JMLR: Workshop and Conference Proceedings 63:270–285, 2016 ACML 2016

Hierarchical Probabilistic Matrix Factorization with Network Topology for Multi-relational Social Network

Haoli Bai BAI@STD.UESTC.EDU.CN
Yingcai Honors College,
University of Electronic Science and Technology of China

Zenglin Xu ZLXU@UESTC.EDU.CN
Bin Liu LIU@STD.UESTC.EDU.CN
Big Data Research Center, School of Computer Science and Engineering,
University of Electronic Science and Technology of China

Yingming Li YINGMING@ZJU.EDU.CN
College of Information Science & Electronic Engineering, Zhejiang University

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

Link prediction in multi-relational social networks has attracted much attention. For instance, we may care the chance of two users being friends based on their contacts of other patterns, e.g., SMS and phone calls. In previous work, matrix factorization models are typically applied in single-relational networks; however, two challenges arise to extend it into multi-relational networks. First, the interaction of different relation types is hard to be captured. The second is the cold start problem, as the prediction of new entities in multi-relational networks becomes even more challenging. In this article we propose a novel method called Hierarchical Probabilistic Matrix Factorization with Network Topology (HPMFNT). Our model exploits the network topology by extending the Katz index into multi-relational settings, which could efficiently model the multidimensional interplay via the auxiliary information from other relationships. We also utilize the extended Katz index along with entitiy attributes to solve the cold-start problem. Experiments on two real world datasets have shown that our model outperforms the state-of-the-art with a significant margin.

Keywords: link prediction, multi-relational social network, probabilistic matrix factorization, network topology.

1. Introduction

Multi-relational social networks are universal in the real world. People usually interact through multiple patterns; e.g., they may contact with friends, or share photos and online posts with different people, forming a multi-relational network. Given such a network, a typical problem is to predict missing links (e.g., friendship interactions). This is known as the link prediction problem, an important branch in relational learning and recommender systems.

Although matrix factorization (MF) models have been wildly studied in link prediction problems, most of them have limitations. Well-known MF methods such as probabilistic matrix factorization (PMF) (Salakhutdinov and Mnih, 2007) are effective for single-relation problems, but are not scalable to multi-relational networks. There are two challenges to extend MF into multi-relational settings. One is to capture the correlation among different relationships. Some previous efforts based on PMF (Zhang et al., 2010) seek to embrace the multidimensional interaction into the covariance matrix of the latent feature variables implicitly. However, network topology, which contains informative structures across diverse relationships explicitly, is seldom considered under the MF framework. The other challenge is how to handle new entities, i.e., the cold-start problem. By exploiting side information from entity attributes, a majority of former tasks (Park et al., 2013)(Simm et al., 2015) have obtained desirable results. Nevertheless, network topology, aside from depicting multidimensional interaction, acts as good supplementary information for sparse networks as well, but has been ignored in the literature.

In this paper, we present a novel MF-based model called hierarchical probabilistic matrix factorization with network topology (HPMFNT) to address the above challenges. Grounded on PMF, a number of other latent factors are introduced to collectively improve its performance. Specifically, to model the interaction among relationships, we investigate the network topology explicitly by exploiting a multi-relational network extension of Katz index (Katz, 1953). To alleviate the cold-start problem, we further exploit the extended Katz index together with entities' attributes. We also sort to variational inference to learn the parameters with reduced computational cost. Furthermore, experiments have demonstrated the advantages of our model over the state-of-the-art. The contributions of our work are two folds:

- We model the interactive pattern of different relationships explicitly, making matrix factorization methods scalable to multi-relational networks.
- We incorporate entity attributes and topological features into the modeling of matrix factorization to collectively solve the cold-start problem.

The rest of the paper is structured as follows: Section 2 gives a brief review of the related work. In Section 3 we demonstrate our model in detail, and Section 4 presents the variational inference for the parameters. In Section 5, the description of experiments and the performances of HPMFNT against baselines are presented. Finally Section 6 concludes the paper and describes the future work.

2. Related Work

Multi-relational link prediction and its relevant problems have a long history with vast literature, and here we first give a snapshot on the matrix factorization approaches, since they are the focus of this paper. Then we introduce some other methodologies on the same issue.

Traditional matrix factorization models such as PMF (Salakhutdinov and Mnih, 2007) and Bayesian PMF (BPMF) (Salakhutdinov and Mnih, 2008) are single-relation targeted,

and could not take advantage of multidimensional knowledge. To extend matrix factorization models into multi-relational settings, various approaches are proposed. One focus of these models is to capture the interactive pattern of relations. Collective matrix factorization (Singh et al., 2008; Li et al., 2016), in which entities evolved in multiple relations, contains have common parameters along with the decomposed matrices. Based on PMF, (Zhang et al., 2010; Krohn-Grimberghe et al., 2012) consider the latent features as matrix-variate Gaussian variables, and embed the interdependence of different relations into the covariance matrix. Another focus is to solve the cold-start problem. A common way is to incorporate entity attributes and side information based on MF models, as shown in (Menon and Elkan, 2011; Park et al., 2013; Simm et al., 2015).

A number of MF approaches focus on symmetric social networks, and propose a symmetric decomposition. For instance, in community-based social networks, the stochastic equivalence assumption requires the decomposed community weight matrix to be symmetric (Holland et al., 1983; Zhou, 2015; Acharya et al., 2015) . However, in our paper we focus on both symmetric and asymmetric social networks, and propose a $\mathbf{UV}$ decomposition. For symmetric settings we expose a constraint of $\mathbf{V} = \mathbf{U}$. There are also similar investigations on asymmetric networks, such as (Hoff, 2008).

