Modeling IP-to-IP Communication using the Weighted Stochastic Block Model

Patrick Kalmbach, Lion Gleiter, Johannes Zerwas, Andreas Blenk, Wolfgang Kellerer
Technical University of Munich

Stefan Schmid
University of Vienna

ABSTRACT

The vision of self-driving networks integrates network measurements with network control. Processing data for each of the tasks comprising network control separately might be prohibitive due to the large volume and waste of computational resources. In this work we make the case of using the Weighted Stochastic Block Model (WSBM), a probabilistic model, to learn a task independent representation. In particular, we consider a case study of real-world IP-to-IP communication. The learned representation provides higher level-features for traffic engineering, anomaly detection, or other tasks, and reduces their computational effort. We find that the WSBM is able to accurately model traffic and structure of communication in the considered trace.

CCS CONCEPTS

• **Networks** → **Network monitoring**;

KEYWORDS

Network Monitoring; Stochastic Block Model; Data Analysis

1 INTRODUCTION

The Context: Emerging self-driving Networks. Self-driving networks [3] integrate data measurement with network control and rely on data analytics and learning to combat the increasing difficulties in network management in today's and future complex networks.

The Problem: Task diversity. Network control features different tasks: network design, traffic engineering, capacity planning or anomaly detection [8]. Tasks that are becoming even more complex due to the trend towards dynamic and flexible network fabrics. Processing raw flow-data for each task might be prohibitive due to its volume. Having specialized systems for each task might prove a brittle solution, due to their sensitivity towards changes, and hinder the transfer of knowledge [6]: Features engineered for an educational data-center network might not hold for a cloud data-center. Hence, an abstraction is needed that can be adapted to a particular system and used for different tasks.

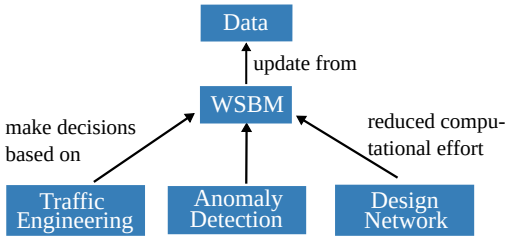


Figure 1: The WSBM is intended to serve as a compressed representation for different network control tasks and is updated from measurements.

The Opportunity: Probabilistic Models (PMs). PMs leverage domain knowledge for rough guidelines, while details are filled in by fitting the model to data [6]. PMs can handle uncertain measurements and answer questions concerning network control: What is the expected future traffic?, What is the expected traffic for a specific node?, behaves this host strange?. We investigate the representational capabilities of the Weighted Stochastic Block Model (WSBM) [1], which is a PM for relational data. Fig. 1 illustrates how the WSBM can be used in the context of self-driving networks. The WSBM is updated with measurements, and reduces the computational effort for different control tasks by providing a compact representation.

Contribution. This poster shows how the WSBM can be used to create a model of an educational data-center network. Using real world traces we investigate how the WSBM captures the network structure and evaluate how well it is able to capture the traffic volume characteristics.

Background: The WSBM. The WSBM separates nodes $\mathcal{N}$ of a graph into k groups and models the connectivity and

SIGCOMM Posters and Demos '18, August 20–25, 2018, Budapest, Hungary

ACM ISBN 978-1-4503-5915-3/18/08…$15.00
https://doi.org/10.1145/3234200.3234245

weights of edges between groups. Since usually $k << | \mathcal{N} |$ a compressed representation is obtained.

Connectivity is modeled as a Bernoulli distribution and edge weights with a distribution from the exponential family. Parameters of the WSBM are: number of groups k, node to group assignment z, probabilities $\theta_e(r, s)$ and parameters for the distribution of weights $\theta_w(r, s)$ of edges between nodes in groups r, s. Given a graph, k and a choice for θ_w, variational inference is used to find the most likely group assignment z, and the parameterization for θ_e and θ_w. A parameter $\alpha \in [0, 1]$ controls the importance of weights and structure during inference. For $\alpha = 0$ the WSBM focuses on the correct modeling of the edge weights, and for $\alpha = 1$ the WSBM becomes the Stochastic Block Model (SBM) [7].

Compared to heuristic graph partitioning approaches (e.g. modularity) allows the WSBM: to reason about future or missing data, to make probabilistic statements about observations, or to indicate the quality of the data description over likelihood scores [6]. Those advantages come with a more complex algorithmic procedure to fit the model to data [1].

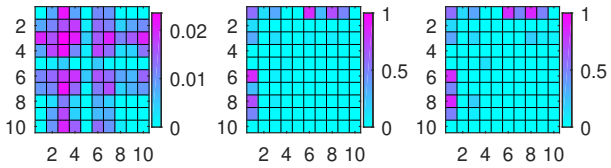
Related Work. Grouping nodes and modeling inter- and intra group connectivity is related to traffic matrix (TM) estimation. Authors in [5] use SNMP link counters and inference methods from Internet Service Provider Networks to obtain Top-of-Rack switch level TMs. Authors in [8] take a similar approach and leverage the structure of data center topologies to cluster switches and obtain an efficient tomography algorithm. The WSBM does not assume a specific network topology, and relies on flow-level information.

