Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems, Publication 4
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2014-Philadelphia, PA

Network traffic anomalies, natural language processing, and random matrix theory

Pedro N. Safier[a]*, Ira S. Moskowitz[b]

[a]*S&J Solutions LLC, 107 S. West St PMB 509, Alexandria, VA 22314, USA*
[b]*Ira S. Moskowitz, Code 5580, Naval Research Laboratory, Washington, DC 20375, USA*

Abstract

Random Matrix Theory (RMT) is an important tool for detecting correlations in multidimensional time series, such as stock market price histories, and origin-destination flows in data networks.
We review the basic theory and propose two novel applications: the detection of traffic anomalies in data networks and natural language processing.
For traffic anomalies the advantage of this approach is that training sets are not necessary. In the case of natural language processing, our approach is a refinement of the standard Latent Semantic Analysis (LSA).
We will demonstrate applications to real traffic from a data network, and present the use in Natural Language Processing.
Directions for future work will be discussed.

Selection and peer-review under responsibility of scientific committee of Missouri University of Science and Technology.

Keywords: Computer Networks; Traffic Anomaly Detection; Random Matrix Theory; Natural Language Processing.

* Corresponding author. Tel.: +1-703-765-5047; fax: +1-703-997-1401.
E-mail address: pedro-safier@sj-solutions.com

Selection and peer-review under responsibility of scientific committee of Missouri University of Science and Technology.

1. Introduction

The increasing practicality of large-scale flow capture makes it possible to conceive of Internet traffic analysis methods that detect and identify a large and diverse set of anomalies. However the challenge of effectively analyzing this massive amount of data for anomaly diagnosis is as yet unmet.

Over the past two decades great effort has been invested in meeting this challenge using an impressive array of different mathematical tools and approaches to harvest and understand network-wide views of traffic in the form of sampled flow data[1]. One of this techniques is Random Matrix Theory (RMT).

Random matrices were introduced in Nuclear Physics to model the spectra of heavy nuclei. Since its introduction, RMT has been used to investigate ultrasonic resonances of structural materials, chaotic systems, the zeros of the Riemann and other zeta functions, and any sufficiently complicated system[2].

Barthelemy et al.[3] showed that RMT can be used to detect correlations among different origin-destination flows. Here we argue that the correlations that are detected can be used to discover network-wide traffic anomalies.

2. Overview of Random Matrix Theory

Let $\mathbf{A}$ be a matrix of size $N \times L$, $L \geq N$, with entries a_{ij} that are i.i.d. random variables drawn from a probability distribution with zero mean and variance σ. Such matrix is called a random matrix.

The most common case is that where the a_{ij} are drawn from a zero-mean Gaussian distribution with variance σ.

Consider now the covariance or Wishart matrix $\mathbf{R}$ of size $N \times N$

$$\mathbf{R} = \frac{1}{N}\mathbf{A} \cdot \mathbf{A}^T, \tag{1}$$

A fundamental result of RMT is that the eigenvalues of $\mathbf{R}$ are distributed according to the probability density function (pdf)

$$P_{\lambda}^{(\mathrm{rmt})} = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \quad \lambda_- \leq \lambda \leq \lambda_+, \tag{2}$$

where

$$\lambda_{\pm} = \sigma^2\left(1 + 1/Q \pm 2\sqrt{1/Q}\right) \tag{3}$$

and

$$Q = \lim_{\substack{N\to\infty \\ L\to\infty}} \frac{L}{N}. \tag{4}$$

In general, some further explanation is needed for the limit in eq. (4) to exist and its value. For the rest of this paper we will safely assume[4] that this limiting value is simply L/N.

Note that the requirement $L \geq N$ in the definition of $\mathbf{A}$ implies that $Q \geq 1$. Hereafter, and without loss of generality, we assume that $\sigma = 1$.

Even though eq. (2) was first derived for the case of a matrix with random elements from a zero-mean Gaussian distribution, recent results[5] show that eq. (2) also applies when the a_{ij} are randomly drawn from distributions that satisfy

$$\mathrm{E}a_{ij} = 0, \tag{5a}$$

$$\mathrm{E}a_{ij}^2 = 1; \tag{5b}$$

where E stands for the expectation value.

