Distributed PCA and Network Anomaly Detection

Ling Huang
Xuanlong Nguyen
Minos Garofalakis
Michael Jordan
Anthony D. Joseph
Nina Taft

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-99
http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-99.html

July 14, 2006

Distributed PCA and Network Anomaly Detection

Ling Huang* XuanLong Nguyen* Minos Garofalakis†
Michael Jordan* Anthony Joseph* Nina Taft†
*UC Berkeley †Intel Research Berkeley
{hling, xuanlong, adj, jordan}@cs.berkeley.edu {minos.garofalakis, nina.taft}@intel.com

Abstract

We consider the problem of network anomaly detection given the data collected and processed over large distributed systems. Our algorithmic framework can be seen as an approximate, distributed version of the well-known Principal Component Analysis (PCA) method, which is concerned with continuously tracking the behavior of the data projected onto the residual subspace of the principal components within error bound guarantees. Our approach consists of a protocol for local processing at individual monitoring devices, and global decision-making and monitoring feedback at a coordinator. A key ingredient of our framework is an analytical method based on stochastic matrix perturbation theory for balancing the tradeoff between the accuracy of our approximate network anomaly detection, and the amount of data communication over the network.

1 Introduction

The area of distributed computing systems provides a promising domain for applications of machine learning methods. One of the most interesting aspects of such applications is that learning algorithms that are embedded in a distributed computing infrastructure are themselves part of that infrastructure and must respect its inherent local computing constraints (e.g., constraints on bandwidth, latency, reliability, etc.), while attempting to aggregate information across the infrastructure so as to improve system performance (or, availability) in a global sense.

Consider, for example, the problem detecting anomalies in a wide-area network. While it is straightforward to embed learning algorithms at local nodes to attempt to detect node-level anomalies, these anomalies may not be indicative of network-level problems. Indeed, in recent work, [10] demonstrated a useful role for Principal Component Analysis (PCA) to detect network anomalies. They showed that the minor components of PCA (the subspace obtained after removing the components with largest eigenvalues) revealed anomalies that were not detectable in any single node-level trace. While their work did not face the distributed data analysis problem (it involved centralized, off-line analysis of blocks of data), it does provide clear motivation for attempting to design a distributed PCA-based system for analyzing network anomalies in real time.

The development of such a design involves facing several challenging problems that have not been addressed in previous work. Naive solutions that continuously push all data to a central analysis site simply cannot scale to large networks or massive data streams. Instead, viable solutions need to process data "in-network" to intelligently control the frequency and size of data communications. The key underlying problem is that of developing a mathematical understanding of how to trade off quantization arising from local bandwidth restrictions against fidelity of the data analysis. We also need to understand how this tradeoff impacts overall detection accuracy. Finally, the implementation needs to be simple if it is to have impact on developers.

In this paper, we present a simple algorithmic framework for approximate distributed PCA tracking, together with supporting theoretical analysis. In brief, the architecture involves a set of local monitors that maintain parameterized sliding filters. These sliding filters yield quantized data streams

that are sent to a coordinator. The coordinator makes global decisions based on these quantized data streams, and also provides feedback to the monitors, allowing them to update the parameters of their filters. Our basic theoretical tool is stochastic matrix perturbation theory. This tool turns out to be quite well suited to our problem, yielding analytical expressions or explicit bounds for many of the quantities that are needed in either our approximate PCA-tracking algorithm or in the analysis of its performance guarantees. The combination of our theoretical tools and local filtering strategies results in a distributed tracking algorithm that can achieve high detection accuracy with low communication overhead; for instance, our experiments show that, by choosing a relative eigen-error of 1.5% (yielding, approximately, a 4% missed detection rate and a 6% false alarm rate), we can filter out more than 90% of the traffic for the original signal.

Prior Work. The original work on PCA-based methods by Lakhina et al. [10] has been extended by [20], who show how to infer network anomalies in both spatial and temporal domains. As with [10], this work is completely centralized. Other initiatives in distributed monitoring, profiling and anomaly detection aims to share information and foster collaboration between widely distributed monitoring boxes to offer improvements over isolated systems [13, 18, 19]. In the setting of simpler statistics such as sums and counts, distributed detection methods related to ours have been explored by [8]. Finally, work in the machine learning literature that combines learning methods with distributed constraints includes work by [12], who introduce a distributed kernel-based classification algorithm and [9], who consider a distributed message passing algorithm in graphical models.

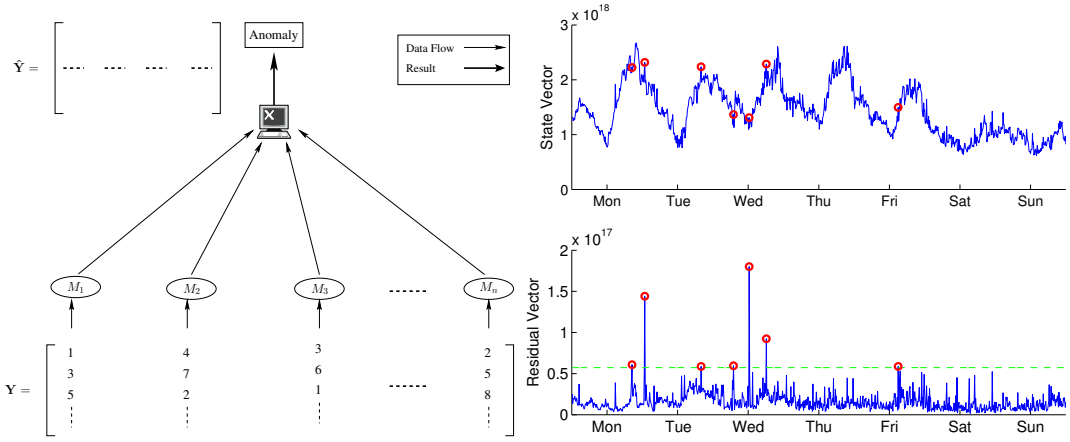
Organization. We start by discussing our system model and background on PCA-based network traffic anomaly detection in Section 2. Section 3 presents our distributed PCA-tracking algorithm and highlights our main analytical results. (Due to space constraints, detailed proof arguments can be found in the appendix.) We present the experimental evaluation of our system in Section 4, and give our conclusions in Section 5.

2 Problem description and background

We consider a monitoring system comprising a set of *local monitor nodes* $M_1, \ldots, M_n$ each of which collects a locally-observed time-series data stream (Fig. 1(a)). For instance, the monitors may collect information on the number of TCP connection requests per second, the number of DNS transactions per minute, or the volume of traffic at port 80 per second. A central *coordinator node* aims to continuously monitor the global collection of time series, and make global decisions such as those concerning matters of network-wide health. Although our methodology is generally applicable, in this paper, we focus on the particular application of detecting *volume anomalies*. A volume anomaly refers to unusual traffic load levels in a network that are caused by anomalies such as worms, DDoS attacks, device failures, misconfigurations, and so on.

Each monitor collects a new data point at every time step and, assuming a naive, "continuous push" protocol, sends the new point to the coordinator. Based on these updates, the coordinator keeps track of a sliding time window of size m (i.e., the m most recent data points) for each monitor time series, organized into a matrix $\mathbf{Y}$ of size $m \times n$ (where the i^{th} column $\mathbf{Y}_i$ captures the data from monitor i, see Fig. 1(a)). The coordinator then makes its decisions based solely on this (global) $\mathbf{Y}$ matrix.

The network-wide volume anomaly detection algorithm of [10] works by local monitors measuring the total volume of traffic (in bytes) on each network link, and periodically (e.g., every 5 minutes) centralizing the data by pushing all recent measurements to the coordinator. The coordinator then performs PCA on the assembled $\mathbf{Y}$ matrix to detect volume anomalies. (Details are given later in this section.) This method has been shown to work remarkably well, in part due to the inherently low-dimensional nature of the underlying data. However, such a "periodic push" approach suffers from inherent limitations: To ensure fast detection, the update periods should be relatively small; unfortunately, small periods also imply increased monitoring communication overheads, which may very well be unnecessary (e.g., if there are no significant local changes across periods). Instead, in our work, we study how the monitors can effectively filter their time-series updates, sending as little as possible, yet enough so as to allow the coordinator to make global decisions accurately. We provide analytical bounds on the errors that occur because decisions are made with incomplete data, and explore the tradeoff between reducing data transmissions (communication overhead) and decision accuracy.



