Structural Analysis of Network Traffic Flows

Anukool Lakhina, Konstantina Papagiannaki, Mark Crovella,
Christophe Diot, Eric D. Kolaczyk, and Nina Taft *

November 10, 2003

BUCS-TR-2003-021 and RR03-ATL-110708

Abstract

Network traffic arises from the superposition of Origin-Destination (OD) flows. Hence, a thorough understanding of OD flows is essential for modeling network traffic, and for addressing a wide variety of problems including traffic engineering, traffic matrix estimation, capacity planning, forecasting and anomaly detection. However, to date, OD flows have not been closely studied, and there is very little known about their properties.

We present the first analysis of complete sets of OD flow timeseries, taken from two different backbone networks (Abilene and Sprint-Europe). Using Principal Component Analysis (PCA), we find that the set of OD flows has small intrinsic dimension. In fact, even in a network with over a hundred OD flows, these flows can be accurately modeled in time using a small number (10 or less) of independent components or dimensions.

We also show how to use PCA to systematically decompose the structure of OD flow timeseries into three main constituents: common periodic trends, short-lived bursts, and noise. We provide insight into how the various constitutents contribute to the overall structure of OD flows and explore the extent to which this decomposition varies over time.

*A. Lakhina and M. Crovella are with the Department of Computer Science, Boston University; email: {anukool,crovella}@cs.bu.edu. K. Papagiannaki is with Sprint Advanced Technology Labs; email: dina@sprintlabs.com. C. Diot is with Intel Research, Cambridge; email: christophe.diot@intel.com. E. D. Kolaczyk is with the Department of Mathematics and Statistics, Boston University; email: kolaczyk@math.bu.edu. N. Taft is with Intel Research, Berkeley; email: nina.taft@intel.com. This work was performed while M. Crovella was at Laboratoire d'Informatique de Paris 6 (LIP6), with support from Centre National de la Recherche Scientifique (CNRS) France. Part of this work was done when N. Taft was at Sprint Advanced Technology Labs and A. Lakhina was at Sprint Advanced Technology Labs and Intel Research, Cambridge. This work was in part supported by a grant from Sprint Labs, ONR award N000140310043 and NSF grants ANI-9986397 and CCR-0325701.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
10 NOV 2003		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Structural Analysis of Network Traffic Flows	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research,One Liberty Center,875 North Randolph Street Suite 1425,Arlington,VA,22203-1995**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**26**	

1 Introduction

Much of the work in network traffic analysis so far has focussed on studying traffic on a single link in isolation. However, a wide range of important problems faced by network researchers today require modeling and analysis of traffic on all links simultaneously, including traffic engineering, traffic matrix estimation [17, 31, 32], anomaly detection [6, 1], attack detection [30], traffic forecasting and capacity planning [19].

Unfortunately, whole-network traffic analysis – *i.e.,* modeling the traffic on all links simultaneously – is a difficult objective, amplified by the fact that modeling traffic on a single link is itself a complex task. Whole-network traffic analysis therefore remains an important and unmet challenge.

One way to address the problem of whole-network traffic analysis is to recognize that the traffic observed on different links of a network is not independent, but is in fact determined by a common set of underlying origin destination (OD) flows. An origin destination flow is the collection of all traffic that enters the network from a common ingress point and departs from a common egress point. The superposition of these point-to-point flows, as determined by routing, gives rise to all link traffic in a network. Thus, instead of studying traffic on all links, a more direct and fundamental focus for whole-network traffic study is the analysis of the network's set of OD flows.

However, even though OD flows are conceptually a more fundamental property of a network's workload than link traffic, analyzing them suffers from similar difficulties. The principal challenge presented by OD flow analysis is that OD flows form a high dimensional multivariate structure. For example, even a moderate-sized network may carry hundreds of OD flows; the resulting set of timeseries has hundreds of dimensions. The high dimensionality of OD flows is in fact a prime source of difficulty in addressing the whole-network analysis problems listed above. Thus the central problem one confronts in OD flow analysis is the so-called "curse of dimensionality" [7].

In general, when presented with the need to analyze a high-dimensional structure, a commonly-employed and powerful approach is to seek an alternate lower-dimensional approximation to the structure that preserves its important properties. It can often be the case that a structure that appears to be complex because of its high dimension may be largely governed by a small set of independent variables and so can be well approximated by a lower-dimensional representation. Dimension analysis and dimension reduction techniques attempt to find these simple variables and can therefore be a useful tool to understand the original structures.

The most commonly used technique to analyze high dimensional structures is the method of *Principal Component Analysis* [11] (PCA, also known as the Karhunen-Loève procedure and singular value decompositon [25]). Given a high dimensional object and its associated coordinate space, PCA finds a new coordinate space which is the best one to use for dimension reduction of the given object. Once the object is placed into this new coordinate space, projecting the object onto a subset of the axes can be done in a way that minimizes error. When a high-dimensional object can be well approximated in this way in a smaller number of dimensions, we refer to the smaller number of dimensions as the object's *intrinsic dimensionality*.

In this paper, we use PCA to explore the intrinsic dimensionality and structure of OD flows using data collected from two different backbone networks: Abilene and Sprint-Europe. Even though both these networks have over a hundred origin-destination pairs, we show that on long

timescales (days to week), their structure can be well captured using remarkably few dimensions. In fact, we find that using between 5 and 10 dimensions, one can accurately approximate the ensemble of OD flows in each network.

In order to explore the nature of this low dimensionality, we introduce the notion of *eigenflows*. An eigenflow, derived from a PCA of OD flows, is a timeseries that captures a particular source of temporal variability (a "feature") in the OD flows. Each OD flow can be expressed as a weighted sum of eigenflows; the weights capture the extent to which each feature is present in the given OD flow. We show that eigenflows fall into three natural classes: (i) deterministic eigenflows, which capture the predictable periodic trends in the OD flow timeseries, (ii) spike eigenflows, which capture the occasional short-lived bursts in OD flows, (iii) noise eigenflows, which account for traffic fluctuations appearing to have relatively time-invariant properties across all OD flows. This taxonomy, systematically and quantitatively unearthed by PCA, can be viewed as being parallel to characteristics observed in various analyses of network traffic in the literature: periodic trends [19, 23], stochastic bursts [24] and fractional Gaussian (or other) noise [16, 20]. Thus, the systematic decomposition of a set of OD flows into its constitutent eigenflows sheds light on the intrinsic structure of OD flows, and consequently on the behavior of the network as a whole.

In fact, by categorizing eigenflows in this manner, we find that we can obtain significant insight into the whole-network properties of data traffic. First of all, we find that each OD flow is well captured by only a small set of eigenflows. Thus, each OD flow has a certain small set of features. Furthermore, these features vary in a predictable manner as a function of the amount of traffic carried in the OD flow. In particular, we show quantitatively that the largest OD flows in both networks are primarily deterministic and periodic; OD flows of moderate strength are generally comprised of both bursts and comparatively, noise; and the weakest OD flows are generally primarily noise. This broad characterization of the nature of OD flows provides a useful basis for organizing and interpreting studies of whole-network traffic.

Finally, from a broader perspective, an important methodological contribution of our work is the application of a dimension analysis technique to analyze the structure of network traffic. Although we concentrate on timeseries of traffic counts, analogous problems arise when studying delay or loss patterns in networks. Examining intrinsic dimensionality and structure in the manner we outline in this paper may be fruitful in studying other network properties as well.

This paper is organized as follows. We begin in Section 2 with a discussion of the high dimensionality of OD flows and provide the necessary foundations of Principal Component Analysis. We outline the steps taken to collect and construct OD flows from both the Sprint-Europe and Abilene networks in Section 3. We then apply PCA to OD flow timeseries from both networks and present evidence of their low dimensionality in Section 4. We elaborate on the notion of eigenflows and show how they can be interpreted, understood and harnessed in Section 5. In Section 6, we examine the temporal stability of the decomposition of OD flows into their constitutent eigenflows. The low intrinsic dimensionality of OD flows at long timescales suggests new approaches to a number of network engineering problems. A discussion of these and our ongoing work is in Section 7, where we also place our work, which we believe to be the the first to systematically examine complete sets of OD flow data, in the context of related work. Concluding remarks are presented in Section 8.