Figure 1 is a plot of eq. (2) for $N = 100$ and different values of L. The key feature that makes this result useful is that if $\mathbf{A}$ is truly a random matrix with elements that satisfy the conditions in eq. (5), then the eigenvalues of its covariance matrix are in the range $[\lambda_-, \lambda_+]$ and follow the pdf in eq. (2). If we do not use the limiting case from eq. (4) the range is approximate[4].

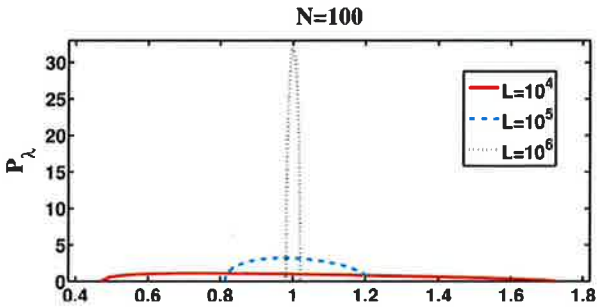

Fig. 1. The RMT eigenvalue probability density function eq. (2)

Why is this result useful? Suppose that we have N variables, and that each of them is observed L times. We would like to know whether the data at hand is random; more precisely, we would like to know whether the observations are uncorrelated.

The null hypothesis is then that the $N \times L$ observations are indistinguishable from $N \times L$ i.i.d. random variables drawn from a distribution with zero mean and unit variance. To answer this question using RMT, define the row vector $\mathbf{x}_i$ of length L that contains the observations of variable i, i.e., x_{ij} is the jth observation of variable i. Assume that $\mathbf{x}_i$ is standardized, i.e. it is zero-mean and has unity standard deviation. In the next step, form the matrix $\mathbf{X}$ using for its rows the vectors $\mathbf{x}_i, i = 1, \ldots, N$; compute its covariance (or Wishart) matrix

$$\mathbf{C} = \frac{1}{L}\mathbf{X} \cdot \mathbf{X}^T, \tag{6}$$

and finally compute the eigenvalues of $\mathbf{C}$.

If the eigenvalues of $\mathbf{C}$ are distributed according to $P_\lambda^{(\mathrm{rmt})}$, then the data are random variables drawn from one or more distributions satisfying eq. (5). Otherwise, we can discover correlations in the data by consider the eigenvalues outside the range $[\lambda_-, \lambda_+]$ and analyzing the corresponding eigenvectors. Of course, we have to keep in mind that we are only approximating eq. (4). This approach has been used successfully to analyze financial data[6] and Internet traffic[3].

3. Application of RMT to Anomalous Traffic Detection

The application to anomalous traffic detection is straightforward.

Suppose that we have a network with N flows, and we wish to detect traffic anomalies; and suppose that L time samples are obtained for each flow. Let $\mathbf{x}_i$ be the vector containing the L time-samples for flow i.

As explained above, we form the matrix $\mathbf{X}$ from the individual flows $\mathbf{x}_i$—we shall call the matrix $\mathbf{X}$ the traffic matrix—and proceed to compute its covariance matrix $\mathbf{C}$ as in eq. (6). However, an additional step is required

here: instead of looking at the vector $\mathbf{x}_i$ it is better to consider the logarithm of the ratio of successive elements of $\mathbf{x}_i$ along the time dimension, i.e., we use

$$\hat{x}_{ij} \triangleq \ln x_{i,j+1} - \ln x_{i,j}\,,\ j = 1,\ldots,L-1. \tag{7}$$

The rationale for $\hat{\mathbf{x}}_i$ using instead of $\mathbf{x}_i$ is the removal of any multiplicative biases in the time series.

