Joint Link Prediction and Multi-Label learning in Heterogeneous networks

Srichandra Chilappagari and Mentor: Sumit Negi

Xerox Research Centre India

1 Introduction

Most real world systems usually consist of a large number of interacting, multi-typed components, such as human social activities, communication systems, and biological networks. Such complex systems can be called as heterogeneous information networks where the interacting components constitute interconnected networks(graphs). In general, whenever a graph has multiple node types or multiple link types, we call it as Heterogenous graph. Consider real social networks where users are connected via different types of social ties. For example, in a mobile communication network, the relationship types could include family, colleagues, and friends. In Flickr, photos are linked together via users, groups, tags and comments. It is well known that the different types of links have essentially different influence between the entities. Most of these networks are highly dynamic, they grow and change quickly over time through the addition of new edges. Basic computational problem lying this evolution is Link prediction(LP), that is, predicting the formation of links in a network in the future or predicting the missing links in a network. Problem of LP in heterogeneous graphs with multiple link types has been considered in [1], [2], [3]. In social networks, different entities (people) are associated with multiple groups and interests. In drug discovery, one molecular drug can bind with multiple protein targets, and researchers would like to predict which protein targets that one chemical compound can bind with in order to discover new drugs for a certain disease. Multi-label learning in heterogeneous networks[4][5][6] has become increasingly important in recent years, where each example can be associated with multiple labels simultaneously.

Link prediction (LP) and multi-label learning (MLL) on graphs are two important and challenging problems which can be applied to diverse fields. These two problems are inherently correlated and appear concurrently. So far they have been mostly considered to be unrelated problems. For example, in a protein-protein interaction graph, the partially known protein function annotations(labels) have been exploited as a major source of information for predicting edges [7]. In [8], functions of protein are determined by the their interaction activity. In [9], problems of LP and MLL are solved alternatively

and output of one method is used to facilitate the other. However, it is not trivial to determine how frequently and to what extent information should be exchanged between LP and MLL. We might even end up propagating errors from the result of one method to other. These alternating iterations between two tasks are not guaranteed to converge. We intend to develop an algorithm to solve these two problems on graphs jointly. Most people in social networks belong to multiple groups and involve in multiple types of activities with different degrees of engagement. This heterogeneity of the network is modelled in the links and labels of users in the connection graphs. By utilizing the richer structure and semantic information of the heterogeneous information networks, we try to solve the problems of LP and MLL jointly.

2 Our Contribution

In [10], a novel approach was introduced combining the problems of MLL and LP problems into a single joint objective function based on marginalized denoising framework [11][12][13]. Both the problems were phrased as instances of graph denoising and co-regularize the predictions with laplacian graph regularization. The objective function is jointly convex and can be optimized very efficiently. However ,this algorithm assumes all links to be of same type and thus limiting the application of this algorithm to homogeneous graphs. By utilizing the richer structure and semantic information of the heterogeneous information networks, we try to solve the problems of LP and MLL jointly in a heterogeneous graph.

3 Formulation:

We consider the problem of LP and MLL on a graph G = (G^1, G^2, ... G^m, Y), where $G^l \in \{0,1\}^{nxn}$ denote the partially observed relational graphs between nodes corresponding to link type l. $G^l_{ij} = 1$ iff we observe a link of type l between nodes i and j and $G^l_{ij} = 0$ otherwise. $Y \in \{0,1\}^{Kxn}$ where k is the number of labels. $Y_{ij} = 1$ iff node i is associated with label K and $Y_{ij} = 0$ otherwise. We denote the i^{th} column of Y, y_i as label vector of i^{th} node. We assume that labels of the first l nodes are given and labels of other nodes are unknown.

