Hypergraph-Based Anomaly Detection in Very Large Networks

Jorge Silva, *Member, IEEE,* and Rebecca Willett, *Member, IEEE*

Abstract

This paper addresses the problem of detecting anomalous interactions or traffic within a very large network using a limited number of unlabeled observations. In particular, consider n recorded interactions among p nodes, where p may be very large relative to n. A novel method based on using a hypergraph representation of the data is proposed to deal with this very high-dimensional, "big p, small n" problem. Hypergraphs constitute an important extension of graphs which allows edges to connect more than two vertices simultaneously. An algorithm for detecting anomalies directly on the corresponding discrete space, without any feature selection or dimensionality reduction, is presented. The algorithm has $O(np)$ computational complexity, making it ideally suited for very large networks, and requires no tuning, bandwidth or regularization parameters.

The distribution of the data is modeled as a two-component mixture, consisting of a "nominal" and an "anomalous" component. The deviance of each observation from nominal behavior, as well as the mixture parameters, are learned using Expectation-Maximization (EM), assuming a multivariate Bernoulli variational approximation. This approach is related to probability mass function level set estimation and is shown to allow False Discovery Rate control. The identifiability of the underlying distribution, the local consistency of the EM algorithm, and the avoidance of singular solutions are proved. The proposed approach is validated on high-dimensional synthetic data and it is shown that, for a useful class of data distributions, it can outperform other state-of-the-art methods.

I. Introduction

Many important problems in networks, be they computer, social, biological, or other types of networks, are commonly tackled using data structures based on graphs and associated theoretical results. Graphs are a well established discrete mathematical tool for representing connectivity data with irregular topology. It is convenient to use graphs for tasks as varied as anomaly detection [1], semi-supervised classification [2], traffic matrix estimation [3], dimensionality reduction [4], and many others.

However, graphs cannot encode potentially critical information about *ensembles* of networked nodes interacting together. Despite a wealth of theoretical work in graph theory (for instance, see [5] and [6]), graphs are simply not a sufficiently rich structure in many contexts. Consider the following example: we make several observations of groups of people meeting, and wish to recognize some pattern in these meetings or detect unusual meetings. Since each meeting consists of several (potentially more than two) people, pairwise connections between people

Jorge Silva and Rebecca Willett are with the Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University. This work was supported by NSF CAREER Award No. CCF-06-43947 and DARPA Grant No. HR0011-07-1-003.

encoded by graphs only represent a portion of the real information collected and available for analysis.

This calls for a new paradigm in network traffic analysis, and this paper proposes an approach based on *hypergraphs*. Hypergraphs [7] are generalizations of graphs, where the notion of an edge is generalized to that of a *hyperedge*, which may connect more than two vertices (e.g. more than two people in a social network or more than two nodes in a route taken by a packet in a network). Many of the main theoretical results for graphs are directly applicable to hypergraphs. In fact, many theorems on graphs are proven using the more general hypergraph structure [7].

Assume that there exist p vertices, corresponding to p network nodes, and that we observe n messages, or interactions where there is co-occurrence of some of the vertices. Then, each interaction can be represented as a hyperedge in the network hypergraph. This paper addresses the problem of detecting anomalous interactions, with special emphasis on the case when $p \gg n$ and when p may be in the hundreds or even thousands, *without intermediate feature selection or dimensionality reduction*.

The definition of "anomaly" can sometimes be taken out of the hands of the learning system, if labeled examples are provided during training: this then becomes a supervised *classification* problem. However, such an approach is not always possible, either because labeled examples are scarce, or because the nature of the anomalies changes rapidly over time, which is relevant in contexts such as network intrusion. Therefore, we will focus on the unsupervised setting where "anomalous" is taken to signify "unusual" or "rare". We are interested in interactions in the network that occur with very low probability. In particular, we want to identify hypergraph edges that have very little probability mass, when the number of observations is small and estimating the probability mass function (pmf) over the entire space of hyperedges is challenging, if not infeasible, in terms of statistical robustness and computational efficiency.

We start, in Section II, by introducing the hypergraph representation and formulating the problem of anomaly detection on the corresponding discrete space. We then provide, in Section III, a brief review of relevant prior work and the current state-of-the-art methods available for anomaly detection on networks. Prior work on the annotation of observations with a measure indicating the degree of anomalousness is highlighted in Section IV; these annotations are based on the positive False Discovery Rate (pFDR) [8] from a hypothesis testing perspective. Next, in Sections V and VI we propose a variational approximation to the pmf and a resulting $O(np)$ variational Expectation-Maximization (EM) algorithm that automatically learns: (i) the parameters of a finite mixture model for the distribution of the observed data; (ii) posterior probabilities of observations being anomalous. We address theoretical issues, such as identifiability of the mixture model and convergence of the EM algorithm, in Section VII. Section VIII shows experimental results that demonstrate the algorithm's performance in comparison to other state-of-the art anomaly detection algorithms. To conclude, in Section IX we discuss the results and take into account how the variational approximation affects the class of distributions (which we denote as $\mathcal{F}$) that can be well estimated. Namely, we look into new directions that aim at enriching $\mathcal{F}$, while retaining the attractive computational properties – such as $O(np)$ complexity – of the variational approximation.

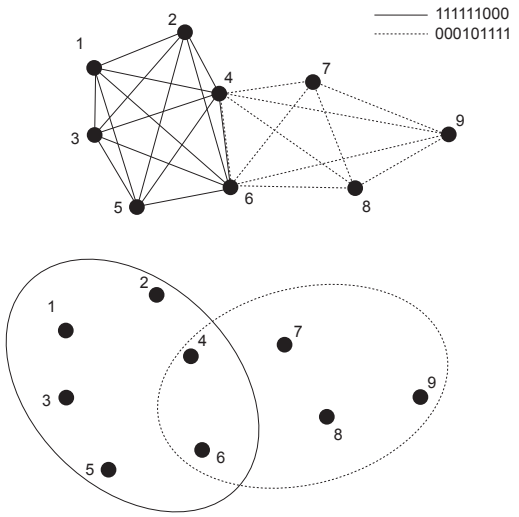


Fig. 1

MODELING TWO OBSERVATIONS, 111111000 AND 000101111, WITH $p = 9$, USING A GRAPH (TOP) AND A HYPERGRAPH (BOTTOM). WITH THE GRAPH, REPRESENTING ONE OBSERVATION OF AN INTERACTION REQUIRES MULTIPLE EDGES. WITH A HYPERGRAPH, ONE HYPEREDGE SUFFICES. THE HYPERGRAPH IS MORE EFFICIENT FOR STORING/REPRESENTING OBSERVATIONS AND MORE INFORMATIVE ABOUT THE REAL STRUCTURE OF THE DATA.

II. ANOMALY DETECTION ON HYPERGRAPHS

Let $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$ be a hypergraph [7] with vertex set $\mathcal{V}$ and hyperedge set $\mathcal{E}$. Each hyperedge, denoted $\boldsymbol{x} \in \mathcal{E}$, can be represented as a binary string of length p. Bits set to 1 correspond to vertices that participate in the hyperedge. In this setting, we may approximately equate $\mathcal{E}$ with $\{0, 1\}^p$, i.e. the binary hypercube of dimension p. (We say "approximately" due to the existence of prohibited hyperedges, namely the origin, $\boldsymbol{x} = \mathbf{0}$, and all $\boldsymbol{x}$ within Hamming distance 1 of the origin, which correspond to interactions between zero or one network nodes. The impact of this precluded set becomes negligible for very large p and is omitted from this paper for simplicity of presentation.) This is a finite set with 2^p elements. We define $g(\boldsymbol{x})$ to be the probability mass function (pmf) over $\mathcal{E}$, evaluated at $\boldsymbol{x}$.

Hypergraphs provide a more natural representation than graphs for multiple co-occurrence data of the type examined in this paper. For example, one could consider using a graph to represent co-occurrence data by having each vertex represent a network node and using weighted edges to connect vertices associated with observed co-occurrences. As Figure 1 illustrates, using a graph in this manner would imply connecting any pair of vertices appearing in an observation with an edge. The edge structure of a graph is usually represented as a $p \times p$ symmetric adjacency matrix with $\frac{p}{2}(p-1)$ distinct elements, so that even converting observations into a collection edge weights could be enormously challenging computationally. As Figure 1 illustrates, two observations can

efficiently be represented using only two hyperedges, but would require significantly more edges in a traditional graph-based representation. Furthermore, the graph data structure as described above does not encapsulate information about how often more than two vertices may interact simultaneously, and instead reduces the data to an overly simple pairwise representation.

If the data consists of a multiset $\mathcal{X}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ containing n observed, and possibly repeated, interactions $\boldsymbol{x}_i$, with each $\boldsymbol{x}_i$ an independent realization of a random variable $X \in \mathcal{E}$, then one might be tempted to simply form the histogram of the $\boldsymbol{x}_i$. However, given that in our problem $p \gg n$, histogram estimation would lead to every single new $\boldsymbol{x}$ not in the training set being branded an anomaly. Since, in a practical setting, the dimension p might be in the thousands, the curse of dimensionality guarantees that $\mathcal{X}_n$ can never adequately cover $\mathcal{E}$. The finite nature of our space of interest confers it no immunity to the curse, which instead appears in the guise of an exponential explosion in the numbers of histogram bins: 2^p, when we have a sample size of $n \ll p$.

