Anomaly Detection in Partially Observed Traffic Networks

Elizabeth Hou, *Student Member, IEEE,* Yasin Yılmaz, *Member, IEEE,* and Alfred O. Hero, *Fellow, IEEE*

***Abstract*—This paper addresses the problem of detecting anomalous activity in traffic networks where the network is not directly observed. Given knowledge of what the node-to-node traffic in a network should be, any activity that differs significantly from this baseline would be considered anomalous. We propose a Bayesian hierarchical model for estimating the traffic rates and detecting anomalous changes in the network. The probabilistic nature of the model allows us to perform statistical goodness-of-fit tests to detect significant deviations from a baseline network. We show that due to the more defined structure of the hierarchical Bayesian model, such tests perform well even when the empirical models estimated by the EM algorithm are misspecified. We apply our model to both simulated and real datasets to demonstrate its superior performance over existing alternatives.**

***Index Terms*—anomaly detection, latent variable model, EM algorithm, minimum relative entropy, hypothesis testing**

I. Introduction

In today's connected world, communication is increasingly voluminous, diverse, and essential. Phone calls, delivery services, and the Internet are all modern amenities that send massive amounts of traffic over immense networks. Thus network security, such as the ability to detect network intrusions or illegal network activity, plays a vital role in defending these network infrastructures. For example, (i) computer networks can protect themselves from malware such as botnets by identifying unusual network flow patterns; (ii) supply chains can prevent cargo theft by monitoring the schedule of shipments or out-of-route journeys between warehouses; (iii) law enforcement agencies can uncover smuggling operations by detecting alternative modes of transporting goods.

Identifying unusual network activity requires a good estimator of the true network traffic, including the anomalous activity, in order to distinguish it from a baseline of what the network should look like. However, often it is not possible to for an external observer to observe the network directly due to constraints such as cost, protocols, or legal restrictions. This makes the problem of estimating the rate of traffic between nodes in a network difficult because the edges between nodes are latent unobserved variables. Network tomography approaches have been previously proposed for estimating network topology or reconstructing link traffic from incomplete measurements and limited knowledge about network connectivity. However for network anomography, the detection of anomalous deviations of traffic in the network, highly accurate estimation of all network traffic may not be necessary. It often suffices to detect perturbations within the network at an aggregate or global scale. This paper addresses the problem of network anomography rather than that of network tomography or traffic estimation.

This work was supported in part by the Consortium for Verification Technology under Department of Energy National Nuclear Security Administration award number DE-NA0002534, in part by the University of Michigan ECE Departmental Fellow, in part by the U.S. National Science Foundation (NSF) under grant CNS-1737598, and in part by the Southeastern Center for Electrical Engineering Education (SCEEE) under grant SCEEE-17-03.

Elizabeth Hou and Alfred Hero are with the EECS Department at University of Michigan, Ann Arbor, MI, Contact email: emhou@umich.edu

Yasin Yılmaz was with the EECS Department at the University of Michigan, Ann Arbor, MI. He is now with the Electrical Engineering Department at the University of South Florida, Tampa, FL.

A. Related Work

Broadly defined, the network tomography problem is to reconstruct complete network properties, e.g., source-destination (SD) traffic or network topology, based on incomplete data. The term "network tomography" was introduced in [1] where the objective is to estimate unknown source destination traffic intensities given observations of link traffic and known network topology. Since the publication of [1], the scope of the term network tomography has been used in a much broader sense (see the review papers [2], [3], [4], and [5]). For example, a variety of passive or active packet probing strategies have been used for topology reconstruction of the Internet, including unicast, multicast, or multi-multicast [6], [7], and [8]; or using different statistical measures including packet loss, packet delay, or correlation [9], [10], [11], and [12].

In the formulation of [1], the network tomography objective is to determine the total amount of traffic between SD pairs given knowledge of the physical network topology and the total amount of traffic flowing over links, called the link data. This leads to the linear model for the observations $\boldsymbol{y}^t = \boldsymbol{A}\boldsymbol{x}^t$ where $\boldsymbol{A}$ is the known routing matrix defining the routing paths, and at each time point t, $\boldsymbol{y}^t$ is a vector of the observed total traffic on the links and $\boldsymbol{x}^t$ is a vector of the unobserved message traffic between SD pairs. Using the model that the elements of $\boldsymbol{x}^t$ are independent and Poisson distributed, an expectation-maximization (EM) maximum likelihood estimator (MLE) and a method of moments estimator are proposed in [1] for the Poisson rate parameters $\boldsymbol{\lambda}$. The authors of [13] propose a Bayesian conditionally Poisson model, which uses a Markov chain Monte Carlo (MCMC) method to iteratively draw samples from the joint posterior of $\boldsymbol{\lambda}$ and $\boldsymbol{x}$. The authors of [14] and [15] assume the message traffic is instead from a Normal distribution, obtaining a computationally simpler estimator of the SD traffic rates. The authors of [16] relax the assumption that the traffic is an independent and identically Poisson distributed sequence and instead consider the network

as a directly observable Markov chain. Under this weaker assumption, they derive a threshold estimator for the Hoeffding test in order to detect if the network contains anomalous activity.

In [17] the authors propose an EM approach for Poisson maximum likelihood estimation when the network topology is unknown; however, their solution is only computationally feasible for very small networks and it does not account for observations of traffic through interior nodes. This has led to simpler and more scalable solutions in the form of gravity models where the rate of traffic between each SD pair is modeled by $x_{sd} = (N_s N_d)/N$ where N_s and N_d are the total traffic out of the source node and into the destination node respectively and N is the total traffic in the network. Standard gravity models do not account for the interior nodes, thus in [18] and [19] tomogravity and entropy regularized tomogravity models were proposed, which incorporate the interior node information in the second stage of their algorithm. The authors of [20] generalize the tomogravity model from a rank one (time periods are independent) to a low rank approximation (time periods are correlated) and allow additional observations on individual SD pairs. Similarly, the authors of [21] and [22] use a low rank model with network traffic maps to incorporate a sparse anomaly matrix, and they solve their multiple convex objectives with the alternating direction method of multipliers (ADMM) algorithm.

Dimensionality reduction has also been used directly for anomaly detection in the SD traffic flows in networks. Under the assumption that traffic links have low rank structure, the authors in [23] and [24] use Principle Component Analysis (PCA) to separate the anomalous traffic from the nominal traffic. This low rank framework is generalized to applying PCA in networks that are temporally low rank or have dynamic routing matrices, in [25]. The authors of [25] also coin the term "network anomography" to reflect the influence of network topology reconstruction, which is a necessary component to detecting anomalies in a network with unknown structure. However, later work in [26] discusses the limitations of PCA for detecting anomalous network traffic, e.g., it is sensitive to (i) the choice of subspace size; (ii) the way traffic measurements are aggregated; (iii) large anomalies. The low rank plus sparse framework is extended to online setting with a subspace tracking algorithm in [27].

Specifically for Internet Protocol (IP) networks, some works prefer to perform anomaly detection on the flows from the IP packets instead of the SD flows. The authors of [28] use PCA to separate the anomalous and nominal flows from sketches (random aggregations of IP flows) while the authors of [29] model the sketches as time series and detect change points with forecasting. The works of [30] and [31] also perform change point detection using windowed hypothesis testing with generalized likelihood ratio or relative entropy respectively.

Because our approach in this paper is based on traffic networks or SD models, these types of approaches were the focus of our related works subsection. However, networks can also be represented as graph models or as features of the network characteristics. This subsection would be incomplete if it did not mention anomaly detection approaches to other types of network models. So, we refer to some survey papers that cover many of the recent techniques in graph based approaches: [32] and [33]. In particular, similar to the low rank approaches for SD networks, there are low rank approaches to graph models such as [34] who assume the inverse covariance matrix of their wireless sensor network data has a graph structure and solve a low rank penalized Gaussian graphical model problem and [35] who impose graph smoothness by a low rank assumption on graph Laplacian of the features of the network. [36] also uses a low rank approach on their KDD intrusion data set, but they directly apply the low rank assumption to the network characteristics of their data.

B. Our Contribution

In this paper, we consider networks where an exterior node (a node in an SD pair) only transmits and receives messages from a few other nodes, but because, as an external observer (one that is not located on a node), we cannot observe network directly, we do not know which SD pairs have traffic and which do not. Thus, we develop a novel framework to detect anomalous traffic in sparse networks with unknown sparsity pattern. Our contributions are the following. 1) In order to estimate the network traffic, we propose a parametric hierarchical model that alternates between estimating the unobserved network traffic and optimizing for the best fit rates of traffic using the EM algorithm. 2) We warm-start the algorithm with the solution to non-parametric minimum relative entropy model that directly projects the rates of traffic onto the nearest attainable sparse network. 3) Since we do not make assumptions of fixed edge structure in our model, it allows us to accommodate the possibility of anomalous edges in the actual network structure because anomalies will never be known in advance. 4) Using our probabilistic model's estimator of actual traffic rates, we test for anomalous network activity by comparing it to a baseline to determine which deviations are anomalies and which are estimation noise. We develop specific statistical tests, based on the generalized likelihood ratio framework, to control for the false positive rate of our probabilistic model, and show that even when our models are misspecified, our tests can accurately detect anomalous activity in the network.

The rest of the paper is organized in the following way. Section II proposes a problem formulation of the network we are interested in and our assumptions about it. Section III describes our proposed hierarchical Bayesian model, which is solved with a generalized EM algorithm and warm-starting the EM with a solution that satisfies the minimum relative entropy principle. Section IV describes our anomaly detection scheme through statistical goodness of fit tests and Section V describes the computational complexity of our method. Section VI contains simulation results of the performance of our proposed estimators and applications to the CTU-13 dataset of botnet traffic and a dataset of NYC taxicab traffic. Finally, Section VII concludes the paper.

II. Proposed Formulation

We give a simple diagram of a notional network in Fig. 1(a). An exterior node, V_i, sends messages, N_{ij}^t, at a rate, Λ_{ij}, to

another exterior node, V_j, at each time point, t. Messages can flow through interior nodes, such as U_1, but the interior nodes do not absorb or create messages. Because the magnitude of flow is just the total number of messages that have been sent from one node to another, network traffic between nodes is a counting process. For tractability, it is common to assume the messages are independent and identically distributed (i.i.d.) and the total number of messages in a time period is from some parametric distribution. The Poisson distribution is the most natural choice because it models events occurring independently with a constant rate, and it is used by [1], [17], [13], [14], and [15] although the latter two works use a Normal approximation to the Poisson for additional tractability.

When the network is observed directly, the edge structure and rates can be easily estimated using a sample of observations at different time points. Under these Poisson process assumptions, the uniformly minimum variance unbiased estimator is simply the maximum likelihood estimator (MLE) of the Poisson distribution. However, this is a very strong and unrealistic assumption because it implies that we, as an external observer, are able to track every single message being passed in the network. Thus, we are interested in the much weaker assumption that we can only monitor the nodes themselves. Fig. 1(b) shows what we can actually observe from the network under this weaker assumption. While we also observe the total amount of traffic, unlike in [1], we do not know the network topology.

Since we can only monitor the nodes, we can only observe the total ingress and egress of the exterior nodes. Thus we know an exterior node, V_i, transmits $N_{i\cdot}^t$ messages and receives $N_{\cdot i}^t$ messages, but we do not know which of the other nodes it is interacting with. We can also observe the flow through interior nodes, but we cannot distinguish where the messages come from or are going to. For instance, in Fig. 1(a), an interior node, such as U_1, will observe all messages, $F_1^t = N_{14}^t + N_{2P}^t$, that flow through it, but it will not be able to distinguish the number of messages from each SD pair or whether all the SD pairs actually send messages.

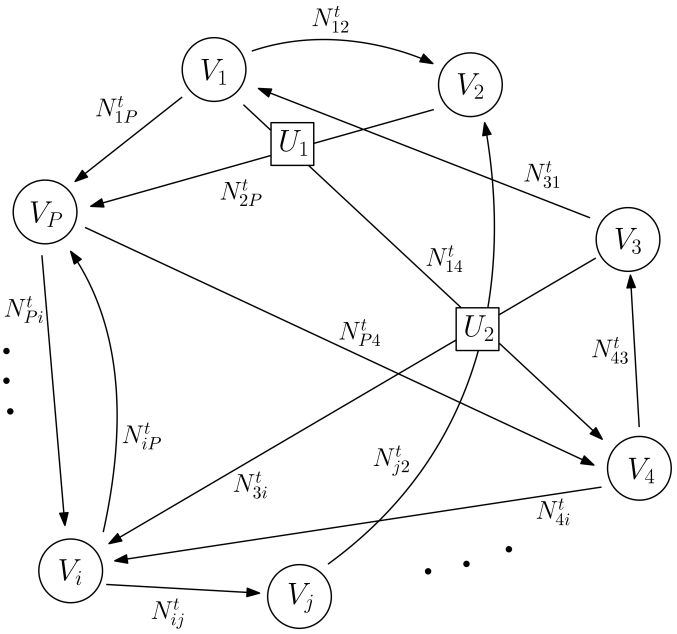
A network with P exterior nodes can naturally be mathematically formulated as a $P \times P$ matrix, which is observed T times. Let $\boldsymbol{N}^t$ be the unobserved traffic matrix at time instance t and let the elements of the matrix, N_{ij}^t, be the amount of traffic between nodes i and j. The row and column sums of the traffic are denoted by $\boldsymbol{R} = [N_{1\cdot} \dots N_{P\cdot}]'$ and $\boldsymbol{C} = [N_{\cdot 1} \dots N_{\cdot P}]'$ respectively, and $\boldsymbol{F} = [F_h]$ are the observed flows through interior nodes, which are indexed by h. The traffic at each time instance t is generated from a distribution with mean $\boldsymbol{\Lambda}$, the true intensity/rate parameter of the matrix, and $\boldsymbol{\Lambda_0}$ is the baseline parameter of a network without any anomalies. This mathematical formulation is shown below.

$$\boldsymbol{N}^t = \begin{bmatrix} 0 & N_{12}^t & N_{13}^t & \cdots & N_{1P}^t \\ N_{21}^t & 0 & N_{23}^t & \cdots & N_{2P}^t \\ N_{31}^t & N_{32}^t & 0 & \cdots & N_{3P}^t \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ N_{P1}^t & N_{P2}^t & N_{P3}^t & \cdots & 0 \end{bmatrix}$$

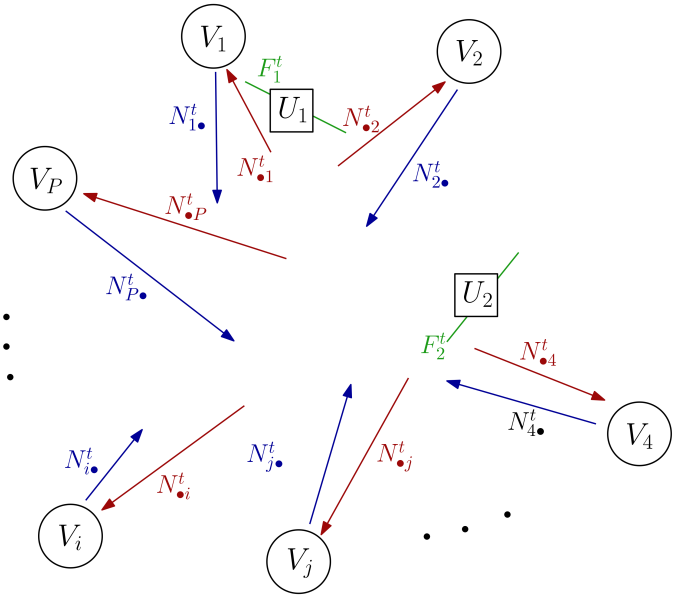
Observations

$N_{\cdot j}^t = \sum_{i=1}^P N_{ij}^t$
$N_{i\cdot}^t = \sum_{j=1}^P N_{ij}^t$
$F_h^t = \sum N_{ij}^t$
N_{ij}^t for some ij

We assume *a priori* that the distribution of the rate matrix is centered around some baseline rate matrix $\boldsymbol{\Lambda_0}$, which are the assumed rates when there is no anomalous activity. We then update this prior distribution using the observations $\mathcal{D} = \{\boldsymbol{R}^t, \boldsymbol{C}^t, \boldsymbol{F}^t\}_{t=1}^T$ in order to get a distribution of the rates $\mathrm{P}(\boldsymbol{\Lambda}|\mathcal{D})$, which does account for potential anomalous activity.



