Modeling Network Structure via Correlated Community Membership

Abstract: In large social networks individuals form connections for a variety of reasons. One such reason is when individuals share a common interest that is not shared by others. Such individuals can be seen as members of a community of interest. In general, the network consists of multitude of such communities. As network formation and interactions depend on the links within and between these communities, it is vital to detect and understand their underlying structure. A large body of literature has developed theories about the role and features of influentials in social networks. However when the network is a blend of smaller communities, we need to revisit the detection of such influentials at the sub-network level rather than globally. 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 of our model 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 role of structure in social influence (Aral, Muchnik, & Sundararajan, 2009; Braun & Bonfrer, 2011). Goldenberg, Han, Lehmann, and Hong (2009), show that actors with higher degree centrality (hubs) can speed up contagion or lead to higher volume for diffusion. Stephen and Toubia (2010) 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. Goel and Goldstein (2013) have shown that social data can prove to be beneficial and complementary to behavioral data in terms of predicting the future contagion. 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, and clustering coefficient (Goldenberg et al., 2009; Hinz, Skiera, Barrot, & Becker, 2011; Katona, Zubcsek, & Sarvary, 2011). Although these measures are still valuable in predicting and studying the diffusion (Iyengar et al., 2011), we argue that sub-network level measures can provide further insights into recognizing finer grained measures. Moreover, marketing science is always interested in finding influential people in a network to be able to target and segment them properly (Stephen & Toubia, 2010; Trusov, Bodapati, & Bucklin, 2010; Aral & Walker 2012). Yet not all centrality measures on the network level correspond to similar notions and could provide contradicting results. Although these measures have been used to indicate the level of importance of individuals in influence literature, some evidence signal their misfit. For instance, Trusov et al. (2010) argue that the mere friend counts in a social network does poorly when it comes to detecting influential users. Hence, delving into more robust constructs at the sub-network level might be a worthwhile investigation into the potential roles of specific individuals. For instance Ansari, Koenigsberg, and Stahl (2011) model several sub-networks of professionals to study the impact of the organizational interventions on the nature of the connections.

As an illustration for the relevance of sub-network level measures, figure 1. shows the graph of degree centrality of three different communities in a publicly available email network of a European Research Institute extracted from SNAP dataset (Leskovec & Krevl, 2014). The connections are formed based on the emails sent between members of this institute, and the ground truth communities are the departments within the institute. As can be seen, nodes that are central (measured by degree) at the community level are not necessarily central at the network level.

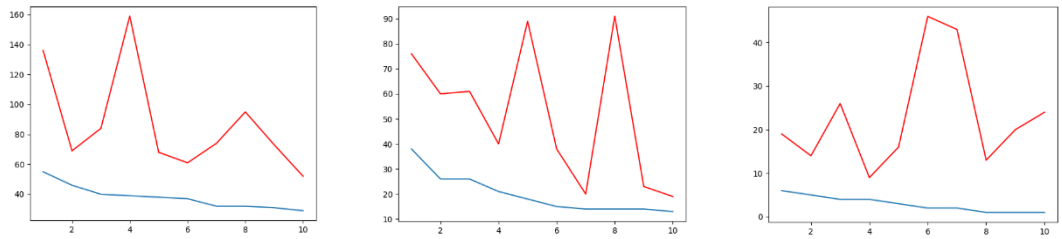


Figure 1. Three different communities, where each community is sorted based on the degree centrality (blue line) and the corresponding network centrality for that node is shown on top (red line)

Many real life networks exhibit smaller grouped structures (Newman, 2006). Sometimes these smaller patterns are more accurately described by multiple interconnected sub-networks instead of one homogeneous network (Lancichinetti, Fortunato, & Kertész, 2009). While theoretically we can expect existing network measures to provide predictive power for observed behaviors and patterns

in a social network (Goel & Goldstein, 2013), empirically this power will be undermined because the network is in fact a mixture of multiple sub-networks.

Understanding how individuals are connected in a network can be crucial in designing marketing strategies. Braun and Bonfrer (2011) argue that finding hidden patterns through latent space modelling in network structure can lead to better insights for devising strategies to reach more relevant individuals in a diffusion process. While latent space models account for additional structure in the network formation process, they do not provide access to network measures for various sub-networks. We employ a model-based approach that allows us to define the network structure according to homophily (McPherson, Smith-Lovin, & Cook, 2001), suggesting that individuals in a social network tend to connect with rather similar people. This phenomenon leads to structural patterns in networks with denser intra-group links and fewer inter-group connections.

