Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Clustering-Based Collaborative Filtering for Link Prediction *

Xiangyu Wang, Dayu He, Danyang Chen, Jinhui Xu
Department of Computer Science and Engineering
State University of New York at Buffalo
{*xiangyuw,dayuhe,danyangc,jinhui*}*@buffalo.edu*

Abstract

In this paper, we propose a novel collaborative filtering approach for predicting the unobserved links in a network (or graph) with both topological and node features. Our approach improves the well-known compressed sensing based matrix completion method by introducing a new multiple-independent-Bernoulli-distribution model as the data sampling mask. It makes better link predictions since the model is more general and better matches the data distributions in many real-world networks, such as social networks like Facebook. As a result, a satisfying stability of the prediction can be guaranteed. To obtain an accurate multiple-independent-Bernoulli-distribution model of the topological feature space, our approach adjusts the sampling of the adjacency matrix of the network (or graph) using the clustering information in the node feature space. This yields a better performance than those methods which simply combine the two types of features. Experimental results on several benchmark datasets suggest that our approach outperforms the best existing link prediction methods.

1 Introduction

Over the last few years, there is a rapidly growing interest in predicting the potential or absent links between nodes in large, complex networks. The link information represents the interactions, relationships, or influences between different nodes. Thus, predicting the likelihood of the unknown links is essential to forecasting the future or determining the hidden relationships between the nodes. For instance, with the surge of social networks in our daily life, link prediction has been widely applied in friendship recommendation and social network marketing (Liben-Nowell and Kleinberg 2003). In this paper, we focus on predicting unknown links in a partially observed network. The observed information can be classified into two categories. One is the set of topological features, such as edges or links, and the other is the set of node features, such as a user's name, gender, location, school, and other profile elements in some networks like Facebook.

*The research of this work was supported in part by NSF under grants IIS-1115220, IIS-1422591, and CCF-1422324.

1.1 Previous Work

A number of methods have been developed to predict the absent links, using either one or both types of features. A straightforward way of link prediction is to first compute the similarity between each pair of nodes and then use it to determine the likelihood of their possible link. Two types of similarity between a pair of nodes are frequently used. One is based on node attributes such as the number of common features shared by them (Lin 1998), and the other is based on graph-topological features such as the paths or neighbors connecting them (Leicht, Holme, and Newman 2006; Lü, Jin, and Zhou 2009). A major challenge encountered by such methods is that the node features are often treated independently of the topological features, and therefore it lacks a consistent and appropriate way to define similarity using the two types of correlated features. As a result, the performance of such methods could vary significantly among different types of similarities. To overcome this obstacle, several Bayesian probabilistic models have been proposed to make predictions by learning a link probability distribution model from the observed network (Wang, Satuluri, and Parthasarathy 2007; Miller, Griffiths, and Jordan 2009; Sarkar, Chakrabarti, and Jordan 2012). In such models, the node features and topological features are represented as random variables, and their relations are estimated via some assumed latent structure from the Bayesian method.

Another approach for link prediction is to formulate it as an adjacency matrix recovery problem, and use some sparse matrix recovery methods or collaborative filtering methods to solve it (Menon and Elkan 2011; Koren, Bell, and Volinsky 2009; Singh and Gordon 2008; Agrawal, Garg, and Narayanam 2013; Ye et al. 2013). Comparing to the statistical models, the matrix recovery approach has a prominent advantage, that is, its stability and accuracy of solutions can be theoretically ensured by the well-developed sparse recovery theory (Zhou et al. 2010; Candès and Plan 2009; Wang and Xu 2012). However, it also suffers from two limitations. One limitation is that the matrix recovery approach uses matrix norm as its measure and thus mainly focuses on the topological features. The other limitation is that the matrix recovery approach searches the missing links along the norm-minimizing direction (Menon and Elkan 2011; Chen 2008) (*e.g.,* similar to the least square problem), and the resulting matrix may not always reflect the real structure

of the network.

Thus it is desirable to combine the advantages of the two approaches, and avoid their respective limitations. This motivates us to propose a new model to predict links by learning the original structure from both node and topological features and guarantee the stability of the solutions.

1.2 Main Ideas

A key assumption of our approach is that the two types of features are correlated and the topological feature is a reflection of some correlated node features. Thus, our approach tries to make use of the node features to help determining the unknown topological features. Particularly, our approach adopts the following main ideas.

Firstly, to ensure stability of predication, our approach adopts the framework of the reliable matrix completion approach. The predication can be treated as recovering the (partially unknown) original link matrix from the observed (*i.e.,* sampled) matrix M with $M_{ij} = 1$ indicating a link between nodes N_i and N_j. An effective strategy of sampling is proposed, in this paper, to guarantee the recovered matrix obeying the real structure of the original network (*i.e.,* the network containing both the observed and unknown links). The newly recovered links are the predications.

