AwareNet: Using WSBMs for Network Traffic Analyis

Maximilian Stephan
maximilian.stephan@tum.de
Technical University of Munich
Munich, Germany

Patrick Krämer
patrick.kraemer@tum.de
Technical University of Munich
Munich, Germany

Wolfgang Kellerer
wolfgang.kellerer@tum.de
Technical University of Munich
Munich, Germany

ABSTRACT

Characterizing network behavior, an essential building block for many network management tasks becomes increasingly difficult for administrators. Reasons for that are, among others, rising traffic volume and dynamicity in modern networks. To automate network behavior characterization, we present AwareNet, a system that uses Weighted Stochastic Block Models (WSBMs). By providing insight into network-inherent dynamics AwareNet supports administrators in handling the rising traffic volume and dynamicity. As an example, we show that AwareNet can detect targeted host scans in a campus network.

CCS CONCEPTS

• **Networks** → **Network monitoring**; *Network management*.

KEYWORDS

network traffic analysis, stochastic block model, anomaly detection

ACM Reference Format:
Maximilian Stephan, Patrick Krämer, and Wolfgang Kellerer. 2022. AwareNet: Using WSBMs for Network Traffic Analyis. In *CoNEXT Student Workshop 2022 (CoNEXT-SW '22), December 9, 2022, Roma, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3565477.3569158

1 INTRODUCTION

Characterizing network behavior becomes increasingly complex because of increases in network traffic volume and dynamicity due to a rising number of network components. This poses a challenge to network management that heavily relies on manual identification of network-inherent dynamics and subsequent decision-making. To allow network growth beyond these limitations in network management and support administrators in their task, the networks have to autonomously gain an understanding of themselves.

We present AwareNet that uses the language of probability theory to formulate a probabilistic model for network behavior and data analysis to fit the model to the data of a network at hand. The fitted model provides insights about the network to the administrator and enables AwareNet to make autonomous decisions. Further, AwareNet uses the model to detect suspicious changes in the communication of nodes. This early study deploys AwareNet in a campus network, lets AwareNet learn an internal representation of the communication, and uses this representation to detect scanning attacks that we initiate.

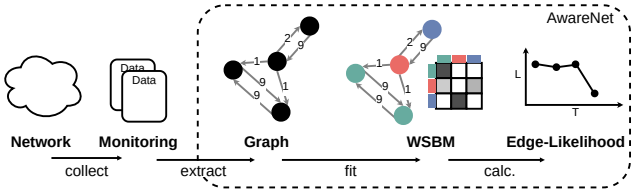


Figure 1: AwareNet creates a graph from network monitoring data, fits a WSBM to this graph, and uses the fitted WSBM to characterize network behavior.

2 BACKGROUND AND RELATED WORK

Probabilistic models of networks describe the communication in a network with probability theory. Examples of such models are Stochastic Block Models (SBMs). SBMs are latent variable models representing (weighted) graphs that assign nodes to individual groups [4]. The groups two nodes belong to, then govern their communication behavior. Given a (weighted) graph, probabilistic inference can discern the most likely parameters of the SBM. The fitted SBM enables the generation of synthetic communication patterns as well as analysis of observed communication. In this context, Kalmbach [2] uses WSBMs to replicate data center network traffic in an offline setting, while NOracle [3] uses unweighted SBMs to detect malware in a testbed network. In contrast to [3], AwareNet can detect anomalies like targeted host scans in a real campus network. Port scans are often part of network attack's probing phases [6]. In contrast to [2], AwareNet operates online and tunes edge weights towards anomaly detection.

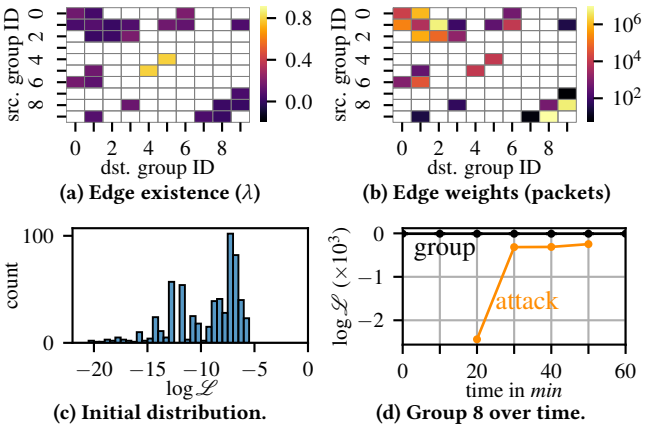
Eltanbouly [1] gives a general overview of approaches for network anomaly detection. AwareNet's unsupervised nature and explainable model properties pose an advantage over listed approaches that require labeled data or black-box models.

3 ARCHITECTURE

Fig. 1 shows the architecture of AwareNet. Given a monitoring system in place that provides flow level data, AwareNet uses 10 min slices of monitoring data to construct a graph of the communication. This graph consists of IP addresses as nodes and edges corresponding to data flows occurring within the 10 min interval. Additionally, edges are annotated with statistics about the flow, e.g., the number of packets. Then, AwareNet fits the parameters of a WSBM using the GraphTool [5] library. The result of the fit is a partition z of all nodes in the graph into k groups, depicted by node coloring in Fig. 1. The partition itself can already facilitate a better understanding of network-inherent structures by pairing nodes with similar communication behavior into the same group. Further, the statistics about

CoNEXT-SW '22, December 9, 2022, Roma, Italy

ACM ISBN 978-1-4503-9937-1/22/12.
https://doi.org/10.1145/3565477.3569158



(a) Edge existence (λ) (b) Edge weights (packets)

(c) Initial distribution. (d) Group 8 over time.

