A block model suitable for sparse graphs

Juuso Parkkinen JUUSO.PARKKINEN@TKK.FI
Helsinki University of Technology, Department of Information and Computer Science

Adam Gyenge ADAMGYENGE@ILAB.SZTAKI.HU
Data Mining and Web Search Research Group, Informatics Laboratory, Computer and Automation Research Institute, Hungarian Academy of Science, Budapest, Hungary

Janne Sinkkonen JANNE.SINKKONEN@TKK.FI
Helsinki University of Technology, and Xtract Ltd., Hitsaajankatu 22, 00810 Helsinki, Finland

Samuel Kaski SAMUEL.KASKI@TKK.FI
Helsinki University of Technology, Department of Information and Computer Science, P.O. Box 5400, FI-02015 TKK, Finland

Keywords: block model, collapsed Gibbs, hierarchical Bayes, latent variables, statistical network analysis.

Abstract

We introduce a new generative block model for graphs. Vertices (nodes) have mixed memberships in *margin components*, and edges arise from a multinomial defined over the cartesian product of the margin components. The model is able to represent block structures of "non-community" type, that is, it is able to model linkage between margin components. Compared to earlier mixed membership stochastic blockmodels which have a Bernoulli parameterization for the generation of links between each margin component pair, in the new model collapsed Gibbs samplers need to represent only those interactions with realized data in them, making possible large and sparse block models.

1. Introduction

Generative models for graphs can be divided to three common subtypes. (1) In latent-space models nodes reside in a continuous latent space, and usually a logistic link produces Bernoulli probabilities for linkage (e.g., Handcock et al., 2007; not discussed here). (2) Only links *within* blocks of nodes are modelled in community models (Hofman & Wiggins, 2008; Sinkkonen et al., 2007). (3) In block models (Airodi et al., 2008), linkage between blocks is also modelled. Community models are suitable for natural social networks and in general where tightly integrated subnetworks exist, or linkage between "communities" is not interesting. Their advantage is simpler parameterization, leading to better estimates.

Full block models have parameters for linkage between all node subgroups or "communities". If node subgroups are presented on margins of a contingency table, each cell of the table corresponds to a potential interactions that needs to be parameterized—hence the name "block model". Airoldi et al, (2008) have introduced a *mixed membership stochastic blockmodel* (MMSB), which assigns each node on multiple subgroups, effectively capturing the fact that nodes in a network may arise from different sources and have different roles. Linkage between node groups is represented by Bernoulli distributions associated to each cell of the block interaction table.

In this paper we introcude an alternative block formulation. Nodes still have mixed memberships, but links arise from a multinomial spanned over the pairs of node subgroups. That is, the Bernoulli parameters of MMSB in the cells of the contingency table are replaced by a multinomial over all the cells. The model is able to generate multiple links for pairs of nodes, but on sparse graphs where the proportion of linked pairs p is small, the number of doubly linked pairs

Presented at *ILP-MLG-SRL*, Leuven, Belgium, 2009.

is on the order of p^2, that is, vanishingly small, and the multinomial parameterization approximately corresponds to the Bernoulli parameterization.

An advantage of the multinomial parameterization is easy estimation with collapsed Gibbs sampling. The implementation does not need bookkeeping of cells with no data, making even quite large models with a high number of components feasible.

In this paper we introduce the model structure and demostrate it with toy data. Comparisons to existing models and applications to larger datasets will appear in the future, as well as a hierarchical mechanism to allow the use of Dirichlet Process priors.

2. Simple interaction block model

A plate diagram of the block model is shown in Figure 1. The key insight is that the *links* can be seen to belong to the latent components z which directly determine a pair of margin components, $z = (z_1, z_2)$, into which the nodes belong. The z correspond to cells of the contingency table between margin component pairs. The nodes are sampled according to node memberships $p(i|z_1) = \phi_{z_1 i}$ and $p(j|z_2) = \phi_{z_2 j}$ (nodes i or j, node subgroups z_1 and z_2). The Dirichlet prior for ϕ_z is parameterized by β, and here β optionally arises from a Gamma distribution.

