Learning latent structure in complex networks

Morten Mørup and Lars Kai Hansen
e-mail: {mm,lkh}@imm.dtu.dk
Informatics and Mathematical Modeling, Technical University of Denmark

Abstract

Latent structure in complex networks, e.g., in the form of community structure, can help understand network dynamics, identify heterogeneities in network properties, and predict ‘missing’ links. While most community detection algorithms are based on optimizing heuristic clustering objectives such as the Modularity, it has recently been shown that latent structure in complex networks is learnable by Bayesian generative link distribution models (Airoldi et al., 2008, Hofman and Wiggins, 2008). In this paper we propose a new generative model that allows representation of latent community structure as in the previous Bayesian approaches and in addition allows learning of node specific link properties similar to that in the modularity objective. We employ a new relaxation method for efficient inference in these generative models that allows us to learn the behavior of very large networks. We compare the link prediction performance of the learning based approaches and other widely used link prediction approaches in 14 networks ranging from medium size to large networks with more than a million nodes. While link prediction is typically well above chance for all networks, we find that the learning based mixed membership stochastic block model of Airoldi et al., performs well and often best in our experiments. The added complexity of the LD model improves link predictions for four of the 14 networks.

1 Introduction

A community is traditionally defined as a densely connected subset of nodes that is sparsely linked to the remaining network. Latent structure in complex networks, e.g., in the form of community structure, can help understand network dynamics, identify link density heterogeneities, and predict ‘missing’ links, therefore a large number of algorithms have been proposed. Most community detection algorithms are based on heuristic clustering objectives such as Hamiltonian and Modularity optimization [3, 10, 9, 8]. The Hamiltonian optimization problem is typically formulated as minimizing $-\operatorname{trace}[\boldsymbol{S}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{S}^\top]$ where $\boldsymbol{A}$ is the $n \times n$ network’s adjacency matrix, $\boldsymbol{B}$ is a background ‘null hypothesis’ matrix of the same size as $\boldsymbol{A}$ and $\boldsymbol{S}$ a $c \times n$ clustering assignment matrix. For Modularity optimization $\boldsymbol{B}_{i,j} = \frac{\boldsymbol{E}_{i,j}}{|\boldsymbol{A}|}k_i k_j$ where $\boldsymbol{E}_{i,j} = 1 - \delta_{i,j}$ and $k_i = \sum_j \boldsymbol{A}_{i,j}$ such that communities are defined by regions of link densities above that expected based on the nodes degree distribution [9, 8]. Other common choices for $\boldsymbol{B}$ include $\boldsymbol{B} = \frac{m}{n^2}\boldsymbol{E}$ where $m = |\boldsymbol{A}|$, i.e. communities are defined by link densities above that expected on average [3, 10]. A drawback of Modularity and Hamiltonian optimization is that they are based on heuristic null hypotheses that are not adapted to the actual network. Recently, generative models for complex networks have been proposed in [5] (here referred to as the H&W model) and in [1] (the mixed membership stochastic block model MMSB). Both approaches are based on links distributed according to independent Bernoulli draws where the probability of a given link between node i and j is conditioned on latent variables representing community structure. The two existing generative models do not take into account node specific properties such as the degree in the clustering process, as is key to some of the heuritics methods. While H&W assumes a within group link probability ρ_c and between group link probability ρ_n the MMSB model assumes specific between group probabilities parameterized by a between group link probability matrix $\boldsymbol{P}$ of size $c \times c$ where c is the number of communities. Both approaches are based on Bayesian inference. Common to the optimization of the Modularity, Hamiltonian, H&W model and MMSB model are that they can be expressed in terms of what we define as the generalized Hamiltonian for graph clustering (GHGC). Thus, the computational bottleneck for all the models is the optimization of GHGC with respect to the clustering assignments having a computational cost of $\mathcal{O}(cm)$. In this paper we demonstrate that the hard assignments optimization problem of the GHGC can be relaxed to efficient standard continuous optimization over the simplex such that hard assignments are recovered at stationarity. Within the GHGC framework we propose a generalization of the MMSB model to take into account node specific properties in the clustering process - in particular, we propose a new link density model (LD) that is able to take node degree into account during the learning process. Based on the exact continuous relaxation of the GHGC we analyze a variety of large scale complex networks and compare the Modularity, Hamiltonian, H&W model, MMSB model and LD model in terms of their capability to infer the inherent structure of networks based on their ability to predict links not seen during the learning process.

2 Methods

We define the generalized Hamiltonian for graph clustering (GHGC) as

$$\mathcal{H}(\boldsymbol{S}) = -\sum_{n=0}^{N} \mathrm{trace}[\boldsymbol{S}\boldsymbol{B}^{(n)}\boldsymbol{S}^\top \boldsymbol{J}^{(n)}] + \sum_j \boldsymbol{h}_j^\top \boldsymbol{s}_j \quad s.t. \quad \boldsymbol{B}_{i,i}^{(n)} = 0 \forall n, i. \tag{2.1}$$

Here $\boldsymbol{J}^{(n)}$ denotes couplings between groups, $\boldsymbol{B}^{(n)}$ denotes couplings between nodes and $\boldsymbol{h}_j$ the potential of node $\boldsymbol{s}_j$ where $\boldsymbol{S}$ is a binary clustering assignment matrix. N we will refer to as the order of the GHGC.

