Rough Chapter 1 and 2

Introduction

Extensive studies of social contagion in marketing, have dealt with the partial role of structure in studying influence or disentangling causality from homophily[3, 4, 27, 71]. However sociometric data could prove to be useful. Goldenberg et al (2009), show that actors with higher degree centrality(hubs) can speed up contagion or lead to higher volume diffusion depending on being innovative or follower[27]. Goel and Goldstein(2014) have shown that regardless of the causal nature of influence and possible selection biases in clustering of behaviors, social data can prove to be beneficial and complementary to behavioral data in terms of predicting the future actions[26]. Hence a good knowledge on the mechanics of structure formation in social networks could be vital in further analyzing the behavioral data. However many of these studies only account for network level characteristics of individuals, such as degree centrality,betweenness, prestige, clustering coefficient, etc[31]. Although these measures are still valuable in predicting and studying the diffusion, and are more important than self-reported measures of opinion leadership[37], in this paper we argue that sub-network level measures can provide further insights into recognizing more fine-grained centrality measures, and into recognizing topic related brokerage roles. Moreover, marketers are always interested in finding influential people in a network to be able to target them and segment them properly[4, 37, 69, 68, 71, 70, 42]; In addition to what previous literature has mostly involved with in recognizing the role of connectors, we believe the role of possible other individuals such as mavens and salesmen have been less attended in the context of social data[25]. Yet not all centrality measures on the network level correspond to similar notions and could provide even opposite results[68, 69]. Hence delving into more robust constructs at the community level might be a worthwhile investigation into the potential roles of specific individuals that otherwise would have not been identified. Detecting these sub-network properties can facilitate understanding the likelihood of information exchange and the attention given to information[80]. Several studies have emerged in the field of marketing and management that pinpoint the importance of communities underlying social networks in better understanding consumer-firm or consumer-consumer relationships. Ansari et al (2011) model a multiplex network of professionals to simultaneously study the impact of the organizational interventions on the nature of the connections[2]. Ma et al (2014) use communities to account for homophily when studying the social influence of decision purchases and timing of individuals in a mobile network [50].

Many real-life social and other types of networks of individuals, however, are more accurately described by multiple interconnected networks instead of one *homogeneous* network. While theoretically we still expect much predictive ability from the network measures studied till now[26], empirically this impact will be blurred because the network is in fact a mixture of multiple networks, leaving the empirical assessment of network measures a shaky foundation. Identifying these constructs could still be valuable even in the absence of behavioral data. For example accounting for shared communities in a directed networks, can signal information about the identification and potential power of information brokers sitting at the edge of the groups[29, 54]. For those individuals belonging to several communities, it is important to account for the size of their incoming connections given their outgoing links. This can shed light on their potential brokerage power on specific topic communities??.

Another missing piece in many network studies in the field of marketing is that most measures are only evaluated in small scale networks. A related stream of diffusion research, studies the contagion phenomenon in new prescription drugs among a network of physicians[37, 70, 71]. Other examples, and larger networks, Aral, ...????.Some skip latent homophily, Aral, Snijders, resolve with experiment Aral, ... why it is important(Shalizi).Among studies that incorporate large scale networks are ([9],?) .Braun and Bonfrer (2011) use a latent space approach to map individual latent traits to a Euclidean space. In latent space modeling, individual characteristics are represented as hidden structures that need to be estimated from the network data, where potentially infinite dimensional individual characteristics are mapped to a lower dimensional Euclidean space; and the similarity is measured by the distance between individual locations in this low dimensional space[33, 9, 2]. Braun and Bonfrer (2011) argue that latent space approach could provide very different insights compared to geodesic distances,especially when evaluating the reach. More individuals could potentially and better semantically be reached through a small radius around a focal node in a low dimensional space rather than accommodating first and second degree connections of that focal node that may end up to be irrelevant.This again suggests how strategies can lead to different results when social data are applied differently through community detection and geodesic reach[9]. While latent (Euclidean) space models[9, 2, 30, 33] account for additional structure in the network formation process, they do not provide access to network measures for various sub-networks as they do not leave much for interpretation when representing the network in a Euclidean space.

Ample amount of research has recognized a prevalent feature that networks exert meaningful smaller groups[24, 21, 58]. These groups namely communities/clusters share the property that within group connections among the vertices of the graph are denser compared to the between group connections.[5, 58].

Traditionally the problem of clustering networks included partitioning graphs into separate groups[56, 67], whereas more recent works represent the embodiment of more natural processes such as shared community memberships[78, 76, 46, 1].Although caution must be taken regarding how one defines communities, as the concept could vary dramatically depending on the given context. A common phenomenon observed in many social networks including multilevel relationships signifies shared community memberships, hence in this paper we define communities as groups of individuals yielding a better understanding of the network connections, where individual vertices can belong to multiple clusters. Many different methods regarding community detection have been developed among which address the problem in terms of algorithmic[57, 60, 80] or model-based(probabilistic) approaches[1, 30, 33].

Implications of finding overlapping communities for network research can be manifold. To gain better understanding of diffusion of ideas(),products[70, 4],medical innovation[71, 17], one has to be able to acknowledge different sources of contagion. Behaviors and decisions made by many individuals in observed networks tend to assimilate both in node space and in time[3]. However disentangling the underlying reasons can become infeasible due to endogenous network formation portrayed by latent homophily[65]. Conveniently addressing latent homophily and using a proxy estimation could improve the estimation of influence versus homophily[65, 64, 18]. We attend to this problem furthermore in chapter 3 and chapter 4.Maybe this paragraph should go before scalability.

On the other hand many evidence from real world networks show signs of evolution of networks in time[12, 39, 66]. This means that at certain times in the network some ties are formed and some are severed. Brot et al(2016) argue that this happens when observing bursts in connection that underlies the change in homophily[11]. Hence trying to explain these rather stochastic changes requires a better understanding of homophily and formation of networks. It is not clear to what extent, commonly known measures such as degree centrality, betweenness, prestige, etc are able to discover influentials, a concept that can depend on how one defines the communities at hand[38]. Chen et al (2017) suggest that incorporating connection specific characteristics in a multi-graph network can yield much improved prediction of diffusion when the seeding strategies are adjusted accordingly rather than exploiting traditional centrality measures[15].

Another important type of detectable communities can arise not from the densely connected groups but rather dense patterns of connections.Yang et al (2014) argues that patterns of connections may also be an indicator of different communities when observing denser intra-cluster compared to inter-cluster connections[79]. not very relevant.

