**Modeling Network Structure via Correlated Community Membership**

**Abstract:In large social networks, individuals form connection to other individuals for variety of reasons. The reason why two individuals connect will likely influence their interaction within the network. One such reason is for individuals to share the same interests that are not shared by others. Such individuals can be seen as members of a community for such interest. In general, the network consists of multitude of these communities. As network formation and more importantly interactions will depend on the links within and between these communities, it will be vital to understand such this structure. A large body of literature has already developed theories about the role and features of influentials in social networks; although when the network is a blend of multiple smaller networks, we will need to revisit the detection of such influentials at the smaller sub-networks rather than in the global network. To achieve this, we develop an extended mixed membership stochastic blockmodel (mmsb), and use a stochastic variational inference on directed networks to detect smaller communities. To assess the performance we use two synthetic networks.**

*Keywords: community detection, influentials, stochastic variational inference*

*Track: Methods, Modelling, &Marketing Analytics*

# 1 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, 9, 73]. Hence sociometric data derived from network structure 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 a follower [27]. Stephan et al assess the importance of different centrality measures, such as in-degree and out degrees of online sellers in the value they create in terms of sales. [73]. 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.[Goldenberg,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 finer grained centrality measures. Moreover, marketing science is always interested in finding influential people in a network to be able to target and segment them properly [4, 37, 72, 71, 74, 73, 42]. Yet not all centrality measures on the network level correspond to similar notions and could provide even contradicting results [71, 72]. Trusov et al argue that the mere friend counts in a social network does poorly when it comes to detecting influential users.[74].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 or been labeled differently. For instance 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 [51].

Figure X shows the graph of degree centrality of three different communities in a publicly available email network of a European Research Institute extracted from SNAP[?]. The connections are formed from the emails between the members of this institute and the ground truth communities are the departments within the institute. As can be seen nodes that are central at the community level are not necessarily central at the network level.

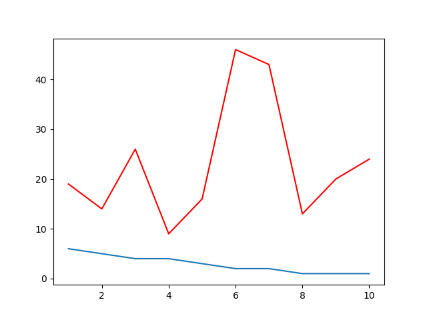
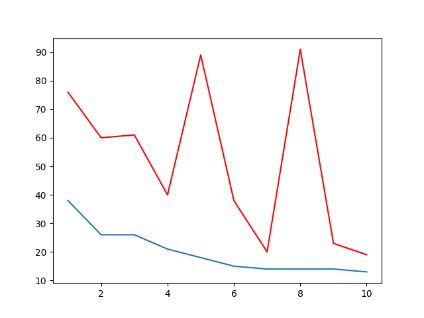
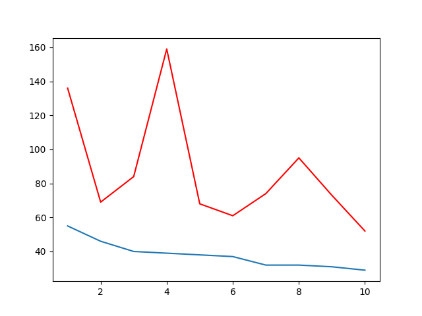


Figure 1 Three different communities, where each community is sorted based on the degree centrality (blue line) and the corresponding network centrality for that node is shown on top (red line)

Many real life networks exhibit smaller grouped structures [5, 60]. Sometimes these smaller patterns are more accurately described by multiple interconnected networks instead of one homogeneous network [81, 79, 46, 1]. While theoretically we can expect existing network measures to provide predictive power for observed behaviors and patterns in a social network [26], empirically this power will be undermined because the network is in fact a mixture of multiple sub-networks. This leaves the empirical assessment of network measures on a shaky foundation.