We note that existing community detection approaches can be expressed by the GHGC. For example the Hamiltonian and Modularity objectives correspond to a zero order GHGC[1], i.e. $N = 0$. In particular, there are only couplings between groups of same cluster-membership. The negative log-likelihood of the H&W model (see also [5]) as well as the MMSB can both be expressed as a first order GHGC[2], i.e. $N = 1$.

Efficient Optimization of $\mathcal{H}(\boldsymbol{S})$
Finding $\boldsymbol{S}$ mounts to optimizing the GHGC which can be achieved through standard annealing approaches and various variants of Gibbs sampling (see also [1, 5, 6] and references therein). The drawback of annealing approaches are that they are dependent on some problem specific cooling scheme while Gibbs sampling can be prohibitively slow. Optimizing $\mathcal{H}(\boldsymbol{S})$ with respect to the assignment matrix $\boldsymbol{S}$ (i.e. MAP-estimation) is in general a NP-hard binary combinatorial optimization problem. Despite this, the following theorem states that we can invoke standard continuous optimization[3] yet recover binary solutions at stationary solutions of the equivalent problem relaxed to the simplex.

Theorem 1. *A (local) optimum of the continuous optimization problem*

$$\arg\min_{\boldsymbol{S}} \mathcal{H}(\boldsymbol{S}) - \delta\, \mathrm{trace}(\boldsymbol{S}\boldsymbol{S}^\top) \quad s.t. \quad \boldsymbol{S} \geq 0, \quad \sum_c s_{c,j} = 1$$

is a binary 1-spin stable configuration for $\delta > 0$.

The full proof is left out for brevity but follows by deriving the stationary solutions of the continuous optimization problem and proving that these are 1-spin stable solutions, i.e. a single flip of assignment form suboptimal configurations. The term $\delta\, \mathrm{trace}(\boldsymbol{S}^\top \boldsymbol{S})$ is merely a technicality invoked to resolve potential ties and needless to say is negligible for small δ. In table 2 comparison between the proposed continuous optimization approach denoted $\mathrm{MAP}_\triangle$ with Gibbs sampling is given. From

[1] $J^{(0)} = \boldsymbol{I}, \boldsymbol{h}_j = 0 \forall j$ $\boldsymbol{B}^{(0)} = \boldsymbol{A} - \frac{m}{n^2}\boldsymbol{E}$ and $\boldsymbol{B}_{i,j}^{(0)} = \boldsymbol{A}_{i,j} - \frac{\boldsymbol{E}_{i,j}}{|\boldsymbol{A}|} k_i k_j$ respectively

[2] $\boldsymbol{B}^{(0)} = \boldsymbol{A}$ and $\boldsymbol{B}^{(1)} = \boldsymbol{E} - \boldsymbol{A}$, $\boldsymbol{J}_{c,c'}^{(0)} = \log \frac{\boldsymbol{P}_{c,c'}}{1-\boldsymbol{P}_{c,c'}}$, $\boldsymbol{J}_{c,c'}^{(1)} = \log(1 - \boldsymbol{P}_{c,c'})$, where $\boldsymbol{P}^{\mathrm{H\&W}} = \rho_c \boldsymbol{I} + \rho_n(\boldsymbol{1} - \boldsymbol{I})$. In both models clustering assignments $\boldsymbol{s}_j$ are drawn from a multinomial parameterized by μ resulting in the potential $\boldsymbol{h}_j = -2\log \mu \forall j$

[3] We used a standard projected gradient approach.

the figure it can be seen that the proposed continuous optimization identify better solutions than traditional MAP estimation based on assigning nodes to their most likely cluster. While modeling the parameter uncertainty in general improves learning of latent structure the proposed continuous optimization is computationally efficient and form also an efficient burn in method for sampling approaches.

The Link Density Model

An attractive property of the models proposed in [1, 5] are their generative nature, i.e. their ability to generate synthetic networks according to the statistical models imposed. However, the models operate with link probabilities parameterized by properties based on the grouping of the network and as such disregard properties at the node level. As such nodes that have low degree tend to cluster together regardless of their link properties. By assuming that the probability of a link between nodes in the network are formed by independent Bernoulli draws given by properties both of the grouping $\boldsymbol{P}_{c(i),c(j)} \in [0;1]$ where $c(i)$ and $c(j)$ denotes the assigned cluster of node i and j according to the MMSB model as well as node to node specific properties $\boldsymbol{R}_{i,j} \in [0;1]$ and that these two properties are independent we get the following parameterization of the probability for a link between node i and j in the network $Bern(\boldsymbol{A}_{i,j}|\boldsymbol{P}_{c(i),c(j)}\boldsymbol{R}_{i,j})$. After some rearrangement and using the fact that $\log(1-\alpha) = -\sum_{n=1}^{\infty} \alpha^n/n$ we find for the log-likelihood of the observed network