2 Background

In order to facilitate discussion in subsequent sections, we first introduce relevant notation. Let p denote the number of OD flows in a network and t denote the number of successive time intervals of interest. In this paper, we study networks which have on the order of hundreds of OD Flows, over long timescales (days to weeks) and over time intervals of 5 and 10 minutes so that $t > p$. Let X be the $t \times p$ measurement matrix, which denotes the timeseries of all OD flows in a network. Thus, each column i denotes the timeseries of the i-th OD flow and each row j represents an instance of all the OD flows at time j. We refer to individual columns of a matrix using a single subscript, so OD flow i is denoted X_i. Note that X thus defined has rank at most p. Finally, all vectors in this paper are column vectors, unless otherwise noted.

2.1 OD Flows

An OD flow consists of all traffic entering the network at a given point, and exiting the network at some other point. Each network ingress and egress point serves a distinct customer population.[1] Thus, each OD flow arises from the activity of a distinct user population.

The traffic actually observed on a network link arises from the superposition of OD flows. The relationship between link and flow traffic can be concisely captured in the *routing matrix* A. The matrix A has size (# links) $\times$ (# flows), where $A_{ij} = 1$ if flow j passes over link i, and is zero otherwise. Then the vector of traffic counts on links ($\mathbf{y}$) is related to the vector of traffic counts in OD flows ($\mathbf{x}$) by $\mathbf{y} = \mathbf{A}\mathbf{x}$. Traffic engineering is the process of adjusting A, given some OD flow traffic $\mathbf{x}$, so as to influence the link traffic $\mathbf{y}$ in some desirable way. Thus accurate traffic engineering and link capacity planning depends on a good understanding of the properties of the OD flow vector $\mathbf{x}$.

In a typical network with n PoPs (points of presence where traffic may enter or exit the network) there are n^2 PoP-pairs, and hence n^2 OD flows. Thus even in a moderate sized network with tens of PoPs, there are hundreds of OD flows, meaning that $\mathbf{x}$ is a vector residing in a high dimensional space. Successive OD flow traffic measurements over time (X) then become a high dimensional multivariate timeseries.

Because each OD flow is the result of activity of distinct user populations, it is not clear to what extent OD flows share common characteristics. That is, it is not clear whether we should expect the columns of X to be related (so that the *effective* rank of X is less than p). A particularly powerful approach to answering these questions quantitatively is dimension analysis via PCA.

2.2 Principal Component Analysis

PCA is a coordinate transformation method that maps the measured data onto a new set of axes. These axes are called the principal axes or components. Each principal component has the property that it points in the direction of maximum *variation* or *energy* (with respect to the Euclidean norm)

[1] We assume for purposes of discussion that routing changes do not affect where traffic for a particular population enters or exits.

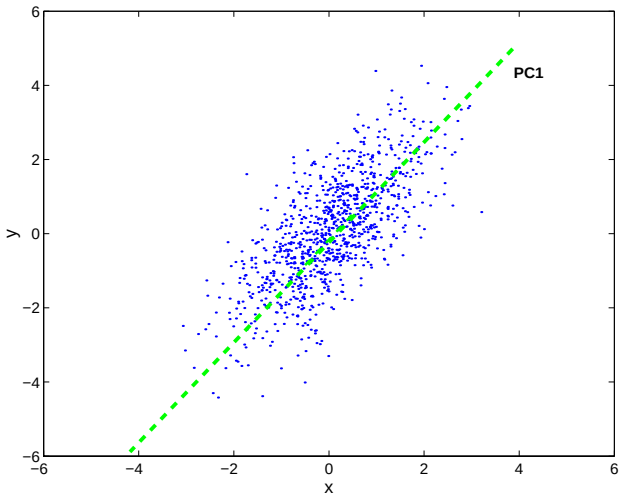


Figure 1: Illustration of PCA on a correlated, 2-D dataset.

remaining in the data, given the energy already accounted for in the preceding components.[2] As such, the first principal component captures the total energy of the original data to the maximal degree possible on a single axis. The next principal components then capture the maximum residual energy among the remaining orthogonal directions. In this sense, the principal axes are ordered by the amount of energy in the data they capture.

The method of PCA can be motivated via a geometric illustration. An application of PCA on a two dimensional dataset is shown in Figure 1. The first principal axis points in the direction of maximum energy in the data. Generalization to higher dimensions, as in the case of X, may be understood as follows. Take the rows of X as points in Euclidean space, so that we have a dataset of t points in $\mathbb{R}^p$. Mapping the data onto the first r principal axes places the data into an r-dimensional hyperplane.

Shifting from the geometric interpretation to a linear algebraic formulation, calculating the principal components is equivalent to solving the symmetric eigenvalue problem for the matrix X^TX. The matrix X^TX is a measure of the covariance between flows. Each principal component v_i is the i-th eigenvector computed from the spectral decomposition of X^TX:

$$X^TXv_i = \lambda_i v_i \quad i = 1, ..., p \tag{1}$$

where λ_i is the eigenvalue corresponding to v_i. Furthermore, because X^TX is symmetric positive definite, its eigenvectors are orthogonal and the corresponding eigenvalues are nonnegative real. By convention, the eigenvectors have unit norm and the eigenvalues are arranged from large to small, so that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$.

To see that calculating the principal components of X is equivalent to computing the eigenvectors of X^TX, consider the first principal component. Let v_1 denote the vector of size p corresponding to the first principal component of X. As mentioned earlier, the first principal axis, v_1, captures the maximum energy of the data:

[2] We will use the terms variation and energy interchangably in the rest of the paper.

$$v_1 = \arg\max_{\|v\|=1} \|Xv\| \tag{2}$$

where $\|Xv\|$ is the energy of the data captured along v. The above equation can be rewritten as:

$$\begin{aligned} v_1 &= \arg\max_{\|v\|=1} \|Xv\| \\ &= \arg\max_{v} \frac{\|Xv\|}{v^T v} \\ &= \arg\max_{v} \frac{v^T X^T X v}{v^T v}. \end{aligned}$$

The quantity being maximized in the last equation above is the *Rayleigh Quotient* of $X^T X$. It can be shown that the eigenvector corresponding to the largest eigenvalue of $X^T X$ (or the first eigenvector) maximizes its Rayleigh quotient (see, for instance [25]). In this way, maximizing the energy of X along the first principal component v_1 is equivalent to computing the first eigenvector of $X^T X$.

Proceeding recursively, once the first $k-1$ principal components have been determined, the k-th principal component corresponds to the maximum energy of the residual. The residual is the difference between the original data and the data mapped onto the first $k-1$ principal axes. Thus, we can write the k-th principal component v_k as:

$$v_k = \arg\max_{\|v\|=1} \|(X - \sum_{i=1}^{k-1} X v_i v_i^T) v\|.$$

By a similar argument, computing the k-th principal component is equivalent to finding the k-th eigenvector of $X^T X$. Thus, in this manner, computing the set of all principal components, $\{v_i\}_{i=1}^{p}$ is equivalent to computing the eigenvectors of $X^T X$.

Once the data have been mapped into principal component space, it can be useful to examine the transformed data one dimension at a time. Considering the data mapped onto the principal components, we see that the contribution of principal axis i as a function of time is given by Xv_i. This vector can be normalized to unit length by dividing by $\sigma_i = \sqrt{\lambda_i}$. Thus, we have for each principal axis i,

$$u_i = \frac{Xv_i}{\sigma_i} \quad i = 1, ..., p \tag{3}$$

The u_i are vectors of size t and orthogonal by construction. The above equation shows that all the OD flows, when weighted by v_i, produce one dimension of the transformed data. Thus vector u_i captures the temporal variation common to all flows along principal axis i. Since the principal axes are in order of contribution to the overall energy, u_1 captures the strongest temporal trend common to all OD flows, u_2 captures the next strongest, and so on. Because the set of $\{u_i\}_{i=1}^{p}$ capture the time-varying trends common to the OD flows, we refer to them as the *eigenflows* of X.

