A Scalable Proximity Measure for Link Prediction via Low-rank Matrix Estimation

Ye Liu, Zhisheng Wang, Jian Yin*
School of Information Science and Technology
Sun Yat-sen University
Guangzhou, China

Yan Pan
School of Software
Sun Yat-sen University
Guangzhou, China

Abstract—**Recent years, the link prediction problem in social network and other complex networks become a popular research field. One of the most significant task in link prediction is to design the proximity measure to calculate the similarities of the nodes in the network. The potential structure of the networks in the link prediction problem can be learned from the network data. In this paper, we propose a data-dependent proximity measure under the low-rank assumption in the social network and many other complex networks, then design a scalable matrix estimation algorithm to figure out the proximity measure. According to our experiment results, the proposed proximity measure can get competitive performance compared with other state-of-the-art methods and can be scalable for complex network link prediction.**

Keywords-link prediction; social network; proximity measure; low-rank estimation; data mining

I. INTRODUCTION

Link mining is a newly emerging research area which focuses on building predictive models of the linked data [17]. The study of the similarity relationship, which is also called link prediction, in the social network and other complex networks is one of the key link mining research field. In the link prediction problem, we can observe part of the relationship in the network and need to design a model to predict the unknown links [1]. The predict task can also be extended to dynamic network in which we know the status of the similarity in the network now and predict the new status of the network with the lapse of time.

In order to predict the similarity relationship of the network, topological methods are chosen to calculate the closeness between the network nodes and extended to use in the scalable systems [10]. The calculation method for the similarity of nodes in the network is called the proximity measure. Most of the supervised [12], semi-supervised [6] and unsupervised [1][2] link prediction framework use the proximity measure directly or as a part of the predicting algorithm. Thus, the most significant task in the link prediction framework is to design and use the proximity measure. With the rapid growth of the social network and other complex networks in the big data era, we can obtain much more data and the traditional proximity measure become computationally challenging [14]. The potential structure can be acquired by learning techniques from network data and can help to form more precise proximity measure. Besides, large network containing noisy data and missing values can also be handled by the learning algorithm of the proximity measure.

In this paper, we propose a novel scalable proximity measure for link prediction with the low-rank network structure assumption. The proposed proximity measure contain the low-rank and sparse constraint which can be learn form the network data for capturing the intrinsic structure and de-noising. Then we propose an optimization algorithm for solving the constraint with the Augmented Lagrangian Multiplier (ALM) scheme [19]. The experiment result shows that our novel scalable proximity measure for link prediction can get competitive performance compared to several state-of-the-art data-dependent proximity measure.

II. RELATED WORK

In the link prediction problem, designing an efficient and effective procedure of the proximity measure to capture the structure of the complex network is a normal and significant task [2]. The low-rank assumption of the network dataset is used for learning the intrinsic structure and de-noising in the previous work [4][9]. It has been proven that using both the sparse and low-rank constraint simultaneously in the matrix estimation is better than only considering the low-rank structure. The work in [4] provide a reliable method to learn the similarity relationship of the network for link prediction. The matrix factorization method which is a popular solution used in recommendation system can also be an acceptable scheme for learning under the low-rank assumption.

The differences between the previous methods and our proposed method can be both in the formulation of the proximity measure and in the optimization procedure. Besides the low-rank and sparse constraint, we select the hinge loss with additional equation constraint. Furthermore, the ALM [19] is chosen as the optimization algorithm which can alternatively update the matrix variances with the closed-form solution.

III. OUR METHOD

A. Problem Formulation

In the link prediction problem, we define V is the set of nodes and E is the set of edges in the network, so the network can be described as graph $G=(V,E)$. Then the similarity relationship can be denoted as the adjacency matrix A, and the element of the adjacency matrix A can be defined as $A[x,y]$. The element $A[x,y]$ equals 1 when node x and node y have similar relationship, and equals 0 if the relationship between nodes x and y is dissimilar or unknown.