When dealing with large scale data, a lot of scholars pay little to no attention to the directedness of connections in social networks under study. Many social and relational data structures arise from the directed connections between the nodes, where the directed edge implies a connection from one node to the other that is not necessarily reciprocated. Examples of this behavior could be observed in networks such as twitter followership[13, 44], co-authorship and citation networks[49, 43], and many more[36].Although many treatments of network connections assume undirectedness[9, 28, 60], we argue that a lot of important features underlying the edges could be misjudged, lost, or lead to biased estimates.Direction can be interpreted as the main denominator of followership, where the same cannot be reciprocated. Additionally, when behavioral data is added, time and direction cannot be separated from the network. In other words contacting an individual via a mobile app, emailing colleagues in an organizational setting, liking a post of brand page or a friend's comments, following a celebrity or a friend in a microblog, cannot be easily treated as a mutual relationships;But time and the nature of communication dictates a specific orientation of a link from one person to another. Hence we allow for directed edges that can further signal more information about the clustering of the network. Also put marketing in here

In chapter 2, we employ a model-based approach that allows us to define the network structure according to a set of hypotheses in line with the context under study and the theory underlying the social formation of friendships. We follow the argument of assortative mixing and homophily[52, 55, 59] suggesting individuals in a social network tend to communicate with rather similar people. This phenomenon leads to patterns of structure in networks where we observe denser groups of alike individuals that have fewer connections to the rest of the network. Affinity of individuals and how to measure the likeness among them is conditional on both context and the availability of information on the individual level. Aral et al(2009) employ 20 individual and network characteristics as a proxy for similarity between friends where the degree of closeness is measured by cosine distance. Using a dynamic matched sample estimation, they found further evidence that mobile application behavior could be partly be explained by homophily rather than mere social influence[3]. More recent works on detection of communities tend to account for overlapping structures that allow individual to belong to several communities[1, 28, 78].Yang et al (2012) propose affiliation graph model that allows for detection of dense overlaps in the community network[78]. Airoldi et al (2008) suggest a mixed-membership-stochastic-blockmodel(MMSB) that allows individuals to belong to multiple groups by trying to estimate community membership strength[1]. We adopt the model of MMSB[1, 28] and extend it to allow for more flexible specification and scalable inference.

similar ideas to mixed membership models in marketing(soft clustering):[73, 72]

mixed membership in marketing:[40]

Estimating MMSB , or in fact any model that estimates the latent factors that drive link formation, also helps in solving the identification problem of disentangling influence from homophily as suggested by [65, 64].

In chapter 2 we propose a model of network formation that incorporates the notion of overlapping communities. It is widely known that individuals in a network form declustered communities, where each individual can belong to several groups. Although another important aspect of these separate/shared communities should be encoded via their correlation, as some may induce friendships while others inhibit such connections. In this paper we account for such co-occurrence via introducing correlations among the community memberships of individuals via logistic normal prior on membership strengths. We argue that accounting for such correlations both complies with the assortative mixing theory and performs better compared to the conventional mixed-membership stochastic block model[1]. We furthermore use stochastic variational methods to offer both fast and efficient inference.

In Chapter 3 we intend to gain some insights about networked behavior via community structure that we developed in chapter 2. Observational data on social networks suggest that individuals make decisions that cluster both in network circles and in time. A major hurdle to identify the causal effect of individuals on each other is imposed by the latent homophily, where similar people with analogous preferences may be responsible for their homogeneous behavior. We introduce a new approach in latent variable models for directed networks, where individuals make decisions according to both their potentially overlapping preferences for network formation and activities, where these preferences are unobserved to the researcher. In other words we jointly model both behavior and network formation by first detecting overlapping communities and and using them as potential latent source of forming behavioral response. Controlling for this latent structure would allow us to disentangle the effects of homophily and causal influence.

Chapter 4 builds on the findings of chapter 2 and 3 and extends these models to a dynamic setting of networks, where we aim for finding evidence of contagion driven homophily. Current studies regarding disentanglement of influence-driven and homophily driven behavior assume that these two phenomena act separately on decision making of individuals. We argue that preferences of individuals mapped to a latent space can indeed change according to their adoption behavior. A dynamic networked data, enables us to monitor the preference changes of individuals when they adopt or refuse to adopt a behavior from their peers. While Chapter 2 gives an insight on how to model both link formation and decisions made by individuals, dynamic nature of contagion, requires us to have a model that allows the latent structure to develop through time. Still controlling for latent proxies we can monitor the bursts of link formation and severance and find evidence that adopting similar behaviors among individuals depending on the intensity of their relationship can lead to either more or less homophilous bondings.

Table 1 provides a summary of chapters.

	Chapter 2	Chapter 3	Chapter 4
Topic	• *Modeling network formation via correlated community structure*	• *Insights from networked behavior via community structure*	• *Contagion driven homophily in large dynamic networks*
Goal	• *overlapping community detection and insights on position*	• *Insights from community level attributes to behavior*	• *Monitoring changes in homophilous structure as a function of behavior*
Model	• *Logistic Normal Mixed-Membership-Stochastic-Blockmodel*	• *Joint LDA and LNMMSB*	• *dynamic Joint LDA and LNMMSB*
Inference	• *Variational Inference* • *Stochastic Variational Inference*	• *Stochastic Variational Inference*	• *Stochastic Variational Inference*
Data	• *Synthetic small* • *Synthetic large* • *political blogs* • *bookmarking network*	• *synthetic small* • *synthetic large* • *bookmarking network*	• *synthetic small* • *synthetic large* • *bookmarking network*
Implications	• *scalable inference for large directed networks* • *accounting for potential correlation between community membership of individuals* • *Allowing for dynamic setup* • *defining community level measures of positioning*	• *scalable inference for large networked behavior data* • *jointly modeling behavior and network* • *importance of community level measures in detecting behavioral differences*	• *dynamic evolution of social networks jointly modeled with behavioral data* • *evidence of behavior affecting network positions* • *using community level measures to redefine opinion leadership*

Table 1: summary of chapters

Chapter 1

It is widely known that individuals in a network form declustered communities, where each individual can belong to several groups. Although another important aspect of these separate/shared communities should be encoded via their correlation, as some may induce friendships while others inhibit such connections. In this paper we account for such co-occurrence via introducing correlations among the community memberships of individuals via logistic normal prior. We argue that accounting for such correlations both complies with the assortative mixing theory and performs better compared to the widely known mixed-membership stochastic block model[1]. We furthermore use stochastic variational methods to offer both fast and efficient inference.

Model

In this paper we propose a model based approach to detect overlapping communities as an extended version of mixed-membership-stochastic-blockmodel(MMSB)[1, 28], that allows for both scalable and efficient inference and more flexible community definitions. Mixed membership models provide tools to define a mixture over each grouped data[20]; a problem that mixture models tend to avoid by clustering data into separable groups that are conditionally independent of each other given their cluster assignment[35]. First introduced in the context of topic discovery in text corpora, Blei et al (2003) defined distributions over the vocabulary, where underlying patterns define the topics, and each document is a distribution over these topics[8].

MMSB first proposed by Airoldi et al (2008)[1], defines a generative setting for the formation of the links in a network. This model has been applied frequently to finding overlapping communities in social networks[1, 16], proteininteraction networks[1, 28],citation networks[16], etc. The generative framework assumes that each individual in the network has different degrees of belonging to a set of K pre-specified potentially overlapping communities. Among each directed pair of nodes(a potential link consisting of a sender and a receiver), sender s activates one of its potential roles according to its membership strengths in different communities when communicating with receiver r. Likewise The receiver r activates one of its roles according to its membership strengths in different communities when contacted by the sender s. This means that each individual can belong to several communities or take up different roles depending on whom they are contacting or being contacted by. According to the pair-based community announcements, a links is formed depending on the strength of the connection between those clusters. The data generating process is as follows:

Algorithm 1 MMSB data generating process

```
• $\forall a \in \mathcal{N}$
    – draw a $K$-dimensional mixed membership vector,$\boldsymbol{\theta}_a \sim Dirichlet(\boldsymbol{\alpha})$
• $\forall (a, b) \in \mathcal{E}$
    – draw one-hot membership indicator vector for $a$ when contacting $b$, $\boldsymbol{z}_{a \to b} \sim Categorical(\boldsymbol{\theta}_a)$
    – draw one-hot membership indicator vector for $b$ when contacted by $a$, $\boldsymbol{z}_{a \leftarrow b} \sim Categorical(\boldsymbol{\theta}_b)$
    – sample a link between $a \to b$ with probability $\boldsymbol{z}_{a \to b}\boldsymbol{B}\boldsymbol{z}_{a \leftarrow b}$, $Y(a, b) \sim Bernoulli(\boldsymbol{z}_{a \to b}\boldsymbol{B}\boldsymbol{z}_{a \leftarrow b})$
```

In the algorithm above the $\boldsymbol{\theta}_a$ represents a K-dimensional simplical vector of membership strength of node a from a Dirichlet distribution, where a is a member of the vertex set $\mathcal{N}$. For each directed pair (a, b) that belong to the edge set $\mathcal{E}$, we acquire the indicator vector $\boldsymbol{z}$,for each contact point from a categorical distribution that is parametrized by their membership strengths. Finally a diagonal Block(compatibility) matrix $\boldsymbol{B}$ determines the strength of inter-community connections based on what role is activated for each node. More generally in a repeated setting where links could frequent, we can replace the categorical distributions with the multinomial, and the block matrix $\boldsymbol{B}$ could entail any asymmetric and non-diagonal elements. Due to assortativity of many real world networks we assume here that $\boldsymbol{B}$ is diagonal. Several methods have been applied to estimate the model parameters, which among them variational inference[41, 1, 28] and MCMC[14, 47] are prevalent. Later in this section we discuss Variational method as it is the approach we take for our inference engine. Both variational methods and MCMC for this specific model have excelled to scale to very large networks through introducing stochastic mini-batch sampling[34, 28, 47].Although widely applied all these models face some practical and technical limitations, that we aim to resolve only some of them in this paper.

One of the main limitations in most scalable formulation of community detection under the assumptions of MMSB is that links are treated as undirected edges. We aim to recover communities by not disregarding the direction of the links;Although this might come with a cost, in the sense that we incorporate more information to process that makes the inference more computationally expensive, but we guarantee a scalable and efficient algorithm through unique mini-batch sampling within stochastic variational inference.

More recently, but still limitations

#VI,collapsed Gibbs

Upon detecting communities as explained by the data generating process in 1, individuals who belong to one community and not to a very similar one fail to connect to corresponding individuals due to small chance of connections between clusters. Although this is one of the more prevalent features of many community structures, it would be sensible if we allow for possibility of these connections.

This may happen if some sort of hierarchy or nested structure exists in the nature of these communities that represent roughly their preferences,tastes, roles, groups, etc. Although through variational inference the simpler model exhibits conjugacy and simplifies the estimation, a more natural way would be to allow for correlated mixed memberships by introducing a Logistic-Normal prior instead of Dirichlet for membership probabilities. This enables us to account for the connections among individuals that share rather similar interests or connections among communities that tend to interact more often compared to the specification in MMSB. In the context of mixed membership models , Lafferty and Blei (2006) introduced correlated topic models(CTM), that captures the correlation between topic proportions realized in a text corpora by incorporating a Logistic-Normal(LN) prior.[45]. In the case of our proposed MMSB variant, this would provide an advantage when moving from static to dynamic setting, where the LN-distributed parameters can change according to a simple autoregressive rule, that was not possible under the assumption of Dirichlet distribution[7, 32, 22, 77]. Our proposed model is closely related to the LN-MMSB in [77], however we take a hierarchical Bayesian perspective that allows for fully Bayesian variational inference, alongside preserving the direction of the links. The model is as follows:

Inference and Estimation

In this section we introduce the variational inference method. Other approaches fitting our Bayesian framework that could be used to handle the intractable posteriors,include Monte Carlo method[53], and its variations including the Gibbs sampling, and Metropolis-Hastings. Although these algorithms provide theoretical guarantees on convergence, they may fail to converge in large parameter spaces as they need to process the full sample, and hence more recent methods

Algorithm 2 Hierarchical-Logistic-Normal-MMSB

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

including collapsed Gibbs sampler[48], or Hamiltonian Monte Carlo[10] were developed to address this issue. For more examples of scalable Monte Carlo methods see[23, 61, 75], and for an example of the use of Stochastic Riemannian Langevin Dynamics Monte Carlo applied to the problem of community detection see[47].

Instead we use the variational inference, widely applied in the realm of probabilistic inference and parameter learning, which transforms the problem of inference to an optimization one, by trying to minimize the Kullback-Leibler divergence between the true posterior distribution p and a simplified proposed variational distribution q. Hence instead of making exact inference through approximation, variational inference tend to offer deterministic approximation to the the model posterior distribution. In its simplest case, the proposed model follows a mean field assumption, where it tries to decouple parameters in a way that we can still have tractable results, close enough to the true posterior. In a fully Bayesian framework, where we set all the latent variable to be $\boldsymbol{Z}$ and all the observed variables to be $\boldsymbol{X}$, we specify a joint probability model $P(\boldsymbol{X}, \boldsymbol{Z})$[1], and our goal is to find an approximation to the true posterior $P(\boldsymbol{Z}|\boldsymbol{X})$ and also our model evidence $P(\boldsymbol{X})$.The log likelihood of the model follows:

$$P(\boldsymbol{X}) = \int_{\boldsymbol{Z}} P(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}$$

We further use the log transformation:

$$\begin{aligned} ln\, P(\boldsymbol{X}) &= ln \int_{\boldsymbol{Z}} P(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} \\ &\overbrace{=}^{\times \frac{q(\boldsymbol{Z})}{q(\boldsymbol{Z})}} ln \int_{\boldsymbol{Z}} P(\boldsymbol{X}, \boldsymbol{Z}) \times \frac{q(\boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\ &= ln\, \mathbb{E}_q \left[\frac{P(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right] \end{aligned} \tag{1}$$

Since the logarithm is a concave function, Jensen equality could be applied to get

$$\begin{aligned} ln\, P(\boldsymbol{X}) &= ln\, \mathbb{E}_q \left[\frac{P(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right] \\ &\geq \mathbb{E}_q \Big[ln\, P(\boldsymbol{X}, \boldsymbol{Z}) \Big] - \mathbb{E}_q \Big[ln\, q(\boldsymbol{Z}) \Big] = \mathcal{L}(q) \end{aligned} \tag{2}$$

Equation 2 is known as the evidence lower bound(ELBO). Note that we can

[1] We have dropped the model parameters θ, only to avoid cluttered notation.

write the equation 11 in the following format following the fact that $P(\boldsymbol{X}, \boldsymbol{Z}) = P(\boldsymbol{Z}|\boldsymbol{X})P(\boldsymbol{X})$

$$\begin{aligned} ln\, P(\boldsymbol{X}) &= ln\, \mathbb{E}_q\left[\frac{P(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})}\right] \\ &= \mathbb{E}_q\Big[ln\, P(\boldsymbol{X}, \boldsymbol{Z})\Big] - \mathbb{E}_q\Big[ln\, q(\boldsymbol{Z})\Big] + \mathbb{E}_q\Big[ln\, q(\boldsymbol{Z})\Big] - \mathbb{E}_q\Big[ln\, P(\boldsymbol{Z}|\boldsymbol{X})\Big] \end{aligned} \tag{3}$$

