Predicting User Replying Behavior on a Large Online Dating Site

Peng Xia
Department of Computer Science
University of Massachusetts Lowell
Lowell, MA 01854, USA

Hua Jiang, Xiaodong Wang
Product Division
Baihe.com
Beijing, China

Cindy Chen, Benyuan Liu
Department of Computer Science
University of Massachusetts Lowell
Lowell, MA 01854, USA

Abstract

Online dating sites have become popular platforms for people to look for potential romantic partners. Many online dating sites provide recommendations on compatible partners based on their proprietary matching algorithms. It is important that not only the recommended dates match the user's preference or criteria, but also the recommended users are interested in the user and likely to reciprocate when contacted. The goal of this paper is to predict whether an initial contact message from a user will be replied to by the receiver. The study is based on a large scale real-world dataset obtained from a major dating site in China with more than sixty million registered users. We formulate our reply prediction as a link prediction problem of social networks and approach it using a machine learning framework. The availability of a large amount of user profile information and the bipartite nature of the dating network present unique opportunities and challenges to the reply prediction problem. We extract user-based features from user profiles and graph-based features from the bipartite dating network, apply them in a variety of classification algorithms, and compare the utility of the features and performance of the classifiers. Our results show that the user-based and graph-based features result in similar performance, and can be used to effectively predict the reciprocal links. Only a small performance gain is achieved when both feature sets are used. Among the five classifiers we considered, random forests method outperforms the other four algorithms (naive Bayes, logistic regression, KNN, and SVM). Our methods and results can provide valuable guidelines to the design and performance of recommendation engine for online dating sites.

1 Introduction

Online dating sites have become popular platforms for people to look for potential romantic partners, offering an unprecedented level of access to potential dates that is otherwise not available through traditional means. According to a recent survey[1], 40 million single people (out of 54 million) in the US have signed up with various online dating sites such as Match.com, eHarmony, etc, and around 20% of currently committed romantic relationships began online, which is more than through any means other than meeting through friends.

An online dating site typically allows a user to create a profile that includes the user's photos, basic demographic information, behavior and interests (e.g., smoking, drinking, hobbies), self-description, and desired characteristics of an ideal partner. After creating a profile, a user can search for other people's profiles, browse other user profiles, and communicate with them. Some sites require a user to complete a questionnaire for evaluating the person's personality type and using it in the matching process.

Many online dating sites provide suggestions on compatible partners based on their proprietary matching algorithms. Unlike in many other recommendation systems where the goal is typically to predict a user's opinion towards given passive items (e.g., books, movies, etc) based on their evaluation of other items, when making recommendation of potential dates to a user on an online dating site, it is important that not only the recommended dates match the user's preference or criteria, but also the recommended users are interested in the user and likely to reciprocate (i.e., reply to contact message from the user) when contacted. Matching users with mutual interest in each other will result in better chances of interactions between them and improved user satisfaction on an online dating site. In this paper we study user reply behavior based on a large scale real-world dataset obtained from a collaboration with a large online dating site in China with a total number of 60 million registered users. In particular, *given a set of user profiles and their communication traces, we seek to accurately predict whether a user will reply to initial contact messages from other users.*

We formulate the above user reply prediction as a link prediction problem of social networks and approach it using a machine learning framework. Link prediction in social networks was first investigated in (Liben-Nowell and Kleinberg 2003) and has since been widely studied. Given a snapshot of a social network, the link prediction problem aims to infer which new interactions among its members are likely to occur in the near future. For our work we model the communications between users in our dataset as a bipartite directed network as the online dating site in our study is for heterosexual dating and only allows communications between male and female users. An edge (or link) in our

[1] http://statisticbrain.com/online-dating-statistics

constructed network represents an initial contact message or a reply to an initial contact message. Therefore, our user reply prediction problem seeks to accurately predict whether a reciprocal link will occur for an initial contact link between two users given the current snapshot of the constructed bipartite directed network.

The characteristics of the online dating network present unique opportunities and challenges to the reply prediction problem. First, there is a rich set of user attributes available in our dataset that can be used in the prediction models, including a user's age, gender, height, weight, education level, income level, house ownership, geographic location, occupation, interests/hobbies, photos, etc. In addition to these personal attributes, there are a variety of online dating specific information such as a users preference in potential dates (age range, height range, education level, income range, geography location, etc) and his/her dating and marriage plan (when to get married, whether to live with parents and have child after marriage, marriage ceremony style, etc). Second, network structure based features (e.g., node degree, graph distance, etc) have been widely used in previous link prediction studies. However, due to the bipartite nature of the online dating network, some network structure features used for homogeneous networks cannot be directly applied. For example, the common neighbors and Jaccards coefficient are not directly applicable since two males with common neighbors (female) will never communicate with each other for online dating. To this end, we need to derive meaningful features appropriate for the specific bipartite directed dating network for our study.

In this paper we extract user-based features from user profiles and graph-based features from the constructed bipartite dating network, apply them in a variety of classification algorithms, and compare the utility of these features and performance of the classifiers in prediction. We adopt the notion of collaborative filtering and introduce features that capture the similarity of user interest and attractiveness. We also revise the definition of some traditional network structure features (e.g., common neighbors, Jaccard's coefficient, etc) to fit the bipartite directed dating network in our study. Our results show that the user-based and graph-based features result in similar performance, and can be used to effectively predict the reciprocal links. Only a small performance gain is achieved when both feature sets are used. Among the five classifiers we considered, Random Forests method outperforms the other four algorithms (Naive Bayes, Logistic Regression, KNN, and SVM). Our methods and results can provide valuable guidelines to the design and performance of recommendation engine for online dating sites.

The rest of the paper is structured as follows. Section 2 describes the related work on the recommendation for online dating and the link prediction problem. The dataset used in our study is described in Section 3. Section 4 presents the problem formulation, data preprocessing, and features derived from user profiles and the bipartite dating graph. We then apply a variety of classification algorithm on the extracted features and present the experiment results in Section 5. Finally, we conclude our paper in Section 6.

