Supplementary File for *Graph Regularized Meta-path Based Transductive Regression in Heterogeneous Information Network*

Mengting Wan* Yunbo Ouyang* Lance Kaplan† Jiawei Han*

1 Theorem Proofs

Recall the constraint optimization problem

$$\begin{aligned}\min\ \mathbf{J}(\mathbf{w};\mathbf{f}) =& 2\sum_{k=1}^{K} w_k \mathbf{f}^T\mathcal{L}^{(k)}\mathbf{f} + \alpha_1(\mathbf{f}_L-\mathbf{y}_L)^T(\mathbf{f}_L-\mathbf{y}_L)\\ &+\alpha_2(\mathbf{f}_U-\tilde{\mathbf{y}}_U)^T\Sigma^{-1}(\mathbf{f}_U-\tilde{\mathbf{y}}_U).\end{aligned}$$

subject to

$$\sum_{k=1}^{K}\exp(-w_k)=1.$$

We provide following theorems and proofs:

Theorem 1.1. *Suppose* $\mathbf{f}$ *is fixed, the objective problem* $\mathbf{J}(\mathbf{w};\mathbf{f})$ *with constraint function* $\delta(\mathbf{w})=0$ *is a convex optimization problem. The global optimal solution is given by*

$$w_k = -log\left(\frac{\mathbf{f}^T\mathcal{L}^{(k)}\mathbf{f}}{\sum_{k=1}^{K}\mathbf{f}^T\mathcal{L}^{(k)}\mathbf{f}}\right). \tag{1.1}$$

Proof. The Hessian matrix for $\delta(\mathbf{w})$ is a diagonal matrix with elements $\{\exp(-w_k)\}_{k=1}^{K}>0$. Besides when $\mathbf{f}$ is fixed, the objective function is a linear function. Hence both the objective function and the constraint function are convex, therefore this problem is a convex optimization problem.

Suppose

$$J_L(\mathbf{w};\mathbf{f}) = J(\mathbf{w};\mathbf{f}) + \lambda\left(\sum_{k=1}^{K}\exp(-w_k)-1\right),$$

and

$$\mathbf{t}=(t_1,...,t_K)^T,\ t_k=\exp(-w_k), k=1,...,K.$$

We have

$$\begin{aligned}J_L(\mathbf{t};\mathbf{f}) =& 2\sum_{k=1}^{K}(-log(t_k))\mathbf{f}^T\mathcal{L}^{(k)}\mathbf{f} + \alpha_1(\mathbf{f}_L-\mathbf{y}_L)^T(\mathbf{f}_L-\mathbf{y}_L)\\ &+\alpha_2(\mathbf{f}_U-\tilde{\mathbf{y}}_U)^T\Sigma^{-1}(\mathbf{f}_U-\tilde{\mathbf{y}}_U)+\lambda\left(\sum_{k=1}^{K}t_k-1\right)\end{aligned}$$

From $\frac{\partial}{\partial t_k}J_L(\mathbf{t};\mathbf{f})=0$, we have

$$2\mathbf{f}^T\mathcal{L}^{(k)}\mathbf{f}=\lambda t_k.$$

Since $\sum_{k=1}^{K}\exp(-w_k)=\sum_{k=1}^{K}t_k=1$, by plugging in we have the above solution.

Theorem 1.2. *Suppose* $\mathbf{w}$ *is fixed, the objective problem* $\mathbf{J}(\mathbf{w};\mathbf{f})$ *is a convex optimization problem. The global optimal solution is given by solving the following linear system:*

(1.2)

$$\begin{aligned}(2\sum_{k=1}^{K}w_k+\alpha_1)\mathbf{f}_L &= 2\sum_{k=1}^{K}w_k(\mathbf{S}_{11}^{(k)}\mathbf{f}_L+\mathbf{S}_{12}^{(k)}\mathbf{f}_U)+\alpha_1\mathbf{y}_L;\\ (2\sum_{k=1}^{K}w_k+\alpha_2\Sigma^{-1})\mathbf{f}_U &= 2\sum_{k=1}^{K}w_k(\mathbf{S}_{21}^{(k)}\mathbf{f}_L+\mathbf{S}_{22}^{(k)}\mathbf{f}_U)+\alpha_2\Sigma^{-1}\tilde{\mathbf{y}}_U.\end{aligned}$$

where

$$\mathbf{S}^{(k)}=\begin{bmatrix}\mathbf{S}_{11}^{(k)} & \mathbf{S}_{12}^{(k)}\\ \mathbf{S}_{21}^{(k)} & \mathbf{S}_{22}^{(k)}\end{bmatrix},$$

is partitioned according labeled and unlabeled objects.

Proof. When $\mathbf{w}$ is fixed, $\mathbf{J}(\mathbf{w};\mathbf{f})$ is a non-negative quadratic function which has Hessian matrix

$$H = 4\sum_{k=1}^{K}w_k\mathcal{L}^{(k)}+\begin{bmatrix}2\alpha_1 I_n & O\\ O & 2\alpha_2\Sigma^{-1}\end{bmatrix} \tag{1.3}$$

For any $\mathbf{f}$ and any k, $\mathbf{f}^T\mathcal{L}^{(k)}\mathbf{f}\geqslant 0$, therefore Hessian matrix is non-negative, which means $\mathbf{J}(\mathbf{w};\mathbf{f})$ is convex when $\mathbf{w}$ is fixed.

