Generating Synthetic Internet- and IP-Topologies using the Stochastic-Block-Model

Patrick Kalmbach, Andreas Blenk, Markus Kluegel and Wolfgang Kellerer
Chair of Communication Networks
Department of Electrical Engineering
Technical University of Munich, Germany
Email: {patrick.kalmbach,andreas.blenk,markus.kluegel,wolfgang.kellerer}@tum.de

***Abstract*—Developing models to generate realistic graphs of communication networks often requires a deep understanding and extensive analysis of the underlying network structure. Since deployed communication networks are dynamic, the findings a generator is based on might lose validity. We alleviate the need for extensive analysis of graphs by estimating parameters of a probabilistic model. The model parameters encode the structure of the graph, which is thus learned in an unsupervised fashion. Synthetic graphs can be generated from the model and will have the structure previously inferred. For this, we use the Stochastic-Block-Model (SBM) and the Degree-Corrected-Block-Model (DCBM), a variant allowing for heavy tailed degree distributions. The models originate in the social sciences and separate a graph into groups of nodes. To show the applicability of the models to the task of synthetic graph generation in the domain of communication networks, we use one router level and one IP-to-IP communication graph. We assert the quality of the generated models by evaluating a number of graph features and comparing our results to those obtained with the network generator Orbis. We find our approach to be on par with, or even outperforming Orbis. Furthermore, the models are able to capture large-scale structure in communication networks.**

I. Introduction

Synthetic generation of graphs with characteristics of communication networks is important in order to develop, test and verify e.g., routing or resource management algorithms. Many models and generators that produce graphs with properties similar to communication networks, mainly the Internet topology, have been proposed. Examples are preferential attachment [1]–[3] models or generators replicating the degree distribution [4], [5] or matching structural patterns [6]–[8]. However, many network generators replicate only a subset of the original graphs' properties [6], [9]. In addition, communication networks change over time or develop new characteristics, which makes a continuous and extensive analysis necessary. Other graph types, e.g. IP-to-IP communication between hosts, which is used for anomaly detection, resource management or traffic classification, are important [10]–[12]. A generator modeling the current Internet topology might fail to generate realistic graphs in this case. Accordingly, challenges in generating synthetic graphs for different use-cases can be summarized as:

- Empirical graphs of different networking domains show and may combine different structural properties [6].
- Communication networks are dynamic and their respective graphs change over time [9].
- Structure of a graph can be influenced by different mechanisms [2] or node properties [10].

Other research domains such as social sciences and biology also heavily rely on methodologies for efficient graph analysis. Here, machine learning has become an important tool to infer structures in complex graphs. Exemplar applications are the detection of communities in social networks and of functional components in protein-protein interaction networks [13], [14]. The used technique is a latent variable model for relational data called *Stochastic-Block-Model (SBM)* [15], representing a parametric probability distribution over graphs. The model parameters encode the graph structure and can be inferred in an unsupervised fashion from empirical data. Furthermore, synthetic graphs with similar structure can be sampled from the inferred distribution. Given the ability to represent various structural patterns, we believe the SBM to have huge potential for simplifying synthetic communication network generation by replacing extensive network analysis with machine learning techniques. In fact, an SBM variant identifying hierarchies in a graph has already been applied to CAIDA-traces in [14]. However, a deep and extensive analysis of the application of SBMs and their variants to varying types of communication networks, and especially their generative capabilities, has not been provided yet.

As a first step towards understanding how block modeling can generally be used in the context of communication networks, we apply the SBM to a router-level and an IP-to-IP communication graph. Since respective graphs express heavy tailed degree distributions, we additionally consider a variant of the SBM, the Degree-Corrected-Block-Model (DCBM), which is able to model such distributions. We compare our results with those obtained using the generator Orbis [5]. The results show that probabilistic models (1) successfully infer large scale structures and (2) can generate synthetic graphs preserving those structures.

The remainder of this paper is organized as follows: In Sec. II we introduce related work. Sec. III introduces the SBM and DCBM. Sec. IV compares DCBM, SBM and Orbis. We conclude and provide ideas for future work in Sec. V

978-3-901882-89-0 @2017 IFIP

II. Related Work

We use the SBM in this work for the *detection of groups of "similar" nodes* and for the *generation of synthetic graphs*. Both tasks have been addressed in the context of communication networks already. We separate the related work into two sections accordingly.