$$\log P(\boldsymbol{A}|\boldsymbol{R},\boldsymbol{P}) = \sum_{i,j} \boldsymbol{A}_{i,j} \log\left(\boldsymbol{R}_{i,j}\boldsymbol{P}_{c(i),c(j)}\right) + \left(\boldsymbol{E}_{i,j} - \boldsymbol{A}_{i,j}\right)\log\left(1 - \boldsymbol{R}_{i,j}\boldsymbol{P}_{c(i),c(j)}\right) =$$
$$\boldsymbol{A}_{i,j} \log\left(\boldsymbol{R}_{i,j}\boldsymbol{P}_{c(i),c(j)}\right) - \left(\boldsymbol{E}_{i,j} - \boldsymbol{A}_{i,j}\right)\sum_{n}^{\infty} \boldsymbol{R}_{i,j}^{n}\boldsymbol{P}_{c(i),c(j)}^{n}/n \equiv \sum_{n=0}^{\infty} \mathrm{trace}[\boldsymbol{S}^{\top}\boldsymbol{B}^{(n)}\boldsymbol{S}\boldsymbol{J}^{(n)}].$$

We approximate $\log(1-\alpha)$ by its second order expansion forming a 3rd order GHGC, i.e. $N = 3$ [4]. The above model is over-complete as $\boldsymbol{R}_{i,j}$ has as many free variables as links and non-links observed in the graph. Therefore strong assumptions have to be imposed to make the model identifiable in general. We use the outer product representation $\boldsymbol{R}_{i,j} = r_i r_j$ to enable a node specific link probability, e.g. as in the Modularity objective, but with r_i a new free parameter. As such, nodes with a very low number of links can potentially be included in a cluster with high link density if this is the cluster of greatest preference of the node. We will in the following denote this model the link density model (LD) as the model both take into account the varying densities between groups (as in the MMSB model) but also node specific link densities (i.e. node degree).

In figure 2 is given a simple graphical representation of the proposed link density model (LD), where the H&W and MMSB are special cases[5] that do not take into account node specific properties in the generative model. The hyper-parameters of the model, i.e. $\boldsymbol{P}$, $\boldsymbol{r}$ and $\boldsymbol{\pi}$ can be inferred through variational Bayesian inference, Gibbs sampling or maximum a posteriori (MAP) estimation. We used in the following if not otherwise stated variational Bayesian inference to estimate the hyper-parameters $\boldsymbol{P}$ and $\boldsymbol{\pi}$ of the H&W, MMSB and LD model by imposing non-informative conjugate Beta and Dirichlet priors on these model parameters respectively and MAP estimation to infer $\boldsymbol{r}$ and $\boldsymbol{S}$.

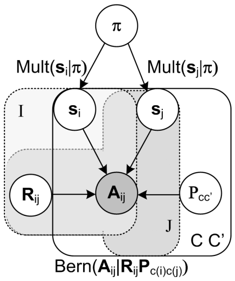

Figure 2: Graphical representation of the link density model (LD). The node assignment is drawn from a multinomial while links and non-links between node i and j ($\boldsymbol{A}_{i,j}$) are drawn according to a bernoulli distribution given by $Bern(\boldsymbol{A}_{i,j}|\boldsymbol{E}_{i,j}\boldsymbol{P}_{c(i),c(j)})$ where $\boldsymbol{P}$ denotes the probability of links occurring between communities $c(i)$ and $c(j)$ according to the MMSB. $\boldsymbol{R}_{i,j}$ gives the probability of links at the node level between node i and j. We assume $\boldsymbol{R}_{i,j} = r_i r_j$ such that node specific link probabilities can be taken into account during the clustering process. The model can be expressed in the form of the GHGC. Priors on all the latent parameters can further be imposed (not shown).

Properties of the MMSB and LD model

Since both the H&W, MMSB and LD models are generative it is fairly transparent what types of network dynamics the models are suitable for. Contrary to the H&W the MMSB and LD models can account for variability in link densities between the groups of nodes and the LD model can further account for potential variations in the degree distributions of the nodes. In particular, the interpretation of each entry $\boldsymbol{P}_{c,c'}$ of $\boldsymbol{P}$ denote the observed density of links between groups of nodes (for

[4] $\boldsymbol{B}_{i,j}^{(0)} = \boldsymbol{A}_{i,j}log(\boldsymbol{R}_{i,j})$, $J^{(0)} = \mathbf{1}$, $\boldsymbol{B}_{i,j}^{(1)} = \boldsymbol{A}_{i,j}$, $\boldsymbol{J}^{(1)} = \log\boldsymbol{P}$, $\boldsymbol{B}^{n+1} = -(\boldsymbol{E}_{i,j} - \boldsymbol{A}_{i,j})\boldsymbol{R}_{i,j}^{n}$, $\boldsymbol{J}_{c,c'}^{(n+1)} = \boldsymbol{P}_{c,c'}^{n}/n$ for $n > 0$

[5] Note that we here for brevity have disregarded potential links generated by noise as described in [1].