Next, each $\hat{\mathbf{x}}_i$ is normalized to have zero mean and unit variance:

$$\hat{\mathbf{z}}_i \triangleq \left(\hat{\mathbf{x}}_i - <\hat{\mathbf{x}}_i>\right)/<\hat{\mathbf{x}}_i^2>, \tag{8}$$

where $\langle\bullet\rangle$ indicates a sample average. Thus, the matrix $\hat{\mathbf{Z}}$ formed from the $\hat{\mathbf{z}}_i$ has dimensions $N\times(L-1)$, while its covariance matrix $\hat{\mathbf{C}}$ has the same dimensions as $\mathbf{C}$, namely $N\times N$.

Here we present two examples of this application: first by considering a simulated traffic matrix, and, then using real network traffic.

3.1. Simulated Network Traffic

The purpose of this simulation is to create traffic that is uncorrelated, and then artificially insert correlations between the different flows to test whether RMT can detect these anomalous correlations.

In principle, given M nodes and N edges, a realistic simulation should take into account the traffic created by each of the nodes, the way this traffic is routed, and the time-dependent behavior of the different protocols that handle this traffic[7]. However, a reasonable shortcut is based on the finding[8] that internet flows can be modeled as α-stable[9] flows .

We used the Matlab routine `stabrnd` to generate 100 α-stable flows ($N=100$) each consisting of 200 samples ($L=200$). For each flow, we chose randomly the distribution parameters α and β from $\alpha\in[0.5,1.8]$ and $\beta\in[-0.5,0.5]$ with uniform probability. Note that we simulated the values of $\hat{\mathbf{x}}_i$, *not* $\mathbf{x}_i$. Furthermore, the values of $\hat{\mathbf{x}}_i$ thus obtained were limited to $\hat{x}_{ij}\in[-20,20]$, i.e., we limited the range of the logarithmic changes to one compatible with observed values.

Next, we computed the matrix $\hat{\mathbf{Z}}$, its covariance matrix $\hat{\mathbf{C}}$ and its eigenvalues. Figure 2a is a plot of the eigenvalue distribution and a comparison with the eigenvalue distribution predicted by RMT, eqs. (2)-(4) with $Q=1.99$.

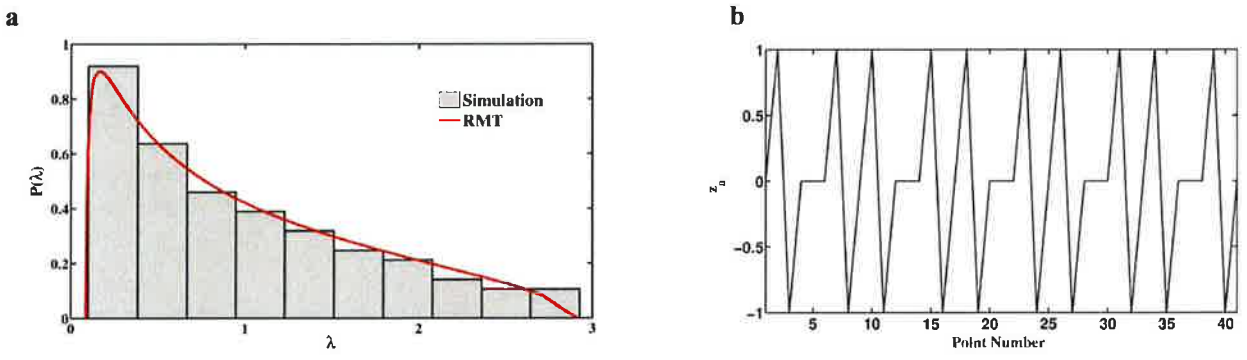

Fig. 2. (a) Comparison of the pdf in eq.(2) with that for our simulation with α-flows; (b) the injected anomaly.

Next, we artificially introduce correlations in the simulated data to test the detection capabilities of RMT. To do so, we randomly chose ten flows and a time interval with forty points, and added the anomaly in Fig. 2b; because the

flows are normalized to unit variance, the anomaly amplitude was set to 2.5, to obtain enough of a signal. The resulting anomalous flows can be seen in Fig. 3.