(a) Proposed Network: V_i - exterior nodes, U_i - interior nodes, N_{ij}^t - messages from node i to node j at time point t



(b) Actual Observed Network: $N_{i\cdot}^t$ - total egress of exterior nodes, $N_{\cdot i}^t$ - total ingress of exterior nodes, F_i^t - total flow through interior nodes

Fig. 1. Diagram of a network with P exterior nodes and 2 interior nodes.

III. Hierarchical Poisson Model with EM

We propose a generative model that assumes a series of statistical distributions govern the generation of the network. We assume that the messages N_{ij}^t passed through the network are Poisson distributed with rates Λ_{ij}. However, because we cannot observe the traffic network directly, we do not have the complete Poisson likelihood and use the EM algorithm. In

the following subsections, we will show a series of generative models with increasing complexity that attain successively higher accuracy. Then we will discuss warm-starting the EM algorithm at a robust initial solution to compensate for its sensitivity to initialization.

A. Proposed Hierarchical Bayesian Model

1) Maximum Likelihood by EM: The simplest hierarchical model assumes all priors are uniform, thus the only distributional assumption is that likelihood $\mathrm{P}(\boldsymbol{N}^1, \ldots, \boldsymbol{N}^T|\boldsymbol{\Lambda})$ is $\prod_{t=1}^{T}\prod_{ij} Poisson(\Lambda_{ij})$. The maximum likelihood estimator for the Poisson rates $\boldsymbol{\Lambda}$ can be approximated by lower bounds of the observed likelihood $\mathrm{P}(\mathcal{D}|\boldsymbol{\Lambda})$ using the maximum likelihood expectation maximization (MLEM) algorithm. The MLEM alternates between computing a lower bound on the likelihood function $\mathrm{P}(\mathcal{D}|\boldsymbol{\Lambda})$, the E-step, and maximizing the lower bound, the M-step. A general expression for the E-step bound can be expressed as:

$$\log \mathrm{P}(\mathcal{D}|\boldsymbol{\Lambda}) \geq \sum_{t=1}^{T} \mathrm{E}_{\mathrm{q}^t}\left(\log \mathrm{P}(\boldsymbol{R}^t, \boldsymbol{C}^t, \boldsymbol{F}^t, \boldsymbol{N}^t|\boldsymbol{\Lambda})\right) + \mathrm{H}(\mathrm{q}^t) \quad (1)$$

where q^t is an arbitrarily chosen distribution of $\boldsymbol{N}^t$, $\mathrm{E}_{\mathrm{q}^t}$ denotes statistical expectation with respect to the reference distribution q^t, and $\mathrm{H}(\mathrm{q}^t)$ is the Shannon entropy of q^t. The choice of q^t that makes the bound (1) the tightest, and results in the fastest convergence of the MLEM algorithm, is $\mathrm{q}^t = \mathrm{P}(\boldsymbol{N}^t|\boldsymbol{R}^t, \boldsymbol{C}^t, \boldsymbol{F}^t, \boldsymbol{\Lambda})$, (see Section 11.4.7 of [37]); however, this is not a tractable distribution. When the observations consist of the row and column sums of the matrix $\boldsymbol{N}^t$, this distribution is the multivariate Fisher's noncentral hypergeometric distribution, and when the flows are also observed the distribution is unknown. Unfortunately, use of this optimal distribution leads to an intractable E-step in the MLEM algorithm due to the coupling (dependence) between the row and column sums of $\boldsymbol{N}^t$. As an alternative we can weaken the bound on the likelihood function (1) by using a different distribution q that leads to an easier E-step. To this aim, we propose to use a distribution q that decouples the row sum from the column sum; equivalent to assuming that each sum is independent, e.g., as if each were computed with different realizations of $\boldsymbol{N}^t$.

Proposition 1. *Assume t_1, t_2 and t_3 are different time points so that observations at these time points are independent*

$$\mathrm{P}(\mathcal{D}|\boldsymbol{\Lambda}) = \prod_{t_1=1}^{T} \mathrm{P}(\boldsymbol{R}^{t_1}|\boldsymbol{\Lambda}) \prod_{t_2=1}^{T} \mathrm{P}(\boldsymbol{C}^{t_2}|\boldsymbol{\Lambda}) \prod_{t_3=1}^{T} \mathrm{P}(\boldsymbol{F}^{t_3}|\boldsymbol{\Lambda}).$$

Then the tightest lower bound of the observed data log likelihood is

$$\log \mathrm{P}(\mathcal{D}|\boldsymbol{\Lambda}) \geq \sum_{\tau=1}^{3} \sum_{t_\tau=1}^{T} \mathrm{H}(q^{t_\tau}) + \mathrm{E}_{q^{t_\tau}}\left(\log \mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{\Lambda})\right)$$

where $q^{t_1} = \mathrm{P}(\boldsymbol{N}^{t_1}|\boldsymbol{R}^{t_1}, \boldsymbol{\Lambda})$, $q^{t_2} = \mathrm{P}(\boldsymbol{N}^{t_2}|\boldsymbol{C}^{t_2}, \boldsymbol{\Lambda})$, and $q^{t_3}(\boldsymbol{N}^{t_3}) = \mathrm{P}(\boldsymbol{N}^{t_3}|\boldsymbol{F}^{t_3}, \boldsymbol{\Lambda})$ are multinomial distributions.

In the EM algorithm, the expectation in the E-step is taken with respect to the distribution estimated using the previous iteration's estimate of the parameter $\hat{\boldsymbol{\Lambda}}^k$, and the M-step does not depend on the entropy terms in the lower bound in Proposition 1, which are constant with respect to $\boldsymbol{\Lambda}$. Since the likelihoods are all Poisson, the E-step reduces to computing the means of multinomial distributions and the M-step for any ij pair is given by the Poisson MLE with the unknown N_{ij}^t terms replaced by their mean values. Explicitly the M-step objective is

$$\hat{\Lambda}_{ij}^{k+1} = \arg\max_{\Lambda_{ij}} \; -\Lambda_{ij} + \log(\Lambda_{ij}) N_{ij}^{total} \quad (2)$$

where $N_{ij}^{total} = \sum_{t_1=1}^{T} \mathrm{E}(N_{ij}^{t_1}|\boldsymbol{R}^{t_1}, \hat{\boldsymbol{\Lambda}}^k)$ $+ \sum_{t_2=1}^{T} \mathrm{E}(N_{ij}^{t_2}|\boldsymbol{C}^{t_2}, \hat{\boldsymbol{\Lambda}}^k) + \sum_{t_3=1}^{T} \mathrm{E}(N_{ij}^{t_3}|\boldsymbol{F}^{t_3}, \hat{\boldsymbol{\Lambda}}^k)$ and the expectations are with respect to the multinomial distributions of Proposition 1 . Thus the Poisson MLE equals $\hat{\Lambda}_{ij}^{k+1} = N_{ij}^{total}/3T$.

2) Maximum a Posteriori by EM: Because there are P^2 unobserved variables and only $\mathcal{O}(P)$ observed variables, the expected log likelihoods have a lot of local maxima. In order to make the EM objective better defined and incorporate the baseline Poisson rate information $\boldsymbol{\Lambda}_0$, a prior can be added to the likelihood model of the previous subsection. The EM objective of this new model is now the expected log posterior and the estimator in the M-step is the maximum a posteriori (MAP) estimator. It is natural to choose a conjugate prior of the form $\mathrm{P}(\boldsymbol{\Lambda}) = \prod_{ij} \mathrm{P}(\Lambda_{ij})$ where each $\Lambda_{ij} \sim Gamma(\epsilon_{ij}\Lambda_{0\,ij} + 1, \epsilon_{ij})$ (shape, rate) as this choice yields a closed form expression for the posterior distribution. These priors have modes at the baseline rates $\Lambda_{0\,ij}$. The hyperparameters ϵ_{ij} can be thought of as the belief we have in the correctness of the baseline so as $\epsilon \to 0$, the prior variance goes to infinity, and the prior becomes non-informative because we have no confidence in the baseline, while as $\epsilon \to \infty$, the prior variance goes to zero, and the prior degenerates into the point $\Lambda_{0\,ij}$ because we are certain the baseline is correct.

Given a matrix of hyperparameters $\boldsymbol{\epsilon}$, the complete data posterior distribution is $\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon}, \boldsymbol{N}^1, \ldots, \boldsymbol{N}^T) = \prod_{ij} \mathrm{P}(\Lambda_{ij}|\epsilon_{ij}, N_{ij}^1, \ldots, N_{ij}^T)$ where each posterior is of the form of $Gamma(\epsilon_{ij}\Lambda_{0\,ij} + 1 + \sum_{t=1}^{T} N_{ij}, \epsilon_{ij} + T)$. Because we can only observe the network indirectly $\mathcal{D} = \{\boldsymbol{R}^t, \boldsymbol{C}^t, \boldsymbol{F}^t\}_{t=1}^{T}$, we again must estimate the mode of this posterior using the EM algorithm, which is very similar to the algorithm for the likelihood model. The only difference is the M-step in which an additional term of the form $\sum_{ij}(\epsilon_{ij}\Lambda_{0\,ij})\log(\Lambda_{ij}) - \epsilon_{ij}\Lambda_{ij}$ is added to (2). Thus at every EM iteration, the entries of the MAP estimator matrix $\hat{\boldsymbol{\Lambda}}^{k+1}$ are

$$\hat{\Lambda}_{ij}^{k+1} = \frac{\epsilon_{ij}\Lambda_{0\,ij} + N_{ij}^{total}}{\epsilon_{ij} + 3T} \quad (3)$$

where N_{ij}^{total} is the same as in (2).

3) Bayesian Hierarchical Model: Choosing the hyperparameters ϵ_{ij} can be difficult because it is not always possible to quantify our belief in the correctness of the baseline rates. We can rectify this by allowing the ϵ_{ij} to be random with hyperpriors $\epsilon_{ij} \sim Uniform(0, \infty)$. We choose uninformative

hyperpriors for $\epsilon_{ij} > 0$. A notional diagram for the proposed hierarchical model is shown in Fig. 2.

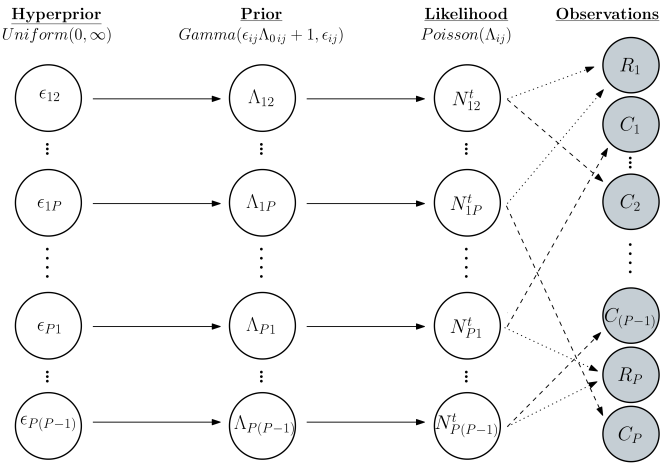


Fig. 2. The statistical process believed to underlie our network.

With these uninformative priors the posterior takes the form

$$\begin{aligned}\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T) &= \int \frac{\mathrm{P}(\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T|\boldsymbol{\Lambda})\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)}\,d\boldsymbol{\epsilon}\\ &= \int \frac{\mathrm{P}(\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T|\boldsymbol{\Lambda})\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon})}{\mathrm{P}(\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T|\epsilon)}\frac{\mathrm{P}(\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T|\boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)}\,d\boldsymbol{\epsilon}\\ &= \int \mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon},\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)\mathrm{P}(\boldsymbol{\epsilon}|\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)\,d\boldsymbol{\epsilon} \qquad (4)\end{aligned}$$

where $\mathrm{P}(\boldsymbol{\epsilon}|\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T) = \int \mathrm{P}(\boldsymbol{\Lambda},\boldsymbol{\epsilon}|\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)\,d\boldsymbol{\Lambda}$. The observed (incomplete data) log posterior $\log \mathrm{P}(\boldsymbol{\Lambda}|\mathcal{D})$ has lower bound proportional to

$$\log\left(\int \exp\left\{\mathrm{E_q}\left(\log \mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon},\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)\right)\right\} \exp\left\{\mathrm{E_q}\left(\log\int \mathrm{P}(\boldsymbol{\Lambda},\boldsymbol{\epsilon}|\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)\,d\boldsymbol{\Lambda}\right)\right\}d\boldsymbol{\epsilon}\right)$$

which is tight when $\mathrm{q} = \mathrm{P}(\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T|\mathcal{D},\boldsymbol{\Lambda})$, as shown in (9) in the Appendix.