non-informative priors). As such the following properties of networks can be modelled:
Communities: Communities constitute regions of high density in the network between nodes with same cluster membership, i.e. $\boldsymbol{P}_{c,c} > \boldsymbol{P}_{c,c'} \forall c' \neq c$.
Satellites: constitute regions where nodes within same cluster have low density but have high density with at least another group c, i.e. $\exists c' : \boldsymbol{P}_{c,c} < \boldsymbol{P}_{c,c'}$.
Overlapping groups: The degree of overlap between groups of nodes are given by $\boldsymbol{P}_{c,c'}$. Furthermore, we define the overlap coefficient $\boldsymbol{\Lambda}_{c,c'} = \frac{\boldsymbol{P}_{c,c'}}{\sqrt{\max_{c''} \boldsymbol{P}_{c,c''} \max_{c''} \boldsymbol{P}_{c',c''}}}$ indicating the between group ratio relative to the most strongly connected group to c and c'. This coefficient takes value between 0 and 1 where 0 indicate no connectivity between the groups whereas 1 indicate that this is the most prominent relation for both groups.
Hierarchy: The matrix $\boldsymbol{P}$ as well as the overlap coefficient $\boldsymbol{\Lambda}$ can be considered a similarity matrix from which group hierarchies can be inferred.
Node degree distributions: r_i is a free parameter that can take into account the node specific degree distribution during the clustering process.
Both the MMSB and the LD model share the first four important properties that are a result of the between and within group densities parameterized by $\boldsymbol{P}$. The properties above are illustrated in figure 3 and demonstrated on a real network in figure 5.

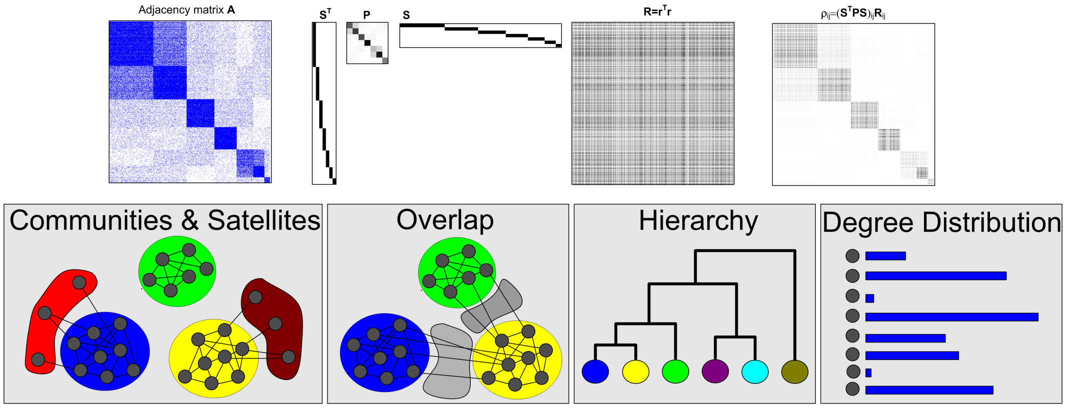

Figure 3: **Top panel:** Illustration of the LD model. Based on the adjacency matrix of the network the MMSB and LD model simultaneously estimate the grouping of nodes $\boldsymbol{S}$ and densities between groups of nodes $\boldsymbol{P}$. As a result the network is modeled as blocks defining regions with varying probabilities for the occurrence of links. The LD model further account for varying degree distribution of the nodes by $\boldsymbol{R} = \boldsymbol{r}^\top \boldsymbol{r}$ resulting in the overall link probabilities within the network given to the right. **Bottom panel:** While communities in the MMSB and LD models corresponds to rows and columns of $\boldsymbol{P}$ where the diagonal element contain the largest density, the MMSB and LD model can account for groups that we denote satellites. Satellites are given by nodes that have little within group connectivity compared to the connectivity to some of the other groups in the network. Overlap between groups of nodes are naturally accounted for by the between group link densities $\boldsymbol{P}_{c,c'}$ and as such the derived between and within group density matrix $\boldsymbol{P}$ can be considered a measure of similarity between groups from which group hierarchies can be inferred. Finally, for the LD model the parameter $\boldsymbol{r}$ can take properties of node degree into account in the clustering process.

Evaluating Network Models by Link Prediction

Cross-validation is a well established framework to evaluate model performance in terms of generalization error. Recently, link-prediction has been proposed for the evaluation of models for complex networks [2, 4]. Within the GHGC link prediction can be inferred by treating some of the links and non-links as missing in the model estimation process. Let $\boldsymbol{W}$ be an indicator matrix of links and non-links treated as missing (for prediction) in the model estimation. These missing values can straightforward be ignored in the model estimation process (i.e., marginalized) by setting $\boldsymbol{B}^{(n)}_{i,j} = \boldsymbol{B}^{(n)}_{i,j} - \boldsymbol{B}^{(n)}_{i,j} \boldsymbol{W}_{i,j}$ in the expression of the GHGC. This changes the computational cost to $\mathcal{O}(c \max(|\boldsymbol{A}|, |\boldsymbol{W}|))$. To get a large number of links for prediction in the testing process without introducing a heavy computational cost in the marginalization process we set the number of missing