Aside from MF, tensor approaches are also an important branch for multi-relational link prediction. Classical tensor factorization models like Candecomp/Parafac (CP) (Harshman, 1970; Carroll and Chang, 1970) could model the relational interaction via a higher dimension of latent features. An increasing number of tensor approaches like (Nickel et al., 2011; Xiong et al., 2010; Sheng et al., 2012) are also based on the classical models like CP and Dedicom (Harshman, 1978), and display competitive performance. Other efforts include extending a number of supervised measurements (Liben-Nowell and Kleinberg, 2007) such as Common neighbors (CN) (Newman, 2001), Adamic/Adar (Adamic and Adar, 2003), Katz (Katz, 1953) into multi-relational settings, as shown in (Davis et al., 2012; Rossetti et al., 2011). Typically these measurements are based on statistical indices of network topology.

3. Model

In this section, we present HPMFNT. Suppose $\mathbf{Y} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, ..., \mathbf{Y}^{(R)}\}$ denotes a multi-relational social network, and R is the number of relations. We use $y_{ij}^{(r)} = 1$ to denote that there exists a link between entity i and j in relation r, and $y_{ij}^{(r)} = 0$ vice versa. $y_{ij}^{(r)} =?$ denotes an unobserved link.

In the following paragraphs, we choose relation r as a representative to formulate our model. We first introduce a set of latent variables to map the binary values of $y_{ij}^{(r)}$ into continuous space, then exploit network topology with an extended Katz index to model the interdependence of relationships, and combine it with hierarchical probabilistic matrix factorization models to form HPMFNT. Finally we present the joint distribution of HPMFNT.

3.1. Probit noise function

Typically, network entries take values in binary space. Since entries recovered from MF approaches are not discrete, a common way is to map the elements of $\mathbf{Y}^{(r)}$ into continuous

space via a latent matrix $\mathbf{X}^{(r)}$. Specifically, we assign a probit function on $x_{ij}^{(r)}$, and assume that the elements of $\mathbf{Y}^{(r)}$ are conditionally independent given $\mathbf{X}^{(r)}$, then we have

$$p(\mathbf{Y}^{(r)}|\mathbf{X}^{(r)}) = \prod_{1\leq i\leq j\leq N} \Phi(x_{ij}^{(r)})^{y_{ij}^{(r)}} \cdot (1-\Phi(x_{ij}^{(r)}))^{1-y_{ij}^{(r)}}, \tag{1}$$

where $\Phi(\cdot)$ is the cumulated distribution function of a standard normal distribution.

However, direct inference on $\mathbf{X}^{(r)}$ is intractable. For the convenience of parameter estimation, similar to (Albert and Chib, 1993) and (Yan et al., 2012), we incorporate another latent matrix $\mathbf{Z}^{(r)} = \{z_{ij}^{(r)}\}$ as an augmented representation of the probit model:

$$p(y_{ij}^{(r)}|z_{ij}^{(r)}) = \delta(y_{ij}^{(r)} = 1)\delta(z_{ij}^{(r)} > 0) + \delta(y_{ij}^{(r)} = 0)\delta(z_{ij}^{(r)} \leq 0), \tag{2}$$

$$p(z_{ij}^{(r)}|x_{ij}^{(r)}) = \mathcal{N}(z_{ij}^{(r)}; x_{ij}^{(r)}, 1), \tag{3}$$

where $\delta(\cdot)$ is the indicator function (i.e. its value is 1 if the statement inside is true, and 0 otherwise). $z_{ij}^{(r)}$ is a normal distribution with its mean $x_{ij}^{(r)}$ and covariance 1. If we marginalize $z_{ij}^{(r)}$ in Equation (2), it could be found that this is an equivalent representation of the probit model.

3.2. Network topology via extended Katz index

Network topology often contains structural information, and is therefore quite helpful in link prediction problems. As discussed in Section 2, a number of supervised measurements are proposed to extract the network topology. Especially, Katz index (Katz, 1953) has been applied widely in link prediction problems for its simplicity and effectiveness, which is defined as

$$score(i,j) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{i,j}^l|, \tag{4}$$

where β is the parameter of the predictor, $|\cdot|$ is the cardinality of the set, and $paths_{i,j}^l$ is the set of length-l paths from i to j.

Our model would extend Katz index into multi-relational networks from the prospective of Bayesian framework. The advantages include that Katz index could be naturally extended into multi-relations, and parameter β would be learned from data rather than a subjective choice. Besides, this removes the restriction of exponential damping of path lengths. Furthermore, as it will be illustrated later, in Katz index the path numbers contribute to the score linearly, while we generalize it to handle nonlinear network interactions.

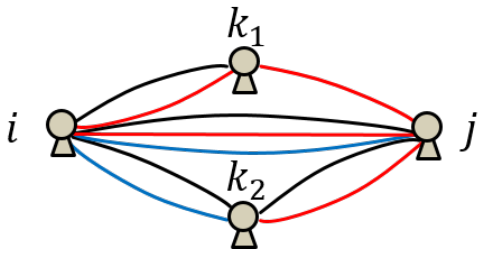


Figure 1: An example of entities i, j and their intermediate entities k_1, k_2 in three relations. The black, red and blue lines correspondingly denote relation 1, 2 and 3 respectively.

Specifically, for an entity pair (i, j) in a multi-relational network, entity i usually takes paths of various patterns via its intermediate entities to the target entity j, as shown in Figure 1. In order to model the mutual interaction of different relations, we introduce latent regression parameters $\mathbf{B}^{(r,l)}$ to augment the path weights β in Katz index, where $\mathbf{B}^{(r,1)}$ is an $N \times 1$ vector, $\mathbf{B}^{(r,2)}$ is an $N \times N$ matrix, and $\mathbf{B}^{(r,l)}$ is an l-way tensor ($l \geqq 3$). In order to incorporate the extended Katz index into the PMF framework, we set Gaussian priors on $\mathbf{B}^{(r,l)}$:

$$p(\mathbf{B}^{(r,l)}) = \prod_{r_1,r_2,...r_l=1}^{R} \mathcal{N}(B^{(r,l)}_{r_1,r_2,...r_l}|b_0, \rho_0^{-1}).$$

We use $path^{(r,l)}(i,j)$ to represent the set of length-l paths from i to j in relation r, and $path^{(l)}(i,j) = path^{(1,l)}(i,j) \bigcup path^{(2,l)}(i,j) \bigcup ... \bigcup path^{(R,l)}(i,j)$ therefore denotes the set of length-l paths from i to j in all relations.

