SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Inferring Explicit and Implicit Social Ties Simultaneously in Mobile Social Networks

Guojie Song[1*], Yuanhao Li[1], Junshan Wang[1] & Lun Du[1]

[1]*Key Laboratory of Machine Perception, Ministry of Education, Peking University, Beijing* 100871*, China*

Appendix A Analysis of Implicit Social Ties

In mobile social networks, members in a family or company usually generate a group where people have the same relations with each other. These relations are categorized into two types: explicit and implicit, according to whether they can be observed. The former can be captured by direct interactions between users, such as calls or messages. As for the latter, they exist in real life but seldom interactions can be observed. For example, when working at a company, two users may have no interaction, but they are still colleagues, or they just choose to use the inner network to communicate so that no interaction is recorded in mobile social networks. People may have few calls or messages with their families because they are usually together. What's more, there are a bunch of online social networks like Facebook nowadays. They gradually replace the traditional mobile phones on account of rapid development of the internet. So, a great portion of relations are hidden from interaction records in mobile social network, which is confirmed statistically in Table A1. About 40% of family relations and more than 80% of colleague relations are implicit. In addition to mobile social networks, implicit social ties also exist in other social networks like Twitter, Facebook, Wechat, Weibo and Taobao. In these social networks, explicit social ties generate when users follow others, become friends with others or join an interest communities. But it's challenging to have users giving all their relations. Many potential relations are hidden under the networks.

	Explicit Ties		Implicit Ties	
	Count	Ratio	Count	Ratio
Family	45689	60.01%	30442	39.99%
Colleague	3200230	12.14%	23159474	87.86%

Table A1 Ratio of explicit and implicit ties from dataset

Appendix B Community Features

Communities with different relation types have different features. These features provide rich information to help detect the community and infer its relation type. They can help infer both explicit and implicit social ties, which are the relation type of the community. Here, we propose two categories of community features: structural features and spatial features, and observe the difference of each feature between each relation type. We choose several typical and important features from each category. Then we introduce them and their effects in the following. We also use traditional edge-level features, such as interaction times, interaction homogeneity and entropy. For lack of space, we will not reproduce them but use them directly in our experiments.

Structural features represent the inherent network topology in a community, which can be calculated by interaction information between users. We choose community size and explicit edge clustering coefficient.

• **Community size**: One of the most simple and most important features is the community size. Community of families and colleagues have a significant difference in size, as shown in Figure B1. Families mainly have 2 to 3 members, and most of colleague cliques have more than 5 members. Despite the significant difference, families and colleagues have similar trends: high distribution at small-size, and low distribution at large-size.

* Corresponding author (email: gjsong@pku.edu.cn)

• **Explicit edge clustering coefficient**: Different from traditional clustering coefficient, what can be calculated in a relation clique is the explicit edge clustering coefficient. We treated those interacted relations as having edges, and the other implicit ties as no edges. Then we calculate the clustering coefficient of each community and summarize as Figure B2. As can be seen, families have high clustering coefficient as most of the family members will be likely to have some interactions. To the contrast, most of the colleagues may even have no idea of each other, let alone having interactions. As the result colleague communities are inclined to have low clustering coefficient.

