Detecting Hierarchical Structure in Networks

Tue Herlau, Morten Mørup, Mikkel N. Schmidt and Lars Kai Hansen
Technical University of Denmark, 2800 Lyngby, Denmark
Email: [tuhe@imm.dtu.dk, mm@imm.dtu.dk, mns@imm.dtu.dk and lkh@imm.dtu.dk]

***Abstract*—Many real-world networks exhibit hierarchical organization. Previous models of hierarchies within relational data has focused on binary trees; however, for many networks it is unknown whether there is hierarchical structure, and if there is, a binary tree might not account well for it. We propose a generative Bayesian model that is able to infer whether hierarchies are present or not from a hypothesis space encompassing all types of hierarchical tree structures. For efficient inference we propose a collapsed Gibbs sampling procedure that jointly infers a partition and its hierarchical structure. On synthetic and real data we demonstrate that our model can detect hierarchical structure leading to better link-prediction than competing models. Our model can be used to detect if a network exhibits hierarchical structure, thereby leading to a better comprehension and statistical account the network.**

I. INTRODUCTION

Graphs of relationships between entities are of interest in practically all fields of science. For example, in sociology graphs are used to represent interaction between people, in economy to represent trade relations and business structures, and in biology to describe interaction between proteins, organisms, and neurons.

Statistical network analysis seeks to discover an underlying structure amongst the vertices in these network. A type of structure of particular interest is that of a hierarchy, where vertices are divided into groups and further subdivided into subgroups and so forth. real-world networks exhibit hierarchical organization, where the groups correspond to known functional units, such as ecological niches in food webs, modules in biochemical networks or communities in social networks[1], [2], [3], [4], [5] and within the cognitive sciences hierarchies have long been proposed as a useful organization of knowledge [6], [7]. Simply put, a hierarchy allows us to easily get a coarse top-level understanding of the data which can then be refined, thus making it easier to understand what the lower levels of the hierarchy mean, contrasted to other structured alternatives such as the mondrian process[8] where the inferred structure is not easily interpretable.

While hierarchical structure in feature data has been a longstanding focus of research (see [9], [10], [11] and references therein), the importance of hierarchical structure in networks have been acknowledged, both using more traditional non-statistical agglomorative techniques (cf. [12]), and more recently in [13], [14], [15]. These models all consider representing the hierarchical structure by a binary tree, and can be considered as models that impose hierarchically structured homogeneities in the stochastic blockmodel [16] and its non-parametric extension, the infinite relational model (IRM) [17], [18].

While existing models of hierarchies in networks have been based on binary trees it has been acknowledged in the literature on hierarchies of feature data that a binary tree structure often is too restricted, potentially producing spurious structure if the binary tree is not well supported by the data [10]. We propose a non-parametric Bayesian generative model that can infer any hierarchical structure in networks while reducing to the IRM model if the data does not support hierarchy.

A. Motivation

As a motivating example, consider a friendship network of high school students. Let us say that there are three high schools in a city (A, B, and C), each with four grades (freshman, sophomore, junior, and senior), see fig.1. Assuming that the structure in the friendship network can be well characterized by intractions between the 3 schools × 4 grades = 12 groups of students, how should these interactions best be modeled?

In the the stochastic block model, the relations are modeled by a parameter for each of the $12 + 12 \times (12 - 1)/2 = 78$ combinations of groups, i.e., there is a parameter that defines how likely it is that a freshman in high school A is friends with another freshman in high school A, how likely a junior in B is friends with a sophomore in C etc. While certainly able to fit the data, this model is unable to capture any hierarchical structure, and the inferred parameters are less well-defined.

To alleviate this problem, several authors have proposed to model hierarchical structure in relational data using a binary tree. In these models, the probability that two students are friends is governed by a parameter located at the lowest common ancestral node in the tree. As any binary tree with 12 leaf nodes has 11 internal nodes, the high school friendship graph would be governed by only $12 + 11 = 23$ parameters. We will argue that this leads to a too restricted model. For example, if at the root of the tree we find a division between high school A on one side and high schools B and C on the other, the binary tree model asserts that students from A have an equal probability of being friends with students from either B or C.