The usual way of exorcizing the curse is to perform some kind of regularization, i.e. to restrict the class of pmf estimates to be "simple" in some sense. This can be accomplished, for example, by defining some smoothness measure on a graph and then favoring smoother pmfs. It is likely that the sample is not sufficiently dense to properly estimate highly varying functions; however, defining such smoothness measures is a difficult problem in its own right. Furthermore, the amount of regularization that strikes the right balance between under- or oversmoothing is not trivial to achieve.

An additional source of difficulty is the *contamination* of the training set. If we assume that $\mathcal{X}_n$ only contains examples of normal, or *nominal* behavior, then our problem is essentially pmf estimation, where we must threshold the estimated pmf, denoted $\widehat{f}_n(\boldsymbol{x})$, at some appropriate level. If, however, $\mathcal{X}_n$ is contaminated with some unknown proportion of anomalies, then it is more appropriate to assume the following mixture model:

$$g(\boldsymbol{x}) = (1-\pi)f(\boldsymbol{x}) + \pi\mu(\boldsymbol{x}) \tag{1}$$

where the overall pmf g of the observed data is a mixture of the nominal distribution f and an anomalous distribution μ with proportion π. This type of mixture model is sometimes called *semi-parametric* in the case when a nonparametric procedure is used to obtain estimates for f. To make it possible to learn this mixture, it is necessary to make assumptions on the component distributions f and μ. It is assumed here that μ is known and equal to the uniform distribution on $\mathcal{E}$. Assuming that μ is uniform can be shown to be the optimal choice, in terms of maximizing the worst-case detection rate, among all possible anomalous distributions [9], [10].

A realization $\boldsymbol{x}$ of $\boldsymbol{X}$ is called an anomaly if it is drawn from the distribution μ instead of the nominal distribution f. We can now define the anomalous set of hyperedges as

$$\mathcal{A}^* = \{\boldsymbol{x} : (1-\pi)f(\boldsymbol{x}) < \alpha\pi\mu(\boldsymbol{x})\},$$

where α is a parameter that controls the tradeoff between false positives and false negatives. This leads to the definition of the (unobserved) binary random variable $Y = I_{\boldsymbol{X}\sim\mu}$, where I denotes the indicator function. Define $\eta(\boldsymbol{x}) \equiv P(Y = 1|\boldsymbol{X} = \boldsymbol{x}, f, \pi)$; we note that it is possible to write $\mathcal{A}^* = \left\{\boldsymbol{x} : \eta(\boldsymbol{x}) > \frac{1}{1+\alpha}\right\}$. Since f and π are unknown, this means that the distribution

of Y, η and $\mathcal{A}^*$ are unknown as well. In this paper, we use the iid data $\mathcal{X}_n$ to estimate η and hence $\mathcal{A}^*$.

III. Related work

When labeled data is available, anomaly detection has often been treated as a classification problem in which no attempt is made to learn anything about the underlying pmf. In this type of approach, sometimes called "discriminative" as opposed to "generative", we could potentially apply an off-the-shelf classifier such as the Support Vector Machine (SVM). For problems where examples are available from one class only, there is a well-known variant: the one-class SVM (OCSVM) [11]. Our setting is somewhat different, since we have unlabeled examples from both classes, although that might be addressed through the use of slack variables in the OCSVM framework, as they would be needed regardless for the purpose of regularization. Note that an approach based on the OCSVM is very sensitive to the choice of kernel and bandwidth parameter, and also that is has, at best, $O(n^2p)$ computational complexity. We will show that it is possible, and more informative, to tackle the problem of detecting anomalies through pmf estimation, even in very high dimensional spaces. In particular, this approach allows us to control performance criteria such as the pFDR.

Another common approach is to estimate the underlying distribution g and then threshold it. When there exists no particular functional form that can be assumed for $f(\boldsymbol{x})$, it is common to use nonparametric methods, kernel pmf estimation (better known as kernel density estimation (KDE) in the continuous case) being one of the most widespread. Recall that a kernel pmf estimate is usually of the form

$$\widehat{f}_n(\boldsymbol{x}) \;=\; \frac{1}{nh^p}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right) \tag{2}$$

where $K(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h})$ is an *a priori* defined kernel function with bandwidth parameter h. In this form, it is assumed that the kernel is radially symmetric, with equal bandwidth in all dimensions and across the entire domain of $\boldsymbol{x}$. For the specific case of the binary hypercube defined by our hyperedge space $\mathcal{E}$, a kernel based on the Hamming distance $|\boldsymbol{x}-\boldsymbol{x}_i|$ has been proposed by [12], leading to a pmf estimate of the form

$$\widehat{f}_n(\boldsymbol{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} h_n^{|\boldsymbol{x}-\boldsymbol{x}_i|}(1-h_n)^{p-|\boldsymbol{x}-\boldsymbol{x}_i|}, \tag{3}$$

where $h_n \in [\frac{1}{2}, 1]$.

This approach, however, has several disadvantages. First, as in the OCSVM, it is hard to estimate the best bandwidth. This can be tackled by cross-validation, but that is typically a computationally expensive procedure. Second, estimating measures of sets under a kernel pmf estimate is computationally intractable for large p and precludes FDR control. Moreover, the error performance of KDE can be poor in the big p small n setting [13]. Finally, the computational complexity of KDE is, at best, $O(n^2p)$.

Kernel pmf estimates are known to be universally, strongly consistent, subject to the conditions that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. These asymptotic, large sample results ensure that the estimate will, eventually, converge to the true pmf with probability one, although that may be of little use with finite sample size, especially in very high dimensions. The condition $h \to 0$ also implies that the bandwidth must change with sample size n. Relaxing either radial symmetry, or the spatial invariance, or both, can improve performance, but this comes at the price of making the choice of bandwidth even harder. Examples of kernel-based pmf estimation approaches for anomaly detection include [9], who work in general Euclidean space with a mixture model similar to our own.

Another type of approach, based on graph theoretic results, can be found in [1], where it is shown that a K-point minimum spanning tree (K-MST) can be used to estimate the nominal set (i.e. the complement of $\mathcal{A}^*$). A K-MST is a tree which connects K vertices and has the least sum of edge weights. Once such a tree has been estimated from the data, using a greedy algorithm (exact computation is computationally intractable), vertices outside of the K-MST can be considered anomalies. A computationally efficient, $O(pn^2 \log n)$, variant of this approach, called a leave-one-out k nearest-neighbor graph (L1O-kNNG), is proposed in [1]. One of the benefits of the L1O-kNNG approach is that bounds on the false alarm level can be derived as a function of K. However, this method is best suited to problems in which the network connectivity graph is known. Moreover, it is not clear how to account for contamination in the dataset.

More thorough surveys of anomaly detection in networks, with an emphasis on network security, may be found in [14] and [15]. Scalability of methods to large p is a significant challenge for the vast majority of these approaches.

IV. Annotations

The related work described above either (a) makes a hard decision about whether each observation is an anomaly and does not provide any additional information or (b) provides an estimate of the distribution underlying the data but does not provide a simple mechanism for detecting anomalies from this pmf estimate. One of the key facets of the approach proposed in this paper is the annotation of observations. The annotations, which we assume to be scalars in the $[0, 1]$ interval, allow the observations to be ranked, and provide some measure of how anomalous they appear to be under the model. Most methods in the existing literature cannot readily accomplish this task.