A. Detection of Groups of Nodes

Detection of groups is used in the context of communication networks to classify traffic [10], to provide host profiles for anomaly detection [12] or content management [16]. Work in [11] provides a general framework for detection of groups in IP-to-IP communication networks

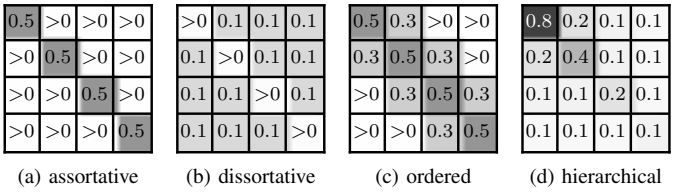
None of the authors highlight the use-case of synthetic graph generation. Also, they do not use a probabilistic framework for group detection and mostly assume assortative mixing. In contrast, an SBM makes no explicit assumption about the mixing pattern.

B. Generation of Synthetic Graphs

In the area of communication networks, popular models for generating synthetic graphs already exist. For instance, in the Waxman model [17], the probability of an edge is a function of the spatial location of nodes. In the preferential attachment model, the probability of an edge between two nodes depends on the nodes' degree. The Watts-Strogatz model generates networks with the small world property [25].

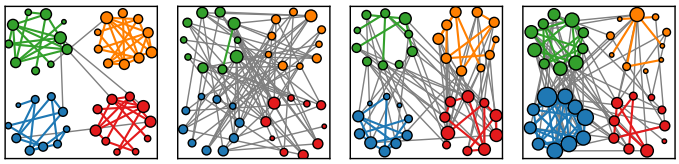
Many network generators exist that build upon the mentioned models. For example, GT-ITM [7] is based on the Waxman model and generates graphs by separately modeling hierarchical levels of the Internet graph. The more recent generator GeoTopo [8] is similar in principle to GT-ITM and incorporates not only hierarchy, but also considers geographic, demographic and economic information for topology generation. BRITE [3] is based on preferential attachment and targets the connectivity pattern of the Internet graph by generating a degree distribution following a power-law. INET [19] is similar to BRITE and also combines structural and degree information. S-BITE [6] separates the Internet graph into different levels, models the connectivity between and within the levels and uses specific degree distributions, as well as preferential attachment and small world mechanisms. In contrast, Orbis [5] generates graphs according to a given (joint) degree distribution. Furthermore, Orbis is able to generate graphs of different sizes by scaling the distribution and is applicable to arbitrary graphs.

Our approach differs in that we learn structural properties of a graph in an unsupervised fashion by estimating the parameters of a probability distribution over graphs. Extensive analysis of empirical network data, which was necessary for the design of existing network generators, is replaced by inference of the model parameters, for which principled methods exist. As a result, our approach is applicable to arbitrary graphs, contrary to most of the existing generators [3], [6]–[8], [19]. Synthetic graphs can be generated by sampling from the inferred distribution.



(a) assortative (b) dissortative (c) ordered (d) hierarchical

Fig. 1. Block-matrices for $k = 4$ groups, representing different forms of large scale structure.



(a) assortative (b) dissortative (c) ordered (d) hierarchical

Fig. 2. Graphs drawn from the distribution defined by the block matrices in Fig. 1. Each group contains ten nodes.

III. Introduction to the Stochastic-Block-Model

The Stochastic-Block-Model (SBM) is a probabilistic generative model for relational data, and has its origins in mathematical sociology [15]. In social sciences, the SBM is used to identify communities of tightly interconnected nodes, but is not limited to this pattern. The SBM defines a parametric probability distribution $P(A \mid M, z)$ over graphs, where the parameters are:

- number of groups $k \in \mathbb{R}$,
- group membership of nodes $z \in \{1, \ldots, k\}^N$ where z_i gives the group of node i and N is the number of nodes,
- a Stochastic-Block-Matrix $M \in \mathbb{R}^{k \times k}$ where each entry M_{rs} gives the probability of a node in group r to connect to a node in group s.

Thus, a (complex) graph is reduced to a set of groups and inter- and intra group connectivity, with the central assumption: *Nodes connect to other nodes solely based on their group membership*. In addition, nodes in each group share the same set of parameters (row/column in M), and are thus stochastically equivalent.