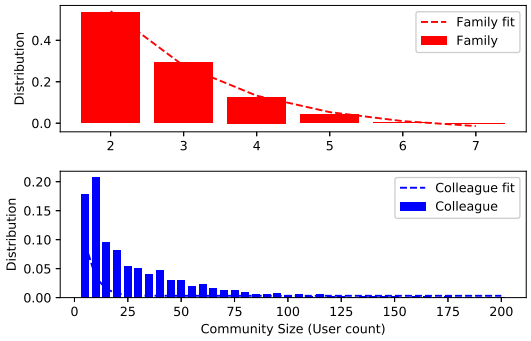

Figure B1 Community size

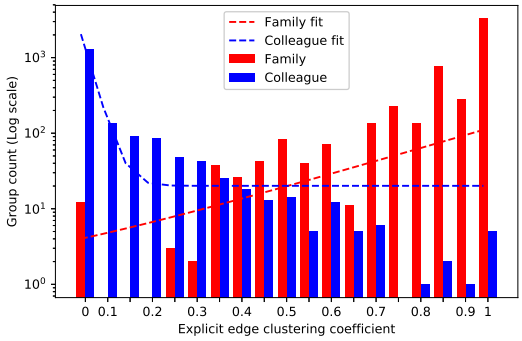

Figure B2 Explicit edge clustering coefficient

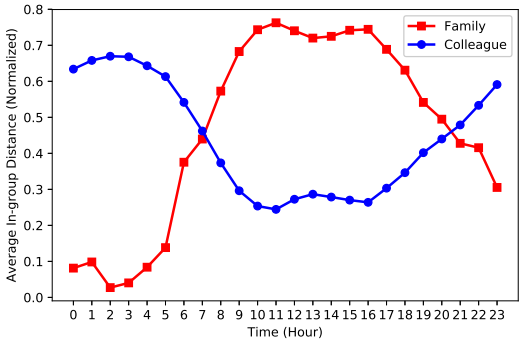

Figure B3 Spatial distance

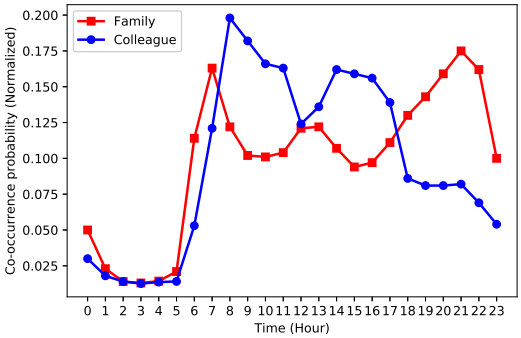

Figure B4 Spatial co-occurrence

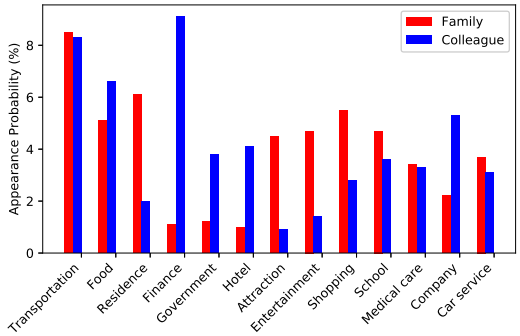

Figure B5 Spatial semantics

Spatial features represent the spatial information of a community, which are obtained by the trajectories of all individuals in this community. We choose three spatial features, which are distance, co-occurrence and semantics.

• **Spatial distance**: We calculate the in-group average distance of each pair of users to capture the distribution of users in a community in space. As there are high contingency of user locations, we only choose the major groups who have a considerable amount of event records and filter the other groups for a better presentation. We normalize the distances by the max possible distance of each two base stations. From Figure B3 it is clear that at family communities have high

distance at day time and low at night. This is because family members usually go to different spots, e.g. schools, offices, shops, and go back to home. Colleagues are on the contrary. At the day time most of the colleagues will stay at the office and go back to their own home at night.

• **Spatial co-occurrence**: Before the analysis, we define the concept of spatial co-occurrence. From the aspect of time, we polymerize our data by hourly granularity. From the aspect of space, we consider the fact that base stations in some places may be densely distributed. In that case, even though two users are in very close proximity, it is possible that they are exposed to different base station, so it is necessary to merge some base stations together. According to other scholars' work on algorithms of merging base stations, we use Voronoi diagrams algorithm [1] to merge adjacent base stations. Finally if two users locate in the same base station in one hour, then we define this phenomenon as spatial co-occurrence. The possibility of spatial co-occurrence is defined as: $C^h(x,y) = \sum_{l \in L} p_x^h(l) * p_y^h(l)$, where L is the assembly of base stations, $p_x^h(l)$ represents the possibility of user x appears at base station l at time h. Figure B4 shows that, family and colleague ties tend to have high co-occurrence in some hours in a day. Families members usually co-occur at the same place in the morning and the evening, at which time most of them will be at home. In contrast, colleagues have high co-occurrence at working hours, and dropped to a low level in the evening. These are also consistent with reality.