Figure 2: WSBM parameters (a,b), edge log $\mathcal{L}$ (c,d)

edge existence and edge weight distribution on a group-to-group level, depicted by grey-scale colors in the matrix in Fig. 1, qualify for calculating edge likelihoods. Given the fitted WSBM model θ, we calculate the log-likelihood of an edge e_{ij} between node i and j with weight w_{ij} as follows:

$$\log \mathcal{L}(e_{ij}|\theta) = \log \text{Pois}_{\lambda_{z_i,z_j}}(1) + \log \mathcal{N}_{\mu_{z_i,z_j},\sigma_{z_i,z_j}}(w_{ij}) \quad (1)$$

In the first part of Eq. 1, we assume a Poisson distribution for edge existence. The related parameter λ is the ratio of observed edges from the group of node i, z_i to the group of node j, z_j and the number of possible edges between the two groups, $|z_i| \cdot |z_j|$. In the second part of Eq. 1, we assume a normal distribution for the edge weights. The parameters μ and σ are the related maximum likelihood estimates given all weights for edges from group z_i to group z_j. Lastly, we assume independence between edge existence and weight. Therefore the total edge log-likelihood given the fitted model is the sum of log-likelihoods for existence and weight. AwareNet applies this approach to subsequent traffic observations.

After the model fit with k groups and n nodes AwareNet has a memory complexity of $O(k^2 + n)$. This reflects the group-to-group relations stored as matrices and the node-to-group mapping. The computational complexity for calculating edge log-likelihoods depends linearly on the number of edges in subsequent observations.

4 PRELIMINARY EVALUATION

We show first results of how AwareNet presents network traffic. In the scenario at hand, AwareNet gets data from a live university campus network with 250 nodes. For data privacy reasons IP Addresses are pseudonymised before usage. Additionally, users are aware of the traffic monitoring system in place. All results describe host behavior on a group level to further protect privacy of individual users. The only exception are hosts specifically created for the experiment described later in this section.

Fig. 2 (a) and (b) represent the initial fit into 10 groups of nodes based on 10 min of data. Fig. 2a shows a heatmap for the parameter λ from Eq. 1. The blank cells represent the case, that no outgoing edges were observed from the source group to the destination group. For the colored cells, a lighter color means a higher λ, and therefore an increased likelihood for edges of the source-destination group pair. Similarly, Fig. 2b shows a heatmap for the parameter μ from Eq. 1. In this case, edge weights are the number of packets. The colors encode the average number of packets seen on edges from the src. group to the dst. group. By not only considering "if" communication was observed between nodes but also "how much", AwareNet takes a holistic approach to model a network's behavior.

Fig. 2c displays the distribution of edge log-likelihoods for the initial data. All the values fall between -5 and -21. This establishes a notion of which values are expected given the model and the initial data. After the model fit, AwareNet calculates the edge log-likelihoods for subsequent 10 min blocks of monitoring data. To artificially create an anomaly a host from group 8 starts a `nmap` host-scan targeting another node. Both hosts are specifically deployed for this experiment, i.e., the experiment does not impair the functioning or security of other nodes in the network. Fig. 2d shows the edge log-likelihoods from source group 8 over time. The black line represents the average values of all "normal" edges. It generally stays above -21, the minimum from Fig. 2c. In contrast, log-likelihood values for the "attack" edge, shown in orange, drop to under -2k and never go higher than -300. This showcases that AwareNet is capable of detecting targeted scans and thereby opens up possibilities for further use of probabilistic models in network traffic analysis.

5 SUMMARY AND FUTURE WORK

This paper presents AwareNet, a system that brings awareness about the structure and amount of communication to networks. AwareNet creates an internal representation of the communication in the network. This representation offers insights regarding the network structure to network administrators and enables the detection of anomalies like targeted host scans.

Based on our initial findings, we plan to do an extensive study about what kind of anomalies can be detected under which circumstances. This includes more complex attack patterns and the involvement of previously unseen hosts. Further, we seek to evaluate how the WSBM's parameters change over longer observation periods, and model these changes. With such information, we are confident that AwareNet can model a broader set of network behaviors and thus become even more useful to administrators.

ACKNOWLEDGMENTS

This work is partially funded by Federal Ministry of Education and Research in Germany (BMBF) as part of the project AI-NET-PROTECT (grant ID 16KIS1294), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 438892507.

REFERENCES

[1] Sohaila Eltanbouly, May Bashendy, Noora AlNaimi, Zina Chkirbene, and Aiman Erbad. 2020. Machine Learning Techniques for Network Anomaly Detection: A Survey. In *ICIoT*.
[2] Patrick Kalmbach, Lion Gleiter, Johannes Zerwas, Andreas Blenk, and Wolfgang Kellerer. 2018. Modeling IP-to-IP communication using the Weighted Stochastic Block Model. In *SIGCOMM Posters and Demos*.
[3] Patrick Kalmbach, Fabian Lipp, David Hock, Wolfgang Kellerer, and Andreas Blenk. 2019. NOracle: Who is communicating with whom in my network?. In *SIGCOMM Posters and Demos*.
[4] Brian Karrer and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical review E* 83, 1 (2011).
[5] Tiago P. Peixoto. 2014. The graph-tool python library. *figshare* (2014). http://figshare.com/articles/graph_tool/1164194
[6] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes, and Andreas Hotho. 2019. A survey of network-based intrusion detection data sets. *Computers & Security* (2019).