Secondly, for each node N_j, our approach views its observed neighbors as a cluster in the node feature space. A node N_i has a link to N_j if it belongs to the cluster of N_j in the node feature space. Thus, the j-th column M_j of the link matrix can be viewed as a vector generated by a Bernoulli distribution model in which the probability P_j is calculated by N_j's cluster. That is, we have $M_{ij} = 1$ with probability P_j (if N_i is inside the cluster of N_j) and $M_{ij} = 0$ with probability $1 - P_j$ (if N_i is outside the cluster). A major advantage of such a model is that it allows us to extend the well-known matrix completion framework (Candès and Tao 2005; Candès and Recht 2009; Zhou et al. 2010; Gross 2011) and prove the stability of our approach. In the original matrix completion framework, the stability is shown based on the assumption that every entry of the matrix is sampled by the same Bernoulli distribution. In our approach, the stability is analyzed under a more practical model in which the columns of the matrix are generated by multiple independent Bernoulli distributions and the probability of each entry depends on the actual relations between the corresponding node and its neighbors. There are also some existing non-uniform sampling models. A representative one is the local coherences model introduced by Chen *et al.* (Chen et al. 2014), in which each entry is sampled based on its own row and column coherence coefficient. Comparing to our model, these methods still depend only on the topological features (*i.e.,* without using the node feature information). Furthermore, computing the local coherences could be quite expensive, which needs to perform Singular Value Decomposition on a large-scale adjacency matrix.

Thirdly, our approach adopts an adaptive calibration strategy to obtain the Bernoulli model generating each column M_j. Hence the partially unknown network structure can be more accurately estimated. It consists of two calibration phases. In the first phase, our approach calculates a unique weight vector for each node N_j. This gives us the tightest cluster after filtering out those node features irrelevant to the cluster. In the second phase, the sampling probability of each column is adjusted based on the obtained clustering information, which serves as a bridge between the node feature space and the topological feature space. This allows us to use the relevant node features to help predicting the topological features.

1.3 Our Contributions

Our proposed approach has the following main contributions.

(1) *A matrix completion model based on multiple independent Bernoulli distributions*: We extend the original matrix completion model for the reconstruction of a partially observed network by using a new data distribution model. A theoretical analysis on the stability is provided to ensure the performance of this model.

(2) *An adaptive calibration strategy based on feature subspace*: We introduce a new way to bridge the node feature space and the topological feature space, which allows us to use relevant node features to predict topological features.

The rest of the paper is organized as follows. In the second section, we present our new model for link prediction. A stability analysis of this model is given in the third section. Experimental results are shown in the fourth section. Conclusion remarks are given in the last section.

2 Proposed Model for Link Prediction

Let $\mathbb{G}$ be a directed graph representing the considered network, and N be the set of n nodes of $\mathbb{G}$ with N_i being the i-th node of $\mathbb{G}$. Let $M \in \mathbb{R}^{n\times n}$ be the observed adjacency matrix of the network, and M_j be its j-th column. $M_{ij} = 1$ if a link from N_i to N_j is observed, and $M_{ij} = 0$ otherwise. Denote Ω the set of indices of all non-zero entries in M. Then predicting unknown links is equivalent to recover the unobserved links in M based on Ω. Let $\tilde{M}$ be the resulting (or recovered) adjacency matrix. The predicted links are stored in $\tilde{M} - M$.

2.1 Problem Description and Preliminary

We now formulate the link prediction problem in the matrix completion framework (Candès and Recht 2009) as follows.

$$\begin{aligned} &\min\ ||\tilde{M}||_* \\ &\text{subject to } \mathcal{P}_\Omega(\tilde{M}) = \mathcal{P}_\Omega(M), \end{aligned} \tag{1}$$

where $||\tilde{M}||_*$ denotes the nuclear norm of $\tilde{M}$, which is a good convex relaxation of *matrix rank*, and $\mathcal{P}_\Omega$ is the sampling projector determined by our multiple independent Bernoulli distribution model. The definition of $\mathcal{P}_\Omega$ will be introduced later. The observed links in M are treated as the links sampled from the unknown original adjacency matrix of $\mathbb{G}$ by $\mathcal{P}_\Omega$. In matrix completion theory, the quality of recovery depends on sampling. Hence, our model focuses on constructing $\mathcal{P}_\Omega(M)$ so that it closely reflects the structure of $\mathbb{G}$. In the following subsections, we explain how to achieve this.

To be consistent with the framework in (Candès and Recht 2009; Gross 2011), below we introduce some necessary definitions and assumptions required by the matrix completion framework.

Let $M = \sum_{k=1}^{r} \delta_k u_k v_k^\top$ be the SVD form of $M$and r be the rank of M. Let U and V be the spaces spanned by $\{u_k\}_{k=1}^r$ and $\{v_k\}_{k=1}^r$, respectively. Assume that M has low rank and satisfies the following coherence condition introduced in (Gross 2011).

