Friendship Prediction using Semi-supervised Learning of Latent Features in Smartphone Usage Data

Yuka Ikebe[1], Masaji Katagiri[1,2] and Haruo Takemura[3,2]
[1]*R&D Center, NTT DOCOMO, Inc., Hikari-no-oka, Yokosuka-shi, Japan*
[2]*Graduate School of Information Science and Technology, Osaka University, Suita-shi, Japan*
[3]*Cybermedia Center, Osaka University, Ibaraki-shi, Japan*

Keywords: Link Prediction, Matrix Factorization, Latent Feature, Semi-supervised Learning, One-class Setting.

Abstract: This paper describes a semi-supervised learning method that uses smartphone usage data to identify friendship in the one-class setting. The method is based on the assumption that friends share some interests and their smartphone usage reflects this. The authors combine a supervised link prediction method with matrix factorization which incorporates latent features acquired from the application usage and Internet access. The latent features are optimized jointly with the process of link prediction. Moreover, the method employs the sigmoidal function to estimate user affinities from the polarized latent user features. To validate the method, fifty university students volunteered to have their smartphone usage monitored for 6 months. The results of this empirical study show that the proposal offers higher friendship prediction accuracy than state-of-the-art link prediction methods.

1 INTRODUCTION

With the prevalence of social networking services (SNSs) such as Facebook, Twitter and MySpace, the mining of social network data is attracting more attention because the results appear promising for increasing the sales of products and services through marketing strategies such as viral marketing.

However, according to a survey (Keller and Fay, 2009), such effects are mostly observed in real face-to-face inter-personal relationships rather than in cyber relationships on SNSs at this moment. This implies that the social networks acquired from SNSs are not sufficient to acquire friend networks. Although other information such as phone calls and/or e-mail records is considered to be useful for identifying the desired friend networks, they are, in practice, very difficult to collect on a massive scale due to privacy concerns. Moreover, they still reflect only a part of the whole network. Fig.1 shows a real example of the friend networks collected by the authors (details are shown in a later section). Solid lines represent the friends captured by monitoring phone calls for 6 months. Dotted lines are showing the ground truth obtained by questionnaires. The figure implies that, in reality, only a small portion is likely to be observable, and thus the capability to predict friendship from

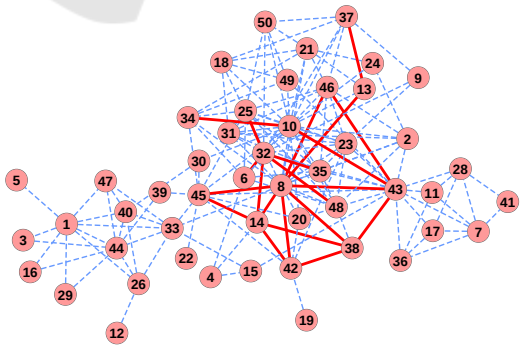

Figure 1: An example of friend network.

a small set of known links is important. This motivated the authors into proposing a semi-supervised method to predict friendship relations by utilizing application execution and web access histories on smartphones, rather than monitoring communication activities directly. The idea is based on the assumption that friends share some interests, a characteristic known as *homophily* (McPherson et al., 2001), and their smartphone usages reflect this.

A straight forward approach to predicting friendship is to apply link prediction methods. Link prediction is a common binary classification task and has been used in many studies (Lü and Zhou, 2011).

Ikebe Y., Katagiri M. and Takemura H..
Friendship Prediction using Semi-supervised Learning of Latent Features in Smartphone Usage Data.
DOI: 10.5220/0004133201990205
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 199-205
ISBN: 978-989-8565-29-7

Friendship prediction is seen as a subtask of link prediction. However, there are three major difficulties with friendship prediction not solved by existing link prediction methods;

One-class Setting. Since it is virtually impossible to monitor all activities of a pair of people, it is possible that they are friends even if no interaction is observed between the pair. Thus, in practice, we cannot distinguish known absent link from unknown link. Therefore, available observations consist of known present links only, i.e. positive samples only, and no negative samples are available for learning.

Sparse Friend Network. In general and practical settings, the number of friends is extremely small compared to the total number of participants. In conjunction with the one-class setting, this yields a small number of known links, i.e. most links remain unknown.