Understanding how individuals are connected in a network can be crucial in designing marketing strategies. Braun and Bonfrer (2011) argue that finding hidden patterns in network structure can lead to better insights for devising strategies to reach more relevant individuals in a diffusion process. They use a latent space approach to map individual latent traits to a Euclidean space. In a latent space model, hidden tie related preferences are mapped to a lower dimensional Euclidean space. Distance between individuals in this space is interpreted as the similarity between them, and more similar nodes are more probable to form connections [33, 9, 2]. This approach could provide very different insights compared to geodesic distances, especially when evaluating the reach in a diffusion process. Individuals can be reached better 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 modeling the latent space versus accounting for geodesic reach [9]. While latent (Euclidean) space models account for additional structure in the network formation process[9, 2, 30, 33] , they do not provide access to network measures for various sub-networks, since they do not leave much for interpretation when representing the network in a Euclidean space. 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 connections. We follow the argument of assortative mixing and homophily [53, 57, 61], suggesting that individuals in a social network tend to connect with rather similar people. This phenomenon leads to structural patterns 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.

More recent work on community detection tends to account for overlapping structures that allow individual to belong to several communities [1, 28, 81]. Yang et al. (2012) propose the affiliation graph model that allows for detection of dense overlaps among the communities in a network[81]. 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 strengths[1]. We adopt the model of MMSB[1, 28] and extend it to allow for more flexible specification and scalable inference. A common missing piece in many network studies in the field of marketing is that most measures are only evaluated in small scale networks[37, 73, 74]. Although several studies have used large networks[9,3,31,27,72], model inference and estimation about latent traits or individual and link level parameters have remained a challenging task. We try to address this problem by using a stochastic variational inference.

Additionally when dealing with large scale data, little to no attention is given to the directedness of connections in the network. Many social and relational data structures arise from the directed connections between the nodes. Examples of this behavior could be observed in networks such as twitter followership [13, 44], co-authorship and citation networks[50, 43]. Although many treatments of network connections assume undirectedness [9, 28, 62], a lot of important features underlying the edges could be misjudged or lost, and ultimately lead to biased estimates. In our model, we treat the network as directed to eliminate these shortcomings.

**2 Model**

In this paper we propose a model-based approach for detecting overlapping communities. Our approach is an extended version of mixed-membership-stochastic-blockmodel(MMSB)[1, 28], that allows for both scalable and efficient inference, as well as more flexible community definitions. 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], protein interaction networks[1, 28], and citation networks[16], among others. 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. The summary of the generative process is shown in algorithm X.

* + )
  + )

Several methods have been applied to estimate the model parameters, among which variational inference[41, 1, 28] and MCMC[14, 48] are prevalent. Later in this section we discuss the variational method, as it is the approach we take for our inference engine. Both variational methods and MCMC for this MMSB have excelled to scale to very large networks through introducing stochastic mini-batch sampling[34, 28, 48]. Although widely applied, all these models face some practical and technical limitations. 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. In our model, we address this limitation by not disregarding the direction of the links. Although this might come with a cost, as we incorporate more information which makes the inference more computationally expensive, we guarantee a scalable and efficient algorithm through mini-batch sampling within stochastic variational inference. The model explained in algX does not account for correlated community memberships. In correlated community membership, individuals who belong to rather similar clusters can also connect to each other. Although the simpler model exhibits conjugacy, which 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. 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 [7, 32, 22, 80](Although investigation of correlations between communities and the dynamic evolution of networks is postponed for future research). We take a hierarchical Bayesian perspective that allows for fully Bayesian variational inference, in addition to preserving the direction of the links. The model is summarized as follows:

)

)

## 2.1 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 posterior, include Stochastic Riemannian Langevin Dynamics Monte Carlo applied to the problem of community detection [48]. Instead we use the variational inference, widely applied to 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 and a simpler proposed variational distribution. Hence, instead of making exact inference through approximation, variational inference tend to offer deterministic approximation to 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 and close enough results to the true posterior.

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. For data and all latent variables and parameters The KL-divergence is as follows:

Where is known as the Evidence Lower Bound(ELBO), and as can be seen the problem of minimizing the KL-divergence translates to the maximization of the ELBO as the maximization is independent of the term .

