Summary of Projects

January 3, 2018

Part I

Community Detection in Social Networks

We motivate the importance of sub-network level in identifying opinion leaders/influentials contrasting it with the common network level metrics such as node centralities. The concept of influential or opinion leaders in social networks arise from the high level of impact that some specific individuals can have on their peers to adopt a behavior or act upon an action. Ample amount of literature has attributed the importance of individuals at the network level to their degree of opinion leadership. Degree centrality, prestige/eigenvector centrality, and betweenness centrality are among common measures that incroporate the importance of individuals based on their network positions. On the other hand, every network is comprised of smaller sub-networks that hereafter we refer to as communities. We argue that analysis of individuals at these smaller communities can shed more light on their real role and importance in the network.

As an illustration for the relevance of sub-network level measures, figure 1. shows the graph of degree centrality of three different communities in a publicly available email network of a European Research Institute extracted from SNAP dataset . The connections are formed based on the emails sent between members of this institute, and the ground truth communities are the departments within the institute. As can be seen, nodes that are central (measured by degree) at the community level are not necessarily central at the network level.

1 More on Communities

Networks usually cluster into smaller sub-networks that exhibit denser inter-community ties compared to intra-communitiy connections. Many community discovery algorithms assume complete disjoint between communities(like in a mixture), however more recent studies, also account for the potential overlap between these structures. Approaches concerning community discovery include both algoirthmic and probabilistic methods. We incorporate and enhance on an exisiting probabilistic model known as mixed-membership-stochastic-blockmodel(MMSB) that allows for such overlaps among the communities.

2 Big Picture

For developing such model we follow the argument of homophily(birds of a feather flock together), where two individuals form a connection with each other if they have enough commonalities that make their interaction more plausible. Following figure demonstrates this idea of tie formation via a simple graphical model.

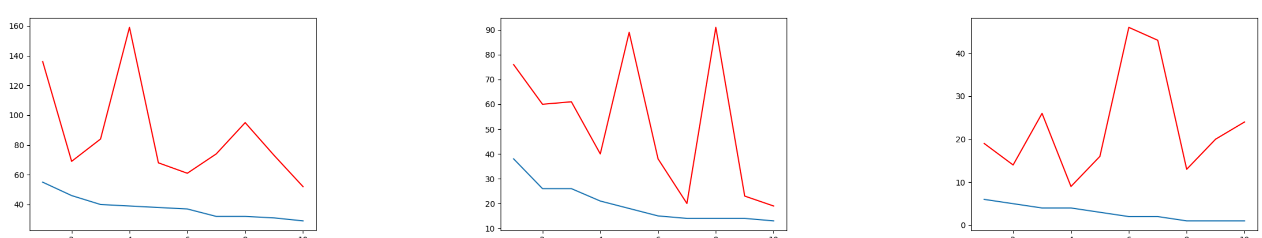

Figure 1: Three different communities, where each community is sorted based on the degree centrality (blue line) and the corresponding network centrality for that node is shown on top (red line)

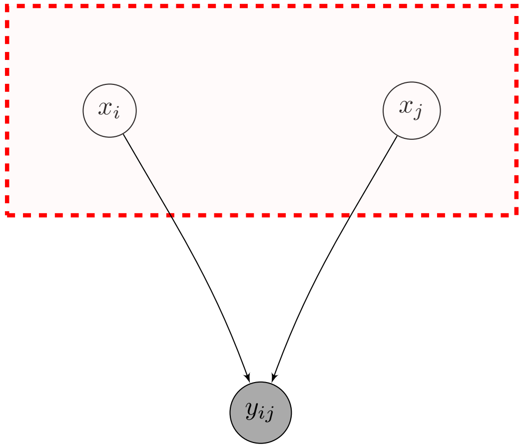

Figure 2: Big Picture-The shaded node y_{ij} represents our observation of connection(0, or 1) between two individuals i, and j. The unshaded nodes x_i and x_j are hidden variables(potentially multidimensional) that represent characteristics or preferences of individuals pertinent to tie formation.

3 Model

Based on the general idea established in the previous section, our main objective for community discivery is to uncover patterns in the latent variables x. To simplify we assume that the social network consists of K predefined, potenitially overlapping communities, and each individual can belong to several communities. This overlapping strucuture is allowed by allowing individuals to activate their specific roles(communities) when potentially interacting with other individuals. This is desired, since main volume of communications in real world networks arise from common interests, and at the same time each individual has several interests with different intensities. In our model these intensities are expressed by a $K-$dimensional membership probability vector for each individual. To realize the role activation of each individual in any contact to/by others, we define an interaction specific parameter(one-hot vector), where the preferred community is announced. However as discussed before, we expect denser patterns of communications within each community in comparison with between communities. To address this, we also employ a $K \times K$ compatibility matrix with large probabilities on the diagonal and small values in the off-diagonal entries.

The conventional MMSB model is described as below:

Algorithm 1 MMSB generative process

$for\ k \in 1, \ldots, K$
$\quad \beta_{kk} = Beta(\eta_0, \eta_1), \quad \beta_{kl_{l \neq k}} = \epsilon$
$for\ a \in \mathcal{N}:$
$\quad \theta_a \sim Dir(\alpha_{[K]})$
$for\ (a, b) \in \mathcal{N} \times \mathcal{N}:$
$\quad z_{a \to b} \sim Mult(\theta_a)$
$\quad z_{a \leftarrow b} \sim Mult(\theta_b)$
$\quad y(a, b) \sim Bern(z_{a \to b}^T B\, z_{a \leftarrow b})$

Note that here θ_a is a membership probability vector, and $z_{a \to b}$ is a one-hot community indicator for individual a in a potential interaction with b, and similarly $z_{a \leftarrow b}$ is a one-hot community indicator for individual b , when b is pontentially contacted by a. Moreover B is the $K \times K$ block matrix that has β on its diagonal and ϵ elsewhere.

There are some caveats concerning the Dirichlet prior in the MMSB generative process, that we try to resolve in our proposed model. We classify the potential loss by assuming Dirichlet prior in three categories:

1. Inability to capture correlation among different individual community memberships–Individuals who belong to distinct yet relevant(similar) communities may not be allowed to communicate with each other.

2. The probability strengths are quite extreme, that tends to encourage very disjoint clusters

3. The forced negative correlation in Dirichlet parameters, does not allow for defining actual correlations among communities and time dependence in the case of network evolution.

Algorithm 2 MMSB with Logistic Normal prior

- $\forall k \in [1, .., K]$
  - draw the diagonal elements of the block matrix B via$\beta_{k,k} \sim Beta(\eta_0, \eta_1)$
- $\forall i \in \mathcal{N}$
  - draw the mean of the logit mixed membership vector through $\boldsymbol{\mu} \sim Normal(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$
  - draw the precision of the logit mixed membership vector through $\boldsymbol{\Lambda} \sim Wishart(\boldsymbol{\ell}_0, \boldsymbol{L}_0)$
  - draw a K-dimensional vector,$\boldsymbol{\theta}_i^* \sim Normal(\boldsymbol{\mu}, \boldsymbol{\Lambda})$
  - construct the simplical mixed membership via logistic transformation , $\boldsymbol{\theta}_{i,k} = \frac{exp(\boldsymbol{\theta}^*_{i,k})}{\sum_l exp(\boldsymbol{\theta}^*_{i,l})}$
- $\forall (i, j) \in \mathcal{E}$
  - draw one-hot membership indicator vector for i when contacting j, $\boldsymbol{z}_{i\to j} \sim Categorical(\boldsymbol{\theta}_i)$
  - draw one-hot membership indicator vector for j when contacted by i, $\boldsymbol{z}_{i\leftarrow j} \sim Categorical(\boldsymbol{\theta}_j)$
  - sample a link between $i \to j$ with probability $\boldsymbol{z}_{i\to j}\boldsymbol{B}\boldsymbol{z}_{i\leftarrow j}$, $Y(i,j) \sim Bernoulli(\boldsymbol{z}_{i\to j}\boldsymbol{B}\boldsymbol{z}_{i\leftarrow j})$

For the first project we only attend to the first two cases, and later on in the third project address the network evolution. To allows for community membership strengths to have correlation, we define instead a logistic normal prior. The hierarchical generative process is explained below.