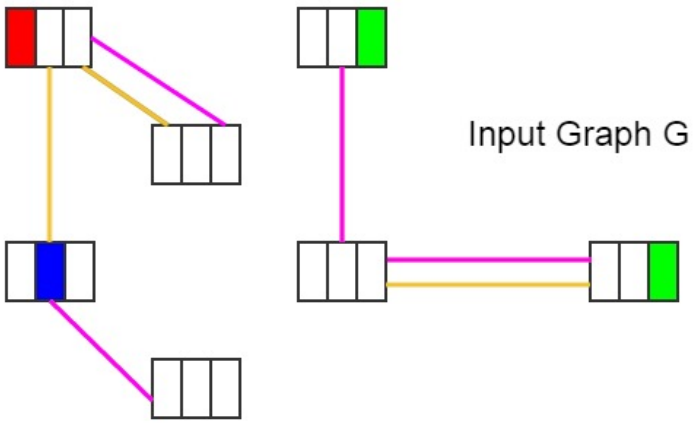

Figure 1: Input connected graph with maximum three labels on each node(Red, Blue and Green) and two link types(Pink and Yellow)

Without loss of generality, we assume two link types (m = 2) and three labels(K = 3) through out this paper. We assume that labels of first u nodes are given(albeit incomplete) and the other labels are unknown. See figure 1 for reference.

4 Approach:

We introduce K new nodes, one for each label and refer to these as label nodes and original nodes as data nodes. We then create links of all mtypes between the i^{th} data node and k^{th} label node if node carries a label k i.e. $Y_{ij} = 1$. See the figure 2 below for a schematic illustration.

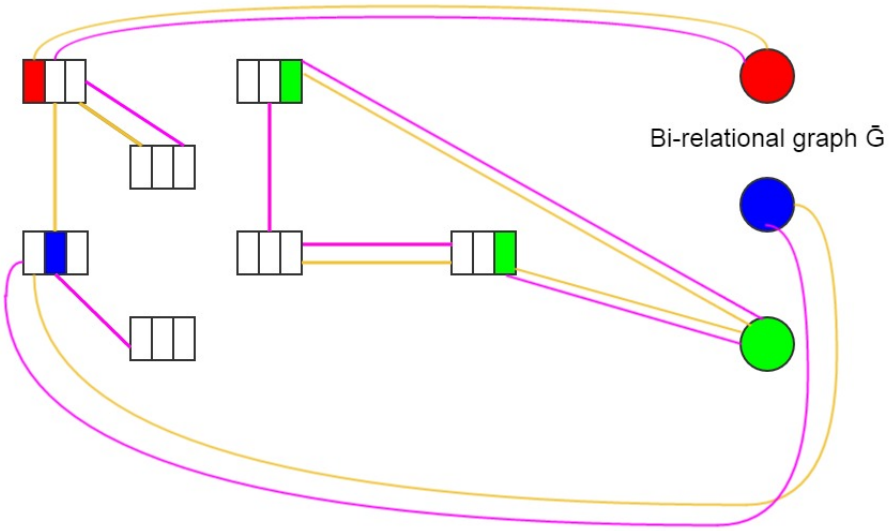

Figure 2: Augmented graph with three label nodes(Red, Blue and Green) and new links

We use $\bar{G}$ to denote the link matrix of the new graph. $\bar{G} \in \{0,1\}^{2n+kx2n+k}$ can be decomposed as

$$\bar{G}= \begin{bmatrix} \mathrm{G}^1 & S & \mathrm{Y}^T \\ \mathrm{S}^T & \mathrm{G}^2 & \mathrm{Y}^T \\ Y & Y & H \end{bmatrix}$$

G^1, G^2 are adjacency matrices of original graphs corresponding to link types 1 and 2 respectively. Y is the label matrix and H is the cosine similarity between observed labels. S captures the similarity between G^1 and G^2, which models the engagement between two nodes via different link types and nodes. We claim to perform a better job of LP and MLL by allowing information flow in both directions based on a marginalized denoising framework[14] and combining the problems into a single joint objective function[10]. See figure 3 for the output and figure 4 for the algorithm.