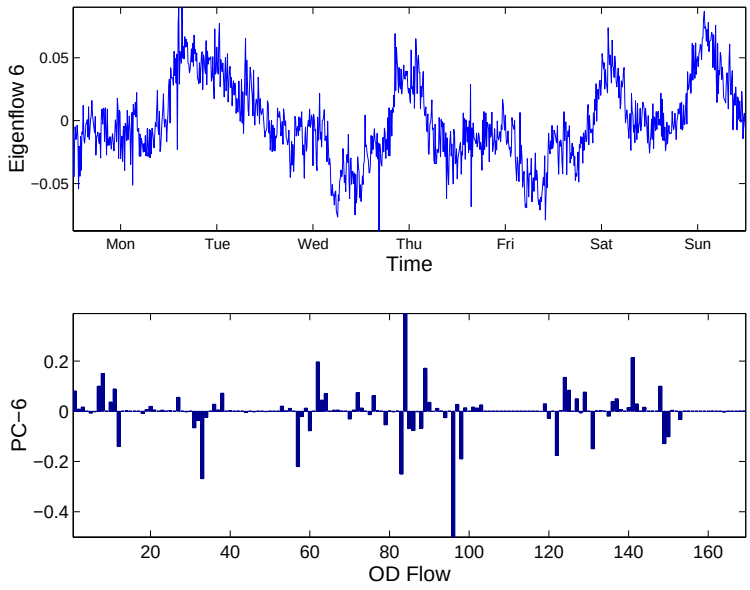


Figure 2: An eigenflow and its corresponding principal component.

The set of principal components $\{v_i\}_{i=1}^{p}$ can be arranged in order as columns of a *principal matrix* V, which has size $p \times p$. Likewise, we can form the $t \times p$ matrix U in which column i is u_i. Then taken together, V, U, and σ_i can be arranged to write each OD flow X_i as:

$$\frac{X_i}{\sigma_i} = U(V^T)_i \quad i = 1, ..., p \tag{4}$$

where X_i is the timeseries of the i-th OD flow and $(V^T)_i$ is the i-th row of V. Equation (4) makes clear that each OD flow X_i is in turn a linear combination of the eigenflows, with associated weights $(V^T)_i$.

In Figure 2 we show typical examples of an eigenflow u_i and its corresponding principal axis v_i. The eigenflow captures a pattern of temporal variation common to the set of OD flows, and the extent to which this particular temporal pattern is present in each OD flow is given by the entries of v_i. In this case, we can see that this eigenflow's feature is most strongly present in OD flow 84 (the strongest peak in v_i).

The elements of $\{\sigma_i\}_{i=1}^{p}$ are called the *singular values*. Note that each singular value is the square root of the corresponding eigenvalue, which in turn is the energy attributable to the respective principal component:

$$\|Xv_i\| = v_i^T X^T X v_i = \lambda_i v_i^T v_i = \lambda_i \tag{5}$$

where the second equality holds from Equation 1, and the last equality follows from the fact that v_i has unit norm. Thus, the singular values are useful for gauging the potential for reduced dimensionality in the data, often simply through their visual examination in a *scree plot*. Specifically, finding that only r singluar values are non-neglegible, implies that X effectively resides on an r-dimensional subspace of $\mathbb{R}^p$. In that case, we can approximate the original X as:

$$X' \approx \sum_{i=1}^{r} \sigma_i u_i v_i^T \quad (6)$$

where $r < p$ is the effective intrinsic dimension of X.

In the next section, we introduce the complete-sets of OD flow timeseries from both networks that we have collected. In the section that follows it (Section 4), we analyze the flows using PCA.

3 Data

3.1 Networks Studied

This analysis of OD-pair flow properties is based on measurements from two different backbone networks. However, it is not specific to backbone networks and can be applied to different types of networks.

Sprint-Europe (henceforth Sprint) is the European backbone of a US tier-1 ISP. This network has 13 Points of presence (PoPs) and carries commercial traffic for large customers (companies, local ISPs, etc.). Abilene is the Internet2 backbone network. It has 11 PoPs and spans the continental USA. The traffic on Abilene is non-commercial, arising mainly from major universities in the US. The networks also have very different topologies.[3]

3.2 Flow Data Collected

Measuring flow data by capturing every packet at high packet rates can overwhelm available processing power. Therefore, we collected sampled flow data from every router in both networks. On the Sprint network, we used NetFlow [5] to collect every 250th packet. Sampling is periodic, and results are aggregated in flows at the network prefix level, every 5 minutes. On Abilene, the sampling rate is random, capturing 1% of all packets using Juniper's Traffic Sampling [27]. The monitored flow granularity is at the 5-tuple level (IP address and port number for both source and destination, along with protocol type) and sampled measurements are reported every minute. We aggregated the Sprint and Abilene flow traffic counts into bins of size 10 minutes and 5 minutes respectively to avoid possible collection synchronization issues.

Using sampled flow data has two major drawbacks. First, when a link is lightly utilized, sampling every N-th packet undersamples some flows. However, we found excellent agreement (within 1%-5% accuracy) between sampled flow bytecounts, adjusted for sampling rate, and the corresponding SNMP bytecounts on links with utlization more than 1 Mbps. Most of the links from both networks fall in this category, and so our sampled flow bytecounts are likely to be accurate. Another problem with measuring flows by sampling packets on any link is that some flows are not sampled altogether. As [8, 10] show, these unsampled flows have a small number of packets, carry very few bytes and so will have negligible impact on our aggregated flow bytecounts.

[3]The Abilene topology can be viewed at www.abilene.iu.edu/images/Ab-IGP-topo.jpg. The Sprint topology can be found at [15].

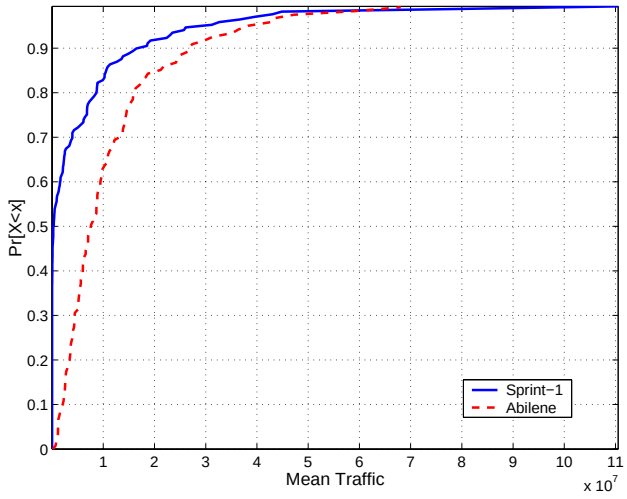


Figure 3: CDF of traffic exchanged between OD pairs.