During graph synthesis, a graph for one specific choice of M and z is generated from the model. Each possible edge A_{ij} is connected according to a Bernoulli experiment with probability $M_{z_i z_j}$ for success. Edges can thus be seen as conditionally independent random variables given the node membership. Edges running between two groups r and s are in addition identically (thus iid) distributed. Fig. 1 shows block matrices and Fig. 2 respectively sampled graphs for specific choices of M and z. The figures illustrate how the parameters of the SBM encode different structural properties.

During inference, i.e., if we do not know M, z and k, the task is to find the most likely parameters that could have generated the observed graph. We adopt the approach in [13] and use a Maximum-Likelihood Estimate (MLE) for the parameters z and M. That is, we want to find the parameter $\hat{M}, \hat{z}$ that maximize the likelihood $P(A \mid M, z)$. Parameter k is a free parameter and must be chosen separately. However, the larger we choose k, the larger the model's likelihood

becomes, approaching 1 as k approaches N. The difficulty is to choose k, such that the model captures relevant structural characteristics, but ignores noise [14].

A. Inference for the Stochastic-Block-Model (SBM)

This section outlines how parameters M and z are estimated given an observed adjacency matrix A. For more details see [13], and [14] for a principled method to choose k.

To simplify mathematical treatment, an undirected multi-graph is usually chosen as graph representation. A directed version can be easily obtained by allowing M and A to be a-symmetric. This changes the model parameters and the generative process given previously:

- adjacency matrix $A \in \mathbb{N}_0^{N\times N}$ with element A_{ij} for $i \neq j$ gives the number of edges between nodes i and j and twice the number of self-edges for $i = j$.
- Stochastic-Block-Matrix $M \in \mathbb{N}_0^{N\times N}$ where M_{rs} is the expected number of edges between groups r and s.
- Edges no longer follow a Bernoulli but a Poisson distribution: $A_{ij} \sim Poi(M_{z_i,z_j})$

By taking an MLE approach, we aim to choose M and z such that the probability $P(A \mid M, z)$, i.e., the a-posteriori likelihood, of generating exactly the observed edges is maximized. The probability is given by:

$$P(A \mid M, z) = \prod_{i=1}^{N-1}\prod_{j=i+1}^{N} \frac{(M_{z_i z_j})^{A_{ij}}}{A_{ij}!}\exp(-M_{z_i z_j}) \times \prod_{i=1}^{N} \frac{(M_{z_i z_i})^{\frac{1}{2}A_{ii}}}{\frac{1}{2}A_{ii}!}\exp(-\frac{1}{2}M_{z_i z_i}). \tag{1}$$

Rearranging terms and taking the logarithm of (1) gives:

$$\log P(A \mid M, z) \propto \sum_{r=1}^{k}\sum_{s=1}^{k}\left(m_{rs}\log M_{rs} - n_r n_s M_{rs}\right), \tag{2}$$

where n_r is the number of nodes in group r and m_{rs} is the total number of edges running between groups r and s, or twice that much if $r = s$. The MLE of M_{rs} is then:

$$\hat{M}_{rs} = \frac{m_{rs}}{n_r n_s}. \tag{3}$$

Taking the result from (3), plugging it into (2) and omitting constant terms, we obtain the unnormalized log likelihood:

$$\log P(A \mid z) \propto \sum_{r=1}^{k}\sum_{s=1}^{k} m_{rs}\log\frac{m_{rs}}{n_r n_s}, \tag{4}$$

depending solely on the group memberships z of the nodes. Eq. (4) can be efficiently computed and gives rise to a heuristic algorithm to obtain $\hat{z}$, which is described in [13]. As the problem is not convex, reestimating parameters with different random assignments is recommended.

One of the issues with the SBM is that it tends to group nodes according to their degree. As [20] demonstrates, this can be an asset when the graph expresses core-periphery structure, which graphs arising in the context of communication networks frequently do [6], [7], [18]. On the other hand, [6] reports the emergence of community structure in the Internet graph, and assortative mixing was reported for IP-to-IP communication graphs [10], [11]. Here, the SBM might fail to correctly infer the actual group structure due to degree heterogeneity within the groups. In Sec. III-B, we thus review inference for the DCBM, a variant of the SBM, which detects groups containing nodes with heterogeneous degrees [13].