We propose a model for learning the structure of relational data, that can represent a hierarchy as a multifurcating tree, and contains the stochastic blockmodel and models based on binary trees as special cases. Each node in the tree is allowed to have two or more child nodes, and each node contains parameters for the probabilities of links between all combinations of the child nodes. If the tree first splits into three

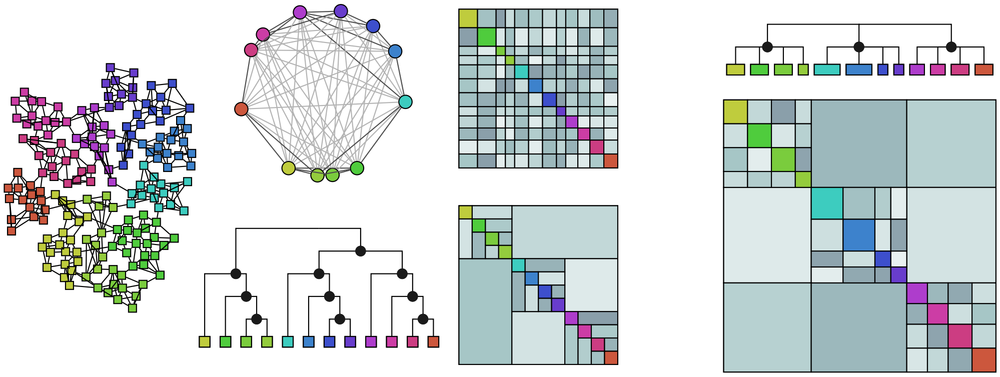


Fig. 1: Illustration of the proposed hierarchical modeling of graphs for the three high schools example. While the IRM model (top middle panel) does not explicitly impose a hierarchical structure in the data the binary tree subdivides the graph into homogeneous regions such that each split defines the interaction between all vertices of the right and left hand side in the subsequent levels of the tree (bottom middle panel). The proposed model is not restricted to binary trees. Therefore, each node in the tree can split the vertices into an arbitrary number of regions that interact homogeneously. This model naturally reduces to the IRM model if the root of the tree splits into all the inferred clusters.

nodes corresponding to schools and next for each school splits into four nodes corresponding to grades, the model would describe the friendship structure using a total of 33 probability parameters.

II. METHOD

The problem that we address in this paper is the following: Given a graph, expressing relations between entities, we wish to infer the hierarchical structure that best accounts for the data. This includes the possibility of inferring that there is no hierarchical structure, a binary tree structure, or a multifurcating tree structure. Our model simultaneously learns the number of clusters as well as the type of hierarchical structure governing their relations.

To avoid confusion, we will allways use either *vertices* or *entities* to denote the objects in the network we try to model, while *nodes* will be reserved for the nodes in the multifurcating tree, eg. the leaf nodes correspond to communities.
In the high-school example. the vertices correspond to children, leaf nodes to school classes, the root node to the entire collection of children and the nodes at intermediate level to schools.

The following outline of a generative process can be used to characterize relational data with a hierarchical cluster structure.