Starting from premises similar to our own, [9] propose first learning the mixing parameter π separately and then assigning annotations γ_i to each $\boldsymbol{x}_i$, equal to

$$\gamma_i = 1 - \text{pFDR}(\mathcal{A}_i), \tag{4}$$

where pFDR is the positive false discovery rate [8] associated with the set $\mathcal{A}_i$, which is defined as

$$\mathcal{A}_i = \arg \max_{\mathcal{A} \subseteq \{0,1\}^p} \left\{ \mathbb{U}(\mathcal{A}) : f(\boldsymbol{x}) < f(\boldsymbol{x}_i), \forall \boldsymbol{x} \in \mathcal{A} \right\}, \tag{5}$$

with $\mathbb{U}$ denoting μ-measure. Note that $\mathcal{A}_i$ can be thought of as largest set of anomalous (i.e. low probability mass) hyperedges which excludes observation $\boldsymbol{x}_i$, so that larger $\mathcal{A}_i$ suggests that

$\boldsymbol{x}_i$ is less anomalous. Further note that $\mathcal{A}_i$ is the *complement* of the minimum volume level set of f that includes $\boldsymbol{x}_i$ and that the $\mathcal{A}_i$'s constitute a collection of nested level sets of f. To see the relationship between the $\mathcal{A}_i$'s and $\mathcal{A}^*$, let $\mathcal{A}_{(k)}$ denote the k^{th} largest $\mathcal{A}_i$ according to the μ-measure (i.e. $\mathcal{A}_{(k)}$ corresponds to the $\boldsymbol{x}_i$ with the k^{th} largest $f(\boldsymbol{x}_i)$). Then there exists some k^* such that

$$\mathcal{A}_{(1)} \supseteq \mathcal{A}_{(2)} \supseteq \cdots \supseteq \mathcal{A}_{(k^*-1)} \supseteq \mathcal{A}^* \supseteq \mathcal{A}_{(k^*)} \cdots \supseteq \mathcal{A}_{(n)};$$

in other words, the level set $\mathcal{A}^*$ contains a nested collection of the $\mathcal{A}_i$'s. The value of k^* depends on α, the parameter which controls for the compromise between false alarms and detection failures, and on the mixture parameters f, π. The pFDR, for some set $\mathcal{A}$, is defined as follows:

$$\text{pFDR}(\mathcal{A}) = P\{\boldsymbol{X} \sim f | \boldsymbol{X} \in \mathcal{A}\}. \tag{6}$$

Thus, if we declare observations $\boldsymbol{x}$ that lie in $\mathcal{A}_i$ to be "discovered" anomalies, then pFDR$(\mathcal{A}_i)$ is the probability that those observations arise from the nominal distribution f. It can be shown that

$$\gamma_i = \pi \mathbb{U}(\mathcal{A}_i)/\mathbb{G}(\mathcal{A}_i), \tag{7}$$

where $\mathbb{G}(.)$ refers to the probability measure associated with g. Denoting the f-measure by $\mathbb{F}(.)$, we can estimate the γ_i by

$$\widehat{\gamma_i} = \frac{\widehat{\pi}\widehat{\mathbb{U}}(\mathcal{A}_i)}{\widehat{\mathbb{G}}(\mathcal{A}_i)} \quad = \quad \frac{\widehat{\pi}\widehat{\mathbb{U}}(\mathcal{A}_i)}{(1-\widehat{\pi})\widehat{\mathbb{F}}(\mathcal{A}_i) + \widehat{\pi}\widehat{\mathbb{U}}(\mathcal{A}_i)}. \tag{8}$$

Since, for non-trivial sets $\mathcal{A}_i$, these empirical probability measures cannot be obtained in practice (because that would require enumerating the 2^p elements of the hypercube), they must be estimated via Monte Carlo methods. If we can obtain m samples $\boldsymbol{z}$ from f, then the empirical measure

$$\widehat{\mathbb{F}}(\mathcal{A}_i) = \frac{1}{m}\sum_{l=1}^{m} I_{\boldsymbol{z}_l \in \mathcal{A}_i}$$

is an estimator of $\mathbb{F}(\mathcal{A}_i)$. The same can be done for $\mathbb{U}$.

Unfortunately, most methods for estimating the pmf do not provide an "easy to sample" form of the pmf. Drawing samples from distributions estimated using nonparametric methods such as kernel estimation require involved MCMC techniques whose convergence is hard to assess. To counteract this issue and other computational complexity bottlenecks, we propose a variational approximation to f which results in a variational EM algorithm for estimating both the mixture components of g and the posterior probabilities ($\eta(\boldsymbol{x}_i)$ for each $i = 1, \ldots, n$). By choosing the variational approximation to be fully factorized, it becomes very easy to obtain samples from both f and μ and hence to estimate the annotations described above. In the following section we describe the variational approximation of f.

V. VARIATIONAL APPROXIMATION

As a computationally efficient alternative to kernel pmf estimation, we choose an estimate of f from the class $\mathcal{F}$ of distributions with the following properties:

(a) each $\widetilde{f} \in \mathcal{F}$ can be expressed as the product of its marginals, so that

$$\widetilde{f}(\boldsymbol{x}) \;=\; \prod_{j=1}^{p} \widetilde{f}_j(x_j), \tag{9}$$

where each $\widetilde{f}_j : \{0,1\} \longrightarrow \mathbb{R}_0^+$ is a bona fide probability mass function that sums to one, with x_j a realization of a binary random variable X_j that corresponds to the participation of the j^{th} network node being observed in an interaction, and

(b) members of $\mathcal{F}$ have no uniform marginals (this condition ensures identifiability, as we discuss below).

Assume, for now, that we know which observations come from f, and that we wish to use them to obtain an estimate $\widehat{f}_n = \prod_{j=1}^{p} \widehat{f}_{n,j}$, where $\widehat{f}_{n,j}(\boldsymbol{x}) \equiv \widehat{f}_{n,j}(x_j)$ are given by

$$\widehat{f}(x_j) \;=\; \frac{\sum_{i=1}^{n} I_{x_{i,j}=x_j}}{n}. \tag{10}$$

In our hypergraph setting, where the space of possible hyperedges can be represented as vertices of the p-dimensional hypercube, the natural way to obtain the estimated marginals $\widehat{f}_{n,j}(x_j)$ is to use two-bin, $\{0,1\}$ histograms, which, like all histograms (subject to conditions on the bin size, which do not apply here) are well-known to be consistent estimators. As for the overall consistency of $\widehat{f}_n(\boldsymbol{x})$, it is unfortunately only verified when X_k is independent of (though not necessarily uncorrelated with) X_j for all $k, j \in 1, \ldots, p, k \neq j$, since in that case, assuming we have training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, which are i.i.d. draws from f, the expectation of $\widehat{f}_n(\boldsymbol{x}) \equiv \widehat{f}_n(\boldsymbol{x}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ becomes

$$\begin{aligned} E\widehat{f}_n(\boldsymbol{x}) \;&=\; \int \widehat{f}_n(\boldsymbol{x}) f(\boldsymbol{x}_1) \cdots f(\boldsymbol{x}_n) d\boldsymbol{x}_1 \cdots d\boldsymbol{x}_n \\ &=\; \int \left(\prod_{j=1}^{p} \widehat{f}_{n,j}(x_j) \right) \left(\prod_{j=1}^{p} f_j(x_{1,j}) \right) \cdots \left(\prod_{j=1}^{p} f_j(x_{n,j}) \right) \\ &\qquad\qquad (dx_{1,1} \cdots dx_{1,p}) \cdots (dx_{n,1} \cdots dx_{n,p}), \end{aligned} \tag{11}$$

where $x_{i,j}$ is the j^{th} bit of the i^{th} observation, $\boldsymbol{x}_i$. Using the separability of f due to independence of the X_j, we have

$$E\widehat{f}_n(\boldsymbol{x}) \;=\; \prod_{j=1}^{p} \int \widehat{f}_{n,j}(x_j) \left(\prod_{i=1}^{n} f_j(x_{i,j}) \right) (dx_{1,j} \cdots dx_{n,j}) \tag{12}$$

$$=\; \prod_{j=1}^{p} E\widehat{f}_{n,j}(x_j) = \prod_{j=1}^{p} f_j(x_j) = f(\boldsymbol{x}). \tag{13}$$

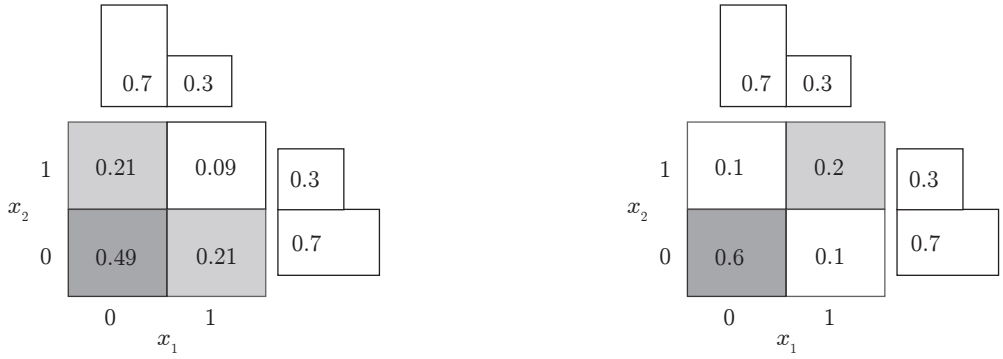


Fig. 2

EXAMPLES OF (A) ORTHOUNIMODAL (I.E UNIMODAL AND AXIS-ALIGNED) AND (B) NON-ORTHOUNIMODAL DISTRIBUTIONS, FOR $p = 2$. THE DISTRIBUTIONS HAVE THE SAME MARGINALS AND MODE $\boldsymbol{m} = 00$.

Approximating the true f^* by members of $\mathcal{F}$ is an example of a *variational approximation*, which has been used in machine learning in several contexts. For example, in Bayesian networks [16], [17], such a factorization of the class conditional densities leads to the well-known "naïve" Bayes classifier. We note the very attractive properties associated with this approximation: (a) Estimating each of the p marginals requires only the sum of $2n$ terms, which is a key factor in achieving $O(np)$ complexity. (b) There are no tuning parameters (such as bandwidth) to select. (c) Protection from overfitting comes as a natural consequence of the restricted class of estimates.