2 Related Work

There has been recently a few studies on the recommendation of potential romantic dates for online dating users (Cai et al. 2010; Pizzato et al. 2010; Zhao et al. 2014). In particular, (Cai et al. 2010) uses collaborative filtering algorithms based on the similarity of user's interest and attractiveness, while both (Pizzato et al. 2010) and (Zhao et al. 2014) consider the reciprocal interactions as an important factor in recommendation. In (Pizzato et al. 2010), a list of reciprocal recommendations is provided with significant improvements over recommendation algorithms that do not take reciprocal interactions into account. In (Zhao et al. 2014), a hybrid collaborative filtering based algorithm taking reciprocal links into consideration is proposed and shown to have good performance in recommending both initial and reciprocal contacts. The work of (Li and Li 2012) considers both local utility (users mutual preference) and global utility (overall bipartite network), and proposes a generalized framework for reciprocal recommendation in online dating sites. The authors in (Tu et al. 2014) propose a two-side matching framework for online dating recommendations and design an Latent Dirichlet Allocation (LDA) model to learn the user preferences from the observed user messaging behavior and user profile features.

Most of the previous studies on link prediction problem fall into one of the following two categories: unsupervised learning algorithms and supervised learning algorithms. Previous unsupervised learning methods for link prediction mainly focus on assigning appropriate scores to the potential link (Liben-Nowell and Kleinberg 2003). Simple unsupervised predictors include the number of common neighbors between two nodes, Jaccard's coefficient (fraction of common neighbors between two nodes over the total number of neighbors), and the preferential attachment (product of the degrees of the two nodes).

Early work on applying supervised learning algorithms to link prediction includes (Hasan et al. 2006), which aims to predict the co-author relationship in datasets including BIOBASE[2] and DBLP[3] using graph topological features. Further work has extended the link prediction problem to online social networks such as Facebook, Twitter, Epinions and Slashdot (Leskovec, Huttenlocher, and Kleinberg 2010; Fire et al. 2011; Sadilek, Kautz, and P.Bigham 2012; Dong et al. 2012). The authors of (Lichtenwalter, Lussier, and Chawla 2010) argue that supervised learning algorithm is a more suitable approach for the link prediction problem. In (Wu, Raschid, and Rand 2011), the authors study the link prediction problem in Blogosphere and show that a method combining both network and content properties of the blog yields better results than those considering just a single property.

A few recent studies extend the traditional link prediction problem to consider more specific aspects of link property. In particular, (Kahanda and Neville 2009) aims to predict the strength of a link, while (Guha et al. 2004) and (Leskovec, Huttenlocher, and Kleinberg 2010) study the link sign pre-

[2] http://www.elsevier.com/elsevier-products/biobase

[3] http://dblp.uni-trier.de

diction in Epinions, Sladshdot, and Wikipedia, which interpret the trust/distrust, friend/foe and vote relations as positive and negative links. In (Hopcroft, Lou, and Tang 2011), the authors study the prediction of the follow back interaction (reciprocal link) in Twitter using TriFG model.

3 Dataset Description

The dataset used in our study is obtained through a collaboration with baihe.com, one of the major online dating sites in China. Our dataset includes the profile information of 200,000 users uniformly sampled from users registered in November of 2011. Of the 200,000 sampled users, 139,482 are males and 60,518 are females, constituting 69.7% and 30.3% of the total number of sampled users respectively. For each user, we have his/her message sending and receiving traces (who contacted whom at what time) in the online dating site and the profile information of the users that he or she has communicated with from the date that the account was created until the end of January 2012. Note that the site is for heterosexual dating and only allows communications between users of opposite sex.

A user's profile provides a variety of information including user's gender, age, current location (city and province), home town location, height, weight, body type, blood type, occupation, income range, education level, religion, astrological sign, marriage and children status, photos, home ownership, car ownership, interests, smoking and drinking behavior, self introduction essay, among others. Each user also provides his/her preferences for potential romantic partners in terms of age, location, height, education level, income range, marriage and children status.

After a user creates an account on the online dating site, he/she can search for potential dates based on information within the profiles provided by the other users including user location, age, etc. Once a potential date has been discovered, the user then sends a message to him/her, which may or may not be replied by the recipient. In this paper we focus on the prediction of whether a user will reply to initial messages sent by other users. Subsequent interactions between the same pair of users do not represent a new sender-receiver pair and can not be used as the only indicator for continuing relationship as users may choose to go off-line from the site and communicate via other channels (e.g., email, phone or meet in person).

Since we only have eight full weeks' worth of online dating interaction records for our sample users, we will consider the activities of each user during the first eight weeks of his/her membership. Table 1 describes the characteristics of the dataset. More detailed description and analysis of the dataset can be found in our recent work (Xia et al. 2013; 2014).

Table 1: Dataset Description

Type	Initial contact links	Reciprocal links (Reply rate)
Male to Sample Female	1,586,059	150,917 (9.5%)
Female to Sample Male	328,645	58,946 (17.9%)

4 Predicting User Reply Behavior

We now consider the problem of predicting whether a user will reply to the initial contact message from another user. We first present the problem formulation and model it as a link prediction problem of a bipartite dating network. We then describe the features we extract and prediction algorithms to be considered in our study.

4.1 Problem Formulation

Here we give a formal description of our problem: Based on the user profiles and their communication traces until time t, we construct a bipartite directed graph $G_t(V_t, E_t), V_t = M_t \cup F_t$ as follows. M_t and F_t represent two disjoint sets of male and female users who have sent or received at least one message until time t, respectively. A directed edge (or link) (u, v) exists if user u has sent a message to user v before time t. In this paper we will use the terms edge and link interchangeably. An edge may correspond to an initial contact message or a reply to an initial contact message. Since the online dating site in our study is for heterosexual dating and only allows communications between users of opposite sex (males and females), all of the edges are between vertexes in M_t and F_t, resulting in a bipartite graph as illustrated in Figure 1.

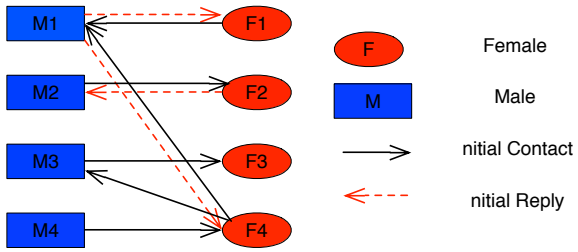


Figure 1: Online dating interactions modeled as a bipartite directed graph.

For an initial contact edge (u, v), where v is from our sampled users, we want to predict whether there is a reciprocal edge (v, u) (reply to the initial contact message) at a future time $t' > t$, i.e., user v replies to the initial contact message from user u. Since we observe significant difference between the male and females online dating behaviors (Xia et al. 2013; 2014), in this paper we will consider the prediction models for the male and female reply prediction problem separately.