• **Spatial semantics**: We also analyze the distribution of spatial semantics of different relation users. Since our data only records the longitude and latitude information, we get nearby Point of Information (POIs) from a major map service provider in China. Through these POIs, we get the semantics of each POI near the base station. As there are multiple POIs around each single base station, we use TF-IDF(Term Frequency-Inverse Document Frequency), where TF defined as the frequency of a specific semantic in POIs nearby the base station and IDF as the logarithmic ratio of total number of base stations and the number of base stations containing a specific semantic. Results are shown in Figure B5. Users of different relationships have different spatial semantics. For example, families have high probability of appearing at residence, shopping and school, while colleagues tend to appear at finance, government and hotel. Spatial semantics reflect the purpose of a pair of people when appearing at a base station together.

Appendix C Community-based Recognition Model

Appendix C.1 Community Factor Graph

In Node Layer, v_i is a user. In Relation Layer, r_{ij} is the relation between user v_i and v_j, and y_{ij} represents its relation type. f and g respectively represent attribute feature function and correlation feature function of relation factor. In Community Layer, c_i is a community, and z_i represents the type of its relations. h and k respectively represent attribute feature function and correlation feature function of community factor.

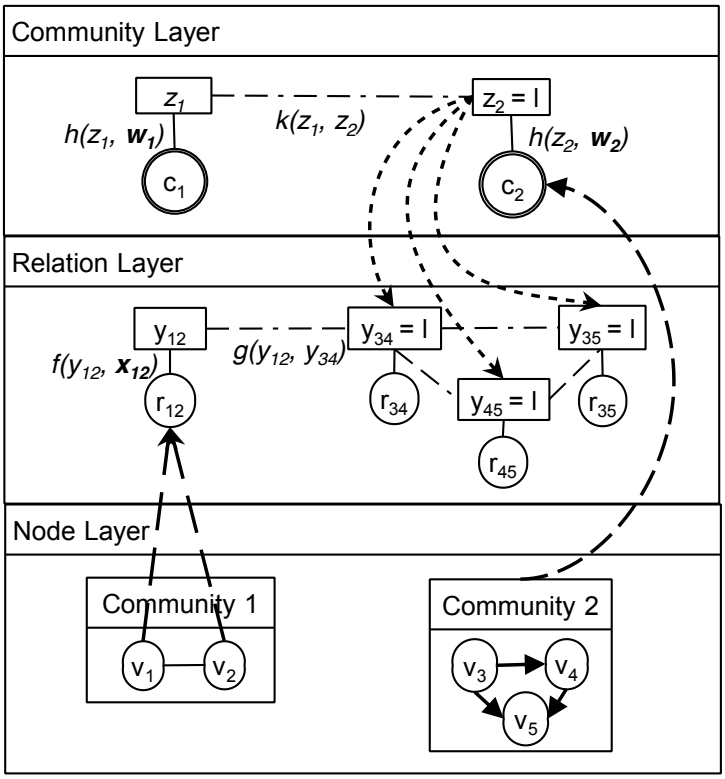

Figure C1 Community Factor Graph.

Appendix C.2 Relation Factor and Community Factor

In this paper we choose exponential-linear functions and we have the relation factor

$$F_R(y_i, \mathbf{x_i}) = f(y_i, \mathbf{x_i}) g(y_i, G(y_i)),$$

where

$$f(y_i, \mathbf{x_i}) = \frac{1}{Z_\lambda} exp\{\lambda^T \mathbf{F}(y_i, \mathbf{x_i})\};$$
$$g(y_i, G(y_i)) = \frac{1}{Z_\alpha} exp\{\sum_{y_t \in G(y_t)} \alpha^T \mathbf{G}(y_i, y_t)\}.$$

Feature function $f(y_i, \mathbf{x_i})$ represents the probability of y_i given the features $\mathbf{x_i}$, and $g(y_i, G(y_i))$ denotes the likelihood value of the correlation between relations, where $G(y_i)$ is the set of correlated relations to y_i. $\mathbf{F}$ is a vector of feature function and $\mathbf{G}$ is a vector of attribute function. We use traditional relation features $\mathbf{x_i}$ of relation r_i, which have been introduced in previous work [2].