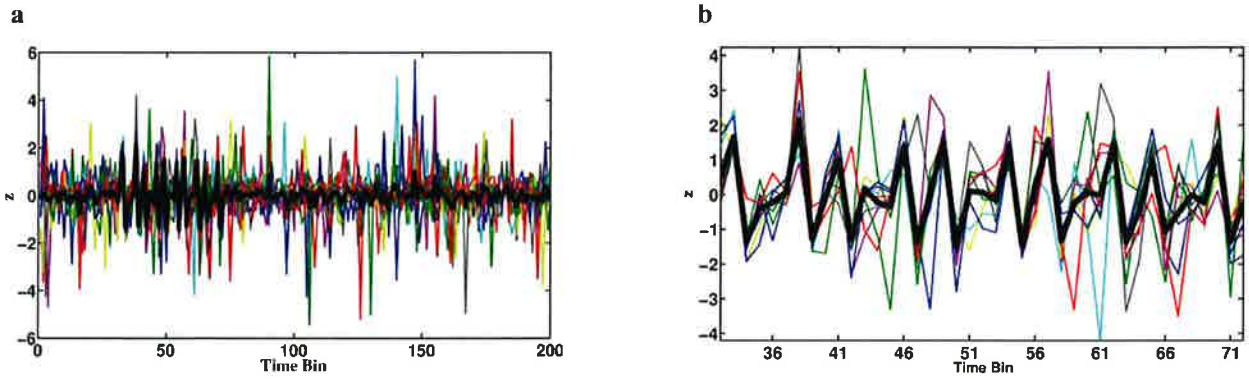

Fig. 3. (a) The ten flows with the injected anomaly. The thick black line is the mean value; (b) zoomed view of (a).

We computed the eigenvalues of the matrix with the anomalous flows, $\hat{\mathbf{Z}}$, and compared their distribution to that predicted by RMT. The comparison can be seen in Fig.4.

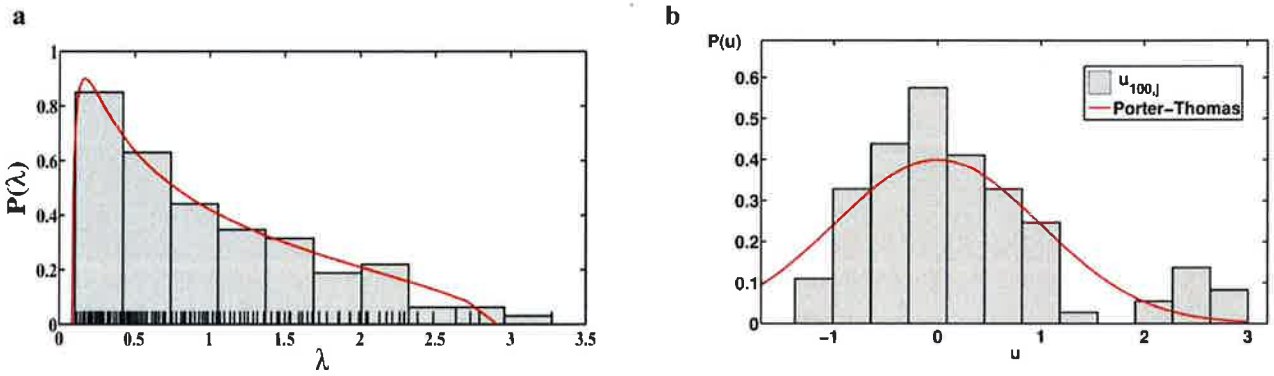

Fig. 4. (a) Deviation from the RMT eigenvalue pdf after anomaly injection; (b) eigenvector components for the largest eigenvalue.

Note that the anomalies introduced as described above appear in Fig. 4a as an outlying eigenvalue $\lambda = 3.3$, while RMT predicts $\lambda_+ = 2.9$. However, these results raise a question: which of the flows are correlated? Fig. 3 shows that even if we knew which flows were anomalous, it is difficult to detect the anomaly by eye (this is why we plotted the mean value to guide the eye).