The proximity measure is to calculate the probability of the similarity relationship between nodes in the network. In many real world network, the similarity relationship of the network is aggregating by groups like clustering. Thus, the

---

*Corresponding Author: Jian Yin (issjyin@mail.sysu.edu.cn)

low-rank assumption can be used to be the constraint $rank(S)$. Moreover, the adjacency matrix of graph G has sparsity that can be the constraint $||E||_0$ which is called the l_0-norm. The relation between the input adjacency matrix A and the target matrix S can be defined as the hinge loss [11] $||B||_+$ which is also used as the loss function in the support vector machine. Then the optimization problem of the proximal measure can be formulated as below:

$$\min_{S,E,B} \lambda_1 rank(S) + \lambda_2 ||E||_0 + \lambda_3 ||B||_+$$
$$s.t \quad S = E$$
$$B = I_{\frac{1}{4}} - (A - I_{\frac{1}{2}}) \circ (D - I_{\frac{1}{2}})$$
$$D = S. \qquad (1)$$

In the formulation, we design three additional constraint beside the constraint $rank(S)$, $||E||_0$ and $||B||_+$. The equation $A=S+E$ defines an input adjacent matrix A is combined with a low-rank matrix S and sparse matrix E. The other two equation constraints are used to define the similarity relationship and make the calculation convenient.

However, solving the constraint $rank(S)$ and constraint $||E||_0$ is NP-hard problem. In order to optimize the objective, the trace norm $||S||_*$ and the l_1-norm $||E||_1$ are always used to replace the constraint $rank(S)$ and $||E||_0$. The formulation can be define as below:

$$\min_{S,E,B} \lambda_1 ||S||_* + \lambda_2 ||E||_1 + \lambda_3 ||B||_+$$
$$s.t \quad S = E$$
$$B = I_{\frac{1}{4}} - (A - I_{\frac{1}{2}}) \circ (D - I_{\frac{1}{2}})$$
$$D = S. \qquad (2)$$

Thus, the optimization problem can be solved by the the Augmented Lagrangian Multiplier framework [19] and all of the constraint $||S||_*$, $||E||_1$ and $||B||_+$ have closed-form solution.

B. Optimization Algorithm

Now, we focus on the detail procedure to solve the above optimization formulation which receives the adjacent matrix A of graph G as the input and outputs the similarity matrix S.

Algorithm: A Novel Scalable Proximity Measure for Link Prediction via Low-rank Matrix Estimation Solving by ALM Framework

---

Input: $\lambda_1, \lambda_2, \lambda_3, A \in \mathcal{R}^{m \times m}$.
Initialize:All zeros for S, E, B, D, iter=0,
$\quad \mu = 10^{-8}, \rho = 1.2, \epsilon = 10^{-8}$.
  Repeat
    iter = iter + 1.
    Updating S.
    Updating E.
    Updating B.
    Updating D.
    Updating Y_1, Y_2, Y_3.
    $\mu \leftarrow \min(\rho\mu, 10^{10})$.
  Until
    $||S - E||_\infty \leq \epsilon$.
    $||B - [I_{\frac{1}{4}} - (A - I_{\frac{1}{2}}) \circ (D - I_{\frac{1}{2}})]||_\infty \leq \epsilon$.
    $||D - S||_\infty \leq \epsilon$.
Output: S.

We design the algorithm above based on the Augmented Lagrangian Multiplier (ALM) scheme, and the optimization problem can be solved by alternating updating procedure until the result meets the condition of convergence.