This indeed comes with some caution, as now the prior is not conjugate to the categorical distribution. We will elaborate more on this when we devise our variational inference engine for estimation of our parameters.

The graphical model for the above algorithm is shown below:

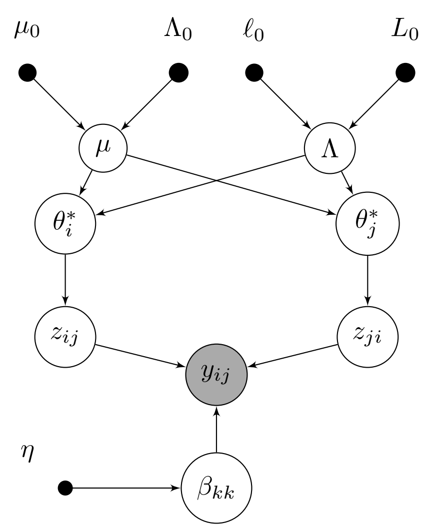

Figure 3: Graphical model for LN-MMSB

4 Variational Inference

We resort to variational inference to uncover intractable distributions, where methods such as MCMC may not be able to recover. In this section we introduce the variational inference (VI) method which transforms the problem of inference to an optimization one, by trying to minimize the Kullback-Leibler divergence between the true posterior distribution and a simpler proposed variational distribution. Hence, instead of making exact inference through stochastic approximation, variational inference uses a deterministic approximation of the model posterior distribution. In its simplest case, the proposed model follows a mean field assumption, which decouples parameters in a way that we can still have tractable and close enough results to the true posterior. For data and all latent variables and parameters, the KL-divergence that is minimized by VI is given by:

$KL\Big(q(Z)||p(Z|X)\Big) = -\mathbb{E}_q\Big[ln\,p(X,Z)\Big] + \mathbb{E}_q\Big[ln\,q(Z)\Big] + ln\,p(X)$

The term $\mathbb{E}_q\Big[ln\,p(X,Z)\Big] - \mathbb{E}_q\Big[ln\,q(Z)\Big]$ is known as the Evidence Lower BOund(ELBO), and since $p(x)$ is independent of $q(z)$ minimizing the KL-divergence is equivalent to a easier optimization problem, maximizing the ELBO.

We can formally define the lower bounds as :

$\mathcal{L} = \mathbb{E}_q\Big[ln\,p(joint)\Big] + H_q[params]$

However this might need the screening of all individual/link level observations for updating the variational parameters. On the other hand, Stochastic Variational Inference (SVI) offers a stochastic search in the parameter space. SVI samples only a small mini-batch, where iterating over the noisy gradients acquired by the sampled batch is proven to converge.There are several sub-sampling schemes, including the link-only sampling which provides efficient inference for undirected networks. Adding community correlation and link direction make the inference problem even more computationally expensive. But using SVI combined with our sampling scheme, allows us to have scalable and efficient inference. Since large networks exhibit very sparse patterns of connections, at each iteration we sample few nodes with all their links and a small proportion of their randomly selected non-links. After rounds of iteration, this assumption both takes into account the information of all links and non-links. The log joint model of data, latent variables and parameters is given below

The log joint probability of the model is defined as

$$\begin{aligned} ln\,p(joint) &= ln\,p(\mu|m_0, M_0) + ln\,p(\Lambda|\ell_0, L_0) + \sum_a ln\,p(\theta_a|\mu,\Lambda) + \sum_a\sum_b ln\,p(z_{a\rightarrow b}|\theta_a) \\ &+ \sum_a\sum_b ln\,p(z_{a\leftarrow b}|\theta_b) + \sum_k ln\,p(\beta_{kk}|\eta) + \sum_a\sum_b ln\,p(y_{ab}|z_{a\rightarrow b}, z_{a\leftarrow b}, \beta) \end{aligned}$$

We further define the variational distribution for each parameter as follows based on the mean field assumption:

$$\begin{aligned} \mu &\sim q(\mu|m, M) \sim \mathcal{N}(\mu|m, M) \\ \Lambda &\sim q(\Lambda|\ell, L) \sim \mathcal{W}(\Lambda|\ell, L) \\ \theta_a &\sim q(\theta_a|\mu_a, \Lambda_a) \sim \mathcal{N}(\theta_a|\mu_a, \Lambda_a) \\ \beta_{kk} &\sim q(\beta_{kk}|b_k) \sim \mathcal{B}(b_{k0}, b_{k1}) \\ z_{a\rightarrow b} &\sim q(z_{a\rightarrow b}|\phi_{a\rightarrow b}) \sim Cat(z_{a\rightarrow b}|\phi_{a\rightarrow b}) \\ z_{a\leftarrow b} &\sim q(z_{a\leftarrow b}|\phi_{a\leftarrow b}) \sim Cat(z_{a\leftarrow b}|\phi_{a\leftarrow b}) \end{aligned}$$

Note that $\phi_{a\rightarrow b}$ and $\phi_{a\leftarrow b}$ are the link level parameters for the Categorical distribution.

To derive the lower bound we first expand the cross entropy term

$$
\begin{aligned}
\mathbb{E}_q\Big[ln\, p(joint)\Big] &= -\tfrac{K}{2} ln\, 2\pi + \tfrac{1}{2} ln\, |M_0| - \tfrac{1}{2}\Big(Tr\, M_0\Big[M^{-1} + (m-m_0)(m-m_0)^T\Big]\Big) \\
&- \tfrac{K(K+1)}{2} ln\, 2 + \tfrac{\ell_0 - K - 1}{2}\psi_K(\tfrac{\ell}{2}) - ln\, \Gamma_K(\tfrac{\ell_0}{2}) - \tfrac{\ell}{2} Tr\,(L_0^{-1} L) - \tfrac{K+1}{2} ln\, |L| + \tfrac{\ell_0}{2} ln\, |L_0^{-1} L| \\
&- \sum_a \tfrac{K}{2} ln\, 2\pi + \tfrac{1}{2}\sum_a \psi_K(\tfrac{\ell}{2}) + \tfrac{1}{2}\sum_a K ln\, 2 + \tfrac{1}{2}\sum_a ln\, |L| \\
&- \tfrac{\ell}{2}\Big(Tr\,\Big[L\Big(\sum_a \big(\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T\big) + \sum_a M^{-1}\Big)\Big\}\Big) \\
&+ \sum_a \sum_b \sum_k \phi_{a\to b,k}\mu_{a,k} - \sum_a \sum_b \mathbb{E}_q[ln\,(\sum_l exp(\theta_{a,l}))] \\
&+ \sum_a \sum_b \sum_k \phi_{a\leftarrow b,k}\mu_{b,k} - \sum_a \sum_b \mathbb{E}_q[ln\,(\sum_l exp(\theta_{b,l}))] \\
&+ \sum_k ln\, \Gamma(\eta_0 + \eta_1) - \sum_k ln\, \Gamma(\eta_0) - \sum_k ln\, \Gamma(\eta_1) + \sum_k (\eta_0 - 1)\psi(b_{k0}) \\
&+ \sum_k (\eta_1 - 1)\psi(b_{k1}) - \sum_k (\eta_0 + \eta_1 - 2)\psi(b_{k0} + b_{k1}) \\
&+ \sum_{a,b \in link} \sum_k \phi_{a\to b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0} + b_{k1}) - ln\, \epsilon\big) + ln\, \epsilon \\
&+ \sum_{a,b \notin link} \sum_k \phi_{a\to b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k1}) - \psi(b_{k0} + b_{k1}) - ln\,(1-\epsilon)\big) + ln\,(1-\epsilon)
\end{aligned}
$$