The distribution of links over the cells of the contigency table is denoted θ, which arises from a Dirichlet with parameter α, and that is again either fixed or arises from a Gamma prior. In summary:

1. Initialization
   (a) Optionally, generate (α, β) from Gamma.
   (b) From $Dir(\alpha)$, draw parameters for the multinomial linkage distribution θ over pairs $z = (z_1, z_2)$ of margin components.
   (c) From $Dir(\beta)$, draw multinomial parameters ϕ_z of the margin node subgroup memberships, one multinomial per one margin component.
2. For each link $l = (i, j)$:
   (a) Draw a component z from θ, which determines margin components z_1 and z_2.
   (b) Draw the link endpoints, or interacting nodes, i and j, from ϕ_{z_1} and ϕ_{z_2}, respectively, and set a directed link between them.

The number of components is fixed. A hierarchical Dirichlet Process arrangement is an interesting alternative, not considered here. Hyperparameter α controls how evenly links are distributed over the node

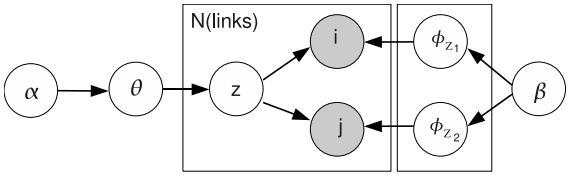

Figure 1. A plate diagram of the block model.

subgroup pairs. Links are often allocated to only a small proportion of all the pairs, because large practical networks are sparse.

We estimate the model with collapsed Gibbs. The sampler iterates over links, and samples a new latent component for each link l at a time, conditional on other links and their assigments:

$$p(z_l|\{z\}^{\neg l}, \{(i,j)\}^{\neg l}, \alpha, \beta) \propto (n_z^{\neg l} + \alpha) \cdot \frac{(q_{z_1 i}^{\neg l} + \beta)(q_{z_2 j}^{\neg l} + \beta)}{(q_{z_1 \cdot}^{\neg l} + M\beta)(q_{z_2 \cdot}^{\neg l} + M\beta + \delta_z)}, \quad (1)$$

where n is the count over the component pairs (bins of θ, cells of the contingency table), and q counts component-node co-occurences. M is the number of nodes, and $\delta_z \in \{0, 1\}$ is one for the diagonal (z_1, z_2).

If hyperparameters are not fixed, they are independently drawn from their posteriors, or alternatively set to their MAP values. This is repeated after each full round over link assigments. Note that the model generates directed links. A modification allows generation of undirected links.

3. Experiments

We applied the model to a toy data with a block structure. Adjacency matrix of the nodes, with notable amount of noise, is shown in Figure 2 (top; 50 nodes, 10,000 links of which about half is noise). Underlying is a structure of five node margin groups, with varying linkage within nine of the potential 25 interactions. The populated interactions can be seen as darker areas in the adjacency node matrix, which counts the links and where the nodes are sorted to make subgroup nodes adjacent. Figure 2 (bottom) shows the block structure correctly inferred with the model, with MAP hyperpriors.

In the second experiment, we apply the model to the famous friendship graph of 18 monks (Breiger et al., 1975), with 88 links describing their relationships. The monks are known to form three factions, but there are also three "waverers" with less clear affiliations. Figure 3 presents the inferred mixed memberships of