4.2 Feature Extraction

We now describe the features used for our prediction models. The features we extract include both individual user attributes and topological features derived from the bipartite dating network. As mentioned in Section 2, some topological measures used in previous studies of link prediction do not directly apply to our bipartite dating network, and in these cases we adapt the measures accordingly to make them appropriate for our specific network model.

User Profile. A typical profile of a user includes the user's age, height, weight, geographic location (city), education level, income level, marriage status (never married, divorced, widowed), house ownership, etc. A user may also post his/her photos on the profile page. These user attributes provide the basic information and give a first impression of a user when people browse the user's profile. The user attributes can be divided into two categories, namely, numeric attributes and discrete attributes. Specifically, numeric features include a users age, height, weight, number of photos, etc. We also computed the geographic distance between two users as a numeric feature. Categorical features include a users education level, income level, home ownership, and marriage status, etc. In total, there are 47 features in this category.

User Preference. On the online dating site, a user can specify the following eight attributes that he/she is looking for in a date, including the age range, geographic location, height range, marriage status (never married, divorced, widowed), education level, income range, house ownership, children status (no children, children living with user, children not living with user). Based on a users stated dating preference, for a given link (u, v), we calculate the user preference match vector $\text{fit}_{attribute}(u, v)$ that describes whether each attribute of user u matches the dating preference of user v.

User Life Style. On the online dating site, a user can describe his/her family situation and life style including whether the user's parents still live, whether the user is the only child in his/her family, the user's housework and cooking habits, and the user's smoking and drinking habits. There are 14 features in this categroy.

User Future Plan. A user's future plan describes the dating and wedding style that the user prefers, which includes when he/she wants to get married, whether the couple will live with their parents after marriage, whether they want children, and the preferred dating and wedding style. In total, there are 12 features in this category.

All of these above features correspond to a user's individual and personal information. In the following we describe the network structure features derived from the constructed bipartite directed dating graph.

Topological Features. Topological features extracted from our constructed bipartite directed dating graph are good indicators representing users' active and popularity levels. Specifically, the in-degree of a node u corresponds to the number of messages received by user u, representing the popularity of user u. The out-degree of a node u corresponds to the number of messages sent out by user u, representing his/her active level. We define $n_{send}(u)$ as the number of initial contact messages sent by u, $n_{receive}(u)$ as the number of initial contact messages received by u, $n_{reply}(u)$ as the number of initial contact messages that are replied to by u, and $n_{replied}(u)$ as the number of initial replies received by u. Clearly, $n_{send}(u)$ and $n_{reply}(u)$ reflect how actively user u is looking for a potential date, while $n_{receive}(u)$ and $n_{replied}(u)$ reflect how popular user u is in the whole network.

When computing topological features, many previous studies do not take time factor into consideration, e.g., (Leskovec, Huttenlocher, and Kleinberg 2010; Fire et al. 2011; Hasan et al. 2006). A major drawback of this approach is that the prediction of current events depends on features computed from future events, which is not suitable for online prediction. Another drawback is that users who have been in the network for a longer time may have unfair advantages over those relatively newer users. A user who sends out a certain number of messages in one year should not be considered as active as a user who sends out the same number of messages in one month. To avoid these drawbacks, we take the time factor into account when extracting topological features from the dating network. Below we give formal definitions of the topological features used in our study.

Given the snapshot of the network at time t, we use superscript Δt to denote the corresponding measures for time interval $(t - \Delta t, t]$ and define the following topological features:

$$\begin{aligned}
\text{send-rate}^t(u) &= \frac{n^{\Delta t}_{send}(u)}{\Delta t} \\
\text{receive-rate}^t(u) &= \frac{n^{\Delta t}_{receive}(u)}{\Delta t} \\
\text{reply-rate}^t(u) &= \frac{n^{\Delta t}_{reply}(u)}{\Delta t} \\
\text{replied-rate}^t(u) &= \frac{n^{\Delta t}_{replied}(u)}{\Delta t}
\end{aligned} \tag{1}$$

The online dating site we study provides the *follow* functionality similar to that in Twitter. A user can follow other users of his/her interest. The number of people followed by a user and following a user also represent the active level and popularity of the user. Let $n^{\Delta t}_{follow}(u)$ and $n^{\Delta t}_{followed}(u)$ denote the number of people that user u follows and number of people that u is followed by during time interval $(t - \Delta t, t]$, respectively. We build the following features:

$$\begin{aligned}
\text{follow-rate}^t(u) &= \frac{n^{\Delta t}_{follow}(u)}{\Delta t} \\
\text{followed-rate}^t(u) &= \frac{n^{\Delta t}_{followed}(u)}{\Delta t}
\end{aligned} \tag{2}$$

where follow-rate$^t(u)$ measures user $u's$ active level, while followed-rate$^t(u)$ measures user $u's$ popularity.

For this category, there are 12 features representing a user's activity and popularity level.

Similarity Feature. Based on the concept of collaborative filtering, if two users send initial contact messages to the same person, we say that they share similar *interest*, and if two users receive initial contact messages from the same person, we say that they have similar *attractiveness*. Denote $E^{\Delta t}_t$ as the set of edges in the graph created during time interval $(t - \Delta t, t]$, we represent the set of users who share similar *interest* with u as $S_i(u) = \left\{v | \exists w, (u, w) \in E^{\Delta t}_t, and\, (v, w) \in E^{\Delta t}_t\right\}$, and the set of

users who have similar attractiveness with u as $S_a(u) = \{v|\exists w,(w,u) \in E_t^{\Delta t}, and\,(w,v) \in E_t^{\Delta t}\}$. Clearly the similarity relationship defined above is symmetric, i.e., if $u \in S_{i/a}(v)$, we also have $v \in S_{i/a}(u)$. We say that user u is similar to user v if they have similar *interest* or *attractiveness*.