The negative entropy involving the variational distribution is

$$
\begin{aligned}
H_q[params] &= \tfrac{K}{2} ln\,(2\pi) + \tfrac{K}{2} - \tfrac{1}{2} ln\, |M| \\
&+ \tfrac{K(K+1)}{2} ln\, 2 + \tfrac{K+1}{2} ln\, |L| - \tfrac{\ell - K - 1}{2}\psi_K(\tfrac{\ell}{2}) + ln\, \Gamma_K(\tfrac{\ell}{2}) + \tfrac{K\ell}{2} \\
&+ \sum_a \tfrac{K}{2} ln\,(2\pi) + \sum_a \tfrac{K}{2} - \sum_a \tfrac{1}{2} ln\, |\Lambda_a| \\
&+ \sum_k ln\, \Gamma(b_{k0}) + \sum_k ln\, \Gamma(b_{k1}) - \sum_k ln\, \Gamma(b_{k0} + b_{k1}) - \sum_k (b_{k0} - 1)\psi(b_{k0}) \\
&- \sum_k (b_{k1} - 1)\psi(b_{k1}) + \sum_k (b_{k0} + b_{k1} - 2)\psi(b_{k0} + b_{k1}) \\
&- \sum_a \sum_b \sum_k \phi_{a\to b,k} ln\, \phi_{a\to b,k} \\
&- \sum_a \sum_b \sum_k \phi_{a\leftarrow b,k} ln\, \phi_{a\leftarrow b,k}
\end{aligned}
$$

Adding both together the ELBO follows:

$$
\begin{aligned}
\mathcal{L} = &-\tfrac{1}{2}\Bigg(K ln\, 2\pi - ln\,|M_0| + tr\, M_0(m-m_0)(m-m_0)^T + tr\, M_0 M^{-1}\Bigg) \\
&+ \tfrac{1}{2}\Bigg(- K(K+1) ln\, 2 + (\ell_0 - K - 1)\sum_i \Psi(\tfrac{\ell-i+1}{2}) - \tfrac{K(K-1)}{2} ln\, \pi - 2\sum_i ln\, \Gamma(\tfrac{\ell_0-i+1}{2}) \\
&- \ell tr\,(L_0^{-1}L) - (K+1) ln\,|L| + \ell_0 ln\,|L_0^{-1}L|\Bigg) \\
&- \tfrac{1}{2}\sum_a \Bigg(K ln\, 2\pi - \sum_i \Psi(\tfrac{\ell-i+1}{2}) - K ln\, 2 - ln\,|L| + \\
&\ell tr\,\Big\{L\big[(\mu_a - m)(\mu_a - m)^T + M^{-1} + \Lambda_a^{-1}\big]\Big\}\Bigg) \\
&+ \sum_a \sum_{b \in sink(a)} \Big(\sum_k \phi_{a\to b,k}\mu_{a,k} - ln\sum_l exp(\mu_{a,l} + \tfrac{1}{2}\Lambda_{a,l}^{-1})\Big) \\
&+ \sum_a \sum_{b \notin sink(a)} \Big(\sum_k \phi_{a\to b,k}\mu_{a,k} - ln\sum_l exp(\mu_{a,l} + \tfrac{1}{2}\Lambda_{a,l}^{-1})\Big) \\
&+ \sum_a \sum_{b \in source(a)} \Big(\sum_k \phi_{b\leftarrow a,k}\mu_{a,k} - ln\sum_l exp(\mu_{a,l} + \tfrac{1}{2}\Lambda_{a,l}^{-1})\Big) \\
&+ \sum_a \sum_{b \notin source(a)} \Big(\sum_k \phi_{b\leftarrow a,k}\mu_{a,k} - ln\sum_l exp(\mu_{a,l} + \tfrac{1}{2}\Lambda_{a,l}^{-1})\Big) \\
&+ \sum_k ln\,\Gamma(\eta_0+\eta_1) - \sum_k ln\,\Gamma(\eta_0) - \sum_k ln\,\Gamma(\eta_1) + \sum_k (\eta_0 - 1)\Psi(b_{k0}) + \sum_k (\eta_1 - 1)\Psi(b_{k1}) \\
&- \sum_k (\eta_0 + \eta_1 - 2)\Psi(b_{k0} + b_{k1}) \\
&+ \sum_a \sum_{b \in sink(a)} \sum_k \Big(\phi_{a\to b,k}\phi_{a\leftarrow b,k}(\Psi(b_{k0}) - \Psi(b_{k0}+b_{k1}) - ln\,\epsilon) + ln\,\epsilon\Big) \\
&+ \sum_a \sum_{b \notin sink(a)} \sum_k \Big(\phi_{a\to b,k}\phi_{a\leftarrow b,k}(\Psi(b_{k1}) - \Psi(b_{k0}+b_{k1}) - ln\,(1-\epsilon)) + ln\,(1-\epsilon)\Big) \\
&+ \tfrac{1}{2}\Big(K ln\, 2\pi + K - ln\,|M|\Big) \\
&+ \tfrac{1}{2}\Big((K+1) ln\,|L| + K(K+1) ln\, 2 + \ell K + \tfrac{1}{2}K(K-1) ln\,\pi \\
&+ 2\sum_i ln\,\Gamma(\tfrac{\ell-i+1}{2}) - (\ell - K - 1)\sum_i \Psi(\tfrac{\ell-i+1}{2})\Big) \\
&+ \tfrac{1}{2}\sum_a \Big(K ln\, 2\pi - ln\,|\Lambda_a| + K\Big) \\
&+ \sum_k \Big(ln\,\Gamma(b_{k0}) + ln\,\Gamma(b_{k1}) - ln\,\Gamma(b_{k0}+b_{k1}) - (b_{k0}-1)\Psi(b_{k0}) \\
&- (b_{k1}-1)\Psi(b_{k1}) + (b_{k0}+b_{k1}-2)\Psi(b_{k0}+b_{k1})\Big) \\
&- \sum_a \sum_{b \in sink(a)} \sum_k \Big(\phi_{a\to b,k} ln\,\phi_{a\to b,k}\Big) \\
&- \sum_a \sum_{b \notin sink(a)} \sum_k \Big(\phi_{a\to b,k} ln\,\phi_{a\to b,k}\Big) \\
&- \sum_a \sum_{b \in sink(a)} \sum_k \Big(\phi_{a\leftarrow b,k} ln\,\phi_{a\leftarrow b,k}\Big) \\
&- \sum_a \sum_{b \notin sink(a)} \sum_k \Big(\phi_{a\leftarrow b,k} ln\,\phi_{a\leftarrow b,k}\Big)
\end{aligned}
$$

More information about deriving the cross entropies and entropies are given in the appendix.