However, marginalizing the joint posterior $\int \mathrm{P}(\boldsymbol{\Lambda},\boldsymbol{\epsilon}|\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)\,d\boldsymbol{\Lambda}$ is often not feasible, so instead it is popular to use empirical Bayes to approximate it with a point-estimate

We propose an empirical Bayes approach to maximizing the log posterior as an alternative to maximization of (4) $\hat{\boldsymbol{\epsilon}} = \arg\max_{\boldsymbol{\epsilon}} \mathrm{P}(\boldsymbol{\epsilon}|\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)$. This empirical Bayes approximation can be embedded in the EM algorithm so that once we have an estimate for $\boldsymbol{\epsilon}$, an estimator for $\boldsymbol{\Lambda}$ is obtained by maximizing the expected log conditional posterior $\mathrm{E_q}\left(\log \mathrm{P}(\boldsymbol{\Lambda}|\hat{\boldsymbol{\epsilon}},\boldsymbol{N}^1,\ldots,\boldsymbol{N}^T)\right)$.

Theorem 1. *Using the time independence in Proposition 1 and the empirical Bayes approximation, the E-step of the EM algorithm for the hierarchal model is*

$$\hat{N}^{t_1}_{ij} = \mathrm{E}(N^{t_1}_{ij}|\boldsymbol{R}^{t_1},\hat{\boldsymbol{\Lambda}}^k) = \frac{\hat{\Lambda}^k_{ij}}{\sum_{j=1}^P \hat{\Lambda}^k_{ij}} R^{t_1}_i$$

$$\hat{N}^{t_2}_{ij} = \mathrm{E}(N^{t_2}_{ij}|\boldsymbol{C}^{t_2},\hat{\boldsymbol{\Lambda}}^k) = \frac{\hat{\Lambda}^k_{ij}}{\sum_{i=1}^P \hat{\Lambda}^k_{ij}} C^{t_2}_j,$$

$$\hat{N}^{t_3}_{ij} = \mathrm{E}(N^{t_3}_{ij}|\boldsymbol{F}^{t_3},\hat{\boldsymbol{\Lambda}}^k) = \frac{\hat{\Lambda}^k_{ij}}{\sum_{ij} \hat{\Lambda}^k_{ij}} F^{t_3}_h \textit{ for any pair } ij,$$

and the M-step is

$$\begin{aligned}\hat{\epsilon}^{k+1}_{ij} &= \arg\max_{\epsilon_{ij}} \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \log\frac{\Gamma(\hat{N}^{t_\tau}_{ij}+\epsilon_{ij}\Lambda_{0\,ij}+1)}{\Gamma(\epsilon_{ij}\Lambda_{0\,ij}+1)}\\ &\quad+\sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T}(\epsilon_{ij}\Lambda_{0\,ij}+1)\log\frac{\epsilon_{ij}}{1+\epsilon_{ij}} - \hat{N}^{t_\tau}_{ij}\log(1+\epsilon_{ij})\end{aligned}$$

and

$$\begin{aligned}\hat{\Lambda}^{k+1}_{ij} &= \arg\max_{\Lambda_{ij}} (\hat{\epsilon}_{ij}\Lambda_{0\,ij})\log(\Lambda_{ij}) - \hat{\epsilon}_{ij}\Lambda_{ij} - 3T\Lambda_{ij}\\ &\quad+\log(\Lambda_{ij})\left(\sum_{t_1=1}^{T}\hat{N}^{t_1}_{ij}+\sum_{t_2=1}^{T}\hat{N}^{t_2}_{ij}+\sum_{t_3=1}^{T}\hat{N}^{t_3}_{ij}\right).\end{aligned}$$

Since the function that lower bounds the observed log likelihood changes after every iteration of the EM algorithm, the prior should also change after every iteration. Intuitively, the earlier iterations of the EM algorithm will have expected log likelihoods that are more misspecified than the later iterations. This suggests spreading the prior distribution in the earlier iterations. The empirical Bayes approximation of Theorem 1 effectively does this by allowing the variance of the prior to be chosen using the data instead of fixing it as a constant. In this manner, the empirical Bayes approximation can be thought of as a Bayesian analog to the regularized EM algorithm of [38].

B. Warm Starting with Minimum Relative Entropy

The EM algorithm is well known to be sensitive to initialization, especially if the objective has a lot of local maxima. Thus if instead of a random initialization, the EM algorithm is warm-started, it is more likely to converge to a good maximum and also potentially converge faster. A good choice for an initialization point is a more robust estimator of the rate matrix such as the solution to a model with fewer distributional assumptions. Thus instead of modeling an explicit generative model, we can instead adopt the minimum relative entropy (MRE) principle [39], [40], [41], and [42]. Geometrically, this reduces to an information projection of the prior distribution, as shown in Fig. 3.

The constrained minimum relative entropy distribution is the density that is closest to a given prior distribution and lies in a feasible set, $\mathcal{P}$. This feasible set is formed from constraints that require their expected values, with respect to the minimum relative entropy distribution, to match properties of the observations, $\mathcal{D}$ (the total ingress, egress, and flows). And because relative entropy is the Kullback-Leibler (KL) divergence between probability distributions, this is used as the metric for closeness. This closeness criterion is well suited to the anomaly detection problem of interest to us because anomalous activity is rare, so the distribution of the actual rates, $\boldsymbol{\Lambda}$, should be similar to the prior distribution $\mathrm{P}_0(\boldsymbol{\Lambda}) = \mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\Lambda}_0)$, which is parameterized by the baselines rates $\boldsymbol{\Lambda_0}$.

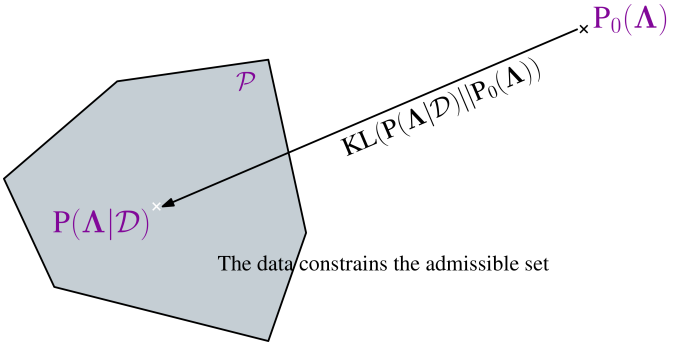


Fig. 3. A projection of the prior, $\mathrm{P}_0(\mathbf{\Lambda})$, onto a feasible set $\mathcal{P}$ of distributions that satisfy the observed data, $\mathcal{D}$.

The MRE objective is

$$\min_{\mathrm{P}(\mathbf{\Lambda}|\boldsymbol{R},\boldsymbol{C},\boldsymbol{F})} \mathrm{KL}\left(\mathrm{P}(\mathbf{\Lambda}|\boldsymbol{R},\boldsymbol{C},\boldsymbol{F})||\mathrm{P}_0(\mathbf{\Lambda})\right)$$

subject to

$$\int \mathrm{P}(\mathbf{\Lambda}|\boldsymbol{R},\boldsymbol{C},\boldsymbol{F})(\mathbf{\Lambda}\mathbf{1}-\bar{\boldsymbol{R}})\,d\mathbf{\Lambda}=\mathbf{0}$$

$$\int \mathrm{P}(\mathbf{\Lambda}|\boldsymbol{R},\boldsymbol{C},\boldsymbol{F})(\mathbf{1}'\mathbf{\Lambda}-\bar{\boldsymbol{C}})\,d\mathbf{\Lambda}=\mathbf{0}$$

$$\int \mathrm{P}(\mathbf{\Lambda}|\boldsymbol{R},\boldsymbol{C},\boldsymbol{F})(\boldsymbol{A}\mathbf{\Lambda}\boldsymbol{B}-\bar{\boldsymbol{F}})\,d\mathbf{\Lambda}=\mathbf{0}$$

where $\mathbf{0}$ and $\mathbf{1}$ are vectors of zeros and ones respectively, $\bar{\boldsymbol{C}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{C}^t$ and $\bar{\boldsymbol{R}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{R}^t$ are the average rates of observed total traffic into and out of each node, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are 0-1 matrices summing the rates that flow through each of the interior nodes with average observations $\bar{\boldsymbol{F}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}^t$. Using the Legendre transform of the Lagrangian to get the Hamiltonian, the optimal density has the form

$$\mathrm{P}(\mathbf{\Lambda}|\boldsymbol{R},\boldsymbol{C},\boldsymbol{F})=\frac{\mathrm{P}_0(\mathbf{\Lambda})}{Z(\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi})}\exp\left\{\boldsymbol{\rho}'(\mathbf{\Lambda}\mathbf{1}-\bar{\boldsymbol{R}})+\boldsymbol{\gamma}'(\mathbf{1}'\mathbf{\Lambda}-\bar{\boldsymbol{C}})+\boldsymbol{\phi}'(\boldsymbol{A}\mathbf{\Lambda}\boldsymbol{B}-\bar{\boldsymbol{F}})\right\} \tag{5}$$

where $\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}$ are Lagrange multipliers that maximize the negative log partition function $-\log(Z(\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}))$.

Proposition 2. *Let* $\mathrm{P}_0(\mathbf{\Lambda})=\prod_{ij}\mathrm{P}_0(\Lambda_{ij})$ *be independent Laplace distributions with mean parameter* $\Lambda_{0\,ij}$ *and scale parameter* 1*, then the constrained mode of the MRE distribution is the solution to*

$$\arg\max_{\mathbf{\Lambda}\in\mathbb{R}^+} -||\mathbf{\Lambda}-\mathbf{\Lambda}_0||_1+\hat{\boldsymbol{\rho}}'(\mathbf{\Lambda}\mathbf{1}-\bar{\boldsymbol{R}})+\hat{\boldsymbol{\gamma}}'(\mathbf{1}'\mathbf{\Lambda}-\bar{\boldsymbol{C}})'+\hat{\boldsymbol{\phi}}'(\boldsymbol{A}\mathbf{\Lambda}\boldsymbol{B}-\bar{\boldsymbol{F}})$$

where $\hat{\boldsymbol{\rho}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\phi}}=\arg\max_{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}} -\log\left(Z(\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi})\right)$.

Maximizing the above expression over $\mathbf{\Lambda}$ (constrained to only positive real numbers) can be seen as a slight relaxation of the more direct objective of minimizing the loss function

$$\arg\min_{\mathbf{\Lambda}\in\mathbb{R}^+}||\mathbf{\Lambda}-\mathbf{\Lambda}_0||_1 \tag{6}$$

$$\text{subject to } \mathbf{\Lambda}\mathbf{1}=\bar{\boldsymbol{R}},\ \mathbf{1}'\mathbf{\Lambda}=\bar{\boldsymbol{C}},\ \boldsymbol{A}\mathbf{\Lambda}\boldsymbol{B}=\bar{\boldsymbol{F}}$$

where $||\cdot||_1$ is the element wise ℓ_1 norm. The loss function in (6) has the advantage that it can be easily implemented in any constrained convex solver such as CVX [43].

The objective in (6) is an easily interpretable formulation for estimating the rate matrix, which does not depend on the unobserved traffic N_{ij}^t. And, because it does not put distributional assumptions on the "likelihood", it is more robust to model mismatch, at the cost of accuracy. The generality of the solution to (6), while not precise enough on its own, makes it a good candidate to be further refined by the EM algorithm in the Hierarchical Poisson model.

IV. Testing For Anomalies

Since the estimators in the previous section are maximizers of probabilistic models, a natural way to test for anomalies in the rate matrix $\mathbf{\Lambda}$ is to compare goodness of fit of the fitted model using hypothesis testing. By testing the null hypothesis $\mathrm{vec}(\mathbf{\Lambda})=\mathrm{vec}(\mathbf{\Lambda}_0)$ against the alternative hypothesis $\mathrm{vec}(\mathbf{\Lambda})\neq\mathrm{vec}(\mathbf{\Lambda}_0)$, we can control the false positive rate (FPR) (Type 1 error), of incorrectly declaring anomalous activity in the rate matrix, using a level-α test. In this section we will represent a statistical model with the notation $\mathcal{M}(\cdot)$, as the results apply for both log likelihood and log posterior models.

Depending on if the statistical models are likelihoods or posteriors, the statistic

$$\psi=-2\sum_{t=1}^{T}\Big(\log(\mathcal{M}_t(\mathbf{\Lambda}_0))-\log(\mathcal{M}_t(\hat{\mathbf{\Lambda}}))\Big) \tag{7}$$

would be either a log likelihood ratio (LR) statistic or a log posterior density ratio (PDR) statistic [44] respectively, where $\hat{\mathbf{\Lambda}}=\arg\max_{\mathbf{\Lambda}\in\mathbb{R}^+}\mathcal{M}(\mathbf{\Lambda})$. Thus testing ψ against a threshold can be seen as a generalized log likelihood ratio test or generalized log posterior ratio test with a composite alternative hypothesis.

Proposition 3. *Under the standard regularity conditions for the log LR statistic or under the sufficient conditions of the Bernstein-von Mises theorem for the log PDR statistic,* ψ *will be asymptotically* $\chi^2_{P^2-P}$ *distributed under the null hypothesis.*

Next we show that the statistic ψ in (7) is a good estimator of the KL divergence between the true model at its maximum and the true model at the baseline. And even if the models are misspecified, the statistic

$$\hat{\psi}=-2\sum_{t=1}^{T}\Big(\log(\hat{\mathcal{M}}_t^k(\mathbf{\Lambda}_0))-\log(\hat{\mathcal{M}}_t^k(\hat{\mathbf{\Lambda}}))\Big)$$

can still be a good estimator for goodness-of-fit, where the k in $\hat{\mathcal{M}}^k(\mathbf{\Lambda}_0)$ and $\hat{\mathcal{M}}^k(\hat{\mathbf{\Lambda}})$ indicates the iteration of the EM algorithm.

Proposition 4. *The statistic* ψ/T *is a consistent estimator for*

$$\Psi=2\,\mathrm{KL}\left(\mathcal{M}(\mathbf{\Lambda}^*)||\mathcal{M}(\mathbf{\Lambda}_0)\right),$$

the KL divergence between the true model and the true model under the null hypothesis. The statistic $\hat{\psi}/T$ is a consistent estimator for

$$2\,\mathrm{KL}\left(\mathcal{M}(\mathbf{\Lambda}^*)||\mathcal{M}(\mathbf{\Lambda}_0)\right) - 2\left(\mathrm{KL}(\mathcal{M}(\mathbf{\Lambda}^*)||\hat{\mathcal{M}}^k(\hat{\mathbf{\Lambda}}^*)) - \mathrm{KL}(\mathcal{M}(\mathbf{\Lambda}_0)||\hat{\mathcal{M}}^k(\mathbf{\Lambda}_0))\right) \tag{8}$$

where $\hat{\mathcal{M}}^k(\hat{\mathbf{\Lambda}}^*)$ is the closest population local maximum at iteration k.