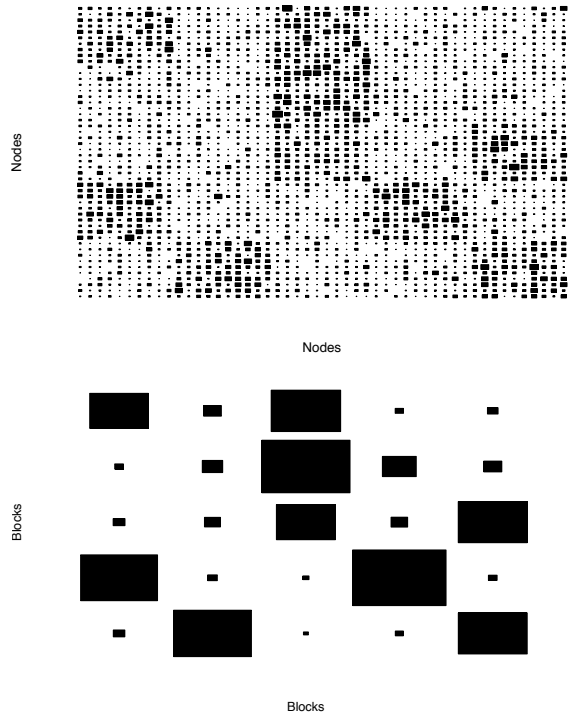

Figure 2. Top: Adjacency matrix, or link counts of node pairs, of the generated data, with the underlying block structure made visible by ordering nodes to make the within-block nodes adjacent. *Bottom:* Inferred block structure.

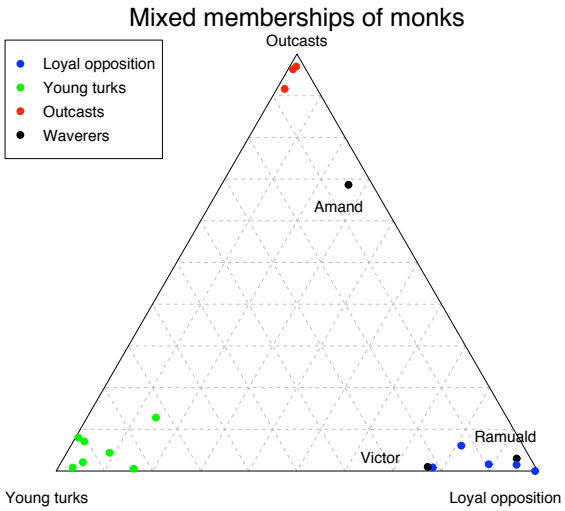

Figure 3. Inferred mixed memberships of the 18 monks. Colors indicate the correct affiliations of the monks, and the names of the three waverers are also shown. Positions indicate inferred membership degrees in the three groups. Underlying social structure, the factions (colors) is separated well by the model. Of the three waveres (black dots), Amand and Victor distinguish themselves while Ramuald sits firmly in "loyal opposition". Hyperparameters were fixed at $\alpha = 0.1$, $\beta = 0.1$.

each monk, projected on a simplex. The model is able to correctly separate the three known factions, and in addition captures some of the waverer nature of the three.

4. Discussion

We introduced a block model formulation for graphs, where nodes have mixed memberships over latent components, and the component interactions generate links. The interactions are parameterized in a multinomial style, which allows sparse representations for inference. Two small experiments showed that the model is able to find both communities and block structures from generated and real datasets.

Acknowledgments

SK, JS and JP belong to the Finnish CoE on Adaptive Informatics Research of the Academy of Finland and Helsinki Institute for Information Technology, and were partially supported by EU Network of Excellence PASCAL2.

References

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

Breiger, R., Boorman, S., & Arabie, P. (1975). An algorithm for clustering relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, *12*, 328–383.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal Of The Royal Statistical Society Series A*, *170*, 301–354.

Hofman, J. M., & Wiggins, C. H. (2008). A Bayesian approach to network modularity. *Physical Review Letters*, *100*, 258701–259900.

Sinkkonen, J., Aukia, J., & Kaski, S. (2007). Inferring vertex properties from topology in large networks. *Working Notes of the 5th International Workshop on Mining and Learning with Graphs (MLG'07)*. Florence, Italy: Universita degli Studi di Firenze.