Affinity Differs from Similarity. General node similarity approaches such as cosine similarity do not always match the characteristics of friendship, especially given that node information is highly dimensional and sparse.

Existing missing link prediction methods can be classified into two types: (1) topological-information-based and (2) node-information-based. The first type uses only known network structure such as common adjacent nodes or paths between nodes. Well-known methods are Jaccard Similarity (Lü and Zhou, 2011) and Adamic/Adar (Adamic and Adar, 2003). These methods are problematic if the known network is sparse. The second type uses information of nodes as well as known network structure; they try to predict links even if a node is isolated from known links. Since the number of known links is small for friendship prediction as mentioned above, the proposed method takes the node-information-based approach. Two state-of-the-art node-information-based link prediction methods were proposed recently.

Latent Feature Model: Menon et al. proposed a supervised learning method to predict links by applying latent feature model (LFL) in (Menon and Elkan, 2010) and (Menon and Elkan, 2011). This method tries to minimize the loss between predicted results and known present/absent links by adjusting latent features. Yang et al. proposed Joint Friendship and Interest Propagation (FIP) which combines latent feature models for user–user and user–item (Yang et al., 2011).

Link Propagation: Kashima et al. proposed the Link Propagation method; it tries to propagate known links using observed node features with pre-specified kernel (Kashima et al., 2009). If observed node features are highly dimensional and sparse, it is not trivial to construct the proper kernel.

Besides the works related to link prediction, several data-oriented approaches have been reported for friendship prediction. The data used are;

1. Location data such as GPS coordinates: (Wang et al., 2011), (Eagle et al., 2009), (Scellato et al., 2011),
2. Bluetooth encounter data: (Quercia et al., 2010), (Eagle et al., 2009),
3. Call records: (Wang et al., 2011), (Mirisaee et al., 2010).

Although location trajectory and encounter data show a strong correlation to friendship, they capture only the relationships that yield frequent physical contacts. Call records are hamstrung by privacy issues and so are impractical for this purpose.

The authors' basic idea to overcoming the one-class setting and the sparsity of friend networks is to incorporate rich user information for friendship prediction. Here, rich user information means application execution and web browsing histories. Application execution and web browsing are universal activities for any smartphone user, and so it is reasonable to expect those histories to be available for most users, unlike friendship links. Since standard operating systems can create the logs needed, it is also practical in terms of deployment. Moreover, application execution and web browsing histories are potentially informative enough, since recent research (Fujimoto et al., 2011) showed that user's interests can be extracted from the web browsing history. However, note that we need to extract interests from histories, since they are expressed by items and are not directly observable.

Yang et al. proposed a strategy similar to that of the FIP model in (Yang et al., 2011); it enables the incorporation of user–item interaction into user–user modeling. However, they assume that the item is the key point of interest and focus in obtaining latent features from the observed node (i.e. user and item) information, not from user–item interaction, because of its different problem setting.

Here, the authors propose to employ matrix factorization to extract latent user features from each user's application execution and Internet access records. The authors believe matrix factorization on user–item interaction is promising as a method to identify latent features, since it is successful in collaborative filtering by extracting users' and items' latent features from

the observed user–item rating matrix (Koren et al., 2009). The supervised link prediction method is modified to incorporate matrix factorization, so that the results of matrix factorization are optimized jointly with link prediction. Moreover, to better model friendship, the authors propose a specifically designed affinity measure that reflects the assumed correlation between friendship and the degree of interest matching. This approach allows the proposed method to optimize latent features (obtained by matrix factorization with all user–item observations) and link prediction (supervised by known links only) simultaneously in terms of the proposed affinity measure through semi-supervised learning. To the best of our knowledge, this is the first paper to propose friendship prediction in the one-class setting with an evaluation conducted on realistic data.

In summary, this work makes the following contributions;

- Proposed semi-supervised friendship prediction method that combines supervised link prediction method with matrix factorization technique in order to acquire latent features from smartphone application execution and web access histories. It also employs a specific affinity measure based on a prior knowledge of the level of interest matching.
- Validation of the method by real monitored data from 50 university students.

This paper is organized as follows. Section 2 explains our formulation. Section 3 describes the proposed method for friendship prediction. Section 4 explains the collected data and evaluation results. Section 5 concludes the paper.