Finally, we can also characterize $\mathcal{F}$: all members of $\mathcal{F}$ are unimodal, axis-aligned distributions (although the converse is not necessarily true). To see this, first note that the marginals $\widehat{f}_{n,j}(x_j)$ are non-uniform Bernoulli distributions and, therefore, strongly unimodal as well as log-concave. We use the definitions of unimodality and log-concavity introduced in [18] for discrete distributions. Furthermore, the product of log-concave functions is log-concave. Thus, all members of $\mathcal{F}$ are log-concave and, equivalently, strongly unimodal. "Axis-aligned" is defined using the notion of orthounimodality described in [19]: if $f(\boldsymbol{x})$ is a distribution and $\boldsymbol{m} = [m_1 \dots m_p]^T$ is a mode of f, then $f(\boldsymbol{x})$ is orthounimodal iff, for $\boldsymbol{x} = [x_1, \dots, x_j, \dots, x_p]^T$ and for all j, f is monotonically non-decreasing in x_j (holding all other coordinates of $\boldsymbol{x}$ fixed) when $x_j < m_j$ and monotonically non-increasing in x_j when $x_j > m_j$. To illustrate with a simple example in the binary setting, in Figure 2 we show two different joint distributions that have the same marginals and the same mode, $\boldsymbol{m} = 00$, but where one is orthounimodal (and in $\mathcal{F}$) and the other is not, meaning that it can not be consistently estimated by our procedure. This variational approximation is used during the M-step in the variational EM algorithm proposed in the next section for estimating the entire mixture distribution g in addition to the posterior probabilities η.

VI. Estimation Method

A. Variational Expectation-Maximization

Since we do not know whether or not a given observation is an anomaly (i.e. whether it was drawn from μ), we may treat that information as the hidden random binary variable $Y \equiv I_{X\sim\mu}$. Given the dataset $\mathcal{X}_n$, the corresponding $\mathcal{Y}_n = \{y_i\}_{i=1,\ldots,N}$ may be treated as missing data and the posterior probabilities $\eta_i \equiv \eta(\boldsymbol{x}_i)$ may be estimated using EM. As is customary in the EM setting, let $\mathcal{L}(\mathcal{X}_n, \mathcal{Y}_n | f, \pi) = \sum_{i=1}^n \log p(\boldsymbol{x}_i, \boldsymbol{y}_i | f, \pi)$ be the log-likelihood, under the joint distribution $p(X, Y | f, \pi)$, of the complete data $(\mathcal{X}_n, \mathcal{Y}_n)$. This cannot be computed, since $\mathcal{Y}_n$ is missing, but the conditional expectation $E_{Y|X}\mathcal{L}$, with respect to $Y|X$, can. Omitting the conditioning on (f, π) for brevity, we have

$$\begin{aligned} E_{Y|X}\mathcal{L} &= E_{Y|X}\sum_{i=1}^n \log p(\boldsymbol{x}_i, \boldsymbol{y}_i) = \sum_{i=1}^n E_{Y|X} \log p(\boldsymbol{x}_i, \boldsymbol{y}_i) \\ &= \sum_{i=1}^n (1-\eta_i)\log p(\boldsymbol{x}_i, Y=0) + \eta_i \log p(\boldsymbol{x}_i, Y=1) \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^n (1-\eta_i)\log\left[p(\boldsymbol{x}_i|Y=0)P(Y=0)\right] + \eta_i \log\left[p(\boldsymbol{x}_i|Y=1)P(Y=1)\right] \\ &= \sum_{i=1}^n (1-\eta_i)\log\left[(1-\pi)f(\boldsymbol{x}_i)\right] + \eta_i \log\left[\pi\mu(\boldsymbol{x}_i)\right]. \end{aligned} \tag{14}$$

Alternating, at each iteration $t+1$, between maximizing $E_{Y|X}\mathcal{L}$ with respect to the η_i and to (f, π), leads to the following E- and M-steps, as first derived in [20]:

- E-step:

$$\widehat{\eta}^{(t+1)}(\boldsymbol{x}_i) = \frac{\widehat{\pi}^{(t)}\mu(\boldsymbol{x}_i)}{(1-\widehat{\pi}^{(t)})\widehat{f}^{(t)}(\boldsymbol{x}_i) + \widehat{\pi}^{(t)}\mu(\boldsymbol{x}_i)} \tag{15}$$

- M-step:

$$\begin{aligned} \widehat{\pi}^{(t+1)} &= \frac{1}{n}\sum_{i=1}^n \widehat{\eta}_i^{(t+1)} \\ \widehat{f}_j^{(t+1)}(x_{i,j}) &= \frac{\sum_{k=1}^n (1-\widehat{\eta}_i^{(t+1)}) I_{x_{k,j}=x_{i,j}}}{\sum_{k=1}^n (1-\widehat{\eta}_i^{(t+1)})} \\ \widehat{f}^{(t+1)}(\boldsymbol{x}_i) &= \prod_{j=1}^p \widehat{f}_j^{(t+1)}(x_{i,j}). \end{aligned} \tag{16}$$

In (16), $x_{i,j}$ denotes the value of the j^{th} bit of pattern $\boldsymbol{x}_i$. If a hard decision is necessary, it is natural to threshold the annotations η_i at $\frac{1}{1+\alpha}$, where α controls the tradeoff between false positives and detection failures. Recall that $\mathcal{A}^* = \left\{\boldsymbol{x} : \eta(\boldsymbol{x}) > \frac{1}{1+\alpha}\right\}$.

B. Computation of annotations

At the conclusion of the EM algorithm, we may use the estimate $\widehat{f}^{(t+1)}$ to compute the annotations γ_i for $i = 1, \ldots, n$ using a very computationally efficient Monte Carlo estimate. In particular, we sample from the fully factorized f and μ distributions – this amounts to sampling from p independent Bernoulli distributions, which can be easily done using MATLAB's `rand` command – and then compute the empirical measures of the $\mathcal{A}_i$, for each of the $\boldsymbol{x}_i$. Afterwards, we plug the estimates into (8), thus obtaining the $\widehat{\gamma}_i$.

C. A note about numerical underflow

When working in very high-dimensions, pmf values tend to become extremely small, eventually leading to underflow problems. Therefore, it is advantageous to work with the logarithm of the pmf, whenever possible. The proposed variational EM algorithm can be formulated almost entirely in terms of the log-probabilities, with the exception of the denominator in the E-step, (15), where we are faced with the sum of posterior probabilities. By using the identity

$$\log \sum_i e^{a_i} \quad = \quad \max_j a_j + \log \sum_i e^{a_i - \max_j a_j}, \tag{17}$$

the problem can be easily circumvented, provided that the absolute difference between the log-probabilities remains much smaller than the absolute value of their maximum.

D. Key advantages of proposed estimation method

The proposed variational EM algorithm has a number of key advantages:

(i) the variational approximation leads to a very computationally efficient M-step;
(ii) the γ_i's and the η_i's can be computed very easily and rapidly;
(iii) the pmf only has to be computed at the n $\boldsymbol{x}_i$ locations, rather than at all 2^p hyperedges, for anomaly detection and observation annotation;
(iv) unlike the OCSVM, the proposed method returns posterior probabilities rather than simple hard decisions;
(v) unlike KDE-based methods, a principled criterion for making a decision about each observation, based on the pFDR, is available.

VII. Theoretical properties

In this section, we build on previous groundwork which guarantees that the problem of identifying model (1) is well posed, in the sense that it has a unique solution (f, π) among all $\widetilde{f} \in \mathcal{F}$ and all $\pi \in [0, 1]$. We then show that consistency of $\widehat{f}_n$ and $\widehat{\pi}$, for the case when the true $f \in \mathcal{F}$, comes from the properties of maximum likelihood estimation and depends on the convergence of the variational EM algorithm to that estimate. For our case, we show that the regularity conditions for *local* consistency are satisfied, which means that the variational EM will converge to the global maximum of the likelihood, provided that it is initialized sufficiently close to that maximum.

A. Identifiability

A mixture model such as (1) is said to be *identifiable* when the vector $\Theta = (f, \mu, \pi)$ of mixture parameters that satisfies (1) exists and is unique. This definition can be generalized to more than two mixture components. Identifiability is not guaranteed in general and requires restrictions on f and μ. The case of mixtures of multivariate Bernoulli distributions has been addressed by, e.g. [21], [22], where it is shown that such mixtures are not, *in general*, identifiable. However, our particular model is identifiable for $p \geq 3$, as the following theorem states.

Theorem 1 (Hall & Zhou [23]): If g has the form $g(\boldsymbol{x}) = (1-\pi)\prod_{j=1}^{p} f_j(\boldsymbol{x}) + \pi \prod_{j=1}^{p} \mu_j(\boldsymbol{x})$, with $p \geq 3$, and if g is irreducible, then, up to exchanging $(1-\pi, f_1, \ldots, f_p)$ with $(\pi, \mu_1, \ldots, \mu_p)$, the parameter vector $\Theta = (f_1, \ldots, f_p, \mu_1, \ldots, \mu_p, \pi)$ is uniquely determined by g.