The answer is found by looking at the eigenvector of the eigenvalue under consideration, in this case the largest one. According to RMT, if the components of eigenvector $\mathbf{u}_i$ are normalized such that $\sum u_{ij}^2 = N$, then they should be distributed as a Gaussian pdf with zero mean and unit variance—in the RMT literature this distribution of eigencomponents is known as the Porter-Thomas distribution.

In our case, however, the components of the eigenvectors that correspond to the eigenvalues that deviate from RMT do *not* obey the Porter-Thomas distribution; therefore, by looking at the pdf of the eigenvector components and their values we can identify the flows that are correlated. This is demonstrated in Fig. 4b for the largest eigenvalue: those components with value greater than ≈ 2 are anomalous.

In Fig. 5a we plotted the absolute value of the eigencomponents for the largest eigenvalue, and two horizontal lines that identify the significance level p of the deviations from the Porter-Thomas distribution in Fig. 4b; the thick lines identify the location of the anomalous flows. Note that at a significance level $p = 0.05$ we correctly identify all the anomalous flows, while at $p = 0.01$ three out of ten anomalous flows are detected.

This example illustrates the power of the technique we propose.

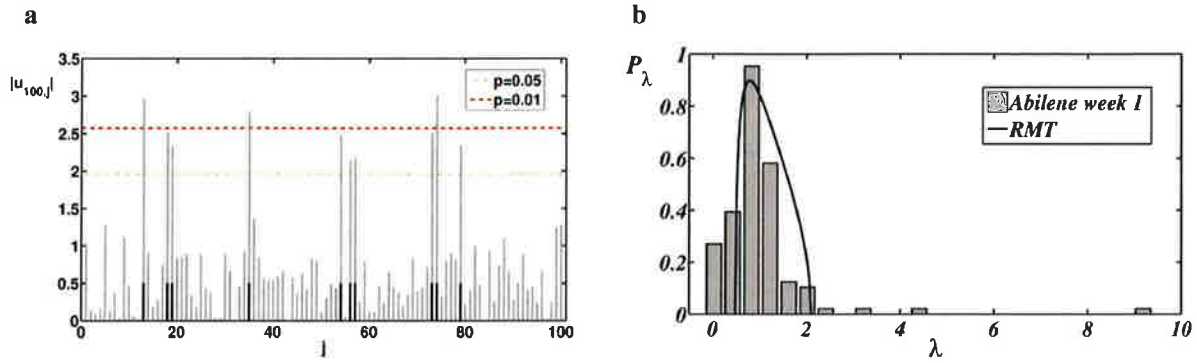

Fig. 5. (a) Eigenvector components for the largest eigenvalue (#100); (b) eigenvalue distribution for the Abilene data.

3.2. Real Network Traffic

To test our methodology we used one of the Abilene network datasets. The Abilene network was a high-performance backbone network created by the Internet2 community in the late 1990s. In 2007 the Abilene network was retired and the upgraded network became known as the "Internet2 network." Abilene was a private network used for education and research, but was not entirely isolated, since its members usually provided alternative access to many of their resources through the public Internet. The network backbone consisted of 11 points of presence (PoPs).

The dataset used here consists of the 121 origin-destination (OD) flows between the 11 PoPs; note that flows originating and ending in the same PoP are included. It is the same used by Lakhina et al.[10] ; these authors collected three weeks of sampled IP-level traffic flow data from every PoP in Abilene for the period 8 December 2003 to 28 December 2003. Sampling was periodic, at a rate of 1 out of 100 packets. The network reported flow statistics every 5 minutes, and therefore the data is binned in 5-minutes bins.

We used the RMT approach on the first week of data (2016 time samples). Fig. 5b shows the eigenvalue distribution and a comparison with that predicted by RMT theory.

Next, we examined the components of the eigenvector corresponding to the largest eigenvalue in Fig. 5b. The absolute value of these components are plotted in Fig 6a. Note the very large values for components 45–55.