3.3 From Raw Flows to OD Flows

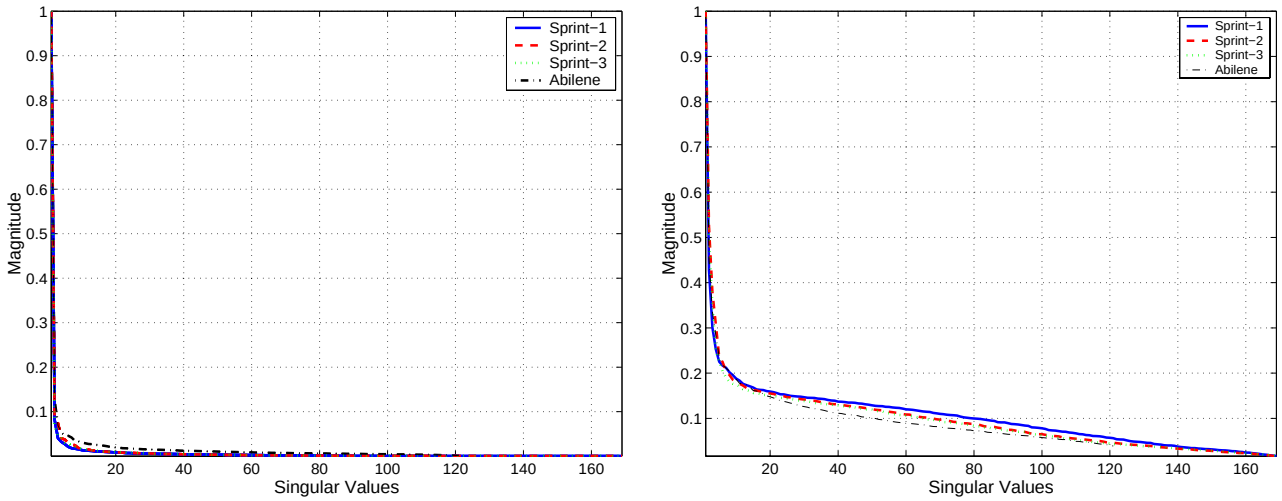
To obtain Origin-Destination flows from the raw flows collected, we have to identify the ingress and egress points of each flow. The ingress points can be identified because we collect data from *each* ingress link in both networks. For egress point resolution, we use BGP and ISIS routing tables as detailed in [9].[4] Using this procedure, we obtained three weeks of complete OD flow data from Sprint and one week from Abilene. Table 1 summarizes our datasets.

	# Pairs	Type	Time Bin	Period
Sprint-1	169	Net. Prefix	10 min	Jul 07-Jul 13
Sprint-2	169	Net. Prefix	10 min	Aug 04-Aug 10
Sprint-3	169	Net. Prefix	10 min	Aug 11-Aug 17
Abilene	121	IP 5-Tuple	5 min	Apr 07-Apr 13

Table 1: Summary of datasets studied.

A preliminary inspection of the amount of traffic exchanged between OD pairs is presented as a CDF in Figure 3. There is a large magnitude of variation in size, a phenomenon that has been termed ‘elephants and mice’ in other analyses of network traffic. Similar findings were noted for OD pair traffic in the AT&T network by [31].

[4]For Sprint, we supplemented routing tables with router configuration files to resolve customer IP address spaces. Also, Abilene anonymizes the last 11 bits of the destination IP address. This is not a significant concern because there are few prefixes less than 11 bits in the Abilene routing tables, and we found very little traffic destined to these prefixes.



(a) OD Flows (b) Normalized OD Flows

Figure 4: Scree plot for OD flows.

4 Analyzing OD Flows

As described in Section 2, the foundation of our approach is to use PCA to decompose an ensemble of OD flows into its constituent set of eigenflows. In this section, we present the results of that process. We first show that only a small set of eigenflows is necessary for reasonably accurate construction of OD traffic – meaning that OD flows in fact form a multivariate timeseries of low effective dimension. Then we examine the structure of OD flows, that is, how each OD flow is decomposed into constituent eigenflows.

4.1 Low Dimensionality of OD Flows

As described in Section 2.2, the energy contributed by each eigenflow to aggregate network traffic is summarized in the scree plot. We form scree plots by applying PCA to the Sprint and Abilene datasets. In Figure 4(a) we show the scree plots for each dataset.

The figure shows the surprising result that the vast majority of traffic variability is contributed by the first few eigenflows; furthermore, this effect is consistent in both networks. Both curves have a very sharp knee, showing that a handful of eigenflows, between 5 and 10, are much more significant in terms of contribution to traffic variability than are the rest. In different terms, this result shows that the OD flow timeseries together form a structure with effective dimension between 5 and 10 – much lower than the number of flows (over 100 in each case).

As an illustration of this low dimensionality of OD flows, we plot a sample of OD flows using a low-dimensional reconstruction. We do so by representing each OD flow using only the first five eigenflows. This construction is given by Equation 6, with $r = 5$. The results are shown in Figure 5. The figure shows that even if we omit over 100 dimensions from the original data, we can capture the temporal characteristics of these OD flows remarkably well.

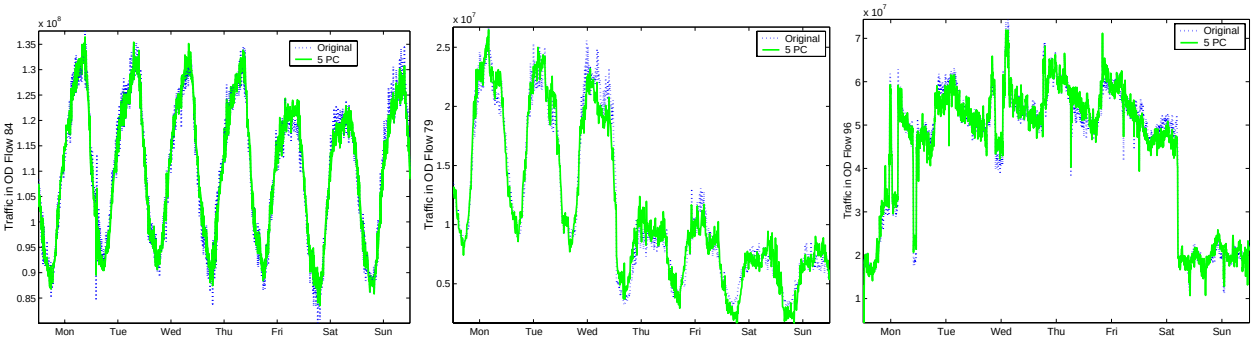


Figure 5: Reconstructing OD flow timeseries with 5 principal components (left and center plots: Sprint-1; right plot: Abilene).

Motivated by Figure 4(a), examination of the data indicates that the top 50 eigenflows in the Sprint data, and the top 90 eigenflows in the Abilene data, contribute 99.9% of the total energy to their respective datasets. All the remaining eigenflows have an insignificant effect. Thus, for simplicity in the rest of the paper, we will often concentrate our analysis only on these top eigenflows.

What is the reason for this low dimensionality in OD flow data? There are at least two ways in which this sort of low-dimensionality can arise. First, if the magnitude of variation among dimensions in the original data differs greatly, then the data may have low effective dimension for that reason alone. This is the case if variation along a small set of dimensions in the original data is dominant. Second, a multivariate timeseries may exhibit low dimensionality if there are common underlying patterns or trends across dimensions – in other words, if dimensions show non-negligible correlation.

We can distinguish these cases in OD flow analysis by normalizing the OD flows before performing PCA. The standard approach is to normalize each dimension to zero mean and unit variance. For OD flow data we have:

$$\bar{X}_i = \frac{(X_i - \mu_i)}{\sqrt{\text{Var}(X_i)}}, \quad i = 1, ..., p$$

where $\mu_i \equiv \mu(X_i)$ and $\text{Var}(X_i)$ are the sample mean and variance of X_i. If we find that OD flows still exhibit low dimensionality after normalization, we can infer that the remaining effect is due to correlation among flows.

The results of applying PCA to normalized versions of all datasets is shown in Figure 4(b). The most striking feature of this figure is that the sharp knee from Figure 4(a) remains, in the same location. It is also clear that the relative significance of the first few eigenflows has diminished somewhat. Taken together, these observations suggest that while differences in flow size contribute to the low-dimensionality of flows, that correlations among flows (common underlying flow patterns) play a significant role. As the discussion in Section 2.2 points out, these common underlying flow patterns are in fact the eigenflows.