This result applies to discrete as well as continuous distributions, as long as the mixture g is *irreducible*, a property we shall define shortly. It does not apply to more than two mixture components, however. As stated, identifiability is assured up to exchanging the roles of $(1-\pi, f)$ and (π, μ), which is called *label switching*. Since, in our setting, μ is known (and, being uniform, factorizes in the form appropriate for the theorem), this particular ambiguity is resolved.

Definition 1 (Irreducibility [23]): The pmf g is *irreducible* if none of its bivariate marginals factorizes into the product of univariate marginals.

In our setting, the irreducibility condition is equivalent to the condition that the component f must not have any uniform marginals, i.e. $f_j \neq \mu_j$ for all $j = 1, \ldots, p$. To see that this condition is a direct result of irreducibility, and presenting the original reasoning in a slightly different format, assume that $f_j = \mu_j$ for some j, say $j = 1$ without loss of generality. Then,

$$g(\boldsymbol{x}) \;=\; \mu_1(x_1)\left[(1-\pi)\prod_{j=2}^{p} f_j(x_j) + \pi \prod_{j=2}^{p} \mu_j(x_j)\right]$$

and, integrating over all other x_j except, for example x_2, we get the bivariate marginal

$$\begin{aligned} g(x_1, x_2) &= \mu_1(x_1) \int \left[(1-\pi)\prod_{j=2}^{p} f_j(x_j) + \pi \prod_{j=2}^{p} \mu_j(x_j)\right] dx_3 \ldots dx_p \\ &= \mu_1(x_1)\left\{(1-\pi)\int \left[\prod_{j=2}^{p} f_j(x_j)\right] dx_3 \ldots dx_p + \pi \int \left[\prod_{j=2}^{p} \mu_j(x_j)\right] dx_3 \ldots dx_p\right\} \\ &= \mu_1(x_1)\left\{(1-\pi) f_2(x_2) + \pi \mu_2(x_2)\right\} \end{aligned}$$

and irreducibility is thus violated.

B. Convergence and consistency

The EM algorithm is well-known, in each iteration, to not decrease the likelihood. Also, if the true nominal density f is in the class $\mathcal{F}$ (and, hence, has no uniform marginals), then the mixture is identifiable. When $p \geq 3$, if the true f is in $\mathcal{F}$, then the maximum likelihood estimates of f and π will tend to the true f and π as $n \longrightarrow \infty$. Therefore, maximum likelihood gives (at least weakly) consistent estimates for π and f, which lead to consistent estimates for

g and for the posterior probabilities η. However, the EM algorithm is also well-known to be vulnerable to local maxima, and even to saddle points if they exist. Unlike the continuous case, in which the log likelihood may be unbounded, having singularities where it tends to $\pm\infty$, our hypergraph-based approach yields a log likelihood which is bounded above [21]; the $-\infty$ case is not problematic because the EM algorithm monotonically increases the log likelihood and hence will not be attracted to these singularities. Therefore, we are left with the problem of stationary points. Even with the variational approximation, the log-likelihood is not necessarily concave, and therefore can have stationary points other than the global maximum; we show this below.

Proposition 1 (Non-concavity of the log-likelihood): Let $g(\boldsymbol{x})$ be of the form (1), with $\boldsymbol{x} \in \mathcal{E} = \{0,1\}^p$, and let $\mathcal{X}_n$ be a sample of size n. Denote the log-likelihood as $\mathcal{L}(\mathcal{X}_n|\Theta)$, with $\Theta = (f, \pi)$. Then, $\mathcal{L}(\mathcal{X}_n|\Theta)$ is a not necessarily a concave function of Θ.

Proof: Recall that a function is log-concave if and only if its logarithm is concave. We may write the log-likelihood as a sum over the hyperedge set $\mathcal{E} = \{0,1\}^p$, as follows:

$$\mathcal{L}(\mathcal{X}_n|\Theta) = \sum_{k=1}^{2^p} n_k \log g(\boldsymbol{b}_k|\Theta),$$

where k is an index over all elements in $\mathcal{E}$, $\boldsymbol{b}_k \in \mathcal{E}$ is the k^{th} element, n_k denotes its frequency in the dataset $\mathcal{X}_n$ and $g(\boldsymbol{b}_k|\Theta)$ is the mixture $g(\boldsymbol{x})$ evaluated at $\boldsymbol{x} = \boldsymbol{b}_k$, as a function of Θ. Because the n_k are non-negative, a sufficient condition for the concavity of $\mathcal{L}$ is that each of the $g(\boldsymbol{b}_k|\Theta)$ must be log-concave. Since $g(\boldsymbol{b}_k|\Theta)$ is the sum of $\pi\mu$, which is linear in Θ and thus concave, plus $(1-\pi)f(\boldsymbol{b}_k|\Theta)$, which is the product of a concave term with f, clearly the concavity of $\mathcal{L}$ hinges on the concavity of f. In our variational approximation setting, we may write f as a product of Bernoulli distributions with parameters $\theta_j = P(x_j = 1)$, i.e.

$$f(\boldsymbol{b}_k|\Theta) = \prod_{j=1}^{p} f_j(b_{k,j}|\Theta) = \prod_{j=1}^{p} \theta_j^{b_{k,j}}(1-\theta_j)^{1-b_{k,j}}.$$

Each of the terms θ_j and $1-\theta_j$ is concave with respect to θ_j. The product of concave functions (i.e. f or $(1-\pi)f$) is not necessarily concave, although it is log-concave. Moreover, the sum with $\pi\mu$ does not preserve log-concavity [24]. Thus, $\mathcal{L}$ is not necessarily concave. ■

Proposition 1 means that we can not guarantee that the outlined EM procedure will always converge to the global maximum likelihood estimate $\widehat{\Theta}_n^{\text{ML}} = (\widehat{f}_n^{\text{ML}}, \widehat{\pi}_n^{\text{ML}})$ for all initializations, because there might be local maxima and/or saddle points. Therefore, even with a unique global maximum, there may be a risk of not reaching it for all initializations. For this reason, the EM algorithm may require random restarts; its consistency, in this case, has been shown for stochastic versions (see [25] and the accompanying discussion and references). Alternatively, we may turn to a slightly weaker form of consistency, called *local* consistency [26], [27]. Local consistency implies that EM will converge to the correct ML solution if it is initialized sufficiently close to the global maximum, subject to regularity conditions involving derivatives of g and $\log g$ up to third order. The regularity conditions are the following [27]:

- *Condition* 1. For all $\boldsymbol{\Theta} = [\Theta_1, \ldots, \Theta_{p+1}] \equiv [\pi, \theta_1, \ldots, \theta_p]^T$ in a neighborhood Ω of the true parameter vector $\boldsymbol{\Theta}^*$ and for almost all $\boldsymbol{x}$, all of the $p+1$ partial derivatives $\partial g(\boldsymbol{x}|\boldsymbol{\Theta})/\partial\Theta_k$,

$k = 1, \ldots, p$, exist and there also exist integrable functions $\phi_i(\boldsymbol{x})$, $\phi_{ij}(\boldsymbol{x})$ and $\phi_{ijk}(\boldsymbol{x})$ such that

$$\left|\frac{\partial g}{\partial \Theta_i}\right| < \phi_i, \qquad \left|\frac{\partial^2 g}{\partial \Theta_i \partial \Theta_j}\right| < \phi_{ij}, \qquad \left|\frac{\partial^3 \log g}{\partial \Theta_i \partial \Theta_j \partial \Theta_k}\right| < \phi_{ijk},$$

for i, j, k in $1, \ldots, p+1$.

- *Condition* 2. The Fisher information matrix

$$I(\boldsymbol{\Theta}) = E_{\boldsymbol{X}}\left[(\nabla_{\boldsymbol{\Theta}} \log g)^T (\nabla_{\boldsymbol{\Theta}} \log g)\right]$$

is well defined and positive definite at $\boldsymbol{\Theta} = \boldsymbol{\Theta}^*$.

It can be readily seen, by differentiating g and $\log g$ (see Appendix), that all the relevant partial derivatives exist, have no singularities, and are bounded. This is enough to satisfy Condition 1. As for Condition 2, first note that $\log g$ is sufficiently well behaved to allow us to write the Fisher information matrix as

$$I(\boldsymbol{\Theta}) = -E_{\boldsymbol{X}}\left[\frac{\partial^2 \log g}{\partial \Theta_i \partial \Theta_j}\right] = -H(\boldsymbol{\Theta})$$

for i, j, k in $1, \ldots, p+1$, where $H(\boldsymbol{\Theta})$ is the Hessian of $\log g$. For $I(\boldsymbol{\Theta}^*)$ to be positive definite, the Hessian must be negative definite at $\boldsymbol{\Theta}^*$, i.e. $\log g$ must be locally concave at $\boldsymbol{\Theta}^*$. The purpose of this condition is to guard against $+\infty$ singularities and against ridges of the log-likelihood, which might contain a continuum of solutions (including $\boldsymbol{\Theta}^*$) within which EM could potentially oscillate without necessarily converging. We have already shown that g is identifiable, which readily precludes any such multiplicity of solutions. Recalling that the log-likelihood also has no $+\infty$ singularities, as mentioned above, it follows that the conditions of Redner and Walker are satisfied and, therefore, EM is locally consistent for our two-component mixture. With this in mind, we note that in every single one of our experiments (described below), a highly accurate solution was reached.