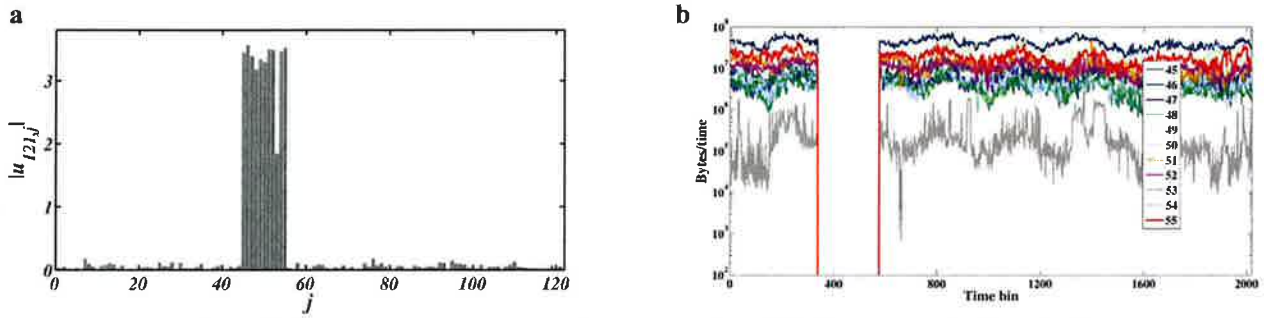

Fig. 6. (a) Eigenvector components for the largest eigenvalue (#121); (b) time series for flows 45–55.

The very large values for flows 45–55 in Fig. 6a point to a strong correlation between these flows, and we proceeded to examine the corresponding time series; these are plotted in Fig. 6b. These results show the ability of RMT to detect an anomaly: there was an outage at the Indianapolis node (the point of origin of flows 45–55) during the first week of data; while the outage lasted, it caused a perfect correlation between the OD flows originating at

the Indianapolis node; and this correlation resulted in a very large eigenvalue—much larger than that predicted by RMT.

4. Natural Language Processing

Natural Language Processing is concerned with the interactions between computers and human (natural) languages. An important example is the analysis of relationships between a set of documents (a corpus) and the terms they contain by producing a set of concepts related to the documents and terms. This type of semantic analysis posits that each document is a mixture of a small number of topics and that the occurrence of every word is attributable to one of the document's topics. The goal is, then, to identify the topics in a corpus.

In particular, Latent Semantic Analysis[11,12] (LSA) is a very important, zero-knowledge technique. The approach of LSA is to first build a matrix, the so-called occurrence matrix, that is formed of word counts per document (rows represent unique words and columns represent each document) constructed from a corpus; and then to look for correlations between the words by using Singular Value Decomposition (SVD) to reduce the number of columns while preserving the similarity structure among rows.

A key step in LSA is dimensionality reduction of the problem; this is achieved by selecting the k -largest singular values obtained by SVD. An open problem[11,12] is how to choose the cut-off rank k_0 below which the singular values are discarded.

We propose the use of RMT to identify k_0, assuming that the elements of the occurrence matrix are drawn from a distribution that satisfies eq. (5); without loss of generality let us assume that they are drawn from a Gaussian distribution with zero-mean and unit variance.

To be specific, Let w_{ij} be the number of times that word i appears in document j. In other words, consider the vector $\mathbf{w}_i$, $i=1,\ldots,V$ with dimension N (the number of documents in the corpus), where V is the number of words considered.

The analysis would then proceed as follows: (1) compute the eigenvalues of the correlation matrix $\mathbf{C}$ for the matrix $\mathbf{W}$ formed with the vectors $\mathbf{w}_i$ after these are normalized to have zero mean and unit variance; (2) set $Q=N/V$ and compute the corresponding RMT eigenvalue pdf, $P_\lambda^{(\mathrm{rmt})}$ [eqs. (2)–(3)]; (3) compare the distribution of eigenvalues of $\mathbf{C}$ with $P_\lambda^{(\mathrm{rmt})}$.