Real world networks feature overlapping structures that allow individual to belong to several groups (Airoldi, Blei, Fienberg, & Xing, 2008; Yang & Leskovec, 2012). Airoldi et al. (2008) suggest a mixed-membership-stochastic blockmodel (MMSB) that allows individuals to belong to multiple groups by estimating community membership strengths. We adopt MMSB, and extend it to allow for more flexible specification and scalable inference. Commonly, many network studies employ measures that are only evaluated in small scale networks to explain different diffusion processes (Iyengar et al., 2011; Van den Bulte & Lilien, 2001). Although studies exist to have used large networks (Aral et al., 2009; Braun & Bonfrer, 2011), model inference and estimation about latent traits at the individual or link level states have remained a challenging task. We address this problem by using a stochastic variational inference, a method that scales to a large parameter space by iterating over small random samples (Hoffman, Blei, Wang, & Paisley, 2013). Additionally, most large scale community discovery models ignore the direction of links. Many social data structures arise from the directed connections between the nodes (Kempe, Kleinberg, & Tardos, 2003). In our model, we treat the network as directed to eliminate these shortcomings.

2.Model

We propose a model-based approach for detecting overlapping communities. Our approach is an extended version of the MMSB (Airoldi et al., 2008), in which we allow for scalable inference and more flexible model specification. MMSB, defines a generative setting for the formation of the links in a network. This model has been applied frequently to finding overlapping communities (Airoldi et al., 2008; Cho, Ver Steeg, Ferrara, & Galstyan, 2016). The summary of the generative process for a network of N individuals in K communities is shown in Algorithm 1.

$\forall a \in N$
$\quad draw\ a\ K-dimensional\ mixed\ membership\ vector, \theta_a \sim Dirichlet(\alpha)$
$\forall (a,b) \in E$
$\quad draw\ one-hot\ membership\ indicator\ for\ 'a'\ when\ contacting\ 'b', z_{a\to b} \sim Categorical(\theta_a)$
$\quad draw\ one-hot\ membership\ indicator\ for\ 'b'\ when\ contacted\ by\ 'a', z_{a\leftarrow b} \sim Categorical(\theta_b)$
$\quad sample\ a\ link\ a \to b\ with\ probability\ z_{a\to b} B z_{a\leftarrow b}, Y(a,b) \sim Bernoulli(z_{a\to b} B z_{a\leftarrow b})$

Algorithm 1. Generative process for MMSB

Several methods have been applied to estimate the model parameters for MMSB, among which variational inference and MCMC have excelled to scale to very large networks through introducing stochastic mini-batch sampling (Gopalan & Blei, 2013; Li et al., 2016). However the model in

algorithm 1. does not account for correlated community memberships. In correlated community memberships, individuals who belong to rather similar communities can also connect to each other. A more natural way to allow for correlated mixed memberships is by introducing a Logistic-Normal prior instead of Dirichlet for membership probabilities. In the case of our proposed MMSB variant, this would provide an advantage when moving from static to dynamic settings, where the LN-distributed parameters can change according to a simple autoregressive rule (Although investigation of correlations between communities and the dynamic evolution of networks is postponed for future research). The model is summarized in Algorithm 2.

$\forall k \in \{1, \cdots, K\}$

$draw\ diagonal\ elements\ of\ the\ block\ matrix\ B\ via\ \beta_{kk} \sim Beta(\eta_0, \eta_1)$

$\forall a \in N$

$draw\ the\ mean\ of\ logit\ mixed\ membership\ vector, \mu \sim Normal(\mu_0, \Lambda_0)$

$draw\ the\ precision\ of\ the\ logit\ mixed\ membership, \Lambda \sim Wishart(\ell_0, L_0)$

$draw\ a\ K - dimensional\ vector\ for\ logit\ mixed\ membership\ \theta_a^* \sim Normal(\mu, \Lambda)$

$construct\ the\ simplical\ mixed\ membership\ via\ softmax, where\ \theta_{a,k} \sim \frac{\exp(\theta_{a,k}^*)}{\sum_l \exp(\theta_{a,l}^*)}$

$\forall (a, b) \in E$

$draw\ one - hot\ membership\ indicator\ for\ a\ when\ contacting\ b, z_{a \to b} \sim Categorical(\theta_a)$

$draw\ one - hot\ membership\ indicator\ for\ b\ when\ contacted\ by\ a, z_{a \leftarrow b} \sim Categorical(\theta_b)$

$sample\ a\ link\ a \to b\ with\ probability\ z_{a \to b} B z_{a \leftarrow b}, Y(a, b) \sim Bernoulli(z_{a \to b} B z_{a \leftarrow b})$

Algorithm 2. Generative process for logistic normal MMSB

3.Inference and Estimation

In this section we introduce the variational inference (VI) method which transforms the problem of inference to an optimization one, by trying to minimize the Kullback-Leibler divergence between the true posterior distribution p and a simpler proposed variational distribution q. Hence, instead of making exact inference through stochastic approximation, variational inference uses a deterministic approximation of the model posterior distribution. In its simplest case, the proposed model follows a mean field assumption, which decouples parameters in a way that we can still have tractable and close enough results to the true posterior.