In the same way we define the community factor

$$F_C(z_j, \mathbf{w_j}) = h(z_j, \mathbf{w_j}) k(z_j, G(z_j)),$$

where

$$h(z_j, \mathbf{w_j}) = \frac{1}{Z_\beta} exp\{\beta^T \mathbf{H}(z_j, \mathbf{w_j})\};$$
$$k(z_j, K(z_j)) = \frac{1}{Z_\gamma} exp\{\sum_{z_t \in K(z_j)} \gamma^T \mathbf{K}(z_j, z_t)\}.$$

Feature function $h(z_j, \mathbf{w_j})$ represents the probability of z_i given the features $\mathbf{w_i}$ and $k(z_j, G(z_j))$ denotes the likelihood value of the correlation between communities, where $G(y_i)$ is the set of correlated communities to z_j. $\mathbf{H}$ is a vector of feature function and $\mathbf{K}$ is a vector of attribute function.

Appendix C.3 Learning

Algorithm C1 CFG Learning Algorithm

Require: CFG G; Learning rate η;
Ensure: Learned parameter θ;
1: Initialize θ;
2: **repeat**
3: **for** Each layer **do**
4: Calculate posterior probability and likelihood value using LBP
5: **end for**
6: Calculate $E_{p_\theta(Y|G)} S$;
7: Calculate gradient of θ according to Equation C3
8: Update θ by η: $\theta_{new} = \theta_{old} - \eta \cdot \frac{\partial O(\theta)}{\partial \theta}$
9: **until** Convergence
10: **return** θ;

Learning CFG is to give an estimate of parameter $\theta = (\lambda, \alpha, \beta, \gamma)$, s.t. the likelihood function reaches the maximum value on labeled input G^L. As the input is given, each explicit or implicit relation are assigned certainly. Thus we can easily solve both the explicit and implicit model learning together, which follows in the latter paragraphs.

For a simple presentation, we concatenate each factor functions to $s(y_i) = (F(y_i, \mathbf{x_i})^T, \sum_{y_t} G(y_i, y_t)^T)^T$, $s(z_j) = (H(z_j, \mathbf{w_j})^T, \sum_{z_t} K(z_j, z_t)^T)^T$. Then the joint probability is written as:

$$p(Y|G) = \frac{1}{Z} \prod_{i,j} exp\{(\lambda, \alpha)^T s(y_i)\} exp\{(\beta, \gamma)^T s(z_j)\} = \frac{1}{Z} exp\{\theta^T S\}, \tag{C1}$$

where $Z = Z_\lambda Z_\alpha Z_\beta Z_\gamma$ is the normalization factor for $p(Y|G)$. S is the aggregation of all the feature functions for all the relation and community nodes. As the input is labeled, Y^L is fixed. The log-likelihood objective function $O(\theta)$ can be defined as:

$$\begin{aligned} O(\theta) &= \ln p(Y^L|G) = \ln \sum_{Y^L} \frac{1}{Z} exp\{\theta^T S\} \\ &= \ln exp\{\theta^T S_{Y^L}\} - \ln \sum_{Y} exp\{\theta^T S\}. \end{aligned} \tag{C2}$$

Having the definition of $O(\theta)$, this problem becomes to find a θ to maximize $O(\theta)$. We apply Newton-Raphson method to solve the objective. Specifically, we calculate the gradient for the parameter vector θ:

$$\begin{aligned}\frac{\partial O(\theta)}{\partial \theta} &= \frac{\partial(\ln exp\{\theta^T S_{Y^L}\} - \ln \sum_Y exp\{\theta^T S\})}{\partial \theta} \\ &= S_{Y^L} - E_{p_\theta(Y|G)} S.\end{aligned} \tag{C3}$$

Though the CFG have multiple layers, as can be seen from C1, the parameters and feature functions in different layers do not affect each other, so that each layer can be calculated individually. Another problem is each layer have a complex structure and may contain several loops and in this paper we use LBP [3]. Algorithm C1 summarize the learning process.