Moreover, note that Katz index assumes the paths between entity i and j linearly contribute to the $score(i,j)$. However, this may be insufficient to capture the complex social interactions, i.e. the probability of the presence of a link may not be proportional to the number of the shared intermediate entities. Therefore, we replace $|path^{(r,l)}(i,j)|$ in Equation (4) by a nonlinear mapping $p_{ij}^{(r,l)} = \phi(|path^{(r,l)}(i,j)|)$ to model the nonlinear network interactions. We further set $\mathbf{p}_{ij}^{(l)} = [p_{ij}^{(1,l)}, p_{ij}^{(2,l)}, ..., p_{ij}^{(R,l)}]$ as the path-counting vector in multi-relational settings.

Consequently, for each entity pair (i,j) in relation r, the Katz index can be extended as

$$score^{(r,L)}(i,j) = \sum_{l=1}^{L} \prod_{m=1}^{l} \mathbf{B}^{(r,l)} \overline{\times}_m \mathbf{p}_{ij}^{(l)}, \tag{5}$$

where $\overline{\times}_m$ is the *n-mode(vector) product* defined in (Kolda and Bader, 2008), and L is the maximum path length to consider. Note that usually for $L \geq 3$. Note that the contribution of paths becomes negligible to the summation while it boosts the computation significantly (Lu et al., 2010), therefore it is often the case to set $L = 1, 2$.

To give a concrete example, Figure 1 shows three relations $\mathbf{Y} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{Y}^{(3)}\}$ denoted by the black, red and blue lines correspondingly. By counting the lines of different relations with different path lengths, Equation (5) can be formulated as:

$$
\begin{aligned}
&score^{(r,2)}(i,j) = \mathbf{p}_{ij}^{(1)\top} \cdot \mathbf{B}^{(r,1)} + \mathbf{p}_{ik}^{(2)\top} \cdot \mathbf{B}^{(r,2)} \cdot \mathbf{p}_{kj}^{(2)} \\
&= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}^{\top} \cdot \begin{pmatrix} B_1^{(r,1)} \\ B_2^{(r,1)} \\ B_3^{(r,1)} \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}^{\top} \cdot \begin{pmatrix} B_{11}^{(r,2)} & B_{12}^{(r,2)} & B_{13}^{(r,2)} \\ B_{21}^{(r,2)} & B_{22}^{(r,2)} & B_{23}^{(r,2)} \\ B_{31}^{(r,2)} & B_{32}^{(r,2)} & B_{33}^{(r,2)} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}.
\end{aligned}
$$

For notation simplicity, we use $\mathbf{B}^{(r)} = \{\mathbf{B}^{(r,1)}, \mathbf{B}^{(r,2)}, ...\mathbf{B}^{(r,L)}\}$ to embrace all the path regression parameters, and $s_{ij}^{(r,L)}$ to replace $score^{(r,L)}(i,j)$ in the following paragraphs.

3.3. Incorporating the extended Katz index into hierarchical probabilistic matrix factorization

We aim at the decomposition of both symmetric and asymmetric social networks, and follow (Salakhutdinov and Mnih, 2007) to introduce $\mathbf{U}^{(r)}$, $\mathbf{V}^{(r)} \in R^{N \times K}$ as latent feature matrices in relation r, where K denotes the decomposition rank. $\mathbf{u}_i^{(r)}$, $\mathbf{v}_j^{(r)}$ are the entity-targeted latent vectors of i and j, respectively. For symmetric networks we expose the constraint of $\mathbf{U}^{(r)} = \mathbf{V}^{(r)}$ to preserve the mutual interaction. The entry of latent matrix $x_{ij}^{(r)}$ is generated by a sum of inner product of user latent variables, regression terms $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ on the entity feature vectors $\mathbf{f_i}$, $\mathbf{f_j}$ and their corresponding path scores $s_{ij}^{(r)}$, plus a noise ε_{ij}:

$$
x_{ij}^{(r)} = [\mathbf{u}_i^{(r)}]^{\top} \mathbf{v}_j^{(r)} + \boldsymbol{\alpha}^{\top} \mathbf{f}_i + \boldsymbol{\beta}^{\top} \mathbf{f}_j + s_{ij}^{(r)} + \varepsilon_{ij},
$$

where $\varepsilon_{ij} \sim \mathcal{N}(\varepsilon_{ij}|0, \sigma^{-1})$, and σ is the precision. Thus

$$
p(\mathbf{X}^{(r)}|\mathbf{U}^{(r)}, \mathbf{V}^{(r)}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}^{(r)}, \sigma) = \prod_{i,j}^{N} \mathcal{N}(x_{ij}^{(r)}|[\mathbf{u}_i^{(r)}]^{\top} \mathbf{v}_j + \boldsymbol{\alpha}^{\top} \mathbf{f}_i + \boldsymbol{\beta}^{\top} \mathbf{f}_j + s_{ij}^{(r)}, \sigma^{-1})^{\delta(y_{ij}^{(r)}=1)}.
$$

