On Learning Mixed Community-specific Similarity Metrics for Cold-start Link Prediction

Linchuan Xu
The Hong Kong Polytechnic University
Kowloon, Hong Kong
cslcxu@comp.polyu.edu.hk

Xiaokai Wei
University of Illinois at Chicago
Chicago, IL,USA
weixiaokai@gmail.com

Jiannong Cao
The Hong Kong Polytechnic University
Kowloon, Hong Kong
csjcao@comp.polyu.edu.hk

Philip S. Yu
University of Illinois at Chicago
Chicago, IL,USA
psyu@uic.edu

ABSTRACT

We study the cold-start link prediction problem where edges between vertices is unavailable by learning vertex-based similarity metrics. Existing metric learning methods for link prediction fail to consider communities which can be observed in many real-world social networks. Because different communities usually exhibit different intra-community homogeneities, learning a global similarity metric is not appropriate. In this paper, we thus propose to learn community-specific similarity metrics via joint community detection. Experiments on three real-world networks show that the intra-community homogeneities can be well preserved, and the mixed community-specific metrics perform better than a global similarity metric in terms of prediction accuracy.

1. INTRODUCTION

In the cold-start link prediction problem where edges between vertices are unavailable, we can only rely on vertex-based similarities. Although learning vertex-based similarity metrics has been studied before for link prediction [3, 2], they fail to consider communities within the networks. However, plenty of studies [1] have revealed that vertices of real world social networks usually tend to form clusters or communities. For example, in academic social networks, co-authorships mostly occur within the same research field. Accordingly, real world social networks actually display a mixture of intra-community homogeneities.

Moreover, it is observed communities tend to overlap in real-world social networks [4] as a result of multiple memberships of a single vertex. Instead of uncovering the overlapping characteristics so as to partition networks, in this paper, we thus propose to learn a mixture of community-specific similarity metrics via joint community detection.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054269

.

The Studied Problem: Given a network $G(V_1, V_2, F_1, E)$, where V_1 and V_2 are sets of vertices, F_1 is the set of attributes, and E is the set of edges among V_1, the objective is to learn attribute-based similarity metrics to infer interactions involving V_2, and $V_2 \cap V_1 = \emptyset$.

The Similarity Metric: The studied similarity metric is weighted attribute similarity (WAS) denoted as follows:

$$\text{WAS}(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{w}) = \frac{\boldsymbol{w}^\top(\boldsymbol{x}_i \otimes \boldsymbol{x}_j)}{\boldsymbol{w}^\top(\boldsymbol{x}_i \oplus \boldsymbol{x}_j)}, \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ denotes the vector of attribute weights, d is the number of attributes, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote the attribute vectors of vertex i and j, respectively, and $x_{ij} = 1$ if vertex i contains attribute j or $x_{ij} = 0$ otherwise. $\otimes$ is a binary function that returns a vector with each element equal to logical "and" of two corresponding elements of input vectors, and $\oplus$ returns a vector with each element equal to logical "or" of two corresponding elements of input vectors. In the rest of paper, we let $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denote $WAS(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

2. GLOBAL PROBABILITY MODEL

The probability of a link between two vertices estimated on the similarity is defined as follows:

$$p(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{w}) = \frac{1}{1 + \exp\{-s(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{w})\}}, \tag{2}$$

To learn an appropriate $\boldsymbol{w}$, both small probabilities of existing links and large ones of non-existing links should be penalized. The logistic loss is used to perform the penalty. With a F_2-norm used as egularization on $\boldsymbol{w}$ to control the complexity, the loss function is formulated as $L(\boldsymbol{w}) =$

$$-\sum_{(i,j)\in E} \log(p(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{w})) - \sum_{(h,k)\notin E} \log(1 - p(\boldsymbol{x}_h, \boldsymbol{x}_k|\boldsymbol{w})) + \eta||\boldsymbol{w}||^2, \tag{3}$$

where $\eta \in \mathbb{R}$. Minimizing Eq. (3) is a convex optimization problem, and we may solve it by gradient descent.