5 ELBO Gradients

5.1 Gradient with respect to m

$$
\begin{aligned}
\mathcal{L}_m &= -\tfrac{1}{2}\Big(Tr\, M_0(m-m_0)(m-m_0)^T\Big]\Big) \\
&\quad -\tfrac{\ell}{2}\Big(Tr\, L\Big(\sum_a(\mu_a-m)(\mu_a-m)^T\Big)\Big) \\
&\propto Tr\, M_0(m-m_0)(m-m_0)^T \\
&\quad +\ell\Big(Tr\, L(\sum_a mm^T+\mu_a\mu_a^T-m\mu_a^T-\mu_a m^T)\Big) \\
&= \\
&\Longrightarrow \\
\nabla_m\mathcal{L}_m &\propto 2M_0(m-m_0)-2\ell L\sum_a(\mu_a-m)=0 \\
&\Longrightarrow \\
&\boxed{m=(M_0+N\ell L)^{-1}(M_0m_0+\ell L\sum_a\mu_a)}
\end{aligned}
$$

In minibatch node sampling this would be

$$
\boxed{m=M^{-1}(M_0m_0+\ell L\frac{N}{\#mbnodes}\sum_{a\in mbnodes}\mu_a)}
$$

5.2 Gradient with respect to M

$$
\begin{aligned}
\mathcal{L}_M &= -\tfrac{1}{2}\Big(Tr\, M_0M^{-1}\Big) \\
&\quad -\tfrac{\ell}{2}Tr\, NLM^{-1} \\
&\quad -\tfrac{1}{2}ln\,|M| \\
&\propto Tr\, M_0M^{-1}+\ell Tr\, NLM^{-1}+ln\,|M| \\
&\Longrightarrow \\
\nabla_{M^{-1}}\mathcal{L}_M &= 0 \\
&= -M_0-N\ell L+M=0 \\
&\quad \boxed{M=M_0+N\ell L}
\end{aligned}
$$

5.3 Gradient with respect to L

$$\begin{aligned}
\mathcal{L}_L &= -\tfrac{\ell}{2}Tr\,(L_0^{-1}L) - \tfrac{K+1}{2}ln\,|L| + \tfrac{\ell_0}{2}ln\,|L_0^{-1}L| \\
&\quad +\tfrac{1}{2}\sum_a ln\,|L| - \tfrac{\ell}{2}\Big(Tr\,\Big[L\Big(\sum_a \big(\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T\big) + \sum_a M^{-1}\Big)\Big\}\Big) \\
&\quad +\tfrac{K+1}{2}ln\,|L| \\
&\propto -\ell Tr\,(L_0^{-1}L)\cancel{-(K+1)ln\,|L|} + \ell_0 ln\,|L_0^{-1}L| \\
&\quad +\sum_a ln\,|L| - \ell\Big(Tr\,\Big[L\Big(\sum_a \big(\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T\big) + \sum_a M^{-1}\Big)\Big\}\Big) \\
&\quad \cancel{+(K+1)ln\,|L|} \\
&\Longrightarrow \\
\nabla_L\mathcal{L}_L &= -\ell L_0^{-1} + \tfrac{1}{2}(\ell_0 + N)L^{-1} - \ell\Big(\sum_a \big(\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T\big) + \sum_a M^{-1}\Big)^T = 0 \\
&\quad \ell(L_0^{-1} + \sum_a \Lambda_a^{-1} + \sum_a (\mu_a - m)(\mu_a - m)^T + NM^{-1}) = (N + \ell_0)L^{-1} \\
&\Longrightarrow \boxed{L = \frac{(N + \ell_0)}{\ell}\Big((L_0^{-1} + \sum_a \big(\Lambda_a^{-1} + (\mu_a - m)(\mu_a - m)^T\big) + \sum_a M^{-1})\Big)^{-1}}
\end{aligned}$$

optimizing simultaeneously with ℓ in the minibatch setting:

$$\boxed{L = \Big((L_0^{-1} + \frac{N}{\#mbnodes}\{\sum_a \Lambda_a^{-1} + \sum_a (\mu_a - m)(\mu_a - m)^T\} + NM^{-1})\Big)^{-1}}$$

5.4 Gradient with respect to ℓ

$$\begin{aligned}
\mathcal{L}_\ell &= revise \\
&\propto \\
&\Longrightarrow \\
&\propto \\
&\Longrightarrow \\
\nabla_\ell\mathcal{L}_\ell &= \\
&\Longrightarrow \\
&hence, \\
&\Longrightarrow \\
&\qquad \boxed{\ell = \ell_0 + N}
\end{aligned}$$

5.5 Gradient with respect to b_k

$$\mathcal{L}_{b_k} = revise$$

$$\begin{aligned}
& \text{simultaenously optimizing } b_{k0}, b_{k1} \\
\implies & \text{Similar to our previous results} \\
\nabla_{b_{k0}} \mathcal{L}_{b_k} &= 0 \\
\implies & \boxed{b_{k0} = \eta_0 + \frac{\#trainlinks}{\#mblinks} \sum_{a,b \in mblinks} \phi_{a\rightarrow b,k}\phi_{a\leftarrow b,k}} \\
\nabla_{b_{k1}} \mathcal{L}_{b_k} &= 0 \\
& \boxed{b_{k1} = \eta_1 + \frac{\#trainnonlinks}{\#mbnonlinks} \sum_{a,b \notin mblinks} \phi_{a\rightarrow b,k}\phi_{a\leftarrow b,k}}
\end{aligned}$$

5.6 Gradient with respect to $\phi_{a\rightarrow b,k}$ for links

$$\begin{aligned}
\mathcal{L}_{\phi_{a\rightarrow b,k}} &= \phi_{a\rightarrow b,k}\mu_{a,k} \\
&\quad + \phi_{a\rightarrow b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big) \\
&\quad - \phi_{a\rightarrow b,k} ln\,\phi_{a\rightarrow b,k} \\
&= \phi_{a\rightarrow b,k}\Big(\mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big) - ln\,\phi_{a\rightarrow b,k}\Big) \\
\nabla_{\phi_{a\rightarrow b,k}} \mathcal{L}_{\phi_{a\rightarrow b,k}} &= \mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big) - ln\,\phi_{a\rightarrow b,k} = 0 \\
& \boxed{\phi_{a\rightarrow b,k} \propto exp\Big\{\mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big)\Big\}}
\end{aligned}$$

5.7 Gradient with respect to $\phi_{a\leftarrow b,k}$ for links

$$\begin{aligned}
\mathcal{L}_{\phi_{a\leftarrow b,k}} &= \phi_{a\leftarrow b,k}\mu_{b,k} \\
&\quad + \phi_{a\rightarrow b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big) \\
&\quad - \phi_{a\leftarrow b,k} ln\,\phi_{a\leftarrow b,k} \\
&= \phi_{a\leftarrow b,k}\Big(\mu_{b,k} + \phi_{a\rightarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big) - ln\,\phi_{a\leftarrow b,k}\Big) \\
\nabla_{\phi_{a\leftarrow b,k}} \mathcal{L}_{\phi_{a\leftarrow b,k}} &= \mu_{b,k} + \phi_{a\rightarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big) - ln\,\phi_{a\leftarrow b,k} = 0 \\
& \boxed{\phi_{a\leftarrow b,k} \propto exp\Big\{\mu_{b,k} + \phi_{a\rightarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big)\Big\}}
\end{aligned}$$