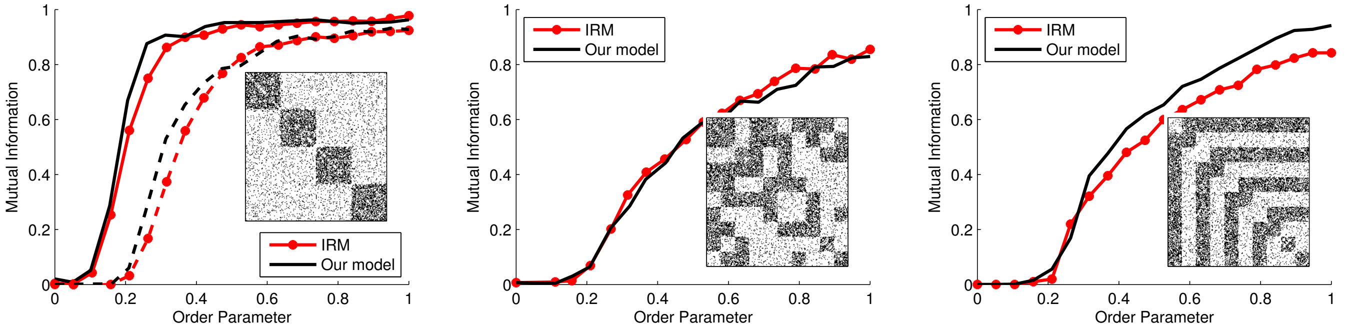
1) Generate a tree and a partitioning of the vertices consistent with the tree.
2) For each node in the tree, generate parameters that describe the probabilities of links between each of its children.
3) For each pair of vertices in the graph, generate a link with the probability given at the lowest common ancestral node in the tree.

Several existing generative network models can be seen as implementing this framework with different choices of priors over partitions, trees, and link probabilities. For example, in [17] the prior over partitions is given by a Chinese restaurant process, and there is no hierarchical structure corresponding to a tree where the root immediately splits into N leaf nodes corresponding to the clusters. In [14], [19], there is no partitioning of the nodes, and the hierarchical structure is a binary tree terminating at individual nodes in the graph. In [13], a joint distribution over binary trees and partitions consistent with the tree is formulated. In [15] the tree is a fixed binary tree, and vertices of the graph are assigned to nodes of the tree, inducing a partition.

A. Generative process

Based on the framework above, we formulate a generative model of hierarchical network structure, that contains the models of Clauset et al. [14], [19] and Kemp et al. [17] as special cases. As in IRM, we choose a Chinese restaurant process prior over vertice partitions, and generate link probabilities i.i.d from a Beta distribution. As a prior over trees, we choose a uniform prior over rooted multifurcating trees. Our model can be described by the following generative process:

1a. Generate a random partitioning of the I vertices into K clusters using a Chinese restaurant process, $\boldsymbol{z} \sim \mathrm{CRP}(\alpha)$.
1b. Generate a tree T with K leaf nodes from a uniform distribution over all rooted multifurcating trees, and assign each cluster to a leaf.
2. Generate a link-probability for each unordered (non-distinct) pair of children (k, l) of each node n, $\theta_{kl}^{(n)} \sim \mathrm{Beta}(\beta, \beta)$.
3. For each pair of vertices (i, j) generate a link $A_{ij} \sim \mathrm{Bernoulli}\big(\theta^{n(i,j)}_{c(i,j),c(j,i)}\big)$, where $n(i,j)$ is the nearest common ancestral node of i and j and $c(i,j)$ and $c(j,i)$ are the children of $n(i,j)$ to which the two nodes are connected.

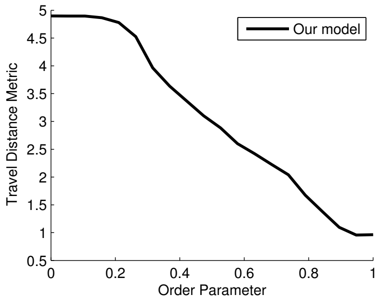


(a) Network supporting hierarchical structure, solid/dashed, $I = 256, 128$

(b) Network with no hierarchical structure

(c) Network with binary cascading tree structure

Fig. 2: Performance of our model and the IRM on three types of graphs (see inserts).



