A simple infinite topic mixture for rich graphs and relational data

Janne Sinkkonen
Helsinki University of Technology, and Xtract Oy
janne.sinkkonen@tkk.fi

Juuso Parkkinen
Helsinki University of Technology
juuso.parkkinen@tkk.fi

Janne Aukia
Xtract Oy
Hitsaajankatu 22, 00810 Helsinki, Finland
janne.aukia@xtract.com

Samuel Kaski
Helsinki University of Technology
Department of Information and Computer Science, P.O. Box 5400, FI-02015 TKK, Finland
samuel.kaski@tkk.fi

Abstract

We propose a simple component or "topic" model for relational data, that is, for heterogeneous collections of co-occurrences between categorical variables. Graphs are a special case, as collections of dyadic co-occurrences (edges) over a set of vertices. The model is especially suitable for finding global components from collections of massively heterogeneous data, where encoding all the relations to a more sophisticated model becomes cumbersome, as well as for quick-and-dirty modeling of graphs enriched with, e.g., link properties or nodal attributes. The model is here estimated with collapsed Gibbs sampling, which allows sparse data structures and good memory efficiency for large data sets. Other inference methods should be straightforward to implement. We demonstrate the model with various medium-sized data sets (scientific citation data, MovieLens ratings, protein interactions), with brief comparisons to a full relational model and other approaches.

1 Introduction

Generative models for generic graphs try to find well-connected [11] or similarly-connected subgraphs [17], homogeneously connected blocks of vertices (e.g., [1, 4]), or a latent space explaining the connectivity [3]. But data sets are often richer and better described as *enriched graphs*, where additional properties are associated either to the edges or to the vertices, or both. From another viewpoint, such data sets are *multi-relational* [6, 14], consisting of heterogeneous co-occurrences between categorical variables: In the relational sense, graphs are co-occurrences (edges) within a single categorical variable (vertices), while for example graphs with associated vertex data have an additional co-occurrence type, between vertices and a nominal variable describing attributes of the vertices.

We propose a simple component model for relational data. The component structure is global, as in Latent Dirichlet Allocation (LDA) and other topic models, with no need to specify latent structure except up to what is already apparent in the types of input data. The model is good for very heterogeneous collections of relational data, where one is not able or willing to set up a complicated

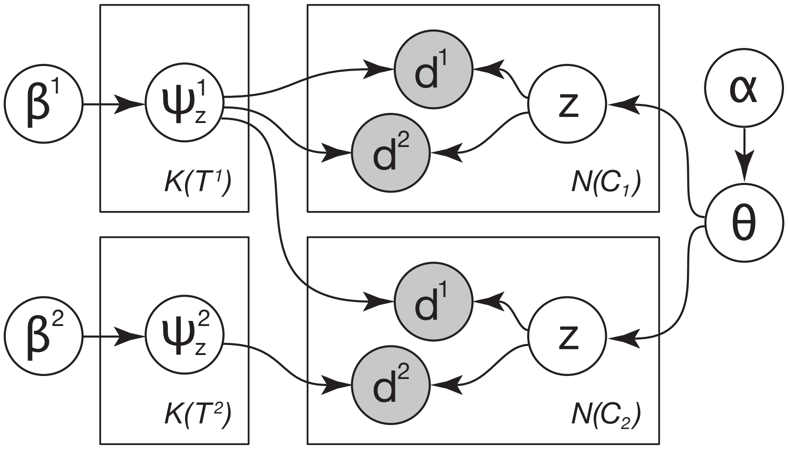

Figure 1: Plate representation for the graph model with nominal vertex data (Section 2.2).

latent structure. It can also be used as a quick-and-dirty model for enriched graphs, as demonstrated in the experimental section.

The model generates relations, that is, co-occurrences, between nominal variables. It finds a global, probabilistic component structure. Within each component, relations are generated from margin probabilities of the nominal variables. All relation types arise from the same process, with the same components.