Definition 1 *The $n \times n$ matrix M has the coherence μ with respect to basis $\{\omega_{ij}\}_{ij}^{n^2}$ if*

$$\max_{ij} ||\mathcal{P}_T(\omega_{ij})||_F^2 \leq \frac{\mu r}{n}, \tag{2}$$

$$\max_{ij} \langle \omega_{ij}, sgn(UV^\top) \rangle \leq \frac{\mu r}{n^2}, \tag{3}$$

where $\text{sgn}(UV^\top)$ is the sign function of $UV^\top$ (denoted by sgn). Let $||M||_2$ be the spectral norm of a matrix M, and $||M||_F$ be its Frobenius norm.

Below we discuss our model.

2.2 Node Clusters

A key assumption used in the original matrix completion framework is that elements in $\mathcal{P}_\Omega(M)$ are generated by a sequence of independent identically distributed $0/1$ Bernoulli random variables. Below, we introduce a new probabilistic model for $\mathcal{P}_\Omega$ based on the correlation between the topological features M and the node features F.

For each node, we assume that all its possible linking nodes aggregate as a cluster. Let $\{M_{ij}\}_{i=1}^k$ be the set of non-zero entries of vector M_j, and $\{N_i\}_{i=1}^k$ be the corresponding nodes which have a link to node N_j in M. $\{N_i\}_{i=1}^k$ forms a neighborhood of N_j (this is based on a commonly used belief that two nodes sharing a link have certain similarity). By our assumption that topological features are reflections of node features, this means that $\{N_i\}_{i=1}^k$ aggregates in the f-dimensional node feature space or its subspace, where f is the number of node features in the network. To measure the quality of the aggregation, we can find a weight vector $W_j \in \mathbb{R}^f$ to make $\{N_i\}_{i=1}^k$ become the smallest weighted cluster of N_j. More specifically, let $F \in \mathbb{R}^{f \times k}$ denote the feature matrix of the k nodes with the i-th column F_i being the feature vector of N_i; the cluster of N_j is estimated by the following

$$\delta_j = \min_{W_j} \frac{1}{k} \sum_{i=1}^{k} ||W_j \otimes F_i - F_j||, \tag{4}$$

where δ_j is the radius of cluster of each N_j, and $\otimes$ is an element-by-element multiplication.[1] Using W_j and δ_j, we can determine whether any other node N_i is inside the cluster.

[1]That is, the result of $W_j \otimes F_i$ is still a f dimensional vector with its q-th entry equaling to the product of the two q-th entries in W_j and F_i respectively.

2.3 Multiple Independent Bernoulli Distribution Model

With the above ideas, we can define the *multiple independent Bernoulli distribution model* in the following way. Let $\{m_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq n}$ be a sequence of random variables independently sampled from $0/1$ Bernoulli's distribution with

$$m_{ij} = \begin{cases} 1 & \text{if } N_i \text{ is inside } N_j\text{'s cluster} \\ 0 & \text{if } N_i \text{ is outside } N_j\text{'s cluster} \end{cases} \tag{5}$$

with the probability

$$\mathbb{P}(m_{ij} = 1) = P_j, \tag{6}$$

where P_j is decided only by the cluster of N_j (*i.e.* W_j and δ_j). Then the operator $\mathcal{P}_\Omega$ in Eq(1) is defined as

$$\mathcal{P}_\Omega(M) = \sum_{i,j \in \Omega} \frac{1}{P_j} \omega_{ij} \langle \omega_{ij}, M \rangle, \tag{7}$$

where $\omega_{ij} = \{e_i e_j\}_{i,j}^n$ and $\{e_i\}_{i=1}^n$ is the standard basis; $\langle , \rangle$ is the matrix inner product defined as $\langle M_1, M_2 \rangle = tr(M_1^\top M_2)$ for two matrices M_1 and M_2. After defining $\mathcal{P}_\Omega$, we now need to calculate the probability P_j.

2.4 Adaptive Calibration

To obtain the sampling probability P_j for the jth column, we use a two-phase adaptive calibration strategy.

The first phase is only for the purpose of estimating the structure of the original network. In this phase, we use a uniform Bernoulli distribution to sample at least $O(\mu n r \log n)$ entries in Ω, which means that for each column we can sample at least $O(\mu r \log n)$ entries. Then, for each j, we obtain the radius δ_j and the feature weight W_j using Eq(4). With these, we can calculate the distances between node N_j and each of its neighbors N_k by the following equation.

$$\delta_{kj} = ||W_j \otimes F_k - F_j||. \tag{8}$$

Note that we only need the feature information of each node, which is already available.

In the second phase, we first set the sampling probability P_j using the clustering information computed in the first phase, and then calculate $\mathcal{P}_\Omega(M)$ by Eq(7). More specifically, we first compare δ_{kj} with δ_j to obtain the total number D_j of nodes inside the cluster of N_j, and then set the sampling probability of the jth column as

$$P_j = \max\left(\frac{D_j}{n}, \frac{\mu r \log n}{n}\right). \tag{9}$$

With this calibration, the stability of the recovery only depends on P_j. As a result, it guarantees that our recovery can preserve the main structure of the original network structure. We analyze the recovery stability in the next section.