VIII. Experiments

In order to validate our algorithm, and to illustrate one type of setting in which it will be useful, we have created a synthetic dataset consisting of a mixture of nominal and anomalous interactions among networked nodes, distributed according to (1). The dataset was split into training and test sets, each of size n. The algorithms were trained using the training set and all results below were obtained for the test set. The nominal samples were generated according to the following rule: let $\mathcal{S}$ be a subset of the vertex set $\mathcal{V}$ of the hypergraph $\mathcal{H}$. Let the elements of $\mathcal{S}$ be nodes that are active with probability p_H. Let $\bar{\mathcal{S}}$ be the complement of $\mathcal{S}$, corresponding to nodes which are active with probability p_L, and assume $p_L < \frac{1}{2} < p_H$. Intuitively, this represents a situation where a main group $\mathcal{S}$ might be active with high probability, while the background group $\bar{\mathcal{S}}$ would have much lower activity. In a social network context, this situation corresponds to the existence of a group $\mathcal{S}$ of highly active individuals and a less active, background group $\bar{\mathcal{S}}$. As another example, in a communication network, members of $\mathcal{S}$ could be high connectivity nodes or routers.

The type of distribution that arises from this situation is a unimodal cluster, centered at

$$\bar{\boldsymbol{x}} \;=\; \underbrace{11\ldots1}_{\#\mathcal{S}}\underbrace{00\ldots0}_{\#\bar{\mathcal{S}}}. \tag{18}$$

where, without loss of generality, we have reordered $\mathcal{V}$ such that the elements of $\mathcal{S}$ and $\bar{\mathcal{S}}$ appear consecutively, and where the cardinalities $\#\mathcal{S}$ and $\#\bar{\mathcal{S}}$ sum to p. To see this, note that the true distribution f is, without normalization,

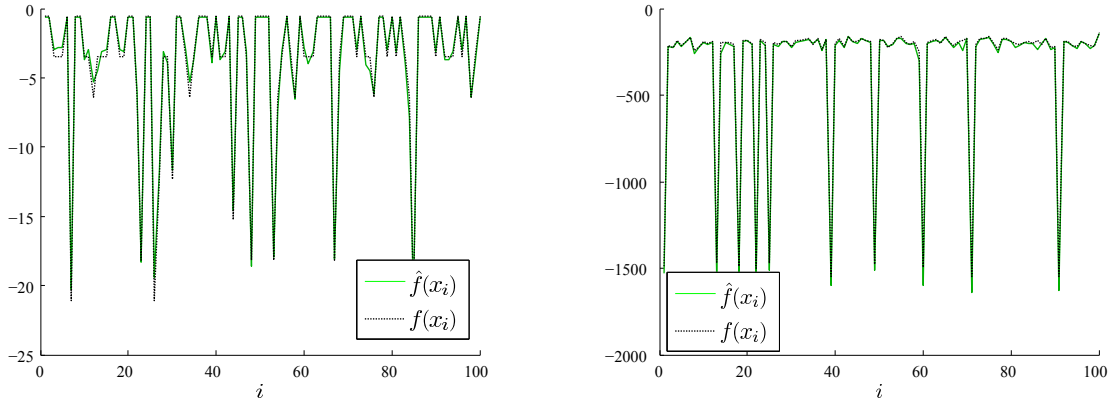
$$f(\boldsymbol{x}) \;\propto\; \prod_{j=1}^{\#\mathcal{S}} p_H^{x_j}(1-p_H)^{1-x_j} \prod_{l=\#\mathcal{S}+1}^{p} p_L^{x_l}(1-p_L)^{1-x_l}. \tag{19}$$

The expression (19) takes its maximum value when $x_j = 1$ for vertices $j \in \mathcal{S}$ and $x_l = 0$ for vertices $l \in \bar{\mathcal{S}}$, i.e., when $\boldsymbol{x} = \bar{\boldsymbol{x}}$. In the particular case when $p_L = 1 - p_H$, and denoting the Hamming distance between $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ as $|\boldsymbol{x} - \bar{\boldsymbol{x}}|$, then (19) reduces to

$$\begin{aligned} f(\boldsymbol{x}) \;&\approx\; \prod_{j=1}^{\#\mathcal{S}} p_H^{x_j}(1-p_H)^{1-x_j} \prod_{l=\#\mathcal{S}+1}^{p} (1-p_H)^{x_l} p_H^{1-x_l} \\ &=\; \prod_{j=1}^{p} p_H^{1-|I_{j\in\mathcal{S}}-x_j|}(1-p_H)^{|I_{j\in\mathcal{S}}-x_j|} \\ &=\; p_H^{p-|\bar{\boldsymbol{x}}-\boldsymbol{x}|}(1-p_H)^{|\bar{\boldsymbol{x}}-\boldsymbol{x}|}, \end{aligned} \tag{20}$$

which is clearly isotropic around $\bar{\boldsymbol{x}}$. In other words, $\bar{\boldsymbol{x}}$ corresponds to a "typical" meeting, and the likelihood of a group of people meeting under f decreases with Hamming distance from $\bar{\boldsymbol{x}}$.

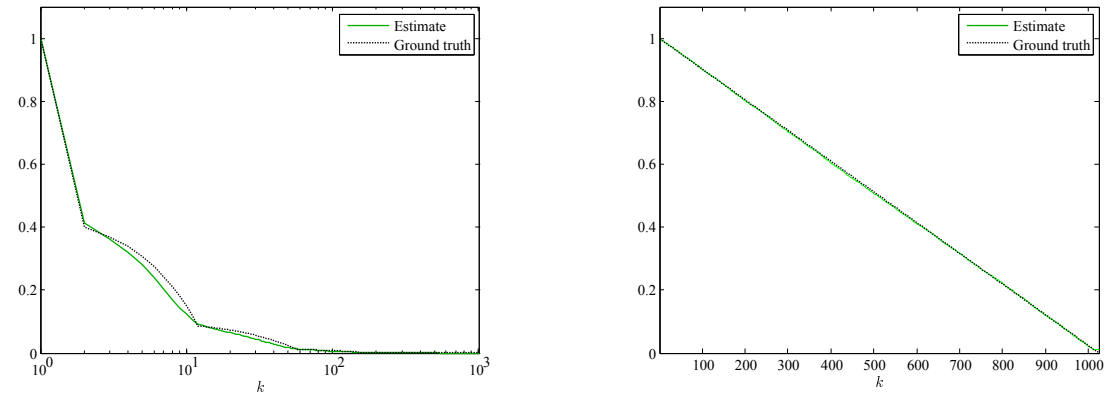
For our experiments, we have used $p = 10$ and 2000, with $n = 100$, $p_H = 0.95$, $p_L = 0.05$ and $\pi = 0.1$. The pmf estimate results from the variational EM algorithm are shown in Figure 3, which depicts the estimated and true densities, $\widehat{f}$ and f, evaluated at the test data locations. The left plot corresponds to $p = 10$ and the right plot to $p = 2000$. Also shown, in Figure 4, are the nominal measures $\mathbb{F}(\mathcal{A}_k)$ their empirical estimates $\widehat{\mathbb{F}}(\mathcal{A}_k)$, obtained by Monte Carlo with 10000 samples of $\widehat{f}_n$. Thanks to the factorized form of $\widehat{f}_n$, no Markov chain was necessary to obtain the samples and the estimates could be computed extremely rapidly. Ground truth probability masses were exhaustively computed for all hypercube vertices $\boldsymbol{b}_k$, $k = 1, \ldots, 2^p$. This was done for $p = 10$ only, since it is impractical to compute ground truth using (19) for $p = 2000$. It is clear that the proposed method is highly successful in estimating both f and the measure of its level sets. For comparison, the OCSVM [11] and L1O-kNNG [1] algorithms were applied to the same dataset, with the Gaussian kernel used for OCSVM. Note that the OCSVM treats the p-dimensional binary data as vectors in the Euclidean space $\mathbb{R}^p$. The model parameters for OCSVM are (ν,γ), where ν controls the amount of regularization and γ is the kernel bandwidth parameter. The results of these methods and the proposed approach, using test data independent from the training set, are displayed in Figure 5, where the first image corresponds to $p = 10$ and the second image corresponds to $p = 2000$. The top plot in each image corresponds to "ground truth", i.e. $y_i = I_{\{\boldsymbol{X}\sim\mu\}}$, so that the large spikes correspond to truly