The model is here introduced with the Dirichlet Process prior for the component probabilities, and estimated by collapsed Gibbs. Although other inference methods might be either faster or produce better results (mix better), the presented approach is both simple and implementable in a sparse form [11] for very large data collections.

2 Global components for heterogeneous co-occurrences

Let the data $\mathcal{D}$ consist of independent *co-occurrences* $\mathcal{D}_i$, $i = 1, \ldots, N$, that can (within a single data set) fall into several co-occurrence classes described by C_i, $i = 1, \ldots, n(C)$. The structure of the co-occurrences is heterogeneous, but fixed within a class C. A co-occurrence of class C_k is a tuple of nominal values, $(d^{(1)}, d^{(2)}, \ldots, d^{(h_k)})$, of size $h_k > 0$. If $h_k = 2$, the co-occurrences are dyadic and presentable as a co-occurrence matrix or a graph. To each variable d we associate a nominal variable type T; the types differ in their domains. Then a tuple $(T_a, T_b, \ldots, T_{h_i})$ of length h_i becomes associated to each C_i. The same variable type T may be shared by several nominal variables d, even within one C.

Note that although the co-occurrences may often be dyadic, the model class includes triplets and higher-order co-occurrences. It also includes independent events, but they are likely to be of less use.

We assume the data are generated from latent components. A latent component z is drawn, from a multinomial with parameters θ, for each co-occurrence $\mathcal{D}_i$. Given the component z for the datum $\mathcal{D}_i$, and its class C, the nominal values $d^{(t)}$ are generated independently from the associated multinomials, having the types T_t. (Figure 3 and Section 4 offer an example with two co-occurrence classes and two nominal variable types.) Denote the parameters of the multinomials by $\psi_z^{(t)}$. Note that all multinomials of type T_t generated by the same component z share the same parameters $\psi_z^{(t)}$; this is the assumption that ties the co-occurrences together.

We have conjugate priors in the model, a Dirichlet or Dirichlet process (DP) prior for the latent components z_i, and Dirichlet priors for $\psi_z^{(t)}$. With the DP prior, the data are generated by

1. $\theta \sim \mathrm{DP}(\alpha)$; $\psi_z^{(t)} \sim \mathrm{Dir}(\beta^{(t)})$, $t = 1, \ldots, n(T)$;
2. For each $i \in 1, \ldots, N$:
    - $z_i \sim \mathrm{Mn}(\theta)$;
    - $d_i^{(j)} \sim \mathrm{Mn}(\psi_{z_i}^{(t_{ij})})$, $j = 1, \ldots, h_{k(C(\mathcal{D}_i))}$;

with the hyper-parameter α controlling the component diversity, and the hyperparameters $\beta^{(t)}$ the evenness of the specific data type distributions. The index t_{ij} simply indicates that each d should be generated from the multinomial (Mn) T to which it is associated via the description of the co-occurrence class $C(\mathcal{D}_i)$. The occurrence of classes C within $\mathcal{D}$ is not modeled—we have a model only for the contents of an occurrence $\mathcal{D}_i$ given its class C_i. That is, the amounts of data of the various types are not modeled either.

All models of the model class can be easily estimated by collapsed Gibbs sampling [7], and the formulas for sampling the latent classes of the various co-occurrence types are simple enough that they can be derived automatically. Such a sampler gives only posterior samples of the latent memberships $\mathcal{Z}_i$ of the co-occurrences; The parameters ψ and θ are marginalized out. The sampler proceeds by removing one co-occurrence from the sampling "urn" at a time, then drawing a new assignment z for the sample, given assignments of other co-occurrences. An example is presented below in Section 2.2.

2.1 Example 1: Model for graph topology

A trivial case of an undirected graph with one object type, $\{C_1 = (T_1, T_1)\}$, is described in [11]. The co-occurrences are edges of an undirected graph, with values of T_1 being vertices.

2.2 Example 2: Model for graphs with nominal vertex data

Another example is a model for two co-occurrence types, $\{C_1 = (T_1, T_1), C_2 = (T_1, T_2)\}$, where $n(C) = 2$. An interpretation is a graph with undirected edges (C_1), and a categorical variable T_2 generating vertex-specific properties (C_2). The corresponding plate model is presented in Figure 1.