5.8 Gradient with respect to $\phi_{a\rightarrow b,k}$ for nonlinks

$$\begin{aligned}
\mathcal{L}_{\phi_{a\rightarrow b,k}} &= \phi_{a\rightarrow b,k}\mu_{a,k} \\
&\quad + \phi_{a\rightarrow b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k1}) - \psi(b_{k0}+b_{k1}) - ln\,(1-\epsilon)\big) \\
&\quad - \phi_{a\rightarrow b,k} ln\,\phi_{a\rightarrow b,k} \\
&= \phi_{a\rightarrow b,k}\Big(\mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k1}) - \psi(b_{k0}+b_{k1}) - ln\,(1-\epsilon)\big) - ln\,\phi_{a\rightarrow b,k}\Big) \\
\nabla_{\phi_{a\rightarrow b,k}} \mathcal{L}_{\phi_{a\rightarrow b,k}} &= \mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k1}) - \psi(b_{k0}+b_{k1}) - ln\,(1-\epsilon)\big) - ln\,\phi_{a\rightarrow b,k} = 0 \\
& \boxed{\phi_{a\rightarrow b,k} \propto exp\Big\{\mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k1}) - \psi(b_{k0}+b_{k1}) - ln\,(1-\epsilon)\big)\Big\}}
\end{aligned}$$

5.9 Gradient with respect to $\phi_{a\leftarrow b,k}$ for nonlinks

$$\begin{aligned}
\mathcal{L}_{\phi_{a\leftarrow b,k}} &= \phi_{a\leftarrow b,k}\mu_{b,k} \\
&\quad +\phi_{a\rightarrow b,k}\phi_{a\leftarrow b,k}\big(\psi(b_{k1})-\psi(b_{k0}+b_{k1})-ln\,(1-\epsilon)\big) \\
&\quad -\phi_{a\leftarrow b,k} ln\,\phi_{a\leftarrow b,k} \\
&= \phi_{a\leftarrow b,k}\Big(\mu_{b,k}+\phi_{a\rightarrow b,k}\big(\psi(b_{k1})-\psi(b_{k0}+b_{k1})-ln\,(1-\epsilon)\big)-ln\,\phi_{a\leftarrow b,k}\Big) \\
\nabla_{\phi_{a\leftarrow b,k}}\mathcal{L}_{\phi_{a\leftarrow b,k}} &= \mu_{b,k}+\phi_{a\rightarrow b,k}\big(\psi(b_{k1})-\psi(b_{k0}+b_{k1})-ln\,(1-\epsilon)\big)-ln\,\phi_{a\leftarrow b,k}=0 \\
&\boxed{\phi_{a\leftarrow b,k}\propto exp\Big\{\mu_{b,k}+\phi_{a\rightarrow b,k}\big(\psi(b_{k1})-\psi(b_{k0}+b_{k1})-ln\,(1-\epsilon)\big)\Big\}}
\end{aligned}$$

5.10 Gradient with respect to μ_a

μ_a and Λ_a are two of the scarier ones.

$$\begin{aligned}
\mathcal{L}_{\mu_a} &= -\tfrac{\ell}{2}\big[(\mu_a-m)^T L(\mu_a-m)\big]+ \\
&\quad \sum_{b\in sink(a)}\phi^T_{a\rightarrow b}\mu_a + \\
&\quad \sum_{b\notin sink(a)}\phi^T_{a\rightarrow b}\mu_a + \\
&\quad \sum_{b\in source(a)}\phi^T_{b\leftarrow a}\mu_a + \\
&\quad \sum_{b\notin source(a)}\phi^T_{b\leftarrow a}\mu_a - \\
&\quad \sum_{b} log\Big(\mathbf{1}^T\underline{\mathrm{f}}(\mu_a,\Lambda_a)\Big)
\end{aligned}$$

where $\underline{\mathrm{f}}(\mu_a,\Lambda_a)=\begin{pmatrix} exp(\mu_{a,1}+\frac{1}{2}\Lambda^{-1}_{a,1}) \\ \vdots \\ exp(\mu_{a,k}+\frac{1}{2}\Lambda^{-1}_{a,k}) \\ \vdots \\ exp(\mu_{a,K}+\frac{1}{2}\Lambda^{-1}_{a,K})\end{pmatrix}$, and we may for convenience interchangably use $\underline{\mathrm{f}}_a$ to refer to $\underline{\mathrm{f}}(\mu_a,\Lambda_a)$:

Hence the geradient is

$$
\begin{aligned}
\nabla_{\mu_a}\mathcal{L}_{\mu_a} = & \quad -\ell L(\mu_a - m)+ \\
& \sum_{b \in sink(a)} \phi_{a\to b} + \sum_{b \notin sink(a)} \phi_{a \to b} + \\
& \sum_{b \in source(a)} \phi_{b \leftarrow a} + \sum_{b \notin source(a)} \phi_{b \leftarrow a} - \\
& \sum_b \frac{\partial \underline{\mathrm{f}}(\mu_a, \Lambda_a)}{\partial \mu_a}(\mathbf{1}) \\
= & \quad -\ell L(\mu_a - m)+ \\
& \sum_{b \in sink(a)} \phi_{a\to b} + \sum_{b \notin sink(a)} \phi_{a \to b} + \\
& \sum_{b \in source(a)} \phi_{b \leftarrow a} + \sum_{b \notin source(a)} \phi_{b \leftarrow a} - \\
& \sum_b \frac{\boldsymbol{J}_{\underline{\mathrm{f}}} \times \mathbf{1}}{\mathbf{1}^T \underline{\mathrm{f}}(\mu_a, \Lambda_a)} \\
= & \quad -\ell L(\mu_a - m)+ \\
& \sum_{b \in sink(a)} \phi_{a\to b} + \sum_{b \notin sink(a)} \phi_{a \to b} + \\
& \sum_{b \in source(a)} \phi_{b \leftarrow a} + \sum_{b \notin source(a)} \phi_{b \leftarrow a} - \\
& \sum_b \frac{\begin{pmatrix}
\frac{\partial \underline{\mathrm{f}}_{a1}}{\partial \mu_{a1}} & \cdots & \frac{\partial \underline{\mathrm{f}}_{a1}}{\partial \mu_{ak}} & \cdots & \frac{\partial \underline{\mathrm{f}}_{a1}}{\partial \mu_{aK}} \\
\vdots & \ddots & & & \vdots \\
\frac{\partial \underline{\mathrm{f}}_{ak}}{\partial \mu_{a1}} & \cdots & \frac{\partial \underline{\mathrm{f}}_{ak}}{\partial \mu_{ak}} & \cdots & \frac{\partial \underline{\mathrm{f}}_{ak}}{\partial \mu_{aK}} \\
\vdots & & & \ddots & \vdots \\
\frac{\partial \underline{\mathrm{f}}_{aK}}{\partial \mu_{a1}} & & \cdots & & \frac{\partial \underline{\mathrm{f}}_{aK}}{\partial \mu_{aK}}
\end{pmatrix}}{\mathbf{1}^T \underline{\mathrm{f}}(\mu_a, \Lambda_a)} \\
= & \quad -\ell L(\mu_a - m)+ \\
& \sum_{b \in sink(a)} \phi_{a\to b} + \sum_{b \notin sink(a)} \phi_{a \to b} + \\
& \sum_{b \in source(a)} \phi_{b \leftarrow a} + \sum_{b \notin source(a)} \phi_{b \leftarrow a} - \\
& \sum_b \underline{\mathrm{sfx}}(a)
\end{aligned}
$$

where $\underline{\mathrm{sfx}}(a) = \begin{pmatrix}
\frac{exp(\mu_{a,1} + \frac{1}{2}\,\Lambda^{-1}_{a,1})}{\sum_l exp(\mu_{a,l} + \frac{1}{2}\,\Lambda^{-1}_{a,l})} \\
\vdots \\
\frac{exp(\mu_{a,k} + \frac{1}{2}\,\Lambda^{-1}_{a,k})}{\sum_l exp(\mu_{a,l} + \frac{1}{2}\,\Lambda^{-1}_{a,l})} \\
\vdots \\
\frac{exp(\mu_{a,1} + \frac{1}{2}\,\Lambda^{-1}_{a,1})}{\sum_l exp(\mu_{a,l} + \frac{1}{2}\,\Lambda^{-1}_{a,l})}
\end{pmatrix}$

so all in all the gradient is :

$$
\nabla_{\mu_a}\mathcal{L}_{\mu_a} = \boxed{-\ell L(\mu_a - m) + \sum_{b \in sink(a)} \phi_{a\to b} + \sum_{b \notin sink(a)} \phi_{a \to b} + \sum_{b \in source(a)} \phi_{b \leftarrow a} + \sum_{b \notin source(a)} \phi_{b \leftarrow a} - \sum_b \underline{\mathrm{sfx}}(a)}
$$