In terms of the prior distribution on the latent variables, we first assign Gaussian distributions on entity latent feature variables:

$$
p(\mathbf{U}^{(r)}|\boldsymbol{\mu}_U, \boldsymbol{\Lambda}_U) = \prod_{i}^{N} \mathcal{N}(\mathbf{u}_i^{(r)}|\boldsymbol{\mu}_U, \boldsymbol{\Lambda}_U^{-1}),
$$

$$
p(\mathbf{V}^{(r)}|\boldsymbol{\mu}_V, \boldsymbol{\Lambda}_V) = \prod_{j}^{N} \mathcal{N}(\mathbf{v}_j^{(r)}|\boldsymbol{\mu}_V, \boldsymbol{\Lambda}_V^{-1}).
$$

Furthermore, the mean and covariance of $\mathbf{u}_i^{(r)}$ and $\mathbf{v}_j^{(r)}$ are assumed to be Gaussian-Wishart distributions:

$$
\begin{aligned}
p(\boldsymbol{\mu}_U, \boldsymbol{\Lambda}_U) &= \mathcal{N}(\boldsymbol{\mu}_U|\boldsymbol{\mu}_0, (\gamma \boldsymbol{\Lambda}_U)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_U|\mathbf{W}_0, \nu_0), \\
p(\boldsymbol{\mu}_V, \boldsymbol{\Lambda}_V) &= \mathcal{N}(\boldsymbol{\mu}_V|\boldsymbol{\mu}_0, (\gamma \boldsymbol{\Lambda}_V)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_V|\mathbf{W}_0, \nu_0),
\end{aligned}
$$

where μ_0, $\mathbf{W}_0$, ν_0 are the hyper parameters. We also set the regression parameters on entity feature as Gaussian variables:

$$p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0, \boldsymbol{\Lambda}_\alpha^{-1}),$$
$$p(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_\beta^{-1}),$$

where similarly $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$ $\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}$ and $\boldsymbol{\Lambda}_\beta$ are the hyper parameters as well.

3.4. Joint distribution

Combining all the variables together, the joint distribution of HPMFNT becomes

$$\begin{aligned} & p(\mathbf{Y}^{(r)}, \mathbf{Z}^r, \mathbf{X}^{(r)}, \mathbf{U}^{(r)}, \mathbf{V}^{(r)}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}^{(r)}) \\ & = \prod_{i,j}^{N} p(y_{ij}^{(r)}|z_{ij}^r, x_{ij}^{(r)}) p(z_{ij}^{(r)}|x_{ij}^{(r)}) \\ & \cdot p(\mathbf{x}_{ij}^{(r)}|\mathbf{u}_i^{(r)}, \mathbf{v}_j^{(r)}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}^{(r)}) p(\mathbf{u}_i^{(r)}|\boldsymbol{\mu}_U, \boldsymbol{\Lambda}_U) p(\mathbf{v}_j^{(r)}|\boldsymbol{\mu}_V, \boldsymbol{\Lambda}_V) \\ & \cdot p(\boldsymbol{\mu}_U, \boldsymbol{\Lambda}_U) p(\boldsymbol{\mu}_V, \boldsymbol{\Lambda}_V) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\mathbf{B}^{(\mathbf{r})}). \end{aligned} \tag{6}$$

Given the observed networks $\mathbf{Y}$, the next task is to estimate all the latent variables, and then predict unknown links. We would like to use the maximum a posteriori (MAP) method by uplifting the lower bound of Equation (6) iteratively, as will be discussed in the next subsection. The graphical presentation of our model is shown in Figure 2.

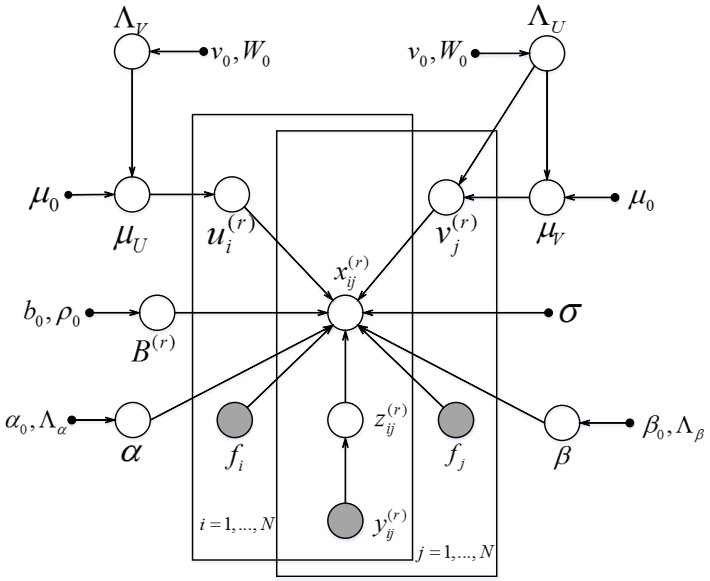


Figure 2: Graphical representation of HPMFNT.

4. Variational inference for parameter estimation

Typically in Bayesian inference problems, sampling and variational inference are two common tools for parameter inference. Due to low efficiency of sampling, here we sort to variational inference to estimate the parameters in the model. The posterior distribution is approximated via a set of latent variables:

$$\mathcal{Z} = \{\mathbf{X}, \mathbf{Z}, \mathbf{U}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\mu_U}, \boldsymbol{\Lambda_U}, \boldsymbol{\mu_V}, \boldsymbol{\Lambda_V}\}.$$

Then our task is to maximize the lower bound

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathcal{Z}) d\mathcal{Z} \geq \int q(\mathcal{Z}) \log \frac{p(\mathbf{Y}, \mathcal{Z})}{q(\mathcal{Z})} d\mathcal{Z} = \mathcal{L}(q).$$