(a) *The system setup* (b) *Abilene network traffic data*

Figure 1: (a) The distributed monitoring system; (b) Data sample ($\|\mathbf{y}\|^2$) collected over one week (top); its projection in residual subspace (bottom). Dashed line represens a threshold for anomaly detection.

Using PCA for centralized volume anomaly detection. As observed by Lakhina et al. [10], due to the high level of traffic aggregation on ISP backbone links, volume anomalies can often go unnoticed by being "buried" within normal traffic patterns (e.g., the circle dots shown in the top plot in Fig 1(b)). On the other hand, they observe that, although, the measured data is of seemingly high dimensionality (n = number of links), normal traffic patterns actually lie in a very low-dimensional subspace; furthermore, separating out this normal traffic subspace using PCA (to find the principal traffic components) makes it much easier to identify volume anomalies in the remaining subspace (bottom plot of Fig. 1(b)).

As before, let $\mathbf{Y}$ be the global $m \times n$ time-series data matrix, centered to have zero mean, and let $\mathbf{y} = \mathbf{y}(t)$ denote a n-dimensional vector of measurements (for all links) from a single time step t. Formally, PCA is a coordinate-transformation method that maps a given set of data points onto principal components ordered by the amount of data variance that they capture. The set of n principal components, $\{\mathbf{v}_i\}_{i=1}^{n}$, are defined as:

$$\mathbf{v}_i = \arg\max_{\|\mathbf{x}\|=1} \|(\mathbf{Y} - \sum_{j=1}^{i-1} \mathbf{Y}\mathbf{v}_j\mathbf{v}_j^T)\mathbf{x}\|$$

and are the n eigenvectors of the estimated covariance matrix $\mathbf{A} := \mathbf{Y}^T\mathbf{Y}$. As shown in [10], PCA reveals that the Origin-Destination (OD) flow matrices of ISP backbones have low intrinsic dimensionality: For the Abilene network with 41 links, most data variance can be captured by the first $k = 4$ principal components. Thus, the underlying normal OD flows effectively reside in a (low) k-dimensional subspace of $\mathbb{R}^n$. This subspace is referred to as the *normal* traffic subspace $\mathcal{S}_{no}$. The remaining $(n-k)$ principal components constitute the *abnormal* traffic subspace $\mathcal{S}_{ab}$.

Detecting volume anomalies relies on the decomposition of link traffic $\mathbf{y} = \mathbf{y}(t)$ at any time step into normal and abnormal components, $\mathbf{y} = \mathbf{y}_{no} + \mathbf{y}_{ab}$, such that (a) $\mathbf{y}_{no}$ corresponds to modeled normal traffic (the projection of $\mathbf{y}$ onto $\mathcal{S}_{no}$), and (b) $\mathbf{y}_{ab}$ corresponds to residual traffic (the projection of $\mathbf{y}$ onto $\mathcal{S}_{ab}$). Mathematically, $\mathbf{y}_{no}(t)$ and $\mathbf{y}_{ab}(t)$ can be computed as

$$\mathbf{y}_{no}(t) = \mathbf{P}\mathbf{P}^T\mathbf{y} = \mathbf{C}_{no}\mathbf{y} \quad \text{and} \quad \mathbf{y}_{ab}(t) = (\mathbf{I} - \mathbf{P}\mathbf{P}^T)\mathbf{y} = \mathbf{C}_{ab}\mathbf{y}$$

where $\mathbf{P} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}]$, is formed by the first k principal components which capture the dominant variance in the data. The matrix $\mathbf{C}_{no} = \mathbf{P}\mathbf{P}^T$ represents the linear operator that performs projection onto the normal subspace $\mathcal{S}_{no}$, and, $\mathbf{C}_{ab}$ projects onto the abnormal subspace $\mathcal{S}_{ab}$.

As observed in [10], a volume anomaly typically results in a large change to $\mathbf{y}_{ab}$; thus, a useful metric for detecting abnormal traffic patterns is the squared prediction error (SPE):

$$\mathbf{SPE} \equiv \|\mathbf{y}_{ab}\|^2 = \|\mathbf{C}_{ab}\mathbf{y}\|^2$$

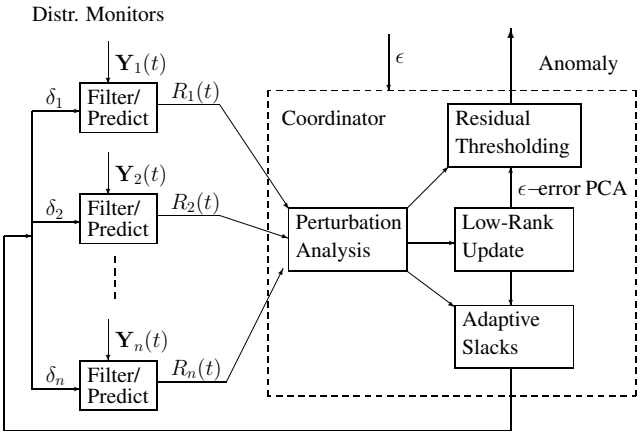


Figure 2: Our distributed tracking and detection framework.

(essentially, a quadratic residual function). More formally, their proposed algorithm signals a volume anomaly, if $\mathbf{SPE} > Q_\alpha$, where Q_α denotes the threshold statistic for the $\mathbf{SPE}$ residual function at the $1-\alpha$ confidence level. Such a statistical test for the $\mathbf{SPE}$ residual function, known as the Q-statistic [7], can be computed as a function $Q_\alpha = Q_\alpha(\lambda_{k+1}, \ldots, \lambda_n)$, of the $(n-k)$ non-principal eigenvalues of the covariance matrix $\mathbf{A}$.

3 Distributed PCA for anomaly detection

We now describe our communication-efficient distributed solution for real-time approximate PCA tracking and detection. In a nutshell, the key idea is to limit monitor-coordinator interactions by installing *local filters* at the monitors, and forcing communication only when such local constraints are violated. Using stochastic matrix perturbation theory to analyze the effect of this local quantization on the global matrix PCA, we propose a simple, adaptive distributed protocol with useful, provable performance guarantees on the communication/detection accuracy tradeoff.

3.1 Overview of our approach

Our approach for distributed, PCA-based anomaly detection comprises two parts: (1) the monitors process their collected data by applying local filtering rules to suppress unnecessary message updates to the coordinator; and, (2) the coordinator makes global detection decisions and provides feedback to the monitors (e.g., local filter parameter settings) based on the observed updates. Fig. 2 illustrates the overall architecture of our distributed anomaly-detection system.

Local Processing at Monitors. The goal of a monitor is to track its local raw data stream, and update the coordinator of any considerable drift that might affect the global decision made by the coordinator. Of course, our goal is to avoid flooding the coordinator with all the raw data. To this end, each monitor i maintains a filtering window $F_i(t)$ of size $2\delta_i$ centered at a value R_i (i.e., $F_i(t) = [R_i(t) - \delta_i, R_i(t) + \delta_i]$). The monitor updates the coordinator with the most recent data $\mathbf{Y}_i(t)$ only if $\mathbf{Y}_i(t) \notin F_i$ (and sends nothing otherwise).

The window parameter δ_i is called the *slack*. Clearly, increased slack implies reduced communication between a monitor and the coordinator at the expense of potential information loss (i.e., poor approximation) at the the coordinator. The center parameter $R_i(t)$ denotes the approximate representation of $\mathbf{Y}_i(t)$ that the coordinator uses; in general, $R_i(t)$ can be based on any type of *filtering/prediction model* for node M_i's behavior over time. (For instance, in our implementation, $R_i(t)$ is simply the average of last five signal values observed locally at monitor i.) Obviously, this implies that the coordinator maintains only an approximate "filtered" version $\hat{\mathbf{Y}}$ of the $\mathbf{Y}$ matrix.