B. Inference for the Degree-Corrected-Block-Model (DCBM)

The authors in [13] propose a degree-corrected variant of the SBM, the Degree-Corrected-Block-Model (DCBM), by introducing a new parameter θ_i for each node. This parameter models the expected degree of the node. We briefly outline the DCBM and refer the reader to [13] for more details.

Again, we assume undirected multi-graphs. Our goal is now to choose values for M, z and θ that maximize the likelihood. As before, k is a free parameter and the problem is not convex. The probability of an undirected multi-graph, given by adjacency matrix A is:

$$P(A \mid M, z, \theta) = \prod_{i=1}^{N-1}\prod_{j=i+1}^{N} \frac{(\theta_i\theta_j M_{z_i z_j})^{A_{ij}}}{A_{ij}!}\exp(\theta_i\theta_j M_{z_i z_j}) \times \prod_{i=1}^{N} \frac{(\theta_i^2 M_{z_i z_i})^{\frac{1}{2}A_{ii}}}{(\frac{1}{2}A_{ii})!}\exp(\theta_i^2 M_{z_i z_i}). \tag{5}$$

Rearranging and taking the logarithm yields:

$$\log P(A \mid M, z, \theta) \propto 2\sum_{i=1}^{N} k_i \log\theta_i + \sum_{r=1}^{k}\sum_{s=1}^{k}(m_{rs}\log M_{rs} - M_{rs}), \tag{6}$$

from which we derive the MLE estimates for θ and M by setting the gradient of Eq. (6) equal to zero:

$$\hat{\theta}_i = \frac{k_i}{\kappa_{z_i}}, \quad \hat{M}_{rs} = m_{rs}, \tag{7}$$

where m_{rs} is the total number of edges running between communities r and s, and κ_{z_i} is the sum of degrees in group z_i. Similar to Section III-A we can write the un-normalized log-likelihood in sole dependence of z by plugging the respective MLE solutions into Eq (6):

$$\log P(A \mid z) \propto \sum_{r=1}^{k}\sum_{s=1}^{k} m_{rs}\log\frac{m_{rs}}{\kappa_r\kappa_s}. \tag{8}$$

The same heuristic algorithm used for the SBM and described in [13] can be used to obtain the group membership of the nodes. Eq. (4) is replaced as objective function by Eq (8).

The DCBM cannot be applied to directed graphs as readily as the SBM. Since nodes in directed graphs have in general different in- and out-degrees, the directed version of the DCBM needs two parameters per node. One parameter to model

the expected in- and the other to model the expected out-degree [21]. The unnormalized log-likelihood for the directed case is then:

$$\log P(A \mid z) \propto \sum_{r=1}^{k}\sum_{s=1}^{k} m_{rs} \log \frac{m_{rs}}{\kappa_r^{in}\kappa_s^{out}}, \tag{9}$$

where κ_r^{out} is the sum of out-degrees, and κ_r^{in} the sum of in-degrees of group r. For a more detailed derivation we refer the reader to [13] and [21].

C. Generating Synthetic Networks

Once a set of parameters is obtained, we can sample graphs from the respective distribution as follows:

1) Draw the number of edges e_{rs} between two groups:
   - SBM: $e_{rs} = \text{Poi}(\hat{M}_{rs} n_r n_s)$.
   - DCBM: $e_{rs} = \text{Poi}(\hat{M}_{rs})$.
   - If $r = s$ and graph undirected, multiply e_{rs} with $\frac{1}{2}$.
2) Draw tail and head from group r and s with:
   - equal probability for the SBM,
   - probability proportional to θ_i for the DCBM.

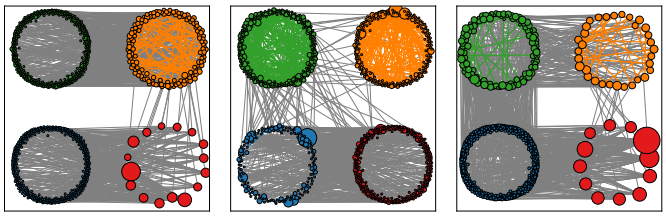
This approach generates a synthetic graph with time linear in the number of edges and the original number of nodes.

The so generated graphs may be unconnected. We therefore randomly rewire edges by taking an edge from a node of the biggest connected component with degree larger one, and place that edge on a node of a smaller component. The new and old end-point of the edge must be in the same group. The nodes are drawn according to the assumptions made by the model. Depending on the generated edges, the algorithm cannot merge all unconnected components, but works well in practice, as our results will show (Tab. I).