The presence of mice and elephants in the original (non-normalized) flows, have many implications for traffic enginnering [18]. We therefore focus all subsequent analysis on these operationally

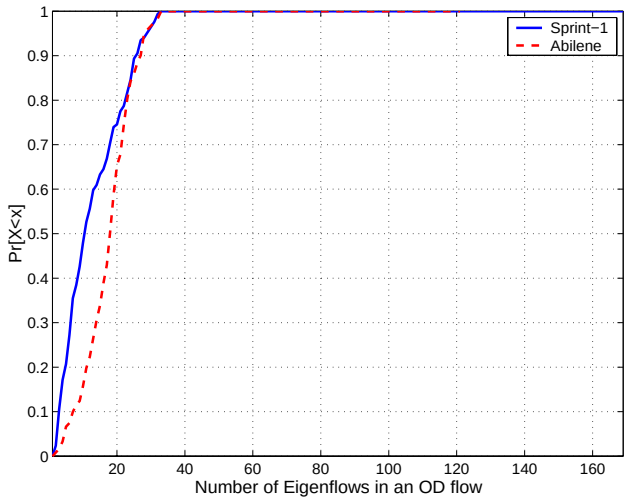


Figure 6: Number of eigenflows that constitute each OD flow (CDF).

more relevant non-normalized flows.

4.2 Structure of OD Flows

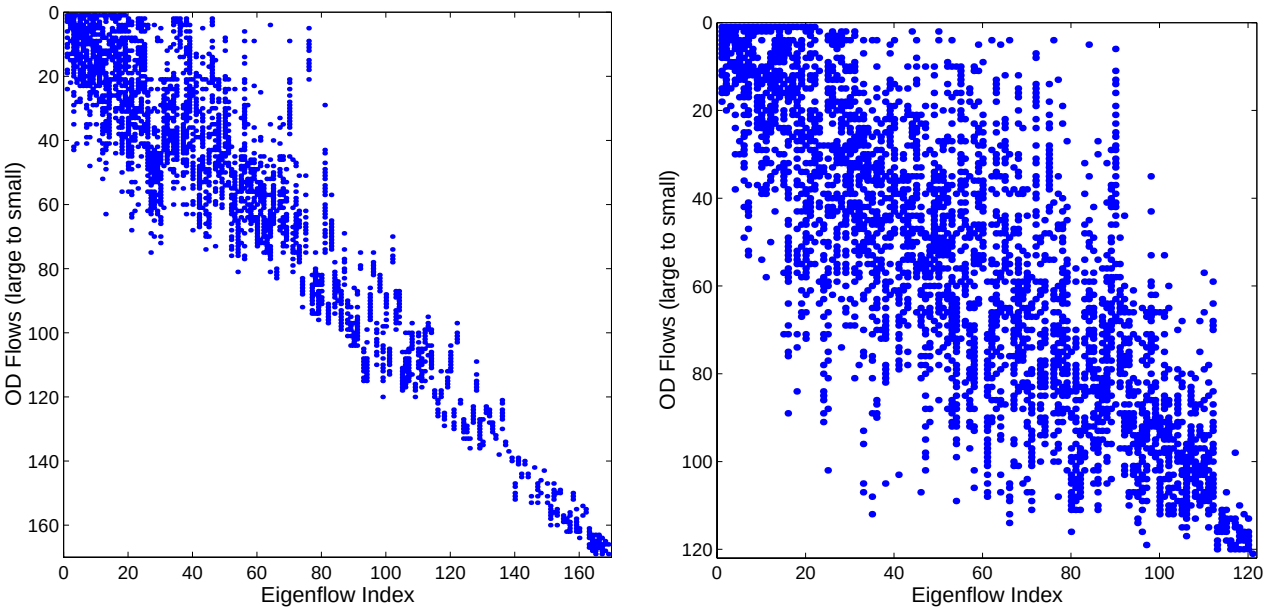
To understand how eigenflows contribute common patterns of variability across OD flows, we return to the discussion of PCA from Section 2.2. A row i of the principal matrix V specifies the extent to which each eigenflow (scaled by its corresponding singular value) contributes to OD flow i. This is summarized in Equation 4. Thus we can examine rows of V to discern the *structure* of the set of OD flows – how each OD flow is composed of eigenflows, and how any two OD flows are similar or dissimilar expressed in terms of eigenflows.

Inspecting the rows of V for a number of our datasets yields some surprising observations about how OD flows are structured in terms of eigenflows.[5] Our first observation is that the each OD flow is comprised of only a handful of significant eigenflows. We demonstate this as follows.

Considering any given row of V, we are interested in how many entries are significantly different from zero. We can make this precise by setting a threshold and counting how many entries in the row exceed this threshold in absolute value. A resonable threshold is $1/\sqrt{p}$, since a perfectly equal mixture of all eigenflows would result in a row of V with all entries equal and, applying this reasoning across all rows simultaneously, the constraints that columns of V have unit norm must be enforced.

In Figure 6, we plot the CDF of the number of entries per row of V that exceed this threshold for our Sprint-1 and Abilene datasets. The figure shows that, regardless of dataset, most rows of V have less than 20 significant entries, and no row has more than 35 significant entries. In terms of OD flows, this means that any given OD flow is composed of no more than 35 significant eigenflows, and generally many fewer. This surprising result means that we can think of each OD

[5]Exhaustive presentation of such voluminous data is impractical in the current context, but the reader is invited to inspect [15] which displays all rows of V for both Sprint-1 and Abilene datasets.



(a) Sprint-1 (b) Abilene

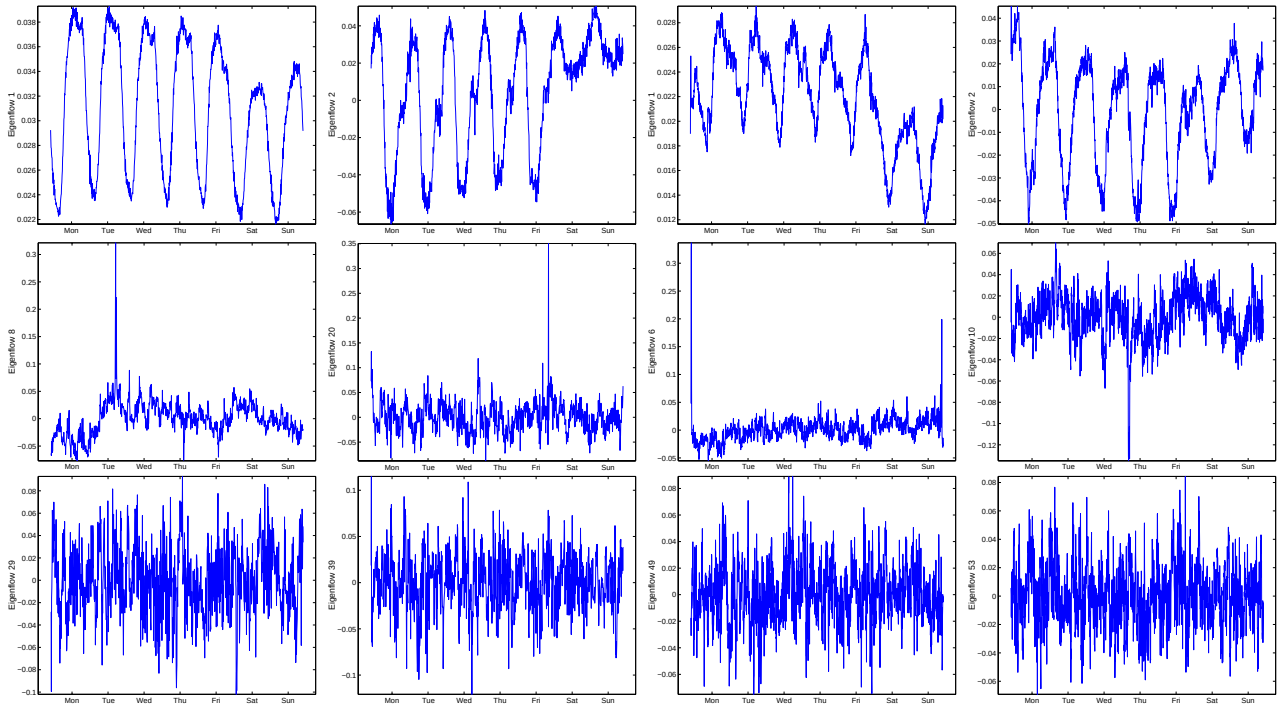
Figure 7: Indices of the eigenflows constituting each OD flow.

flow as having only a small set of "features." Thus, we should expect different OD flows to differ considerably in the nature of the temporal variation that they exhibit.