The sampling formulas for the two object types are[1]

$$p(z|\mathcal{D}_i) \propto \frac{\{n_z, \alpha\}}{N + \alpha} \times \begin{cases} g_{z,l_1}^{(1)} g_{z,l_2}^{(1)} / (g_{z,\cdot}^{(1)} (g_{z,\cdot}^{(1)} + 1)) & \text{for } \mathcal{D}_i \in C_1 \ , \\ g_{z,l_1}^{(1)} g_{z,l_2}^{(2)} / (g_{z,\cdot}^{(1)} (g_{z,\cdot}^{(2)})) & \text{for } \mathcal{D}_i \in C_2 \ . \end{cases}$$

All counts, g, n, and N, in the sampling formulas are *with the object removed* for which we are drawing the latent component. The total number of objects is denoted by N, while n_z is the number of objects (co-occurrences) associated to the latent component z. The first factor arises from the DP prior, with the case $n_z = 0$ corresponding to a new component, and we define $\{n_z, \alpha\} = \alpha$ for $n_z = 0$, otherwise n_z.

A matrix of counts $g_{z,l}^{(t)}$ exists for each type T_t, counting atomic events d assigned to a latent z. The index l is over the bins of the multinomial $\psi_z^{(t)}$. In the sampling formula associated to a co-occurrence class C_k, the indices $l_1, l_2, \ldots, l_{h_k}$ refer to the atomic events d within that type of co-occurrence. Priors β are included in the counts g as virtual data. The dot notation is used for summation.

In the general case of multiple object types, there is one sampling formula similar to those above for each co-occurrence class, and the structure with the g counters closely follows the structure of the object type.

3 Experiments

First, we demonstrate the ability of the model of Section 2.2 to find meaningful structure from the Cora and Citeseer citation data sets [9]. We compare results obtained with full paper attribute

[1] We have assumed no self-links in C_1, the citation network. If papers were citing themselves, $g_{z,l_2}^{(1)}$ in the numerator of first formula would need to be $g_{z,l_2}^{(1)} + \delta_{l_1,l_2}$.

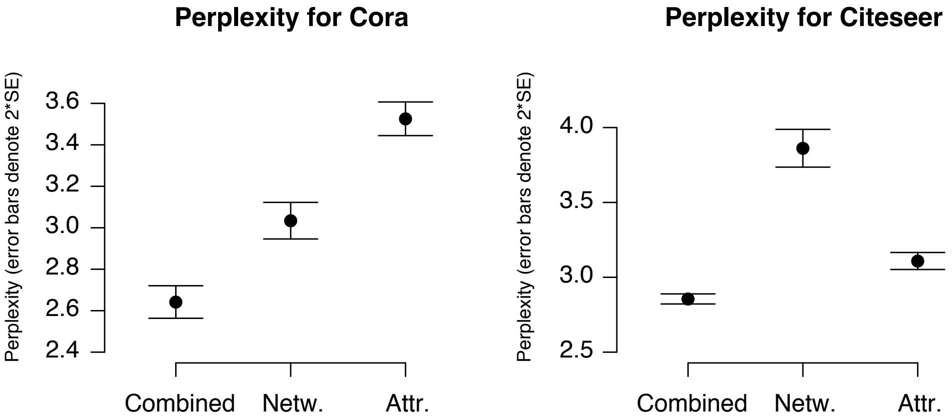

Figure 2: In terms of perplexity, the subject categories of Cora and Citeseer citation sets are best recovered with the model of Section 2.2, which is able to combine citation and content information ("Combined"; lower perplexity is better). The other candidates are the model of Section 3 ("Netw."), and a similar model for the article content only ("Attr."). The 2SE error bars are over ten runs. The models are with Dirichlet priors that work well in cases with a known number of categories. Note that the models are unsupervised—perplexities are not assumed to beat those from supervised models. With a sparse implementation with Java (with some overhead from handling the sparse data structures), the slowest model for Figure 2 ran in 1.25 hours, with a conservative number of iterations (50,000) to assure convergence. The sizes of the Cora and Citeseer sets are 2708 and 3312 vertices, 5429 and 4732 edges, and 1433 and 3703 indicators for the existence of unique words, respectively. At the time of writing this, the data sets, with more detailed descriptions, are available at `http://www.cs.umd.edu/~sen/lbc-proj/LBC.html`.