Similarly the Hessian will be as follows:

$$
\begin{aligned}
\nabla^2_{\mu_a}\mathcal{L}_{\mu_a} = & \quad -\ell L- \\
& \sum_b \frac{\partial \underline{\text{sfx}}(a)}{\partial \mu_a^T} \\
= & \quad \ell L- \\
& \sum_b \boldsymbol{J}_{\underline{\text{sfx}}(a)} \\
= & \quad -\ell L- \\
& \sum_b \begin{pmatrix} \frac{\partial \underline{\text{sfx}}_{a1}}{\partial \mu_{a1}} & \dots & \frac{\partial \underline{\text{sfx}}_{a1}}{\partial \mu_{ak}} & \dots & \frac{\partial \underline{\text{sfx}}_{a1}}{\partial \mu_{aK}} \\ \vdots & \ddots & & & \vdots \\ \frac{\partial \underline{\text{sfx}}_{ak}}{\partial \mu_{a1}} & \dots & \frac{\partial \underline{\text{sfx}}_{ak}}{\partial \mu_{ak}} & \dots & \frac{\partial \underline{\text{sfx}}_{ak}}{\partial \mu_{aK}} \\ \vdots & & & \ddots & \vdots \\ \frac{\partial s\underline{\text{fx}}_{aK}}{\partial \mu_{a1}} & & \dots & & \frac{\partial \underline{\text{sfx}}_{aK}}{\partial \mu_{aK}} \end{pmatrix} \\
= & \quad -\ell L- \\
& \sum_b \begin{pmatrix} \underline{\text{sfx}}_{a1} - \underline{\text{sfx}}^2_{a1} & \dots & -\underline{\text{sfx}}_{a1}\underline{\text{sfx}}_{ak} & \dots & \underline{\text{-sfx}}_{a1}\underline{\text{sfx}}_{aK} \\ \vdots & \ddots & & & \vdots \\ \underline{\text{-sfx}}_{a1}\underline{\text{sfx}}_{ak} & \dots & \underline{\text{sfx}}_{ak} - \underline{\text{sfx}}^2_{ak} & \dots & \underline{\text{-sfx}}_{ak}\underline{\text{sfx}}_{ak} \\ \vdots & & & \ddots & \vdots \\ -\underline{\text{sfx}}_{a1}\underline{\text{sfx}}_{aK} & & \dots & & \underline{\text{-sfx}}_{aK} - \underline{\text{sfx}}^2_{aK} \end{pmatrix} \\
= & \quad \boxed{-\ell L - \sum_b \Big(diagm(\underline{\text{sfx}}_a) - \underline{\text{sfx}}_a\underline{\text{sfx}}_a^T\Big)}
\end{aligned}
$$

5.11 Gradient with respect to Λ_a

similarly assuming that Λ_a is a diagonal matrix(or a column vector).

$$
\begin{aligned}
\mathcal{L}_{\Lambda_a^{-1}} = & \quad -\tfrac{\ell}{2}diag\,(L)'\Lambda_a^{-1} + \tfrac{1}{2}ln\,|diagm(\Lambda_a^{-1})| - \sum_b log\,\Big(\mathbf{1}^T\underline{\text{f}}(\mu_a,\Lambda_a)\Big) \\
= & \\
\nabla_{\Lambda_a^{-1}}\mathcal{L}_{\Lambda_a^{-1}} = G_{\Lambda_a^{-1}} = & \quad \boxed{-\tfrac{\ell}{2}diag(L) + \tfrac{1}{2}(\Lambda_a) - \tfrac{1}{2}\sum_b(\underline{\text{sfx}}(a))}
\end{aligned}
$$

$\square$

$$
\nabla^2_{\Lambda_a^{-1}}\mathcal{L}_{\Lambda_a^{-1}} = H_{\Lambda_a^{-1}} \propto \quad \boxed{-\tfrac{1}{2}diagm(\Lambda_a \odot \Lambda_a) - \tfrac{1}{4}\sum_b\Big(diagm(\underline{\text{sfx}}_a) - \underline{\text{sfx}}_a\underline{\text{sfx}}_a^T\Big)}
$$

We use only the first moment and use Adagrad to optimize for this parameter.

The above derivations are for the general case of directed networks. With minor tweaks, we can apply the same model for the simpler case of undirected graphs. In this scenario, there is no difference between $\phi_{a\to b,k}$ and $\phi_{a\leftarrow b,k}$ for links and hence we represent them as $\phi_{ab,k}$. The gradient update for $\phi_{ab,k}$for this case results in

$$
\phi_{ab,k} \propto exp\Bigg\{\mu_{a,k} + \mu_{b,k} + \psi(b_{k0}) - \psi(b_{k0} + b_{k1})\Big)\Bigg\}
$$

We should also note that for nonlinks, $\phi_{a\to b,k} = \phi_{b\leftarrow a,k}$ and $\phi_{b\to a,k} = \phi_{a\leftarrow b,k}$

Part II
Social Influence and Latent Homophily

In this project we aim to exploit the behavioral data in better detecting communities and gaining insights into labeling them. We can furthermore assess the improve the prediction of a specific behavior. This can also shed lights in redefining opinion leadership in social networks. In this scenario the latent space is not just driven by the preferences for tie formation but also by behavioral tendencies. Hence the latent space of behavioral and structural preferences can overlap. For this, we suggest a joint modeling approach for both behavior and link formation. The idea of this latent space is shown below.

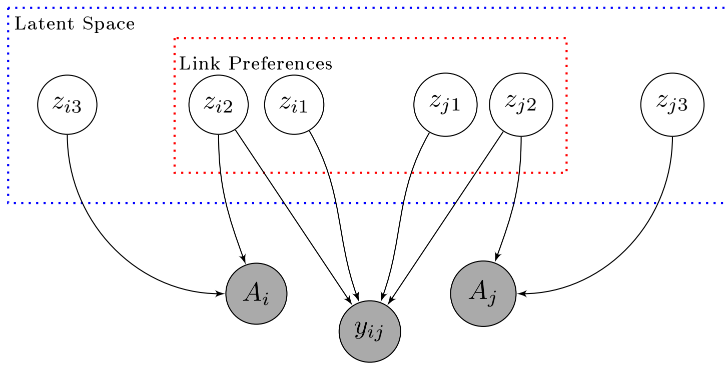

Figure 4: Graphical representation of network structure and behavior

Behaviors cluster in space and in time. This can be attributed to both social influence of individuals and also homophily phenomenon(and also external shocks). In the presence of the social influence the graph is modified in the following way.