Appendix C.4 Inferring

Algorithm C2 Explicit Tie Inferring Based on LBP & SA

Require: CFG G; Model parameter θ; Initial temperature T_0; Random variation amount k;
Ensure: Node label configuration Y and coefficient P;

Initialize label Y;
Without using community features, calculate $E_{p_\theta(Y|G)}S$ using LBP ONLY on EDGE LAYER;
Calculate node labels: $y_i = argmax(p_\theta(y_i|G, U_i))$ $C_j = (V_j, E_j), V_j \subset V, E_j \subset E$, where if $e_i \in E_j$, then $y_i = z_j$
Initialize Temperature $T = T_0$:
repeat
  Randomly select $1\tilde{}k$ nodes and exchange them to other communities;
  Calculate new label Y' using LBP and Calculate joint distribution according to Equation C1
  if new solution Y' is better **then**
    Keep the new solution $Y = Y'$
  else
    Let the amplitude be $\delta = \frac{p(Y|G) - p'(Y|G)}{p(Y|G)}$, accept the new solution by the probability of $e^{-\frac{\delta}{T}}$;
  end if
  Update the temperature $T(t) = \frac{T_0}{ln(1+t)}$
until $T = 0 (T < \epsilon)$
return Y, P;

Algorithm C3 Implicit Tie Inferring Based on LBP & SA

Require: CFG G; Model parameter θ; Initial temperature T_0; Random variation amount k;
Ensure: Node label configuration Y and coefficient P;

Initialize label Y;
Calculate node labels similar to explicit tie inferring.
for each communities c_i **do**
  Add all e_j **to** E **and** E_i **if** e_j **is not currently in** E**.**
end for
Initialize Temperature$T = T_0$:
repeat
  Randomly select $1\tilde{}k$ nodes and exchange them to other communities;
  for each changed communities **do**
    Add all e_i **to** E **if** e_i **is not currently in** E**.**
  end for
  Calculate new label Y' using LBP;
  SA similar to explicit tie inferring.
until $T = 0 (T < \epsilon)$
return Y, P;

Appendix D PR-Curve

Figure D1 shows the Precision-Recall Curve of our model, analyzing the model sensitivity and showing that our model has good performance with different thresholds. Figure D2 shows the Precision-Recall Curve of our model, analyzing the model sensitivity and showing that the performance of our model is stable with different thresholds.

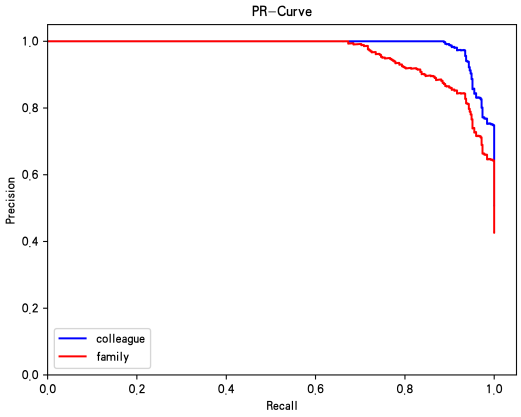

Figure D1 Precision-Recall Curve of our model for explicit tie inferring

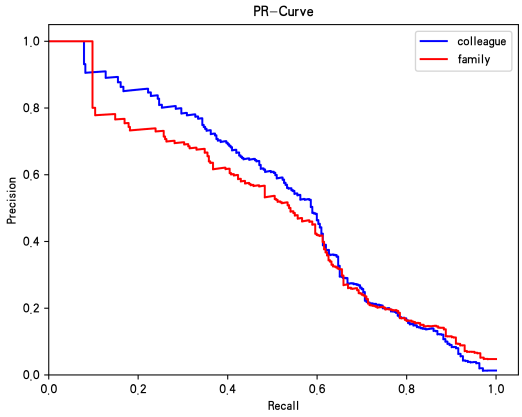

Figure D2 Precision-Recall Curve of our model for implicit tie inferring

Appendix E Feature Analysis

We conduct the experiment to prove that each kind of features is useful and make contribution to the improvement of F1-score, for both explicit and implicit social ties inferring of both family and colleague.

We evaluate our model by adding each kind of features step by step. The baseline is to learn the factor graph only with interaction features of edges. Based on this, the next two versions are to learn the factor graph with either structural features or spatial features, which are community features. And the final version is our model, which contains all edge features and community features.