3. MIXED PROBABILITY MODEL

The mixed probability model extends the global one by considering the communities which the vertices belong to and which community-specific similarity metric should be

Global probability Model	Mixed community-weighted probability Model					
	Database	1&2	**Machine learning**	2&3	**Networking**	1&3
ranking	**XML**	ranking	**tracking**	query	**internet**	tracking
retrieval	**TREC**	clustering	**shape**	distributed	**routing**	localization
routing	**language**	privacy	**clustering**	stream	**mobility**	predictive
internet	**ranking**	stream	**face**	streams	**traffic**	informative
clustering	**test**	matrix	**stereo**	peer-to-peer	**packet**	scene

Table 1: Top 5 features from probability models

used. Considering the fact that each vertices may belong to multiple communities as a result of overlapping communities in real world social networks, we adopt the soft community assignment, which means that every vertex is assigned to a certain community with an probability. This soft community assignment can be achieved by the commonly used multi-class logistic regression model defined as follows:

$$\pi_a(\boldsymbol{x}_i) = \frac{\exp\{\boldsymbol{v}_a^\top \boldsymbol{x}_i + b_a\}}{\sum_{t=1}^{A} \exp\{\boldsymbol{v}_t^\top \boldsymbol{x}_i + b_t\}}, \qquad (4)$$

where $\pi_a(\boldsymbol{x}_i)$ is the probability that vertex x_i belongs to community a, $\boldsymbol{v}_a \in \mathbb{R}^d$ is the vector of centroid of community a, A is the number of communities, and $b \in \mathbb{R}$ is a bias term.

Then to compute the probability of a link, every possible community of each vertex should be considered. Two vertices may belong to the same community, and may belong to different communities as well. To facilitate the computation of the probabilities that two vertices belong to different communities, we introduce a special kind of community, which is the overlapping of two normal communities. The special community captures commonalities of the two communities, and we assign a corresponding similarity metric to it.

Hence, the probability of a link using a mixture of community-specific similarities is quantified as follows:

$$p(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{W}) = \sum_a \sum_b \pi_a(\boldsymbol{x}_i) \pi_b(\boldsymbol{x}_j) p(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{w}_{ab}), \qquad (5)$$

where $\boldsymbol{W}$ is a set of community-specific metrics, $p(\boldsymbol{x}_i, \boldsymbol{x}_j | \boldsymbol{w}_{ab})$ is the probability of a link between vertex x_i from community a and vertex x_j from community b. Since $\boldsymbol{w}_{ab}$ when $a \neq b$ is the similarity metric for the community resulting from overlapping of community a and b, $\boldsymbol{w}_{ba}$ is equal to $\boldsymbol{w}_{ab}$.

The logistic loss is employed to formulate the loss function like Eq. (3). The different is that the objective has to be optimized w.r.t W and community centroids. We can solve the variables iteratively by gradient descent.

4. EMPIRICAL EVALUATION

Experiment Settings: Regularization coefficients are set as 1, backtracking line search is employed to learn decent rates, and relative loss of 0.001 is set as the converge criterion. For the mixed probability model, the number of communities is set as the ground truth number, and centroids are pre-trained by k-means clustering on attributes vectors. The global probability model is used as the baseline since no existing methods are applicable to the studied problem.

AUC	Blog	PubMed	DBLP
Global Probability Model	60.32	85.12	88.81
Mix Probability Model	**63.98**	**88.02**	**91.08**

Table 2: Performance comparison in link prediction

Datasets: One BlogCatalog [7] social network (5567 vertices, 21775 friendships, and 8675 attributes) is sampled from four communities, Art, Technology, Development and Growth, and Finance. One citation network (19717 papers, 44338 citations, and 500 words) is PubMed Diabetes [5]. Another DBLP [6] citation network (11512 papers, 11996 citations, and 8172 words from abstract) is sampled from popular conferences, including SIGMOD, VLDB, ICDE from database, AAAI, ICML, NIPS from machine learning, and SIGCOMM, GLOBECOM, INFOCOM from networking during the time span from the year 2000 to 2012.

4.1 Overview of Community-Specific Metrics

Top 5 attributes ranked according to the learned weights on the DBLP dataset are presented in Table 1. Firstly, it shows for each community, the top 5 attributes are all domain attributes. Hence, the community-specific metrics preserve the community structure well. Secondly, attributes of the global model is a mixture of attributes from the three communities. This may suggest that a global similarity metric is not appropriate as top attributes of one domain may mislead the inference of new links in another domain, which is demonstrated in the following link prediction.

4.2 Link Prediction

We use stratified 90% links as training links, and the rest as test links. Also, the same number of negative links are randomly sampled for the evaluation purpose. The performance on AUC score is presented in Table 2, where numbers have been multiplied by 100%, which shows the mixed probability model performs better in all datasets.

5. REFERENCES

[1] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[2] N. Li, X. Feng, S. Ji, and K. Xu. Modeling relationship strength for link prediction. In *Intelligence and Security Informatics*, pages 62–74. Springer, 2013.

[3] L. Lü and T. Zhou. Role of weak ties in link prediction of complex networks. In *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, pages 55–58. ACM, 2009.

[4] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[5] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[6] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[7] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *the 10th IEEE International Conference on Data Mining series (ICDM2010)*, Sydney, Australia, December 14 - 17 2010.