2 FORMULATION

2.1 Supervised Link Prediction

A supervised link prediction method using known node information $\mathbf{X}$ is formulated as identifying unknown parameters θ that minimize the following objective function;

$$\min_{\theta} \left\{ \sum_{(i,j)\in O} \ell(G_{ij}, \hat{G}_{ij}(\mathbf{X},\theta)) + \lambda\Omega(\theta) \right\}, \quad (1)$$

where O represents the set of known links. G is an $n \times n$ adjacency matrix that indicates the graph structure, in which each element takes one of three values, 0, 1 and ? representing a known absent link, a known present link, and an unknown link, respectively. Here, n represents the number of nodes. Note that in the one-class setting, the value of G_{ij} takes either 1 or ?. $\hat{G}$ represents the predicted graph structure. ℓ is a loss function, Ω is a regularization term (e.g. $\ell 2$ norm) and λ is a weight parameter for the regularization term. $\mathbf{X}$ represents a set of known node information such as observed features. θ represents the set of unknown parameters to be acquired.

During the training phase, unknown parameters, θ, are obtained by optimizing the objective function Eq.1. Upon completion of training, $\hat{G}$ gives predictive results for unknown links using the identified θ and the known node information $\mathbf{X}$.

2.2 Matrix Factorization and Latent Feature

In the context of friendship prediction, let $\mathbf{X}$ be a user–item matrix which represents observed frequency of item usage on each user. The dimension of $\mathbf{X}$ is $U \times I$, where U and I represent the number of users and items, respectively. The equation below describes the matrix factorization of $\mathbf{X}$;

$$\mathbf{X} \approx \mathbf{WH}. \quad (2)$$

Matrices $\mathbf{W}$ and $\mathbf{H}$ are the outcome of matrix factorization. $\mathbf{W}$ becomes a latent user feature matrix whose dimension is $U \times L$, and $\mathbf{H}$ becomes a latent item feature matrix with dimension of $L \times I$. Here L is a given constant positive integer which defines the number of dimensions for the latent features. Eq.3 is the typical objective function used,

$$\min_{\mathbf{W},\mathbf{H}} \left\{ \|\mathbf{X} - \mathbf{WH}\|^2 + \lambda'\Omega'(\mathbf{W},\mathbf{H}) \right\}, \quad (3)$$

where Ω' is a regularization term (e.g. $\ell 2$ norm) and λ' is a weight parameter of the regularization term.

3 PROPOSED METHOD

3.1 Affinity Measure

To predict friendship, identifying the proper covariate is one of the key requirements. Cosine similarity of node features is a commonly used conventional measure. However, the possibility of being friends does not always follow such "similarity". For instance, suppose user A has several interests such as baseball, fishing and camping. User A's friend, user B, does not necessarily like all of them. Typically, user B shares only a few of the interests. Here, intuitively we can see that the key factor in friendship is existence of common topics of interest. In other words,

Table 1: Parameters in the proposed method.

Parameter	Type	Explanation
α	Controlled	A weight parameter for the supervised term in Eq.8
β	Controlled	A weight parameter for the matrix factorization term in Eq.8
λ'''	Controlled	A weight parameter for the regularization term in Eq.8
g	Controlled	Gain of sigmoid function in Eq.5
th	Controlled	Flexion point of sigmoid function in Eq.5 = threshold
L	Controlled	The number of latent feature dimensions = Number of columns of matrix W = Number of rows of matrix H
W,H	Inferred	Latent user / item matrix

the strength of interest is much less important in predicting friendship. Thus, based on this consideration, the authors propose to model affinity between user i and j, denoted by f_{ij}, as the inner product of polarized latent user features using the following sigmoidal functions;

$$f_{ij}(\mathbf{W}) = \frac{\sigma(\mathbf{w}_i) \cdot \sigma(\mathbf{w}_j)}{L}, \quad (4)$$

where,

$$\sigma(\mathbf{w}_i) = \left(\frac{1}{1+e^{-g(w_{i1}-th)}}, \cdots, \frac{1}{1+e^{-g(w_{iL}-th)}} \right). \quad (5)$$

Here, $\mathbf{w}_i$ represents the latent user feature for user i; it is a vector extracted from the ith row of matrix $\mathbf{W}$, $\mathbf{w}_i = (w_{i1}, \cdots, w_{iL})$. th and g represent a threshold and a gain of sigmoid function, respectively. $\sigma(\mathbf{w}_i)$ polarizes each component of vector $\mathbf{w}_i$ using sigmoid function, thus the value of f_{ij} ranges $0 \leq f_{ij} \leq 1$.