(d) Distance between inferred and true binary cascading tree

B. Inference

According to the generative model, the joint distribution of data and parameters can be expressed as

$$p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\theta}, T|\alpha, \beta) = \\ p(\boldsymbol{A}|\boldsymbol{z}, \boldsymbol{\theta}, T)p(\boldsymbol{\theta}|\beta, T)p(T|\boldsymbol{z})p(\boldsymbol{z}|\alpha). \tag{1}$$

Analytically integrating away the link probabilities, $\boldsymbol{\theta}$, we arrive at

$$\begin{aligned} p(\boldsymbol{A}, \boldsymbol{z}, T|\alpha, \beta) &= \\ &\int p(\boldsymbol{A}|\boldsymbol{z}, \boldsymbol{\theta}, T)p(\boldsymbol{\theta}|\beta, T)p(T|\boldsymbol{z})p(\boldsymbol{z}|\alpha)\mathrm{d}\boldsymbol{\theta} \\ = &\prod_{\substack{n\in\text{nodes}\\(k,l)\in\text{children(n)}}} \frac{\text{Beta}(N^{+}_{n,k,l}+\beta, N^{-}_{n,k,l}+\beta)}{\text{Beta}(\beta,\beta)} \\ &\times \left(\frac{1}{C_K}\right)\left(\frac{\alpha^K\Gamma(\alpha)}{\Gamma(I+\alpha)}\prod_{k=1}^{K}\Gamma(M_k)\right) \end{aligned} \tag{2}$$

where I is the number of vertices, C_K is the number of rooted multifurcating trees with K leaf nodes which can be computed by a simple recursive formula [20], and M_k is the number of vertices in the kth cluster. $N^{+}_{n,k,l}$ and $N^{-}_{n,k,l}$ denote the number of links and nonlinks between vertices belonging to child k and child l of node n.

We use Gibbs sampling to jointly sample from the posterior distribution over partitions and trees. In each Gibbs move, we sample a new placement for a vertex, either in an existing cluster or as a new cluster placed in any possible position in the tree. A vertex can either be a child of any of the $|T|$ nodes in the graph (including leaf-nodes) or split one of the $|T|-1$ edges in the tree. Thus, the Gibbs sampler requires the evaluation of $2|T|-1$ posterior distributions. Although this Gibbs sampler is sufficient to sample from the joint posterior distribution over partitions and trees, we also consider a less incremental move, that conditions on the current partition $\boldsymbol{z}$ and modifies only the tree. We choose a node in the tree uniformly at random, and compute the probability of inserting the node and its subtree at all possible edges in the remaining tree, and sample a modified tree from this distribution.

C. Performance measures

As measure of performance we will use the *normalized mutual information* (NMI) between a planted and estimated community structure defined as

$$\text{NMI}(\boldsymbol{z}, \widetilde{\boldsymbol{z}}) = \frac{2I(\boldsymbol{z}, \widetilde{\boldsymbol{z}})}{H(\boldsymbol{z}) + H(\widetilde{\boldsymbol{z}})}, \text{ where}$$

$$I(\boldsymbol{z}, \widetilde{\boldsymbol{z}}) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

where $H(\boldsymbol{z}) = I(\boldsymbol{z}, \boldsymbol{z})$ is the entropy. Notice It is trivial to see that if the two variables are independent $p(x,y) = p(x)p(y)$, the normalized mutual information is zero and if they are identical $p(x,y) = \delta(x-y)p(x)$ then the normalized mutual information becomes 1. NMI is a commonly used metric for partitions of data sets [21], [22], and has been shown to be an efficient measure of partitions in the type of problem considered here [23]. It has become a standard in assessing the quality of communities in artificial relational data [24], [25], [26]. The joint distribution $p(x,y)$ is simply the probability a random observation which is in community x in the actual community structure is in community y in the detected, $p(x,y)$ will therefore in general form a rectangular matrix since the number of communities need not be the same. The measure