Global Decision-Making at the Coordinator. The role of the coordinator is twofold. First, it makes global anomaly-detection decisions based upon the received updates from monitors. Secondly, it provides feedback to the monitors in order to adjust their filtering windows' slack and

center parameters. The global detection task is the same as the centralized detection scheme described in Section 2 using the $\mathbf{SPE}$ statistic based on the projection of a filtered signal $\hat{\mathbf{y}}$ on the residual subspace represented by matrix $\mathcal{S}_{ab}$. In contrast to the centralized setting, however, the coordinator does not have an exact version of the raw data matrix $\mathbf{Y}$; instead, PCA is performed and $\mathcal{S}_{ab}$ is computed on the *perturbed/filtered* version of the covariance matrix $\hat{\mathbf{A}} := \mathbf{A} - \mathbf{\Delta}$, where the magnitude of the perturbation matrix $\mathbf{\Delta}$ is decided by the slack variables δ_i ($i = 1, \ldots, M$).

3.2 Distributed tracking of eigenvalues and the residual subspace

A key ingredient of our framework is a practical method for choosing slack parameters δ_i that effectively balance the tradeoff between the desirable loss of detection accuracy (i.e., due to the use of $\hat{\mathbf{Y}}$ instead of $\mathbf{Y}$) and the savings in data communication. The mathematical tool that we employ to resolve this issue is *stochastic matrix perturbation theory*. This theory quantifies the effects of the perturbation of a matrix on key quantities such as eigenvalues and the eigen-subspaces, which in turn affects the detection accuracy. In the remainder of this section, we highlight several key results of our analysis; due to space constraints, the complete details can be found in the appendix.

In our framework, the coordinator's view of the data matrix is the perturbed matrix $\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{W}$, where all elements of the column vector $\mathbf{W_i}$ are bounded within interval $[-\delta_i, \delta_i]$. Let λ_i and $\hat{\lambda}_i$ ($i = 1, \ldots, n$) denote the eigenvalues of the covariance matrix $\mathbf{A} = \mathbf{Y}^T\mathbf{Y}$ and its perturbed version $\hat{\mathbf{A}} := \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$. Applying classical theorems of Mirsky and Weyl [17], we obtain bounds on the eigenvalue perturbation in terms of the Frobenius norm $\|.\|_F$ and the spectral norm $\|.\|_2$ of $\mathbf{\Delta} := \mathbf{A} - \hat{\mathbf{A}}$, respectively:

$$\epsilon_{eig} := \sqrt{\sum_{i=1}^{n} \frac{1}{n}(\hat{\lambda}_i - \lambda_i)^2} \leq \|\mathbf{\Delta}\|_F/\sqrt{n} \quad \text{and} \quad \max_i |\hat{\lambda}_i - \lambda_i| \leq \|\mathbf{\Delta}\|_2 \tag{1}$$

Applying the sin theorem and results on bounding the angle of projections to subspaces [3, 17] (see the appendix for more details), we also obtain bounds on the perturbation of the residual subspace $\mathbf{C}_{ab}$ in terms the Frobenius norm of $\mathbf{\Delta}$:

$$\|\mathbf{C}_{ab} - \hat{\mathbf{C}}_{ab}\|_F \leq \frac{\sqrt{2}\|\mathbf{\Delta}\|_F}{\nu} \tag{2}$$

where ν denotes the eigengap between the k^{th} and $(k+1)^{th}$ eigenvalues of the estimated covariance matrix $\hat{\mathbf{A}}$.

The issue is to obtain practically useful bound on the norms of $\mathbf{\Delta}$. To this end, we obtain expectation bounds instead of worst case bounds. We make the following assumptions on the error matrix $\mathbf{W}$:

1. The column vectors $\mathbf{W_1}, \ldots, \mathbf{W_n}$ are independent and radially symmetric m-dim vectors.
2. For each $i = 1, \ldots, n$, all elements of column vector $\mathbf{W_i}$ are i.i.d. random variables with mean 0, variance $\sigma_i^2 := \sigma_i^2(\delta_i)$ and fourth moment $\mu_i^4 := \mu_i^4(\delta_i)$.

Note that the independence assumption is on the error only – this by no means implies that the signals received by different monitors are statistically independent. We define the *aggregated variance* across monitors as $\sigma := \sum_{\mathbf{i=1}}^{\mathbf{n}} \sigma_{\mathbf{i}}^{\mathbf{2}}$. Under the above assumption, we can show that $\|\mathbf{\Delta}\|_F/\sqrt{n}$ is upper bounded by the following quantity with high probability when m and n are large:

$$Tol_F = 2\sqrt{\frac{\sigma}{n}\sum_{i=1}^{n}\hat{\lambda}_i} + \sqrt{\frac{m^2}{n}\sum_{i=1}^{n}\sigma_i^4 + \frac{m}{n}\sum_{i=1}^{n}(\mu_i^4 - \sigma_i^4) + \frac{m}{n}\sigma^{\mathbf{2}}} \tag{3}$$

Similar results can be obtained for the spectral norm as well. In practice, these upper bounds are very tight because σ tends to be small compared to the top eigenvalues.

In summary, given the tolerable perturbation Tol_F on the norm of the covariance matrix, we can decide the amount of slack for each monitor via Equation. (3) (e.g., by dividing the overall tolerance across monitors either uniformly or in proportion to their observed local variance). At the same time, the bounds (1) and (2) allow us to measure the amount of perturbation ϵ_{eig} in terms of eigenvalues, which affects the detection accuracy, which we discuss next.

3.3 Effects on detection procedure at the coordinator

The coordinator performs online detection based upon the (filtered) data stream $\hat{\mathbf{Y}}$. We now address the impact of approximation in PCA on the detection procedure. Specifically, we examine how the filtering on $\mathbf{Y}$ impacts the residual projection statistic $\mathbf{SPE} = \|\mathbf{C}_{ab}\mathbf{y}\|^2$ and the threshold Q_α. As discussed, using filtering slacks computed by Equation 3, the coordinator can bound the deviation of its computed $\|\hat{\mathbf{C}}_{ab}\|$ and $\hat{Q}_\alpha$ from their true values, which gives us tools to analyze and bound the false alarm rate on the detection at the coordinator when using condition $\|\hat{\mathbf{C}}_{ab}\hat{\mathbf{y}}\|^2 > \hat{Q}_\alpha$. First, note that

$$\begin{aligned} |\|\hat{\mathbf{C}}_{ab}\hat{\mathbf{y}}\| - \|\mathbf{C}_{ab}\mathbf{y}\|| &\leq \|(\hat{\mathbf{C}}_{ab} - \mathbf{C}_{ab})\hat{\mathbf{y}}\| + \|\mathbf{C}_{ab}(\mathbf{y} - \hat{\mathbf{y}})\| \leq \frac{\sqrt{2}\|\mathbf{\Delta}\|_F\|\hat{\mathbf{y}}\|}{\nu} + \|\mathbf{C}_{ab}\|_2 \sum_{i=1}^{n} \delta_i^2 \\ &\leq \frac{\sqrt{2}\|\mathbf{\Delta}\|_F\|\hat{\mathbf{y}}\|}{\nu} + \left(\|\hat{\mathbf{C}}_{ab}\| + \frac{\sqrt{2}\|\mathbf{\Delta}\|_F}{\nu}\right) \sum_{i=1}^{n} \delta_i^2 \quad =: \eta_1(\hat{\mathbf{y}}) \end{aligned}$$

It is simple to obtain an upper bound on the perturbation of $\mathbf{SPE}$ as:

$$\eta_2(\hat{\mathbf{y}}) = \eta_1(\hat{\mathbf{y}})(2\|\hat{\mathbf{C}}_{ab}\hat{\mathbf{y}}\| + \eta_1(\hat{\mathbf{y}})).$$

Turning to the threshold Q_α, which is given as a function of $\lambda_{k+1}, \ldots, \lambda_n$, the eigenvalues of $\mathbf{A}$ [7]:

$$Q_\alpha = \phi_1 \left[\frac{c_\alpha \sqrt{2\phi_2 h_0^2}}{\phi_1} + 1 + \frac{\phi_2 h_0 (h_0 - 1)}{\phi_1^2} \right]^{\frac{1}{h_0}}$$

where c_α is the $(1-\alpha)$-percentile of the standard normal distribution, $h_0 = 1 - \frac{2\phi_1\phi_3}{3\phi_2^2}$, $\phi_i = \sum_{j=k+1}^{n} \lambda_j^i$ for $i = 1, 2, 3$. The perturbation in $\lambda_{k+1}, \ldots, \lambda_n$ directly impacts the change of Q_α. In our application, it is observed that the change of ϕ_1 usually dominates the change. Furthermore, increasing ϕ_1 results in decreasing Q_α. Since the addition of random perturbation to a matrix tends to increase the non-principal eigenvalues $\lambda_{k+1}, \ldots, \lambda_n$, thus increasing ϕ_1. This implies that Q_α decreases as the amount of perturbation increases.

