OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

DISCRIMINATIVE RELATIONAL TOPIC ANALYSIS AND ROLE DISCOVERY IN SOCIAL NETWORKS

S. PUNITHAMARY[1], S. NAGENDIRA[2]

Asst. Professor, Department of Computer Applications, Idhaya College for Women, Sarugani, India

Asst. Professor, Department of Computer Science, Idhaya College for Women, Sarugani, India

Abstract: **Many knowledge sets may be described as a sequence of interactions between entities—for example communications between people in a very social network, protein-protein interactions or DNA-protein interactions in a very biological context, or vehicles' journeys between cities. In these contexts, there is typically interest in creating predictions concerning future interactions, admire who can message whom. A preferred approach to network modeling in a very theorem context is to assume that the discovered interactions may be explained in terms of some latent structure. As an instance, traffic patterns may be explained by the scale and importance of cities, and social network interactions may be explained by the social teams and interests of people. Unfortunately, whereas elucidating this structure may be helpful, it typically doesn't directly translate into a good prophetical tool. Further, several existing approaches don't seem to be applicable for thin networks, a category that has several fascinating real-world things. During this paper, we tend to develop models for thin networks that mix structure elucidation with prophetical performance. we tend to use a theorem statistic approach, that permits North American country to predict interactions with entities outside our coaching set, and permits the all the latent spatiality of the model and therefore the range of nodes within the network to grow in expectation as we tend to see a lot of knowledge. We tend to demonstrate that we are able to capture latent structure whereas maintaining prophetical power, and discuss do able extensions.**

Keywords: Dirichlet Process, Networks, Bayesian Non Parametrics, Gibbs Sampling, Hierarchical Modeling

I INTRODUCTION

We are frequently absorbed in describing and predicting the interactions between objects, be they characters within an organization, proteins within a cell, or transportation hubs within a region. We can signify these objects as nodes in a network, with the non-zero edges of the network describing the interactions between nodes. For example, we can characterize a social network as a binary network, where every node corresponds to a separate, and an edge between nodes corresponds to a friendship between individuals. Patterns of email communication can be modeled using an integer-valued network, with integer-valued edges representing the number of emails sent from one individual to additional. Interactions between proteins can be signified using a real-valued network, where the nodes correspond to proteins and the edges correspond to interaction strength. A number of statistical models for such networks have been planned. Numerous of these representations fall under the stochastic block model (SB) framework, where each node is supposed to belong to one of K latent groups, and the interaction between two nodes depends only on their group assignments. This basic model can be extended by permitting the number of latent collections to be unbounded, as in the infinite interpersonal model (IRM, Kemp et al., 2006), or by permitting each node to demonstration association in multiple latent groups, as in the mixed membership stochastic block model (MMSB, Airoldi et al., 2008). One thing that these models have in common is that they treat nodes as exchangeable, and assume that there exists a fixed, stationary network between these nodes. Every node is represented by the entirety of its interactions with other nodes, and we use this information to cluster (or, in the case of the MMSB, co-cluster) the data into separate groups.

In this paper we follow a different approach: we treat the interactions, rather than the nodes, as data points, and construct an exchangeable sequence of directed binary links. Each link agrees to a single interaction—such as “friending” or “liking” in a social network, or sending a single email—and is characterized in terms of an ordered pair of nodes. We may observe various links between two nodes; this corresponds to repeated interactions (for example, sending multiple emails).

This method has a number of advantages. Unlike the stochastic block model family, the approach described in this paper allows us to model sparse graphs, where the number of non-zero entries grows as O(M), where M is the numeral of nodes. Sparsity is a property of many real-life networks, which tend to exhibit small-world behavior (Caron and Fox, 2015; Orbanz and Roy, 2014).

Additional advantage is that our model is openly considered for the prediction task. In numerous developments, we might be interested in what the next interaction will be: who will email whom, for example. Stochastic block representations aim to model a fully observed network, where the absence of an observed edge is interpreted as an explicitly observed zero. In this setting, any predictions necessity directly contradict these observed zeros. While it is possible to openly mark edges as “missing”, we can only do this for a small subset of unobserved edges—if we assume all zero edges in a stochastic block model are in fact unobserved, the maximum probability network will have all the edges equal to one. Equally, by assembling an integer-valued network via an exchangeable sequence of links, we frame our problematic in a method that directly delivers a predictive distribution over the location of the next link, and allows us to continuously update our posterior predictive distribution in the face of new data. Further, by choosing to place a nonparametric distribution over the sequence of links, we can easily incorporate before unseen nodes, without any prior information of the number of such nodes.