	♯ nodes	♯ links	r	c	L	ρ
H&W model	8,800	687,272	0.7027	0.039	2.741(0.026)	8.9e-3
MMSB model	8,800	1,181,692	0.8612	0.062	2.623(0.032)	15.3e-3
LD model	8,800	1,484,592	0.4589	0.096	2.623(0.070)	19.2e-3
Yeast Network	2,284	13,292	-0.0991	0.134	4.388(0.244)	2.5e-3
US Power	4,941	13,188	0.0035	0.080	19.883(0.668)	5.4e-4
Erdos02	5,534	16,944	-0.0399	0.079	3.881(0.004)	5.5e-4
Free Assoc.	10,617	127,576	-0.0720	0.119	3.908(0.132)	1.1e-3
Reuters911	13,314	296,076	-0.1090	0.368	3.060(0.098)	1.7e-3
Wordnet3	31,867	240,798	-0.0911	0.029	7.196(0.319)	2.4e-4
Dictionary28	39,327	178,076	0.7080	0.222	7.809(0.507)	1.2e-4
CondPhys2005	39,577	351,386	0.1863	0.650	4.559(0.505)	2.2e-4
Internet	124,650	387,240	-0.0078	0.062	11.476(0.630)	2.5e-5
IMDB	896,308	115,025,018	0.2002	0.790	3.589(0.125)	1.4e-4
Patents	3,774,768	29,941,533	0.1071	-	8.5683(0.2259)	2.1e-6

	$corr(\boldsymbol{k}, \boldsymbol{r}^{\mathrm{LD}})$
H&W model	0.713(3)
MMSB model	0.690(2)
LD model	0.762(5)
Yeast Network	0.399(10)
US Power	0.505(12)
Erdos02	0.401(21)
Free Assoc.	0.351(5)
Reuters911	0.309(9)
Wordnet3	0.140(11)
Dictionary28	0.510(8)
CondPhys2005	0.126(7)
Internet	0.171(11)
IMDB	0.100(15)
Patents	-

Table 1: **Left Table:** Properties of the analyzed networks. r denotes the networks assortativity, c denotes the clustering coefficient [11], L the average shortest path and ρ the density of the network. The average shortest path measure was calculated as the average of 10 samples of up to 10,000 links in the network disregarding non-existing paths between nodes (in parenthesis is given the standard deviation of this mean over the samples). **Right Table:** Correlation between node degree distribution and the estimated node specific parameter $\boldsymbol{r}$ of the LD model. Given are the average correlation over 10 model estimations as well as the standard deviation on the last digit. Clearly, there is a significant correlation for all the networks.

links and non-links to be the same. This can potentially introduce a small model bias as the link to non-link ratio of the full network and training network no longer will be exactly the same. In the next section we compare the performance of the Modularity, Hamiltonian, H&W, MMSB and LD models in terms of their ability to predict links. For completion we include the following commonly used non-parametric link prediction scoring methods:

Degree Product:	$\alpha_{i,j} = k_i k_j$	***Shortest Path:***	$\alpha_{i,j} = min_p\{(\boldsymbol{A}^p)_{i,j} > 0\}$
Common Neighbour:	$\alpha_{i,j} = a_i^\top a_j$	***Jaccard:***	$\alpha_{i,j} = a_i^\top a_j/(k_i + k_j - a_i^\top a_j)$

As performance measure of the various link prediction approaches we will use the area under curve (AUC) of the receiver operating characteristic (ROC) proposed in [2] which is a measure that is invariant to the ratio of links to non-links used for prediction.

3 Results

We analyzed a variety of benchmark network data given in table 1. The H&W, MMSB and LD datasets were generated according to their respective generative model using 7 clusters of varying sizes ranging from 500 to 2000 nodes. Directed as well as weighted networks were turned undirected and un-weighted for the analysis disregarding link directions and weights [6]. It is an open problem how many clusters to expect in the data we set $c_{max} = 50$ for all the analysis. The algorithms were terminated when there was no change of assignment in $\boldsymbol{S}$ or when 500 iterations had progressed.

From the right of table 1 it can be seen that there is a significant positive correlation between node degree and the estimated value of $\boldsymbol{r}$ in the LD model for all the networks analyzed by the model.

In table 3 the mean link prediction performance (i.e. mean AUC value) over 10 random splits for the various link prediction approaches[7] using a total of up to $c_{max} = 50$ clusters is given. In each of the analysis 1 % of links as well as an equivalent number of non-links (both randomly chosen) were treated as missing[8]. In half of the analysis the MMSB and LD model are the best performing methods for predicting links and when outperformed this is by non-parametric methods such as the shortest path method. In particular, when comparing table 3 with table 1 this is the case when the average shortest path length between nodes in the network is large. Thus, the MMSB and LD model perform

[6] We note that the proposed approach readily generalize to the analysis of directed and weighted networks. Directed networks can potentially be modelled by asymmetric link densities(i.e., $\boldsymbol{P} \neq \boldsymbol{P}^\top$ whereas weighted networks can be considered networks where multiple links have been drawn between the nodes of the network. These types of analysis is however out of the scope of the present paper.