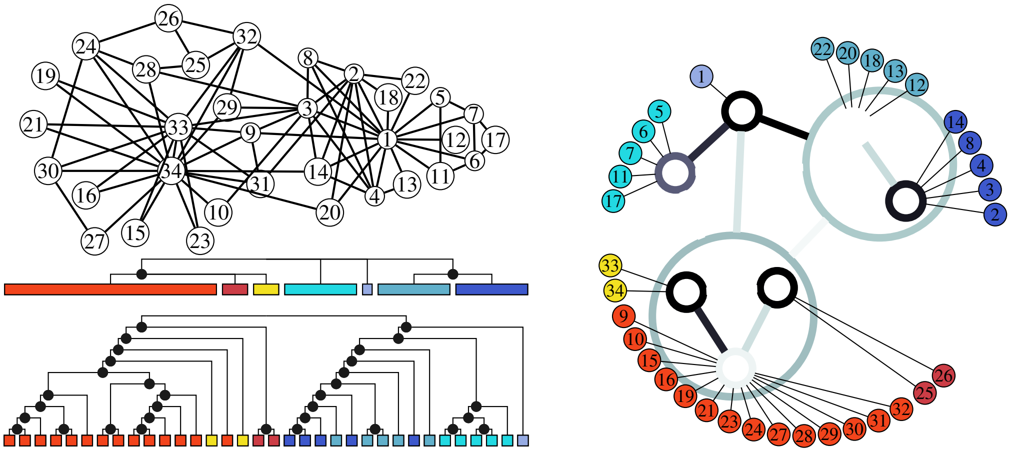


Fig. 3: The karate-club network. Below is both our discovered clustering, and the one reported in [14], indicating that our model gives a simpler representation. The right panel illustrates the link-density of our clustering. The non-numbered circles represent nodes in the tree and the colors indicate the link-density within (circles) or between (edges) clusters.

can be interpreted as *"What fraction of the total amount of information in the communities can be learned from knowing one community"* [23]. An important aspect in the analysis of tree structured data is to determine if the algorithm can properly detect the underlying tree. To evaluate this we use a distance measure for rooted trees based on the travel-distance between vertices [27]: For any two communities μ, ν, define $D(\mu, \nu)$ as the shortest path between the two nodes in the tree, and for any tree T define the $I \times I$ matrix $M^T_{ij} = D(z_i, z_j)$. For two trees a metric $D(T_1, T_2)$ is defined as the average distance between vertices in the tree given by the sum of the upper-triangular part of the absolute distance between the matrices M^{T_1} and M^{T_2} divided by $\frac{1}{2}I(I-1)$. $D(T_1, T_2)$ is zero provided the two trees are equal. It is important to emphasize that this measure has defects. For instance, a singleton cluster inserted "in the middle" of a tree will have a disproportional large effect, as will a local IRM-like structure in the tree. These defects are however inherent to travel-distance type metrics [28], [29], [30].

III. RESULTS

A. Synthetic relational data

We investigate whether our model is able to infer the presence of hierarchical structure on synthetic data by comparing our model to the performance of the IRM model that does not assume hierarchical structure. In the generated graphs we will consider a block-lige structure of the link-probabilities defined either by the density $\rho_0 = 0.1$ or the density $\rho_1 = 0.8$. We will vary the noise-level by introducing the *order parameter* $\lambda \in [0, 1]$ setting $\rho_1' = \rho_0(1-\lambda) + \rho_1\lambda$. When $\lambda = 0$ detection of the planted structure is impossible while for $\lambda = 1$ detection is *easiest*.