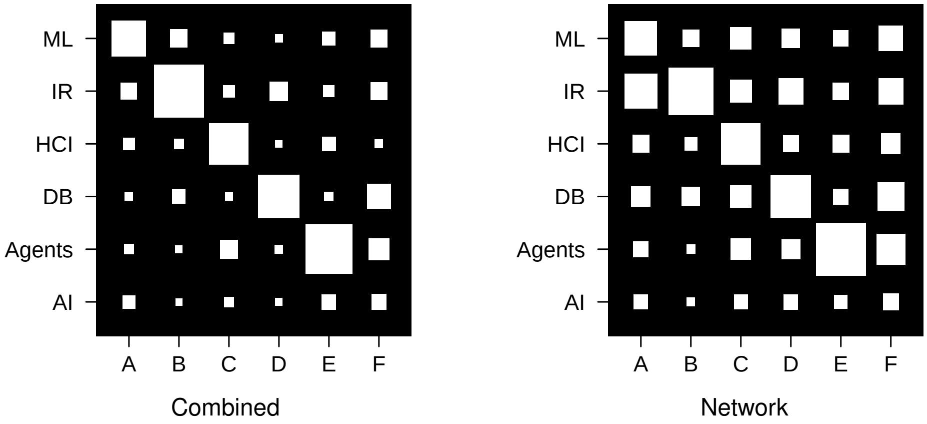

Figure 3: The average confusion matrices between real and computed clusters of Citeseer. *Left:* model for combined citation and content. *Right:* model for the citation information only. The model for combined data recovers the original subject categories except for Artificial Intelligence (AI) that is mixed with Machine Learning (ML) and Agents. Content information is helpful overall, but especially in separating Information Retrieval (IR) from ML.

and citation data to those obtained with simpler models, seeing either the citation graph or vertex attributes only. As expected, the rich data provides the best reconstruction of the original, manual classification of the data sets (Fig. 2).

The second experiment is on collaborative filtering, or predicting ratings in the MovieLens data set [8] on the basis of known ratings of other movies. Our model was compared to the IHRM [14, 15], in a setting similar to that in [15, 16], with either some movie and user attributes included or not.