The rank of the eigenvalues of $\mathbf{C}$ at which the empirical distribution deviates from $P_\lambda^{(\mathrm{rmt})}$ should identify the correct value of k_0 to use. Furthermore, the statistical significance of k_0 chosen thus can be evaluated with the help of the Tracy-Widom distribution[4].

It might be argued that words in a document follow a Poisson rather than a Gaussian distribution, but note that for a Poisson parameter $\nu \gtrsim 10$ the distribution is nearly Gaussian. Otherwise, before standardizing the observations, a variance-stabilization transformation[13] can be performed. In particular, using the transformation

$$x \longrightarrow \sqrt{x+\frac{3}{8}} \tag{9}$$

a Poisson-distributed variable is transformed into a Gaussian-distributed one with zero mean and variance 1/4.

To our knowledge, this technique to choose k_0 has never been tried before; we intend in the near future to investigate the practicality of this approach.

5. Summary and Future Work

We have shown that the application of Random Matrix Theory holds great promise in detecting network traffic anomalies in the form of anomalous correlations between origin-destination flows.

The results presented here are exploratory and several questions remain to be studied:

1. We have detected a network outage using RMT. However, this kind of anomaly introduces a perfect correlation for a finite time span. How can we detect less than perfect correlations?

2. Some of the correlations between flows are benign, for example, the 24hr variations due to changes in human activity. How can these be filtered out from the analysis?
3. What are the correlation signatures of other type of anomalies, in particular anomalies of malicious origin?
4. What is the optimal time-window for analysis?
5. So far, we have focused on the largest eigenvalue of the covariance matrix; what can be learned from the second-largest and subsequent-largest eigenvalues?
6. Last, but not least, in our analysis we used the OD flows instead of the raw packet data. Can we apply RMT to the raw data? This question is relevant for the implementation of automatic, real-time systems for anomaly detection.

As for Natural Language Processing, we have proposed a new method based on RMT to reduce the dimensionality of the problem when tackled with LSA. A still to be proven advantage of our proposal is the systematic way to choose the cut-off rank for the relevant dimensions.

Acknowledgments

We are most indebted to Dr. Paul Hyden for introducing us to the subject of Natural Language Processing and for his many comments and suggestions.

References

1. Bhuyan M, Bhattacharyya D, Kalita J. Network anomaly detection: methods, systems and tools. *Communications Surveys & Tutorials, IEEE*, 2014; **16**:303–336.
2. Mehta M L. *Random matrices*. Boston:Academic Press; 2004.
3. Barthelemy M, Gondran B, Guichard E. Large scale cross-correlations in internet traffic. *Phys Rev E* 2002; **66**:056110.
4. Tracy C A, Widom H. The distributions of random matrix theory and their applications. In: Sidoravicius, V, editor. *New Trends in Mathematical Physics*. New York: Springer; 2009. p. 753–765.
5. Soshnikov A. Universality at the edge of the spectrum in wigner random matrices. *Comm Math Phys* 1999; **207**:697–733.
6. Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Guhr T, Stanley HE. A random matrix approach to cross-correlations in financial data. *Phys Rev E* 2001; **65**:066126.
7. Yuan J, Mills K. A cross-correlation-based method for spatial–temporal traffic analysis. *Performance Evaluation* 2005;**61**: 163–180.
8. Yu J, Petropulu AA, Sethu H. Rate-limited eafrp—a new improved model for high-speed network traffic. *IEEE Trans Sig Proc* 2005;**53**:505–522.
9. Samoradnitsky G, Taqqu MS. *Stable non-gaussian random processes: stochastic models with infinite variance*. Boca Raton:CRC Press; 1994.
10. Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions. *ACM SIGCOMM Comp Comm Rev* 2005;**35**:217–228.
11. Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA. Indexing by latent semantic analysis. *J Am Soc Inf Sci* 1990;**41**:391–407.
12. Dumais ST. Latent semantic analysis. *Ann Rev Inf Sci Tech* 2004;**38**:188–230.
13. Anscombe FJ. The transformation of poisson, binomial and negative-binomial data. *Biometrika* 1948;**35**:246–254.