Note that $\mathbb{E}_q\Big[ln\, q(\boldsymbol{Z})\Big] - \mathbb{E}_q\Big[ln\, P(\boldsymbol{Z}|\boldsymbol{X})\Big]$ in equation3 is equivalent to the Kullback-Leibler divergence of the proposed variational distribution $q(\boldsymbol{Z})$ and the true posterior $P(\boldsymbol{Z}|\boldsymbol{X})$. Hence we can rewrite the log marginal as the following:

$$ln\, P(\boldsymbol{X}) = \mathcal{L}(q) + KL\Big(q(\boldsymbol{Z})||P(\boldsymbol{Z}|\boldsymbol{X})\Big) \tag{4}$$

To simplify the Kullback-Leibler divergence we can rewrite

$$\begin{aligned} KL\Big(q(\boldsymbol{Z})||P(\boldsymbol{Z}|\boldsymbol{X})\Big) &= -\Bigg(\mathbb{E}_q\Big[ln\, P(\boldsymbol{X}, \boldsymbol{Z})\Big] - \mathbb{E}_q\Big[ln\, q(\boldsymbol{Z})\Big]\Bigg) + ln\, P(\boldsymbol{X}) \\ &= -\mathcal{L}(q) + ln\, P(\boldsymbol{X}) \end{aligned} \tag{5}$$

According to equation 5, minimization the KL-divergence between the variational distribution and the true posterior translates to maximization the ELBO as the marginal distribution $P(\boldsymbol{X})$,does not depend on the variational distribution q. We generally in practice tackle the simpler problem of maximizing the lower bound instead of minimizing the KL divergence;however as can be seen in equation 5 this exactly corresponds to our goal of minimizing the distance between the true intractable posterior and the simpler variational distribution[2]. Hence the problem of finding the posterior has been reduced to one that tightens the lower bound. The closeness depends on our approximating distribution and how well it can represent the true posterior and yet is tractable. There is much discussion about the performance of the approximation through variational inference and exact methods(?,?,?). However in practice variational methods have shown to perform at least as good as other methods, and yet they can easily be adjusted to scale to very large datasets[63][6](??).

On asymptotic properties of the variational inference???

In the following section we derive the variational updates for our current model of Logistic-Normal MMSB described above, and later we adjust the algorithm to scale to larger data sets by using stochastic search.

Variational Algorithm for Directed LNMMSB

In this section we develop the variational algorithm for the case of directed network Logistic-Normal MMSB. For this we start by writing down the ELBO which entails the variational expectation of the joint and approximating distributions. The log joint distribution of data, latent variables and model parameters as suggested by Algorithm2 follows:

$$\begin{aligned} ln\, p(joint) &= ln\, p(\mu|m_0, M_0) + ln\, p(\Lambda|\ell_0, L_0) + \sum_a ln\, p(\theta_a|\mu, \Lambda) + \sum_a \sum_b ln\, p(z_{a\rightarrow b}|\theta_a) \\ &+ \sum_a \sum_b ln\, p(z_{a\leftarrow b}|\theta_b) + \sum_k ln\, p(\beta_{kk}|\eta_0, \eta_1) + \sum_a \sum_b ln\, p(y_{ab}|z_{a\rightarrow b}, z_{a\leftarrow b}, \beta) \end{aligned} \tag{6}$$

[2] For other bounds see ?

Furthermore we assume that our approximating distributions follows a mean field assumption[74, 41], where the probability distribution is factorized. The variational parameters are distributed as follows:

$$\begin{aligned}
\mu &\sim q(\mu|m, M) \sim \mathcal{N}ormal(\mu|m, M) \\
\Lambda &\sim q(\Lambda|\ell, L) \sim \mathcal{W}ishart(\Lambda|\ell, L) \\
\theta_a &\sim q(\theta_a|\mu_a, \Lambda_a) \sim \mathcal{N}ormal(\theta_a|\mu_a, \Lambda_a) \\
\beta_{kk} &\sim q(\beta_{kk}|b_k) \sim \mathcal{B}eta(b_{k0}, b_{k1}) \\
z_{a\rightarrow b} &\sim q(z_{a\rightarrow b}|\phi_{a\rightarrow b}) \sim Categorical(z_{a\rightarrow b}|\phi_{a\rightarrow b}) \\
z_{a\leftarrow b} &\sim q(z_{a\leftarrow b}|\phi_{a\leftarrow b}) \sim Categorical(z_{a\leftarrow b}|\phi_{a\leftarrow b}) \qquad (7)
\end{aligned}$$

Following shows the logarithmic transformation of the factorized variational distribution:

$$\begin{aligned}
ln\, q(.) = ln\, q(\mu|m, M) + ln\, q(\Lambda|\ell, L) + \sum_a ln\, q(\theta_a|\mu_a, \Lambda_a) + \sum_a \sum_b ln\, q(z_{a\rightarrow b}|\phi_{a\rightarrow b}) \\
+ \sum_a \sum_b ln\, q(z_{a\leftarrow b}|\phi_{a\leftarrow b}) + \sum_k ln\, q(\beta_{kk}|b_{k0}, b_{k1}) \qquad (8)
\end{aligned}$$

Taking the variational expectation for the log joint probability distribution of the model leads to the negative cross entropy for the corresponding distribution and the variational expectation of the factorized distribution leads to the corresponding negative entropies. To see how cross entropies and entropies are derived refer to Appendix(?). After taking the relevant expectation for each probability distribution the ELBO becomes:

$$
\begin{aligned}
\mathcal{L} \;\; = \;\; & \mathbb{E}_q[ln\, p(\mu|m_0, M_0)] + \mathbb{E}_q[ln\, p(\Lambda|\ell_0, L_0)] + \sum_a \mathbb{E}_q[ln\, p(\theta_a|\mu, \Lambda)] + \\
& \sum_a \sum_b \mathbb{E}_q[ln\, p(z_{a\rightarrow b}|\theta_a)] + \sum_a \sum_b \mathbb{E}_q[ln\, p(z_{a\leftarrow b}|\theta_b)] + \\
& \sum_k \mathbb{E}_q[ln\, p(\beta_{kk}|\eta)] + \sum_a \sum_b \mathbb{E}_q[ln\, p(y_{ab}|z_{a\rightarrow b}, z_{a\leftarrow b}, \beta)] \\
& + H_q[\mu] + H_q[\Lambda] + H_q[\theta] + H_q[\beta] + H_q[z_{\rightarrow}] + H_q[z_{\leftarrow}] \qquad (9)
\end{aligned}
$$

In the extension of above full ELBO in Appendix(?), we differentiate between the ϕ's for links versus for non-links. This can come handy when we use mini-batch sampling and extend the algorithm to be stochastic.

Explanation of the parameter names...N corresponds to the number of nodes and K to the number of communities.

Note that in line 2 of equation 9, we can use the Jensen equality to bound the expectation of log sum of exponentials $\mathbb{E}_q\Bigg[ln\ \sum_l exp(\theta_{a,l})\Bigg]$, since the concavity of the logarithm allows us to write

$$
\begin{aligned}
\mathbb{E}_q\Bigg[ln \sum_l exp(\theta_{a,l})\Bigg] & \leq ln\left(\sum_l \mathbb{E}_q\big[exp(\theta_{a,l})\big]\right) \\
& = ln \sum_l exp(\mu_{a,l} + \tfrac{1}{2}\Lambda^{-1}_{a,l}) \qquad (10)
\end{aligned}
$$

For further details about the variational expectation of this terms refer to Appendix(??).This is one of the caveats of using the Logistic-Normal prior instead of the Dirichlet distribution, introducing non-conjugacy that hinders us from deriving fully closed-form solution for the variational parameters involved.