Figure E1 shows that for explicit tie recognition, both structural and spatial features make contribution, but structural features are more important. And Figure E2 shows that for implicit tie recognition, all community features make contribution, too, but spatial features are more important. Both structural and spatial features provide useful and extra information to detect communities and infer more social ties in our model. And for explicit tie inferring, the structural features are more important, while for implicit tie inferring, inferring the relation between two users without interactions relies more on spatial information.

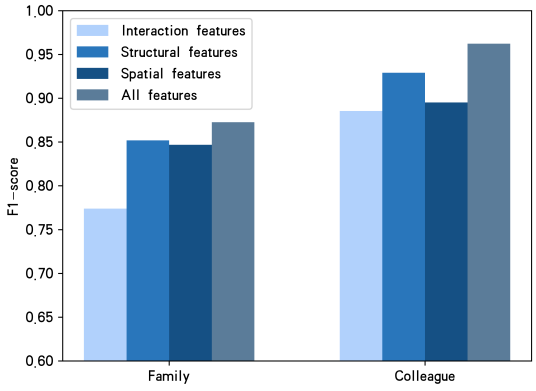

Figure E1 Feature analysis for explicit tie inferring

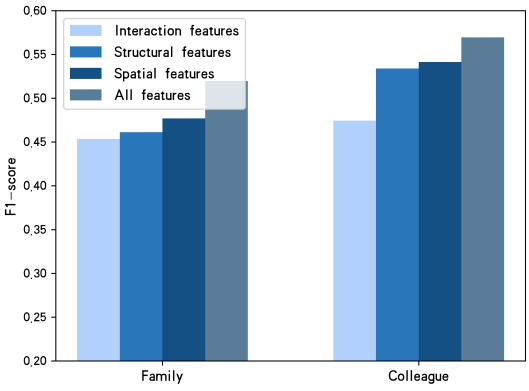

Figure E2 Feature analysis for implicit tie inferring

References

1 Aurenhammer F. Voronoi diagramsa survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR), 1991, 23(3): 345-405.Aurenhammer F. Voronoi diagramsa survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR), 1991, 23(3): 345-405.

2 Tang J, Lou T, Kleinberg J. Inferring social ties across heterogenous networks. In: Proceedings of the fifth ACM international conference on Web search and data mining, 2012: 743-752.

3 Granville V, Krivnek M, Rasson J P. Simulated annealing: A proof of convergence. IEEE transactions on pattern analysis and machine intelligence, 1994, 16(6): 652-656.

4 Sadilek A, Kautz H A, Silenzio V. Modeling Spread of Disease from Social Interactions. In: ICWSM. 2012: 322-329.

5 Xu J J, Chen H. CrimeNet explorer: a framework for criminal network knowledge discovery. ACM Transactions on Information Systems (TOIS), 2005, 23(2): 201-226.

6 Domingos P. Mining social networks for viral marketing. IEEE Intelligent Systems, 2005, 20(1): 80-82.

7 Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011: 1082-1090.

8 Crandall D J, Backstrom L, Cosley D, et al. Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences, 2010, 107(52): 22436-22441.

9 Eagle N, Pentland A S, Lazer D. Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences, 2009, 106(36): 15274-15278.

10 Wang D, Pedreschi D, Song C, et al. Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011: 1100-1108.

11 Dong Y, Tang J, Wu S, et al. Link prediction and recommendation across heterogeneous social networks. In: Data mining (ICDM), 2012 IEEE 12th international conference on, 2012: 181-190.

12 Ma H. An experimental study on implicit social recommendation. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013: 73-82.

13 Krebs V E. Mapping networks of terrorist cells. Connections, 2002, 24(3): 43-52.

14 Diehl C P, Namata G, Getoor L. Relationship identification for social network discovery. In: AAAI. 2007, 22(1): 546-552.

15 Wang C, Han J, Jia Y, et al. Mining advisor-advisee relationships from research publication networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010: 203-212.

16 Tang W, Zhuang H, Tang J. Learning to infer social ties in large networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg, 2011: 381-397.

17 Tang J, Lou T, Kleinberg J. Inferring social ties across heterogenous networks. In: Proceedings of the fifth ACM international conference on Web search and data mining, 2012: 743-752.