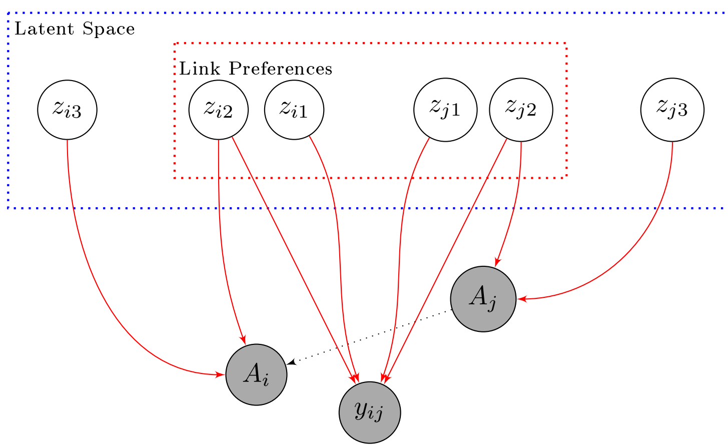

Figure 5: Presence of social influence

However the effect of the social influence cannot be identified due to the confounding path. Our solution with joint modelling of the latent space allows us to find observed proxies for the behavioral and structural preferences, and controlling for that would also allow us to identify the peer effect. The general idea is graphically shown below:

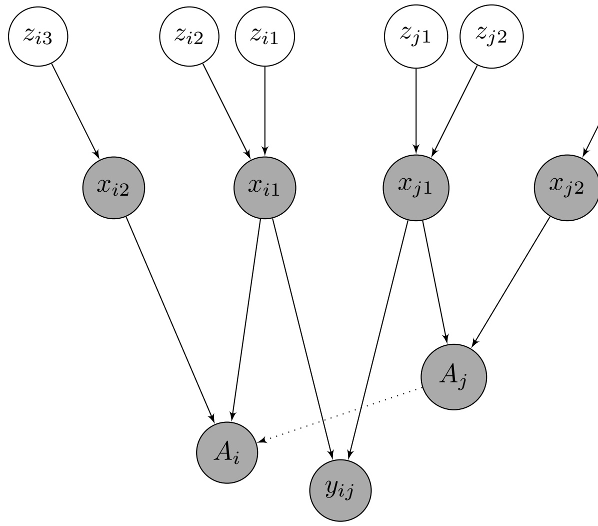

Figure 6: Solution to the confounding path

Specifically, in practice we modify the graph as follows:

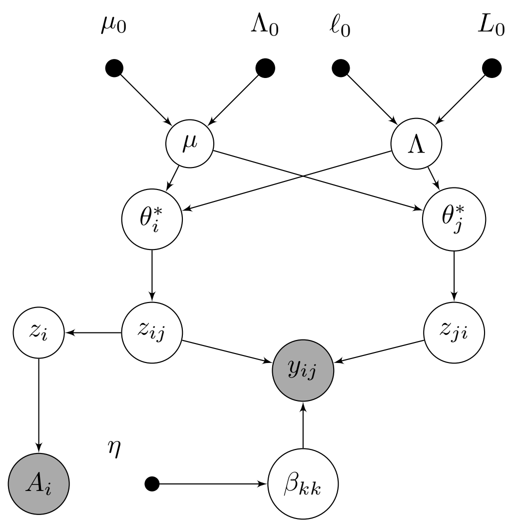

Figure 7: Modified graph for identification

Part III

Dynamic Network evolution of structure and behavior

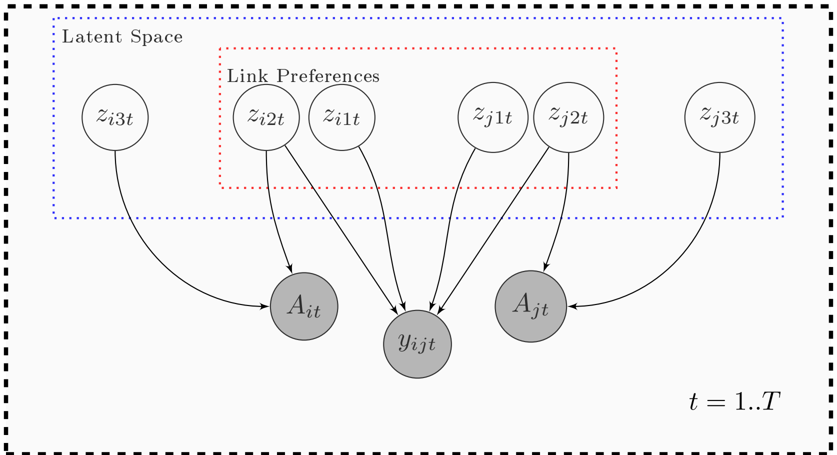

Figure 8: Dynamic evolution of network and behavior

Appendix

A Negative cross entropies

A.1 Two Normals

Note:All the normals are parametrized using the precision matrix.
$q \sim \mathcal{N}(x|m, L)$
$p \sim \mathcal{N}(x|\mu, \Lambda)$

$$\begin{aligned}
\int q(x) ln\, p(x) dx &= \int \mathcal{N}(x|m, L)\Bigg(-\tfrac{K}{2} ln\, 2\pi + \tfrac{1}{2} ln\, |\Lambda| - \tfrac{1}{2}\Big(Tr\, \Lambda\{(x-\mu)(x-\mu)^T\}\Big)\Bigg) dx \\
&= -\tfrac{K}{2} ln\, 2\pi + \tfrac{1}{2} ln\, |\Lambda| + \int \mathcal{N}(x|m, L)\Bigg(-\tfrac{1}{2}\Big(Tr\, \Lambda\{(x-\mu)(x-\mu)^T\}\Big)\Bigg) dx \\
&= -\tfrac{K}{2} ln\, 2\pi + \tfrac{1}{2} ln\, |\Lambda| + \int \mathcal{N}(x|m, L)\Bigg(-\tfrac{1}{2}\Big(Tr \Lambda\{xx^T + \mu\mu^T - x\mu^T - \mu x^T\}\Big)\Bigg) dx
\end{aligned}$$

We should note that $\mathbb{E}_q\Big[xx^T\Big] = Cov_q + \mathbb{E}_q\Big[x\Big]\mathbb{E}_q\Big[x\Big]^T$
$\mathbb{E}_q\Big[x\Big] = m$ and $Cov_q = L^{-1}$

$$\begin{aligned}
\int \mathcal{N}(x|m, L)\Bigg(-\tfrac{1}{2}\Big(Tr\, \Big[\Lambda\{xx^T + \mu\mu^T - x\mu^T - \mu x^T\}\Big]\Big)\Bigg) dx &= -\tfrac{1}{2} Tr\, \Big[(\Lambda L^{-1} + \Lambda m m^T) + \Lambda(mm^T - \mu m^T - m\mu^T)\Big] \\
&= -\tfrac{1}{2}\Big(Tr\, \Big[\Lambda L^{-1}\Big] + (m-\mu)^T \Lambda (m-\mu)\Big)
\end{aligned}$$

Hence we have:

$$\boxed{\mathbb{E}_q[ln\, p(x)] = -\tfrac{K}{2} ln\, 2\pi + \tfrac{1}{2} ln\, |\Lambda| - \tfrac{1}{2}\Big(Tr\, \Big[\Lambda L^{-1}\Big] + (m-\mu)^T \Lambda (m-\mu)\Big)}$$