Our second observation concerns *how* OD flows differ. We note that, in general, there is a relationship between the size of an OD flow (its mean rate) and the eigenflows that comprise it. To examine this relationship, we can inspect where the above-threshold entries of the V matrix occur. Figure 7 shows the above-threshold entries of the V matrix for the Sprint-1 and Abilene datasets. In the figure, there is a dot for each entry in the V matrix that exceeds $1/\sqrt{p}$ in absolute value. Note that the columns of the V matrix are organized by convention in decreasing singular value order, and we have ordered the rows in order of decreasing OD flow rate as well. Thus the top row in each plot indicates the eigenflows that are significant in forming the strongest OD flow, and the bottom row indicates the significant eigenflows for the weakest OD flow.

The figure shows two things: first, in general, the significant entries in most rows of V are clustered in a restricted range (this effect is more pronounced in the Sprint data than in the Abilene data). Second, larger flows tend to be comprised mainly of the most significant eigenflows, and smaller flows tend to be comprised mainly of less significant eigenflows.

In some ways, the results shown in Figure 7 are not surprising. The largest OD flows will tend to dominate the definition of the most significant eigenflows, and so the steady downward trend in the plot is more or less to be expected. However the tight clustering of the significant eigenflows for any OD flow means that if there are qualitative differences between eigenflows in different ranges, then these qualitative differences will be reflected in the OD flows. Indeed, in the next section we show that this is in fact the case.



(a) Sprint-1 (b) Abilene

Figure 8: Eigenflow examples. Top Row: Deterministic Eigenflows; Middle Row: Spike Eigenflows; Bottom Row: Noise Eigenflows.

5 Understanding Eigenflows

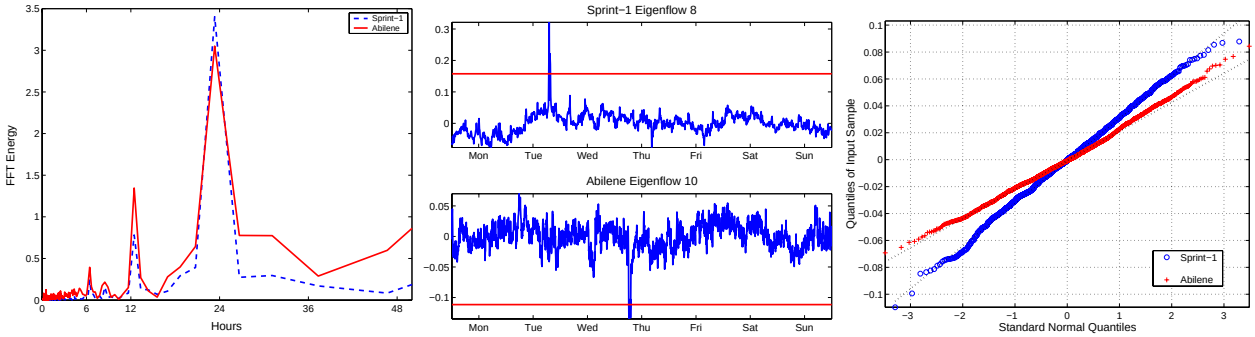
The analysis of OD flows presented in the last section has emphasized the central role of eigenflows in understanding OD flow properties. Thus we turn now to eigenflows; we inspect them, describe the three types most often seen, and show how understanding those types in light of the results in the previous section can yield general insight into OD flow properties.

5.1 A Taxonomy of Eigenflows

We start by inspecting the complete sets of eigenflows for a number of our datasets.[6] Surprisingly, across all of the eigenflows we examined, there appear to be only three distinctly different types. Representative examples of each eigenflow type from both the Sprint-1 and Abilene datasets are shown in Figure 8.

The top row shows examples of eigenflows that exhibit strong trends and periodicities. The periodicities clearly reflect diurnal activity, as well as the difference between weekday and weekend activity. Because these eigenflows appear to be relatively predictable, we refer to them as *d-eigenflows* (for "deterministic").

[6] As in Section 4.2, the raw data is too voluminous to present, but plots of the complete set of eigenflows for both datasets are available at [15].



(a) d-eigenflow (periodicity) (b) s-eigenflow (5σ threshold) (c) n-eigenflow (qq-plot)

Figure 9: Classifying eigenflows by using three tests.

The second row of Figure 8 shows examples of eigenflows that exhibit strong, short-lived excursions from the mean. These *s-eigenflows* (for "spike") show isolated values that can be many standard deviations (e.g., 4 or 5 standard deviations) from the eigenflow mean. These clearly capture the occasional traffic bursts and dips that are a common feature of network data traffic.

Finally, the lowest row of Figure 8 shows examples of eigenflows that appear roughly stationary and Gaussian. These *n-eigenflows* (for "noise") capture the remaining random variation that arises as the result of multiplexing many individual traffic sources. The majority of eigenflows in both datasets appear to be of this type.

These categories of eigenflows are only heuristically distinguished. It is not our intent to suggest that any eigenflow can be unambiguously categorized in this manner. Nonetheless, we observe that these categories are in fact very distinct, and that almost all eigenflows can be easily placed into one of these categories.

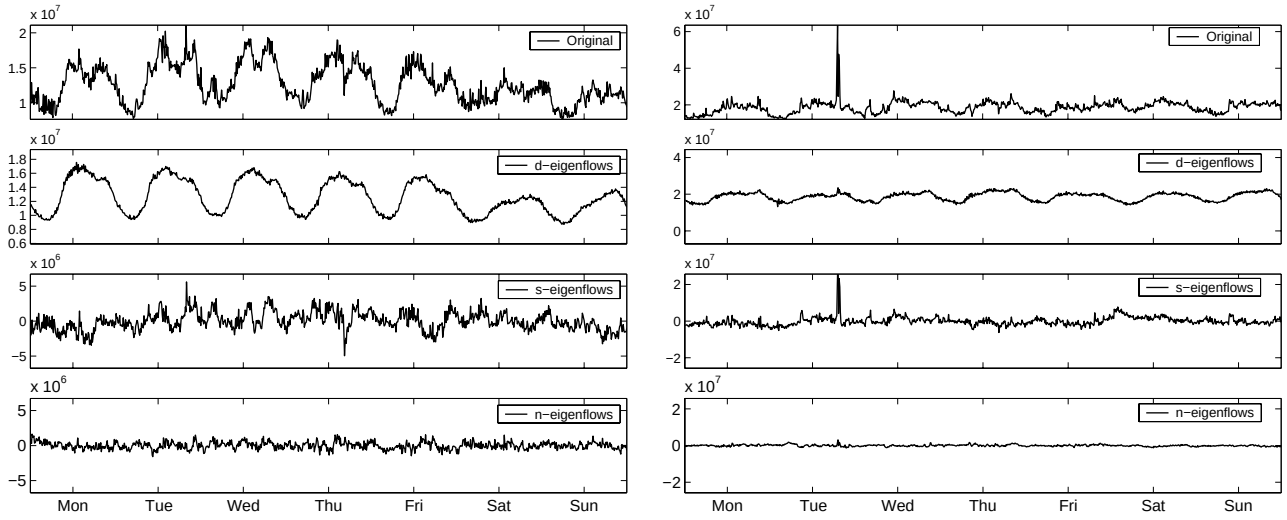
To demonstrate that these categories are distinct and that most eigenflows fall clearly into one of the three categories, we evaluate each flow according to the following criteria:

1. Does the eigenflow have a strong peak in its Fourier spectrum at 12 or 24 hours? A strong peak is defined here as a power value at that frequency greater than any other value in the power spectrum.