To derive each variational parameter, we maximize the ELBO with respect to that parameter. For the full derivation of each parameter refer to the Appendix(?). Variational equations boil down to updates for global and local parameters. Global variational parameters are model specific(m, M, ℓ, L, b), whereas local parameters are individual and observation specific$(\phi_{a\to b}, \phi_{a\leftarrow b}, \mu_a, \Lambda_a)$.

Global Parameters

To derive the variational mean and precision of the mean membership strength we optimize the ELBO with respect to m and M to get

$$\begin{aligned} m &= (M_0 + N\ell L)^{-1}(M_0 m_0 + \ell L \sum_a \mu_a) \\ M &= M_0 + N\ell L \end{aligned} \tag{11}$$

Additionally optimizing the degrees of freedom and scale matrix of the Wishart distributed precision simultaneously, we have:

$$\begin{aligned} L =& \left((L_0^{-1} + \sum_a \Lambda_a^{-1} + \sum_a (\mu_a - m)(\mu_a - m)^T + NM^{-1}) \right)^{-1} \\ \ell =& \ell_0 + N \end{aligned} \tag{12}$$

Simultaneously optimizing the ELBO with respect to b_{k0} and b_{k1}:

$$\begin{aligned} b_{k0} &= \eta_0 + \sum_{a,b \in links} \phi_{a\to b,k} \phi_{a\leftarrow b,k} \\ b_{k1} &= \eta_1 + \sum_{a,b \notin links} \phi_{a\to b,k} \phi_{a\leftarrow b,k} \end{aligned} \tag{13}$$

Local Parameters

ϕ parameters correspond to the membership strength in the variational specification, and as in the generative process within each community they have to add to unity(i,e$\sum_l \phi_{a\rightarrow b,l}$). We can only maximize the ELBO with respect to these parameters up to a scale. Hence after finding the non-normalized ϕ's we make a normalizing transformation to make sure that they sum to one.

For the link specific updates:

$$\phi_{a\rightarrow b,k} \propto exp\left\{ \mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big)\right\} \tag{14}$$

$$\phi_{a\leftarrow b,k} \propto exp\left\{ \mu_{b,k} + \phi_{a\rightarrow b,k}\big(\psi(b_{k0}) - \psi(b_{k0}+b_{k1}) - ln\,\epsilon\big)\right\} \tag{15}$$

Similarly for the non-links:

$$\phi_{a\rightarrow b,k} \propto exp\left\{ \mu_{a,k} + \phi_{a\leftarrow b,k}\big(\psi(b_{k1}) - \psi(b_{k0}+b_{k1}) - ln\,(1-\epsilon)\big)\right\} \tag{16}$$

$$\phi_{a\leftarrow b,k} \propto exp\left\{ \mu_{b,k} + \phi_{a\rightarrow b,k}\big(\psi(b_{k1}) - \psi(b_{k0}+b_{k1}) - ln\,(1-\epsilon)\big)\right\} \tag{17}$$

There is no closed form solution for either μ_a or Λ_a due to the choice of the non-conjugate prior. Hence we derive the gradients of the ELBO with respect to both parameters and use Adagrad[19].

$\nabla_{\mu_a}\mathcal{L}_{\mu_a} =$

$$-\ell L(\mu_a - m) + \sum_{b \in sink(a)} \phi_{a \to b} + \sum_{b \notin sink(a)} \phi_{a \to b}$$

$$+ \sum_{b \in source(a)} \phi_{b \leftarrow a} + \sum_{b \notin source(a)} \phi_{b \leftarrow a} - \sum_{b} \underline{\text{sfx}}(a) \tag{18}$$

where $\underline{\text{sfx}}(a)$ refers to a vector function of dimension K, and each element represents the soft-max transformation of $(\mu_a + \frac{1}{2}\Lambda_a^{-1})$

Instead of optimizing with respect to the precision we maximize the ELBO with respect to the diagonal covariance as follows:

$\nabla_{\Lambda_a^{-1}} \mathcal{L}_{\Lambda_a^{-1}} =$

$$-\frac{\ell}{2} diag(L) + \frac{1}{2}(\Lambda_a) - \frac{1}{2}\sum_{b}(\underline{\text{sfx}}(a)) \tag{19}$$

For further details on the gradients of the ELBO with respect to μ_a and Λ_a see Appendix(?).

The variational procedure is summarized in Algorithm 3:

Stochastic Variational Algorithm

Variational inference offers a fast approximation of the posterior distribution by optimizing the ELBO, however this might need the screening of the whole individual(link) level observations both for updating the variational parameters and also evaluating the ELBO. A more recent method in variational inference offers a stochastic search in the parameter space suggested by [34]. Stochastic Variational Inference(SVI) instead of the full sample, samples only a small mini-batch from the data and reweighs the parameter updates according to a decay rate that satisfies the conditions of stochastic search by Robbins and

Algorithm 3 Variational Algorithm

```
• Until ELBO improves
    – Initialize the variational parameters
    – update local parameters
        ∗ $\forall (a, b) \in \mathcal{E}$
            · update $\phi_{a \rightarrow b,k}$ and $\phi_{a \leftarrow b,k}$ according to 14,and 15
        ∗ $\forall (a, b) \notin \mathcal{E}$
            · update $\phi_{a \rightarrow b,k}$ and $\phi_{a \leftarrow b,k}$ according to 16,and 17
        ∗ $\forall a \in \mathcal{N}$
            · update $\mu_a$ and $\Lambda_a$ using the gradients 18, and19 via Adagrad
    – update global parameters
        ∗ update the global parameters $m, M, \ell, L, b_0, b_1$ according to
          11,12,13
• end
```

Monro[62]. According to SVI, our algorithm only needs to iterate between sub-sampling the data to acquire a noisy gradient of the objective function and only update according to the sub-sampled mini-batch. Under the updating steps conditions the algorithm can provably converge to the optimum[34]. For further examples of this approach in topic modeling refer to[16, 28, 34]. Related to our work, Gopalan et al (2013) offers several sub-sampling schemes, including the link-only sampling which provides both efficient and reasonably appropriate simplification for undirected networks[28]. However for the case of directed network the assumption of link-only sampling although may offer more convenience, can be too simplistic and result in biased estimates. In the following section we introduce our sampling scheme.

Mini-batch Sampling

In our model, we sample our network using only nodes, and since large networks exhibit a very sparse patterns of connections, each time we sample few nodes with all their links and equal proportion(or a small multiple) of their randomly selected non-links. After rounds of iteration this assumption both takes into account the information of all links, and actual non-links. Moreover the random selection of the non-links allows different non-links to be visited. We argue that averaging method of [28, 47] for the non-link ϕ's can introduce biases regarding our estimates, as with directed networks, the amount of information in the direction of links and non-links cannot easily be ignored.