For a user v in the similar user set of user u, i.e., $v \in S_i(u)$ or $v \in S_a(u)$, we define the similarity scores between u and v as follows:

$$\begin{aligned} s_i^{\Delta t}(u,v) = & \frac{|\{w|(u,w) \in E_t^{\Delta t}\ and\ (v,w) \in E_t^{\Delta t}\}|}{|\{w|(u,w) \in E_t^{\Delta t}\ or\ (v,w) \in E_t^{\Delta t}\}|} \\ s_a^{\Delta t}(u,v) = & \frac{|\{w|(w,u) \in E_t^{\Delta t}\ and\ (w,v) \in E_t^{\Delta t}\}|}{|\{w|(w,u) \in E_t^{\Delta t}\ or\ (w,v) \in E_t^{\Delta t}\}|} \end{aligned} \tag{3}$$

where $s_i^{\Delta t}(u,v)$ is the fraction of users who receive messages from both u and v among all users who receive messages from either u or v, representing the interest similarity between u and v. Similarly, $s_a^{\Delta t}(u,v)$ measures the attractiveness similarity between u and v.

Consider two users u and v of opposite genders, communications between users who are similar to them suggest that they may also be interested in each other. For example, if a user with similar attractiveness to sender u receives a lot of messages from users with similar interest to receiver v, it is likely that v is also interested in and may reply to u when contacted. Specifically, in this paper we consider the following two scanarios:

- Interactions between users similar to u and receiver v as illustrated in Figure 2(a).

$$\begin{aligned} GG_{i,i}(u,v) = & \{(x,y)|x \in S_i(u), y \in S_i(v)\} \\ GG_{i,a}(u,v) = & \{(x,y)|x \in S_i(u), y \in S_a(v)\} \\ GG_{a,i}(u,v) = & \{(x,y)|x \in S_a(u), y \in S_i(v)\} \\ GG_{a,a}(u,v) = & \{(x,y)|x \in S_a(u), y \in S_a(v)\} \end{aligned} \tag{4}$$

- Interactions between sender u and users similar to receiver v and interactions between receiver v and users similar to sender u as illustrated in Figure 2(b).

$$\begin{aligned} UG_i(u,v) = & \{(u,y)|y \in S_i(v)\} \\ UG_a(u,v) = & \{(u,y)|y \in S_a(v)\} \\ GU_i(u,v) = & \{(x,v)|x \in S_i(u)\} \\ GU_a(u,v) = & \{(x,v)|x \in S_a(u)\} \end{aligned} \tag{5}$$

For each type of interactions described in Figures 2(a) and 2(b), we sum them up with the similarity scores as defined in equation (3) and build the following features that measure the aggregate similarity between the sender u and receiver v.

$$\begin{aligned} WGG_{i/a,i/a}(u,v) = & \sum_{(x,y) \in GG_{i/a,i/a}(u,v)} s_{i/a}^{\Delta t}(x,u) * s_{i/a}^{\Delta t}(y,v) \\ WUG_{i/a}(u,v) = & \sum_{y \in UG_{i/a}(u,v)} s_{i/a}^{\Delta t}(y,v) \\ WGU_{i/a}(u,v) = & \sum_{x \in GU_{i/a}(u,v)} s_{i/a}^{\Delta t}(x,u) \end{aligned} \tag{6}$$

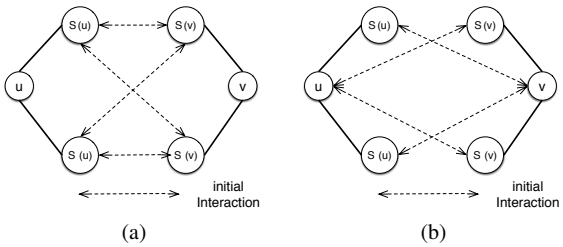


Figure 2: (a) Interactions between users similar to two users u and v of opposite genders. (b) Interactions between users similar to two users u and v of opposite genders.

Note for each item in equations (4), (5) and (6), there are four different types of interactions depending on the direction of the messages relative to node u and v. In later analysis, we will use different superscripts to represent these four types of interactions: s for $send$, r for $receive$, re for $reply$ and rd for $replied$. Here we do not include the superscripts for simplicity. In total, there are 32 features for this category.

Neighbor Features. In conventional online social networks such as Facebook and Twitter, the more common neighbors two nodes share, the higher chances these two nodes will be connected. Thus common neighbor is an important feature for the link prediction problem for these social networks. However, in our bipartite dating network, since links only exist between nodes of different genders, nodes with common neighbors must be of the same gender, which is not consistent with our goal of predicting the reciprocal links between users of different genders. Therefore, traditional neighbor-based features such as common neighbots and Jaccard's coefficient do not directly apply to our reply prediction problem. In the following, we adapt these neighbor-based features to fit our bipartite dating graph.

We first define the set of neighbors for user u as

$$\begin{aligned} \Gamma_{in}(u) = & \{v|(v,u) \in E_t^{\Delta t}\} \\ \Gamma_{out}(u) = & \{v|(u,v) \in E_t^{\Delta t}\} \end{aligned} \tag{7}$$

where $\Gamma_{in}(u)$ is the set of neighbors who send initial contact messages to u, while $\Gamma_{out}(u)$ is the set the neighbors who received initial contact messages from u. We now introduce several features appropriate for our bipartite dating graph.

- *Common Neighbors*: We revise the traditional common neighbor definition $|\Gamma_{in/out}(u) \cap \Gamma_{in/out}(v)|$ to $|\Gamma_{in/out}(u) \cap S_{i/a}(v)|$ and $|\Gamma_{in/out}(v) \cap S_{i/a}(u)|$, which measures the overlap between neighbors of u and users similar to v, and vice versa. Figure 3 depicts two example of the revised common neighbors features.
- *Jaccard's Coefficient*: Jaccard's Coefficient has been widely used in previous link prediction studies, measuring the similarity between two sample sets. Similar to the revision of common neighbor feature as described above, we customize the Jaccard's coefficient as $\frac{|\Gamma_{in/out}(u) \cap S_{i/a}(v)|}{|\Gamma_{in/out}(u) \cup S_{i/a}(v)|}$ and $\frac{|\Gamma_{in/out}(v) \cap S_{i/a}(u)|}{|\Gamma_{in/out}(v) \cup S_{i/a}(u)|}$.