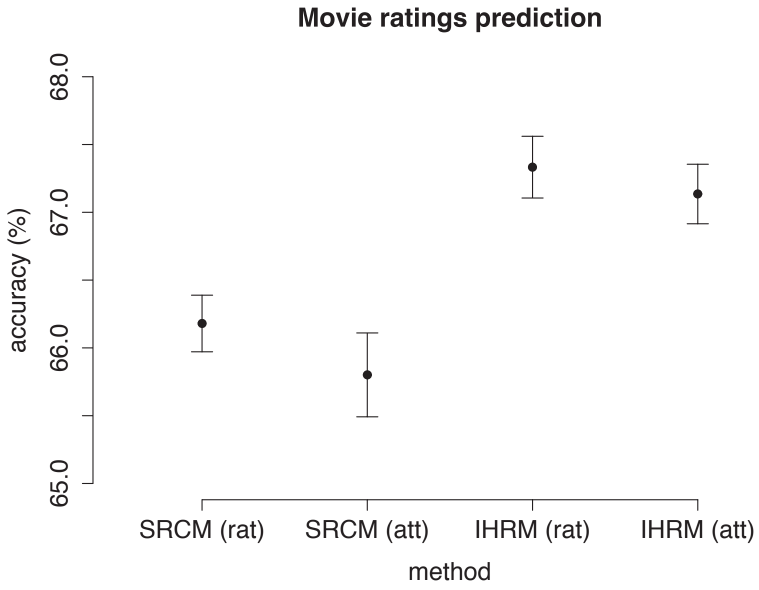

Figure 4: Rating prediction accuracies for our simple relational component model (SRCM) and IHRM of Xu et al. [15], either with data containing ratings only (rat) or with additional movie and user attributes (att). The 2SE error bars are over ten runs. IHRM is better but the less complex SRCM performs well too. Both methods have considerable variance between runs, that is, convergence to a local area of the posterior mass. This experiment does not give evidence of any predictive benefit from adding margin attributes of users and movies to either of the models, but see [15] which reports a benefit from the attributes, for the IHRM. There were 702 users and 603 movies, with about 112 ratings per user (on average). Ratings were binarized, the threshold being user average. For held-out users, 156 of the 702, twenty ratings were used to predict the rest, and the overall average accuracy of these predictions is reported. In the attribute setup, year and genre of the movies, and age, gender, and occupation of the users were added to the model as independent (movie, attribute) or (user, attribute) co-occurrences.

The more complex IHRM performed better, but only slightly[2] (Fig. 4). Both models outperform classic collaborative filtering (by Pearson correlation or SVD; [15]).

In the third experiment, we combined protein interaction data with gene expression to find functional gene modules. A simple relational model, modified to emit normally distributed vertex attributes[3], is able to better recover clusters derived from Gene Ontology [2] than two state-of-the-art methods, HMoF [10] and Matisse [13], that both also integrate the protein and gene expression data sets (Fig. 5).

4 Discussion

We present an infinite topic model for multi-relational data and demonstrate it at work with various collections of data. Although the original motivation for the model is to find communities, or global components, from enriched large social networks, the model is likely to be more widely applicable to relational data. It could be especially suitable for clustering very heterogeneous collections of relations, where defining a sensible latent structure *a priori* over all the variables would be hard.

[2] IHRM inference is with mean field, but difference in performance to Gibbs samplers is supposed to be negligible [15].

[3] The gaussian response, although it fits the continuous data, does make the experiment less relevant for the current paper, but we expect the results to replicate with a binomial or multinomial response.

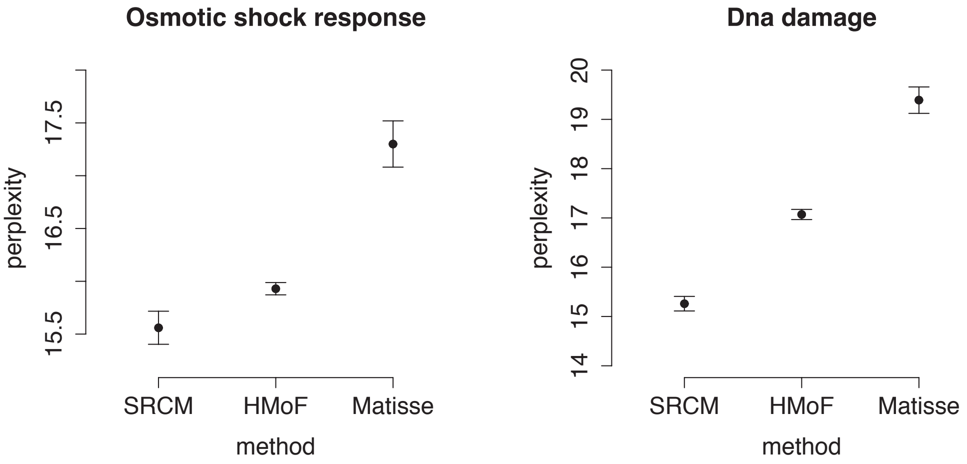

Figure 5: Perplexity of predictions of known functional protein complexes by combined protein interaction and gene expression data, with our simple relational model (SRCM), and with HMoF and Matisse (see text). The 2SE error bars are over 20 runs. The result was replicated for two expression data sets (*left*, *right*). Smaller perplexity is better. The relatively small variation of SRCM results, compared to other experiments in this paper, results from good mixing of the MCMC chains with these smaller data set.