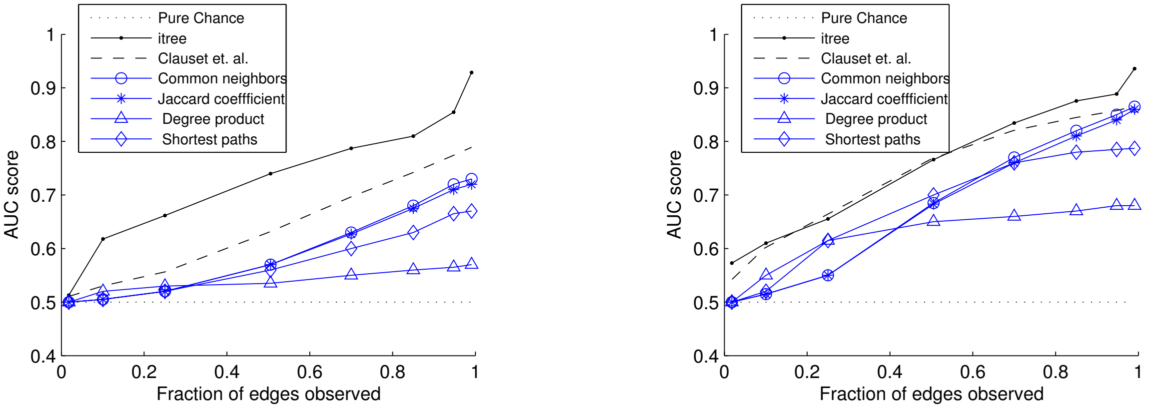
Initially we consider the diagonal community model in the top left panel of figure 2. We let the order parameter λ defined above vary equidistantly between 0 and 1 in 20 steps, and for each such value of λ we generate 10 random graphs with the corresponding ρ_0 and ρ_1' values and for each graph restart the simulation 10 times. For all $10 \times 10 \times 20$ instances, the simulation is run with a burnin time of 50 iterations and simulation time of $T = 50$ iterations, and for each configuration the NMI is extracted. Since an arbitrary hierarchical structure will use fewer parameters than the IRM we expect if our model can identify hierarchical structure at all it should outperform the IRM model. For the top-left panel we choose, in line with what is common in the litterature [26] to let $K = 4$ and vary I between $128, 256$ vertices, re-starting the simulation at each instance. The other plots are for $K = 10$ and $I = 200$ to allow for a non-trivial hierarchical structure.

The parameters for the beta-prior was choosen in line with [17] as $(5, 1)$ within communities and $(1, 5)$ between. The total computation time was about 12 hours, the graph tree model being about a factor 8 slower than the IRM model. Indeed we see that our model outperform the IRM model and as the system-size becomes larger and the noise level decrease, the cluster-structure both become easier to detect and the transition from no detection to detection of structure becomes steeper, approaching the idealized phase-transition described in [26].

In the bottom left panel of figure 2 we tested the behavior of our model on a graph with a hierarchical structure corresponding to a binary cascading tree. We set $n = 200$ and $K = 10$ and ran the simulations with the same parameters as before. Again we see that our model is better at detecting the planted structure than the IRM model and as the SNR of the graph improves we find that our model is close to recovering the true binary tree. The distance metric converges to one rather than decrease all the way to zero since the small clusters of the binary tree can be arbitrarily defined in our model.

We finally compared the performance of our model and IRM model on a graph that is non-hierarchical (right panel of figure 2). We see that our model performs as well as the IRM model, thus is able to reduce to the IRM model when no hierarchical structure is present in the data.

The synthetic data analysis indeed supports that our model



(a) Grassland species (b) Terrorist network

Fig. 4: Comparison of link prediction methods. Averaged AUC statistics as a function of fraction of observed connections known to the algorithm.

is able to detect hierarchical structure when a hierarchy is present and is on par with IRM in detecting the underlying cluster structure when no hierarchical structure exists.

B. *Zachary's Karate Club network*

The first real network we consider is a well-studied friendship network [31] gathered between 1970 and 1972 by Wayne Zachary of friendship between members of a karate club. The friendship network is constructed by noting which of the members of the club interacted regularly at activities outside the club. This included such things as associating at academic classes at the local university, at a local bar, open karate tournaments, at the universities rathskeller etc. 34 members of the karate club which regularly participated in some of the above events were selected (corresponding to vertices in the network) and an edge was added between them if they interacted in at least one of these contexts. This resulted in a binary network with 34 vertices. In the network Mr. Hi corresponds to vertex 1 and Mr. John[1] to vertex 34. Near the beginning of the study, the clubs principal instructor, Mr. Hi, wanted to raise fees, while the clubs administrator, Mr. John, wanted to maintain them at the current level. As time passed the club became more and more divided over the issue, and while never named or explicitly recognized by their members, two factions began to form. The conflict culminated with Mr. Hi being laid off. The supporters of Mr. Hi responded by resigning and forming a new club headed by Mr. Hi.