3 Stability Analysis

In this section, we show stability of our prediction approach by analyzing the stability of the equivalent matrix completion approach under the multiple independent Bernoulli distribution model.

Theorem 1 *Let M be an $n \times n$ matrix of rank r sampled from the multiple independent Bernoulli distributions. If Ω is generated by a sequence of independent $0/1$ Bernoulli random variables $\{m_{ij}\}$ with probability $\mathbb{P}(m_{ij} = 1) = P_j \geq \frac{\mu r \log n}{n}$, then there exist numerical constants C_0 and β, such that the minimizer to the problem (1) is unique and equal to M with probability at least $1 - C_0 n^{-\beta}$.*

To prove this theorem, we first recall how stability is shown in the original matrix completion model (Candès and Recht 2009; Gross 2011). Let T be the linear space spanned by both $\{u_k\}_{k=1}^r$ and $\{v_k\}_{k=1}^r$. Define the orthogonal projector $\mathcal{P}_T$ onto T as

$$\mathcal{P}_T(M) = \mathcal{P}_U M + \mathcal{P}_V M - \mathcal{P}_U M \mathcal{P}_V. \tag{10}$$

Denote by $\mathcal{P}_{T^c}$ the orthogonal projector onto the orthogonal compliment subspace of T.

Our proof strategy follows the architecture of (Candès and Recht 2009; Gross 2011), that is, we need to construct a dual certificate matrix $Y \in \text{range}(\mathcal{P}_\Omega)$ satisfying the following Lemma.

Lemma 1 *There exists a matrix M of rank r which is the solution of the problem (1), if the following two conditions hold:*

1. *$\mathcal{P}_\Omega$ restricted to elements in T is injective.*
2. *There exists a dual certificate $Y \in$ range$(\mathcal{P}_\Omega)$ obeys*

$$||\mathcal{P}_T(Y) - sgn||_F \leq \frac{1}{C_R n^3},\ C_R > 1. \tag{11}$$

$$||\mathcal{P}_{T^c}(Y)||_2 < 1. \tag{12}$$

Below we show that the two conditions indeed hold.

3.1 The Injectivity

To prove the injectivity, a key step is to show the operator $\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T$ has a small spectral norm with a large probability.

Lemma 2 *Let Ω be the set of indices sampled using the multiple independent Bernoulli distribution model, there exists a numerical constants C_1 such that for all $t < 1$,*

$$||\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T - \mathcal{P}_T||_2 \leq t \tag{13}$$

with probability at least $1 - 2n^{-C_1 t^2}$, if $\min P_j \geq \frac{\mu r \log n}{n}$.

Proof: Firstly, for any matrix X, we have

$$\begin{aligned} & [\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T - \mathcal{P}_T](X) \\ = & \sum_j \left[\sum_i (\frac{m_{ij}}{P_j} - 1) \langle \omega_{ij}, \mathcal{P}_T(X) \rangle \mathcal{P}_T(\omega_{ij}) \right]. \end{aligned} \tag{14}$$

Define a new family of operators as $Z_{ij}(X) = (\frac{m_{ij}}{P_j} - 1)\langle \omega_{ij}, \mathcal{P}_T(X)\rangle \mathcal{P}_T(\omega_{ij})$. Then

$$\begin{aligned} & || \sum_{ij} EZ_{ij}(X)^2 ||_2 \\ = & || \sum_{ij} (\frac{m_{ij}}{P_j} - 1) \langle \omega_{ij}, \mathcal{P}_T(\omega_{ij}) \rangle Z_{ij}(X) ||_2 \\ \leq & (\frac{1}{\min P_j} - 1) ||\mathcal{P}_T(\omega_{ij})||_F^2 || \sum_{ij} EZ_{ij}(X) ||_F \\ \leq & \frac{1 - \min P_j}{\min P_j} ||\mathcal{P}_T(\omega_{ij})||_F^2 || \sum_{ij} E \mathcal{P}_T \mathcal{P}_{\Omega_{ij}} \mathcal{P}_T(X) ||_F \\ \leq & (\frac{1}{\min P_j} - 1) ||\mathcal{P}_T(\omega_{ij})||_F^2 ||\mathcal{P}_T(X)||_F, \end{aligned} \tag{15}$$

where $\langle \omega_{ij}, \mathcal{P}_T(\omega_{ij}) \rangle = ||\mathcal{P}_T(\omega_{ij})||_F^2$. Then, based on Eq(2), we have

$$|| \sum_{ij} EZ_{ij}^2 ||_2 \leq (\frac{1}{\min P_j} - 1) \frac{\mu r}{n}. \tag{16}$$

