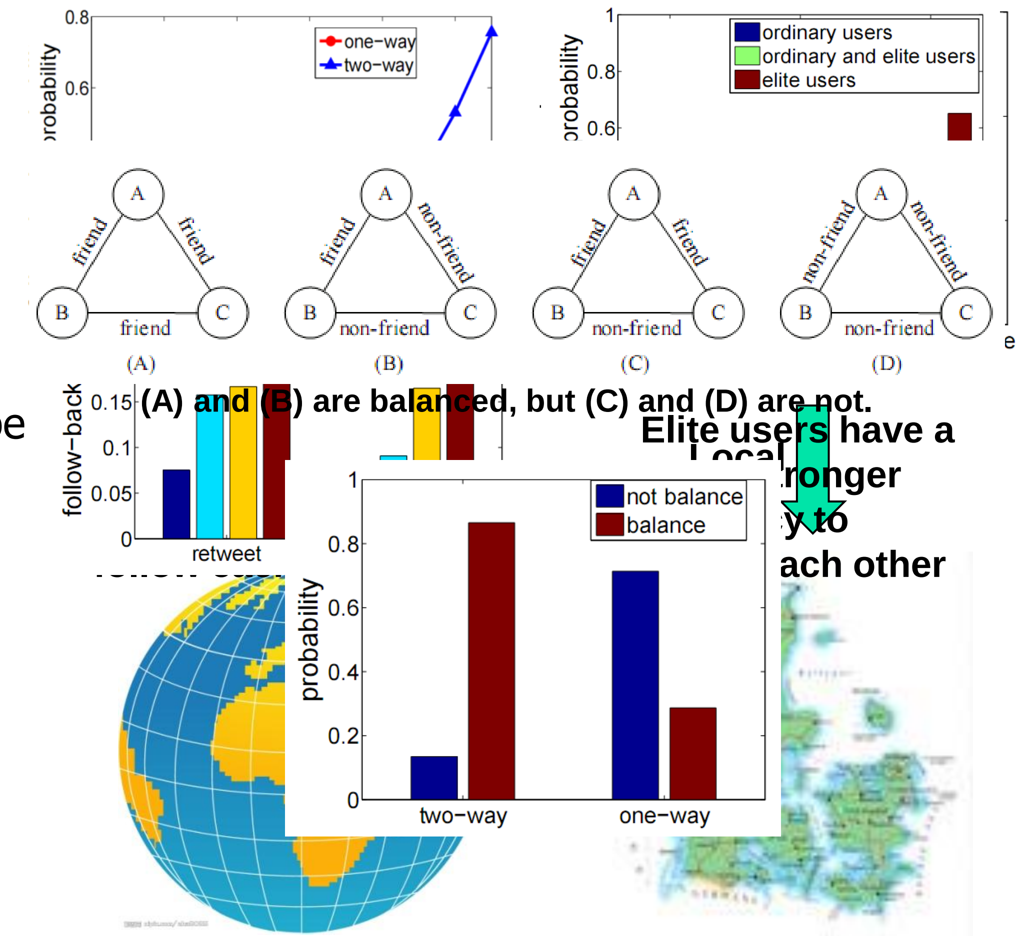
Learning algorithm(TriFG model)

Input : network G^t, learning rateη
Output : estimated parametersθ

Initalize θ= 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 μ_k according to :
    $d\theta / d\mu_k = E[h_k(Y_c)] - E_{P\mu^k(Yc|X,\ G)}[h_k(Y_c)]$
  Update parameter θ with the learning rate η:
    $\theta_{new} = \theta_{old} + \eta d\,\theta$
Until Convergence;