The second term in (8) can be seen as the difference between the true model misspecification error and the model misspecification error of the null hypothesis. So if conditions are satisfied so that the EM algorithm converges to the global maximum as the number of iterations $k \to \infty$ or if the model is equally as misspecified under the truth as under the null hypothesis such that the differences in the second term in (8) cancel to 0, then the statistic $\hat{\psi}/T$ is also a consistent estimator of Ψ. The justification for using misspecified models can also be geometrically interpreted as follows. Because the models estimated from the EM algorithm are from the correct parametric family of distributions, the misspecified models still lie on the same Riemannian manifold as the correct models. Below, we provide an algorithm for performing hypothesis testing on the statistic $\hat{\psi}$.

Algorithm 1: Anomaly Test

Input: models $\hat{\mathcal{M}}_1^k, \ldots, \hat{\mathcal{M}}_T^k$, critical value $c = F^{-1}(\alpha)$ where F is $\chi^2_{p^2-p}$ CDF, α is test level

Solve $\hat{\mathbf{\Lambda}} = \arg\max_{\mathbf{\Lambda}\in\mathbb{R}^+} \sum_{t=1}^T \log(\hat{\mathcal{M}}_t^k(\mathbf{\Lambda}))$

$\hat{\psi} = -2\sum_{t=1}^T \left(\log(\hat{\mathcal{M}}_t^k(\mathbf{\Lambda}_0)) - \log(\hat{\mathcal{M}}_t^k(\hat{\mathbf{\Lambda}}))\right)$

if $\hat{\psi} > c$ **then**

  Reject $\mathrm{vec}(\mathbf{\Lambda}) = \mathrm{vec}(\mathbf{\Lambda}_0)$

else

  Do not reject $\mathrm{vec}(\mathbf{\Lambda}) = \mathrm{vec}(\mathbf{\Lambda}_0)$

end if

Return: Reject or Not

Algorithm 1 calculates the statistic $\hat{\psi}$ as a log ratio of the modes of the model under the null and alternative hypothesis. It then tests $\hat{\psi}$ against a critical value c, which is related to the false positive level.

Under the null hypothesis, the statistic $\hat{\psi}$ can be decomposed as sampling error $-2\sum_{t=1}^T \log \hat{\mathcal{M}}_t^k(\hat{\mathbf{\Lambda}}^*) - \max_{\mathbf{\Lambda}\in\mathbb{R}^+} \log \hat{\mathcal{M}}_t^k(\mathbf{\Lambda})$ plus model error $-2\sum_{t=1}^T \log \hat{\mathcal{M}}_t^k(\mathbf{\Lambda}_0) - \log \hat{\mathcal{M}}_t^k(\hat{\mathbf{\Lambda}}^*)$. Thus for the level-$\alpha$ test $\mathrm{P}(\hat{\psi} > c|\mathcal{H}_0) = \alpha$, a Type-I error can occur due to either sampling error or model error or a combination of both. Since typically the finite sample distribution of the statistic ψ is unknown, the asymptotic distribution described in Proposition 3 can be used to choose the critical value c of $\mathrm{P}(\psi > c|\mathcal{H}_0) = \alpha$. Assuming the model error is small, or small relative to the sampling error, we can also use Proposition 3 to choose the critical value of a test with a misspecified statistic $\mathrm{P}(\hat{\psi} > c|\mathcal{H}_0) = \alpha$. In the following section, we will show in simulations that the asymptotic distribution of the correct statistic ψ is adequate for choosing the critical value of a test using the misspecified statistic $\hat{\psi}$.

V. Computational Complexity

In Algorithm 2, we present our hierarchical Poisson EM model warm started at the MRE estimator and analyze its computational complexity.

Algorithm 2: HP-MRE

Input: observations $\mathcal{D} = \{\boldsymbol{R}^t, \boldsymbol{C}^t, \boldsymbol{F}^t\}_{t=1}^T$, test level α

Initialize: $\hat{\mathbf{\Lambda}}$ as the solution to (6)

repeat

  E-Step: Calculate $\hat{N}_{ij}^{t_1}, \hat{N}_{ij}^{t_2}, \hat{N}_{ij}^{t_3}$ for all i, j in Theorem 1

  M-Step: Solve for $\hat{\epsilon}_{ij}^{k+1}$ and $\hat{\Lambda}_{ij}^{k+1}$ for all i, j in Theorem 1

until convergence

Test: Calculate $\hat{\psi}$ and reject if it is greater than critical value c

Return: Reject or Not

Warm starting the EM algorithm at the MRE solution (6) requires using interior-point methods, which have polynomial complexity in the number of variables. Since the MRE objective has P^2 linear variables and $2P^2$ second order cone problem variables, the computational cost is of order $\mathcal{O}(\#IPiter(3P^2)^r)$ where r is the polynomial degree (often 3) and $\#IPiter$ is the number of iterations of the interior point algorithm.

The E-Step consists of calculating the multinomial means using the observed data. Assume that the number of flows in the interior nodes are roughly P, so that each of the row sums, column sums, and interior node flows are the summation of P values. Then for each independent time instance t_τ, there are P summations of P values in denominator and a multiplication and division operation on each of the P^2 entries in the numerator. The total computational cost of the E-step is of order $\mathcal{O}(\tau TP^2)$ where τ is the number of different time points in Proposition 1 (2 + number of interior nodes).

In the M-step, the estimator $\hat{\epsilon}_{ij}^{k+1}$ can only be solved numerically because the score function of the negative binomial distribution is a non-linear equation. Because we can derive the gradient of the score function, we can use a trust-region method with a Newton conjugate gradient subproblem (each subproblem has linear complexity in time points). Given $\hat{\epsilon}_{ij}^{k+1}$, the estimator $\hat{\Lambda}_{ij}^{k+1}$ can be solved in closed form (3) with scalar operations, making its complexity linear in time points. Thus the total computational cost of the M-step is of order $\mathcal{O}((1 + \#CGiter)TP^2)$ where $\#CGiter$ is the number of conjugate gradient iterations.

Given the final iterations EM estimators, evaluating the models at each i, j entry only involves scalar operations, and getting the log ratio statistic $\hat{\psi}$ requires summing over all i, j entries and the T time points; so the total complexity of the anomaly test statistics is of order $\mathcal{O}(TP^2)$. Thus, overall Algorithm 2 has computational complexity of order $\mathcal{O}(\#IPiter(3P^2)^r + \#EMiter((\tau + 1)TP^2 + \#CGiterTP^2))$. Note that our choice in algorithms for the numerical optimizations were based more on convenience (using popular standard packages e.g. CVX, Matlab's fsolve) than optimal performance, so the computational complexities listed in this section are certainly not the best case scenarios. Nonetheless, even using non-optimal numerical algorithms, we show, in the following section, that our method can run in a

reasonable amount of time in both simulations and large real world problems.

VI. Simulation and Data Examples

In this section, we model network traffic in both simulated and real datasets as hierarchical Poisson posteriors to get estimators of the true network traffic rates. These estimators, from the hierarchical Poisson posteriors where the EM algorithm is initialized randomly or at the MRE estimator (Rand-HP or MRE-HP), are tested against baseline rates to detect anomalous activity in the network, as shown in Algorithm 1. We compare the performance of our proposed models to the maximum likelihood EM (MLEM) model of [17] (with the same time independence assumptions of Proposition 1 for feasibility), the Traffic and Anomaly Map (TA-Map) method of [22], and an "Oracle" that unrealistically observes the network directly. The "Oracle" estimator is the uniformly minimum variance unbiased estimator and achieves the Cramer-Rao lower bound [45].

The Traffic and Anomaly Map method is the state-of-the-art for estimating the rates in networks with traffic anomalies. Specifically for the TA-Map method we use the objective of (P1) in [22], but with the low rank decomposition of (P4) in [22] where $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{Q}'$ and $\boldsymbol{Q} = \mathbf{1}$ is a vector of ones because the rates do not change over time. Since the anomalies also do not change over time, they can be expanded as $\boldsymbol{A}\boldsymbol{Q}'$ where $\boldsymbol{A}$ is a $P^2 \times 1$ vector of rates of anomalous activity. We use $\boldsymbol{\Lambda}_0$ to form the routing matrix for the vector of nominal rates $\boldsymbol{L}$ and a full routing matrix for the vector of anomalous rates $\boldsymbol{A}$ since we do not know any structural knowledge about them. Additionally, converting the notation of [22] to the notation of this paper, $\boldsymbol{Y} = [\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{F}]$, $\boldsymbol{Z}_\Pi$ are defined as the edges that are observed, and $\boldsymbol{L} + \boldsymbol{A} = \text{vec}(\boldsymbol{\Lambda})$, where $\boldsymbol{L}$ and $\boldsymbol{A}$ are solved using CVX on (P1) in [22]. We empirically choose the penalty parameters $\lambda_\star = 0.5$ and $\lambda_1 = 0.1$.

A. Simulation Results

We simulate networks where the baseline rate matrix has 10 exterior nodes and 2 interior nodes. The probability of an edge between any two nodes in the baseline network is 0.65, the baseline rates $\Lambda_{0\,ij}$ are drawn from $Gamma(1.75, 1)$ distributions, and each interior node observes the total flow of a random 7 edges. We consider scenarios where anomalous activity can take place in either the edges or the nodes. In the first scenario, the anomalous activity can cause increases in the rates of some of the edges, new edges to appear or disappear, or both. So, the rates of anomalous activity $\Lambda_{ij} - \Lambda_{0\,ij}$ are drawn from $Gamma(0.75, 1)$ distributions where the probability of anomalous activity between any two nodes is 0.2. In the second scenario, there is a hidden node that is interacting with the other nodes, thus affecting the observed total flows of the known nodes. So the entries of the true rate matrix are drawn from $Gamma(1.75, 1)$ distributions, but the true rate matrix has 11 exterior nodes and the baseline rate matrix is the 10×10 submatrix of known nodes. Like in the first scenario, the probability of an edge between the hidden node and another node is 0.2. All simulations contain 200 trials, with anomalous activity in approximately half of them.

In Fig. 4 we explore the accuracy of correctly identifying anomalous activity as a function of the percentage of observed edges, where we observe $T = 100$ time points (samples). We measure accuracy as $\frac{\#TP+\#TN}{\#Trials}$ where the number of true positives (TP) and true negatives (TN) are the number of times a method correctly detects that there is anomalous activity or no anomalous activity respectively. For the probabilistic models (MLEM, Rand-HP, MRE-HP) , we use the likelihood or posterior density ratio tests described in Section IV where the critical value is calculated using the inverse cumulative distribution function of the $\chi^2_{P^2-P}$ distribution at 0.05. The Traffic and Anomaly Map method uses a threshold on the maximum (absolute) value of the anomaly matrix $\boldsymbol{A}$ where the threshold is chosen so that it has 0.05 Type-I error. While the accuracy of all the probabilistic models increases as the percentage of observed edges increases, the MLEM has low accuracy unless over 80% of the network is observed whereas the two Hierarchical Poisson models have high accuracy even when no part of the network is directly observed. The TA-Map method also has poor performance at all percentages of the network observed. This may due to issues the TA-Map method has at separating $\boldsymbol{L}$ and $\boldsymbol{A}$ into the correct separate matrices even when the total estimator $\boldsymbol{L} + \boldsymbol{A}$ is accurate.

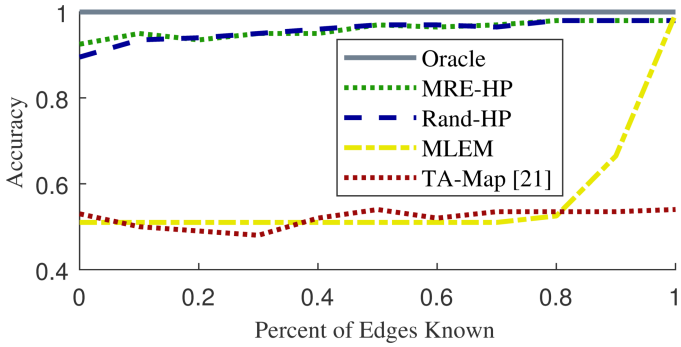


Fig. 4. The network has 10 exterior nodes, 2 interior nodes, 35% sparsity, and a 0.5 probability of having anomalous activity, where $T = 100$ samples are observed. The accuracy of correctly detecting if the network has anomalous activity increases as the number of edges observed increases. The proposed Rand-HP, and MRE-HP models outperform the state-of-the-art TA-Map anomaly detector.

While the Rand-HP and MRE-HP models have approximately the same accuracy at detecting anomalies (MRE-HP does slightly better when only a few of the edges are observed), initializing the EM algorithm of the Hierarchical Poisson model at the MRE solution has additional benefits. Fig. 5 shows that the EM algorithm in the Hierarchical Poisson model with random initialization takes longer to converge than if it is initialized at the MRE solution. This is because, if the EM algorithm is initialized in a place where likelihood is very noisy, it may have difficultly deciding on the best of the nearby local maxima, but the MRE solution is often already close to a good local maximum.

Fig. 6 shows the mean squared error (MSE) of the estimated rate matrices $||\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}||_F^2$. The MRE-HP model gains some of the advantages of the MRE estimator making its MSE much

lower than that of the Rand-HP model. As the percentage of observed edges in the network increases, all estimators' errors decrease to the Oracle estimator's error, which is the lowest possible MSE among all unbiased estimators. However, both the TA-Map method and the MLEM model do not have good performance except when almost all of the network is observed, at which point every estimator performs well. Note that estimating the traffic is not the end goal in the considered anomaly detection problem. We demonstrate this by comparing Fig. 6 to Fig. 4, where we can see that estimating the traffic well (having low MSE) does not guarantee the method high accuracy. Low MSE implies that a method's estimates do not have a large difference with the true rates, however depending on where the differences occur, it can be enough to cause the method to incorrectly detect anomalous activity.

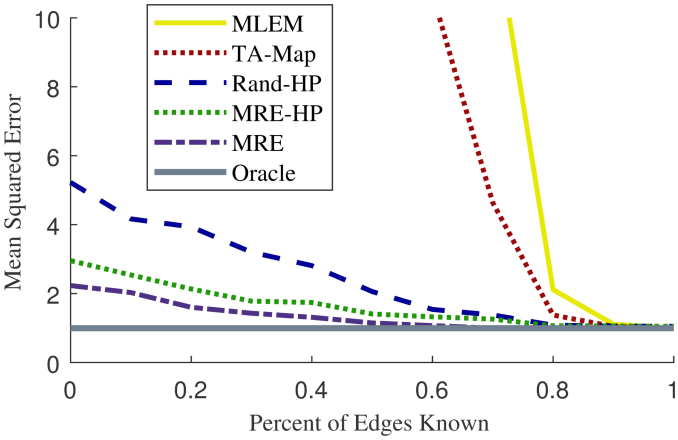


Fig. 6. The MSE decreases as the number of edges observed increases. The proposed MRE, Rand-HP, and MRE-HP models outperform the state-of-the-art TA-Map method.