[7] Due to the large sizes of the IMDB and Patents networks we only ran 3 runs for these two networks and did not include a LD analysis for the Patents network.

[8] For the small networks Yeast , US Power and Erdos02 we treated 5% instead of 1% as missing to have a reasonable validation data set size

AUC	**MAP**	**MAP$_\triangle$**	**Gibbs**	**MAP$_\triangle$ $\rightarrow$ Gibbs**	**Gibbs^{All}**	**MAP$_\triangle$ $\rightarrow$ Gibbs^{All}**
Reuters911	0.937(3)	0.944(2)	0.944(2)	0.946(2)	0.944(2)	0.947(2)
Wordnet3	0.712(9)	0.740(8)	0.820(7)	0.804(5)	0.832(5)	0.805(6)
Dictionary28	0.780(8)	0.813(5)	0.851(8)	0.836(6)	0.859(5)	0.835(5)
CondPhys2005	0.837(8)	0.892(2)	0.907(1)	0.908(2)	0.909(2)	0.907(2)

cpu-time (seconds)	**MAP**	**MAP$_\triangle$**	**Gibbs**	**MAP$_\triangle$ $\rightarrow$ Gibbs**	**Gibbs^{All}**	**MAP$_\triangle$ $\rightarrow$ Gibbs^{All}**
Reuters911	44.55(5.29)	27.15(1.91)	334.51(3.40)	104.68(2.08)	388.46(4.12)	116.37(2.22)
Wordnet3	37.27(4.17)	39.36(2.20)	763.80(5.70)	184.53(2.40)	813.92(7.43)	187.43(2.58)
Dictionary28	34.84(1.90)	41.61(1.67)	1068.13(7.19)	244.38(2.59)	1147.90(9.00)	251.03(2.84)
CondPhys2005	84.70(7.30)	130.76(5.63)	1413.74(14.59)	420.51(5.80)	1548.65(21.55)	430.98(6.20)

Table 2: Comparison of the proposed **MAP$_\triangle$** estimation approach based on continuous optimization to regular **MAP** estimation (i.e. assigning nodes to their most likely cluster) and sampling of $\boldsymbol{S}$ and all model parameters (i.e. **Gibbs** and **Gibbs^{all}**) for the MMSB model [1] for the four medium sized networks using different random initialization. Given are the mean AUC values as well as their standard deviations on the last digit given in parenthesis across 10 data splits with randomly chosen links and non-links treated as missing (for the sampling approaches we used 400 iterations for burn in and 100 iterations for parameter estimation). At the bottom the mean Matlab cpu-time and standard deviation (in seconds) for each of the methods can be found. **MAP** and **MAP$_\triangle$** converged in general in less than 100 iterations.

in general very well except when the nodes of the graphs are far apart rendering all the community detection approaches unable to well detect group structure.

In figure 4 examples of the results obtained by permuting graphs according to the various community detection approaches are given. Comparing the correlation between the median node degree within the nodes of each cluster with the node density of each cluster there is a significant difference between the obtained correlations of the MMSB and LD model. Thus, as expected the LD model has a less tendency to cluster nodes together based on their degree properties as this can be accounted for by the additional parameter $\boldsymbol{r}$. In figure 5 we illustrate how hierarchies and group interactions between communities and so-called satellites can be derived from the estimation of $\boldsymbol{P}$ for the MMSB and LD models. Since the LD model to a lesser extend group nodes according to their node degree fewer satellites are observed in the analysis.

	Degr. Prod.	**Short. Path**	**Com. Neigh.**	**Jaccard**	**Hamilt.**	**Modularity**	**H&W**	**MMSB**	**LD**
H&W model	0.627(5)	0.788(7)	0.827(8)	0.828(8)	0.869(2)	0.867(3)	**0.873(2)**	0.868(2)	0.872(1)
MMSB model	0.621(5)	0.786(5)	0.840(5)	0.845(5)	0.866(2)	0.856(6)	0.861(3)	**0.892(1)**	0.888(1)
LD model	0.690(2)	0.688(2)	0.834(3)	0.825(3)	0.866(2)	0.866(2)	0.870(1)	0.904(1)	**0.905(1)**
Yeast Network	0.783(6)	0.795(9)	0.675(9)	0.675(9)	0.640(13)	0.599(14)	0.573(9)	**0.836(7)**	0.794(9)
US Power	0.449(10)	**0.795(6)**	0.474(13)	0.474(13)	0.479(12)	0.489(13)	0.544(13)	0.407(8)	0.506(9)
Erdos02	0.586(9)	0.580(9)	0.584(5)	0.559(8)	0.530(15)	0.460(11)	0.377(12)	**0.954(3)**	0.873(22)
Free Assoc.	0.845(6)	0.877(5)	0.856(6)	0.854(6)	0.766(11)	0.632(8)	0.609(6)	**0.902(5)**	0.872(4)
Reuters911	0.928(3)	0.892(4)	0.929(3)	0.910(3)	0.728(8)	0.601(5)	0.766(4)	**0.942(2)**	0.935(2)
Wordnet3	0.602(6)	0.613(6)	0.356(4)	0.356(4)	0.507(8)	0.463(7)	0.481(6)	**0.795(8)**	0.658(8)
Dictionary28	0.808(5)	**0.917(4)**	0.776(4)	0.744(5)	0.644(6)	0.633(4)	0.678(6)	0.865(6)	0.894(3)
CondPhys2005	0.790(3)	0.963(2)	0.964(1)	**0.965(1)**	0.737(8)	0.689(5)	0.463(12)	0.909(2)	0.924(1)
Internet	0.604(5)	**0.745(4)**	0.501(4)	0.501(4)	0.662(4)	0.672(5)	0.607(8)	0.695(5)	0.663(5)
IMDB	0.918(1)	0.980(5)	**0.998(0)**	0.997(1)	0.864(2)	0.857(2)	0.861(2)	0.965(2)	0.974(1)
Patents	0.766(1)	**0.946(3)**	0.743(5)	0.743(5)	0.770(1)	0.768(1)	0.696(14)	0.889(1)	-