Next, we have

$$\begin{aligned} ||Z_{ij}||_2 & \leq \frac{1}{P_j} ||\mathcal{P}_T \mathcal{P}_{\Omega_{ij}} \mathcal{P}_T||_2 \\ & \leq \frac{1}{\min P_j} ||\mathcal{P}_T(\omega_{ij})||_F^2 \\ & \leq \frac{\mu r}{(\min P_j) n}. \end{aligned} \tag{17}$$

Since $EZ_{ij} = 0$, by the *Matrix Berstein Inequality* (Tropp 2012) (*i.e.,* Theorem 2), we have the following inequality for any $t \leq 1/2$

$$\Pr(||\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T - \mathcal{P}_T||_2 \geq t) \leq 2n^{-C_1 t^2} \tag{18}$$

if $\min P_j \geq \frac{\mu r \log n}{n}$, where $C_1 > 0$ is a constant coefficient. □

3.2 The Dual Certificate

To construct the dual certificate Y, we use the strategy of golfing scheme (Gross 2011) introdued by Gross. The golfing scheme first partitions Ω into l non-overlapping groups, *i.e.,* $\Omega = \bigcup_{1 \leq k \leq l} \Omega_k$, then set $Y_0 = 0$ and

$$Y_k = Y_{k-1} + \mathcal{P}_{\Omega_k} \mathcal{P}_T(\text{sgn} - Y_{k-1}), \tag{19}$$

where $\mathcal{P}_{\Omega_k}$ has a parameter $Pr(m_{ij}^{(k)} = 1) = P_j^{(k)}$, and sgn stands for the sign function of M. This series $\{Y_k\}_{k=1}^l$ will converge to the expected dual certificate Y_l.

Let $Z_0 = \text{sgn}$. Then we have

$$Z_k = \mathcal{P}_T(\text{sgn} - Y_{k-1}), \tag{20}$$

$$Y_k = \sum_{i=1}^k \mathcal{P}_{\Omega_i} Z_{i-1}. \tag{21}$$

The following two lemmas are key components for proving the conditions (*i.e.,* Eq(11) and Eq(12)) for the dual certificate Y in Lemma 1.

Lemma 3 *Suppose $Z \in T$, then*

$$Pr(||(\mathbf{1} - \mathcal{P}_T\mathcal{P}_{\Omega'}\mathcal{P}_T)Z||_\infty > \alpha||Z||_\infty) \leq 2n^{-C_2\alpha^2}, \quad (22)$$

with $\alpha, C_2 > 0$, if $\min P'_j \geq \frac{r\mu \log n}{n}$.

The proof is shown below. By a similar argument given in (Gross 2011), we also know that the series $\{Y_k\}_{k=1}^l$ converges to the expected dual certificate Y_l from this lemma.

Proof: For any $Z \in T$, we have $\mathcal{P}_T\mathcal{P}_{\Omega'}\mathcal{P}_T Z = \mathcal{P}_T\mathcal{P}_{\Omega'}Z$. Then we know that

$$Z - \mathcal{P}_T\mathcal{P}_{\Omega'}\mathcal{P}_T Z = \sum_i \sum_j (1 - \frac{m'_{ij}}{P'_j})\langle \omega_{ij}, Z\rangle \mathcal{P}_T\omega_{ij}. \quad (23)$$

For any pair of $\{i', j'\} \in \Omega'$, let

$$\begin{aligned} X_{ij} &= \langle (1 - \frac{m'_{ij}}{P'_j})\langle \omega_{ij}, Z\rangle \mathcal{P}_T\omega_{ij}, \omega_{i'j'}\rangle \\ &= (1 - \frac{m'_{ij}}{P'_j})\langle \omega_{ij}, Z\rangle \langle \mathcal{P}_T\omega_{ij}, \omega_{i'j'}\rangle. \end{aligned} \quad (24)$$

Then, we have $|Z - \mathcal{P}_T\mathcal{P}_{\Omega'}\mathcal{P}_T Z|_{i'j'} = \sum_{ij} X_{ij}$ and

$$\begin{aligned} &\sum_{ij} EX_{ij}^2 \\ &\leq (\frac{1}{\min P'_j} - 1)\sum_{ij} |\langle \omega_{ij}, Z\rangle|^2 |\langle \mathcal{P}_T\omega_{ij}, \omega_{i'j'}\rangle|^2 \\ &\leq (\frac{1}{\min P'_j} - 1)||Z||_\infty^2 ||\mathcal{P}_T\omega_{i'j'}||_F^2 \\ &\leq (\frac{1}{\min P'_j} - 1)\frac{\mu r}{n}||Z||_\infty^2. \end{aligned} \quad (25)$$

Also, we know that

$$\begin{aligned} |X_{ij}| &\leq (\frac{1}{\min P'_j} - 1)||Z||_\infty |\langle \mathcal{P}_T\omega_{ij}, \omega_{i'j'}\rangle| \\ &\leq (\frac{1}{\min P'_j} - 1)\frac{\mu r}{n}||Z||_\infty. \end{aligned} \quad (26)$$

Thus, if $\min P'_j \geq \frac{r\mu \log n}{n}$, by Bernstein's inequality, we know that there exists a constant $C_2 > 0$ such that

$$\begin{aligned} &\Pr(||(\mathbf{1} - \mathcal{P}_T\mathcal{P}_{\Omega'}\mathcal{P}_T)Z||_\infty > \alpha||Z||_\infty) \\ &\leq 2\exp\left(-C_2\frac{\alpha^2 n \min P'_j}{\mu r}\right). \end{aligned} \quad (27)$$

This completes the proof of the lemma. □

Next lemma is similar to Lemma 9 of (Gross 2011), it states as follows.