1.1 Notation

We will use the notation Z to represent an M ×M network, with elements $z_{sr} \in N$ indicating the relationship between nodes s and r. If $z_{sr} \in \{0, 1\}$, then a non-zero value indicates the presence of a relationship. If z_{sr} is allowed to take on arbitrary non-negative integer values, we take this to indicate the number of interactions (for example, emails in a social network, packages in a computer network) between nodes s and r. Unless otherwise specified, we will assume Z to be a directed network, where Z 6= Z T . It will sometimes be more convenient to represent the matrix Z as a sequence of interactions Y = y1, y2, . . . , where each interaction yi consists of an ordered pair of nodes. We can reconstruct the matrix Z by setting zsr = P i I(yi = (s, r)), where I(·) represents an indicator function, that returns one iff the statement it refers to is true.

1.2 Nonparametric Models for Networks

A stochastic block models assume a fixed, fully observed network, where zero-valued entries are taken to represent the observed absence of an interaction, and model the network by clustering these nodes. We take a different approach: We model a network as a sequence of observed interactions, and aim to predict the locations of future interactions by explicitly clustering the interactions, rather than the nodes.

To do so, we consider distributions over a sequence of links connecting a set of nodes. Each link, therefore, is associated with an (ordered) pair of nodes experimented from some distribution over such pairs; we may have numerous links associated with a given pair. To allow the network to expand over time, and to facilitate out-of-sample prediction, we let this set of nodes be countable infinite and use a Bayesian nonparametric distribution to assign probabilities to potential pairs.

1.3 Dirichlet Network Distributions

A simple way of constructing an integer-valued network with an unbounded number of nodes is to place a probability distribution G over a countable infinite number of actors. We can represent such a network as a sequence of (sender, receiver) pairs; each pair might, for example, correspond to a single email from a sender to a receiver, or a single journey between two cities. The value of a (directed) edge from a “sender” s to a “receiver” r is the number of times we have seen the pair (s, r). We call each individual pair in the sequence a link; the value of an edge between two nodes is the number of links between them.

II COMPARISON METHODS

We compare the mixture of Dirichlet network distributions to a single symmetric Dirichlet network distribution; to integer-valued variants of the mixed-membership stochastic blockmodel (MMSB, Airoldi et al., 2008) and the infinite relational model (IRM, Kemp et al., 2006); and to two baseline methods.

1. Symmetric Dirichlet network distribution

We modeled the data using a single symmetric DND as described in Section 3.1, with Dirichlet process concentration parameter τ = 1.

2. Infinite relational model

Kemp et al. (2006) describe a variant of the IRM appropriate for integer-valued data. Each pair of clusters (i, j) is associated with a positive real-valued parameter θij , and the N links are assigned to clusters according to a multinomial distribution parameterized by the θij . Inference in this model is performed using existing C code released by the authors of Kemp et al. (2006). This code was not able to handle the number of nodes present in the Enron data sets.

3. Mixed-membership stochastic block model

While the MMSB is designed for binary-valued networks, it can trivially be extended to integer-valued networks by replacing the Bernoulli distributions with Poisson distributions, and placing gamma priors on the Poisson parameters. While, to the best of our knowledge, this extensions has not yet been explored in the literature, it is a natural, and easily implementable, extension. We perform inference in this model, with K = 50 clusters, using Gibbs sampling, via an existing R package (Chang, 2012) that was modified to replace the beta/Bernoulli pairs with gamma/Poisson pairs. Since inference in this model was significantly slower than inference in the IRM and the DNM, we only compared with the gamma/Poisson MMSB on our smallest data set.

III CONCLUSION AND FUTURE WORK

We have presented a new Bayesian nonparametric model, the mixture of Dirichlet network distributions, for integer-valued networks where the number of nodes is unbounded and grows in expectation with the number of binary links. This model allows us to capture sparse networks with latent structure. Existing network models focus either on latent structure—capturing the fact that each node will have a different pattern over which nodes it connects with—or on capturing sparsity; this is, to our knowledge, the first model that combines these two goals. Further, unlike most existing Bayesian network models, this model is explicitly designed for prediction. We can use the mixture of Dirichlet network distributions to obtain an explicit predictive distribution over the nodes associated with an as-yet unseen observation, even if we have not observed these nodes in our training set; we have shown good predictive and qualitative performance on a variety of data sets.