Table 3: Link prediction performance of the various methods on the 14 networks. Given are the mean AUC values as well as their standard deviations on the last digit given in parenthesis across 10 data splits with randomly chosen links and non-links treated as missing. In bold black is given the best performing approach and in underline the best performing community detection approach. The Hamiltonian was based on average link density as the imposed null hypothesis, i.e. $\boldsymbol{B} = \frac{m}{n^2}\boldsymbol{E}$. All clusters were modelled with $c_{max} = 50$.

4 Discussion

We have demonstrated that a wide range of existing community detection approaches for complex networks can be posed as an optimization of the generalized Hamiltonian for graph clustering (GHGC). We further proved how the optimization of GHGC with respect to the clustering assignment matrix $\boldsymbol{S}$ can be efficiently solved by continuous optimization based on an exact relaxation to the simplex of the GHGC. We compared a variety of community detection approaches in their ability to predict links within the GHGC framework including the proposed link density model (LD) that can

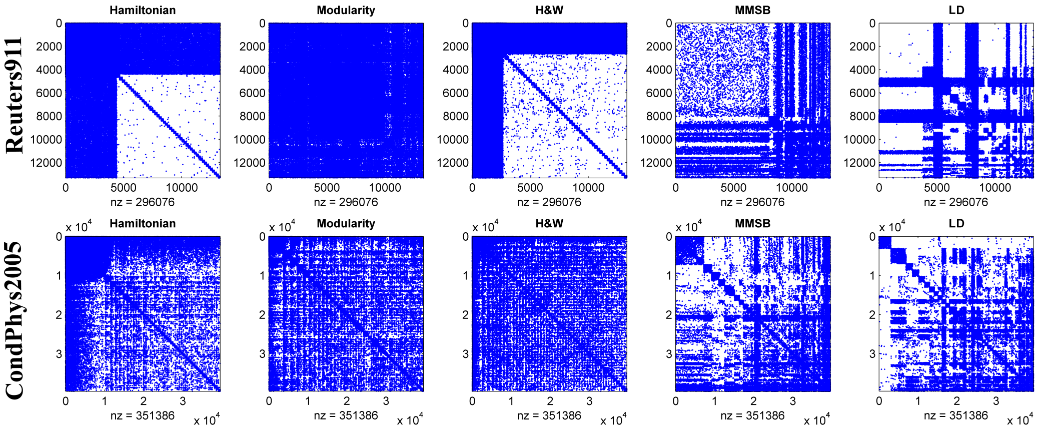

Figure 4: Permuted graphs obtained from the clustering assignments derived from the various community detection approaches shown for the Reuters911 and CondPhys2005 networks. Clearly, both the MMSB and LD models are able to split the graph into blocks of varying link densities. It further appears that the LD model to a lesser extend group nodes with low link degree together. As such the average correlation between median degree of the nodes and the estimated link density of the group for the MMSB model vs. the LD model are 0.900(33) vs. 0.704(33) and 0.874(4) vs. 0.686(16) for the two displayed networks respectively. Notice, despite that it appears the LD model has better split the Reuters911 network into separate parts the model does not predict links as well as the MMSB model.

account for node degree in the clustering process. This analysis demonstrated that the MMSB model and the proposed LD model are superior in extracting inherent structures of networks compared to Modularity and Hamiltonian optimization as well as the H&W model given in [5]. This performance we believe to be due to the models ability to account for overlap between groups of nodes, so-called satellites formed by nodes with little inter-connectivity relative to the connectivity to other groups in the network as well as the ability to model cluster hierarchies, i.e. the MMSB and LD model explicitly quantify cluster similarities. The LD model is able to account for node degree in the clustering process, however, compared to the MMSB the proposed LD model improve the predictions only for a few of the networks but results for the majority of the networks in a less adequate description of the data. Whether this is a result of the model being more prone to local minima due to the extra set of parameters to be estimated is an open question.

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, 2008.

[2] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[3] Y. Fu and P.W. Anderson. Application of statistical mechanics to np-complete problems in combinatorial optimisation. *J. Phys. A: Math. Gen.*, 19:1605–1620, 1986.