IV. Evaluation

We compare the generative capability of SBM, DCBM and Orbis. Additionally, we introduce the SBM+DCBM approach, which uses the SBM to infer group structure and the DCBM to generate synthetic graphs. SBM+DCBM uses the structure inferred with the SBM on the one hand, and the degree correction of the DCBM on the other hand. As graphs we use the HOT graph [23] and an IP-to-IP communication graph, obtained from traces of the Lawrence Berkeley National Laboratory (LBNL) [24] to which we will refer as LBNL-graph. For each graph, we firstly discuss the inferred structure of SBM and DCBM, secondly discuss differences between graphs generated by SBM and DCBM and, thirdly, compare our results to those obtained with Orbis. We choose Orbis as reference generator as it is also applicable to arbitrary graphs.

To quantify generative capabilities, we generate 100 synthetic graphs with each approach and calculate average value and standard deviation for different graph metrics on the largest connected component. We also consider the complementary cumulative distribution function (CCDF) of the node degree K for one synthetic graph per model. Generally, a good generator should yield synthetic graphs with metrics similar to those of the empirical graph, as different graph features encode different aspects of the graph [6], [25].



(a) SBM, $k = 4$ (b) DCBM, $k = 4$ (c) Orbis+SBM, $k = 4$

Fig. 3. Figures 3a and 3b visualize the HOT graph according to the inferred community structure. Figure 3c shows the structure inferred on a graph generated with Orbis. The size of the nodes is proportional to their logarithmic degree.

We estimate parameters for SBM and DCBM by inferring M, z and θ ten times per model on each graph for different values of k. We take the parameters yielding the largest likelihood for each value of k and visually inspect the inferred structure to select a value for k. Synthetic graphs with Orbis are generated using the joint degree distribution of each graph. Orbis generates synthetic graphs by randomly rewiring nodes, such that the given joint degree distribution is preserved.

A. Evaluation of the HOT-Graph

The HOT graph is distributed together with Orbis and accessible at [23]. Since router level topologies are more engineered and less random, the introduced models should be able to infer meaningful structure.

a) Inference: Fig. 3 visualizes the inferred structure by grouping nodes according to group membership. As Fig. 3a shows, the HOT graph nicely separates into four groups for the SBM. The lower left and upper left group contain leaf-nodes, whereas the lower right group serves as aggregation layer and the upper right group provides connectivity.

This is not the case for the structure inferred by the DCBM with $k = 4$ in Fig. 3b. Actually, the separation found by the SBM yields a larger likelihood when used in conjunction with the DCBM. Let z_{SBM} and z_{DCBM} be the inferred group membership for SBM and DCBM respectively, a DCBM using z_{SBM} is more likely to have generated the HOT graph than a DCBM with z_{DCBM}. The reason lies in the objective function and the algorithm used to find the parameter z. In case of the DCBM, it gets caught in a local optima.

b) Generation: Fig. 4 shows the CCDF of the degree distribution for the HOT graph and one synthetic graph per model. The x-Axis depicts the (node) degree K, and the y-axis the probability of $K \geq k$ for some $k \in \mathbb{N}$. The SBM has difficulties to generate the heavy tail of the degree distribution. As mentioned in Sec. III, nodes are stochastically equivalent given z. Nodes in the same group are thus equally likely to receive an edge, and therefore have the same expected degree. This flattens out the heterogeneous degrees of the lower right group in Fig. 3a and leads to the step visible in Fig. 4.

The DCBM fits the degree distribution better than the SBM, despite failing to capture the underlying structure. This is expected, as the DCBM models the expected degrees of the

TABLE I
GRAPH METRICS FOR SBM, DCBM AND ORBIS ON THE HOT GRAPH.