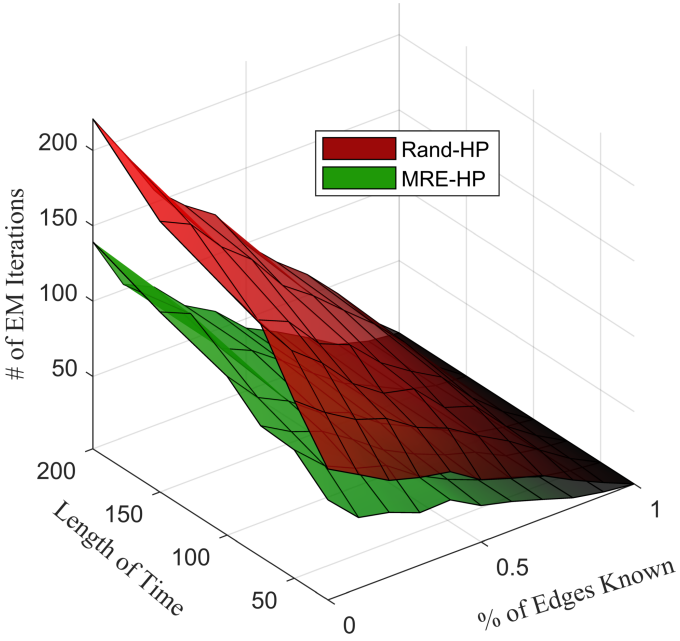


Fig. 5. The number of iterations required for the EM algorithm to converge as the observation time and number of edges observed vary. By warm-starting the EM algorithm at the MRE estimator, the number of iteration is much fewer everywhere because it is already close to a good local maximum.

Fig. 7 shows the ROC curves of the anomaly detection performance of the MRE-HP, MLEM, and TA-Map methods for both the anomalous rates and the hidden node scenarios, where only 20% of the edges are observed. The accuracy of the MRE-HP model increases with the total observation time T, and it can detect anomalous activity almost perfectly with only 100 time points, as evidenced by its area under the curve (AUC) being very close to 1. The stars over the lines are the FPR vs TPR when using the critical values found by calculating the inverse cumulative distribution function of the $\chi^2_{P^2-P}$ distribution at 0.05. The ROC curve for testing a misspecified LR test statistic using the MLEM is just the point at $(1, 1)$ because the Poisson MLE model is so misspecified, it always rejects the null hypothesis. The TA-Map method, while it does not always rejects the null hypothesis like the MLEM model, performs about as bad as random guessing (a diagonal line from $(0, 0)$ to $(1, 1)$). These results are consistent with the accuracy results shown in Fig. 4.

In Table I, we show the corresponding CPU timings of each method in the two scenarios used in Fig. 7. The algorithms were run on an Intel Xeon E5-2630 processor at 2.30GHz without any explicit parallelization; however some of the built-in Matlab functions are by default multi-threaded (such as ones that call BLAS or LAPACK libraries). While the MRE-HP is slower than the competing methods, its computation time is still very fast and on average less than half a minute. Also, note the significant performance improvement provided by MRE-HP in the considered anomaly detection problem (see Fig. 4 and Fig. 7).

TABLE I
FIG. 7 CPU TIMES (IN SECONDS) OVER 200 TRIALS

	Increase in Rates		Hidden Node	
	Average	Standard Dev.	Average	Standard Dev.
MRE-HP	18.594	28.611	18.901	30.785
MLEM	0.0398	0.0112	0.0380	0.0119
TA-Map	3.1860	0.1347	3.1861	0.1912

B. CTU-13 Dataset

The proposed model was applied to botnet traffic networks from the CTU-13 dataset, which come from 13 different scenarios of botnets executing malware attacks captured by CTU University, Czech Republic, in 2011 [46]. The dataset contains real botnet traffic mixed with normal traffic and background traffic and the authors of [46] processed the captured traffic into bidirectional NetFlows and manually labeled them. Because the objective is to detect if there is botnet traffic among the regular users, we will only use the sub-network of nodes that are being used for normal traffic, but the traffic on this sub-network can be of any type: normal, background, or botnet. Thus, baseline traffic on the network is either normal or background traffic and the anomalous traffic is from botnets. And because the botnet traffic originates and also potentially

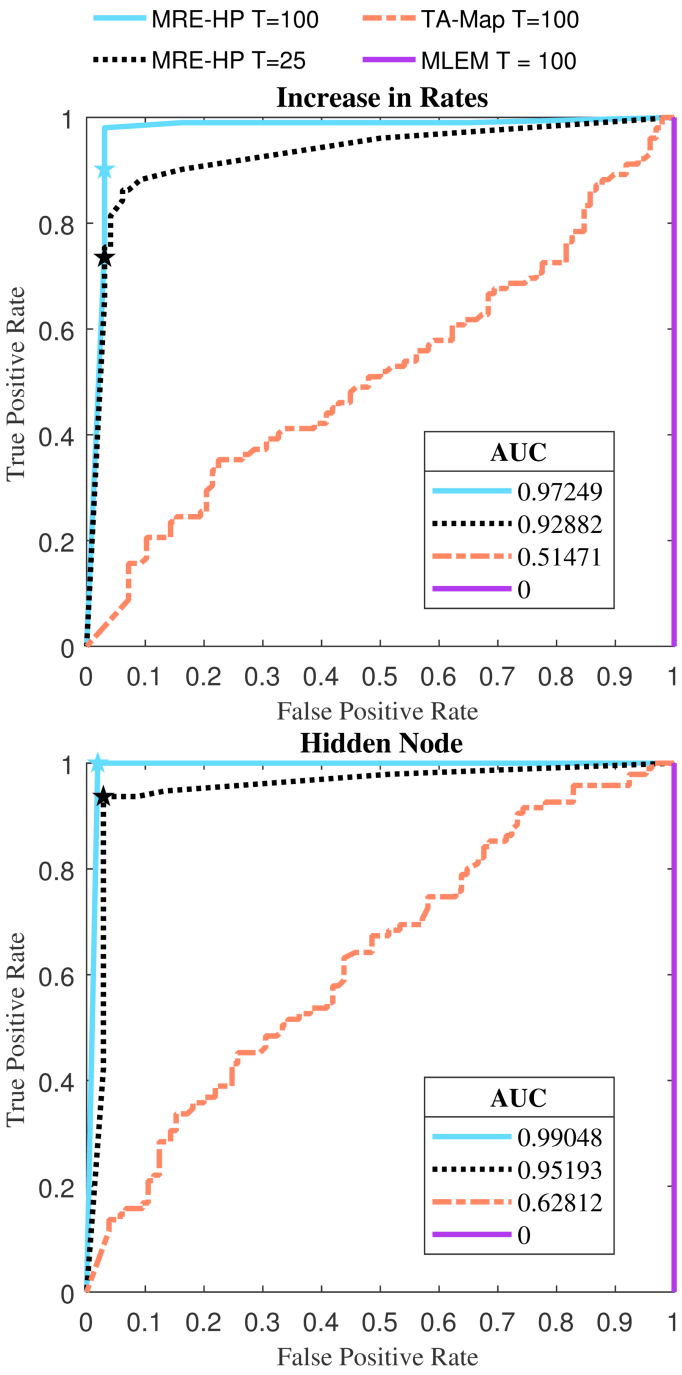


Fig. 7. ROC curves where 20% of edges in the network are observed and roughly half of the networks have anomalous activity. The proposed MRE-HP model can detect anomalous activity almost perfectly while the TA-Map and MLEM methods have poor performance.

ceases from nodes that are not the regular users, the anomalous activity is due to unobserved hidden nodes.

The observations consist of the total ingress and egress of each node along with the total flows of 10 interior nodes, where each interior node receives flow from $0.7P$ other nodes, in addition to observing 20% of the edges in the network. An observation or sample is all the traffic that occurs in a one-hour time period. For each of the scenarios, we test the probabilistic models at an alpha level of 0.05 under both regimes where the null hypothesis is true (no botnet traffic) and not true (botnet traffic). For the TA-Map method of [22], we use the ROC curves from the simulations to choose the threshold that yields a Type-I error equal to 0.05. Table II summarizes the characteristics of each of the 13 difference scenarios.

TABLE II
CTU NETWORK CHARACTERISTICS

Scenario	Time T (Hours)	# of Nodes P	# of Edges Normal Traffic	# of Hidden Nodes	# of Edges Botnet Traffic
1	7	510	1566	2280	4428
2	6	114	249	283	337
3	68	333	977	2463	2466
4	5	414	1737	9	27
5	2	246	652	59	67
6	3	200	380	2	5
7	2	93	161	11	14
8	20	3031	8799	57	106
9	6	485	1799	706	3372
10	6	260	1088	25	131
11	1	53	162	7	19
12	2	290	697	861	1829
13	17	272	814	267	345

Table III shows that the Hierarchical Poisson model initialized at the MRE solution always correctly rejects the null hypothesis when it is not true. However, the model incorrectly rejects the null hypothesis in Scenario 3. This scenario has far more nodes than any of the other scenarios, and as the number of nodes increase, the number of entries that must be estimated, $\mathcal{O}(P^2)$, vastly outweigh the number of observations, $\mathcal{O}(P)$. This gives rise to a large model misspecification error in this scenario, which would negatively impact the accuracy of Algorithm I. Like in the simulations, the Poisson MLE model always rejects the null hypothesis due to its massive model misspecification error and the TA-Map method also has poor performance in the scenarios that are computationally feasible for the method (the ones marked NA are too computationally expensive). Overall MRE-HP has good performance detecting anomalous activity, especially compared to the other methods.

TABLE III
CTU NETWORK TEST

Scenario	When $\mathcal{H}_0$ is True			When $\mathcal{H}_A$ is True		
	MRE-HP	MLE	TA-Map	MRE-HP	MLE	TA-Map
1	✓	×	NA	✓	✓	NA
2	✓	×	×	✓	✓	✓
3	×	×	NA	✓	✓	NA
4	✓	×	NA	✓	✓	NA
5	✓	×	NA	✓	✓	NA
6	✓	×	NA	✓	✓	NA
7	✓	×	×	✓	✓	✓
8	✓	×	NA	✓	✓	NA
9	✓	×	NA	✓	✓	NA
10	✓	×	NA	✓	✓	NA
11	✓	×	×	✓	✓	✓
12	✓	×	NA	✓	✓	NA
13	✓	×	NA	✓	✓	NA

In Table IV, we show the CPU timings of the algorithms for the 13 scenarios in the CTU-13 dataset under both hypothesis, where the algorithms are run on the same processor described in the simulations. Even for scenario 8, the computational times of MRE-HP are feasible despite running on a rather out-of-date processor with a low clock speed. Again we mark NA for the scenarios that are computationally infeasible for the TA-Map method (the memory requirements are above 32GB even for scenario 6). The MRE-HP method despite being slower than the TA-Map on smaller networks (see Table I), scales much more efficiently to larger networks.

TABLE IV
CTU NETWORK CPU TIMES (IN SECONDS)

Scenario	When $\mathcal{H}_0$ is True			When $\mathcal{H}_A$ is True		
	MRE-HP	MLE	TA-Map	MRE-HP	MLE	TA-Map
1	513.72	25.381	NA	1303.2	62.148	NA
2	19.258	5.2586	426.69	206.71	1.0702	683.26
3	790.60	70.706	NA	468.39	55.564	NA
4	2038.0	10.834	NA	3607.5	30.863	NA
5	539.85	2.0568	NA	263.56	3.2602	NA
6	69.452	1.6095	NA	58.427	12.556	NA
7	10.164	0.8487	360.09	17.043	0.6761	366.83
8	62602	8071.2	NA	55591	2087.8	NA
9	5439.3	97.082	NA	903.46	51.762	NA
10	648.88	7.8440	NA	174.66	2.9925	NA
11	4.2550	0.7101	55.126	17.645	0.4735	56.367
12	1864.8	3.6154	NA	514.97	5.5562	NA
13	355.20	17.620	NA	792.81	76.237	NA

C. Taxi Dataset

The proposed model was applied to a dataset consisting of yellow and green taxicabs rides from the New York City Taxi and Limousine Commission (NYC TLC) [47] and [48]. For every NYC taxicab ride, the dataset contains the pickup and drop-off locations as geographic coordinates (latitude and longitude). Green taxicabs are not allowed to pickup passengers below West 110th Street and East 96th Street in Manhattan, but occasionally they risk the chance of getting punished and ignore the regulations. In an article on June 10th 2014, the New York Post explains how the city began hiring more TLC inspectors to catch illegal pickups and enforce the location rules [49]. Thus we are interested in identifying if there are green taxicabs operating in lower Manhattan when we only know the yellow taxicab network. We treat the 18 Neighborhood Tabulation Areas (NTA) in lower Manhattan as nodes and associate any pickups or drop-offs within an NTA's boundaries as traffic entering or leaving the node. We form edges from only frequently occurring routes of traffic, which we define as having activity at least an average of every 20 minutes for yellow taxicabs and twice a month for green taxicabs. For samples, we use the yellow and green taxicab rides from between January and May of 2014 and aggregate them into daily totals.

Like in the previous example, we indirectly observe samples of the total ingress and egress of each node, and the total flows of 10 interior nodes that each observe the flows of $0.7P$ nodes. This creates a total traffic network with $P = 18$ nodes and 187 non-zero edges (39% sparsity) where the baseline network (yellow taxicab rides) has 163 of the edges. There is anomalous activity (green taxicab rides) on 56 of the edges, where 32 of these edges are also in the baseline network and 24 are not. We observe the network for a total of $T = 150$ days. Fig. 8 shows the baseline network formed from yellow taxicab rides and the unknown anomalous activity due to illegal pickups from green taxicabs.