[4] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[5] J. M. Hofman and C. H. Wiggins. A bayesian approach to network modularity. *Phys. Rev. Lett.* 100:258701, 2008.

[6] M. Mørup and L. K. Hansen. An exact relaxation of clustering. *Technical Report*, 2009.

[7] M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.

[8] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, 103(23):8577–8582, 2006.

[9] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113–1–15., 2004.

[10] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(1):016110, 2006.

[11] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

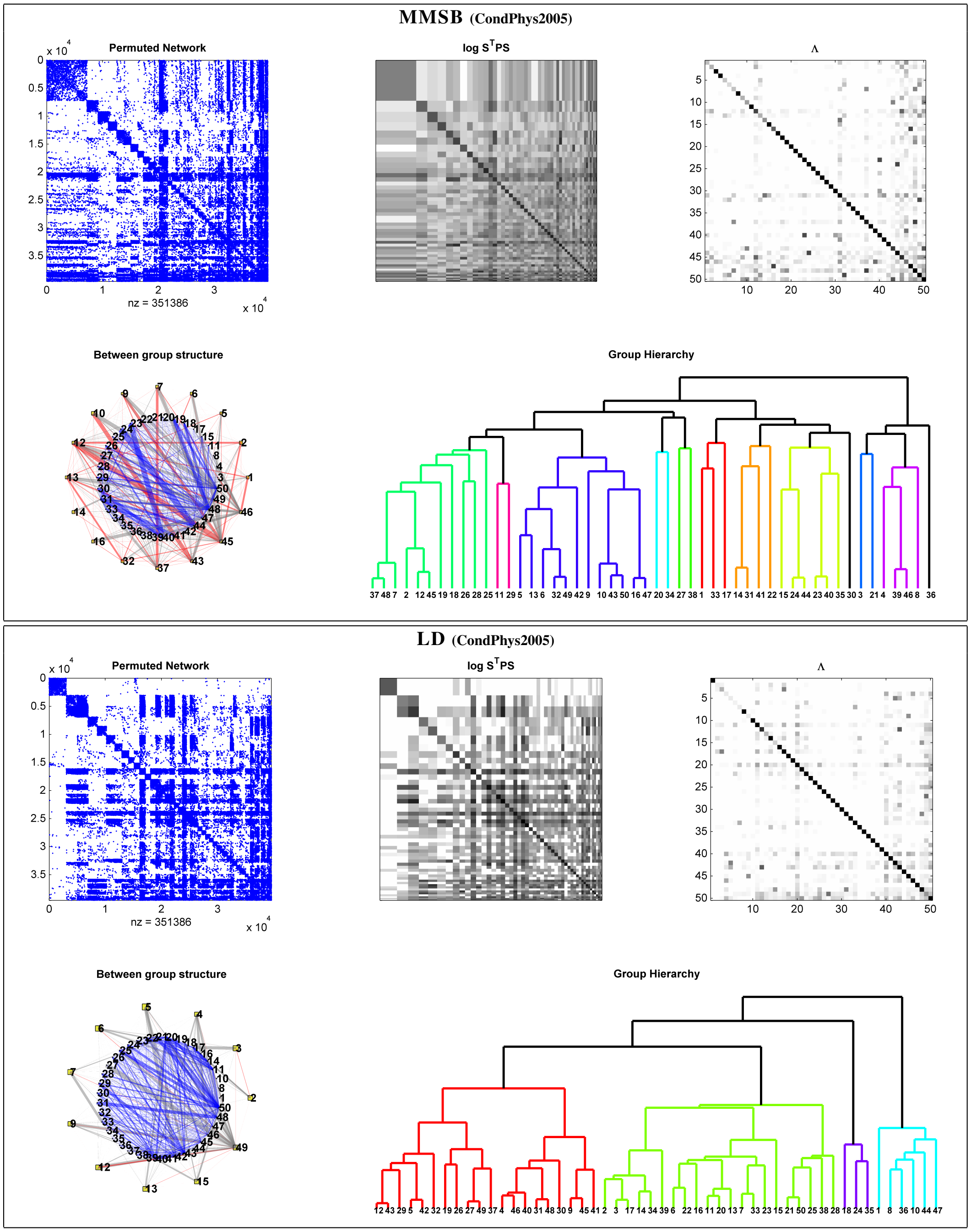

Figure 5: Visualization of the MMSB (at the top) and LD (at the bottom) results for the CondPhys2005 collaboration network. Given are in the top panels the permuted network adjacency matrix, the log link probabilities as well as the derived **Λ** defining communities and satellites in the networks. In the bottom panels are shown the derived between group dynamics. Here outer boxes indicate groups that are satellites, inner circles groups that form communities, and blue links between community relations, red links within satellite connection and gray links between community and satellite relations. Width of connections indicate their strength whereas the sizes of the boxes and circles the log size of the satellites and communities respectively. To the bottom right is given the derived group hierarchies. As can be seen the LD models ability to take into account the nodes degree distribution has reduced the number of satellites extracted as nodes that have few links are not necessarily forced into the same low density group in the clustering process.