Assume density function $q(\mathcal{Z})$ is fully factorized by

$$q(\mathcal{Z}) = q(\mathbf{X})q(\mathbf{Z})q(\mathbf{U})q(\mathbf{V})q(\boldsymbol{\alpha})q(\boldsymbol{\beta})q(\mathbf{B})q(\mu_\mathbf{U}|\boldsymbol{\Lambda_U})q(\boldsymbol{\Lambda_U})q(\mu_\mathbf{V}|\boldsymbol{\Lambda_V}\boldsymbol{\Lambda_V}).$$

It is known in (Bishop, 2006) that when maximizing the lower bound $\mathcal{L}(q)$, the distribution of each variable $q_j(\mathbf{Z}_j)$ can be derived by

$$\log q_j(Z_j) = E_{i \neq j}[\log p(\mathbf{Y}, \mathcal{Z})] + const, \tag{7}$$

where $E_{i \neq j}[\cdot]$ represents an expectation with respect to the variational posterior distributions over all variables except Z_j. By successively applying Equation (7) on each latent variable, we could iteratively maximize the lower bound like the coordinate ascent algorithm, and (Boyd and Vandenberghe, 2004) proved this method would finally converge since the bound is convex. Note that variables that share a similar derivation are skiped in the paper.

The variational posterior of $x_{ij}^{(r)}$ is a normal distribution with precision $\sigma^* = \sigma + 1$ and mean

$$\langle x_{ij}^{(r)} \rangle = \frac{1}{\sigma^*}(\sigma(\langle [\mathbf{u}_i^{(r)}]^\top \mathbf{v}_j^{(r)} \rangle + \langle \boldsymbol{\alpha} \rangle^\top \mathbf{f_i} + \langle \boldsymbol{\beta} \rangle^\top \mathbf{f_j} + \langle s_{ij}^{(r)} \rangle) + \langle z_{ij}^{(r)} \rangle),$$

where $\langle \cdot \rangle$ denotes the regarded expectation. The variational posterior of $z_{ij}^{(r)}$ is a truncated normal distribution

$$q(z_{ij}^{(r)}) = \mathcal{N}(z_{ij}^{(r)}; \langle x_{ij}^{(r)} \rangle, 1)(\delta(y_{ij}^{(r)} = 1)\delta(z_{ij}^{(r)} \geq 1) + \delta(y_{ij}^{(r)} = 0)\delta(z_{ij}^{(r)} < 0)),$$

and hence the mean of $z_{ij}^{(r)}$ takes the form

$$\langle z_{ij}^{(r)} \rangle = (2y_{ij}^{(r)} - 1)\mathcal{N}(\langle x_{ij}^{(r)} \rangle; 0, 1) + \langle x_{ij}^{(r)} \rangle \Phi((2y_{ij}^{(r)} - 1)\langle x_{ij}^{(r)} \rangle).$$

To derive $q(\boldsymbol{\mu}_U)$ and $q(\boldsymbol{\Lambda}_U)$, we have the Gaussian-Wishart distribution

$$q(\boldsymbol{\mu}_U, \boldsymbol{\Lambda}_U) = \mathcal{N}(\boldsymbol{\mu}_U | \boldsymbol{\mu}_0^*, (\lambda_0^* \boldsymbol{\Lambda}_U)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_U | \mathbf{W}_0^*, \nu_0^*),$$

where $\boldsymbol{\mu}_0^* = \frac{\lambda_0 \boldsymbol{\mu}_0 + N\overline{\mathbf{u}}}{\lambda_0 + N}$, $\nu_0^* = \nu_0 + N$, $\lambda_0^* = \lambda_0 + N$, $\overline{\mathbf{u}} = \frac{1}{N}\sum_{i=1}^{N} \langle \mathbf{u}_i^{(r)} \rangle$, $\overline{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N} \langle \mathbf{u}_i^{(r)} [\mathbf{u}_i^{(r)}]^\top \rangle$ and

$$(\mathbf{W}_0^*)^{-1} = \mathbf{W}_0^{-1} + N\overline{\mathbf{S}} + \frac{\lambda_0 N}{\lambda_0 + N}(\boldsymbol{\mu}_0 - \overline{\mathbf{u}})(\boldsymbol{\mu}_0 - \overline{\mathbf{u}})^\top.$$

To update the entity latent features $\mathbf{U}^{(r)}$, we have

$$q(\mathbf{U}^{(r)}) = \prod_{i=1}^{D} \mathcal{N}(\mathbf{u}_i^{(r)} | \boldsymbol{\mu}_{U,i}^{(r)*}, (\boldsymbol{\Lambda}_{U,i}^{(r)*})^{-1}),$$

where $\boldsymbol{\Lambda}_{U,i}^{(r)*} = \boldsymbol{\Lambda}_U + \sigma \sum_{j=1}^{N} \langle \mathbf{v}_j \mathbf{v}_j^\top \rangle$, and

$$\boldsymbol{\mu}_{U,i}^{(r)*} = (\boldsymbol{\Lambda}_U^{(r)*})^{-1}(\sum_{j=1}^{N} \langle \mathbf{v}_j \rangle^{(r)} (x_{ij}^{(r)} - \langle \boldsymbol{\alpha}^\top \rangle \mathbf{f}_i - \langle \boldsymbol{\beta}^\top \rangle \mathbf{f}_j - \langle s_{ij}^{(r)} \rangle)\sigma + \langle \boldsymbol{\Lambda}_U \rangle \langle \boldsymbol{\mu}_U \rangle).$$

To solve the regression parameter $\boldsymbol{\alpha}$ on entity attributes, we have

$$q(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha} | \boldsymbol{\alpha}_0^*, (\boldsymbol{\Lambda}_\alpha{}^*)^{-1}),$$

where $\boldsymbol{\Lambda}_\alpha{}^* = \langle \boldsymbol{\Lambda}_\alpha \rangle + \sum_{i}^{N} \sigma \mathbf{f}_i \mathbf{f}_i^\top$, and

$$\boldsymbol{\alpha}_0^* = (\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^*)^{-1}(\sum_{i,j}^{N} \mathbf{f}_i (x_{ij}^{(r)} - \langle \mathbf{v}_j^{(r)} \rangle - \langle \boldsymbol{\beta}^\top \rangle \mathbf{f}_j - \langle s_{ij}^{(r)} \rangle)\sigma + \boldsymbol{\Lambda}_\alpha \boldsymbol{\alpha}_0).$$