Table 2: Performance of classifiers with different feature sets to predict reciprocal links from females to males

Algorithms	Measure	User-based Feature				Graph-based Feature			All
		Profile	Preference	Life Style	Future Plan	Topological	Similarity	Neighbor	
Naive Bayes	Precision	0.621	0.557	0.582	0.581	0.523	0.612	0.576	0.656
	Recall	0.617	0.557	0.574	0.574	0.501	0.566	0.547	0.656
	F-measure	0.614	0.557	0.563	0.565	0.348	0.516	0.500	0.656
	AUC	0.667	0.576	0.602	0.606	0.605	0.589	0.571	0.708
Logistic Regression	Precision	0.642	0.560	0.582	0.589	0.700	0.654	0.583	0.722
	Recall	0.642	0.560	0.582	0.589	0.666	0.635	0.566	0.722
	F-measure	0.642	0.560	0.581	0.589	0.651	0.623	0.543	0.722
	AUC	0.694	0.584	0.614	0.622	0.747	0.698	0.588	0.796
Random Forests	Precision	0.727	0.564	0.624	0.614	0.730	0.675	0.588	0.762
	Recall	0.727	0.563	0.624	0.613	0.730	0.668	0.587	0.762
	F-measure	0.727	0.560	0.624	0.612	0.730	0.649	0.586	0.762
	AUC	0.801	0.585	0.665	0.647	0.802	0.709	0.617	0.841
KNN	Precision	0.735	0.565	0.620	0.614	0.707	0.638	0.557	0.744
	Recall	0.729	0.563	0.619	0.614	0.707	0.634	0.557	0.738
	F-measure	0.728	0.560	0.618	0.614	0.706	0.631	0.557	0.736
	AUC	0.787	0.585	0.666	0.658	0.777	0.686	0.573	0.819
SVM	Precision	0.505	0.550	0.569	0.575	0.681	0.666	0.563	0.580
	Recall	0.505	0.549	0.568	0.574	0.675	0.630	0.543	0.574
	F-measure	0.504	0.548	0.567	0.573	0.672	0.610	0.505	0.566
	AUC	0.505	0.549	0.568	0.574	0.718	0.675	0.543	0.574

- *Preferential Attachment*: This feature has been previously used in link prediction problem (Liben-Nowell and Kleinberg 2003) with the premise that the probability of a new link involves a node is proportional to the degree of the node. We use the following features to capture the joint effect of the sender and receiver's active level and popularity, $\left|\Gamma_{in/out}(u)\right| * \left|\Gamma_{in/out}(v)\right|$.

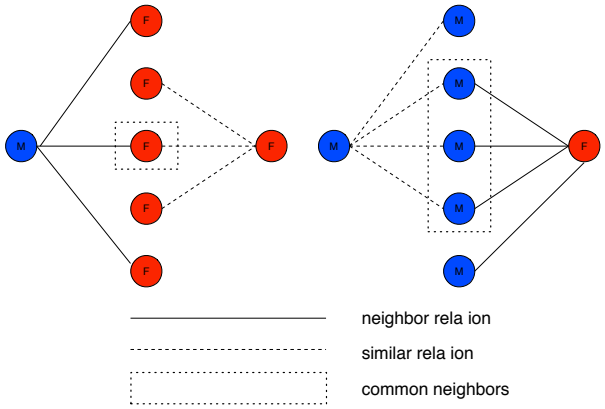


Figure 3: Revised commmon neighbors.

In total, there are 20 features in the Neighbor Feature category.

In summary, we build the following feature sets from user profile information and the constructed bipartite dating graph.

- *User Profile* contains the basic profile information of sender and receiver (including age, height, weight, education level, etc, as described before).
- *User Preference* contains the information about whether a sender's attributes match the stated preference of the receiver.
- *User Life Style* contains the life style information of sender and receiver.
- *User Future Plan* contains the future plans of sender and receiver.
- *Topological Feature* contains the average number of messages associated with sender or receivers. We also include the average number of users following/followed by sender and receiver.
- *Similarity Feature* contains the interactions between similar groups of sender and receiver as well as the interactions between sender/receiver and the corresponding similar groups.
- *Neighbor Feature* contains the revised versions of common neighbors, Jaccard's coefficient, and preferential attachment.

Based on how the aforementioned features are extracted, We further group these features into two broad categories:

- *User-based features* include the first four feature sets (user profile, user preference, user life style, user future plan), representing various information of individual users.
- *Graph-based features* contain the last three feature sets (topological feature, similarity feature, and neighbor feature), as they are derived from the bipartite directed dating network.

In Section 5 we will compare the performance resulting from different feature sets.

4.3 Classification Algorithms

There are a variety of classification algorithms for supervised learning. Some of these algorithms are more suitable than others for a specific dataset or problem instance. In this paper we experiment with five different classification

Table 3: Performance of classifiers with different feature sets to predict reciprocal links from males to females

Algorithm	Measure	User-based Feature				Graph-based Feature			All
		Profile	Preference	Life Style	Future Plan	Topological	Similarity	Neighbor	
Naive Bayes	Precision	0.582	0.547	0.597	0.598	0.665	0.662	0.666	0.678
	Recall	0.582	0.547	0.597	0.598	0.529	0.538	0.536	0.556
	F-measure	0.582	0.546	0.597	0.598	0.406	0.428	0.424	0.465
	AUC	0.623	0.556	0.628	0.630	0.629	0.600	0.604	0.675
Logistic Regression	Precision	0.603	0.547	0.600	0.602	0.675	0.641	0.595	0.678
	Recall	0.603	0.547	0.599	0.601	0.665	0.597	0.584	0.678
	F-measure	0.603	0.547	0.598	0.600	0.660	0.564	0.571	0 .678
	AUC	0.651	0.558	0.633	0.633	0.733	0.674	0.631	0.743
Random Forests	Precision	0.691	0.545	0.622	0.610	0.707	0.631	0.605	0.743
	Recall	0.690	0.545	0.622	0.610	0.706	0.626	0.604	0.74
	F-measure	0.690	0.544	0.621	0.610	0.706	0.623	0.603	0.739
	AUC	0.765	0.557	0.653	0.643	0.780	0.691	0.652	0.816
KNN	Precision	0.699	0.545	0.617	0.613	0.683	0.607	0.586	0.714
	Recall	0.693	0.545	0.616	0.612	0.681	0.605	0.586	0.708
	F-measure	0.691	0.544	0.615	0.611	0.680	0.603	0.586	0.705
	AUC	0.711	0.557	0.661	0.653	0.751	0.658	0.627	0.783
SVM	Precision	0.502	0.546	0.596	0.600	0.641	0.614	0.531	0.543
	Recall	0.502	0.546	0.595	0.598	0.641	0.594	0.530	0.543
	F-measure	0.493	0.546	0.594	0.595	0.641	0.576	0.525	0.543
	AUC	0.502	0.546	0.595	0.598	0.641	0.594	0.530	0.543