In figure 3 is given an analysis by our model as well as the binary tree model of [14], [19] of the karate club network. As can be seen from the figure our model has identified a tree with depth 2 whereas the binary network of [19] has extracted a tree which exhibit cascading behavior forming a hallucinated structure due to the models restriction to binary trees that terminate at the single vertice.

Inspecting the results of our model in the right panel of the figure, the bubble at the 11 O'clock position contains only Mr. Hi (1). We see that he is strongly connected to his faction formed by the two other bubbles at the 10 and 2 O'clock position which in turn has almost nothing to do with each other. Moving on to the group who broke out at the 7 O'clock position, Mr. John (34) and the trainer who worked closely with him (33) is identified as a cluster which interacts strongly with the majority of the people in the group at the bottom-most cluster indicating they play a key role.

C. *Networks of Terrorists and Grassland Species*

We compared our proposed framework to the binary tree of [14] on the grassland species and terrorist network data where the method of [14] was the best performing model for link prediction. We consider the same link-prediction setup as in [14] where a given fraction of the links are removed from the network and treated as non-links during model estimation. Notice that this setup is different from the traditional link prediction approach in networks where links and non-links are treated as missing, forming a test set [32]. The area under curve (AUC) of the receiver operator characteristic is used to evaluate how well the model discriminate between "true" links treated as non-links relative to all the "true" non-links in the network which makes the setup comparable to the link scoring approaches in [33].

In figure 4 the AUC score of our proposed model and the binary tree model of [14] are given. From the graphs it can be seen that our model is on par with Clauset et al.'s model for the terrorist data, while our model outperforms Clauset et al.'s model for the grassland network.

In figure 5 is given the estimated MAP hierarchical structures identified by our model. In the terrorist network, six out of seven nodes in the tree defines a binary split supporting the similar link prediction results of our model to that of Clauset et al. For the grass-land network, only two out of five nodes define binary splits. We attribute the superior link predictive performance of our model to its ability to model more complex hierarchical structure than that formed by restricting the modeling to binary trees.

---

[1] The names are made up and adopted from the original study

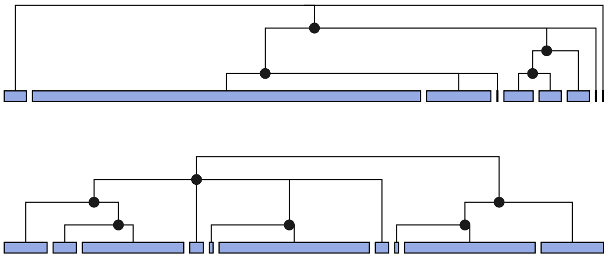


Fig. 5: Top: MAP estimate of tree structure in grassland network. Bottom: Same for terrorist network.

In addition to the above results, the hierarchical model was evaluated on a number of large ($\sim 2-5000$ vertices) networks[2], indicating the method scale well. A prominent feature of the detected hierarchies are the surprising degree of binary splits, indicating the IRM model tend to introduce to many parameters than neccesary to describe the network, but results are omitted due to space limitations.

IV. CONCLUSION

We proposed a model for detecting hierarchical structure in networks, that can interpolate between arbitrary hierarchical structures. On synthetic data we demonstrated that the model was indeed able to infer hierarchies while performing on par with the non-hierarchical IRM model when no hierarchical structure was planted in synthetic data. On the Zachary network, our model detected an interpretable hierarchical structure in compliance with ground truth. We further found that the model was on par with [14] when the hierarchical structure in the data could well be represented by a binary tree whereas our model outperformed [14] for the grass-land species network.

In many real world networks, the presence or absence of hierarchical structure is not known. We have formulated a generative model that includes flat hierarchies, binary trees, and everything in between in its hypothesis space. Our model can thereby be used to investigate if hierarchical structure is present in networks, and in turn lead to a better comprehension and account of network structure in general.

REFERENCES

[1] M. Cosentino Lagomarsino, P. Jona, B. Bassetti, and H. Isambert, "Hierarchy and feedback in the evolution of the escherichia coli transcription network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 13, p. 5516, 2007.