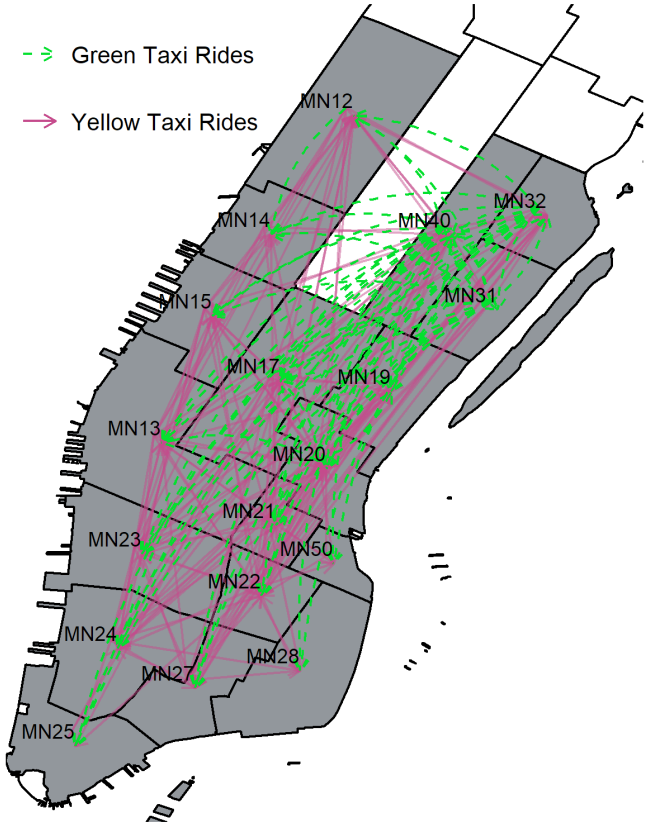


Fig. 8. A network of taxicab rides in lower Manhattan where the nodes are the 18 NTAs. The traffic from yellow taxicab rides (solid purple lines) form the baseline network and the traffic from green taxicab rides (dashed green lines) are anomalous activity in the network.

Table V shows, for different percentages of edges observed, whether the correct decision (reject or not) is made when the null hypothesis is true (no green taxi traffic) and when it is not true (green taxi traffic). The Hierarchical Poisson model initialized at the MRE solution always makes the correct decision while the Poisson MLE model, except for when the network can be directly observed, always rejects the null hypothesis. These two models are tested at an alpha level of 0.05. The Traffic and Anomaly Map method, which has a 0.05 Type-I error threshold chosen from the ROC curves of the simulations, also has poor performance.

From the results of Table V, we know the Hierarchical Poisson model initialized at the MRE solution is always able to detect changes in the network at a global scale, but we are also interested in the recovery of the individual green taxicab routes. When 70% of the network is observed, the model is able to detect 52 of the 56 edges that contain anomalous

TABLE V
TAXI NETWORK TEST

% Edges	When $\mathcal{H}_0$ is True			When $\mathcal{H}_A$ is True		
	MRE-HP	MLE	TA-Map	MRE-HP	MLE	TA-Map
0	✓	×	✓	✓	✓	×
10%	✓	×	✓	✓	✓	×
20%	✓	×	✓	✓	✓	×
30%	✓	×	✓	✓	✓	×
40%	✓	×	✓	✓	✓	×
50%	✓	×	✓	✓	✓	×
60%	✓	×	×	✓	✓	×
70%	✓	×	×	✓	✓	✓
80%	✓	×	×	✓	✓	✓
90%	✓	×	×	✓	✓	✓
100%	✓	✓	×	✓	✓	✓

activity with only a 2% false positive rate. The 4 missed edges and 5 false alarms are shown in Fig. 9.

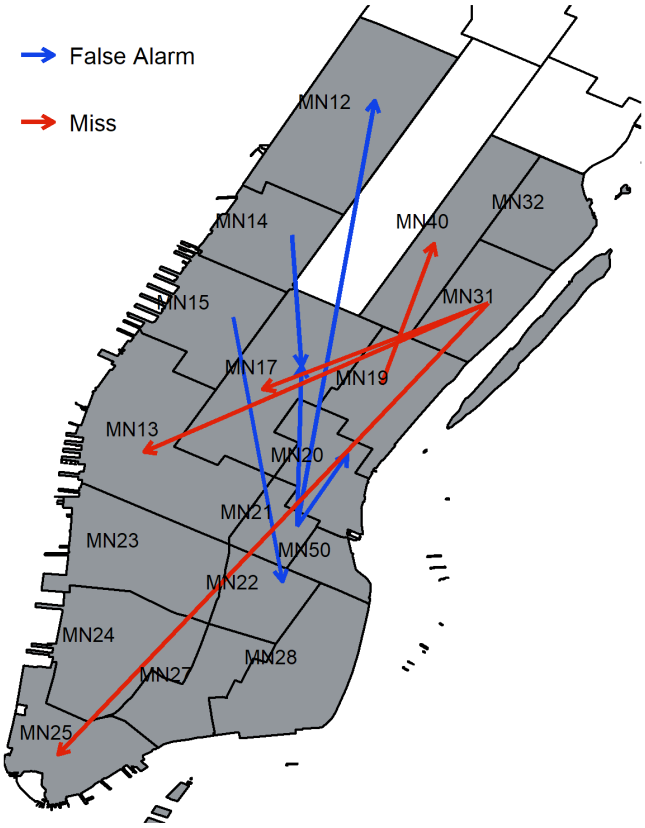


Fig. 9. A miss (red line) is an edge that the MRE-HP model fails to identify as containing anomalous activity and a false alarm (blue line) is an edges that is incorrectly identified as containing anomalous activity. The majority of the misses depart from MN31 (Lenox Hill and Roosevelt Island), which may contain legal activity because green taxis are allowed to pick up passengers from Roosevelt Island.

Out of the 4 misses, 3 of them are from green taxicab pickups from MN31, which contains the Lenox Hill and Roosevelt Island areas. Green taxicabs are allowed to pick up passengers from Roosevelt Island, but not from Lenox Hill, so some of the traffic on these 3 routes could be legal and not anomalous activity. The other miss, from MN19 to MN40, only had 11 rides in 150 days, making it harder to distinguish from just perturbation noise in the samples.

VII. CONCLUSION

We have developed a framework and a probabilistic model for detecting anomalous activity in the traffic rates of sparse networks. Our framework is realistic and robust in that, at minimum, it only requires observing the total egress and ingress of the nodes. Because it imposes no fixed assumptions of edge structure, our framework allows the estimator to handle noisy observations and anomalous activity. Our simulation results show the advantages of our model over competing methods in detecting anomalous activity. Through application of our models to the CTU-13 botnet datasets, we show that the model is scalable and robust to various scenarios, and with the NYC taxi dataset, we show an application of our model and framework to an already identified real-world problem.

APPENDIX

Proof of Proposition 1. By Jensen's inequality, $\log\left(\mathrm{P}(\mathcal{D}|\mathbf{\Lambda})\right)$

$$
\begin{aligned}
&= \log\left(\prod_{t_1=1}^{T}\mathrm{P}(\boldsymbol{R}^{t_1}|\mathbf{\Lambda})\prod_{t_2=1}^{T}\mathrm{P}(\boldsymbol{C}^{t_2}|\mathbf{\Lambda})\prod_{t_3=1}^{T}\mathrm{P}(\boldsymbol{F}^{t_3}|\mathbf{\Lambda})\right)\\
&\geq \sum_{t_1=1}^{T}\mathrm{E}_{\mathrm{q}^{t_1}}\left(\log\mathrm{P}(\boldsymbol{R}^{t_1};\boldsymbol{N}^{t_1}|\mathbf{\Lambda})\right)-\mathrm{E}_{\mathrm{q}^{t_1}}\left(\log\mathrm{q}(\boldsymbol{N}^{t_1})\right)\\
&+\sum_{t_2=1}^{T}\mathrm{E}_{\mathrm{q}^{t_2}}\left(\log\mathrm{P}(\boldsymbol{C}^{t_2};\boldsymbol{N}^{t_2}|\mathbf{\Lambda})\right)-\mathrm{E}_{\mathrm{q}^{t_2}}\left(\log\mathrm{q}(\boldsymbol{N}^{t_2})\right)\\
&+\sum_{t_3=1}^{T}\mathrm{E}_{\mathrm{q}^{t_3}}\left(\log\mathrm{P}(\boldsymbol{F}^{t_3};\boldsymbol{N}^{t_3}|\mathbf{\Lambda})\right)-\mathrm{E}_{\mathrm{q}^{t_3}}\left(\log\mathrm{q}(\boldsymbol{N}^{t_3})\right)\\
&=\sum_{t_1=1}^{T}\mathrm{E}_{\mathrm{q}^{t_1}}\left(\log\mathrm{P}(\boldsymbol{R}^{t_1}|\boldsymbol{N}^{t_1};\mathbf{\Lambda})\right)+\sum_{t_2=1}^{T}\mathrm{E}_{\mathrm{q}^{t_2}}\left(\log\mathrm{P}(\boldsymbol{C}^{t_2}|\boldsymbol{N}^{t_2};\mathbf{\Lambda})\right)\\
&+\sum_{t_3=1}^{T}\mathrm{E}_{\mathrm{q}^{t_3}}\left(\log\mathrm{P}(\boldsymbol{F}^{t_3}|\boldsymbol{N}^{t_3};\mathbf{\Lambda})\right)\\
&+\sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T}\mathrm{E}_{\mathrm{q}^{t_\tau}}\left(\log\mathrm{P}(\boldsymbol{N}^{t_\tau}|\mathbf{\Lambda})\right)+\mathrm{H}\left(\mathrm{q}(\boldsymbol{N}^{t_\tau})\right)
\end{aligned}
$$

and $\mathrm{P}(\boldsymbol{R}^{t_1}|\boldsymbol{N}^{t_1};\mathbf{\Lambda}) = \mathrm{P}(\boldsymbol{C}^{t_2}|\boldsymbol{N}^{t_2};\mathbf{\Lambda}) = \mathrm{P}(\boldsymbol{F}^{t_3}|\boldsymbol{N}^{t_3};\mathbf{\Lambda}) = 1$. The inequality is tight (by KL divergence) when $\mathrm{q}(\boldsymbol{N}^{t_1}) = \mathrm{P}(\boldsymbol{N}^{t_1}|\boldsymbol{R}^{t_1};\mathbf{\Lambda})$, $\mathrm{q}(\boldsymbol{N}^{t_2}) = \mathrm{P}(\boldsymbol{N}^{t_2}|\boldsymbol{C}^{t_2};\mathbf{\Lambda})$, and $\mathrm{q}(\boldsymbol{N}^{t_3}) = \mathrm{P}(\boldsymbol{N}^{t_3}|\boldsymbol{F}^{t_3};\mathbf{\Lambda})$ are multinomial distributions. □

Proof of Theorem 1.
Define $\mathcal{N} = \{\boldsymbol{N}^{t_\tau}: \forall t_\tau = 1,\ldots,T \text{ and } \tau = 1,\ldots,3\}$ as the set of all network traffic at different time points t_τ for the entire sample window $1,\ldots,T$. So $\cap\mathcal{N}$ is the intersection

of the set and $\mathrm{P}(\cap\mathcal{N})$ is its joint probability. By Jensen's inequality, $\log \mathrm{P}(\boldsymbol{\Lambda}|\mathcal{D})$

$$
\begin{aligned}
&= \log \int \mathrm{P}(\boldsymbol{\Lambda}, \boldsymbol{\epsilon}|\mathcal{D})\, d\boldsymbol{\epsilon} = \log \int \frac{\mathrm{P}(\mathcal{D}|\boldsymbol{\Lambda}, \boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\mathcal{D})}\, d\boldsymbol{\epsilon} \\
&= \log \int \left(\int \cdots \int \mathrm{P}\left(\mathcal{D}, \cap\mathcal{N}|\boldsymbol{\Lambda}, \boldsymbol{\epsilon}\right) \frac{\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\mathcal{D})} d\mathcal{N} \right) d\boldsymbol{\epsilon} \\
&= \log \int \mathrm{E}_{\mathrm{q}} \left(\frac{\mathrm{P}(\mathcal{D}, \cap\mathcal{N}|\boldsymbol{\Lambda}, \boldsymbol{\epsilon})}{\prod_{\tau=1}^{3}\prod_{t_\tau=1}^{T} \mathrm{q}(\boldsymbol{N}^{t_\tau})} \frac{\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\mathcal{D})} \right) d\boldsymbol{\epsilon} \\
&\geq \log \int \exp \Bigg\{ \mathrm{E}_{\mathrm{q}} \Bigg(\log \mathrm{P}(\mathcal{D}, \cap\mathcal{N}|\boldsymbol{\Lambda}, \boldsymbol{\epsilon}) + \log \mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{\epsilon}) \\
&\qquad + \log \frac{\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\mathcal{D})} - \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \log \mathrm{q}(\boldsymbol{N}^{t_\tau}) \Bigg) \Bigg\} d\boldsymbol{\epsilon} \\
&= \log \int \exp \left\{ \mathrm{E}_{\mathrm{q}} \left(\log \frac{\mathrm{P}(\mathcal{D}, \cap\mathcal{N}, \boldsymbol{\Lambda}|\boldsymbol{\epsilon})}{\mathrm{P}(\mathcal{D}, \cap\mathcal{N}|\boldsymbol{\epsilon})} \frac{\mathrm{P}(\mathcal{D}, \cap\mathcal{N}|\boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\mathcal{D})} \right) \right\} d\boldsymbol{\epsilon} \\
&+ \log \left(\exp \left\{ -\mathrm{E}_{\mathrm{q}} \left(\sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \log \mathrm{q}(\boldsymbol{N}^{t_\tau}) \right) \right\} \right) \\
&= \log \int \exp \left\{ \mathrm{E}_{\mathrm{q}} \left(\log \mathrm{P}(\boldsymbol{\Lambda}| \cap \mathcal{N}, \boldsymbol{\epsilon}) \right) + \mathrm{E}_{\mathrm{q}} \left(\log \mathrm{P}(\cap\mathcal{N}, \boldsymbol{\epsilon}|\mathcal{D}) \right) \right\} d\boldsymbol{\epsilon} \\
&+ \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{H}\left(\mathrm{q}(\boldsymbol{N}^{t_\tau}) \right) \qquad (9) \\
&= \log \int \exp \left\{ \mathrm{E}_{\mathrm{q}} \left(\log \mathrm{P}(\boldsymbol{\Lambda}| \cap \mathcal{N}, \boldsymbol{\epsilon}) \right) + \mathrm{E}_{\mathrm{q}} \left(\log \mathrm{P}(\boldsymbol{\epsilon}| \cap \mathcal{N}) \right) \right\} d\boldsymbol{\epsilon}
\end{aligned}
$$