For this purpose we use informative stratified random node sampling, where we define two types of sets for each node:informative, and non-informative sets. Informative set consists of all the links that involve a specific node, and non-links of that node that at most h hops away.Non-informative set on the other hand consists of S sets that that partitions the rest of the non-links that involve that node. For the sake of mini-batch sampling, we first sample a node at random, and with a high chance select the informative set of that node, and with a very low probability we choose one of its S non-informative non-link sets. Regarding each update we use this information to reweigh the computation of each expectation in the updates. For example to compute the variational expectation of the link probabilities, we have to reweigh that with $p_{high} \times \frac{1}{2N}$ if it comes from a link set or $(1 - p_{high}) \times \frac{1}{2NS}$ if it comes from the non-link set.

At each iteration we sample a small number of nodes at random, include all their links and a equal number of randomly selected non-links of those nodes, and accordingly we reweigh our previous parameter updates based on the mini-batch sizes.

Parameter Initialization

Data

We start our algorithm by using the variational inference on a small synthetic network, and then we apply the stochastic version of the algorithm to this data. Due to the inability of traditional variational inference to comply with large networks, we apply our stochastic variational inference to a very large synthetic network. To extend our findings in real network datasets, we use a relatively large network of political blog citations, and a very large bookmarking blog.

Training set and holdout set

At each configuration of datasets, we divide our data to training and validation sets, where in the training set, we employ the learning and see the performance in the validation set, and also use the validation to learn the model parameters. Except in the traditional variational inference with the synthetic networks, where we use normalized mutual information(NMI) to assess the performance between our estimates and the ground truth values. Our holdout set consists of 2.5% of links and the equal number of non-links, and the rest would be considered our training data.

Synthetic Data

small network

We construct a small synthetic network, consisting of 250 individuals in 8 relatively dense communities portrayed in figure 1.

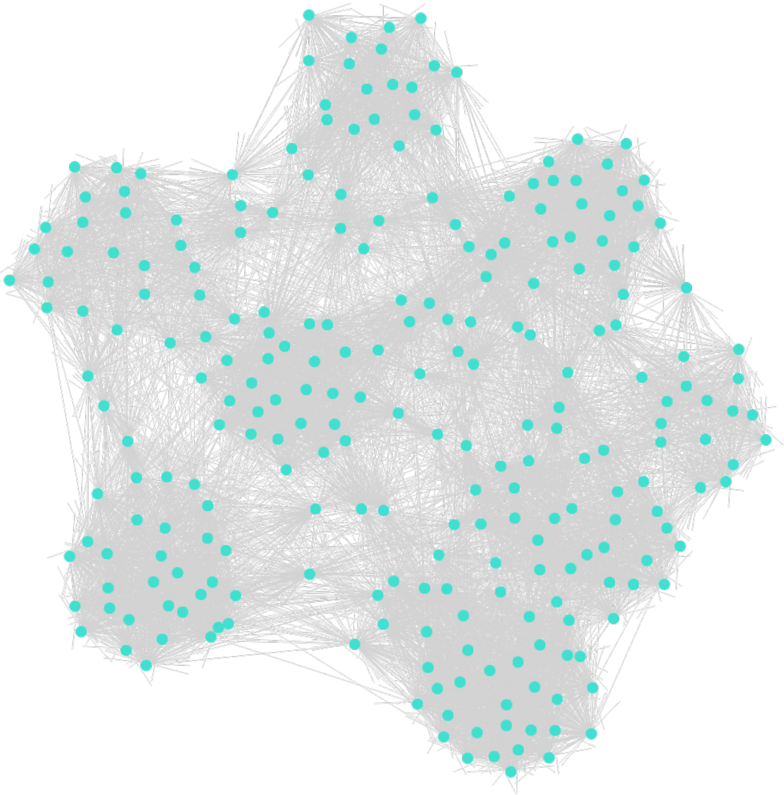


Figure 1: small synthetic network visualization

large network

Real World Data

political blog citation

bookmarking network

Results

synthetic

Small network

Performance Checking

Figure 2 show the improvement of the evidence lower bound as we maximize it, which translates to minimizing the KL-divergence of the two distribution. As we can see the model stays steady and flat after 100 rounds of iterations, which means that the engine has already converged. In the case where we can evaluate the ELBO in a small network, checking the behavior of the lower bound is the first step towards the performance check of the model. This might still be a steady state where convergence has arrived at a non-escapable settings in the parameter space. The monotonic increase in the ELBO signifies the correct direction of the steps taken in the inference.

We estimate the membership probabilities using the variational distribution and compare it with the ground truth values. Although the initialization algorithm starts by detecting 11 communities, we further investigate those clusters after estimation, which yields 3 empty clusters. For better visual comparison, we excluded those empty communities and plotted the membership probabilities of each individual in figure 3. The plot visually provides a reasonable closeness of our estimates and the ground truth for a small network. However for further

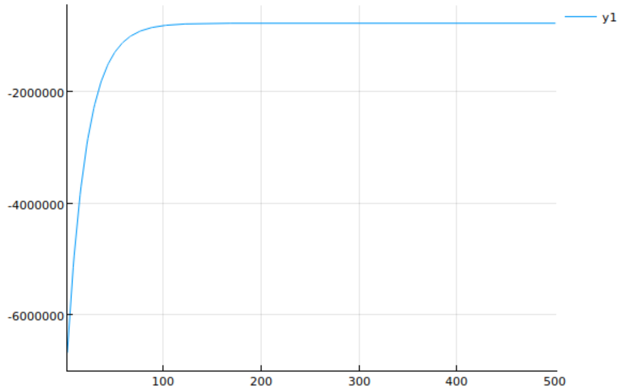
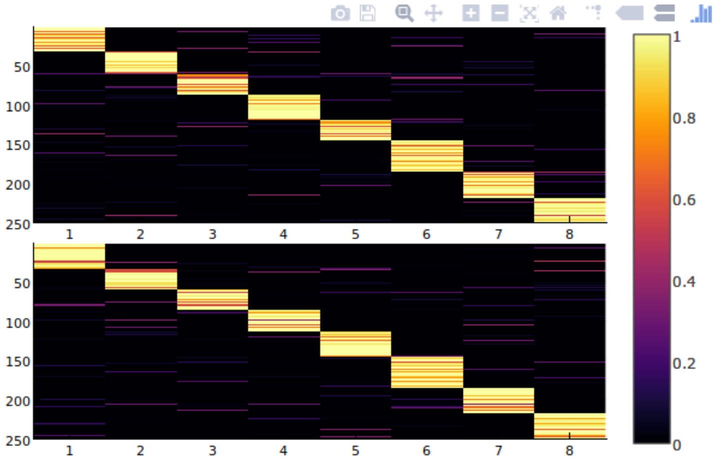


Figure 2: ELBO of small synthetic network

assessments we need to quantify this performance. Next section introduces the Normalized Mutual Information, that is commonly used in evaluating the performance of overlapping community detection algorithms provided the ground truth.



or

Figure 3: θ estimates from VI(top), and the true θ's(bottom)

Normalized Mutual Information(NMI)

For the data with ground truth we can also use normalized mutual information(NMI). NMI has been used vastly in assessing the performance of algorithms involving overlapping clustering[51, 46].The table 2 shows the NMI values for different set of comparisons. As we can see the in the last row, our estimated clusters shows prominent improvement from the well initialized algorithm.

case	NMI
truth vs. *initialized*	0.438348
estimated vs. *initialized*	0.361077
estimated vs. *truth*	**0.804357**

Table 2: Normalized Mutual Information for small synthetic network

Large network.