The variational posterior of $B_{r_1,r_2,\ldots r_l}^{(r,l)}$ is a Gaussian distribution

$$q(B_{r_1,r_2,\ldots r_l}^{(r,l)}) = \mathcal{N}(B_{r_1,r_2,\ldots r_l}^{(r,l)} | \mu_{r_1,r_2,\ldots r_l}^{(r,l)*}, (\rho_{r_1,r_2,\ldots r_l}^{(r,l)*})^{-1}),$$

with precision $\rho_{r_1,r_2,\ldots r_l}^{(r,l)*} = \rho_0 + \sum_{i,j}^{N} (p_{ij}^{(r_1,l)} \cdot p_{ij}^{(r_2,l)} \cdots p_{ij}^{(r_l,l)})^2 \sigma$, and mean in a bit more complicated form:

$$\begin{aligned}\mu_{r_1,r_2,\ldots r_l}^{(r,l)*} =& (\rho_{r_1,r_2,\ldots r_l}^{(r,l)*})^{-1}(\rho_0 b_0 + \sum_{i,j}^{N} p_{ij}^{(r_1,l)} \cdot p_{ij}^{(r_2,l)} \cdots p_{ij}^{(r_l,l)} \\ & \cdot \sigma(x_{ij}^{(r)} - \langle \mathbf{u}_i^{(r)} \rangle^\top \langle \mathbf{v}_j^{(r)} \rangle - \langle \boldsymbol{\alpha}^\top \rangle \mathbf{f}_i - \langle \boldsymbol{\beta}^\top \rangle \mathbf{f}_j - \langle s_{ij}^{(r)} \rangle_{\setminus B_{r_1,r_2,\ldots r_l}^{(r,l)}})),\end{aligned}$$

where $\langle s_{ij}^{(r)} \rangle_{\setminus B_{r_1,r_2,\ldots r_l}^{(r,l)}}$ denotes the expectation with all random variables except $B_{r_1,r_2,\ldots r_l}^{(r,l)}$.

5. Experiment

In this section, we demonstrate the performance of our model on two real world datasets: **Reality Mining** (Pentland et al., 2009) and **Social Evolution** (Dong et al., 2011). We compare our HPMFNT with several other models:

- Probabilistic matrix factorization (PMF) (Salakhutdinov and Mnih, 2007). PMF is a well-known MF model, which utilizes latent feature variables to capture network structures. PMF utilizes stochastic gradient descent (SGD) to update the parameters.
- Bayesian probabilistic matrix factorization (BPMF) (Salakhutdinov and Mnih, 2008). BPMF uses the Gibbs sampler to estimate the parameters in PMF. Moreover, BPMF is a hierarchical model, providing a wider space for latent parameters to fit data. We choose PMF and BPMF to verify the improvement by the network topology and entity attributes.
- Candecomp/Parafac (CP) (Harshman, 1970) factorization. CP is a classical tensor factorization approach. Unlike PMF and BPMF, CP can capture the interaction of different relations via its higher dimension. With the analytical solution of each latent matrix, parameters of CP can be updated iteratively.

We shall report the AUC performance of these models.

5.1. Dataset description

- **Reality Mining** dataset is to infer the structures of social network. There are $R = 3$ relations in the network: friendship, in-lab interactions and out-lab interactions, with 13,395 observed links in total. Each relation consists of $N = 94$ entities. Moreover, the entities' features are also recorded in the data, e.g., satisfaction level of the current social circle, etc. We choose 25 features and form a feature matrix $F \in R^{25\times 94}$.
- **Social Evolution** dataset is to track the everyday life of undergraduates on campus, with $N = 84$ entities and 8,964 observed links in total. We choose $R = 7$ relations in the data: social activity, political discussion, friendship, shared photos, shared online activities, voice calls and proximity network. We omit some other relations because either it's too sparse or it contains too much unknown outside entities. Some links (e.g., voice calls) repeats multiple times and are deduplicated. Entities' features are recorded, like grade, living floor etc, forming a feature matrix $F \in R^{2\times 84}$.

5.2. Experiment setup

First, we randomly split each relation of the network into training set and validation set, and evaluate the performance on each relation. The calculation of path score is only based on the training set. The split ratio is also varied to evaluate the robustness of these models.

Then we compare our model with PMF, BPMF and CP. Our HPMFNT is trained for each relation together with the auxiliary information of other relationships, which predicts the unobserved links based on the mean of $y_{ij}^{(r)}$ in Equation (1). For PMF and BPMF, we also separately implement them to each relation, and make prediction from the learned entity latent features. Since CP can model the interactive pattern via the higher dimension, it needs to be learned only once to make prediction for different relationships.