Experiments demonstrated the overall usability with enriched graphs, with two scientific citation data sets. On a collaborative filtering task, the proposed model performed only slightly worse than a state-of-the-art relational model. The lack of evidence of benefit from margin attributes in the predictive task is interesting, especially considering the slight, $\sim 1\%$ benefit reported by Xu et al. [15]. This could be a statistical anomaly, given the large variation between runs. The tests should also be replicated with more sparse rating data, including users now omitted. If the effect is real for the proposed model, or for both models, behavior of users is then much more informative than static margin traits, or a more discriminative modeling approach is needed to fully benefit from the more 'distant' relational data in the model.

The good side of the proposed inference method is its technical scalability. If the counters g of the collapsed Gibbs sampler are represented sparsely, the proposed inference method is highly scalable with respect to the data set size and the number of components. The number of co-occurrence types can be very high, even on the order of the data set size.

But collapsed Gibbs is known to mix poorly, especially for larger data sets, and this is apparent also in the large deviations between single runs of our experiments. Mixing could be improved for example by the stick-break sampler [5, 15], or by annealing. On the other hand, inference could be made faster and deterministic by adapting a variational technique, such as the collapsed mean field [12]. In pilot experiments, the model performance was sensitive to the order of magnitude of the hyper-parameter values, at least that of β, in a data-set-specific way. Sampling of hyper-parameters may help and is relatively easy to implement.

Acknowledgments

This work was supported by Academy of Finland, grant number 119342. SK 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 PASCAL. We thank Zhao Xu (Fraunhofer IAIS), Achim Rettinger (Technical University of Munich) and Volker Tresp (Siemens Corporate Technology) for providing us an IHRM implementation, and for help with its use.

References

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

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, and H. Butler. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25:25–29, 2000.

[3] Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal Of The Royal Statistical Society Series A*, 170(2):301–354, 2007.

[4] Jake M. Hofman and Chris H. Wiggins. A Bayesian approach to network modularity. *Physical Review Letters*, 100(25):258701–259900, 2008.

[5] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, pages 161–173, 2001.

[6] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In Y. Gil and R Mooney, editors, *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, Menlo Park, USA, 2006. AAAI Press.

[7] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[8] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Analysis of recommender algorithms for e-commerce. In *Proc. ACM E-Commerce Conference*, pages 158–167. ACM, 2000.

[9] Prithviraj Sen and Lise Getoor. Link-based classification. Technical Report CS-TR-4858, University of Maryland, College Park, USA, 2007.

[10] M. Shiga, I. Takigawa, and H. Mamitsuka. Annotating gene function by combining expression data with a modular gene network. *Bioinformatics*, 23:i468–i478, 2007.

[11] Janne Sinkkonen, Janne Aukia, and Samuel Kaski. Component models for large networks. *ArXiv e-prints*, 2008. arXiv:0803.1628.

[12] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for Latent Dirichlet Allocation. NIPS 19, 2006.

[13] Igor Ulitsky and Ron Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1:8, 2007.

[14] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Infinite hidden relational models. In *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, 2006.

[15] Zhao Xu, Volker Tresp, Shipeng Yu, and Kai Yu. Nonparametric relational learning for social network analysis. In *The 2nd SNA-KDD Workshop (SNA-KDD) '08*, 2008. Las Vegas, Nevada, USA.

[16] Zhao Xu, Volker Tresp, Shipeng Yu, Kai Yu, and Hans-Peter Kriegel. Fast inference in infinite hidden relational models. In *Working Notes of the 5th International Workshop on Mining and Learning with Graphs (MLG'07)*, Florence, Italy, 2007. Universita degli Studi di Firenze. Extended Abstract.

[17] Haizheng Zhang, Baojun Qiu, C. Lee Giles, Henry C. Foley, and John Yen. An LDA-based community structure discovery approach for large-scale social networks. In *Intelligence and Security Informatics (ISI) 2007*, pages 200–207. IEEE, 2007.