Performance check

Perplexity of a model q, given a stochastic process p (here a joint generative model that can produce infinite streams of data) is defined as:

$perplexity(p, q) \triangleq 2^{H(p,q)}$

where the cross entropy

$$H(p, q) \triangleq lim_{N\to\infty} - \frac{1}{N} \sum_{y_{1:N}} p(y_{1:N}) log\, q(y_{1:N})$$

and perplexity the lower the better. But we use the empirical distribution for p. $p_{emp}(y_{1:N}) = \delta_{y^*_{1:N}}(y_{1:N})$,where y^* is a single long test sequence. Then $H(p_{emp}, q) = -\frac{1}{N} log\, q(y^*_{1:N})$, so the perplexity becomes:

$$perplexity(p_{emp}, q) = \qquad q(y^*_{1:N})^{1/N} =^N \sqrt{\prod_{i=1}^{N} \frac{1}{q(y^*_i | y^*_{1:i-1})}}$$

which is the geometric mean of the inverse of the predictive probability.

We could also think of it as exponential of the negative average log likelihood of the data. average perplexity on a test set is:

$$perp_{avg}(Test|params) = \quad exp\left(- \frac{\sum_{(a,b)\in Test} log\,(\frac{1}{T}) \sum_{t=1}^{T} p(y_{ab}|params)}{|Test|} \right)$$

We can set for example a 1% of the links and the non-links as the test set.

Real data

political blog citation

bookmarkig network

References

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[2] Asim Ansari, Oded Koenigsberg, and Florian Stahl. Modeling multiple relationships in social networks. *Journal of Marketing Research*, 48(4):713–728, 2011.

[3] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[4] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[5] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

[7] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[9] Michael Braun and André Bonfrer. Scalable inference of customer similarities from interactions data using dirichlet processes. *Marketing Science*, 30(3):513–531, 2011.

[10] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[11] Hilla Brot, Lev Muchnik, Jacob Goldenberg, and Yoram Louzoun. Evolution through bursts: Network structure develops through localized bursts in time and space. *Network Science*, 4(3):293–313, 2016.

[12] Jan K Brueckner and Oleg Smirnov. Workings of the melting pot: Social networks and the evolution of population attributes. *Journal of Regional Science*, 47(2):209–228, 2007.

[13] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining*, page 4. ACM, 2010.

[14] Jonathan Chang. lda: Collapsed gibbs sampling methods for topic models, 2011.

[15] Xi Chen, Ralf van der Lans, and Tuan Q Phan. Uncovering the importance of relationship characteristics in social networks: Implications for seeding strategies. American Marketing Association, 2017.

[16] Yoon-Sik Cho, Greg Ver Steeg, Emilio Ferrara, and Aram Galstyan. Latent space model for multi-modal social data. In *Proceedings of the 25th International Conference on World Wide Web*, pages 447–458. International World Wide Web Conferences Steering Committee, 2016.

[17] James Samuel Coleman, Elihu Katz, and Herbert Menzel. *Medical innovation: A diffusion study.* Bobbs-Merrill Co, 1966.

[18] Joseph Davin. *Essays on the Social Consumer: Peer influence in the adoption and engagement of digital goods.* PhD thesis, 2015.

[19] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[20] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5220–5227, 2004.

[21] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[22] Wenjie Fu, Le Song, and Eric P Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th annual international conference on machine learning*, pages 329–336. ACM, 2009.

[23] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

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

[25] Malcolm Gladwell. The tipping point: How things can make a big difference. *New York*, 2000.

[26] Sharad Goel and Daniel G Goldstein. Predicting individual behavior with social networks. *Marketing Science*, 33(1):82–93, 2013.

[27] Jacob Goldenberg, Sangman Han, Donald R Lehmann, and Jae Weon Hong. The role of hubs in the adoption process. *Journal of marketing*, 73(2):1–13, 2009.

[28] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

[29] Prem K Gopalan, Sean Gerrish, Michael Freedman, David M Blei, and David M Mimno. Scalable inference of overlapping communities. In *Advances in Neural Information Processing Systems*, pages 2249–2257, 2012.

[30] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[31] Oliver Hinz, Bernd Skiera, Christian Barrot, and Jan U Becker. Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75(6):55–71, 2011.

[32] Qirong Ho, Le Song, and Eric P Xing. Evolving cluster mixed-membership blockmodel for time-varying networks. 2011.

[33] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[34] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

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

[36] Norman P Hummon and Patrick Dereian. Connectivity in a citation network: The development of dna theory. *Social networks*, 11(1):39–63, 1989.

[37] Raghuram Iyengar, Christophe Van den Bulte, and Thomas W Valente. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, 2011.

[38] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[39] Matthew O Jackson and Alison Watts. The evolution of social and economic networks. *Journal of Economic Theory*, 106(2):265–295, 2002.

[40] Bruno JD Jacobs, Bas Donkers, and Dennis Fok. Model-based purchase predictions for large assortments. *Marketing Science*, 35(3):389–404, 2016.

[41] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[42] Zsolt Katona, Peter Pal Zubcsek, and Miklos Sarvary. Network effects and personal influences: The diffusion of an online social network. *Journal of marketing research*, 48(3):425–443, 2011.

[43] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[44] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[45] John D Lafferty and David M Blei. Correlated topic models. In *Advances in neural information processing systems*, pages 147–154, 2006.

[46] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[47] Wenzhe Li, Sungjin Ahn, and Max Welling. Scalable mcmc for mixed membership stochastic blockmodels. In *Artificial Intelligence and Statistics*, pages 723–731, 2016.

[48] Jun S Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[49] Xiaoming Liu, Johan Bollen, Michael L Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Information processing & management*, 41(6):1462–1480, 2005.

[50] Liye Ma, Ramayya Krishnan, and Alan L Montgomery. Latent homophily or social influence? an empirical analysis of purchase within a social network. *Management Science*, 61(2):454–473, 2014.

[51] Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.

[52] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[53] Radford M Neal. Probabilistic inference using markov chain monte carlo methods. 1993.

[54] Tamás Nepusz, Andrea Petróczi, László Négyessy, and Fülöp Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1):016107, 2008.

[55] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[56] Mark EJ Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[57] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[58] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[59] Mark EJ Newman and Michelle Girvan. Mixing patterns and community structure in networks. *Statistical mechanics of complex networks*, pages 66–87, 2003.

[60] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *arXiv preprint physics/0506133*, 2005.

[61] Sam Patterson and Yee Whye Teh. Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

[62] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[63] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd*

International Conference on Machine Learning (ICML-15), pages 1218–1226, 2015.

[64] Cosma Rohilla Shalizi and Edward McFowland III. Controlling for latent homophily in social networks through inferring latent locations. *arXiv preprint arXiv:1607.06565*, 2016.

[65] Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40(2):211–239, 2011.

[66] Tom Snijders, Christian Steglich, and Michael Schweinberger. *Modeling the coevolution of networks and behavior.* na, 2007.

[67] Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[68] Andrew T Stephen and Olivier Toubia. Deriving value from social commerce networks. *Journal of marketing research*, 47(2):215–228, 2010.

[69] Michael Trusov, Anand V Bodapati, and Randolph E Bucklin. Determining influential users in internet social networks. *Journal of Marketing Research*, 47(4):643–658, 2010.

[70] Christophe Van den Bulte and Yogesh V Joshi. New product diffusion with influentials and imitators. *Marketing Science*, 26(3):400–421, 2007.