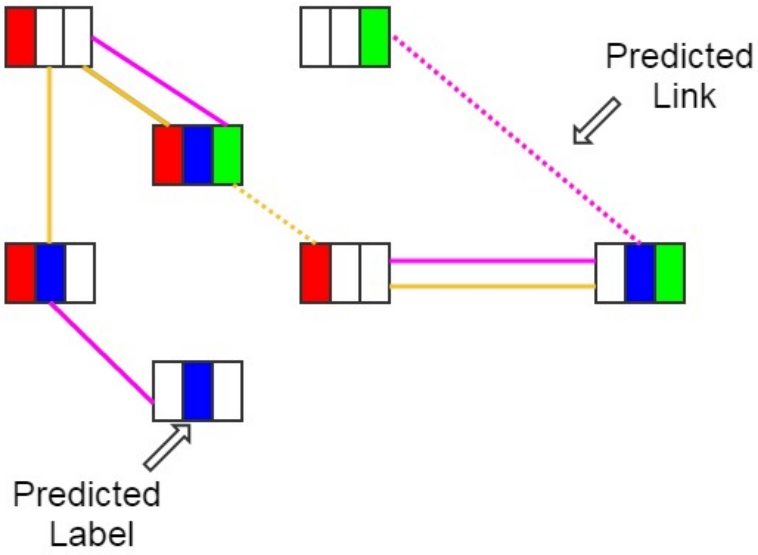

Figure 3: Output graph

Inputs: Y, Graphs G^1, G^2 , Parameters α, $\beta > 0$ and level of corruption: $0 < p < 1$
Output: Link prediction scores $\hat{G}^1$, $\hat{G}^2$ and label prediction scores $\hat{Y}$.

- Set $H_{ij} = \frac{y_i^T y_j}{\|y_i\| \|y_j\|}$, where $y_i \in \{0,1\}^{n \times 1}$ is the i^{th} column of matrix
- Set $S_{ij} = \frac{g_i^1 (g_j^2)^T}{\|g_i^1\| \|g_j^2\|}$, where $g_i^1 \in \{0,1\}^{1 \times n}$ is i^{th} row of matrix G^1 and $g_j^2 \in \{0,1\}^{1 \times n}$ is i^{th} row of matrix G^2.
- $\bar{G} = [G^1, S, Y^T; S^T, G^2, Y^T; Y^1, Y^2, H]$, L^1/ L^2 ← graph Laplacian of G^1/ G^2
- $Q = (1 - p)^2 \bar{G}\bar{G}^T + p(1 - p)\, diag(\bar{G}\bar{G}^T)$
- $P = (1 - p) * W\bar{G}\bar{G}^T$. Initialize $\hat{Y}^1$, $\hat{Y}^2$ and $W \in \{0,1\}^{(2n+k) \times (2n+k)}$ randomly.
- Let $\bar{G}_1$, $\bar{G}_2$, $\bar{G}_3$ represent the first n, second n and last K columns of matrix $\bar{G}$. Let W_1 , W_2, W_3 represent the first n, second n and last K rows of matrix W.
- $Z_1 = (\frac{\alpha}{4} \bar{G}_1 \bar{G}_1^T + \beta Q)^{-1}$
- $Z_2 = (\frac{\alpha}{4} \bar{G}_2 \bar{G}_2^T + \beta Q)^{-1}$
- $Z_3 = (\frac{\alpha}{4} \bar{G}_3 \bar{G}_3^T + \beta Q)^{-1}$
- $Z_4 = (L + \alpha I) - 1$
- $Z_{objective} = Tr(\hat{Y}^1 L^1 \hat{Y}^1) + Tr(\hat{Y}^2 L^2 \hat{Y}^2) + \alpha \| \hat{Y}^1 - 0.5(W_3 \bar{G}_1 + \bar{G}_3^T W_1^T) \|_F^2 + \alpha \| \hat{Y}^2 - 0.5(W_3 \bar{G}_2 + \bar{G}_3^T W_2^T) \|_F^2 + \beta Tr(WQW^T - 2(1 - p) W\bar{G}\bar{G}^T)$
- Repeat
- $W_3^1 \leftarrow (\frac{\alpha}{2}(\hat{Y}^1 - \bar{G}_3^T W_1^T)\bar{G}_1^T + \beta P_3) Z_1$
- $W_3^2 \leftarrow (\frac{\alpha}{2}(\hat{Y}^2 - \bar{G}_3^T W_2^T)\bar{G}_2^T + \beta P_3) Z_2$
- $W_3 \leftarrow \frac{1}{2}(W_3^1 + W_3^2)$
- $W_1 \leftarrow (\frac{\alpha}{2}((\hat{Y}^1)^T - \bar{G}_1^T W_3^T)\bar{G}_3^T + \beta P_1) Z_3$
- $W_2 \leftarrow (\frac{\alpha}{2}((\hat{Y}^2)^T - \bar{G}_2^T W_3^T)\bar{G}_3^T + \beta P_2) Z_3$
- $\hat{Y}^1 \leftarrow (\frac{\alpha}{2}(W_3 \bar{G}_1 + \bar{G}_3^T W_1^T) Z_4$
- $\hat{Y}^2 \leftarrow (\frac{\alpha}{2}(W_3 \bar{G}_2 + \bar{G}_3^T W_2^T) Z_4$
- Until convergence
- $\hat{Y} = \hat{Y}^1$ & $\hat{Y}^2$, $\hat{G}^1 = \frac{1}{2}(W_1 \bar{G}_1 + \bar{G}_1^T W_1^T)$, $\hat{G}^2 = \frac{1}{2}(W_2 \bar{G}_2 + \bar{G}_2^T W_2^T)$
- Return $\hat{Y}$, $\hat{G}^1$, $\hat{G}^2$