3.2 Embedding Latent Features to Supervised Link Prediction

In the one-class setting, G_{ij} is always 1 for any $(i,j) \in O$. In addition, affinity f_{ij} acts as the covariate of $\hat{G}_{ij}$ for friendship prediction. Thus, Eq.1 can be modified into;

$$\min_{\mathbf{W}} \left\{ \sum_{(i,j)\in O} \ell(1, f_{ij}(\mathbf{W})) + \lambda'' \Omega''(\mathbf{W}) \right\}. \quad (6)$$

In order to combine affinity measure with supervised link prediction, Eq.3 and Eq.6 need to be optimized simultaneously. Therefore, by merging Eq.3 and Eq.6, the objective function is defined as;

$$\min_{\mathbf{W},\mathbf{H}} \left\{ \alpha \sum_{(i,j)\in O} \ell(1, f_{ij}(\mathbf{W})) + \beta \|\mathbf{X} - \mathbf{WH}\|^2 + \lambda''' \Omega''' \right\}, \quad (7)$$

where α and β are mixture weight parameters ($\alpha, \beta > 0$). In Eq.7, the matrix factorization process, i.e., the second term, provides additional constraints regarding the latent features for the supervised link prediction process, i.e., the first term.

If the loss function ℓ employs simple absolute loss, Eq.7 can be modified to;

$$\min_{\mathbf{W},\mathbf{H}} \left\{ \alpha \sum_{(i,j)\in O} (1 - f_{ij}(\mathbf{W})) + \beta \|\mathbf{X} - \mathbf{WH}\|^2 + \lambda''' \Omega''' \right\}. \quad (8)$$

Here, standard $\ell 2$ norm is typically used for regularization term Ω''';

$$\Omega''' = \|W\|^2 + \|H\|^2. \quad (9)$$

3.3 Parameter Estimation

To perform an empirical study, the steepest descent method was applied to optimize Eq.8 because of its ease of implementation. Table 1 lists the parameters used in the proposed method. Control parameters were determined experimentally based on extensive parameter search so that the parameters achieved the best optimization of the objective function (Eq.8).

After completion of optimization, high f_{ij} values indicate that user pair (i,j) are more likely to be friends.

4 EXPERIMENTAL EVALUATION

4.1 Collected Data

Fifty university students voluntarily contributed by becoming monitored subjects; proper informed consent was given. Every student knew at least one of the other participating students. Each student received a smartphone Xperia® on which an application to monitor user activities was installed. They were instructed to use it freely for about 6 months (February 2011 – September 2011). The collected logs include two

types of records; (1) application execution, and (2) access to Internet contents through the browser. Each log record included timestamp, terminal ID, and associated strings depending on its type, that are listed in Table 2.

Table 2: Collected logs.

Type	Log
Application exec.	Application package name
Internet access	Access URL

Fig.2 and Fig.3 represent statistical data for "Application execution" and "Internet access". Fig.2 shows the daily transition in the volume of collected log records. Fig.3 shows the daily transition in the number of unique users.

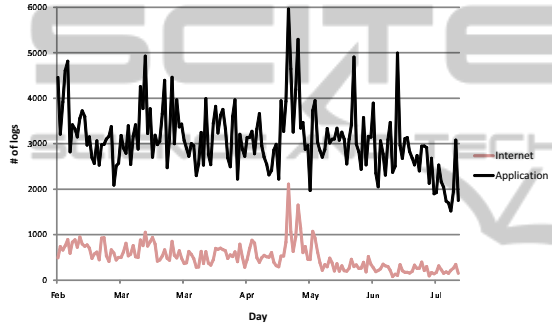

Figure 2: The volume of collected log records.