[71] Christophe Van den Bulte and Gary L Lilien. Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, 106(5):1409–1435, 2001.

[72] Sajeev Varki and Pradeep K Chintagunta. The augmented latent class model: Incorporating additional heterogeneity in the latent class model for panel data. *Journal of Marketing Research*, 41(2):226–233, 2004.

[73] Sajeev Varki, Bruce Cooil, and Roland T Rust. Modeling fuzzy data in qualitative marketing research. *Journal of Marketing Research*, 37(4):480–489, 2000.

[74] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[75] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

[76] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):43, 2013.

[77] Eric P Xing, Wenjie Fu, Le Song, et al. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566, 2010.

[78] Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1170–1175. IEEE, 2012.

[79] Jaewon Yang, Julian McAuley, and Jure Leskovec. Detecting cohesive and 2-mode communities indirected and undirected networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 323–332. ACM, 2014.

[80] Peter Pal Zubcsek, Imran Chowdhury, and Zsolt Katona. Information communities: the network structure of communication. *Social Networks*, 38:50–62, 2014.

Appendix

A Negative cross entropies

A.1 Two Normals

Note:All the normals are parametrized using the precision matrix.

$q \sim \mathcal{N}(x|m, L)$

$p \sim \mathcal{N}(x|\mu, \Lambda)$

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

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 $Cov_q = L^{-1}$

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

Hence we have:

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

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:

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

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

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

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

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

Note that $\mathbb{E}_q[ln\,\beta] = \psi(b_0) - \psi(b_0+b_1)$

so :

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

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

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

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

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,

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

C Variational ELBO

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

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

$$H_q[params] \quad = \quad H_q[\mu] + H_q[\Lambda] + H_q[\theta] + H_q[\beta] + H_q[z_{\rightarrow}] + H_q[z_{\leftarrow}]$$

Furthermore,

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

We parametrize the variational distribution as follows:

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

Using the results from above regarding the negative cross entropies:

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

$\mathbb{E}_q[\Lambda] = \ell L$

$\mathbb{E}_q[ln\,|\Lambda|] = \psi_K(\frac{\ell}{2}) + Kln\, 2 + ln\,|L|$

$$\begin{aligned}
-\sum_a \tfrac{K}{2}ln\, 2\pi + \sum_a \tfrac{1}{2}\mathbb{E}_q\Big\{ln\,|\Lambda|\Big\} \\
-\sum_a \tfrac{1}{2}\Big(Tr\,\Big[\mathbb{E}_q\Big\{\Lambda\Big\}\Lambda_a^{-1}\Big] + \mathbb{E}_q\Big\{(\mu_a-\mu)^T\Lambda(\mu_a-\mu)\Big\}\Big) = \\
-\sum_a \tfrac{K}{2}ln\, 2\pi + \sum_a \psi_K(\tfrac{\ell}{2}) + \sum_a Kln\, 2 + \sum_a ln\,|L| \\
-\tfrac{\ell}{2}\Big(Tr\,\Big[L\Big(\sum_a (\Lambda_a^{-1} + (m-\mu_a)(m-\mu_a)^T)\Big)\Big\}\Big)
\end{aligned}$$

For the expression $\mathbb{E}_q[ln\,(\sum_l exp(\theta_{a,l}))]$, we use the Jensen's inequality to acquire:

$$
\begin{aligned}
\mathbb{E}_q[ln\,(\sum_l exp(\theta_{a,l}))] &\leq ln\,(\sum_l \mathbb{E}_q[exp(\theta_{a,l})]) \\
&= ln\,(\sum_l exp(\mu_{a,l} + \tfrac{1}{2} diag(\Lambda_a^{-1})_{ll}))
\end{aligned}
$$

We can introduce another bound that introduces a new variational parameter per individual:

$$
\mathbb{E}_q[ln\,(\sum_l exp(\theta_{a,l}))] \leq \zeta_a^{-1} \sum_l exp(\mu_{a,l} + \tfrac{1}{2} diag(\Lambda_a^{-1})_{ll}) + ln\,\zeta_a - 1
$$

Moreover, using the entropies from above:

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

Note that here I assume the following for the hyperparameters:

$$
\begin{aligned}
m_0 &= \mathbf{0} \\
M_0 &= 10 \times \boldsymbol{I} \\
\ell_0 &= K + 2 \\
L_0 &= \frac{.1}{\ell_0} \boldsymbol{I} \\
\eta_0 &> 1 = 9 \\
\eta_1 &= 1
\end{aligned}
$$

Note that here I assume the following for the variational parameters:

$$
\begin{aligned}
m &= \mathbf{0} \\
M &= 10 \times \boldsymbol{I} \\
\ell &= K + 2 \\
L &= \frac{.1}{\ell_0} \boldsymbol{I} \\
b_0 &> 1 = 9 \\
b_1 &= 1
\end{aligned}
$$

Finally, we have the following:

$$
\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\rightarrow 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\rightarrow 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\rightarrow 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\rightarrow 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\rightarrow b,k} ln\, \phi_{a\rightarrow b,k}\Big) \\
& -\sum_a \sum_{b\notin sink(a)} \sum_k \Big(\phi_{a\rightarrow b,k} ln\, \phi_{a\rightarrow 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 \sum \sum \Big(\phi_{a\leftarrow b,k} ln\, \phi_{a\leftarrow b,k}\Big)
\end{aligned}
$$

D ELBO Gradients

D.1 Gradient with respect to m

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

In minibatch node sampling this would be

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

D.2 Gradient with respect to M

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

D.3 Gradient with respect to L

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

optimizing simultaeneously with ℓ in the minibatch setting:

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

D.4 Gradient with respect to ℓ

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

D.5 Gradient with respect to b_k

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

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

D.6 Gradient with respect to $\phi_{a\to b,k}$ for links

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

D.7 Gradient with respect to $\phi_{a\leftarrow b,k}$ for links

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

D.8 Gradient with respect to $\phi_{a\rightarrow b,k}$ for nonlinks

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

D.9 Gradient with respect to $\phi_{a\leftarrow b,k}$ for nonlinks

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

D.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

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

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

so all in all the gradient is :

$\nabla_{\mu_a}\mathcal{L}_{\mu_a} =$

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

Similarly the Hessian will be as follows:

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

The newton step would look like:

$\mu_{a,k} = \mu_{a,k} - H^{-1}_{\mu_{a,k}} G_{\mu_{a,k}}$

D.11 Gradient with respect to Λ_a

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

$$\mathcal{L}_{\Lambda_a^{-1}} = \quad -\tfrac{\ell}{2} diag\,(L)'\Lambda_a^{-1} + \tfrac{1}{2} ln\,|diagm(\Lambda_a^{-1})| - \sum_b log\left(\mathbf{1}^T \underline{\mathrm{f}}(\mu_a, \Lambda_a)\right)$$

$$=$$

$$\nabla_{\Lambda_a^{-1}} \mathcal{L}_{\Lambda_a^{-1}} = G_{\Lambda_a^{-1}} = \quad \boxed{-\tfrac{\ell}{2} diag(L) + \tfrac{1}{2}(\Lambda_a) - \tfrac{1}{2}\sum_b(\underline{\mathrm{sfx}}(a))}$$

□

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

The newton step would look like:

$\Lambda_a^{-1} = \Lambda_a^{-1} - H^{-1}_{\Lambda_a^{-1}} G_{\Lambda_a^{-1}}$