Lemma 4 *Suppose $Z \in T$, then*

$$Pr(||\mathcal{P}_{T^c}\mathcal{P}_\Omega Z||_2 > t) \leq 2\exp\left(-\frac{t^2 n \min P_j}{C_3 \max ||Z||_2^2 \mu}\right), \quad (28)$$

for $t \leq \frac{\max ||Z||_2}{4\sqrt{r}}$

Proof: Let $X_{ij} = \mathcal{P}_{T^c}\frac{m_{ij}}{P_j}\langle \omega_{ij}, Z\rangle \omega_{ij}$. Then $\mathcal{P}_{T^c}\mathcal{P}_\Omega Z = \sum_{ij} X_{ij}$ and $EX_{ij} = 0$.

By the fact that

$$||\mathcal{P}_{T^c}\omega_{ij}||_F^2 = \sup_{\psi \in T^c \ ||\psi||_2 = 1} \langle \omega_{ij}, \psi\rangle, \quad (29)$$

and Lemma 8 of (Gross 2011), we have

$$\begin{aligned} \sum_{ij} \mathrm{Var}(X_{ij}) &\leq \sum_{ij} \frac{1}{\min P_j} |\langle \omega_{ij}, Z\rangle|^2 ||\mathcal{P}_{T^c}\omega_{ij}||_F^2 \\ &\leq \frac{1}{\min P_j} \sup_\psi \sum_{ij} |\langle \omega_{ij}, Z\rangle|^2 \langle \psi, \omega_{ij}^2 \psi\rangle \\ &\leq \frac{\mu \max ||Z||_2^2}{n \min P_j}. \end{aligned} \quad (30)$$

Also, we know that

$$|X_{ij}| \leq \frac{1}{\min P_j} |\langle \omega_{ij}, Z\rangle| ||\mathcal{P}_{T^c}\omega_{ij}||_F \leq \frac{\mu \max ||Z||_2}{n \min P_j}. \quad (31)$$

By Bernstein's inequality, the lemma follows. □

3.3 Matrix Berstein Inequality

Theorem 2 *(Matrix Berstein Inequality (Tropp 2012)) Let $(Y_k)_{k \geq 1}$ be independent matrices in $\mathbb{R}^{m \times n}$ satisfying*

$$EY_k = 0 \text{ and } ||Y_k||_2 \leq R. \quad (32)$$

Define the variance parameter δ as

$$\delta^2 = \max(||\sum_k EY_k Y_k^\top||_2, ||\sum_k EY_k^\top Y_k||_2). \quad (33)$$

Then, for all $t \geq 0$

$$Pr(||\sum_k Y_k||_2 \geq t) \leq (m+n)\exp\left(\frac{-t^2}{3\delta^2 + 2Rt}\right). \quad (34)$$

3.4 Proof of Lemma 1

Now we show that the constructed dual certificate Y_l satisfies the two conditions Eq(11) and Eq(12) in Lemma 1.

Proof: (*a*) Let $Z_k = (\mathbf{1} - \mathcal{P}_T\mathcal{P}_{\Omega'}\mathcal{P}_T)Z_{k-1}$, then by Lemma 3, we have

$$||Z_k||_\infty \leq \alpha||Z_{k-1}||_\infty, \quad (35)$$

with large probability.

Then we have

$$||Z_k||_\infty \leq \alpha^k||\mathrm{sgn}||_\infty. \quad (36)$$

It follows that

$$||Z_k||_F \leq \alpha^k||\mathrm{sgn}||_F = \alpha^k\sqrt{r}. \quad (37)$$

Set $\alpha = 1/2$ and $l = \lceil \log_2(2n^3\sqrt{r})\rceil$, we get

$$||\mathcal{P}_T(Y_l) - \mathrm{sgn}||_F = \sum_{k=1}^l (\frac{1}{2})^{k-1} \leq \frac{1}{C_R n^3}, \quad (38)$$

with $C_R > 1$.