To assess the perturbation in terms of false alarm rate, we only need to bound the difference $\hat{c} - c$, where $\hat{c}$ is a perturbed version of:

$$c = \frac{\phi_1[(\mathbf{SPE}/\phi_1)^{h_0} - 1 - \phi_2 h_0(h_0 - 1)/\phi_1^2]}{\sqrt{2\phi_2 h_0^2}}.$$

Let η_c denote the bound on $|\hat{c} - c|$. The change to the false alarm rate is approximated as $P(c_\alpha - \eta_c < U < c_\alpha + \eta_c)$, where U is a standard normal random variable.

4 Evaluation

We implemented our algorithm and developed a trace-driven simulator to validate our methods. We used a one-week trace collected from the the Abilene network[1]. The traces containes per-link traffic loads measured every 10 minutes, for all 41 links of the Abilene network. With a time unit of 10 minutes, data was collected for 1008 time units. This data was used to feed in the simulator. There are 7 anomalies in the data that were detected by the centralized algorithm (and verified by hand to be true anomalies). In our experiments, we injected 70 synthetic anomalies into this dataset using the method described in [10], so that we would have sufficient data to compute error rates. We used a threshold Q_α corresponding to an $1 - \alpha = 99.5\%$ confidence level. Due to space limitations, we only present results for the case of uniform monitor slack $\delta_i = \delta$.

The input parameter for algorithm is the tolerable relative error of eigenvalues (relative eigen error in short), which acts as a tuning knob.[2] Given this parameter and the input data we can compute

[1] Abilene is an Internet2 high-performance backbone network that interconnects a large number of universities as well as a few other research institutes.

[2] Precisely, it is $Tol_F / \sqrt{\frac{1}{n}\sum \lambda_i^2}$, where Tol_F is defined in Eqn (3).

the filtering slack δ for the monitors using Eqn (3). We then feed in the data to run our protocol in the simulator with the computed δ. The simulator outputs a set of results including: 1) the actual relative eigen errors and the relative errors on the detection threshold Q_α; 2) the missed detection rate, false alarm rate and communication cost when using our protocol for distributed tracking and anomaly detection. The *missed-detection rate* is defined as the fraction of missed detections over the total number of real anomalies, and the *false-alarm rate* as the fraction of false alarms over the total number of detected anomalies by our protocol. The communication cost is computed as the fraction of number of messages that actually get through the filtering window to the coordinator.

The results are shown in Fig. 3. In all plots, x-axis is the tolerable relative eigen error. In Fig. 3(a) we plot the relationship between the tolerable eigen error and filtering slack δ when assuming filtering errors are uniformly distributed on interval $[-\delta, \delta]$. With this model, the relationship between the tolerable eigen error and the slack is determined by a simplified version of Eqn (3). The results make intuitive sense. As we increase our error tolerance, we can filter more at the monitor and send less to the coordinator. The slack increases almost linearly with the tolerable eigen error because the first term in the right hand side of Eqn (3) dominates all other terms.

In Fig. 3(b) we compare the tolerable relative eigen error (which is the tuning parameter) to the actual accrued relative eigen error (precisely defined as $\epsilon_{eig}/\sqrt{\frac{1}{n}\sum \lambda_i^2}$, where ϵ_{eig} is defined in Eqn (1)). These were computed using the slack parameters δ as computed by our coordinator. We can see that the real accrued eigen errors are always less than the tolerable eigen errors. The plot shows a tight upper bound, indicating that it is safe to use our model's derived filtering slack δ. In other words, the achieved eigen error always remains below the requested tolerable error specified as input, and the slack chosen given the tolerable error is close to being optimal. In Fig. 3(c) we show the relationship between the tolerable eigen error and the relative error of detection threshold Q_α[3]. It confirmed our analysis that the threshold for detecting anomalies decreases as we tolerate more and more eigen errors. In these experiments, an error of 2% in the eigenvalues, leads to an error of approximately 6% in our estimate of the appropriate cutoff threshold.

We are now ready to examine the false alarm rates achieved. In Fig. 3(d) the curve with triangles represents the upper bound on the false alarm rate as estimated by the coordinator (as discussed in section 3.3). The curve with circles is the actual accrued false alarm rate achieved by our scheme. It is worth noting that the upper bound on the false alarm rate is fairly close to true values, especially when the slack is small. The false alarm rate increases with increasing eigen error because as the eigen error increases, the corresponding detection threshold Q_α will decrease, which in turn causes the protocol to raise an alarm more often. (If we had plotted $\hat{Q}$ rather than the relative threshold difference, we would obviously see a decreasing $\hat{Q}$ with increasing eigen error.) To the best of our knowledge, this is among the first work to provide a bound on false alarm rates for a network anomaly detector in a distributed setting. In Fig. 3(e) we present the missed detection rates. For varying choice of communication overhead the missed detection rate remains below 4%.

Finally we examine the communication overhead in Fig. 3(f). As we can tolerate larger errors, we can consequently reduce the overhead. Using these last three plots (d,e,f) together, we can observe the tradeoffs that occur. For example, when the relative eigen error is 1.5%, our algorithm reduces the data sent through the network by more than 90%. This gain is achieved at the cost of approximately a 4% missed detection rate and a 6% false alarm rate. This is a large reduction in communication for a small increase in detection errors. This illustrates that our distributed protocol can achieve high detection accuracy with low communication overhead.

5 Conclusion

We have presented a distributed algorithmic framework for network anomaly detection using the PCA method. Our framework consists of a simple protocol for local data processing at the monitoring devices and global decision making and feedback at the coordinator. Using tools for stochastic matrix perturbation theory, we provided an analysis for the tradeoff between the detection accuracy

[3]Precisely, it is $1 - \hat{Q}_\alpha/Q_\alpha$, where $\hat{Q}_\alpha$ is computed from $\hat{\lambda}_{k+1}, \ldots, \hat{\lambda}_n$.