(a) $p = 10$ (b) $p = 2000$

Fig. 3

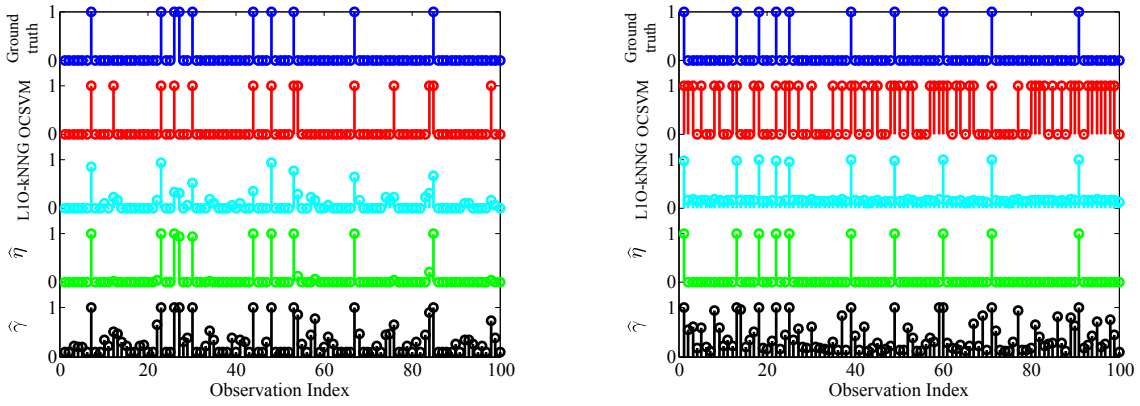
TRUE PMFS AND VARIATIONAL EM PMF ESTIMATES, WITH $n = 100$ AND $p_H = 0.95$. LEFT: $p = 10$; RIGHT: $p = 2000$. THE PLOTS SHOW $f(\boldsymbol{x_i})$ AND $\widehat{f}_n(\boldsymbol{x}_i)$, IN LOG SCALE, FOR $\boldsymbol{x}_i$ IN THE TEST SET.



(a) $\mathbb{F}(\mathcal{A}_k)$ (b) $\mathbb{U}(\mathcal{A}_k)$

Fig. 4

TRUE MEASURES $\mathbb{F}(\mathcal{A}_k)$ (LEFT) AND $\mathbb{U}(\mathcal{A}_k)$ (RIGHT) VERSUS OBSERVATION INDEX k, TOGETHER WITH EMPIRICAL ESTIMATES $\widehat{\mathbb{F}}(\mathcal{A}_k)$ AND $\widehat{\mathbb{U}}(\mathcal{A}_k)$. 10000 MONTE CARLO SAMPLES OF $\widehat{f}_n$ AND μ WERE USED, WITH $p = 10$. GROUND TRUTH WAS EXHAUSTIVELY COMPUTED FOR ALL HYPERCUBE VERTICES $\boldsymbol{b}_k$, BY USING (19).



(a) $p = 10$ (b) $p = 2000$

Fig. 5

SIMULATION RESULTS ON THE TEST SET, FOR $p = 10$ (LEFT) AND $p = 2000$ (RIGHT). HORIZONTAL AXIS IS OBSERVATION INDEX i. TOP PLOT: GROUND TRUTH, $y_i = I_{\boldsymbol{x}_i \sim \mu}$. SECOND PLOT: $\widehat{y}_i$ ESTIMATED BY OCSVM, WHICH CONTAINS A SIGNIFICANT NUMBER OF FALSE ALARMS (WITH BOTH $p = 10$ AND $p = 2000$) AND ZERO MISSED DETECTIONS. THIRD PLOT: ANOMALOUSNESS SCORES ESTIMATED BY L1O-KNNG, WHICH, WHEN THRESHOLDED AT 0.5, CONTAINS ZERO FALSE ALARMS AND ZERO MISSED DETECTIONS WITH $p = 2000$, AND A SMALL NUMBER OF MISSED DETECTIONS WITH $p = 10$. FOURTH PLOT: $\widehat{\eta}_i$ COMPUTED BY THE PROPOSED VARIATIONAL EM ALGORITHM; SETTING $\widehat{y}_i = I_{\{\widehat{\eta}_i > 1/2\}}$ RESULTS IN ZERO FALSE ALARMS AND ZERO MISSED DETECTIONS WITH BOTH $p = 10$ AND $p = 2000$. FIFTH PLOT: $\widehat{\gamma}_i$.

anomalous observations. To compare with the performance of OCSVM, we have selected the best pair (ν^{CV},γ^{CV}), with respect to 5-fold cross-validation performance, using a two-dimensional grid search over the ranges $\gamma \in \{2^{-45}, 2^{-44}, \ldots, 2^4, 2^5\}$ and $\nu \in \{2^{-20}, 2^{-19}, \ldots, 2^{-1}, 2^0\}$. As illustrated in Figure 5, the OCSVM succeeds in detecting most of the anomalies, but, unlike the variational EM algorithm, it does so at the cost of a high number of false positives. Also, if one includes the cross-validation grid search, the OCSVM is orders of magnitude slower that the proposed approach, even with a relatively small $n = 100$ (recall that the OCSVM has $O(n^2 p)$ computational complexity).

As Figure 5 shows, the L1O-kNNG algorithm performs better than the OCSVM, achieving essentially the same performance as our variational EM, in the considered dataset, for $p = 2000$, while performing slightly worse for $p = 10$. It should be taken into account, however, that the computational complexity of L1O-kNNG is, at best, $O(pn^2 \log n)$, compared to $O(np)$ for the proposed variational EM approach. Also, while the L1O-kNNG returns scalar "scores" in the $[0, 1]$ interval, they do not have a clear interpretation, particularly when the training data are contaminated as in our example. This is in contrast to the annotations computed using our proposed approach, which can be directly linked to the pFDR detection performance measure.

IX. Conclusions

This paper addresses the problem of detecting anomalous multi-node interactions in very large networks with p nodes, given a limited number, n, of recorded interactions as training data, and taking into account the possibility that an unknown proportion of the training sample may be contaminated with anomalies. We have shown that it is advantageous to use a hypergraph representation for the data, rather than the more commonly used connectivity graph, and that it is possible to overcome the curse of dimensionality, even when $p \gg n$, by restricting the class of estimates using a variational approximation.

We have proposed a scalable algorithm that detects anomalies through pmf estimation on the p-dimensional hypercube, with only $O(np)$ computational complexity. The algorithm models the data as a two-component mixture, and learns all the parameters of the mixture using Expectation-Maximization with a multivariate Bernoulli variational approximation. We have investigated the theoretical properties of the algorithm, and have demonstrated that, unlike the general mixture model case, our model is identifiable, under very mild assumptions, and that the proposed EM algorithm enjoys local consistency and is guaranteed to avoid singularities of the log-likelihood. Additionally, we have established a relationship between the posterior probabilities estimated in the E-step and other widely used measures of anomalous behavior, such as the pFDR. The proposed algorithm allows annotations (γ_i's) related to the pFDR to be computed more efficiently than alternative procedures such as methods based on kernel pmf estimation. Furthermore, the algorithm improves upon classification-based approaches like the OCSVM by providing more information than simply an all-or-nothing decision, and on kernel pmf estimation by providing principled criteria for choosing a decision threshold.

The proposed procedure has been validated on a *very* high-dimensional example dataset, and compared favorably with other state-of-the-art methods. As the results show, our method can outperform alternatives in terms of estimation error, for a useful class of distributions, while computationally scaling considerably better – in fact, linearly – with both p and n.

An interesting observation is that the proposed pmf estimation algorithm actually performs *better* for higher p, as can be seen in Figure 3. We attribute this fact to a *blessing of dimensionality*: the nominal and anomalous distributions are more separable in high dimensions, i.e. the measure of the set where f and μ have similar values vanishes when p increases, which means that a given observation $\boldsymbol{x}$ will, with high probability, be either unambiguously anomalous or unambiguously nominal. This type of phenomenon has been successfully exploited in other contexts, namely kernel-based SVM classification. As a consequence, not only did the accuracy of the estimates improve with higher p, but the number of EM iterations needed for convergence also decreased dramatically for higher dimensions. With $p > 1000$, convergence never took more than one EM step. We therefore conjecture that the behavior of the likelihood function may, in our setting, become more benign with increasing p. This is reflected in the fact that the η's quickly tend to either zero or one, although this behavior is less pronounced when p_H is set closer to 0.5 than in the reported experiments (since this makes f and μ more similar).

For further research, we identify two main issues: firstly, due to the variational approximation, the algorithm is only consistent within a family $\mathcal{F}$ of distributions, all members of which are unimodal. In order to enrich $\mathcal{F}$ so that is can consistently estimate multimodal distributions, an

evident approach would consist of adding more components to the mixture model. The impact on identifiability and consistency of using more than two components needs to be investigated, however. The second issue is that of developing an online version of the anomaly detector, which would be particularly useful when observations of interactions arrive sequentially and also when the nature of the anomalies changes over time.