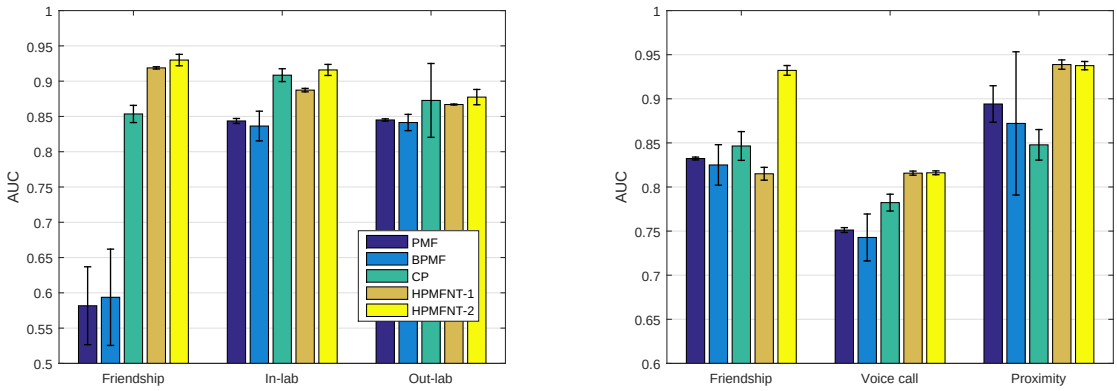
To set proper hyper parameters, we choose the learnning rate $\epsilon = 0.1$ and hold the rest parameters as default for PMF. For BPMF and CP, most parameters remain unchanged. Then we initialize our model based on user latent matrices $\mathbf{U}$ and $\mathbf{V}$ obtained by PMF.

Similar to BPMF, we also choose $\mu_0 = 0$, $\nu_0 = K$, and $\mathbf{W}_0 = I$, and beyond that, we set $b_0 = 0$, $\boldsymbol{\alpha}_0 = \boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{\Lambda}_{\boldsymbol{\alpha}} = \boldsymbol{\Lambda}_{\beta} = \mathbf{I}$. Moreover, grid search is applied with presicion λ in $\{0.01, 1, 100\}$ and ρ_0 in $\{0.01, 1, 100\}$. We consider the cases of $L = 1$ and $L = 2$ (denoted as HPMFNT-1, HPMFNT-2 respectively), since paths of $L \geq 3$ contribute very little to the score. We set sigmoid function $\sigma(\cdot)$ as the nonlinear mapping in Equation (5).

5.3. Results

We first compare the AUC with PMF, BPMF and BPTF with 80% obersved links in the training set and 20% in the validation set in each dataset. Due to the space limitation, we randomly pick three relations to display the results.

From Figure 3, we could find that both HPMFNT-1 and HPMFNT-2 perform generally better than PMF and BPMF. This validates the effectiveness of the network topology and entity attributes in MF framework. Moreover, HPMFNT also beats CP for a considerable degree, since HPMFNT-1 is better or at least comparable to CP, and HPMFNT-2 generally achieves a higher AUC than that of CP. This shows the advantage of our HPMFNT over CP-based tensor factorization methods. We could also observe that both HPMFNT-1 and HPMFNT-2 have lower standard bar than the rest models, which indicates the stability and robustness of our method. To have a better interpretation of the result, the test networks and the predicted networks are also visualized by Gephi, presented in Appendix A.



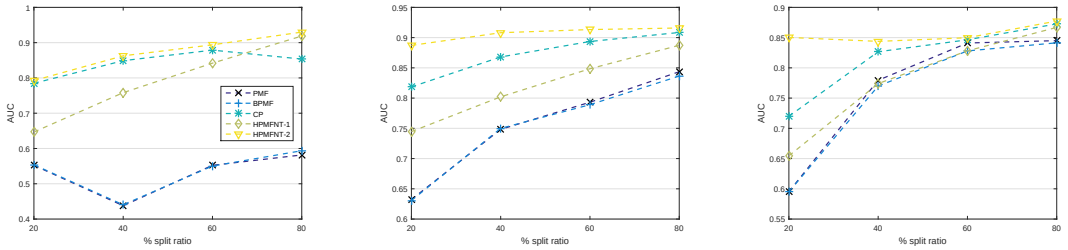
(a) Reality mining

(b) Social evolution

Figure 3: AUC performance on two datasets. The colorbars in panel (b) have the same meaning as those in (a). Note that each model keeps the decomposition rank $K = 10$ for both datasets, while maximum iteration is 50, and each experiment is repeated for 10 times. Standard error bar is plotted.

Since as the network gets more sparse, the cold-start problem becomes even more challenging. Therefore, we vary the split ratio on Reality mining dataset to evaluate the robustness of our model. As shown in Figure 4, both HPMFNT-1 and HPMFNT-2 achieve better results than those of PMF and BPMF. Although HPMFNT-1 is generally lower than

CP, HPMFNT-2 outperforms CP at each split rate. Furthermore, HPMFNT-2 obtains considerable results even if the training set and the validation set are heavily imbalanced (i.e., the split ratio is 20%). This verifies that a larger path length could better exploit the network topology as supplementary information for the sparse network, and hence alleviate the influence of the cold-start problem.



(a) Friendship (b) In-lab (c) Outlab

Figure 4: AUC performance with varying split ratio on Reality mining dataset. The colors in panel (b) and (c) have the same meaning as those in (a).

6. Conclusion

In this paper we have presented HPMFNT, a new method to make MF models scalable to multi-relational networks. By extending Katz index into multi-relational settings, we could efficiently utilize the auxiliary networks and hence successfully model the interactive pattern of different relationships. Moreover, incorporating network topology and entity attributes into matrix factorization modeling considerably solves the cold-start problems. The experimental results illustrate that our model outperforms the competing models.

We consider several branches for future directions. First, aside from Katz index, we could attempt to introduce other supervised measurement into the Bayesian MF framework. Second, to make HPMFNT more suitable to heterogeneous attributes of entities, we will seek a more comprehensive attribute-encoding approach. Finally, since tensor-based relational learning has been a raising topic, we may also investigate the function of network topology in tensor factorization models.

References

Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices. *The 18th International Conference on Artificial Intelligence and Statistics (AISTATS-2015)*, 38, 2015.

Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25 (3):211–230, 2003.

James H. Albert and Siddhartha Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 128. Springer Press, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 35(3): 283–319, 1970.

Darcy Davis, Ryan Lichtenwalter, and Nitesh V. Chawla. Supervised methods for multi-relational link prediction. *Social Network Analysis and Mining*, 3(2):127–141, 2012.