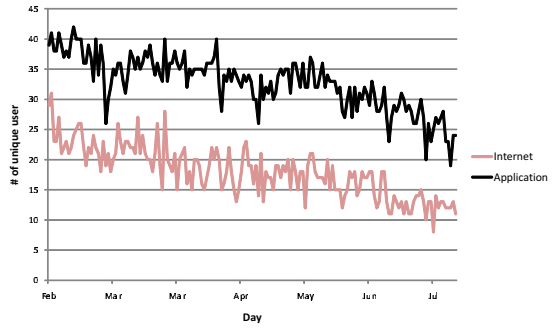

Figure 3: The number of unique users.

User–item matrix $\mathbf{X}$ was acquired from the collected logs. The column dimension of matrix $\mathbf{X}$ became $4,974$, which is the total number of application package names and URLs appearing in the logs. Each element of $\mathbf{X}$ indicates the frequency of item usage during the period, i.e. total number of times the corresponding application was launched or the corresponding URL was accessed.

In addition to the log records collected, an exit questionnaire was conducted in order to obtain the ground truth of friendships at the end of monitoring. The authors asked students whom he or she had called or received call(s) from during the experimental period regardless of the phone used. Fig.1 shows the results. According to the results, there were 157 pairs of students who made phone calls during the period. Here the authors treat these user pairs as the ground truth of friendship.

4.2 Analysis of Collected Data

First, the authors tested whether simple affinity measures such as cosine similarity were capable of identifying friendship from the raw behavioral data collected. Let $\mathbf{x}_i$ denote the vector that is the row of matrix $\mathbf{X}$ for user i; its length is $4,974$. Similarity measures shown below were calculated for each user pair (i, j).

- Cosine similarity (CS):

$$CS_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}. \tag{10}$$

- Dot product of binarized vectors (DB):

$$DB_{ij} = \mathbf{x'}_i \cdot \mathbf{x'}_j, \tag{11}$$

where,

$$\mathbf{x'}_i = (x'_{i1}, \cdots, x'_{iL}), \tag{12}$$

$$x'_{ik} = \begin{cases} 1 & \text{if } x_{ik} > 0, \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

- Dot product of log normalized vectors (DL):

$$DL_{ij} = \mathbf{x''}_i \cdot \mathbf{x''}_j, \tag{14}$$

where,

$$\mathbf{x''}_i = (x''_{i1}, \cdots, x''_{iL}), \tag{15}$$

$$x''_{ik} = \frac{\log(x_{ik}+1)}{\max_{i'}\{\log(x_{i'k}+1)\}}. \tag{16}$$

In addition, by borrowing the concept of LSA (Latent Semantic Analysis (Deerwester et al., 1990)), possible user features were calculated by SVD (Singular Values Decomposition) with designated rank number l. Cosine similarities were calculated based on the outcome. *CS_RD* denotes this result.

By sorting the user pairs according to calculated similarity (descending order), recall factors were obtained from the top-k, see Fig.4. The dashed diagonal line shows the baseline of random sampling. The figure reveals that the similarity measures offered no discrimination ability on either the raw data or the dimension-reduced data yielded by SVD. This implies that non-supervised approaches cannot be predicted friendships from such behavioral data.

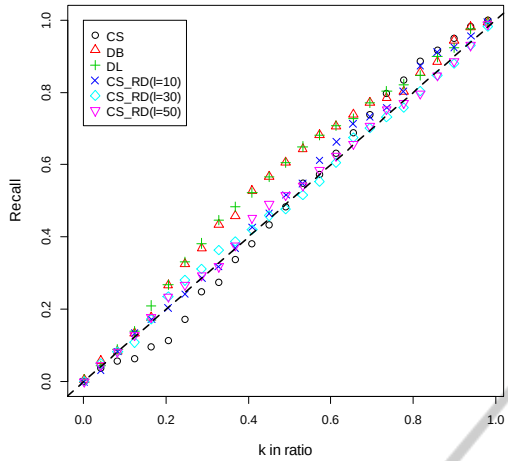

Figure 4: Top-*k* recall factors based on similarity of behavioral data.

4.3 Performance Evaluation

Performance of the proposed method was compared with that of the LFL model (Menon and Elkan, 2011), a state-of-the-art link prediction method. Menon et al. described in their paper a method to combine their LFL model with observable side-information such as node-information, as a possible extension to the model. However, their approach assumes that the side-information independently contributes to the prediction performance as a source of performance improvement. As we saw in the previous section, behavioral data is not expected to directly improve the prediction performance in our case. Thus here, the authors employed the LFL model without using behavioral data as a performance reference. The authors are not aware of any existing method which can utilize such behavioral data for link prediction.

