Title

It is widely known that individuals in a network form smaller communities, where each individual can belong to several groups. 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. Traditionally the problem of clustering networks included partitioning graphs into separate groups, whereas more recent works represent the embodiment of more natural processes such as shared community memberships. Common approaches in community detection do not treat the relationship between them. In this paper we allow for both overlap between communities and correlations between them. We embed these correlations among individual membership strengths via introducing a logistic normal prior for the membership simplex. 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. We further investigate the ability of our model to detect overlapping communities in both synthetic and real world large scale networks, and use the results to highlight the superiority of community level measures in comparison to conventional network level centrality measures to detect influential.

Keywords: up to three

Although I did not talk about stochastic variational inference, non-conjugacy, subsampling, and influentials much

Introduction

Many real life networks exhibit smaller grouped structures [5, 60]. Some of these smaller groups are portrayed as almost disconnected sub-networks [58, 70]. More recently these smaller patterns are more accurately described by multiple interconnected networks instead of one homogeneous network [81, 79, 46, 1]. While theoretically we still expect much ability from the network measures studied till now to predict behaviors and patterns observed in a social network [26], empirically this impact will be blurred because the network is in fact a mixture of multiple networks. This leaves the empirical assessment of network measures on a shaky foundation. Identifying sub-networks could still be valuable even in the absence of behavioral data. For example accounting for shared communities in a directed networks, can shed light on detection of important individuals such as information brokers sitting at the edge of the groups [29, 56]. Latent variable models have opened a new pathway in deriving intuition from the underlying patterns in network structure. 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; distance in this space, is interpreted as the similarity between individuals., where more similar nodes are more probable to form connections[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 in a diffusion process. 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 modeling the latent space versus accounting for 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, 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 friendships. We follow the argument of assortative mixing and homophily [53, 57, 61] 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 peer 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, 81].Yang et al (2012) propose affiliation graph model that allows for detection of dense overlaps in the community 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. For similar ideas of soft clustering and mixed membership in marketing refer to [76, 75, 40]. A common missing piece in many network studies in the field of marketing is that most measures are only evaluated in small scale networks or do not offer scalable methods to tackle the problem at hand. A related stream of diffusion research, studies the contagion phenomenon in new prescription drugs among a network of physicians [37, 73, 74]. Other examples, with larger networks include [31] that use a large network of mobile users, [27] with a Korean social network,[42] using a large European social network,[3] that use a large IM network,[72] that exploits both large simulation and field data. ...????. Although large networks have been in place but a lot of the recent studies face challenges in estimation when it comes to making inference about latent traits or individual and link level parameters. We try to address this problem by using a stochastic variational inference. Among studies that incorporate large scale networks are ([9],?) .

Additionally when dealing with large scale data, little to no attention is given 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 [50, 43], and many more [36]. Although many treatments of network connections assume undirectedness [9, 28, 62], 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. The distinction is important when it comes to detecting specific patterns, such as finding stars and fans [?], or experts and novices[], in networks.

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].

Somewhere here or after this, show comparison of network vs community level measures.

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, 74]. 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 finer 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, 72, 71, 74, 73, 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 contradicting results [71, 72]. 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 labeled differently. Detecting these subnetwork properties can facilitate understanding the likelihood of information exchange and the attention given to information [82]. 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 [51].

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 separate 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], protein interaction 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:

alg1

In the algorithm above the \_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 N. For each directed pair (a; b) that belong to the edge set E, we acquire the indicator vector z;for each contact point from a categorical distribution that is parametrized by their membership strengths.

Finally a diagonal Block(compatibility) matrix B determines the strength of inter-community connections based on what role is activated for each node. Due to assortativity of many real world networks we assume here that B is diagonal. Several methods have been applied to estimate the model parameters, which among them variational inference[41, 1, 28] and MCMC[14, 48] 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, 48].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 mini-batch sampling within stochastic variational inference. Upon detecting communities as explained by the data generating process in alg 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. 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, 80]. Our proposed model is closely related to the LNMMSB in [80], 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:

Algo2

The data generating process for potential link formation between two individuals a and b in 2 is shown in DAG in figure 4. In this figure filled dots indicate model hyper-parameters, and shaded node yab represents observed connection between two nodes. Directed arrows indicate the direction of the influence.

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[55], 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 including collapsed Gibbs sampler [49], or Hamiltonian Monte Carlo[10] were developed to address this issue by using small mini-batches. For more examples of scalable Monte Carlo methods see[23, 63, 78], and for an example of the use of Stochastic Riemannian Langevin Dynamics Monte Carlo applied to the problem of community detection see[48]. 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 simpler proposed variational distribution q. 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 a fully Bayesian framework, where we set all the latent variable to be Z and all the observed variables to be X, we specify a joint probability model P(X;Z)1, and our goal is to find an approximation to the true posterior P(ZjX) and also our model evidence P(X):The log likelihood of the model follows:

eq0

We further use the log transformation:

Eq1

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

eq2

Equation 2 is known as the evidence lower bound(ELBO). Note that we can write the equation 11 in the following format following the fact that P(X;Z) =P(ZjX)P(X)

Eq3

Note that xx in equation3 is equivalent to the Kullback-Leibler divergence of the proposed variational distribution q(Z) and the true posterior P(ZjX). Hence we can rewrite the log marginal as the following:

eq4

To simplify the Kullback-Leibler divergence we can rewrite

Eq5

According to equation 5, minimization the KL-divergence between the variational distribution and the true posterior translates to maximization of the ELBO as the marginal distribution P(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.

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[66][6](??).

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.

Maybe directly jump to minibatch sampling

Minibatch sampling:

Stochastic Variational Algorithm for Directed LNMMSB

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 Monro[65]. According to SVI, our algorithm only needs to iterate between subsampling 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, 48] for the non-link \_'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,

XXXX

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 5.

large network

Real World Data

political blog citation

bookmarking network

Results

synthetic

Small network

Performance Checking

Figure 6 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 \_at 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 7. The plot visually provides a reasonable closeness

of our estimates and the ground truth for a small network. However for further 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.

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[52, 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.

Performance check

Perplexity