where this bound is tight (by KL divergence) when $\mathrm{q} = \mathrm{P}(\cap\mathcal{N}|\mathcal{D}, \boldsymbol{\Lambda}, \boldsymbol{\epsilon})$
$= \prod_{\tau=1}^{3}\prod_{t_\tau=1}^{T} \mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{R}^{t_\tau}, \boldsymbol{\Lambda})\mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{C}^{t_\tau}, \boldsymbol{\Lambda})\mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{F}^{t_\tau}, \boldsymbol{\Lambda})$ are multinomial distributions. And, maximizing $\mathrm{E}_{\mathrm{q}} \left(\log \mathrm{P}(\cap\mathcal{N}, \boldsymbol{\epsilon}|\mathcal{D}) \right)$

$$
\begin{aligned}
&= \mathrm{E}_{\mathrm{q}} \left(\log \mathrm{P}(\mathcal{D}| \cap \mathcal{N}, \boldsymbol{\epsilon}) + \log \mathrm{P}(\cap\mathcal{N}, \boldsymbol{\epsilon}) - \log \mathrm{P}(\mathcal{D}) \right) \\
&= \mathrm{E}_{\mathrm{q}} \left(\log(1) + \log \mathrm{P}(\cap\mathcal{N}|\boldsymbol{\epsilon}) \right) + \log \mathrm{P}(\boldsymbol{\epsilon}) - \log \mathrm{P}(\mathcal{D}) \\
&= \log \mathrm{P}(\boldsymbol{\epsilon}) - \log \mathrm{P}(\mathcal{D}) + \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} \log \mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{\epsilon})
\end{aligned}
$$

is equivalent to maximizing a lower bound of $\log \mathrm{P}(\boldsymbol{\epsilon}|\mathcal{D})$

$$
\begin{aligned}
&= \log \frac{\prod_{t_1=1}^{T}\prod_{t_2=1}^{T}\prod_{t_3=1}^{T} \mathrm{P}(\boldsymbol{R}^{t_1}, \boldsymbol{C}^{t_2}, \boldsymbol{F}^{t_3}|\boldsymbol{\epsilon})\mathrm{P}(\boldsymbol{\epsilon})}{\mathrm{P}(\mathcal{D})} \\
&= \log \mathrm{P}(\boldsymbol{\epsilon}) - \log \mathrm{P}(\mathcal{D}) + \log \prod_{t=1}^{T} \mathrm{E}_{\mathrm{q}^{t_1}} \left(\frac{\mathrm{P}(\boldsymbol{R}^{t_1}, \boldsymbol{N}^{t_1}|\boldsymbol{\epsilon})}{\mathrm{q}^{t_1}(\boldsymbol{N}^{t_1})} \right) \\
&+ \log \prod_{t_2=1}^{T} \mathrm{E}_{\mathrm{q}^{t_2}} \left(\frac{\mathrm{P}(\boldsymbol{C}^{t_2}, \boldsymbol{N}^{t_2}|\boldsymbol{\epsilon})}{\mathrm{q}^{t_2}(\boldsymbol{N}^{t_2})} \right) + \log \prod_{t_3=1}^{T} \mathrm{E}_{\mathrm{q}^{t_3}} \left(\frac{\mathrm{P}(\boldsymbol{F}^{t_3}, \boldsymbol{N}^{t_3}|\boldsymbol{\epsilon})}{\mathrm{q}^{t_3}(\boldsymbol{N}^{t_3})} \right) \\
&\geq \log \mathrm{P}(\boldsymbol{\epsilon}) - \log \mathrm{P}(\mathcal{D}) + \sum_{t_1=1}^{T} \mathrm{E}_{\mathrm{q}^{t_1}} (\log \mathrm{P}(\boldsymbol{R}^{t_1}|\boldsymbol{N}^{t_1})\boldsymbol{\epsilon}) \\
&+ \sum_{t_2=1}^{T} \mathrm{E}_{\mathrm{q}^{t_2}} \left(\log \mathrm{P}(\boldsymbol{C}^{t_2}|\boldsymbol{N}^{t_2}, \boldsymbol{\epsilon}) \right) + \sum_{t_3=1}^{T} \mathrm{E}_{\mathrm{q}^{t_3}} \left(\log \mathrm{P}(\boldsymbol{F}^{t_3}|\boldsymbol{N}^{t_3}, \boldsymbol{\epsilon}) \right) \\
&+ \sum_{\tau=1}^{3}\sum_{t^{(\tau)}=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} \left(\log \mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{\epsilon}) \right) - \mathrm{E}_{\mathrm{q}^{t_\tau}} \left(\log \mathrm{q}(\boldsymbol{N}^{t_\tau}) \right) \\
&\propto \log \mathrm{P}(\boldsymbol{\epsilon}) - \log \mathrm{P}(\mathcal{D}) + \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} \left(\log \mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{\epsilon}) \right)
\end{aligned}
$$

for any distributions of $\mathrm{q}(\boldsymbol{N}^{t_1}), \mathrm{q}(\boldsymbol{N}^{t_2}), \mathrm{q}(\boldsymbol{N}^{t_3})$.

Since $\boldsymbol{N}_{ij}^{t_\tau}|\epsilon_{ij} \sim NegBin(\epsilon_{ij}\Lambda_{0\,ij} + 1, \frac{1}{1+\epsilon_{ij}})$ is the negative binomial distribution and $\epsilon_{ij} \sim Unif(0, \infty)$, the M-step is $\hat{\epsilon_{ij}}$

$$
\begin{aligned}
&= \arg\max_{\epsilon_{ij}} \log \mathrm{P}(\epsilon_{ij}) + \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} \left(\log \mathrm{P}(\boldsymbol{N}_{ij}^{t_\tau}|\epsilon_{ij}) \right) \\
&\propto \arg\max_{\epsilon_{ij}} \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} \left(\log \Gamma(N_{ij}^{t_\tau} + \epsilon_{ij}\Lambda_{0\,ij} + 1) \right) \\
&+ \log(\epsilon_{ij}) 3T(\epsilon_{ij}\Lambda_{0\,ij} + 1) - \log(1 + \epsilon_{ij}) 3T(\epsilon_{ij}\Lambda_{0\,ij} + 1) \\
&- 3T \log \Gamma(\epsilon_{ij}\Lambda_{0\,ij} + 1) - \log(1 + \epsilon_{ij}) \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} (N_{ij}^{t_\tau}) \\
&\geq \arg\max_{\epsilon_{ij}} 3T \left((\epsilon_{ij}\Lambda_{0\,ij} + 1) \log \frac{\epsilon_{ij}}{1 + \epsilon_{ij}} - \log \Gamma(\epsilon_{ij}\Lambda_{0\,ij} + 1) \right) \\
&+ \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \log \Gamma(\mathrm{E}_{\mathrm{q}^{t_\tau}} (N_{ij}^{t_\tau}) + \epsilon_{ij}\Lambda_{0\,ij} + 1) \\
&- \log(1 + \epsilon_{ij}) \sum_{\tau=1}^{3}\sum_{t^\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} (N_{ij}^{t_\tau})
\end{aligned}
$$

and given estimates of the hyperparameters $\hat{\epsilon}_{ij}$, estimators for the rates $\hat{\Lambda}_{ij}$

$$
\begin{aligned}
&= \arg\max_{\Lambda_{ij}} \mathrm{E}_{\mathrm{q}} \left(\log \mathrm{P}(\cap\mathcal{N}|\boldsymbol{\Lambda}, \hat{\boldsymbol{\epsilon}}) + \log \mathrm{P}(\boldsymbol{\Lambda}|\hat{\boldsymbol{\epsilon}}) - \log \mathrm{P}(\cap\mathcal{N}) \right) \\
&\propto \arg\max_{\Lambda_{ij}} \log \mathrm{P}(\boldsymbol{\Lambda}|\hat{\boldsymbol{\epsilon}}) + \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} \left(\log \mathrm{P}(\boldsymbol{N}^{t_\tau}|\boldsymbol{\Lambda}) \right) \\
&\propto \arg\max_{\Lambda_{ij}} (\hat{\epsilon}_{ij}\Lambda_{0\,ij}) \log(\Lambda_{ij}) - \hat{\epsilon}_{ij}\Lambda_{ij} - 3T\Lambda_{ij} \\
&+ \sum_{\tau=1}^{3}\sum_{t_\tau=1}^{T} \mathrm{E}_{\mathrm{q}^{t_\tau}} (N_{ij}^{t_\tau})
\end{aligned}
$$

Thus when $\mathrm{E}_{\mathrm{q}^{t_1}}(N_{ij}^{t}) = \mathrm{E}(N_{ij}^{t_1}|\boldsymbol{R}^{t_1}, \hat{\boldsymbol{\Lambda}}^{k})$ where $\hat{\boldsymbol{\Lambda}}^{k}$ are the previous iterations' estimators for the rate matrix, the lower bound will push up against the observed log posterior $\log \mathrm{P}(\boldsymbol{\Lambda}|\mathcal{D})$. This makes the E-step just the means of the independent Multinomial distributions $\prod_{i=1}^{P} Multi(R_i^{t_1}, \frac{\hat{\Lambda}_{i1}^{k}}{\sum_{j=1}^{P}\hat{\Lambda}_{ij}^{k}}, \ldots, \frac{\hat{\Lambda}_{iP}^{k}}{\sum_{j=1}^{P}\hat{\Lambda}_{ij}^{k}})$ like in the previous models. The same holds when given the column sums $\boldsymbol{C}^{t_2}$ or flows $\boldsymbol{F}^{t_3}$. □

Proof of Proposition 2. The positive estimator $\hat{\boldsymbol{\Lambda}}$ that maximizes the MRE distribution is the solution to

$$
\begin{aligned}
&= \arg\max_{\boldsymbol{\Lambda} \in \mathbb{R}^{+}} \log \left(\mathrm{P}(\boldsymbol{\Lambda}|\boldsymbol{R}, \boldsymbol{C}, \boldsymbol{F}) \right) \\
&= \arg\max_{\boldsymbol{\Lambda} \in \mathbb{R}^{+}} \log(\prod_{ij} \exp\{-|\Lambda_{ij} - \Lambda_{0ij}|\}) - \log(Z(\boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\phi})) \\
&+ \log(\exp\{\hat{\boldsymbol{\rho}}'(\boldsymbol{\Lambda}\mathbf{1} - \bar{\boldsymbol{R}}) + \hat{\boldsymbol{\gamma}}'(\mathbf{1}'\boldsymbol{\Lambda} - \bar{\boldsymbol{C}}) + \hat{\boldsymbol{\phi}}'(\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{B} - \bar{\boldsymbol{F}})\}) \\
&= \arg\max_{\boldsymbol{\Lambda} \in \mathbb{R}^{+}} -\sum_{ij} |\Lambda_{ij} - \Lambda_{0ij}| + \hat{\boldsymbol{\rho}}'(\boldsymbol{\Lambda}\mathbf{1} - \bar{\boldsymbol{R}}) + \hat{\boldsymbol{\gamma}}'(\boldsymbol{\Lambda}'\mathbf{1} - \bar{\boldsymbol{C}})
\end{aligned}
$$

$$
\begin{aligned}
&+ \hat{\boldsymbol{\phi}}'(\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{B} - \bar{\boldsymbol{F}}) \\
&= \underset{\boldsymbol{\Lambda}\in\mathbb{R}^+}{\arg\min} ||\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_0||_1 - \hat{\boldsymbol{\rho}}'(\boldsymbol{\Lambda}\mathbf{1} - \bar{\boldsymbol{R}}) - \hat{\boldsymbol{\gamma}}'(\boldsymbol{\Lambda}'\mathbf{1} - \bar{\boldsymbol{C}}) \\
&- \hat{\boldsymbol{\phi}}'(\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{B} - \bar{\boldsymbol{F}})
\end{aligned}
$$

where $||\cdot||_1$ is the element wise ℓ_1 norm and the optimal Lagrange multipliers $\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}$ are the solution to

$$
\begin{aligned}
&= \underset{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}}{\arg\max} \; -\log\left(Z(\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi})\right) \qquad (10) \\
&= \underset{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}}{\arg\max} \sum_{i=1}^{P} \rho_i \bar{R}_i + \sum_{j=1}^{P} \gamma_j \bar{C}_j + \sum_h \phi_h \bar{F}_h - \log 2 \\
&- \sum_{ij} \Lambda_{0ij}(\rho_i + \gamma_j + \sum_h \phi_h A_{hi} B_j) \\
&+ \log(1 + LM_{ij}) + \log(1 - LM_{ij}) \\
&= \underset{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}}{\arg\max} \sum_{i=1}^{P} \rho_i (\bar{R}_i - \sum_{j=1}^{P} \Lambda_{0ij}) + \sum_{j=1}^{P} \gamma_j (\bar{C}_j - \sum_{i=1}^{P} \Lambda_{0ij}) \\
&+ \sum_h \phi_h (\bar{F}_h - \sum_{ij} A_{hi} \Lambda_{0ij} B_j) + \sum_{ij} \log(1 - LM_{ij}^2)
\end{aligned}
$$

where $LM_{ij} = \rho_i + \gamma_j + \sum_h \phi_h A_{hi} B_j$.

The Lagrangian of the loss function in (6) is $||\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_0||_1 + \boldsymbol{\rho}'(\boldsymbol{\Lambda}\mathbf{1} - \bar{\boldsymbol{R}}) + \boldsymbol{\gamma}'(\mathbf{1}'\boldsymbol{\Lambda} - \bar{\boldsymbol{C}}) + \boldsymbol{\phi}'(\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{B} - \bar{\boldsymbol{F}})$ with optimal Lagrange multipliers that are the solution to dual problem

$$
\begin{aligned}
&= \underset{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}}{\arg\max} \; -\sum_{ij} f^*(-\rho_i - \gamma_j - \sum_h \phi_h A_{hi} B_j) - \sum_{i=1}^{P} \rho_i \bar{R}_i \\
&- \sum_{j=1}^{P} \gamma_j \bar{C}_j - \sum_h \phi_h \bar{F}_h \\
&= \underset{\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\phi}}{\arg\max} \sum_{ij} \Lambda_{0ij}(LM_{ij}) - \sum_{i=1}^{P} \rho_i \bar{R}_i - \sum_{j=1}^{P} \gamma_j \bar{C}_j - \sum_h \phi_h \bar{F}_h \\
&\qquad \text{subject to } |LM_{ij}| < 1 \;\forall i, j
\end{aligned}
$$

because $f^*(-\rho_i - \gamma_j - \sum_h \phi_h A_{hi} B_j)$ are the convex conjugates defined as

$$
\begin{aligned}
&= \max_{\Lambda_{ij}} - \Lambda_{ij}(\rho_i + \gamma_j + \sum_h \phi_h A_{hi} B_j) - |\Lambda_{ij} - \Lambda_{0ij}| \\
&= \max_{\Lambda_{ij}} \begin{cases} \Lambda_{0ij} - \Lambda_{ij}(1 + LM_{ij}) & \text{if} \quad \Lambda_{ij} \geq \Lambda_{0ij} \\ \Lambda_{ij}(1 - LM_{ij}) - \Lambda_{0ij} & \text{if} \quad \Lambda_{ij} < \Lambda_{0ij} \end{cases} \\
&= \begin{cases} \infty & \text{if} \quad |\rho_i + \gamma_j + \sum_h \phi_h A_{hi} B_j| > 1 \\ -\Lambda_{0ij}(\rho_i + \gamma_j + \sum_h \phi_h A_{hi} B_j) & \text{otherwise.} \end{cases}
\end{aligned}
$$

The dual can be relaxed with log barrier terms to an unconstrained problem that is equivalent to (10) making minimizing the Lagrangian of (6) for $\boldsymbol{\Lambda}$ equivalent to maximizing the MRE distribution.