The SBM has been used to generate synthetic network topologies [4] and to identify bots [7]. We generalize previous work by additionally considering edge weights.

2 PROBLEM AND APPROACH

We want to obtain a model of a data-center network using the WSBM that can answer questions about the structure and amount of communication in the network. We create a Traffic Dispersion Graph (TDG) from a time window with duration Δt and model the TDG as a weighted directed Graph, where the edge weights are the sum of transmitted bytes. We then infer the parameters of the WSBM using variational inference [1]. We take the Log-Normal distribution as possible edge weight distributions θ_w [2, 5]. The number of groups k is selected using the Bayes Factor [1].

To evaluate how well the WSBM models edge weights we use the Kullback-Leibler (KL) Divergence between original weights and weights generated from the WSBM. The smaller the KL-divergence the closer the distributions and the better the representation. The baseline is the KL-divergence between observed- and generated edge weights from the chosen edge weight distribution fitted to all values.



(a) $\alpha = 0$ **(b)** $\alpha = 0.5$ **(c)** $\alpha = 1$

Figure 2: Probability of edges between groups for three values of α. Note the change in the colorbar.

3 EVALUATION

We used the data-center trace UNIV2 [2], and took the first 5 minutes to create a TDG. We used values in $\{0, 0.5, 1\}$ for α, and chose the Number of groups k from $\{i\}_{i=2}^{10}$, resulting in a value of $k = 10$. A maximal-value of 10 was selected based on insights about the structure gained during the project. To assess the quality of the edge weights we generated 100 random graphs from the WSBM and report each KL-divergence.

Group Structure. Fig. 2 shows the probability of edges between inferred groups. For $\alpha = 0$ Fig. 2a shows that the probability of intra- and inter group edges is with at most 0.025 very low. Almost all groups connect to many other groups, which is different from the structure obtained with $\alpha = 0.5$ and $\alpha = 1$ visible in Fig. 2b and Fig. 2c. The inferred structure for $\alpha = 0.5$ and $\alpha = 1$ are very similar: Sparse connectivity featuring one central group with 10 resp. 11 nodes, to which most other groups connect with high probability. This structure reflects the client-server architecture that underlies the applications in the UNIV2 trace.

Edge Weights: The smallest KL-divergence with mean 0.03 is obtained for the WSBM with $\alpha = 0$. The WSBM with $\alpha = 0.5$ results in a mean of 0.08, and for $\alpha = 1$ in a mean of 0.16. This is close to the baseline having an average value of 0.18. With $\alpha = 0.5$ the WSBM strikes a balance between inferring a reasonable structure and modeling edge weights.

4 NEXT STEPS

This paper shows that WSBMs can capture the communication structure and model the traffic distribution of a data-center network. WSBMs are useful for various decision making tasks relying on network representations: global routing decisions, anomaly detection etc. At the current time we investigate the groups inferred by the WSBM more closely, and plan to implement the WSBM in a testbed and utilize it for traffic engineering and anomaly detection.

ACKNOWLEDGEMENT

This work is part of a project that has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (grant agreement No 647158 - FlexNets).

REFERENCES

[1] C. Aicher, A. Z. Jacobs, and A. Clauset. 2014. Learning Latent Block Structure in Weighted Networks. *ArXiv e-prints* (April 2014). arXiv:stat.ML/1404.0431

[2] Theophilus Benson, Aditya Akella, and David A. Maltz. 2010. Network Traffic Characteristics of Data Centers in the Wild. In *Proc. ACM IMC*. ACM, New York, NY, USA, 267–280.

[3] Nick Feamster and Jennifer Rexford. 2017. Why (and How) Networks Should Run Themselves. *CoRR* abs/1710.11583 (2017).

[4] P. Kalmbach, A. Blenk, M. Kluegel, and W. Kellerer. 2017. Generating synthetic Internet- and IP-topologies using the Stochastic-Block-Model. In *2017 IFIP/IEEE IM*. 911–916.

[5] Srikanth Kandula, Sudipta Sengupta, Albert Greenberg, Parveen Patel, and Ronnie Chaiken. 2009. The Nature of Data Center Traffic: Measurements & Analysis. In *Proc. ACM IMC*. ACM, New York, NY, USA, 202–208.

[6] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

[7] Patrick Kalmbach, Andreas Blenk, Wolfgang Kellerer, and Stefan Schmid. 2018. Themis: A Data-Driven Approach to Bot Detection. In *IEEE INFOCOM*. Honolulu, Hawaii, USA.

[8] Yan Qiao, Zhiming Hu, and Jun Luo. 2013. Efficient traffic matrix estimation for data center networks. In *2013 IFIP Networking*. 1–9.