Since the alternating updating procedure is the most significant part of the proximity measure, we figure out the step-by-step updating rules below. Firstly, the corresponding Lagrangian function is defined according to the ALM framework below:

$$L(S, E, B, D, Y_1, Y_2, Y_3; \mu)$$
$$= \lambda_1 ||S||_* + \lambda_2 ||E||_1 + \lambda_3 ||B||_+$$
$$+ <Y_1, S - E> + \frac{\mu}{2} ||S - E||_F^2$$
$$+ <Y_2, B - [I_{\frac{1}{4}} - (A - I_{\frac{1}{2}}) \circ (D - I_{\frac{1}{2}})]>$$
$$+ \frac{\mu}{2} ||B - [I_{\frac{1}{4}} - (A - I_{\frac{1}{2}}) \circ (D - I_{\frac{1}{2}})]||_F^2$$
$$+ <Y_3, D - S> + \frac{\mu}{2} ||D - S||_F^2 \qquad (3)$$

Then the each of the matrix S, E, B, D, Y_1, Y_2, Y_3 can be updating repeatedly when others are fixed according to our proposed algorithm. The detail derivation of the updating rules are showed below.

Updating S:

When other matrix variables are known, the procedure of updating the matrix S can be:

$$\min_S \lambda_1 ||S||_*$$
$$+ <Y_1, S - E> + \frac{\mu}{2} ||S - E||_F^2$$
$$+ <Y_3, D - S> + \frac{\mu}{2} ||D - S||_F^2$$
$$\Leftrightarrow \min_S \lambda_1 ||S||_*$$
$$+ \frac{\mu}{2} ||S - E + \frac{Y_1}{\mu}||_F^2 + \frac{\mu}{2} ||D - S + \frac{Y_3}{\mu}||_F^2$$
$$\Leftrightarrow \min_S \lambda_1 ||S||_*$$
$$+ \frac{\mu}{2} ||S - E + \frac{Y_1}{\mu}||_F^2 + \frac{\mu}{2} ||S - D - \frac{Y_3}{\mu}||_F^2$$
$$\Leftrightarrow \min_S \lambda_1 ||S||_*$$
$$+ \mu ||S - \frac{1}{2}(E + D - \frac{Y_1}{\mu} + \frac{Y_3}{\mu})||_F^2 \qquad (4)$$

The closed-form solution of this formulation can be calculated by the shrinkage operator [19] with the singular value threshold method [18].

Updating E:

The subproblem of calculating the matrix E is similar to the procedure of updating the matrix S. From the Lagrangian function, we have:

$$\min_E \lambda_2 ||E||_1$$
$$+ <Y_1, S - E> + \frac{\mu}{2} ||S - E||_F^2$$
$$\Leftrightarrow \min_E \lambda_2 ||E||_1$$
$$+ \frac{\mu}{2} ||S - E + \frac{Y_1}{\mu}||_F^2$$
$$\Leftrightarrow \min_E \lambda_2 ||E||_1$$
$$+ \frac{\mu}{2} ||E - S - \frac{Y_1}{\mu}||_F^2$$
$$\Leftrightarrow \min_E \lambda_2 ||E||_1$$
$$+ \frac{\mu}{2} ||E - (S + \frac{Y_1}{\mu})||_F^2 \qquad (5)$$

The formulation of updating E also has the closed-form solution. Let $S_\mu(X)=max(X+\mu,0)+min(X-\mu,0)$, the closed-form solution is $E=S_\eta(S+Y_1/\mu)$ and $\eta=\lambda_2/\mu$.

Updating B:

The hinge loss is used as a constraint in the procedure of updating matrix B as the constraint penalty between the original input matrix A and the expectation matrix D:

$$
\begin{aligned}
\min_B \ & \lambda_3||B||_+ \\
& + <Y_2, B-[I_{\frac{1}{4}}-(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})]> \\
& + \frac{\mu}{2}||B-[I_{\frac{1}{4}}-(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})]||_F^2 \\
\Leftrightarrow \min_B \ & \lambda_3||B||_+ \\
& + \frac{\mu}{2}||B-[I_{\frac{1}{4}}-(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})]+\frac{Y_2}{\mu}||_F^2 \\
\Leftrightarrow \min_B \ & \lambda_3||B||_+ \\
& + \frac{\mu}{2}||B-[I_{\frac{1}{4}}-(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})-\frac{Y_2}{\mu}]||_F^2 \quad (6)
\end{aligned}
$$