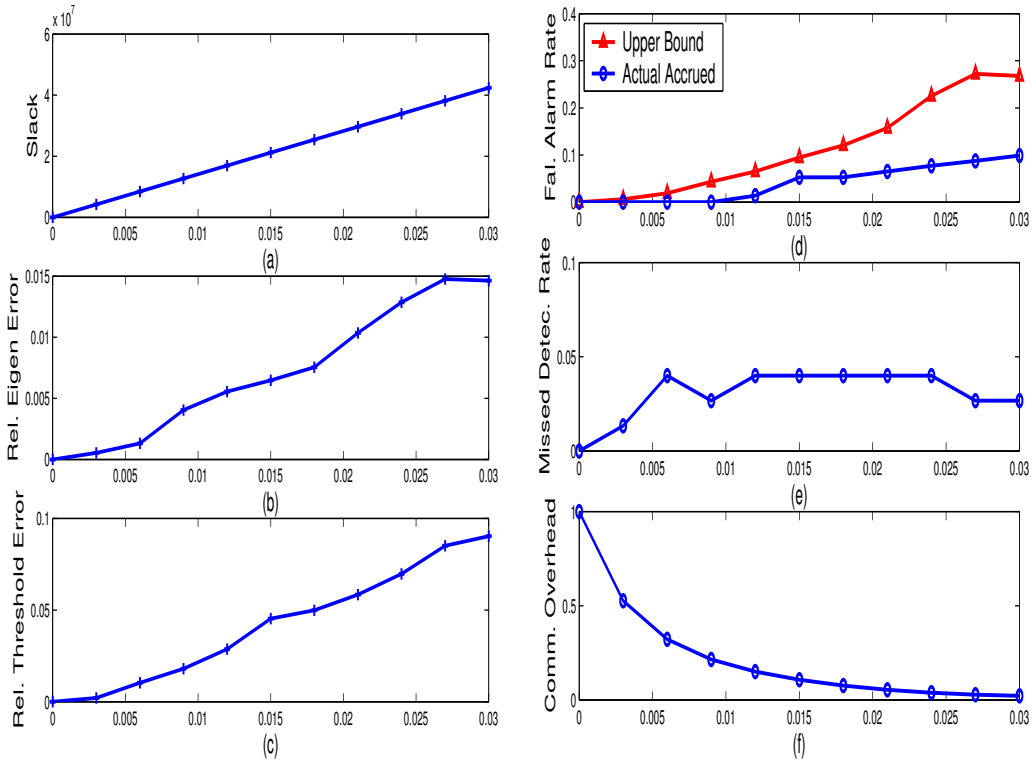


Figure 3: Parameters design, communication overhead and accrued error.

and the data communication overhead. In particular, using relative eigen error as a tuning knob, we were able to control the amount of data overhead, and as well as provide upper bounds on the false alarm rate. Our algorithm is simple to implement, and is empirically shown to yield high accuracy of anomaly detection in Abilene network despite using very little data.

References

[1] ALON. N., KRIVELEVICH, M., AND VU, V. H. On the concentration of eigenvalues of random symmetric matrices. In *Israel J. Math*, 131 (2002), pp. 259-267.

[2] BOTTCHER, A. AND GRUDSKY, S. The norm of the product of a large matrix and a random vector. In *Electronic Journal of Probability*, Vol. 8 (2003), pages 1-29.

[3] DRMAC, Z. On principal angles between subspaces of euclidean of space In *SIAM J. Matrix Anal. Appl.*, 22(1) 2000.

[4] GEMAN, S. A limit theorem for the norm of random matrices. In *Ann. Probab.*, 8 (1980), pp. 252-261.

[5] HOLMES, R. B. On Random Correlation Matrices. IN *SIAM J. Matrix Anal. Appl.*, 12(2) 1991.

[6] HUANG, L., GAROFALAKIS, M., JOSEPH, A. AND TAFT, N. Communication-efficient tracking of distributed cumulative triggers. Under submission, May 2006.

[7] JACKSON, J. E. AND MUDHOLKAR, G. S. Control procedures for residuals associated with principal component analysis. In *Technometrics*, pages 341-349, 1979.

[8] KERALAPURA, R., CORMODE, G. AND RAMAMIRTHAM, J. Communication-efficient distributed monitoring of thresholded counts To appear in *ACM SIGMOD* (2006).

[9] KREIDL, P. O., WILLSKY, A. Inference with Minimal Communication: a Decision-Theoretic Variational Approach. In *NIPS 18* (2006).

[10] LAKHINA, A., CROVELLA, M. AND DIOT, C. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM*, (2005).

[11] LAKHINA, A., PAPAGIANNAKI, K., CROVELLA, M., DIOT, C., KOLACZYK, E. D. AND TAFT, N. Structural analysis of network traffic flows. In *ACM SIGMETRICS*, (2004).

[12] NGUYEN, X., WAINWRIGHT, M. AND JORDAN, I. M. Nonparametric decentralized detection using kernel methods. In *IEEE Trans. Signal Processing* 53(11), (2005).

[13] PADMANABHAN, V. N., RAMABHADRAN, S., AND PADHYE, J. Netprofiler: Profiling wide-area networks using peer cooperation. In *IPTPS* (2005).

[14] RUBINSTEIN, R. Generating random vectors uniformly distributed inside and on the surface of different regions. In *Europ. J. Oper. Res.*, 10 (1982), pp. 205-209.
[15] SPRING, N., WETHERALL, D., AND ANDERSON, T. Scriptroute: A facility for distributed internet measurement. In *USITS* (2003).
[16] STEWART, G. W. *Perturbation theory for the sigular value decomposition.* UMIACS-TR-90-123, 1990.
[17] STEWART, G. W., AND SUN, J.-G. *Matrix Perturbation Theory*. Academic Press, 1990.
[18] XIE, Y., KIM, H.-A., O'HALLARON, D. R., REITER, M. K., AND ZHANG, H. Seurat: A pointillist approach to anomaly detection. In *RAID* (2004).
[19] YEGNESWARAN, V., BARFORD, P., AND JHA, S. Global intrusion detection in the domino overlay system. In *NDSS* (2004).
[20] ZHANG, Y., GE, Z.-H., GREENBERG, A., AND ROUGHAN, M. Network anomography. In *IMC* , (2005).

6 Appendix

In this Appendix we develop a more detailed analysis of the impact of the slackness parameter $(\delta_1, \ldots, \delta_n)$ on the eigenvalues and eigen subspaces on the principal components using matrix perturbation theory. Some of the main results presented herein are summarized in Section 3. We begin with a brief background description of known results from matrix perturbation theory, and then proceeds to its application on our problem.

6.1 Background

Matrix perturbation theory is concerned with measuring the impact of small perturbation on matrices on relevant quantities such as the eigenvalues and eigenvectors.

Eigenvalue perturbation bounds The basic perturbation bounds for eigenvalues of a matrix are due to Weyl and Mirsky with following two theorems [16]. Let matrix $\mathbf{A}$ has eigenvalues λ_i, and its perturbated matrix, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{\Delta}$, has eigenvalues $\hat{\lambda}_i$, for $i = 1, \ldots, n$. We have:

Theorem 1 ***(Weyl)*** $\max_i |\hat{\lambda}_i - \lambda_i| \leq \|\mathbf{\Delta}\|_2$.

Theorem 2 ***(Mirsky)*** $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\lambda}_i - \lambda_i)^2} \leq \frac{\|\mathbf{\Delta}\|_F}{\sqrt{n}}$.

Here $\|.\|_2$ and $\|.\|_F$ denote the spectral 2-norm and the Frobenius norm (cf. [17]).

Invariant subspace perturbation. While eigenvalues are quite stable under matrix perturbation, the individual eigenvectors are not. Instead one needs to look at the perturbation of subspaces spanned by the eigenvectors. Subspaces spanned by eigenvectors are an example of *invariant* subspaces, which are known to be stable [4]

Let $\mathcal{L}(\cdot)$ denote the set of eigenvalues of a matrix, $\mathcal{S}(\cdot)$ denote the subspace spanned by a matrix, and $\mathbf{\Theta}$ denote the matrix of canonical angle between two subspaces (cf. [17]). Then the perturbation of an invariant subspace spanned by eigenvectors can be quantify by the sin of the canonical angle by the following $sin\ \Theta$ theorem [17]:

Theorem 3 *Let* $\mathbf{A}$ *have the spectral resolution*

$$\begin{bmatrix} \mathbf{X_1^T} \\ \mathbf{X_2^T} \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{X_1} & \mathbf{X_2} \end{bmatrix} = \begin{bmatrix} \mathbf{L_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{L_2} \end{bmatrix}$$

where $\begin{bmatrix} \mathbf{X_1} & \mathbf{X_2} \end{bmatrix}$ *is unitary with* $\mathbf{X_1} \in \mathbf{C^{n \times k}}$. *Let* $\mathbf{Z} \in \mathbf{C^{n \times k}}$ *have orthonormal columns, and for any symmetric* $\mathbf{M}$ *of order* k*, let*

$$\mathbf{R} = \mathbf{AZ} - \mathbf{ZM}$$

suppose that $\mathcal{L}(\mathbf{M}) \subset [\mathbf{a}, \mathbf{b}]$ *and for some eigengap* $\nu > 0$*,*

$$\mathcal{L}(\mathbf{L_2}) \subset \mathbb{R} \backslash [\mathbf{a} - \nu, \mathbf{b} + \nu]$$

[4] A subspace $\mathcal{X}$ is invariant of transformation A if $A\mathcal{X} \subset \mathcal{X}$.