A.2 Two Wisharts

$\Lambda \sim q \sim \mathcal{W}(v, W)$
$\Lambda \sim p \sim \mathcal{W}(n, S)$

$$\begin{aligned}
\int q(\Lambda) ln\, p(\Lambda) d\Lambda &= \mathbb{E}_q[ln\, p(\Lambda)] \\
&= \mathbb{E}_q\left[ln\, \frac{|\Lambda|^{\frac{n-K-1}{2}} exp(-\frac{1}{2}Tr\,(S^{-1}\Lambda)}{2^{\frac{nK}{2}}|S|^{n/2}\Gamma_p(\frac{n}{2})}\right] \\
&= \mathbb{E}_q\Bigg[-\tfrac{nk}{2}ln\,2 - \tfrac{n}{2}ln\,|S| - ln\,\Gamma_K(\tfrac{n}{2}) \\
&\quad + \tfrac{n-K-1}{2}ln\,|\Lambda| - \tfrac{1}{2}Tr\,(S^{-1}\Lambda)\Bigg] \\
&= -\tfrac{nk}{2}ln\,2 - \tfrac{n}{2}ln\,|S| - ln\,\Gamma_K(\tfrac{n}{2}) \\
&\quad + \tfrac{n-K-1}{2}\Big(\psi_K(\tfrac{v}{2}) + Kln\,2 + ln\,|W|\Big) - \tfrac{v}{2}Tr\,(S^{-1}W)
\end{aligned}$$

Note that:
$\mathbb{E}_q[\Lambda] = vW$
$\mathbb{E}_q[ln\,|\Lambda|] = \psi_K(\frac{v}{2}) + Kln\,2 + ln\,|W|$
$\psi_K(\frac{v}{2}) = \sum_{i:1}^{K}\psi(\frac{v-i+1}{2})$
$ln\,\Gamma_K(\frac{n}{2}) = \frac{K(K-1)}{4}ln\,\pi + \sum_{i:1}^{K} ln\,\Gamma(\frac{n-i+1}{2})$

$$\begin{aligned}
\mathbb{E}_q[ln\,p(\Lambda)] &= -\tfrac{K(K+1)}{2}ln\,2 + \tfrac{n-K-1}{2}\psi_K(\tfrac{v}{2}) - ln\,\Gamma_K(\tfrac{n}{2}) \\
&\quad -\tfrac{v}{2}Tr\,(S^{-1}W) + \tfrac{n-K-1}{2}ln\,|W| - \tfrac{n}{2}ln\,|S|
\end{aligned}$$

so we have:
$\boxed{\mathbb{E}_q[ln\,p(\Lambda)] = -\frac{K(K+1)}{2}ln\,2 + \frac{n-K-1}{2}\psi_K(\frac{v}{2}) - ln\,\Gamma_K(\frac{n}{2}) - \frac{v}{2}Tr\,(S^{-1}W) + \frac{n-K-1}{2}ln\,|W| - \frac{n}{2}ln\,|S|}$
or
$\boxed{\mathbb{E}_q[ln\,p(\Lambda)] = -\frac{K(K+1)}{2}ln\,2 + \frac{n-K-1}{2}\psi_K(\frac{v}{2}) - ln\,\Gamma_K(\frac{n}{2}) - \frac{v}{2}Tr\,(S^{-1}W) - \frac{K+1}{2}ln\,|W| + \frac{n}{2}ln\,|S^{-1}W|}$

A.3 Two Betas

$\beta \sim q \sim Beta(b)$
$\beta \sim p \sim Beta(\eta)$

$$\begin{aligned}
\mathbb{E}_q[ln\,p(\beta)] &= \mathbb{E}_q\Big[ln\,\Gamma(\eta_0+\eta_1) - ln\,\Gamma(\eta_0) - ln\,\Gamma(\eta_1) + (\eta_0-1)ln\,\beta + (\eta_1-1)ln\,(1-\beta)\Big] \\
&= ln\,\Gamma(\eta_0+\eta_1) - ln\,\Gamma(\eta_0) - ln\,\Gamma(\eta_1) + (\eta_0-1)\big(\psi(b_0)-\psi(b_0+b_1)\big) + (\eta_1-1)\big(\psi(b_1)-\psi(b_0+b_1)\big) \\
&= ln\,\Gamma(\eta_0+\eta_1) - ln\,\Gamma(\eta_0) - ln\,\Gamma(\eta_1) + (\eta_0-1)\psi(b_0) + (\eta_1-1)\psi(b_1) - (\eta_0+\eta_1-2)\psi(b_0+b_1)
\end{aligned}$$

Note that $\mathbb{E}_q[ln\,\beta] = \psi(b_0) - \psi(b_0+b_1)$
so :
$\boxed{\mathbb{E}_q[ln\,p(\beta)] = ln\,\Gamma(\eta_0+\eta_1) - ln\,\Gamma(\eta_0) - ln\,\Gamma(\eta_1) + (\eta_0-1)\psi(b_0) + (\eta_1-1)\psi(b_1) - (\eta_0+\eta_1-2)\psi(b_0+b_1)}$

B Entropies

B.1 Normal

$q(x) \sim \mathcal{N}(m, M)$
$\boxed{H[q] = \frac{K}{2}ln\,(2\pi) + \frac{K}{2} - \frac{1}{2}ln\,|M|}$

B.2 Wishart

$\Lambda \sim q \sim \mathcal{W}(v, W)$

$$\begin{aligned} H[q] &= -\tfrac{v-K-1}{2}\mathbb{E}_q ln|\Lambda| - (-\tfrac{1}{2}\mathbb{E}_q Tr\,(W^{-1}\Lambda)) + \tfrac{v}{2}ln\,|W| + \tfrac{vK}{2}ln\,2 + ln\,\Gamma_K(\tfrac{v}{2}) \\ &= -\tfrac{v-K-1}{2}(\psi_K(\tfrac{v}{2}) + \tfrac{Kv}{2} + Kln\,2 + ln\,|W|) + \tfrac{v}{2}ln\,|W| + \tfrac{vK}{2}ln\,2 + ln\,\Gamma_K(\tfrac{v}{2}) \\ &= \tfrac{K(K+1)}{2}ln\,2 + \tfrac{K+1}{2}ln\,|W| - \tfrac{v-K-1}{2}\psi_p(\tfrac{v}{2}) + ln\,\Gamma_K(\tfrac{v}{2}) + \tfrac{Kv}{2} \end{aligned}$$

so

$$\boxed{H[q] = \tfrac{K(K+1)}{2}ln\,2 + \tfrac{K+1}{2}ln\,|W| - \tfrac{v-K-1}{2}\psi_K(\tfrac{v}{2}) + ln\,\Gamma_K(\tfrac{v}{2}) + \tfrac{Kv}{2}}$$

B.3 Beta

$\beta \sim q \sim Beta(b)$

$$\begin{aligned} H[q] &= ln\,\Gamma(b_0) + ln\,\Gamma(b_1) - ln\,\Gamma(b_0 + b_1) - (b_0 - 1)\mathbb{E}_q[ln\,\beta] - (b_1 - 1)\mathbb{E}_q[ln\,(1-\beta)] \\ &= ln\,\Gamma(b_0) + ln\,\Gamma(b_1) - ln\,\Gamma(b_0 + b_1) - (b_0 - 1)\psi(b_0) - (b_1 - 1)\psi(b_1) + (b_0 + b_1 - 2)\psi(b_0 + b_1) \end{aligned}$$

So,

$$\boxed{H[q] = ln\,\Gamma(b_0) + ln\,\Gamma(b_1) - ln\,\Gamma(b_0 + b_1) - (b_0 - 1)\psi(b_0) - (b_1 - 1)\psi(b_1) + (b_0 + b_1 - 2)\psi(b_0 + b_1)}$$

B.4 Multinomial(,1) or Categorical

$z \sim q \sim Cat(\phi)$

$$H[q] = -\sum_k \mathbb{E}_q[z_k] ln\,\phi_k$$

so,

$$\boxed{H[q] = -\sum_k \phi_k ln\,\phi_k}$$