Given $\mathbf{w}$, we can derive $\mathbf{f}$. If

$$\mathbf{S}^{(k)}=\begin{bmatrix}\mathbf{S}_{11}^{(k)} & \mathbf{S}_{12}^{(k)}\\ \mathbf{S}_{21}^{(k)} & \mathbf{S}_{22}^{(k)}\end{bmatrix},$$

we have

$$\mathcal{L}^{(k)}=\begin{bmatrix}\mathbf{I}_n-\mathbf{S}_{11}^{(k)} & -\mathbf{S}_{12}^{(k)}\\ -\mathbf{S}_{21}^{(k)} & \mathbf{I}_m-\mathbf{S}_{22}^{(k)}\end{bmatrix}.$$

*University of Illinois at Urbana-Champaign
†U.S. Army Research Laboratory

Thus

$$\Omega(\mathbf{w};\mathbf{f}) = 2\sum_{k=1}^{K} w_k \begin{bmatrix}\mathbf{f}_L^T & \mathbf{f}_U^T\end{bmatrix} \begin{bmatrix}\mathbf{I}_n - \mathbf{S}_{11}^{(k)} & -\mathbf{S}_{12}^{(k)} \\ -\mathbf{S}_{21}^{(k)} & \mathbf{I}_m - \mathbf{S}_{22}^{(k)}\end{bmatrix} \begin{bmatrix}\mathbf{f}_L \\ \mathbf{f}_U\end{bmatrix}$$

$$= 2\sum_{k=1}^{K} w_k (\mathbf{f}_L^T(\mathbf{I}_n - \mathbf{S}_{11}^{(k)})\mathbf{f}_L + 2\mathbf{f}_U^T(-\mathbf{S}_{21}^{(k)})\mathbf{f}_L + \mathbf{f}_U^T(\mathbf{I}_m - \mathbf{S}_{22}^{(k)})\mathbf{f}_U).$$

Then from $\frac{\partial}{\partial \mathbf{f_L}} J_L(\mathbf{w};\mathbf{f}) = 0$ and $\frac{\partial}{\partial \mathbf{f_U}} J_L(\mathbf{w};\mathbf{f}) = 0$, we have the above linear system.

2 Meta-path Selection

Notice that we only use some of the meta-path candidates in our experiment. Meta-path selection is necessary in our framework because different meta-paths have different meanings and different contributions to the interaction between objects with respect to the target response variable. Indeed, the user-guided meta-path selection model has been studied in [1]. However, its probabilistic approach requires user-specified seeds and has a relatively high computational cost. In our framework, we suggest an intuitive selection framework to identify the role of each meta-path using *Lasso* [2], the L_1-constrained selection method for linear regression. For any two different labeled objects x_u and x_v, we can calculate a distance between y_u and y_v. For example, we can calculate the distance using $dist_{u,v} = (y_u - y_v)^2$. This distance can be regarded as response and $R_{u,v}^{(k)}$ can be regarded as predictors, where $k = 1, ..., K$ and $u, v = 1, 2, ..., n$. Let $\hat{dist}_{u,v} = b_0 + \sum_{i=1}^{K} b_k R_{u,v}^{(k)}$. The target is to minimize

$$\sum_{(u,v)} (dist_{u,v} - \hat{dist}_{u,v})^2 + \lambda \sum_{k=1}^{K} ||b_k||,$$

where λ can be determined by cross-validation. Intuitively, a good meta-path P_k should be selected and its corresponding coefficient b_k should be negative.

We also notice that when we have a large number of meta-path candidates, this method can achieve meta-path selection automatically and efficiently due to the special property of *Lasso*. Unfortunately, real-world heterogeneous information networks usually contain a large number of links. Regression model thus cannot be efficiently computed. However, since the target of this framework is only to provide an intuition of each meta-path's contribution, we can randomly sample some of them and apply *Lasso* regression on them. Specifically, we randomly select 10 sets of 50000 edges. The final meta-paths used in our *Grempt* model should be selected by more than half of these 10 models and associated coefficients should be negative.

The original meta-path candidates for IMDb data are movie-actor-movie (M-A_1-M), movie-actress-movie (M-A_2-M), movie-director-movie (M-D-M), movie-genre-movie (M-G-M) ,movie-studio-movie (M-S-M) and movie-writer-movie (M-W-M), and the original meta-path candidates for DBLP data are author-paper-author (A-P-A), author-term-author (A-T-A), author-venue-author (A-V-A),author-paper-(cite)-paper-(cited by)-paper-author (A-P→P←P-A) and author-paper-(cited by)-paper-(cite)-paper-author (A-P←P→P-A). For the IMDb data, result shows that M-A_1-M, M-A_2-M, M-D-M, M-G-M and M-W-M have significant contribution to the network based consistency of $log(box\ office)$. The fact that M-S-M has not been selected shows that movies produced by the same studio may have significantly different box office values. For the DBLP data, we notice that A-P-A, A-V-A and A-P←P→P-A are selected by *Lasso*, which means that authors who collaborated with each other, who published papers in same venues or whose papers have been cited by same papers tend to have similar academic influence, which is represented by $log(\#citation + 1)$.

References

[1] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1348–1356.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.