The mixture of Dirichlet network distributions is based on a simpler network model that we refer to as a Dirichlet network distribution. In the symmetric setting—where a common distribution is used for both senders and receivers—this corresponds to a special case of the integer-valued network models of Caron and Fox (2015) and Crane and Dempsey (2016). While these models can be used to obtain desirable properties such as network sparsity and power law degree distribution, they are unable to capture community-type structure in the network. By using a mixture of these networks, we can capture multiple modalities of interaction between nodes; by using a nonparametric hierarchical framework we ensure that both the number of nodes is unbounded, and that nodes can interact as part of multiple clusters. The MDND therefore increases the modeling flexibility of this class of models, while retaining desirable sparsity properties.

The mixture of Dirichlet network distributions is an exchangeable model: It is invariant to permutations of the order in which we observe links. While this is computationally appealing and leads to a straightforward predictive distribution, it does not allow us to capture network dynamics in integer-valued networks. In practice, such dynamics may be important: an individual’s level of activity within a topic may differ over time, and the overall acceptance of topics may change. A amount of authors have found that adding temporal dynamics to network models improves performance (Ishiguro et al., 2010; Xing et al., 2010; Xu and Hero III, 2013). In the case of Dirichlet network distributions, similar temporal dynamics could be incorporated by replacing some or all of the component Dirichlet processes with dependent Dirichlet processes (MacEachern, 2000; Lin et al., 2010; Ren et al., 2008); we intend to explore this in a future work.

In addition to the base model for integer-valued networks, we also deliberated extensions to binary networks. The methods considered comprise truncating the exchangeable integer valued network; while it is possible to achieve exchangeable binary networks related to those considered by Caron and Fox (2015) and Veitch and Roy (2015), we maintained that a dynamic truncation, yielding a non-exchangeable model, is additional appropriate for a temporally expanding binary network. Unfortunately, inference in such truncated models is trickier than the integer-valued case. While we can analytically model the censored observations as auxiliary variables and recover the integer network, the amount of censored observations grows allowing to a coupon-collector problem with the number of observed links, making this method infeasible for large data sets. An interesting avenue for future research is to develop scalable inference methods for this setting.

REFERENCES

1. E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9:1981–2014, 2008.

2. A.A. Amini, A. Chen, P.J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. The Annals of Statistics, 41(4):2097–2122, 2013.

A.Brix. Generalized gamma measures and shot-noise Cox processes. Advances in Applied Probability, 929–953, 1999.

3. F. Caron and E.B. Fox. Sparse graphs using exchangeable random measures. arXiv:1401.1137 [stat.ME], 2015.

4. J. Chang. lda: Collapsed Gibbs sampling methods for topic models., 2012. URL http: //CRAN.R-project.org/package =lda. R package version 1.3.2.

5. H. Crane and W. Dempsey. Edge exchangeable models for network data. arXiv:1603.04571 [math.ST], 2016.
6. E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states. Technical Report P-2777, Massachusetts Institute of Technology, 2007.
7. Q. Ho, J. Yin, and E.P. Xing. On triangular versus edge representations—towards scalable modeling of networks. In Advances in Neural Information Processing Systems (NIPS), pages 2132–2140, 2012.
8. P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. Social networks, 5(2):109–137, 1983.
9. K. Ishiguro, T. Iwata, N. Ueda, and J.B. Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. In Advances in Neural Information Processing Systems (NIPS), pages 919–927, 2010.
10. S. Jain and R.M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics, 13(1):158– 182, 2004.
11.C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In National Conference on Artificial Intelligence (AAAI), pages 381–388, 2006.
12. B. Klimmt and Y. Yang. Introducing the Enron corpus. In Conference on Email and Anti-Spam (CEAS), 2004.
A. Lijoi, I. Pr¨unster, and S.G. Walker. Investigating nonparametric priors with Gibbs structure. Statistica Sinica, 18(4):1653–1668, 2008.
13. D. Lin, E. Grimson, and J.W. Fisher III. Construction of dependent Dirichlet processes based on Poisson processes. In Advances in Neural Information Processing Systems (NIPS), pages 1396–1404, 2010.
14. J. Lloyd, P. Orbanz, Z. Ghahramani, and D.M. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In Advances in Neural Information Processing Systems (NIPS), pages 1007–1015, 2012.
15. S.N. MacEachern. Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University,