(b) By Lemma 4, and setting $\max ||Z_k||_2 = \frac{\sqrt{r}}{2^k}$ (by Eq(37)), we have

$$||\mathcal{P}_{T^c}\mathcal{P}_\Omega Z_k||_2 \leq \frac{1}{4}2^{-k}. \quad (39)$$

It follows that

$$||\mathcal{P}_{T^c}\mathcal{P}_\Omega Y_l||_2 \leq \frac{1}{4}\sum_{k=1}^{l}(\frac{1}{2})^{k-1} < \frac{1}{2}. \quad (40)$$

This means that Y_l satisfies both conditions. □

Now with $\max ||Z_k||_2 = \frac{\sqrt{r}}{2^k}$ as above and $\min P_j \geq \frac{\mu r \log n}{n}$, we can bound the probabilities in Lemmas 2,3,4 by

$$1 - C_0 n^{-\beta}, \quad (41)$$

where $C_0, \beta > 0$ are some constant coefficients. Thus Theorem 1 follows. This indicates that our prediction model has guaranteed performance with $|\Omega| \geq O(\mu r n \log^2 n)$.

4 Experiments

To evaluate the performance of our proposed approach, we implement our clustering-based collaborative filtering method (CBCF) and compare it with some existing link prediction methods. Experimental results are reported as the area under the ROC curve (AUC). The regularization parameters of our model are selected following the same rules in (Goldfarb and Ma 2011). In our experiments, we use the proximal gradient algorithm (APG) in (Cai, Candès, and Shen 2010; Shen, Toh, and Yun 2011; Goldfarb and Ma 2011) as the numerical solver. As for the running time, the first phase of the adaptive calibration procedure takes $O(f\mu rn \log n)$ time, and the second phase takes $O(fn^2)$ time, where f is the number of node features and is in general much smaller than n. Since APG needs to perform SVD on matrix M, which will dominate the time of the adaptive calibration, our model thus has asymptotically the same time complexity as APG. The readers are referred to the reference of APG for more details.

Firstly, we compare our method with a matrix factorization and bilinear regression model (FactBLR) proposed by (Menon and Elkan 2011). The FactBLR extracts edge information using matrix factorization and makes use of explicit node information through bilinear regression. Its prediction is made based on the linear combination of both types of information. Thus, it outperforms the popular link prediction methods which used only one type of information. As comparisons with our approach, we perform these methods on two datasets: the protein interaction dataset (Protein) from (Tsuda and Noble 2004) and the metabolic pathway interaction dataset (Metabolic) from (Yamanishi, Vert, and Kanehisa 2005). Protein dataset contains 2617 proteins and each protein has a 76 dimensional feature. Metabolic dataset contains 668 nodes and each node has a 325 dimensional feature. To be consistent with FactBLR experiments in (Menon and Elkan 2011), we uniformly select 10% of the interactions as training set for each dataset and repeat the whole evaluation process 10 times. We use the popular unsupervised scoring methods Katz and Shortest-Path (ShP) as the standard baseline (Menon and Elkan 2011). Results are reported in Table 1. From the table, it is clear that (1) CBCF significantly outperforms the methods which use only the topological features or only the node features, since CBCF scores a link depending on more information; (2) CBCF still shows advantages compared to FactBLR, which linearly combines the topological feature and node feature. This is reasonable, since there could be different data types and value systems between the topological feature space and the node feature space. For example, the value of topological feature is usually 1 or 0, but the value of node feature could be either a very large number or a very small number. Thus a score generated from a linear combination of these two feature spaces could be weighted unnecessarily more for one feature.

Table 1: Mean AUC scores for Katz, Shortest Path, FactBLR and CBCF

DATASET	KATZ	SHP	FACTBLR	CBCF
PROTEIN	0.727	0.726	0.813	0.831
METABOLIC	0.608	0.626	0.763	0.776

Table 2: Mean AUC scores for Katz, SPLR and CBCF with different corruption fraction δ

δ	KATZ	SPLR	CBCF
5%	0.9298	0.9293	0.9453
10%	0.9189	0.9221	0.9411
20%	0.8941	0.8997	0.9248

Secondly, we compare CBCF with the sparse and low rank matrices recovery method (SPLR) proposed by (Savalle, Richard, and Vayatis 2012). Low rank and sparseness are usually the characters of the social network dataset, especially when the whole dataset is obtained by sampling. SPLR has demonstrated its ability in processing such type of datasets. SPLR is based on the original matrix completion model which assumes that the observed data is sampled from identical independent Bernoulli distributions. Following the experiments of (Savalle, Richard, and Vayatis 2012), we use the Facebook100 dataset proposed by (Traud, Mucha, and Porter 2011), which contains the Facebook friendship relations between students in one hundred universities. We choose the same university dataset, which has 41554 users and each user has a 7 dimensional feature. For consistency, we first filter the data to retain only 10% of nodes which have the highest number of friends. Then uniformly corrupt the entries of the adjacency matrix with a fixed fraction δ. Also Katz method is used as the baseline. Results are reported in Table 2.The related ROC graph is shown in Figure1. From this table, we can make two remarks. (1) This comparison shows that our new matrix model performs better than the original matrix completion model, which seems to suggest that the multiple independent Bernoulli distribution model is more suitable for social network data. (2) This experiment shows that our approach can also be applied to low rank and sparse datasets.