2. Does the eigenflow contain at least one outlier that exceeds 5 standard deviations from its mean?

3. Does the eigenflow have a marginal distribution that appears to be nearly Gaussian? We judge whether an eigenflow meets this criterion by examining its distribution on a qq-plot, which plots quantiles of the data against quantiles of the Normal distribution; a straight line indicates a close fit of the data to the Normal.

Examples of applying these criteria to eigenflows from both datasets are shown in Figure 9. Figure 9(a) shows that the eigenflows that we visually identify as d-eigenflows indeed have a distinct power spectrum peak at 24 hours. In Figure 9(b) we show visually identified s-eigenflows



(a) Sprint-1 OD Flow 153 (b) Sprint-1 OD Flow 148

Figure 10: Decomposition of OD flow timeseries into the sum of its three constituent eigenflows.

that have 5-sigma excursions from the mean. And in Figure 9(c) we show visually identified n-eigenflows that appear to have marginal distributions that are nearly Gaussian.

We used these tools to examine all eigenflows from both datasets.[7] We found that only a very small number of eigenflows had the property that more than one of the three questions above could be answered affirmatively. Eigenflows for which more than one of the criterion above held true were categorized as "indeterminate." In the Sprint-1 dataset, out of the 50 top eigenflows, only 4 were indeterminate (contributing 0.012% to overall energy); in the Abilene dataset, out of the top 90 eigenflows, only 2 were indeterminate (contributing 0.26% to overall energy). For all of the remaining eigenflows, one and only one criterion above held true.

Thus, by using the criteria above, we can (again, heuristically) place almost every eigenflow into one of the three categories. When we do so, we find that we can obtain considerable insight into the properties of the OD flows.

A clear benefit of this categorization is that it cleanly decomposes any given OD flow into its principal features. That is, we can reconstruct each OD flow in terms of three constituents: the contributions made by d-eigenflows, s-eigenflows, and n-eigenflows. When we do so, each constituent tends to capture a distinct feature of the OD flow: its (deterministic) mean, its sharp bursts away from the mean, and its apparently-stationary random variation. An example of this decomposition is shown in Figure 10. The figure shows the original flow along with its three constituent features as captured by its component eigenflows. The separation of bursts and random noise from the nonstationary variation of the mean is quite sharp. Furthermore, the isolation of bursts from background noise is also quite distinct. While a similar result could likely have been obtained through application of (probably sophisticated) timeseries models, we note that we have made no modeling assumptions here other than the simple categorization of eigenflows. Rather,

[7] Plots similar to Figure 9 for each eigenflow can be found at [15].

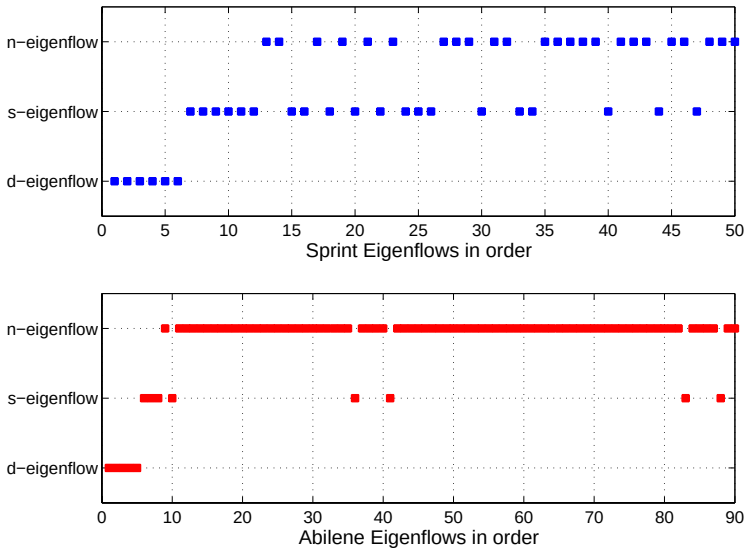


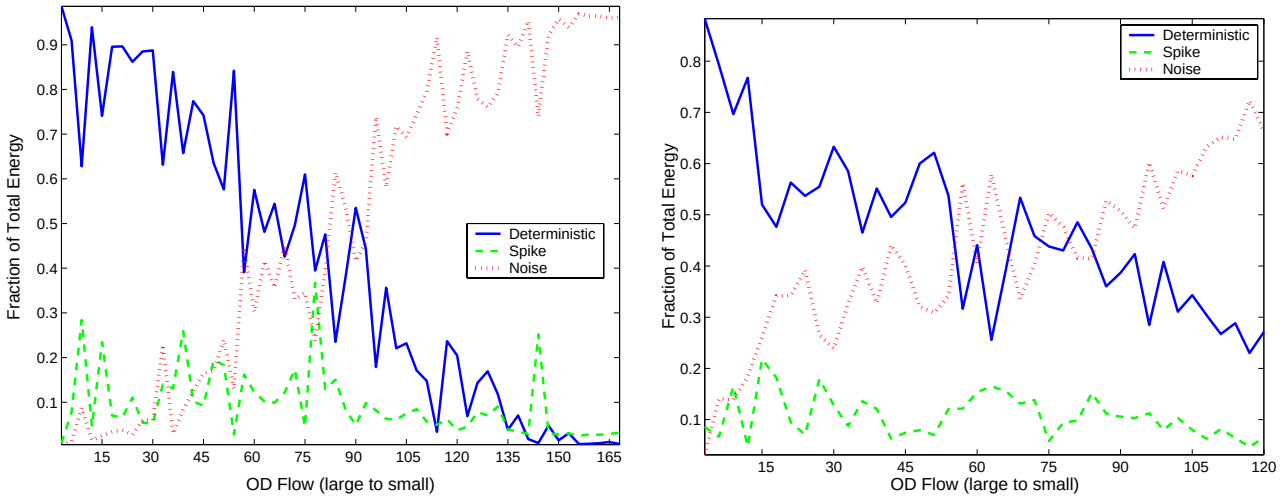
Figure 11: Occurance of eigenflow type in order of importance. Top: Sprint-1; Bottom: Abilene.

the power behind this method comes from the extraction of common variation patterns *across* OD flows as the information needed to identify and separate different kinds of variability *within* a single OD flow.

The features isolated in distinct eigenflows conform to characteristics that have been found in studies of other network traffic. Specifically, the presence of diurnal trends has been noted in [19, 23] for SNMP link data, the presence of stochastic bursts has been found in IP flow data by [24] and finally, the well-known fractional gaussian noise structure was first found in link level traffic by [16]. While previous studies have generally concentrated on identifying and describing these features from a model-based standpoint, this result shows that systematically isolating the common patterns of flow variability without recourse to elaborate modeling results in essentially the same set of features.

Given the apparent power deriving from categorizing eigenflows, it is worth investigating the relative role that the three types play in decomposing OD traffic. As a first step, we note that the different eigenflow types appear in different regions when the eigenflows are ordered by overall importance (*i.e.,* by singular value). To illustrate this effect, we show in Figure 11 the classification for each eigenflow in the Sprint-1 and Abilene datasets. The figure shows that in both datasets, d-eigenflows mainly appear as approximately the first six eigenflows. The next 5-6 eigenflows in order tend to be s-eigenflows, and the n-eigenflows tend to be the least significant. The only significant difference between the datasets is in nature of the least significant eigenflows (eigenflows numbered 12 and beyond): in Abilene, the least significant eigenflows are almost all of the noise type, while in Sprint-1 the least significant eigenflows are more evenly split between spike-type and noise-type. We leave an exploration of these differences for future work.

Figure 11 provides insight into the relative roles played by different sources of variability in our OD flow data. The figure shows that the most important source of variation is the nonstationary changes in the mean due to periodic trends. After these periodic trends, traffic bursts or spikes are



(a) Sprint-1 (b) Abilene

Figure 12: Fraction of total energy captured by eigenflow type for all OD flows.

next in importance. Finally, the least significant contribution to traffic variability in these datasets comes from noise. These conclusions are confirmed in a more quantitative way by the data in Table 2, which shows the fraction of total energy in each dataset that can be assigned to each of the three eigenflow types.