Then for any unitarily invariant norm

$$\| \sin\Theta\left[\mathcal{S}(\mathbf{X_1}), \mathcal{S}(\mathbf{Z})\right] \| \leq \frac{\|\mathbf{R}\|}{\nu}$$

Note that this theorem applies to any unitarily invariant norm such as the spectral norm $\|.\|_2$ and Frobenius norm $\|.\|_F$. Applying this result to the eigen subspaces for (symmetric) covariance matrix $\mathbf{A}$ and its purturbed version $\hat{\mathbf{A}}$, assume that $\hat{\mathbf{A}}$ has the following the spectral resolution

$$\begin{bmatrix} \mathbf{Z_1^T} \\ \mathbf{Z_2^T} \end{bmatrix} \hat{\mathbf{A}} \begin{bmatrix} \mathbf{Z_1} & \mathbf{Z_2} \end{bmatrix} = \begin{bmatrix} \mathbf{M_1} & 0 \\ 0 & \mathbf{M_2} \end{bmatrix}$$

where $[\ \mathbf{Z_1}\ \ \mathbf{Z_2}\ ]$ is unitary with $\mathbf{Z_1} \in \mathbf{C^{n\times k}}$. Then we have $\mathbf{Z_1}^{\mathbf{T}}\hat{\mathbf{A}}\mathbf{Z_1} = \mathbf{M_1}$ and $\hat{\mathbf{A}}\mathbf{Z_1} = \mathbf{Z_1}\mathbf{M_1}$. Let $\mathbf{R} = \mathbf{A}\mathbf{Z_1} - \mathbf{Z_1}\mathbf{M_1} = \mathbf{A}\mathbf{Z_1} - \hat{\mathbf{A}}\mathbf{Z_1} = \mathbf{\Delta}\mathbf{Z_1}$. For any unitarily invariant norm, there holds $\|\mathbf{R}\| = \|\mathbf{\Delta}\mathbf{Z_1}\| = \|\mathbf{\Delta}\|$. As a result, we have:

$$\| \sin\Theta\left[\mathcal{S}(\mathbf{X_1}), \mathcal{S}(\mathbf{Z_1})\right] \| \leq \frac{\|\mathbf{R}\|}{\nu} = \frac{\|\mathbf{\Delta}\|}{\nu}$$

Finally, there is a close relationship between the perturbation of the projection operator onto invariant subspaces and the canonical angle of the subspace perturbation. Let $\mathbf{P}_X$ and $\mathbf{P}_Z$ be the orthogonal projections onto $\mathcal{S}(\mathbf{X})$ and $\mathcal{S}(\mathbf{Z})$. There holds [17]:

$$\|\mathbf{P}_X - \mathbf{P}_Z\|_F \quad = \quad \sqrt{2}\| \sin\mathbf{\Theta}\left[\mathcal{S}(\mathbf{X}), \mathcal{S}(\mathbf{Z})\right] \|_F \leq \frac{\|\mathbf{\Delta}\|_\mathbf{F}}{\nu}.$$

In summary, in order to assess the perturbation in eigenvalues and eigensubspace, we need to estimate the upper bounds given in terms of the Frobenius norm and the spectral norm of Δ.

6.2 Error matrix analysis

For the remainder of this appendix we shall present bounds and estimation of the Frobenius norm and spectral norm of the perturbation. Recall that $\mathbf{A} = \mathbf{Y}^T\mathbf{Y}$ and $\hat{\mathbf{A}} = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$, where $\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{W}$. $\mathbf{W}_i$ is a column vector of filtering error at each monitor i and $\mathbf{W}$ is the filtering (perturbation) error on the distributed matrix $\mathbf{Y}$. Each element e_{ji} of vector $\mathbf{W_i}$ is assumed to be bounded within $[-\delta_i, \delta_i]$. The norm of the perturbation error matrix $\mathbf{\Delta} = \mathbf{A} - \hat{\mathbf{A}}$ can be bounded as follows:

$$\|\mathbf{\Delta}\| = \|\mathbf{Y}^T\mathbf{W} + \mathbf{W}^T\mathbf{Y} + \mathbf{W}^T\mathbf{W}\| \leq \|\mathbf{Y}^T\mathbf{W}\| + \|\mathbf{W}^T\mathbf{Y}\| + \|\mathbf{W}^T\mathbf{W}\|.$$

Our strategy is to obtain bounds for each terms in the RHS of this inequality. It is possible to derive absolute bounds in terms of the absolute error $\delta_i (i = 1, \ldots, n)$. However, such bounds would be too loose for practical purposes. Instead, we appeal to *stochastic* perturbation theory. The basic idea is to assume that the error matrix $\mathbf{W}$ is random according to a certain distribution with estimated mean and higher-order moments. Instead of estimating the absolute upper bound for $\|\Delta\|$, we focus on estimating or bounding $\mathbb{E}\|\Delta\|$. This is done by bounding the expectation of the terms on the RHS of the above inequality.

Our assumption on the random distribution of $\mathbf{W}$ is given as follows:

1. The column vectors $\mathbf{W_1}, \ldots, \mathbf{W_n}$ are independent and radially symmetric m-dim vectors.
2. For each $i = 1, \ldots, n$, all elements of column vector $\mathbf{W_i}$ are i.i.d. random variables with mean 0, variance $\sigma_i^2 := \sigma_i^2(\delta_i)$ and fourth moment $\mu_i^4 := \mu_i^4(\delta_i)$.

6.2.1 Analysis of Frobenius norm

Computation of $\mathbb{E}\|\mathbf{Y}^T\mathbf{W}\|_\mathbf{F}^\mathbf{2}$. We exploit results from [2]: For any m-dimensional random vector $\mathbf{v}$ uniformly distributed on the unit sphere $\mathbb{S}^{m-1}$, and given a $m \times n$ matrix $\mathbf{Y}$, there hold:

$$\mathbb{E}(\|\mathbf{Y}^T\mathbf{v}\|^2) = \frac{\|\mathbf{Y}^T\|_F^2}{m}, \quad \mathrm{Var}(\|\mathbf{Y}^T\mathbf{v}\|^2) \leq \frac{2}{m+2}$$

As observed in [14], since $\mathbf{W}_i$ is assumed to be radially symmetric m-dimensional random vector, its projection on the unit sphere as $\mathbf{W_i} = \mathbf{v_i} \cdot \|\mathbf{W_i}\|$, where $\mathbf{v_i}$ is uniformly distributed on $\mathbb{S}^{m-1}$, and is independent with $\|\mathbf{W_i}\|$. Then we have

$$\begin{aligned}
\mathbb{E}(\|\mathbf{Y}^T\mathbf{W}_i\|^2) &= \mathbb{E}(\|\mathbf{Y}^T\mathbf{v}_i\|^2 \cdot \|\mathbf{W}_i\|^2) = \mathbb{E}(\|\mathbf{Y}^T\mathbf{v}_i\|^2) \cdot \mathbb{E}(\|\mathbf{W}_i\|^2) \\
&= \|\mathbf{Y}\|_F^2 \cdot \frac{\mathbb{E}(\|\mathbf{W}_i\|^2)}{m} = \|\mathbf{Y}\|_F^2 \cdot \sigma_i^2 \\
\mathbb{E}(\|\mathbf{Y}^T\mathbf{W}\|_F^2) &= \mathbb{E}(\|\mathbf{Y}^T\mathbf{W}\|_F^2) = \mathbb{E}(\sum_{i=1}^{n} \|\mathbf{Y}^T\mathbf{W}_i\|_F^2) = \sum_{i=1}^{n} \mathbb{E}(\|\mathbf{Y}^T\mathbf{W}_i\|_F^2) \\
&= \sum_{i=1}^{n} \|\mathbf{Y}\|_F^2 \cdot \sigma_i^2 = \|\mathbf{Y}\|_F^2 \cdot \sum_{i=1}^{n} \sigma_i^2 \\
&= \mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) \cdot \sum_{\mathrm{i=1}}^{\mathrm{n}} \sigma_{\mathrm{i}}^2 = \sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_{\mathrm{i}} \cdot \sum_{\mathrm{i=1}}^{\mathrm{n}} \sigma_{\mathrm{i}}^2 = \sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_{\mathrm{i}} \cdot \sigma,
\end{aligned}$$

where $\lambda_i's$ are eigenvalues of covariance matrix $\mathbf{A} = \mathbf{Y}^T\mathbf{Y}$.[5]