For data X and all latent variables and parameters Z, the KL-divergence that is minimized by VI is given by:

$$KL(q(Z)||p(Z||X)) = -\big(E_q[\ln p(X, Z)] - E_q[\ln q(z)]\big) + \ln p(X) \qquad (1)$$

3.1 Stochastic variational inference algorithm

VI offers a fast approximation of the posterior distribution by maximizing the first two terms in the right hand side of equation (1), known as the Evidence Lower Bound (ELBO). However this might need the screening of *all* individual (link) level observations for updating the variational parameters. On the other hand, Stochastic Variational Inference (SVI) offers a stochastic search in the parameter space suggested by Hoffman et al. (2013). SVI samples only a small mini-batch, where iterating over the noisy gradients acquired by the sampled batch is proven to converge. Gopalan and Blei (2013) offer several sub-sampling schemes, including the link-only sampling

which provides efficient inference for undirected networks. Adding community correlation and link direction make the inference problem even more computationally expensive. But using SVI combined with our sampling scheme, allows us to have scalable and efficient inference. Since large networks exhibit very sparse patterns of connections, at each iteration we sample few nodes with all their links and a small proportion of their randomly selected non-links. After rounds of iteration, this assumption both takes into account the information of all links and non-links. The log joint model of data, latent variables and parameters is given below:

$$\begin{aligned}\ln p(joint) = {} & \ln p(\mu|\mu_0, \Lambda_0) \\ & + \ln p(\Lambda|\ell_0, L_0) + \sum_a \ln p(\theta_a|\mu, \Lambda) + \sum_{a,b} \ln p(z_{a\to b}|\theta_a) \\ & + \sum_{a,b} \ln p(z_{a\leftarrow b}|\theta_b) + \sum_k \ln p(\beta_{kk}|\eta_0, \eta_1) + \sum_{ab} \ln p(y_{ab}|z_{a\to b}, z_{a\leftarrow b}, \beta)\end{aligned} \quad (2)$$

The corresponding log joint probability of proposal distribution follows:

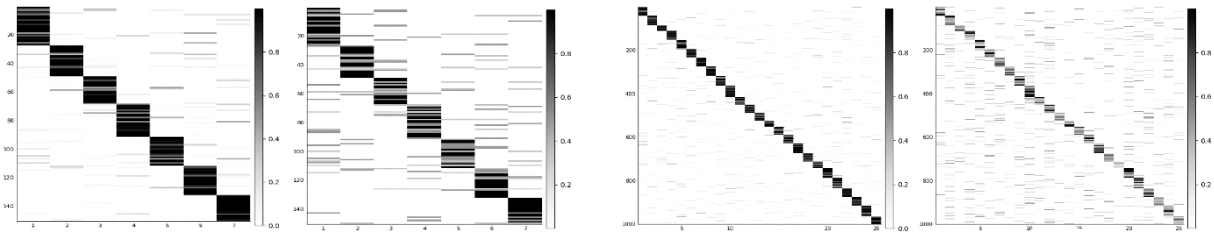
$$\begin{aligned}\ln q(joint) = {} & \ln q(\mu|m, M) \\ & + \ln p(\Lambda|\ell, L) + \sum_a \ln p(\theta_a|\mu_a, \Lambda_a) + \sum_{a,b} \ln p(z_{a\to b}|\phi_{a\to b}) \\ & + \sum_{a,b} \ln p(z_{a\leftarrow b}|\phi_{a\leftarrow b}) + \sum_k \ln p(\beta_{kk}|b_{k0}, b_{k1})\end{aligned} \quad (3)$$

Maximizing the ELBO according to each variational parameter in the mini-batch sampling we get the final updates for parameters by either closed form, or numerical solutions that are modified according to a decaying learning rate (Hoffman et al., 2013).

4.Data Analysis

To evaluate our model we apply the model inference to two synthetic networks, where we can compare the results with ground truth communities. The first network consists of 150 nodes, 7 communities, and 2780 links, and the other has 1000 nodes, 25 communities, and 36744 links.

Figures 2. shows the community membership strengths against the true strengths for the synthetic networks, where the estimated memberships collide very closely with the ground truth membership vectors.



(a) (b) (c) (d)

Figure 2. True (a, c) versus estimated (b, d) community membership strengths for two networks with 150 node and 7 communities (a, b), and 1000 node and 25 communities (c, d). Nodes on vertical axis, and communities on horizontal axis. The heatmap shows the strength of belonging to each group.