Here, according to the formulation above, we define the matrix $C=I_{1/4}-(A-I_{1/2})\circ(D-I_{1/2})-Y_2/\mu$ and the notation $\circ$ means the Hadamard product. When other variances are fixed, the element of the matrix B and C can be defined as $B[x,y]$ and $C[x,y]$. Thus, we can get $B[x,y]=C[x,y]$ if $C[x,y]<0$, and otherwise $B[x,y]=max(0,C[x,y]-\lambda_3/\mu)$.

Updating D:

The matrix D is the auxiliary variable in the updating procedure and we can calculate the derivative when setting the formulation to zero. The formulation which is related to the matrix D is below:

$$
\begin{aligned}
\min_D \ & <Y_2, B-[I_{\frac{1}{4}}-(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})]> \\
& + \frac{\mu}{2}||B-[I_{\frac{1}{4}}-(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})]||_F^2 \\
& + <Y_3, D-S> + \frac{\mu}{2}||D-S||_F^2 \\
\Leftrightarrow \min_D \ & \frac{\mu}{2}||B-[I_{\frac{1}{4}}-(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})]+\frac{Y_2}{\mu}||_F^2 \\
& + \frac{\mu}{2}||D-S+\frac{Y_3}{\mu}||_F^2 \\
\Leftrightarrow \min_D \ & \frac{\mu}{2}||B-I_{\frac{1}{4}}+(A-I_{\frac{1}{2}})\circ(D-I_{\frac{1}{2}})+\frac{Y_2}{\mu}||_F^2 \\
& + \frac{\mu}{2}||D-S+\frac{Y_3}{\mu}||_F^2 \\
\Leftrightarrow \min_D \ & \frac{\mu}{2}||B-I_{\frac{1}{4}}+(A-I_{\frac{1}{2}})\circ D-(A-I_{\frac{1}{2}})\circ I_{\frac{1}{2}}+\frac{Y_2}{\mu}||_F^2 \\
& + \frac{\mu}{2}||D-S+\frac{Y_3}{\mu}||_F^2 \\
\Leftrightarrow \min_D \ & \frac{\mu}{2}||(A-I_{\frac{1}{2}})\circ D-[I_{\frac{1}{4}}+(A-I_{\frac{1}{2}})\circ I_{\frac{1}{2}}-B-\frac{Y_2}{\mu}]||_F^2 \\
& + \frac{\mu}{2}||D-S+\frac{Y_3}{\mu}||_F^2 \quad (7)
\end{aligned}
$$

Let the element of the matrix D be $D[x,y]$, then the result of the matrix D is $D[x,y]=D_m[x,y]/D_z[x,y]$. Here, the matrix D_z and D_m is described below:

$$D_z = [(A-I_{\frac{1}{2}})\circ(A-I_{\frac{1}{2}})+I_1]\circ D \quad (8)$$

$$D_m = [I_{\frac{1}{4}}+(A-I_{\frac{1}{2}})\circ I_{\frac{1}{2}}-B-\frac{Y_2}{\mu}]\circ(A-I_{\frac{1}{2}})+(S-\frac{Y_3}{\mu}) \quad (9)$$

Updating Y_1, Y_2, Y_3:

The updating procedure of the matrix Y_1, Y_2 and Y_3 is an normal step in the ALM framework with the parameter μ which is initialized at the beginning of the algorithm.

IV. EXPERIMENTS

In this section, we implement our proposed proximity measure and use the evaluation metric to validate the performance of our algorithm by comparing it to other state-of-the-art methods on several real-world network datasets.