Figure 4: Algorithm

5 Experiments(WIP):

We intend to evaluate our algorithm on the real-world dataset extracted from the DBLP Bibliography data where:

- Nodes represent Authors
- Link type can be co-author relationship or shared community(journal, conference)
- Labels are the Research areas

References

[1] Darcy Davis, Ryan Lichtenwalter, and Nitesh V. Chawla. Multi-relational link prediction in heterogeneous information networks. *ASONAM*, pages 281–288, 2011.

[2] David Liben-Nowell and Jon Kleinbergr. The link prediction problem for social networks. *CIKM*, pages 556–559, 2003.

[3] Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring social ties across heterogenous networks. *WSDM*, pages 743–752, 2012.

[4] Xiangnan Kong, Bokai Cao, and Philip S. Yu. Multi-label classification by mining label and instance correlations from heterogeneous information networks. *KDD*, pages 614–622, 2013.

[5] Tanwistha Saha, Huzefa Rangwala, and Carlotta Domeniconi. Multi-label collective classification using adaptive neighborhoods. *ICML*, pages 427–432, 2012.

[6] Yang Zhou and Ling Liu. Activity-edge centric multi-label classification for mining heterogeneous information networks. *KDD*, pages 1276–1285, 2014.

[7] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J. Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F. Greenblatt, and Mark Gerstein.

[8] Yiannis A. I. Kourmpetis, Aalt D. J. van Dijk, Marco C. A. M. Bink, Roeland C. H. J. van Ham, and Cajo J. F. ter Braak. Bayesian markov random field analysis for protein function prediction based on network data. *PLoS One*, 2010.

[9] Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor. Combining collective classification and link prediction. *ICDM*, pages 381–386, 2007.

[10] Zheng Chen, Minmin Chen, Kilian Q. Weinberger, and Weixiong Zhang. Marginalized denoising for link prediction and multi-label learning. *AAAI*, pages 1707–1713, 2015, Febraury.

[11] Zheng Chen and Weixiong Zhang. A marginalized denoising method for link prediction in relational data. *SIAM International Conference on Data Mining*, 2014. doi: http://dx.doi.org/10.1137/1.9781611973440.34.

[12] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha.

[13] Minmin Chen, Alice Zheng, and Kilian Weinberger. Marginalized denoising for link prediction and multi-label learning. *ICML*, 28:1274–1282, 2013.

[14] Minmin Chen, Kilian Weinberger, Fei Sha, and Yoshua Bengio. Marginalized denoising auto-encoders for nonlinear representations. *ICML*, pages 767–774, 2012.