algorithms on our datasets and compare the performance of these classifiers. The algorithms we choose include Naive Bayes (NB), Logistic Regression (LR), Random Forests (RF), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). We use Weka (Hall et al. 2009) for our experiments since it implements many machine learning algorithms for supervised learning and has become a widely used tool for link prediction problem (Lichtenwalter, Lussier, and Chawla 2010; Yang et al. 2013). For SVM, we use the LibLinear implemented by (Fan et al. 2008). Although LibSVM (Chih-Chung and Chih-Jen 2011) with radial basis function kernel may yield better results than LibLinear which uses linear kernel, it takes a long time for training (more than one month for 10-fold cross validation on a powerful PC with Intel Core CPU i7-3770K and 32GB memory). For each algorithm, we conduct a number of experiments to find the optimal parameters when applicable. For KNN, K is set 10 to get the best performance, and for Random Forests, the number of trees is set to 50. We report the experimental results of each algorithm with 10-fold cross validation.

5 Experiments

5.1 Datasets

Using the same method as in (Leskovec, Huttenlocher, and Kleinberg 2010), i.e., for each reciprocal link, we randomly select a link from the non-reciprocal class and build the dataset with equal number of reciprocal and non-reciprocal links, as shown in Table 4.

Table 4: Finalized Data Set

Type	Reciprocal Links	Non-reciprocal Links
Male to Sample Female	150,917	150,917
Female to Sample Male	58,946	58,946

5.2 Evaluation Method

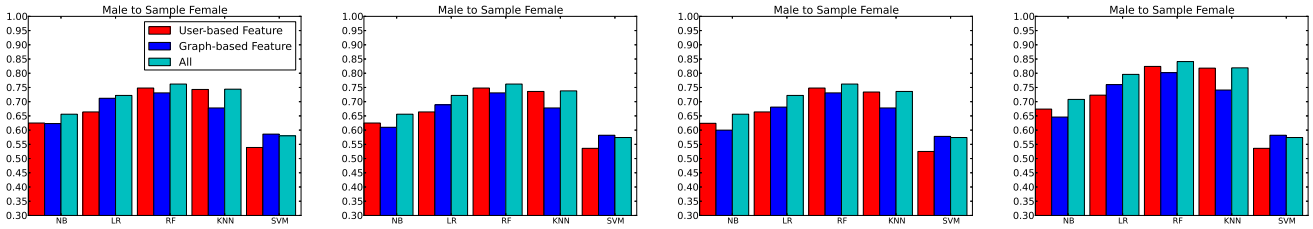
Precision, recall, F-measure and accuracy are commonly used evaluation metrics for binary classification problem. In addition to these fixed-threshold evaluation metrics, threshold curves such as Receiver Operating Characteristic (ROC) Curve and Precision-Recall curve have been proposed and used in recent research on link prediction (Lichtenwalter, Lussier, and Chawla 2010; Cai et al. 2012; Yang et al. 2013; Fire et al. 2011). Specifically, ROC curve illustrates the performance of a binary classification algorithm as its discrimination threshold is varied, and describes the true positive rate over the false positive rate at various threshold. AUC is the area under the ROC curve, which is equivalent to the probability of randomly selecting a positive instance over randomly selecting a negative instance.

For each feature set, we evaluate the algorithms discussed above using the 10-fold cross validation method, and report the result in precision, recall, F-measure, and AUC.

5.3 Feature Category Comparison

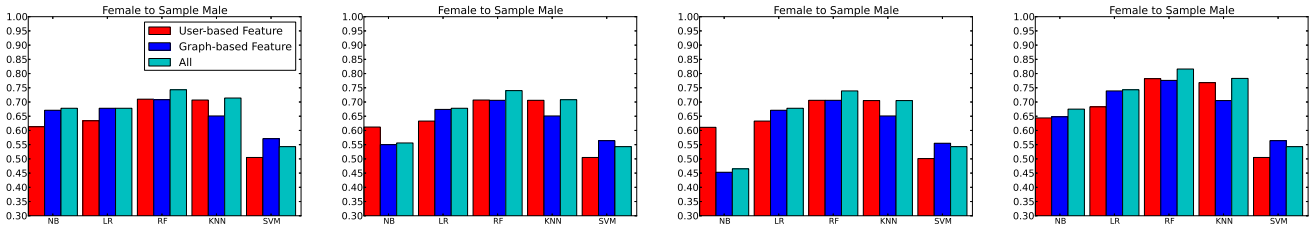
Our first set of experiments evaluates the utility of each feature set applied to different classifiers. The feature sets include *user profile*, *user preference*, *user life style*, *user future plan*, *topological feature*, *similarity feature*, *neighbor feature*. Table 2 and 3 describe the average performance (precision, recall, F-measure, and AUC) of difference classifiers on each feature set.

From classification algorithm perspective, the Random Forests algorithm outperforms all other classifiers, while the Naive Bayes and SVM algorithms are the bottom two performers. From feature perspective, the user profile features (age, income, education level, height, weight, location, photo count, etc.) result in the best performance (with AUC > 0.76 for female to sample male, and AUC > 0.8 for male to sample female under Random Forest model) among all user-based features, while the topological features (send-



(a) Precision (b) Recall (c) F-measure (d) AUC

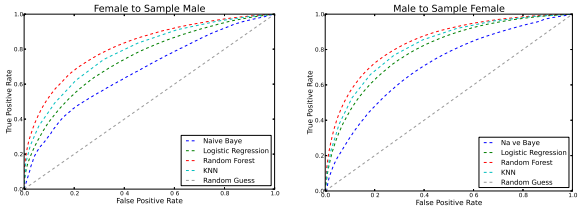
Figure 4: Performance of classifiers with user-based, graph-based, and all features to predict reciprocal links from females to males.



(a) Precision (b) Recall (c) F-measure (d) AUC

Figure 5: Performance of classifiers with user-based, graph-based, and all features to predict reciprocal links from males to females.

rate, receive-rate, reply-rate, replied-rate) yield the best performance (with AUC $>$ 0.78 for female to sample male, and AUC $>$ 0.8 for male to sample female to male under the Random Forest model) among all graph-based features. These results indicate that these two feature sets are most influential among all feature sets. It is interesting to note that the performance of user preference feature is significantly worse than other feature sets with AUC in the range of 0.5 - 0.6. This is consistent from our previous observation that a user's actual dating behavior deviates significantly from his/her stated preference (Xia et al. 2014).