A. Datasets and Baselines

We select three real-world network datasets Facebook, Flickr and Climate to test the proximity measure for link prediction in the contrast experiment. The detail information is showed in the statistic table below:

TABLE I: STATISTICS OF DATASETS

Dataset	Number of Nodes	Number of Edges
Facebook	2790	7760
Flickr	4532	233828
Climate	2660	77392

Since the proximity measure for link prediction is widely used in the social network, we firstly choose the Facebook dataset [13] which contain the relationship of the users, then we select and filter one university from the whole dataset to the input adjacent matrix. Then we use the meta dataset from Flickr [3] and pick one of the attributes which can be aggregated in groups to form the relationship matrix. For verifying the effectiveness of our proposed proximity measure for link prediction, we choose the Climate dataset [5][15] for predicting the weather condition relationship to examine the adaptiveness to other complex network besides the traditional social network. In all the datasets, we form an input adjacent matrix A with notating existed relationship in the network for $A[x,y]=1$ and $A[x,y]=0$ otherwise.

The state-of-the-art methods [4][9] mentioned above is implemented carefully as the baselines. The work with the low-rank and sparse constraint [4] which is solved by the Incremental Proximal Descent (IPD) [8] algorithm is called the LRSP-IPD method in our paper. And we call the matrix factorization for low-rank [9] optimized by the Stochastic Gradient Descent (SGD) [7] method MF-SGD in short. Moreover, We implement our proposed proximity measure with a matrix divide-and-conquer framework as a kind of Generalized Nystrom Method mentioned in previous work [20][21][22] to boost the calculation speed and make the operation of the adjacent matrix scalable for large network.

B. Evaluation Measures

The AUC metric [16] which is also called the area under the receiver operating characteristic curve is the standard assessment method to evaluate the quality of the proximity measure in link prediction problem [2]. We randomly select ten percent of the nodes and the corresponding similarity relationship in the matrix as the input to the proximity measure and then run the algorithm ten times to get the average result with the AUC metric in the experiment.

C. Experimental Results

In the experiment, we use the coarse-to-fine method to select the parameters of all the proximity measures.

TABLE II: THE AUC RESULT ON THE FACEBOOK DATASET

Noise	5%	10%	15%	20%
LRSP-IPD ($\lambda_1=\lambda_2=\lambda_3=1$)	0.963	**0.912**	0.819	0.804
MF-SGD (k=55)	0.902	0.767	0.728	0.751
MF-SGD (k=65)	0.969	0.847	0.764	0.706
MF-SGD (k=75)	0.955	0.861	0.787	0.781
Our Method ($\lambda_1=1,\lambda_2=0,\lambda_3=0.5$)	0.964	0.910	**0.842**	0.804
Our Method ($\lambda_1=1,\lambda_2=10^{-6},\lambda_3=0.5$)	0.965	**0.912**	0.825	**0.811**
Our Method ($\lambda_1=1,\lambda_2=10^{-5},\lambda_3=0.5$)	**0.971**	0.902	0.828	0.810

TABLE III: THE AUC RESULT ON THE FLICKR DATASET

Noise	5%	10%	15%	20%
LRSP-IPD ($\lambda_1=\lambda_2=\lambda_3=1$)	0.967	0.921	0.882	0.813
MF-SGD (k=55)	0.950	0.860	0.813	0.716
MF-SGD (k=65)	0.946	0.857	0.804	0.751
MF-SGD (k=75)	0.898	0.874	0.806	0.749
Our Method ($\lambda_1=1,\lambda_2=0,\lambda_3=0.5$)	**0.970**	0.924	**0.886**	**0.815**
Our Method ($\lambda_1=1,\lambda_2=10^{-6},\lambda_3=0.5$)	**0.970**	**0.926**	0.884	**0.815**
Our Method ($\lambda_1=1,\lambda_2=10^{-5},\lambda_3=0.5$)	0.969	0.924	0.882	**0.815**