Performance was evaluated by precision and recall on the top *k* predicted user pairs using the 3-fold cross validation approach. As for 3-fold cross validation, to simulate a realistic problem setting, only one third of the positive links were used as known links for training, that means all of the rest were treated as unknown for training. The other two thirds of the positive links and all of the non-friend pairs were included in the test set. For both methods, the best performing set of control parameters were used in the evaluation; the parameters were determined from the results of a preliminary parameter scan. Table 3 lists values of some control parameters used in the evaluation. To draw a Precision-Recall graph, *k* was swept from 10 to 50 in steps of 10. Results are shown in Fig.5. The proposed method demonstrated higher performance than the LFL model.

Table 3: Parameter settings for performance comparison.

Parameter	LFL	Proposed
Latent feature dimension L	50	100
Loss function	log	mae
Regularization weight λ	0.0	0.0

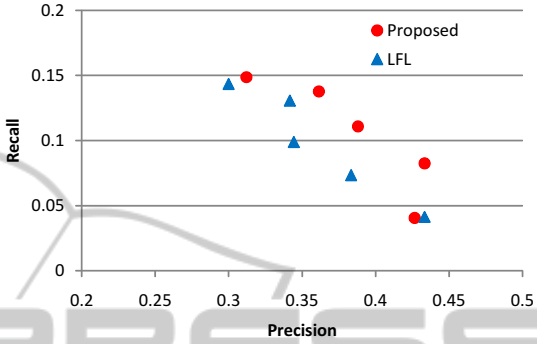

Figure 5: Precision-Recall graph.

5 CONCLUSIONS

This paper proposed a semi-supervised method for friendship prediction. The method utilizes matrix factorization for the user–item matrix to acquire latent user features. The latent user features are embedded into the process of supervised link prediction, so that the latent user features are optimized jointly in terms of known friendship links and observed user item interactions. An affinity measure specifically designed for latent user features on friendship prediction was also proposed. An extensive empirical evaluation was conducted using real data collected from university students. The results confirmed that the proposed method outperforms an existing state-of-the-art method.

Possible extensions and further study items include;

- Employing stochastic gradient descent method in the optimization process to improve its scalability
- Versatility should be studied by using different types of data
- Model extension to include other aspects such as FIP model

REFERENCES

Adamic, L. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278.

Fujimoto, H., Etoh, M., Kinno, A., and Akinaga, Y. (2011). Web user profiling on proxy logs and its evaluation in personalization. In *Proceedings of the 13th Asia-Pacific web conference on Web technologies and applications*, APWeb'11, pages 107–118, Berlin, Heidelberg. Springer-Verlag.

Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M., and Tsuda, K. (2009). Link propagation: A fast semi-supervised learning algorithm for link prediction. In Park, H., Parthasarathy, S., and Liu, H., editors, *SDM 2009*, pages 1099–1110, Society for Industrial and Applied Mathematics. Max-Planck-Gesellschaft, Philadelphia, PA, USA.

Keller, E. and Fay, B. (2009). The Role of Advertising in Word of Mouth. *Journal of Advertising Research*, 49(2):154–158.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(9):30–37.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A Statistical Mechanics and its Applications*, 390:1150–1170.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Menon, A. K. and Elkan, C. (2010). A log-linear model with latent features for dyadic prediction. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 364–373, Washington, DC, USA. IEEE Computer Society.

Menon, A. K. and Elkan, C. (2011). Link prediction via matrix factorization. In *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part II*, ECML PKDD'11, pages 437–452, Berlin, Heidelberg. Springer-Verlag.

Mirisaee, S. H., Noorzadeh, S., and Sami, A. (2010). Mining friendship from cell-phone switch data. In *Proceedings of the 3rd international conference on Human-Centric Computing*, HumanCom 2010, pages 1–5.

Quercia, D., Ellis, J., and Capra, L. (2010). Using mobile phones to nurture social networks. *IEEE Pervasive Computing*, 9(3):12–20.

Scellato, S., Noulas, A., and Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1046–1054, New York, NY, USA. ACM.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA. ACM.

Yang, S.-H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., and Zha, H. (2011). Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 537–546, New York, NY, USA. ACM.