(a) ROC Curve (b) ROC Curve

Figure 6: ROC Curve for each algorithm performed on all features

5.4 Overall Classification

After checking the performance of each single feature set, we grouped the features into user-based, graph-based and all of the features, and report the results of these three feature sets in our second set of experiments. The results obtained from 10 fold cross-validation are shown in Figures 4 and 5.

The SVM algorithm still remains the worst performer (with AUC $<$ 0.6) while the Random Forests algorithm yields the best results (with AUC $>$ 0.77). We observe that user-based features result in similar (for most classifiers slightly better) performance to graph-based features. There is only a small gain in performance when we use both user-based and graph-based features. The best performance (with AUC of 81.6% for female to sample male, and 84.1% for male to sample female) is achieved when we apply Random Forests algorithm on all of the features.

Figure 6 plots the ROC curves when all features are included in the prediction, Note that LibLinear gives no probability output, so we exclude it from the curves. The Naive Bayes and SVM algorithms achieve less than 70% AUC, while Logistic Regression, and KNN achieve around or more than 75% AUC. Among all these algorithms, the Random Forests algorithm achieve the best performance with more than 80% AUC, and thus can be considered as a good predictor for our problem.

5.5 Feature Ranking

In our study there is a total number of 145 extracted features out of which 81 are user-based features while 64 are graph-based features. To evaluate the relative importance of these features we perform feature selection to calculate each feature's ability to distinguish between positive and negative examples using Information Gain (IG) and Chi-Square (χ^2) statistic. We consider the user-based features and graph-based features separately.

Table 5 lists the top 15 user-based features. Note that rankings obtained using the information gain method and

Table 5: User-based Feature Rankings

Male to Sample Female				Female to Sample Male			
Feature Category	Feature Description	IG	χ^2	Feature Category	Feature Description	IG	χ^2
user profile	age of sender	1	1	life style	rank in siblings of receiver	1	1
user profile	income of sender	2	2	life style	parents status of receiver	2	2
user profile	age of receiver	3	3	future plan	receiver's plan about children	3	3
future plan	receiver's plan about children	4	4	future plan	receiver's plan about marriage	4	4
user profile	house status of sender	5	5	future plan	desirable qualities valued by potential partner	5	5
life style	parents status of sender	6	6	future plan	whether receiver wants to live with sender's parents	6	6
life style	receiver's life style	7	8	life style	receiver's life style	7	7
future plan	receiver's plan about wedding	8	7	life style	receiver's drinking habit	8	8
future plan	receiver's plan about marriage	9	9	future plan	receiver's plan about wedding	9	9
life style	receiver's attitude about housework	10	10	life style	receiver's attitude about housework	10	10
life style	receiver's cooking style	11	11	future plan	receiver's dating plan	11	11
user profile	receiver's marriage status	12	12	user profile	receiver's smoking habit	12	12
life style	rank in siblings of receiver	13	13	life style	receiver's cooking style	13	13
user profile	receiver's smoking habit	14	14	user profile	sender's love type	14	14
future plan	whether receiver wants to live with sender's parents	15	15	user profile	height of sender	15	15

Chi-Square statistic are very similar to each other. For male to sample female messages, the top 15 features influencing whether a female will reply to an initial contact message include features from the user profile, life style and future plan categories. Note that these top 15 features do not include any feature in the user preference category, which is consistent with our previous observation that features in this category have the worst prediction power among all features. On the other hand, when a male decides whether to reply to a female (female to sample male messages), life style and future plan features play a more important role than user profile features. For example, while age, income, and house situation of a sender are important factors for female receivers, they are not among the top 15 features for male receivers.

Table 6 lists the top 15 graph-based features. Again rankings obtained using the information gain method and Chi-Square statistic are very similar to each other. Note that the receiver's reply rate feature ranks first for both males and females. This may be intuitive since a receiver's reply rate is a direct measure of how likely the receiver will reply to a message. For both male and female receivers, the top 15 graph-based features include a mixture of topological, similarity and neighbor features which all play an important role in predicting whether a receiver will reply to an initial contact message. For female to sample male messages, topological features are higher ranked than other features.

We also conduct experiments where only the top 15 and top 30 user-based and graph-based features are supplied to the prediction models. For male to sample female messages, the top user-based features yield a slightly better performance than graph-based features, with the best result of AUC = 74.2% for top 15 features and AUC = 79.5% for top 30 features compared to 84.1% AUC when all features are used. For female to sample male messages, the top network-based features yield a slightly better performance than user-based features, with the best result of AUC = 73.5% for top 15 features and AUC = 76.3% for top 30 features compared to 81.6% AUC when all features are used.

6 Conclusion

Matching users with mutual interest in each other is an important task for online dating sites. In this paper we study the reply prediction problem, i.e., whether a user is likely to reciprocate (i.e., reply to the initial contact message) when contacted by another user. Our study is based on a large dataset from a major online dating site in China. We formulate the reply prediction as a link prediction problem of social networks and approach it using a machine learning framework. We extract user-based features from user profiles and graph-based features that are appropriate for the bipartite dating network, apply them in a variety of classification algorithms, and compare the utility of the features and performance of the classifiers. Our results show that the user-based and graph-based features result in similar performance, and can be used to effectively predict the reciprocal links. Only a small performance gain is achieved when both feature sets are used. Among the five classifiers we considered, Random Forests method outperforms the other four algorithms (Naive Bayes, Logistic Regression, KNN, and SVM). Our methods and results can provide valuable guidelines to the design and performance of recommendation engine for online dating sites.

References

Cai, X.; Bain, M.; Krzywicki, A.; Wobcke, W.; Kim, Y. S.; Compton, P.; and Mahidadia, A. 2010. Collaborative filtering for people to people recommendation in social networks. In *Austrlian Joint Conference on Artifical Intelligence*, 476–485. Springer.

Cai, X.; Bain, M.; Krzywicki, A.; Wobcke, W.; Kim, Y. S.; Compton, P.; and Mahidadia, A. 2012. Reciprocal and heterogeneous link prediction in social networks. In *PAKDD*, 193–204.

Chih-Chung, C., and Chih-Jen, L. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Dong, Y.; Tang, J.; Wu, S.; Tian, J.; Chawla, N. V.; Rao, J.; and Cao, H. 2012. Link prediction and recommendation across heterogeneous social network. In *ICDM*. IEEE.