APPENDIX

As stated in VII-B, local consistency requires that g satisfy regularity conditions involving the partial derivatives $\partial g/\partial\Theta_1 \dots \partial g/\partial\Theta_{p+1}$, as well as the second-order derivatives of g and the third-order derivatives of $\log g$. Straightforward (though tedious) differentiation of (1) yields, for $j, l, k = 1, \dots p$,

$$\begin{aligned}
\frac{\partial g}{\partial \pi} &= \mu(\boldsymbol{x}) - f(\boldsymbol{x}) \\
\frac{\partial g}{\partial \theta_j} &= \begin{cases} (1-\pi)\prod_{l\neq j}\theta_l^{x_l}(1-\theta_l)^{1-x_l}, & \text{if } x_j = 1 \\ -(1-\pi)\prod_{l\neq j}\theta_l^{x_l}(1-\theta_l)^{1-x_l}, & \text{if } x_j = 0 \end{cases}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 g}{\partial \pi^2} &= 0 \\
\frac{\partial^2 g}{\partial \theta_j \partial \theta_l} &= \begin{cases} 0, & \text{if } l = j \\ (1-\pi)\prod_{k\neq l, k\neq j}\theta_k^{x_k}(1-\theta_k)^{1-x_k}, & \text{if } l \neq j \text{ and } x_j = x_l \\ -(1-\pi)\prod_{k\neq l, k\neq j}\theta_l^{x_k}(1-\theta_k)^{1-x_k}, & \text{if } l \neq j \text{ and } x_j \neq x_l \end{cases} \\
\frac{\partial^2 g}{\partial \pi \partial \theta_j} = \frac{\partial^2 g}{\partial \theta_j \partial \pi} &= \begin{cases} -\pi\prod_{l\neq j}\theta_l^{x_l}(1-\theta_l)^{1-x_l}, & \text{if } x_j = 1 \\ \pi\prod_{l\neq j}\theta_l^{x_l}(1-\theta_l)^{1-x_l}, & \text{if } x_j = 0 \end{cases},
\end{aligned}$$

and, for $\log g$,

$$\begin{aligned}
\frac{\partial^3 \log g}{\partial \pi^3} &= 2\left(\frac{\mu(\boldsymbol{x}) - f(\boldsymbol{x})}{g(\boldsymbol{x})}\right)^3 \\
\frac{\partial^3 \log g}{\partial \theta_j \partial \theta_l \partial \theta_k} &= \frac{1}{g(\boldsymbol{x})^3}\left[\left(\frac{\partial^3 g}{\partial \theta_j \partial \theta_l \partial \theta_k} g(\boldsymbol{x}) + \frac{\partial^2 g}{\partial \theta_j \partial \theta_l}\frac{\partial g}{\partial \theta_k} - \frac{\partial^2 g}{\partial \theta_j \partial \theta_k}\frac{\partial g}{\partial \theta_l} - \frac{\partial g}{\partial \theta_j}\frac{\partial^2 g}{\partial \theta_l \partial \theta_k}\right) g(\boldsymbol{x}) \right. \\
&\quad \left. - \left(\frac{\partial^2 g}{\partial \theta_j \partial \theta_l} g(\boldsymbol{x}) - \frac{\partial g}{\partial \theta_j}\frac{\partial g}{\partial \theta_l}\right) 2\frac{\partial g}{\partial \theta_k}\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^3 \log g}{\partial\theta_j \partial\pi^2} &= \frac{1}{g(\boldsymbol{x})^3}\left[\left(\frac{\partial^3 g}{\partial\theta_j\partial\pi^2}g(\boldsymbol{x}) + \frac{\partial^2 g}{\partial\theta_j\partial\pi}\frac{\partial g}{\partial\pi} - \frac{\partial^2 g}{\partial\theta_j\partial\pi}\frac{\partial g}{\partial\pi}\right) g(\boldsymbol{x})\right. \\
&\quad \left. - \left(\frac{\partial^2 g}{\partial\theta_j\partial\pi}g(\boldsymbol{x}) - \frac{\partial g}{\partial\theta_j}\frac{\partial g}{\partial\pi}\right) 2\frac{\partial g}{\partial\pi}\right] \\
\frac{\partial^3 \log g}{\partial\theta_j \theta_l \partial\pi} &= \frac{1}{g(\boldsymbol{x})^3}\left[\left(\frac{\partial^3 g}{\partial\theta_j\theta_l\partial\pi}g(\boldsymbol{x}) + \frac{\partial^2 g}{\partial\theta_j\partial\theta_l}\frac{\partial g}{\partial\pi} - \frac{\partial^2 g}{\partial\theta_j\partial\pi}\frac{\partial g}{\partial\theta_l} - \frac{\partial g}{\partial\theta_j}\frac{\partial^2 g}{\partial\theta_l\partial\pi}\right) g(\boldsymbol{x})\right. \\
&\quad \left. - \left(\frac{\partial^2 g}{\partial\theta_j\partial\theta_l}g(\boldsymbol{x}) - \frac{\partial g}{\partial\theta_j}\frac{\partial g}{\partial\theta_l}\right) 2\frac{\partial g}{\partial\pi}\right],
\end{aligned}$$

where the last two lines hold regardless of the order of differentiation. The crucial point is that, in all derivatives that involve a quotient, the denominator is a power of g. In the discrete hypercube, $g(\boldsymbol{x})$ is bounded above. In order to bound it below, away from zero, we impose the additional condition $\pi > \epsilon$, where $\epsilon > 0$. This corresponds to the very mild assumption that the anomalous distribution has a non-zero proportion in the mixture. Thus, we have $\frac{\epsilon}{2^p} < g(\boldsymbol{x}) < 1$ and all the derivatives listed above are bounded above and below, therefore satisfying the conditions in VII-B.

REFERENCES

[1] A. O. Hero, "Geometric entropy minimization (GEM) for anomaly detection and localization," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 585–592.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[3] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," in *Proceedings of the ACM SIGCOMM*, 2003.

[4] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[5] R. Diestel, *Graph Theory*. Springer-Verlag, 2005.

[6] B. Bollobás, *Random Graphs*. Cambridge University Press, 2001.

[7] C. Berge, *Hypergraphs: combinatorics of finite sets*. North-Holland, 1989.

[8] J. Storey, "The positive false discovery rate: a Bayesian interpretation of the q-value," *Annals of Statistics*, vol. 31, no. 6, pp. 2013–2035, 2003.

[9] C. Scott and E. Kolaczyk, "Nonparametric assessment of contamination in multivariate data using Minimum Volume sets and FDR," *unpublished*, April 2007.

[10] R. El-Yaniv and M. Nisenson, "Optimal single-class classification strategies," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[11] B. Schölkopf, J. C. Platt, J. Shawne-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1471, 2001.

[12] J. Aitchison and C. G. G. Aitken, "Multivariate binary discrimination by the kernel method," *Biometrika*, vol. 63, pp. 413–420, 1976.

[13] D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, 1992.

[14] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in *Proceedings of the Third SIAM International Conference on Data Mining*, San Francisco, CA, May 2003.

[15] T. Ahmed, B. Oreshkin, and M. Coates, "Machine learning approaches to network anomaly detection," in *Proceedings of the Second Workshop on Tackling Computer Systems Problems with Machine Learning (SysML)*, Cambridge, MA, April 2007.

[16] A. McCallum and K. Nigam, "A comparison of event models for naïve bayes text classification," AAAI-98 Workshop on Learning for Text Categorization, Tech. Rep. WS-98-05, 1998.

[17] K. Humphreys and D. M. Titterington, "Improving the mean-field approximation in belief networks using Bahadur's reparameterisation of the multivariate binary representation," *Neural Processing Letters*, vol. 12, pp. 183–197, 2000.
[18] J. Keilson and H. Gerber, "Some results for discrete unimodality," *Journal of the American Statistical Association*, vol. 66, no. 334, pp. 386–389, 1971.
[19] L. Devroye, "Random variate generation for multivariate unimodal densities," *ACM Transactions on Modeling and Computer Simulation*, vol. 7, no. 1, pp. 447–477, 1997.
[20] J. H. Wolfe, "Pattern clustering by multivariate mixture analysis," *Multivariate Behavioral Research*, vol. 5, pp. 329–350, 1970.
[21] M. Á. Carreira-Perpiñán and S. Renals, "Practical identifiability of finite mixtures of multivariate bernoulli distributions," *Neural Computation*, vol. 12, pp. 141–152, 2000.
[22] M. Gyllenberg, T. Koski, E. Reilink, and M. Verlaan, "Non-uniqueness in probabilistic numerical identification of bacteria," *Journal of Applied Probability*, vol. 31, pp. 542–548, 1994.
[23] P. Hall and X.-H. Zhou, "Nonparametric estimation of component distributions in a multivariate mixture," *Annals od Statistics*, vol. 31, no. 1, pp. 201–224, 1982.
[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[25] X. L. Meng and D. van Dyk, "The em algorithm - an old folk-song sung to a fast new tune (with discussion)," *Journal of the Royal Statistical Society. Series B (methodological)*, vol. 59, pp. 511–567, 1997.
[26] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.
[27] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, pp. 195–239, 1984.