### 2.1.1 Stochastic Variational Algorithm

Variational inference offers a fast approximation of the posterior distribution by optimizing the Evidence Lower BOund (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), samples only a small mini-batch, where iterating over the noisy gradients acquired by the sampled batch is proven to converge. 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]. We first provide the proposal distribution based on the mean field assumption, and then provide the parameter updates according to stochastic variational inference. Since large networks exhibit a very sparse patterns of connections, at each iteration we sample few nodes with all their links and equal proportion 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.

The log joint model of data and latent variables and parameters and the corresponding proposal distribution follows:

# Maximizing the ELBO according to each variational parameter in the mini-batch sampling we get the updates according to the algorithm in X.

* + - * + using numerical optimization

# 3 Data

To evaluate our model we apply the model inference to a two synthetic networks, where we can see the ground truth communities and compare the results. The first network consists of 150 nodes within 7 communities, and 2780 links, and the second network consists of 1000 nodes, 25 communities, and 36744 links. Detecting communities and learning parameters for email network is in progress.

# 4 Results

Figure X,Y shows the community membership strengths against the true strengths for the synthetic network. As can be seen in figure X, the estimated memberships collide with the truth, however to find out how well the community assignment of each node is presented we compute the normalized mutual information as prescribed by [52, 46]. NMI scores of closer to 1 represent the higher overlaps between the estimated and true community assignment. We found for both networks NMI of higher than 0.80 in all of our experiments.

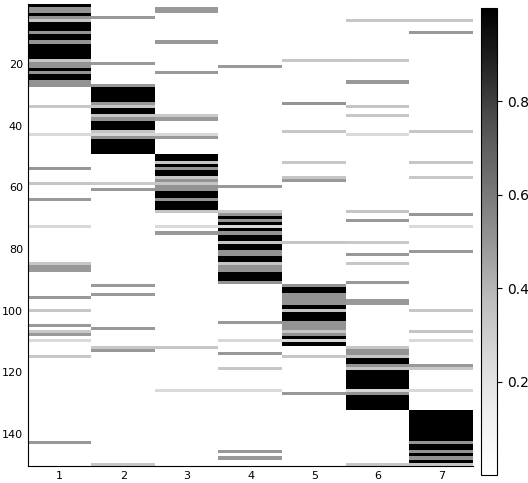
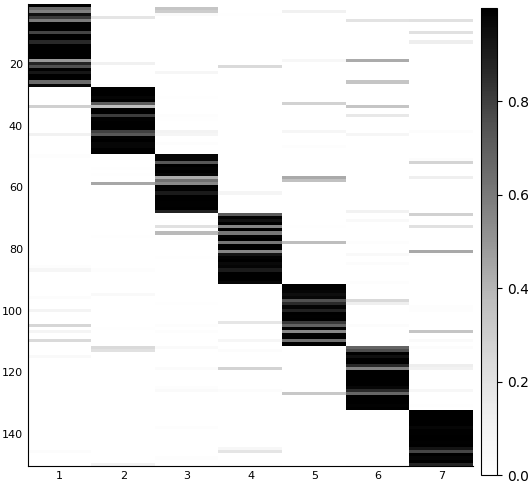


Figure 2 True(left) versus estimated(right) community membership strengths for a network with 150 node and 7 communities. The vertical axis represents individuals, and the horizontal axis represents different communities. The heatmap shows the strength of belonging to each group.