Table 6: Graph-based Feature Rankings

Male to Sample Female				Female to Sample Male			
Feature Category	Feature Description	IG	χ^2	Feature Category	Feature Description	IG	χ^2
topological feature	reply-rate^t $(receiver)$	1	1	topological feature	reply-rate^t $(receiver)$	1	1
similarity feature	$WGU_a^s\,(sender, receiver)$	2	2	topological feature	send-rate^t $(receiver)$	2	2
similarity feature	$WGU_a^r\,(sender, receiver)$	3	3	topological feature	send-rate^t $(sender)$	3	4
similarity feature	$WGU_a^{re}\,(sender, receiver)$	4	4	topological feature	replied-rate^t $(receiver)$	4	3
neighbor feature	$\lvert\Gamma_{out}(sender)\rvert * \lvert\Gamma_{in}(receiver)\rvert$	5	5	similarity feature	$WUG_i^{re}\,(sender, receiver)$	5	5
topological feature	send-rate^t $(receiver)$	6	6	similarity feature	$WGU_a^{re}\,(sender, receiver)$	7	6
neighbor feature	$\frac{\lvert\Gamma_{in}(receiver)\cap S_a(sender)\rvert}{\lvert\Gamma_{in}(receiver)\cup S_a(sender)\rvert}$	7	7	neighbor feature	$\lvert\Gamma_{in}(sender)\cap S_a\,(receiver)\,\rvert$	6	9
neighbor feature	$\lvert\Gamma_{in}(receiver)\cap S_a\,(sender)\,\rvert$	8	9	neighbor feature	$\lvert\Gamma_{in}(receiver)\cap S_a\,(sender)\,\rvert$	8	8
topological feature	followed-rate^t $(sender)$	9	9	similarity feature	$WUG_a^{re}\,(sender, receiver)$	9	7
similarity feature	$WUG_a^{re}\,(sender, receiver)$	10	10	neighbor feature	$\lvert\Gamma_{in}(sender)\rvert * \lvert\Gamma_{out}(receiver)\rvert$	10	11
topological feature	receive-rate^t $(sender)$	11	11	neighbor feature	$\frac{\lvert\Gamma_{in}(receiver)\cap S_a(sender)\rvert}{\lvert\Gamma_{in}(receiver)\cup S_a(sender)\rvert}$	12	10
topological feature	replied-rate^t $(sender)$	12	12	similarity feature	$WGG_{i,i}^{re}\,(sender, receiver)$	11	12
similarity feature	$WGG_{i,a}^{r}\,(sender, receiver)$	13	13	similarity feature	$WGU_a^{rd}\,(sender, receiver)$	14	13
neighbor feature	$\lvert\Gamma_{in}(sender)\rvert * \lvert\Gamma_{in}(receiver)\rvert$	14	15	neighbor feature	$\frac{\lvert\Gamma_{out}(receiver)\cap S_i(sender)\rvert}{\lvert\Gamma_{out}(receiver)\cup S_i(sender)\rvert}$	13	15
neighbor feature	$\lvert\Gamma_{out}(sender)\cap S_i\,(receiver)\,\rvert$	15	14	similarity feature	$WGG_{i,a}^{re}\,(sender, receiver)$	15	14

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. In *Journal of Machine Learning Research*, 1871–1874. ACM.

Fire, M.; Tenenboim, L.; Lesser, O.; Puzis, R.; Rokach, L.; and Elovici, Y. 2011. Link prediction in social networks using computationally efficient topological features. In *SocialCom*, 73–80. IEEE.

Guha, R.; Kumar, R.; Baghavan, P.; and Tomkin, A. 2004. Propagation of trust and distrust. In *WWW*, 403–412. ACM.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. In *KDD*. ACM.

Hasan, M. A.; Chaoji, V.; Salem, S.; and Zaki, M. 2006. Link prediction using supervised learning. In *SDM*.

Hopcroft, J.; Lou, T.; and Tang, J. 2011. Who will follow you back? reciprocal relationship prediction. In *CIKM*. ACM.

Kahanda, I., and Neville, J. 2009. Using transactional information to predict link strength in online social networks. In *ICWSM*. AAAI.

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010. Predicting positive and negative links in online social networks. In *WWW*. ACM.

Li, L., and Li, T. 2012. Meet: a generalized framework for reciprocal recommender systems. In *CIKM*. ACM.

Liben-Nowell, D., and Kleinberg, J. 2003. The link prediction problem for social netowrks. In *CIKM*, 556–559. ACM.

Lichtenwalter, R. N.; Lussier, J. T.; and Chawla, N. V. 2010. New perspectives and method in link prediction. In *KDD*. ACM.

Pizzato, L.; Rej, T.; Chung, T.; Korprinska, I.; and Kay, J. 2010. Recon: A reciprocal recommender for online dating. In *RecSys*. ACM.

Sadilek, A.; Kautz, H.; and P.Bigham, J. 2012. Finding your friends and following them to where you are. In *WSDM*. ACM.

Tu, K.; Ribeiro, B.; Jensen, D.; Towsley, D.; Liu, B.; Jiang, H.; and Wang, X. 2014. Online dating recommendations: Matching markets and learning preferences. In *the 5th International Workshop on Social Recommender Systems (SRS 2014), in conjunction with 23rd International World Wide Web Conference (WWW)*.

Wu, S.; Raschid, L.; and Rand, W. 2011. Future link prediction in the blogosphere for recommendation. In *ICWSM*. AAAI.

Xia, P.; Ribeiro, B.; Chen, C.; Liu, B.; and Towsley, D. 2013. A study of user behaviors on an online dating site. In *ASONAM*. ACM.

Xia, P.; Tu, K.; Ribeiro, B.; Jiang, H.; Wang, X.; Chen, C.; Liu, B.; and Towsley, D. 2014. Who is dating whom: Characterizing user behaviors of a large online dating site. In *arXiv:1401.5710*. arXiv.

Yang, Y.; Chawla, N. V.; Basu, P.; Prabhala, B.; and Porta, T. L. 2013. Link preiction in human mobility networks. In *ASONAM*. ACM.

Zhao, K.; Wang, X.; Yu, M.; and Gao, B. 2014. User recommendation in reciprocal and bipartite social networks –a case study of online dating. In *Intelligent Systems*. IEEE.