[2] R. Guimera and L. Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, no. 7028, pp. 895–900, 2005.

[3] R. Guimera, S. Mossa, A. Turtschi, and L. Amaral, "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 22, p. 7794, 2005.

[4] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, p. 066111, 2004.

[5] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, p. 1551, 2002.

[6] F. Keil, *Semantic and conceptual development: An ontological perspective*. Harvard University Press Cambridge, MA, 1979.

[7] A. Collins and M. Quillian, "Retrieval time from semantic memory1," *Journal of verbal learning and verbal behavior*, vol. 8, no. 2, pp. 240–247, 1969.

[8] Teh, Yee Whye and Roy, Daniel, "The Mondrian Process," *Technology*, vol. 21, no. 1979, pp. 1–8, 2009.

[9] K. A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering," in *In Twenty-second International Conference on Machine Learning*, 2005, pp. 297–304.

[10] Y. W. Teh, H. D. Iii, and D. Roy, "Bayesian agglomerative clustering with coalescents," in *In Advances in Neural Information Processing Systems*, 2008.

[11] C. Blundell, Y. W. Teh, and K. A. Heller, "Discovering non-binary hierarchical structures with Bayesian rose trees," in *Mixture Estimation and Applications*, K. Mengersen, C. P. Robert, and M. Titterington, Eds. John Wiley & Sons, 2011.

[12] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[13] D. Roy, C. Kemp, V. Mansinghka, and J. Tenenbaum, "Learning annotated hierarchies from relational data," *Advances in neural information processing systems*, vol. 19, p. 1185, 2007.

[14] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[15] Y. Park, C. Moore, and J. S. Bader, "Dynamic networks from hierarchical bayesian graph clustering," *PLoS ONE*, vol. 5, no. 1, 2010.

[16] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1077–1087, 2001.

[17] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *National AAAI Conference*, 2006.

[18] Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel, "Learning infinite hidden relational models," *Uncertainity in Artificial Intelligence*, 2006.

[19] Clauset, Aaron and Moore, Cristopher and Newman, Mark, "Structural Inference of Hierarchies in Networks," in *Statistical Network Analysis: Models, Issues, and New Directions*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007, vol. 4503, pp. 1–13.

[20] J. Felsenstein, "The Number of Evolutionary Trees," *Systematic Biology*, vol. 27, no. 1, pp. 27–33, 1978.

[21] A. Fred and A. Jain, "Robust data clustering," 2003.

[22] L. Kuncheva and S. Hadjitodorov, "Using diversity in cluster ensembles," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 2. IEEE, 2004, pp. 1214–1219.

[23] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, p. P09008, 2005.

[24] A. Lancichinetti, "Community detection algorithms: a comparative analysis," *Physical Review E*, vol. 80, no. 5, p. 056117, 2009.

[25] G. Orman and V. Labatut, "A comparison of community detection algorithms on artificial networks," in *Discovery Science*. Springer, 2009, pp. 242–256.

[26] J. Reichardt and J. Reichardt, *Structure in complex networks*. Springer Verlag, 2009.

[27] H. Bunke, "Graph matching: Theoretical foundations, algorithms, and applications," in *Proc. Vision Interface*, 2000, pp. 82–88.

[28] P. Bille, "A survey on tree edit distance and related problems," *Theoretical computer science*, vol. 337, no. 1-3, pp. 217–239, 2005.

[29] K. Zhang, "A new editing based distance between unordered labeled trees," in *Combinatorial Pattern Matching*. Springer, 1993.

[30] K. Zhang, J. Wang, and D. Shasha, "On the editing distance between undirected acyclic graphs and related problems," in *Combinatorial Pattern Matching*. Springer, 1995, pp. 395–407.

[31] W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[32] K. T. Miller, T. L. Griffiths, and M. I. Jordan, "Nonparametric latent feature models for link prediction," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1276–1284.

[33] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

---

[2] See `http://www.cise.ufl.edu/research/sparse/mat/`