Model	Edges	Closeness	Assortativity Coefficient	Clustering Coef. $[e^{-3}]$	Path Length	Nodes	Knn	Degree	Coreness	Neighbour Degree
Original	988.00	0.15	-0.22	0.00	6.81	939.00	4.00	2.10	1.16	19.00
DCBM	**996.65**±20.97	0.12±0.02	-0.16±0.02	1.86±1.41	9.44±2.13	**939.00**±0.00	4.37±0.55	**2.12**±0.04	1.10±0.03	11.82±1.39
SBM	1003.78±21.96	0.12±0.01	-0.45±0.01	0.80±0.82	8.41±1.08	**939.00**±0.00	3.38±0.25	2.14±0.05	**1.17**±0.04	14.47±0.50
SBM+ DCBM	1012.24±27.82	**0.14**±0.01	-0.25±0.02	**0.79**±0.08	7.42±0.87	**939.00**±0.00	**4.03**±0.39	2.16±0.06	1.13±0.04	**20.13**±1.08
Orbis2k	822.11±20.57	0.16±0.00	**-0.24**±0.01	1.81±0.28	**6.23**±0.08	754.25±21.25	4.26±0.02	2.18±0.01	1.12±0.01	21.53±0.70

Average of respective graph feature plus/minus one standard-deviation calculated over 100 samples. Closest value is indicated in bold font.

TABLE II
GRAPH METRICS FOR SBM, DCBM AND ORBIS ON THE LBNL GRAPH.

Model	Edges	Closeness	Assortativity Coefficient	Clustering Coefficient	Path Length	Knn	Degree	Coreness	Neighbour Degree
Original	12329	0.19	-0.31	0.10	2.87	62.96	9.55	6.71	155
DCBM	12356.70±116.45	0.16±0.00	**-0.32**±0.00	0.07±0.03	2.10±0.04	47.49±1.15	9.58±0.09	**5.42**±0.06	**111.72**±2.47
SBM	**12334.71**±113.48	**0.17**±0.00	-0.57±0.01	0.02±0.00	**2.34**±0.04	39.22±1.20	**9.56**±0.09	5.21±0.05	65.19±1.33
SBM+DCBM	12293.33±116.0	0.15±0.00	-0.33±0.01	0.04±0.01	1.84±0.03	56.58±1.16	9.53±0.09	5.31±0.05	101.92±3.00
Orbis2k	14724±12.22	0.30±0.00	-0.34±0.00	**0.09**±0.00	3.38±0.01	**58.24**±0.32	11.41±0.01	8.56±0.01	226.86±0.75

Average of respective graph feature plus/minus one standard-deviation calculated over 100 samples. Closest value is indicated in bold font. All models generated graphs with 2581±0.00 nodes, which is the original value. We therefore omitted the column due to space.

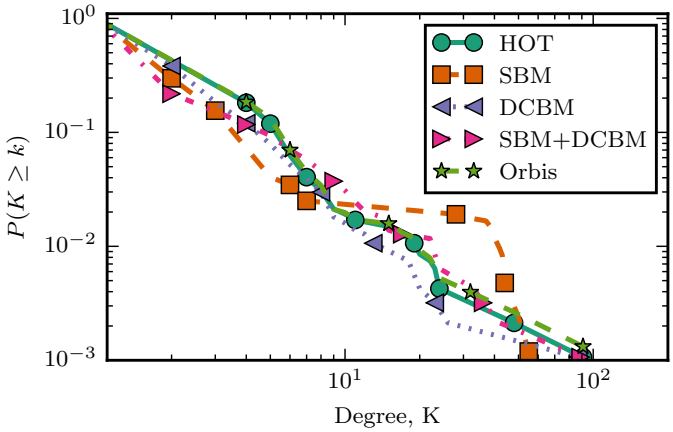


Fig. 4. CCDF of the degree distributions of HOT- and synthetic graphs.

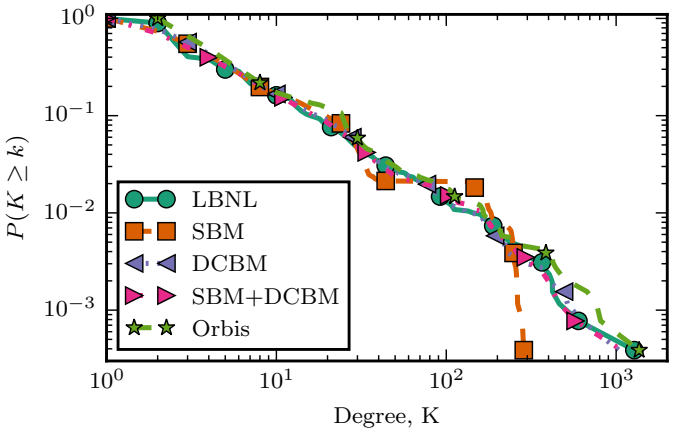


Fig. 5. CCDF of the degree distributions of LBNL- and synthetic graphs.

nodes and is thus able to better reproduce heavy tailed degree distributions. But also regarding other graph metrics in Table I, is the DCBM closer to the original graph. This follows from the insight of graph properties arising to a certain extend due to heterogeneous degrees.