Computation of $\mathbb{E}(\|\mathbf{W}^T\mathbf{W}\|_F^2)$ This is a high order term, and its value is generally dominated by $\mathbb{E}\|\mathbf{Y}^T\mathbf{W}\|_{\mathbf{F}}^{\mathbf{2}}$. Our computation relies on the assumption that the error vectors $\mathbf{W}_1, \ldots, \mathbf{W}_n$ are independent. In addition, we use the following fact from [5]: if $\mathbf{u}$, $\mathbf{v}$ are independently and uniformly distributed column vectors on $\mathbb{S}^{m-1}$, then there hold:

$$\mathbb{E}(\mathbf{u}^T \cdot \mathbf{v}) = 0, \quad \mathbb{E}[(\mathbf{u}^T \cdot \mathbf{v})^2] = \frac{1}{m}, \quad \mathrm{Var}[(\mathbf{u}^T \cdot \mathbf{v})^2] = \frac{2(m-1)}{m^2(m+2)}$$

For $i \neq j$, we have

$$\begin{aligned}
\mathbb{E}[(\mathbf{W}_i^T\mathbf{W}_j)^2] &= \mathbb{E}\left[\left(\frac{\mathbf{W}_i^T}{\|\mathbf{W}_i\|} \cdot \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|}\right)^2 \cdot \|\mathbf{W}_i\|^2 \cdot \|\mathbf{W}_j\|^2\right] = \frac{1}{m} \cdot \mathbb{E}(\|\mathbf{W}_i\|^2 \cdot \|\mathbf{W}_j\|^2) \\
&= \frac{1}{m} \cdot \mathbb{E}(\|\mathbf{W}_i\|^2) \cdot \mathbb{E}(\|\mathbf{W}_j\|^2) = \frac{m^2\sigma_i^2\sigma_j^2}{m} = m\sigma_i^2\sigma_j^2
\end{aligned}$$

Define $z_i := \mathbf{W}_i^T\mathbf{W}_i = \sum_{j=1}^{m} e_{ji}^2$. We have

$$\mathbb{E}(e_{ji}^2) = \sigma_i^2, \quad \mathrm{Var}(e_{ji}^2) = \mathbb{E}(e_{ji}^4) - (\mathbb{E}(e_{ji}^2))^2 = \mu_i^4 - \sigma_i^4$$

Then we have

$$\begin{aligned}
\mathbb{E}(z_i) &= \mathbb{E}(\sum_{j=1}^{m} e_{ji}^2) = \sum_{j=1}^{m} \mathbb{E}(e_{ji}^2) = m\sigma_i^2 \\
\mathrm{Var}(z_i) &= \mathrm{Var}(\sum_{j=1}^{m} e_{ji}^2) = \sum_{j=1}^{m} \mathrm{Var}(e_{ji}^2) = m(\mu_i^4 - \sigma_i^4) \\
\mathbb{E}(z_i^2) &= (\mathbb{E}(z_i))^2 + \mathrm{Var}(z) = m^2\sigma_i^4 + m(\mu_i^4 - \sigma_i^4)
\end{aligned}$$

In sum, we have

$$\begin{aligned}
\mathbb{E}(\|\mathbf{W}^T\mathbf{W}\|_F^2) &= \sum_{i=1}^{n} \mathbb{E}[(\mathbf{W}_i^T\mathbf{W}_i)^2] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathbb{E}[(\mathbf{W}_i^T\mathbf{W}_j)^2] \\
&= m^2\sum_{i=1}^{n} \sigma_i^4 + m\sum_{i=1}^{n}(\mu_i^4 - \sigma_i^4) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} m\sigma_i^2\sigma_j^2
\end{aligned}$$

[5]For simplicity, we typically suppress the dependence on δ in our notations, such as using σ instead of $\sigma(\delta)$.

Expectation bounds An application of Jensen's inequality yields $\mathbb{E}(x) \leq \sqrt{\mathbb{E}(x^2)}$. Then we can upper bound $E(\|\mathbf{\Delta}\|_F)$ as follows

$$\begin{aligned}
&\mathbb{E}(\|\mathbf{\Delta}\|_F) \leq 2\mathbb{E}(\|\mathbf{Y^T W}\|_F) + \mathbb{E}(\|\mathbf{W^T W}\|_F) \leq 2\sqrt{\mathbb{E}(\|\mathbf{Y^T W}\|_F^2)} + \sqrt{\mathbb{E}(\|\mathbf{W^T W}\|_F^2)} \\
= \;& 2\sqrt{\sum_{i=1}^{n}\lambda_i \cdot \sum_{i=1}^{n}\sigma_i^2} + \sqrt{m^2\sum_{i=1}^{n}\sigma_i^4 + m\sum_{i=1}^{n}(\mu_i^4 - \sigma_i^4) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} m\sigma_i^2\sigma_j^2} \\
\approx \;& 2\sqrt{\sum_{i=1}^{n}\lambda_i \cdot \sum_{i=1}^{n}\sigma_i^2} + \sqrt{m^2\sum_{i=1}^{n}\sigma_i^4 + mn\sum_{i=1}^{n}\sigma_i^4} := \sqrt{n}\cdot Tol_F
\end{aligned}$$

Combining with Mirsky's theorem, we have that

$$\mathbb{E}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\lambda}_i - \lambda_i)^2} \leq \mathbb{E}\left(\frac{\|\mathbf{\Delta}\|_F}{n}\right) \leq Tol_F,$$

where Tol_F is given by our foregoing analysis.

Computation of variances The variances of the terms analyzed above can also be computed analytically. Using the following identity for independent variables X and Y that

$$\mathrm{Var}(XY) = \mathrm{Var}(X)\mathrm{Var}(Y) + (\mathbb{E}Y)^2\mathrm{Var}(X) + (\mathbb{E}X)^2\mathrm{Var}(Y),$$

we obtain

$$\begin{aligned}
\mathrm{Var}(\|\mathbf{Y^T W}_i\|^2) &= \mathrm{Var}(\|\mathbf{Y^T v}_i\|^2\|\mathbf{W}_i\|^2) \\
&= \mathrm{Var}(\|\mathbf{Y^T v}_i\|^2)\mathrm{Var}(\|\mathbf{W}_i\|^2) + (\mathbb{E}\|\mathbf{W_i}\|^\mathbf{2})^\mathbf{2}\mathrm{Var}\|\mathbf{Y^T v_i}\|^\mathbf{2} + (\mathbb{E}\|\mathbf{Y^T v_i}\|^\mathbf{2})^\mathbf{2}\mathrm{Var}\|\mathbf{W_i}\|^\mathbf{2} \\
&\leq \frac{2}{m+2}\mathrm{Var}(\|\mathbf{W_i}\|^\mathbf{2}) + \frac{\mathbf{2}}{\mathbf{m+2}}(\mathbb{E}\|\mathbf{W_i}\|^\mathbf{2})^\mathbf{2} + \frac{\|\mathbf{Y^T}\|_\mathbf{F}^\mathbf{4}}{\mathbf{m^2}}\mathrm{Var}(\|\mathbf{W_i}\|^\mathbf{2}) \\
&= \frac{2m}{m+2}\mathrm{Var}(e_{1i}^2) + \frac{2m^2}{m+2}(\mathbb{E}e_{1i}^2)^2 + \frac{1}{m}\|\mathbf{Y}\|_F^4\mathrm{Var}(e_{1i}^2).
\end{aligned}$$