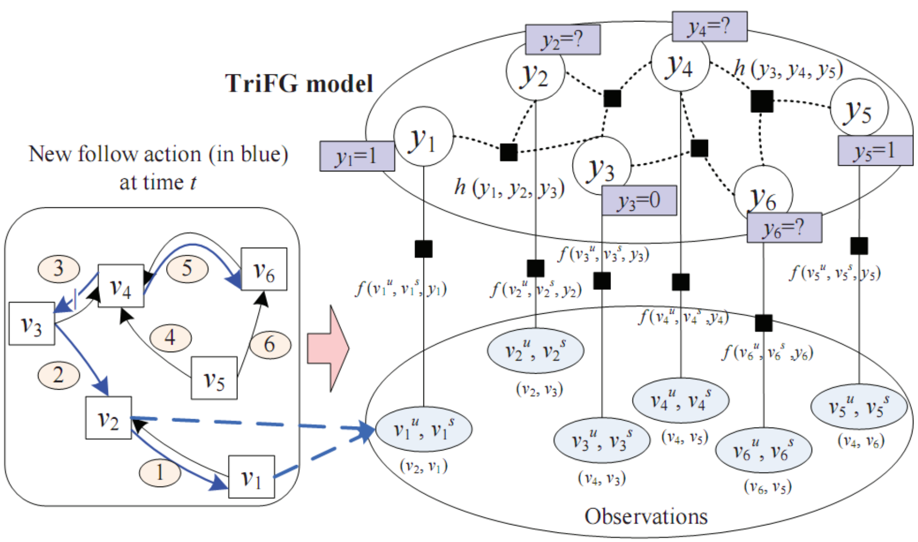
Prediction features

- 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%).



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

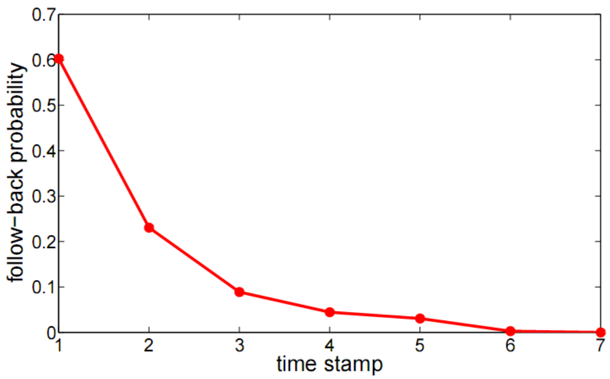
Prediction performance

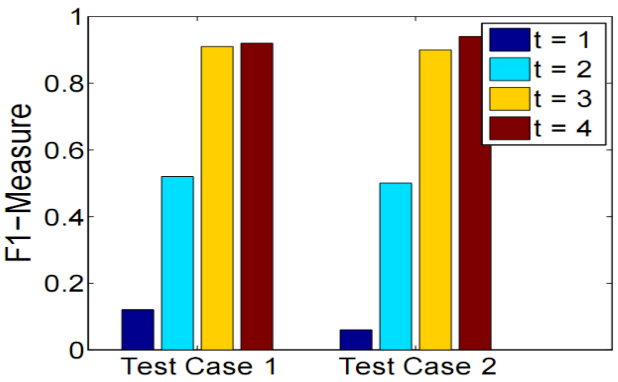
- Baseline algorithms
  - SVM & LRC & CRF
- Accurately infer 90% of reciprocal relationships in twitter.

Data	Algotithm	Precision	Recall	F1Measure	Accuracy
Test Case 1	SVM	0.6908	0.6129	0.6495	0.9590
	LRC	0.6957	0.2581	0.3765	0.9510
	CRF	1.0000	0.6290	0.7723	0.9770
	TriFG	1.0000	0.8548	**0.9217**	**0.9910**
Test Case 2	SVM	0.7323	0.6212	0.6722	0.9534
	LRC	0.8333	0.3030	0.4444	0.9417
	CRF	1.0000	0.6333	0.7755	0.9717
	TriFG	1.0000	0.8788	**0.9355**	**0.9907**

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

- [BL11] L.Backstrom and J.Leskovec. Supervised random walks : predicting and recommending links in social networks. In WSDM'11
- [C10] D.J.Crandall, L.Backstrom, D. Cosley, S.Suri, D.Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. PNAS, Dec. 2010
- [W10] C.Wang, J. Han, Y.Jia, J.Tang, D.Zhang, Y. Yu and J.Guo. Mining advisor-advisee relationships from research publication networks. In KDD'10.
- [N01]M.E.J. Newman. Clustering and preferential attachment in growing networks. Phys. Rev. E, 2001
- [L10] J.Leskovec, D.Huttenlocher, and J.Kleinberg. Predicting positive and negative links in online social networks. In WWW10.
- [T10] C.Tan, J. Tang, J. Sun, Q.Lin, and F.Wang. Social action tracking via noise tolerant time-varying factor graphs. In KDD10
- [T09] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In KDD09.

Reference

- [J07]A. Java, X.Song, T.Finin, and B.L. Tseng. Why we twitter : An analysis of a microblogging community. In KDD2007.
- [K10]H. Kwak, C.Lee, H.Park, and S.B. Moon. What is twitter, a social network or a news media? In WWW2010.
- [M10]M.Mathioudakis and N.Koudas. Twittermonitor : trend detection over the twitter stream. In SIGMOD10.
- [S10]T. Sakaki, M. Okazaki, and Y.Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In WWW10.