TABLE IV: THE AUC RESULT ON THE CLIMATE DATASET

Noise	5%	10%	15%	20%
LRSP-IPD ($\lambda_1=\lambda_2=\lambda_3=1$)	0.964	**0.922**	0.865	**0.810**
MF-SGD (k=55)	0.954	0.806	0.756	0.713
MF-SGD (k=65)	0.957	0.798	0.766	0.728
MF-SGD (k=75)	0.942	0.919	0.763	0.725
Our Method ($\lambda_1=1,\lambda_2=0,\lambda_3=0.5$)	0.965	0.921	0.868	**0.810**
Our Method ($\lambda_1=1,\lambda_2=10^{-6},\lambda_3=0.5$)	0.966	0.920	0.869	0.808
Our Method ($\lambda_1=1,\lambda_2=10^{-5},\lambda_3=0.5$)	**0.968**	0.920	**0.873**	0.806

From the experiment result above, we see that our proposed method can gain competitive result under the Gaussian noise condition from five percent to twenty percent.

V. CONCLUSION

In this paper, we develop a novel scalable proximity measure for link prediction with low-rank constraint and propose an algorithm to calculate the result based on the ALM scheme. Extensive experiments on several real-world network datasets for link prediction show that our proposed proximity measure which is used for calculating the similarity relationship in the link prediction problem can achieve competitive performance compared with other date-of-the-art methods. And the optimization problem of the proximity measure can be solved robustly and effectively by the proposed algorithm.

REFERENCES

[1] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031.

[2] Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[3] McAuley J, Leskovec J. Image labeling on a network: using social-network metadata for image classification[M]. Computer Vision‐ECCV 2012. Springer Berlin Heidelberg, 2012: 828-841.

[4] Richard E, Savalle P A, Vayatis N. Estimation of simultaneously sparse and low rank matrices[J]. arXiv preprint arXiv:1206.6474, 2012.

[5] Davis D, Lichtenwalter R, Chawla N V. Multi-relational link prediction in heterogeneous information networks[C]. Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011: 281-288.

[6] Raymond R, Kashima H. Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs[M]. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2010: 131-147.

[7] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[8] Bertsekas D P. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey[J]. Optimization for Machine Learning, 2011, 2010: 1-38.

[9] Hsieh C J, Chiang K Y, Dhillon I S. Low rank modeling of signed networks[C]. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 507-515.

[10] Song H H, Cho T W, Dave V, et al. Scalable proximity estimation and link prediction in online social networks[C]. Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009: 322-335.

[11] Gentile C, Warmuth M K. Linear hinge loss and average margin[C]. NIPS. 1998, 11: 225-231.

[12] Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]. SDM'06: Workshop on Link Analysis, Counter-terrorism and Security. 2006.

[13] Traud A L, Mucha P J, Porter M A. Social structure of facebook networks[R]. arXiv, cs.SI:1102.2166, 2011.

[14] Sarkar P, Moore A W, Prakash A. Fast incremental proximity search in large graphs[C]. Proceedings of the 25th international conference on Machine learning. ACM, 2008: 896-903.

[15] Pelan A, Steinhaeuser K, Chawla N V, et al. Empirical comparison of correlation measures and pruning levels in complex networks representing the global climate system[C]. Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE, 2011: 239-245.

[16] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern recognition, 1997, 30(7): 1145-1159.

[17] Getoor L, Diehl C P. Link mining: a survey[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12.

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

[19] Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices[J]. arXiv preprint arXiv:1009.5055, 2010.

[20] Mackey L W, Talwalkar A, Jordan M I. Divide-and-Conquer Matrix Factorization[C]. NIPS. 2011: 1134-1142.

[21] Williams C, Seeger M. Using the Nyström method to speed up kernel machines[C]. Advances in Neural Information Processing Systems 13. 2001.

[22] Goreinov S A, Tyrtyshnikov E E, Zamarashkin N L. A theory of pseudoskeleton approximations[J]. Linear Algebra and Its Applications, 1997, 261(1): 1-21.