□

Proof of Proposition 3. Using Remark 1.7 of [50], then for regular models, the MAP estimator will have the same asymptotic properties as the MLE. Thus, the standard proof for the asymptotic distribution for the log likelihood ratio [51] applies to the log posterior density ratio. □

Proof of Proposition 4. Let $\mathcal{M}(\boldsymbol{\Lambda}^*)$ be the true model, then the test statistic ψ

$$
\begin{aligned}
&= -2 \sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda}_0)) - \log(\mathcal{M}_t(\hat{\boldsymbol{\Lambda}})) \\
&= -2 \left(\sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda}_0)) - \max_{\boldsymbol{\Lambda}\in\mathbb{R}^+} \sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda})) \right) \\
&= 2 \sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda}^*)) - \log(\mathcal{M}_t(\boldsymbol{\Lambda}_0)) \\
&- 2 \min_{\boldsymbol{\Lambda}\in\mathbb{R}^+} \sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda}^*)) - \log(\mathcal{M}_t(\boldsymbol{\Lambda}))
\end{aligned}
$$

and as $T \to \infty$, ψ/T

$$
\begin{aligned}
&\to 2\,\mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}^*)||\mathcal{M}(\boldsymbol{\Lambda}_0)\right) - 2 \min_{\boldsymbol{\Lambda}\in\mathbb{R}^+} \mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}^*)||\mathcal{M}(\boldsymbol{\Lambda})\right) \\
&= 2\,\mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}^*||\mathcal{M}(\boldsymbol{\Lambda}_0))\right) = \Psi
\end{aligned}
$$

The misspecified test statistic $\hat{\psi}$

$$
\begin{aligned}
&= -2 \sum_{t=1}^{T} \log(\hat{\mathcal{M}}_t^k(\boldsymbol{\Lambda}_0)) - \log(\hat{\mathcal{M}}_t^k(\hat{\boldsymbol{\Lambda}})) \\
&= -2 \sum_{t=1}^{T} \log(\hat{\mathcal{M}}_t^k(\boldsymbol{\Lambda}_0)) - \max_{\boldsymbol{\Lambda}\in\mathbb{R}^+} \sum_{t=1}^{T} \log(\hat{\mathcal{M}}_t^k(\boldsymbol{\Lambda})) \\
&= 2 \sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda}^*)) - \log(\mathcal{M}_t(\boldsymbol{\Lambda}_0)) \qquad (11) \\
&\quad + 2 \sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda}_0)) - \log(\hat{\mathcal{M}}_t^k(\boldsymbol{\Lambda}_0)) \\
&\quad - 2 \sum_{t=1}^{T} \log(\mathcal{M}_t(\boldsymbol{\Lambda}^*)) - \log(\hat{\mathcal{M}}_t^k(\hat{\boldsymbol{\Lambda}}^*)) \\
&\quad - 2 \min_{\boldsymbol{\Lambda}\in\mathbb{R}^+} \sum_{t=1}^{T} \log(\hat{\mathcal{M}}_t^k(\hat{\boldsymbol{\Lambda}}^*)) - \log(\hat{\mathcal{M}}_t^k(\boldsymbol{\Lambda}))
\end{aligned}
$$

and as $T \to \infty$, $\hat{\psi}/T$

$$
\begin{aligned}
&\to 2\,\mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}^*)||\mathcal{M}(\boldsymbol{\Lambda}_0)\right) + 2\,\mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}_0)||\hat{\mathcal{M}}^k(\boldsymbol{\Lambda}_0)\right) \\
&- 2\,\mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}^*)||\hat{\mathcal{M}}^k(\hat{\boldsymbol{\Lambda}}^*)\right) - 2 \min_{\boldsymbol{\Lambda}\in\mathbb{R}^+} \mathrm{KL}\left(\hat{\mathcal{M}}^k(\hat{\boldsymbol{\Lambda}}^*)||\hat{\mathcal{M}}^k(\boldsymbol{\Lambda})\right) \\
&= \Psi - 2\left(\mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}^*)||\hat{\mathcal{M}}^k(\hat{\boldsymbol{\Lambda}}^*)\right) - \mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}_0)||\hat{\mathcal{M}}^k(\boldsymbol{\Lambda}_0)\right)\right)
\end{aligned}
$$

where $\Psi = 2\,\mathrm{KL}\left(\mathcal{M}(\boldsymbol{\Lambda}^*||\mathcal{M}(\boldsymbol{\Lambda}_0))\right)$ and $\hat{\mathcal{M}}^k(\hat{\boldsymbol{\Lambda}}^*)$ is the closest population local maximum at iteration k. If as $k \to \infty$, the EM model $\hat{\mathcal{M}}^k$ converges to the true model $\mathcal{M}$, then $\hat{\psi}/T \to \Psi$

□

REFERENCES

[1] Y. Vardi, "Network tomography: Estimating source-destination traffic intensities from link data," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 365–377, 1996.

[2] A. Coates, A. O. H. III, R. Nowak, and B. Yu, "Internet tomography," *IEEE Signal processing magazine*, vol. 19, no. 3, pp. 47–65, 2002.

[3] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, "Traffic matrix estimation: Existing techniques and new directions," in *Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, 2002, pp. 161–174.

[4] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu, "Network tomography: Recent developments," *Statistical science*, pp. 499–517, 2004.

[5] E. Lawrence, G. Michailidis, V. N. Nair, and B. Xi, "Network tomography: A review and recent developments," in *Frontiers in statistics*, 2006, pp. 345–366.

[6] M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, and Y. Tsang, "Maximum likelihood network topology identification from edge-based unicast measurements," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 30, no. 1, 2002, pp. 11–20.

[7] R. Cáceres, N. G. Duffield, J. Horowitz, and D. F. Towsley, "Multicast-based inference of network-internal loss characteristics," *IEEE Transactions on Information theory*, vol. 45, no. 7, pp. 2462–2480, 1999.

[8] M. Rabbat, R. Nowak, and M. Coates, "Multiple source, multiple destination network tomography," in *INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, 2004, pp. 1628–1639.

[9] Y. Tsang, M. Coates, and R. D. Nowak, "Network delay tomography," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2125–2136, 2003.

[10] M.-F. Shih and A. O. Hero, "Unicast-based inference of network link delay distributions with finite mixture models," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2219–2228, 2003.

[11] ——, "Hierarchical inference of unicast network topologies based on end-to-end measurements," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 1708–1718, 2007.

[12] N. Duffield, "Network tomography of binary network performance characteristics," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5373–5388, 2006.

[13] C. Tebaldi and M. West, "Bayesian inference on network traffic using link count data," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 557–573, 1998.

[14] J. Cao, S. V. Wiel, B. Yu, and Z. Zhu, "A scalable method for estimating network traffic matrices from link counts," Tech. Rep., 2000.

[15] J. Cao, D. Davis, S. V. Wiel, and B. Yu, "Time-varying network tomography: Router link data," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1063–1075, 2000.

[16] J. Zhang and I. C. Paschalidis, "Statistical anomaly detection via composite hypothesis testing for markov models," *IEEE Transactions on Signal Processing*, vol. 66, no. 3, pp. 589–602, Feb 2018.

[17] R. J. Vanderbei and J. Iannone, "An EM approach to OD matrix estimation," Tech. Rep., 1994.

[18] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast accurate computation of large-scale ip traffic matrices from link loads," in *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '03, 2003, pp. 206–217.

[19] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," in *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '03, 2003, pp. 301–312.

[20] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices (extended version)," *Networking, IEEE/ACM Transactions on*, vol. 20, no. 3, pp. 662–676, June 2012.

[21] M. Mardani, G. Mateos, and G. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *Information Theory, IEEE Transactions on*, vol. 59, no. 8, pp. 5186–5205, Aug 2013.

[22] M. Mardani and G. Giannakis, "Estimating traffic and anomaly maps via network tomography," *Networking, IEEE/ACM Transactions on*, vol. 24, no. 3, pp. 1–15, June 2016.

[23] A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 201–206.

[24] ——, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 4, 2004, pp. 219–230.

[25] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, 2005, pp. 30–30.

[26] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of pca for traffic anomaly detection," in *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2007, pp. 109–120.

[27] H. Kasai, W. Kellerer, and M. Kleinsteuber, "Network volume anomaly detection and identification in large-scale networks based on online time-structured traffic tensor tracking," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 636–650, Sept 2016.

[28] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina, "Detection and identification of network anomalies using sketch subspaces," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, 2006, pp. 147–152.

[29] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-based change detection: methods, evaluation, and applications," in *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, 2003, pp. 234–247.

[30] M. Thottan and C. Ji, "Anomaly detection in ip networks," *IEEE Transactions on signal processing*, vol. 51, no. 8, pp. 2191–2204, 2003.

[31] Y. Gu, A. McCallum, and D. Towsley, "Detecting anomalies in network traffic using maximum entropy estimation," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*. USENIX Association, 2005, pp. 32–32.

[32] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, "Anomaly detection in dynamic networks: a survey," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 3, pp. 223–247, 2015.

[33] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[34] H. E. Egilmez and A. Ortega, "Spectral anomaly detection using graph-based filtering for wireless sensor networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1085–1089.

[35] M. Khatua, S. H. Safavi, and N. Cheung, "Sparse laplacian component analysis for internet traffic anomalies detection," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 4, pp. 697–711, Dec 2018.

[36] Y.-J. Lee, Y.-R. Yeh, and Y.-C. F. Wang, "Anomaly detection via online oversampling principal component analysis," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 7, pp. 1460–1470, 2013.

[37] K. P. Murphy, *Machine Learning: A Pobabilistic Perspective*. MIT press, 2012.

[38] X. Yi and C. Caramanis, "Regularized em algorithms: A unified framework and statistical guarantees," in *Advances in Neural Information Processing Systems*, 2015, pp. 1567–1575.

[39] S. Kullback, *Information theory and statistics*, 1997.

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2006.

[41] Y. Altun and A. Smola, "Unifying divergence minimization and statistical inference via convex duality," in *International Conference on Computational Learning Theory*, 2006, pp. 139–153.

[42] O. Koyejo and J. Ghosh, "A representation approach for relative entropy minimization with expectation constraints," in *ICML WDDL workshop*, 2013.

[43] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar 2014, http://cvxr.com/cvx.

[44] S. Basu, "Bayesian hypotheses testing using posterior density ratios," *Statistics & probability letters*, vol. 30, no. 1, pp. 79–86, 1996.

[45] G. Casella and R. Berger, *Statistical Inference*, ser. Duxbury advanced series. Duxbury Thomson Learning, 2002.

[46] S. Garcia, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," pp. 100–123, 2014, http://mcfp.weebly.com/the-ctu-13-dataset-a-labeled-dataset-with-botnet-normal-and-background-traffic.html.

[47] NYC Taxi & Limousine Commission, "TLC trip record data," http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

[48] T. W. Schneider, "Unified new york city taxi and uber data," Github, 2017, https://github.com/toddwschneider/nyc-taxi-data.

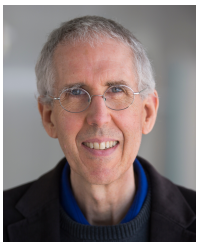
[49] R. Harshbarger, "Tlc cracking down on drivers who illegally pick up street hails," *New York Post*, June 10 2014. [Online]. Available: https://nypost.com/2014/06/10/tlc-cracking-down-on-drivers-who-illegally-pick-up-street-hails/

[50] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009, vol. 25.

[51] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

Elizabeth Hou received her B.A. degree in Statistics from the University of California, Berkeley, in 2012 and her M.A. degree in Statistics from the University of Michigan, Ann Arbor, in 2015. She is currently pursuing a Ph.D. degree in Electrical and Computer Engineering at the University of Michigan, Ann Arbor. She is a Consortium for Verification Technology fellow. She has had research internships at Los Alamos National Laboratory in 2015, 2016, 2017, and 2018. Her research interests include statistical machine learning, sequential learning, Bayesian inference, and anomaly detection.

Yasin Yılmaz (S11-M14) received the Ph.D. degree in Electrical Engineering from Columbia University, New York, NY, in 2014. He is currently an Assistant Professor of Electrical Engineering at the University of South Florida, Tampa. His research interests include statistical signal processing, machine learning, and their applications to cybersecurity, cyber-physical systems, IoT networks, communication systems, energy systems, transportation systems, social and environmental systems. He received the Collaborative Research Award from Columbia University in 2015, and Research Initiation Award from the Southeastern Center for Electrical Engineering Education in 2017.



Alfred O. Hero III is the John H. Holland Distinguished University Professor of Electrical Engineering and Computer Science and the R. Jamison and Betty Williams Professor of Engineering at the University of Michigan, Ann Arbor. He is also the Co-Director of the Universitys Michigan Institute for Data Science (MIDAS) . His primary appointment is in the Department of Electrical Engineering and Computer Science and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. He received the B.S. (summa cum laude) from Boston University (1980) and the Ph.D from Princeton University (1984), both in Electrical Engineering. He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE). He has served as President of the IEEE Signal Processing Society and as a member of the IEEE Board of Directors. He has received numerous awards for his scientific research and service to the profession including several best paper awards, the IEEE Signal Processing Society Technical Achievement Award in 2013 and the 2015 Society Award, which is the highest career award bestowed by the IEEE Signal Processing Society. He has received a Rackham Distinguished Faculty Achievement Award in 2011 and the 2017 Stephen S. Attwood Excellence in Engineering Award, from the University of Michigan. Alfred Heros recent research interests are in high dimensional spatio-temporal data, multi-modal data integration, statistical signal processing, and machine learning. Of particular interest are applications to social networks, network security and forensics, computer vision, and personalized health.