As we have seen, the DCBM has problems inferring the underlying structure, while the SBM fails to model the heavy tail of the degree distribution. The SBM+DCBM combines the respective strengths, resulting in an even tighter fit to the original CCDF. Also, most of the graph features are closer to the original values (Tab. I). This highlights the importance of incorporating structural and degree information.

c) Comparison to Orbis: Orbis fits the degree distribution best (Fig. 4). As Orbis generates graphs with the same joint degree distribution, the degree distribution itself is also preserved. Despite the almost perfect match, the synthetic graphs generated with the combined approach still outperform Orbis in terms of graph features in Table I. The worse fit reflects the structural changes visible in Fig. 3c. The depicted structure is inferred using the SBM with $k = 4$ on the largest connected component of a synthetic graph generated by Orbis. The inferred structure is similar to the original one in Fig. 3a. However, the lower left and upper left group are now tightly connected, and the upper left group contains now inter-community edges. The structural changes are introduced as Orbis randomly rewires nodes and is constraint only by the given joint degree distribution. SBM and DCBM also randomly rewire nodes, however with respect to the inferred parameters, which encode the structure of the graph. Thus SBM and SBM+DCBM generate graphs with the structure shown in Fig. 3a. Depending on the use-case, graphs generated with the SBM might thus be preferable, despite lacking in the replication of statistical measures.

B. Evaluation of an IP-to-IP Communication Graph

The LBNL graph is obtained from traces of port-003 at the 10.09.2007 of the LBNL-dataset [24]. Each node in the graph represents an IP-address and each edge a source/destinaton-IP pair. The resulting graph is directed and unconnected. We take the largest connected component.

a) Inference: The LBNL Graph has a more complex structure than the HOT graph. Still, we find 10 groups ade-

quate to describe the structure of the graph. The large number of groups prohibits a visualization as in Fig. 3, as well as a detailed description of the groups.

The SBM finds a more reasonable split as the DCBM from a structural point of view. The DCBM accommodates high- and low degree nodes in one group. The inferred split is not unreasonable, but for the usage outlined in this paper, the split by the SBM is to be preferred. Contrary to earlier, the split found by the SBM does not lead to a larger likelihood when used with the DCBM. Thus, under the DCBM, the split of the SBM is less likely to generate the observed graph.

b) Generation: The SBM cannot generate the observed heavy tail as the CCDF in Fig. 5 shows. The DCBM and the SBM+DCBM approach have an almost perfect fit. Surprisingly, the SBM replicates graph features in Tab. II better than DCBM and SBM+DCBM.

c) Comparison with Orbis: SBM and DCBM reproduce the structure they inferred previously. Interestingly, Orbis also has slight problems with the heavy tail of the LBNL graph (Fig. 5) and is outperformed by the DCBM and SBM+DCBM. This may be due to Orbis being able to generate undirected graphs only, and thus being not able to handle asymmetric communication, which the SBM and DCBM with their directed variants can capture. This also explains the relative large number of edges and average neighbor degree for synthetic graphs generated by Orbis in Tab. II.

V. Conclusion and Future Work

In this work we show that probabilistic models, namely the Stochastic-Block-Model (SBM) and the Degree-Corrected-Block-Model (DCBM), greatly assist in the analysis of communication networks and their synthetic generation. The models are able to generate synthetic graphs from varying network domains. The inferred structure of the SBM results in a compressed, yet well interpretable model, capturing essential roles for nodes, as shown for a router-level graph, and preserving those roles in synthetically generated graphs. With respect to inference of structure, the SBM has proven itself superior to the DCBM in our set-up. However, the additional capabilities of the DCBM to model individual node degrees are important to generate graphs with observed heavy tailed degree distributions, as shown for an IP-to-IP communication graph. By joining the inference capabilities of SBM and the generative capabilities of DCBM, we show that the models outperform the Orbis network generator.