5 Conclusion

In this paper we proposed a link prediction approach based on a new multiple independent Bernoulli matrix completion

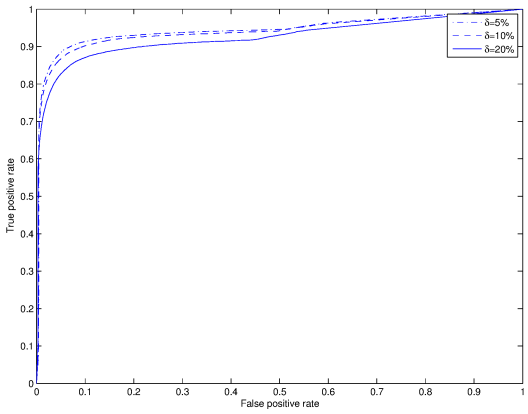

Figure 1: ROC graph for CBCF with different δ.

model. Our approach has a few advantages over the existing ones. Firstly, it is more general and a better fit for predicting unobserved links in many real-world networks such as social networks. This is because in such networks, each node often selects its linked nodes independently and follows only its own interest. Secondly, it allows us to ensure the stability of the solutions, which is given in the third section. Thirdly, it builds a bridge between the topological feature space and the node feature space through clustering, which allows us to collaboratively achieve better solutions by utilizing the two types of features and can deal with networks with insufficient node features.

Experimental results on several benchmark datasets suggest that our approach outperforms the original matrix completion model as well as the matrix factorization and bilinear regression method.

References

Agrawal, P.; Garg, V. K.; and Narayanam, R. 2013. Link label prediction in signed social networks. In *IJCAI*.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Candès, E. J., and Plan, Y. 2009. Matrix completion with noise. *CoRR* abs/0903.3131.

Candès, E. J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* 9(6):717–772.

Candès, E. J., and Tao, T. 2005. Decoding by linear programming. *CoRR* abs/math/0502327.

Chen, Y.; Bhojanapalli, S.; Sanghavi, S.; and Ward, R. 2014. Coherent matrix completion. In *ICML*, 674–682.

Chen, P. 2008. Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix. *International Journal of Computer Vision* 80(1):125–142.

Goldfarb, D., and Ma, S. 2011. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics* 11(2):183–210.

Gross, D. 2011. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory* 57(3):1548–1566.

Koren, Y.; Bell, R. M.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *IEEE Computer* 42(8):30–37.

Leicht, E. A.; Holme, P.; and Newman, M. E. J. 2006. Vertex similarity in networks. *Physical Review E* 73.

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

Lin, D. 1998. An information-theoretic definition of similarity. In *ICML*, 296–304.

Lü, L.; Jin, C.-H.; and Zhou, T. 2009. Similarity index based on local paths for link prediction of complex networks. *Physical Review E* 80:046122.

Menon, A. K., and Elkan, C. 2011. Link prediction via matrix factorization. In *ECML/PKDD (2)*, 437–452.

Miller, K.; Griffiths, T.; and Jordan, M. 2009. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems 22*, 1276–1284.

Sarkar, P.; Chakrabarti, D.; and Jordan, M. I. 2012. Nonparametric link prediction in dynamic networks. In *ICML*.

Savalle, P.-A.; Richard, E.; and Vayatis, N. 2012. Estimation of simultaneously sparse and low rank matrices. In *ICML*.

Shen, Z.; Toh, K.-C.; and Yun, S. 2011. An accelerated proximal gradient algorithm for frame-based image restoration via the balanced approach. *SIAM J. Imaging Sciences* 4(2):573–596.

Singh, A. P., and Gordon, G. J. 2008. A unified view of matrix factorization models. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, number 16 in ECML PKDD '08, 358–373. Springer-Verlag.

Traud, A. L.; Mucha, P. J.; and Porter, M. A. 2011. Social structure of facebook networks. *CoRR* abs/1102.2166.

Tropp, J. 2012. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* 12(4):389–434.

Tsuda, K., and Noble, W. S. 2004. Learning kernels from biological networks by maximizing entropy. In *ISMB/ECCB (Supplement of Bioinformatics)*, 326–333.

Wang, Y.-X., and Xu, H. 2012. Stability of matrix factorization for collaborative filtering. In *ICML*.

Wang, C.; Satuluri, V.; and Parthasarathy, S. 2007. Local probabilistic models for link prediction. In *ICDM '07*, 322–331.

Yamanishi, Y.; Vert, J.-P.; and Kanehisa, M. 2005. Supervised enzyme network inference from the integration of genomic data and chemical information. In *ISMB (Supplement of Bioinformatics)*, 468–477.

Ye, J.; Cheng, H.; Zhu, Z.; and Chen, M. 2013. Predicting positive and negative links in signed social networks by transfer learning. In *WWW*, 1477–1488.

Zhou, Z.; Li, X.; Wright, J.; Candès, E. J.; and Ma, Y. 2010. Stable principal component pursuit. *CoRR* abs/1001.2363.