Wen Dong, Bruno Lepri, and Alex Sandy Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 134–143, 2011.

Richard a Harshman. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16(10): 1– 84, 1970.

Richard A Harshman. Models for analysis of asymmetrical relationships among n objects or stimuli. In *First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology, McMaster University, Hamilton, Ontario*, volume 5, 1978.

Peter D Hoff. Multiplicative latent factor models for description andprediction of social networks. *Computational and Mathematical Organization Theory*, 15(4):261–272, 2008.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1): 39–43, 1953.

Tamara G Kolda and Brett W Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 2008.

Artus Krohn-Grimberghe, Lucas Drumond, Christoph Freudenthaler, and Lars Schmidt-Thieme. Multi-relational matrix factorization using bayesian personalized ranking for social network data. *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, page 173, 2012.

Yingming Li, Ming Yang, and Zhongfei (Mark) Zhang. Multi-view representation learning: A survey from shallow methods to deep methods. *CoRR*, abs/1610.01206, 2016.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

Zhengdong Lu, Berkant Savas, Wei Tang, and Inderjit S. Dhillon. Supervised Link Prediction Using Multiple Sources. *2010 IEEE International Conference on Data Mining*, pages 923–928, 2010.

Ak Menon and Charles Elkan. Link prediction via matrix factorization. *MLKDD*, 6912: 437–452, 2011.

M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 809–816, 2011.

Sunho Park, Yong Deok Kim, and Seungjin Choi. Hierarchical bayesian matrix factorization with side information. In *International Joint Conference on Artificial Intelligence*, pages 1593–1599, 2013.

A Pentland, N Eagle, and D Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36):15274–15278, 2009.

Giulio Rossetti, Michele Berlingerio, and Fosca Giannotti. Scalable link prediction on multidimensional networks. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 979–986. IEEE, 2011.

Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Neural Information Processing Systems*, volume 21, 2007.

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008.

GAO Sheng, Ludovic Denoyer, Patrick Gallinari, and GUO Jun. Probabilistic latent tensor factorization model for link pattern prediction in multi-relational networks. *The Journal of China Universities of Posts and Telecommunications*, 19:172–181, 2012.

Jaak Simm, Adam Arany, Pooya Zakeri, Tom Haber, Jörg K. Wegner, Vladimir Chupakhin, Hugo Ceulemans, and Yves Moreau. Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC. *Arxiv*, pages 1–13, 2015.

Amit Pratap Singh, Girijesh Kumar, and Rajeev Gupta. Relational learning via collective matrix factorization. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, Usa, August*, pages 650–658, 2008.

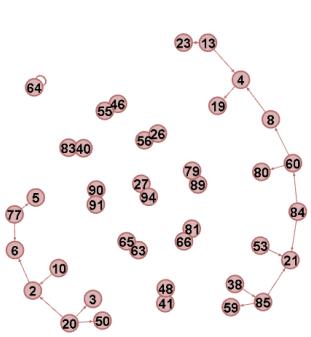
Liang Xiong, Xi Chen, Tzu-kuo Huang, Jeff Schneider, and Jaime G Carbonell. Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization. *Proceedings of the SIAM International Conference on Data Mining*, pages 211—-222, 2010.

Feng Yan, Zenglin Xu, and Yuan Qi. Sparse matrix-variate gaussian process blockmodels for network modeling. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-2012)*, 2012.

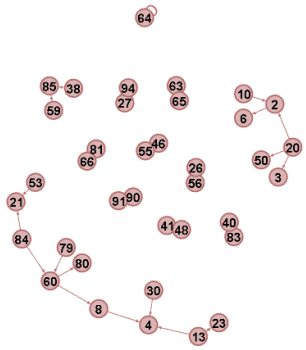
Yu Zhang, Bin Cao, and Dit Yan Yeung. Multi-domain collaborative filtering. In *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, Ca, Usa, July*, 2010.

Mingyuan Zhou. Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction. *Artificial Intelligence and Statistics (AISTATS)*, 38:1–11, 2015.

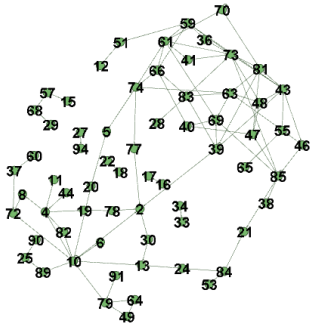
Appendix A. Visualization of Reality mining



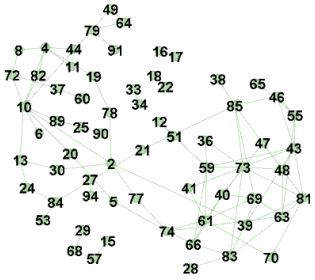
(a) Test network of friendship



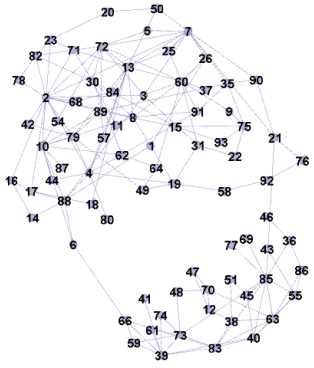
(b) Predicted network of friendship



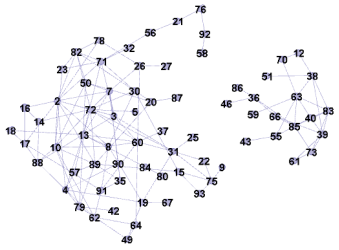
(c) Test network of in-lab



(d) Predicted network of in-lab



(e) Test network of out-lab



(f) Predicted network of out-lab

Figure 5: Visualization of Reality mining social network. Since the network partitions are random when imported, the arrangement of entities may be different between test network and predicted network.