An open issue is the choice of the number of groups for the investigated models. If no prior knowledge exists, model selection using statistical and information theoretic measures could be employed [14]. Also more complex models exist, which might learn the number of groups from data, as well as vary the number of nodes in synthetic graphs [14], [21], [22].

Acknowledgment

This work has been performed in part in the framework of the CELTIC EUREKA project SENDATE-PLANETS (Project ID C2015/3-1) and is partly funded by the German BMBF (Project ID 16KIS0473), and in part in the framework of the EU project FlexNets funded by the European Research Council under the European Unions Horizon 2020 research and innovation program (grant agreement No 647158 - FlexNets). The authors alone are responsible for the content of the paper.

References

[1] S. Zhou, "Characterising and modelling the internet topology – The rich-club phenomenon and the PFP model," *BT Technology Journal*, vol. 24, no. 3, pp. 108–115, 2006.

[2] K. Kim et al., "Effect of Homophily on Network Evolution," Seoul National University; Technology Management, Economics, and Policy Program (TEMEP), Tech. Rep., 2015.

[3] A. Medina et al., "Brite: An approach to universal topology generation," ser. MASCOTS '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 346–.

[4] P. Mahadevan et al., "Systematic topology analysis and generation using degree correlations," in *ACM SIGCOMM CCR*, vol. 36. ACM, 2006, pp. 135–146.

[5] ——, "Orbis: rescaling degree correlations to generate annotated internet topologies," in *ACM SIGCOMM CCR*, vol. 37. ACM, 2007, pp. 325–336.

[6] G. Accongiagioco, E. Gregori, and L. Lenzini, "S-BITE: A Structure-Based Internet Topology gEnerator," *Computer Networks*, vol. 77, pp. 73–89, Feb. 2015.

[7] K. L. Calvert et al., "Modeling internet topology," *IEEE Communications magazine*, vol. 35, no. 6, pp. 160–163, 1997.

[8] Y. Hu, F. Zhang, K. K. Ramakrishnan, and D. Raychaudhuri, "GeoTopo: A PoP-level Topology Generator for Evaluation of Future Internet Architectures," in *ICNP*, Nov. 2015, pp. 90–99.

[9] R. Rossi, S. Fahmy, and N. Talukder, "A multi-level approach for evaluating internet topology generators," in *Proc. IFIP Networking Conference*, May 2013, pp. 1–9.

[10] M. Iliofotou et al., "Profiling-by-association: A resilient traffic profiling solution for the internet backbone," ser. Co-NEXT '10. New York, NY, USA: ACM, 2010, pp. 2:1–2:12.

[11] W. Aiello et al., "Analysis of communities of interest in data networks," in *International Workshop on Passive and Active Network Measurement*. Springer, 2005, pp. 83–96.

[12] A. Jakalan et al., "Profiling IP hosts based on traffic behavior," in *Proc. IEEE ICCSN*. IEEE, 2015, pp. 105–111.

[13] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, p. 016107, 2011.

[14] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, p. 011047, 2014.

[15] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[16] M. Andrews et al., "Clustering and server selection using passive monitoring," in *Proc. IEEE Infocom*, vol. 3. IEEE, 2002, pp. 1717–1725.

[17] B. M. Waxman, "Routing of multipoint connections," *Selected Areas in Communications, IEEE Journal on*, vol. 6, no. 9, pp. 1617–1622, 1988.

[18] R. Albert and A.-L. Barabsi, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[19] J. Winick and S. Jamin, "Inet-3.0: Internet topology generator," Tech. Rep.

[20] X. Zhang, T. Martin, and M. E. J. Newman, "Identification of core-periphery structure in networks," *CoRR*, vol. abs/1409.4813, 2014.

[21] Y. Zhu, X. Yan, and C. Moore, "Oriented and degree-generated block models: Generating and inferring communities with inhomogeneous degree distributions," *CoRR*, vol. abs/1205.7009, 2012.

[22] T. Herlau, M. N. Schmidt, and M. Mrup, "Infinite-degree-corrected stochastic block model," *pre*, vol. 90, no. 3, p. 032819, Sep. 2014.

[23] P. Mahadevan. Analyzing and generating network topologies with orbis. [Online]. Available: http://www.sysnet.ucsd.edu/~pmahadevan/topo_research/topo.html

[24] Enterprise tracing project. [Online]. Available: http://www.icir.org/enterprise-tracing/

[25] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, Mar. 2010.