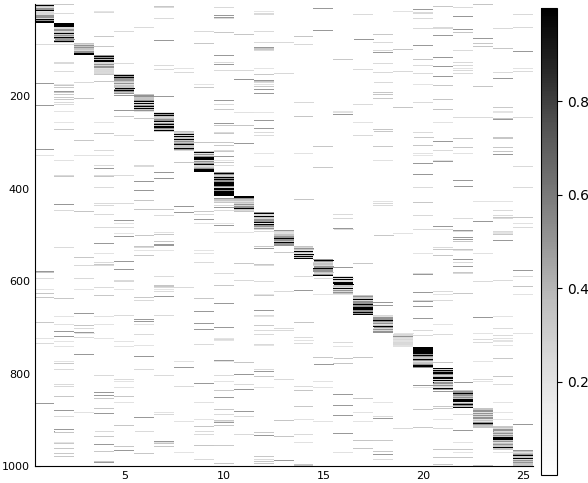
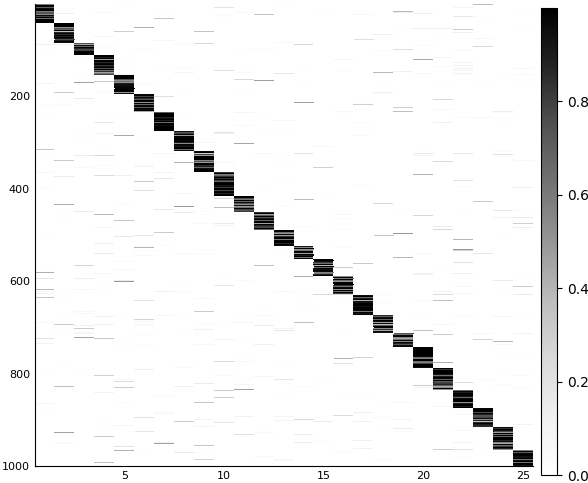


Figure True (left) versus estimated(right) community membership strengths for a network with 1000 node and 25 communities.

# Furthermore, for the synthetic network with 1000 nodes, we observe that more influential nodes are not necessarily influential in different communities. Figure x depict this distinction by graphing the top influential individuals based on their out-degree centrality (blue) compared to their centrality in network (red) for some select communities.

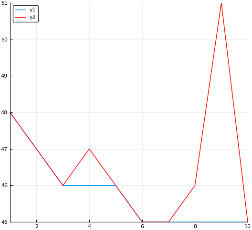
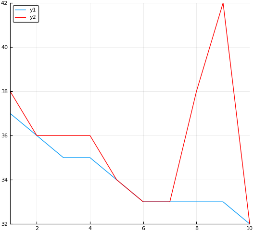
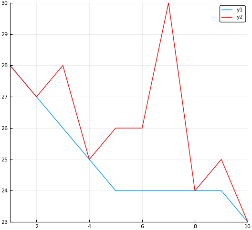
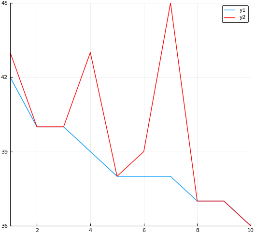
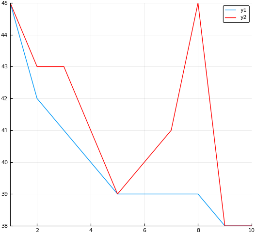


Figure graph of community level (blue) versus network level (red) centralities for different estimated communities

# 5 Discussion

Implications of finding overlapping communities can be manifold. To gain better understanding of diffusion of ideas(),products[73, 4],medical innovation[74,17], 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[68]. Conveniently addressing latent homophily and using a proxy estimation could improve the estimation of influence versus homophily[68, 67, 18]. 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 [68, 67].

Community detection proves imperative in finding counter-intuitive patterns of individual importance in a social network. However in many observed networks the actual memberships are not quite obvious and models that try to find these patterns are fundamental in improving our understanding about detecting influentials and opinion leaders. In contrast with a lot of studies related to detecting influentials where the use of sociometric data is limited to network level measurements, we suggest that looking at multiplex of subnetworks comprising the full network can provide better insights regarding opinion leadership.

# 6 Future Research

Connected to this work we are trying to understand how these smaller patterns translate to behavioral decisions. Adding behavior allows us to jointly model the network structure and behavior together and detect better communities. Furthermore controlling for the latent space simplifies the disentanglement of influence from homophily. This can be crucial in the study of social influence. Furthermore our formulation, especially with adding the logistic normal prior, makes it possible to study the trajectory of network and behavior in a dynamic setting. As diffusion by nature is a dynamic phenomenon, this would be vital in understanding the shift of behavioral and topological related preferences.

**References**