Furthermore, we observe that the most influential nodes at the network level, are not necessarily influential in different communities. Figure 3. depicts this distinction by graphing the out-degree centrality in estimated (blue), and ground truth community (green), versus in network (red) for top influential individuals in a few communities. The estimated and truth degrees are closely related

and follow the same patterns, however network level measures do not correspond with the degree of nodes in these communities.

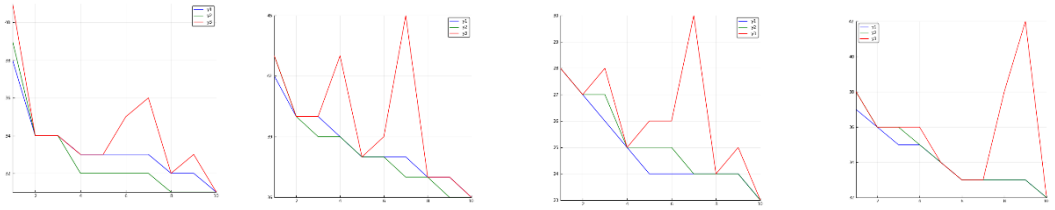


Figure 3. Estimated community (blue) versus truth (green), and network level (red) out-degrees for some estimated communities.

5.Discussion

Community detection proves imperative in finding counter-intuitive patterns of individual influence. 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 studies related to detecting influentials using network level measure, we suggest that looking at multiplex of sub-networks can provide better insights regarding opinion leadership.

Implications of finding overlapping communities can be manifold. Behaviors and decisions made by many individuals in observed networks tend to assimilate both in node space and in time (Aral et al., 2009). However disentangling the underlying reasons for such observations can become infeasible due to endogenous network formation (Shalizi & Thomas, 2011). Addressing latent homophily and using a proxy estimation for it could improve the estimation of influence (Davin, 2015). Any model that estimates the latent factors that drive link formation, helps in solving the identification problem of disentangling influence from homophily.

6.Future Research

Detecting communities and learning parameters for the larger real world networks such the one referenced in Figure 1 is still in progress. As a continuation of this work we aim to understand how sub-networks translate to behavioral patterns. Adding behavioral data allows us to jointly model the network structure and behavior and better detect communities. Furthermore controlling for the latent space simplifies the disentanglement of influence from homophily. This can be crucial in the study of social influence. Moreover, adding the logistic normal prior, we intend to study the network and behavior evolution.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, *9*(Sep), 1981–2014.

Ansari, A., Koenigsberg, O., & Stahl, F. (2011). Modeling multiple relationships in social networks. *Journal of Marketing Research*, *48*(4), 713–728.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), 21544–21549.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, *337*(6092), 337–341.

Braun, M., & Bonfrer, A. (2011). Scalable inference of customer similarities from interactions data using Dirichlet processes. *Marketing Science*, *30*(3), 513–531.

Cho, Y.-S., Ver Steeg, G., Ferrara, E., & Galstyan, A. (2016). Latent space model for multi-modal social data. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 447–458). International World Wide Web Conferences Steering Committee.

Davin, J. (2015). *Essays on the Social Consumer: Peer influence in the adoption and engagement of digital goods*.

Goel, S., & Goldstein, D. G. (2013). Predicting individual behavior with social networks. *Marketing Science*, *33*(1), 82–93.

Goldenberg, J., Han, S., Lehmann, D. R., & Hong, J. W. (2009). The role of hubs in the adoption process. *Journal of Marketing*, *73*(2), 1–13.

Gopalan, P. K., & Blei, D. M. (2013). Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, *110*(36), 14534–14539.

Hinz, O., Skiera, B., Barrot, C., & Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, *75*(6), 55–71.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, *14*(1), 1303–1347.

Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, *30*(2), 195–212.

Katona, Z., Zubcsek, P. P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research*, *48*(3), 425–443.

Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, *11*(3), 033015.

Leskovec, J., & Krevl, A. (2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. Retrieved from http://snap.stanford.edu/data

Li, W., Ahn, S., & Welling, M. (2016). Scalable MCMC for mixed membership stochastic blockmodels. In *Artificial Intelligence and Statistics* (pp. 723–731).

McDaid, A. F., Greene, D., & Hurley, N. (2011). Normalized mutual information to evaluate overlapping community finding algorithms. *ArXiv Preprint ArXiv:1110.2515*.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2), 211–239.

Stephen, A. T., & Toubia, O. (2010). Deriving value from social commerce networks. *Journal of Marketing Research*, *47*(2), 215–228.

Trusov, M., Bodapati, A. V., & Bucklin, R. E. (2010). Determining influential users in internet social networks. *Journal of Marketing Research*, *47*(4), 643–658.

Van den Bulte, C., & Lilien, G. L. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, *106*(5), 1409–1435.

Yang, J., & Leskovec, J. (2012). Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on* (pp. 1170–1175). IEEE.