18 Easley D, Kleinberg J. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, 2010.

19 Roth M, Ben-David A, Deutscher D, et al. Suggesting friends using the implicit social graph. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010: 233-242.

20 Gupte M, Eliassi-Rad T. Measuring tie strength in implicit social networks. In: Proceedings of the 4th Annual ACM Web Science Conference, 2012: 109-118.

21 LibenNowell D, Kleinberg J. The linkprediction problem for social networks. journal of the Association for Information Science and Technology, 2007, 58(7): 1019-1031.

22 Backstrom L, Leskovec J. Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on Web search and data mining, 2011: 635-644.

23 Wang C, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, 2007: 322-331.

24 Guimer R, Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences, 2009, 106(52): 22073-22078.

25 Fortunato S. Community detection in graphs. Physics reports, 2010, 486(3-5): 75-174.

26 Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.

27 Newman M E J. Fast algorithm for detecting community structure in networks. Physical review E, 2004, 69(6): 066133.

28 Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm. In: Advances in neural information processing systems, 2002: 849-856.

29 Aurenhammer F. Voronoi diagramsa survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR), 1991, 23(3): 345-405.

30 Granville V, Krivnek M, Rasson J P. Simulated annealing: A proof of convergence. IEEE transactions on pattern analysis and machine intelligence, 1994, 16(6): 652-656.

31 Zhou T, L L, Zhang Y C. Predicting missing links via local information. The European Physical Journal B, 2009, 71(4): 623-630.

32 Tang J. Computational Models for Social Network Analysis: A Brief Survey. In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017: 921-925.

33 Du L, Wang Y, Song G, et al. Dynamic Network Embedding: An Extended Approach for Skip-gram based Network

Embedding[C]//IJCAI. 2018: 2086-2092.
34 Tang J, Lou T, Kleinberg J, et al. Transfer learning to infer social ties across heterogeneous networks. ACM Transactions on Information Systems (TOIS), 2016, 34(2): 7.
35 Rozenshtein P, Tatti N, Gionis A. Inferring the Strength of Social Ties: A Community-Driven Approach. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 1017-1025.
36 Taheri S M, Mahyar H, Firouzi M, et al. Extracting implicit social relation for social recommendation techniques in user rating prediction. In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017: 1343-1351.
37 Ali I, Hong J, Kim S W. Exploiting implicit and explicit signed trust relationships for effective recommendations. In: Proceedings of the Symposium on Applied Computing, 2017: 804-810.
38 Fazeli S, Loni B, Bellogin A, et al. Implicit vs. explicit trust in social matrix factorization. In: Proceedings of the 8th ACM Conference on Recommender systems, 2014: 317-320.
39 Wang P, Xu B W, Wu Y R, et al. Link prediction in social networks: the state-of-the-art. Science China Information Sciences, 2015, 58(1): 1-38.
40 Du L, Lu Z, Wang Y, et al. Galaxy Network Embedding: A Hierarchical Community Structure Preserving Approach[C]//IJCAI. 2018: 2079-2085.
41 Gong J, Gao X, Cheng H, et al. Integrating a weighted-average method into the random walk framework to generate individual friend recommendations. Science China Information Sciences, 2017, 60(11): 110104.
42 Resnick P, Varian H R. Recommender systems. Communications of the ACM, 1997, 40(3): 56-58.
43 Qin Y, Yu Z, Wang Y, et al. Detecting micro-blog user interest communities through the integration of explicit user relationship and implicit topic relations. Science China Information Sciences, 2017, 60(9): 092105.
44 Guojie Song, Xiabing Zhou, Yu Wang, Kunqing Xie. Influence Maximization on Large-Scale Mobile Social Network: A Divide-and-Conquer Method. IEEE Transactions on TPDS2015:1379-1392.
45 Yu Wang, Gao Cong, Guojie Song, Kunqing Xie. Community-based greedy algorithm for mining top-K influential nodes in social networks. KDD:2010:1039-1048.
46 Qin Tian, Wufan Shangguan, Guojie Song, Jie Tang. Spatio-Temporal Routine Mining on Mobile Phone Data. ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 12 Issue 5, July 2018, Article No. 56.