Noting that $\mathbf{W_1}, ..., \mathbf{W_n}$ are independent, each element e_{ji} has the forth moment μ_i^4, then we have $\mathrm{Var}(e_{ji}^2) = E(e_{ji}^4) - (E(e_{ji}^2))^2 = \mu_i^4 - \sigma_i^4$. Thus,

$$\begin{aligned}
\mathrm{Var}(\|\mathbf{Y^T W}\|_F^2) &= \mathrm{Var}\left(\sum_{i=1}^{n}\|\mathbf{Y^T W}_i\|_F^2\right) = \sum_{i=1}^{n}\mathrm{Var}(\|\mathbf{Y^T W}_i\|^2) \\
&\leq \frac{2m}{m+2}\cdot\sum_{i=1}^{n}\mathrm{Var}(e_{1i}^2) + \frac{2m^2}{m+2}\sum_{i=1}^{n}\sigma_i^4 + \frac{1}{m}\|\mathbf{Y}\|_F^4\sum_{i=1}^{n}\mathrm{Var}(e_{1i}^2) \\
&= \frac{2m}{m+2}\cdot\sum_{i=1}^{n}(\mu_i^4 - \sigma_i^4) + \frac{2m^2}{m+2}\sum_{i=1}^{n}\sigma_i^4 + \frac{1}{m}\|\mathbf{Y}\|_F^4\sum_{i=1}^{n}(\mu_i^4 - \sigma_i^4)
\end{aligned}$$

The variance of $\|\mathbf{W}^T\mathbf{W}\|_F^2$ can also be computed analytically using result from [5]. The computation is tedious, so we omit the procedure here.

Note that our computation of means and variances can be simplied signficantly by using further assumption on the distribution of the error elements e_{ji} of matrix $\mathbf{W}$, so that the result depend directly on the slack parameters $\delta_i (i = 1, \ldots, n)$. For example, if e_{1i} is uniformly distributed on $[-\delta_i, \delta_i]$, we have $\mathrm{Var}(e_{1i}^2) = \mu_i^4(\delta_i) - \sigma_i^4(\delta_i) = \frac{\delta_i^4}{5} - \frac{\delta_i^4}{9} = \frac{4\delta_i^4}{45}$, and so on. On the ther hand, if $e_{1i} \sim N(0, \sigma_i^2(\delta_i))$, we have $\mathrm{Var}(e_{1i}^2) = \mu_i^4(\delta_i) - \sigma_i^4(\delta_i) = 3\sigma_i^4(\delta_i) - \sigma_i^4(\delta_i) = 2\sigma_i^4(\delta_i)$ and so on.

To compare the variance with the expectation, we use Gaussian distribution and assum all $\sigma_i^2 = \sigma^2$ for analysis. Ingoring the high order term in the expectation, we get

$$2\sqrt{\sum_{i=1}^{n}\lambda_i \cdot n\sigma^2} \approx \sqrt{n}\cdot Tol_F \quad \longrightarrow \quad \sigma^2 \approx \frac{Tol_F^2}{4\sum\lambda_i}$$

The variance is dominated by its last term, which is

$$\begin{aligned}\mathrm{Var}\left(\frac{\|\mathbf{Y^T W}\|_F^2}{n}\right) &\approx \frac{1}{mn^2}\|\mathbf{Y}\|_F^4 \sum_{i=1}^{n} 2\sigma_i^4 = \frac{1}{mn^2}\left(\sum \lambda_i\right)^2 \cdot 2n\sigma^4 \\ &= \frac{1}{mn^2}\left(\sum \lambda_i\right)^2 \cdot \frac{2n \cdot Tol_F^4}{16\left(\sum \lambda_i\right)^2} = \frac{Tol_F^4}{8mn}\end{aligned}$$

which should go to zero as n goes to infinite. We can obtain similar results when using other distributions. So we can conclude that expectaion bounds are actually the worst case bounds with high probability.

6.2.2 Analysis of spectral norm

In this subsection, we turn to the estimation of the spectral norm of the perturbation error matrix $\mathbf{\Delta}$. This quantity provides a tighter upper bound for the eigenvalue perturbation (via Weyl's theorem). Unfortunately, it is also difficult to bound. For many applications, it suffices to replace a bound on $\|.\|_2^2$ by its expectation $\mathbb{E}\|.\|_2^2$. In the following derivations, we rely on the concentration of eigenvalues of random symmetric matrices [1]. This result is applicable to matrices whose elements are independent or weakly correlated.

Let $\mathcal{L}_{max}(\cdot)$ denote the maximum eigenvalue of a matrix. Then we have

$$\begin{aligned}\mathbb{E}(\|\mathbf{W^T Y}\|_2^2) &= \mathbb{E}(\mathcal{L}_{max}(\mathbf{Y^T W W^T Y})) \approx \mathcal{L}_{max}(\mathbb{E}(\mathbf{Y^T W W^T Y})) \\ &= \mathcal{L}_{max}(\mathbf{Y}^T\mathbb{E}(\mathbf{WW})^T\mathbf{Y}) = \mathcal{L}_{max}(\mathbf{Y}^T[\sum_{i=1}^{n}\sigma_i^2\mathbf{I}]\cdot\mathbf{Y}) = \mathcal{L}_{max}(\mathbf{Y}^T\mathbf{Y})\cdot\sum_{i=1}^{n}\sigma_i^2 \\ &= \lambda_{max}\cdot\sum_{i=1}^{n}\sigma_i^2.\end{aligned}$$

Likewise, we have

$$\begin{aligned}\mathbb{E}(\|\mathbf{Y^T W}\|_2^2) &= \mathbb{E}(\mathcal{L}_{max}(\mathbf{W^T Y Y^T W})) \approx \mathcal{L}_{max}(\mathbb{E}(\mathbf{W^T Y Y^T W})) \\ &= \mathcal{L}_{max}\left(\mathbb{E}\left[\mathbf{W}_i^T\mathbf{Y}\mathbf{Y}^T\mathbf{W}_j\right]_{1\le i,j\le n}\right) = \mathcal{L}_{max}\left(\mathbb{E}\left[\sum_{k,l}^{m} e_{ik}(\mathbf{YY}^T)_{kl}e_{jl}\right]\right)\end{aligned}$$

Because the elements e_{ji} of matrix $\mathbf{W}$ are independent with mean 0, the matrix inside $\mathcal{L}_{max}$ is a diagonal matrix. As a result,

$$\begin{aligned}\mathbb{E}(\|\mathbf{Y^T W}\|^2) &= \mathcal{L}_{max}\left(\mathbb{E}\left[\sum_{k=1}^{m}\sigma_i^2(\mathbf{YY}^T)_{kk}\right]_{1\le i\le n}\right) \\ &= \max_i\left\{\sigma_i^2\sum_{k=1}^{m}(\mathbf{YY}^T)_{kk}\right\} = \sigma_{max}^2\sum_{k=1}^{m}(\mathbf{YY}^T)_{kk} \\ &= \sigma_{max}^2\mathrm{tr}(\mathbf{YY^T}).\end{aligned}$$

A remaining term is $\mathbb{E}\|\mathbf{W^T W}\|_\mathbf{2}$, which is generally dominated by $\mathbb{E}(\|\mathbf{Y^T W}\|_2) + \mathbb{E}(\|\mathbf{W^T Y}\|)$ and is omitted in our analysis. Thus we have the following *approximate* upper bound on expected spectral norm of the perturbation error matrix:

$$\mathbb{E}\|\mathbf{\Delta}\|_2 \lesssim Tol_2,$$

where

$$Tol_2 = \sqrt{\lambda_{max}\cdot\sum_{i=1}^{n}\sigma_i^2} + \sqrt{\sigma_{max}^2\mathrm{tr}(\mathbf{YY^T})}.$$

By Weyl's theorem, there holds

$$\mathbb{E}\max_i|\lambda_i - \hat{\lambda}_i| \lesssim Tol_2.$$