Eigenflow type	Sprint-1	Abilene
d-eigenflow	92.17%	69.79%
s-eigenflow	5.59%	18.60%
n-eigenflow	2.24%	11.61%

Table 2: Contribution of eigenflow type to overall traffic.

5.2 Decomposing OD Flows

We can refine our understanding of the nature of variability in OD traffic by applying this decomposition at the level of OD flows themselves. In fact, we can use this decomposition to gain insight into how traffic features vary from one OD flow to the next.

To do so, we determine the relative contribution of each eigenflow type to each OD flow. The results are shown in Figure 12. In this figure, OD flows are ordered by mean rate, decreasing from left to right. For each flow we plot the fraction of its energy contributed by d-, s-, and n-eigenflows. (We have averaged adjacent values in this figure to improve legibility.)

The figure shows that the PCA based decomposition of OD flows exposes how the properties of OD flows vary. We can see that high-volume OD flows are dominated by periodic, deterministic

trends. As we move to the right of the figure, the relative contribution of deterministic components decreases; that is, the lower-volume an OD flow, the more it tends to be dominated by noise. Finally, the figure shows that regardless of the volume of the OD flow, a relatively constant proportion of its energy is due to traffic bursts. Thus, we can relate the statistical properties (temporal features) of an OD flow in a particularly simple way to the flow's overall traffic volume.

These results provide a powerful organizing tool for thinking about collections of OD flows. They draw attention to the significant statistical differences between high-volume and low-volume OD flows. They suggest that a simple model may not be appropriate for all OD flows across a network. And they allow researchers and engineers to relate the properties of OD flows to the nature of the source and destination user or customer populations, through those populations' influences on OD flow traffic volume.

6 Temporal Stability of Flow Structure

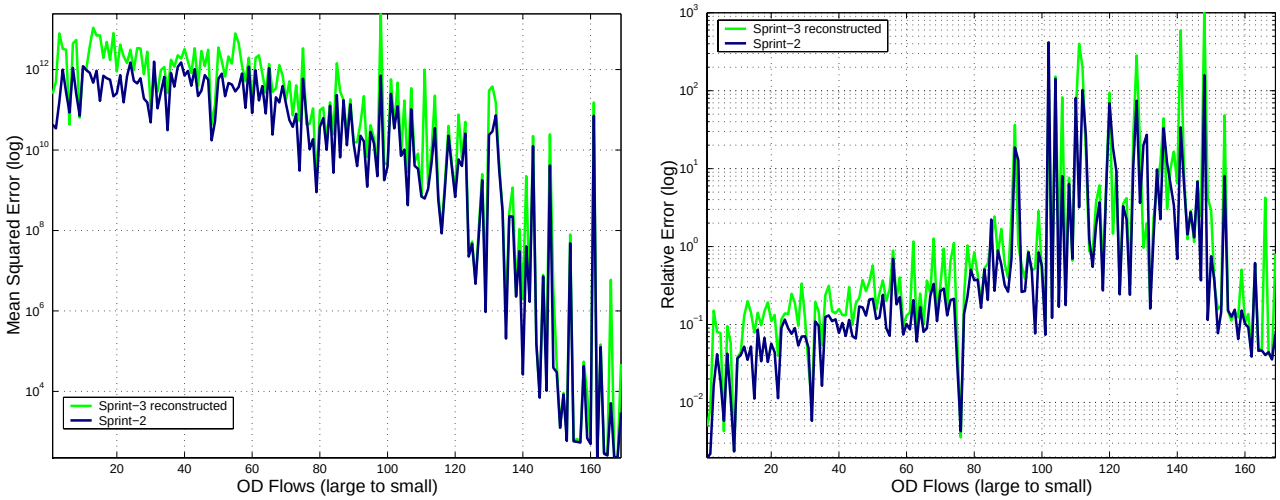
The previous sections have shown that PCA can unearth important structure in OD flow data. For many practical applications, it will be important to know the extent to which this structure varies in time.

The question we are concerned with in this section is whether the decomposition of OD flows into eigenflows, as determined by the set of pricipal components, is useful for analyzing data that was *not* part of the input to the PCA procedure. In general, we envision applications that may benefit from using PCA in an online manner as follows. Given OD flow data observed over some time period $[t_0, t_1)$, obtain the principal components $\{v_i\}$. Subsequently, at some time $t_2 > t_1$, use $\{v_i\}$ to decompose a new set of OD flow observations into eigenflows. Does the subsequent decomposition preserve useful properties of the eigenflows? We can ask two specific versions of this question: First, does the subsequent decomposition still have relatively low effective dimensionality? And second, if the original decomposition has categorized eigenflows by type, is that categorization still useful in the subsequent decomposition? Although space does not permit us to answer these questions thoroughly, we give some initial results here.

To answer the first question, we proceed as follows. One way to assess whether a set of OD flows has low effective dimension is to measure the error resulting from approximating the set of flows using a small number of dimensions. Using two consecutive weeks of OD flow data X_1 and X_2, we start by analyzing X_1 using PCA and obtaining its pricipal components $\{v_i\}$. We use $\{v_i\}$ to construct the top 20 eigenflows for X_1, and we *also* use $\{v_i\}$ in the same way to construct a corresponding set of 20 pseudo-eigenflows for X_2.[8] In each case, we form approximate versions of the original data using only the top 20 (pseudo-)eigenflows, yielding X_1' and X_2'. We then measure the per-flow sum of squared error of each approximation:

$$SSE_1 = ||X_1 - X_1'|| \quad \text{and} \quad SSE_2 = ||X_2 - X_2'||$$

[8] We use the term *pseudo-eigenflows* for the linear combinations of the OD flows of X_2 obtained using the $\{v_i\}$ of X_1, to remind us that they are not the result of applying PCA directly to X_2, but may still approximately have the desirable properties of the eigenflows of X_2.



(a) MSE (b) Relative Error

Figure 13: Exploring the temporal stability of Principal Components.

and the mean relative error of each approximation:

$$R_1 = avg(|X_1 - X_1'|/X_1) \quad \text{and} \quad R_2 = avg(|X_2 - X_2'|/X_2).$$

Based on the results in Section 4.1, we expect the error for X_1 to be small in general, because we know that OD flows can be accurately approximated using a small number of eigenflows. Furthermore, we expect the per-flow error for X_2 to be larger than the corresponding error for X_1, since the $\{v_i\}$ used in approximating X_2 were not necessarily optimal. However, what is not clear is how much worse the error will be for X_2 than for X_1.

We performed this analysis on datasets from the Sprint network, with X_1 consisting of data for the week of 04 August to 10 August (Sprint-2 dataset) and X_2 consisting of data for the next week, *i.e.*, 11 August to 17 August (Sprint-3 dataset). The results are shown in Figure 13. Figure 13(a) shows the sum of squared error per OD flow, with flows ordered by decreasing mean rate from left to right. Figure 13(b) shows the mean relative error per OD flow.

The plots show that overall, the error induced by using the previous week's principal components to analyze the current week's OD flows is not great. The relative approximation error for X_1 for the 20 or so heaviest (most important) flows is in the range of 5%. The relative approximation error for X_2, using the principal components of X_1, is in the range (for the same flows) of approximately 10%. Thus, the first week's principal components appear to remain good choices for forming a low-dimensional representation of the subsequent consecutive week.

The second question we ask is whether the categorization of eigenflows remains consistent enough from week to week to be useful. To answer this question we again decompose X_2 into pseudo-eigenflows, and we designate a pseudo-eigenflow a d-eigenflow if it was a d-eigenflow in the decomposition of X_1. This allows us to "detrend" X_2 without applying PCA to it directly. Detrending a particular set of flows is then accomplishing through a simple matrix multiplication.

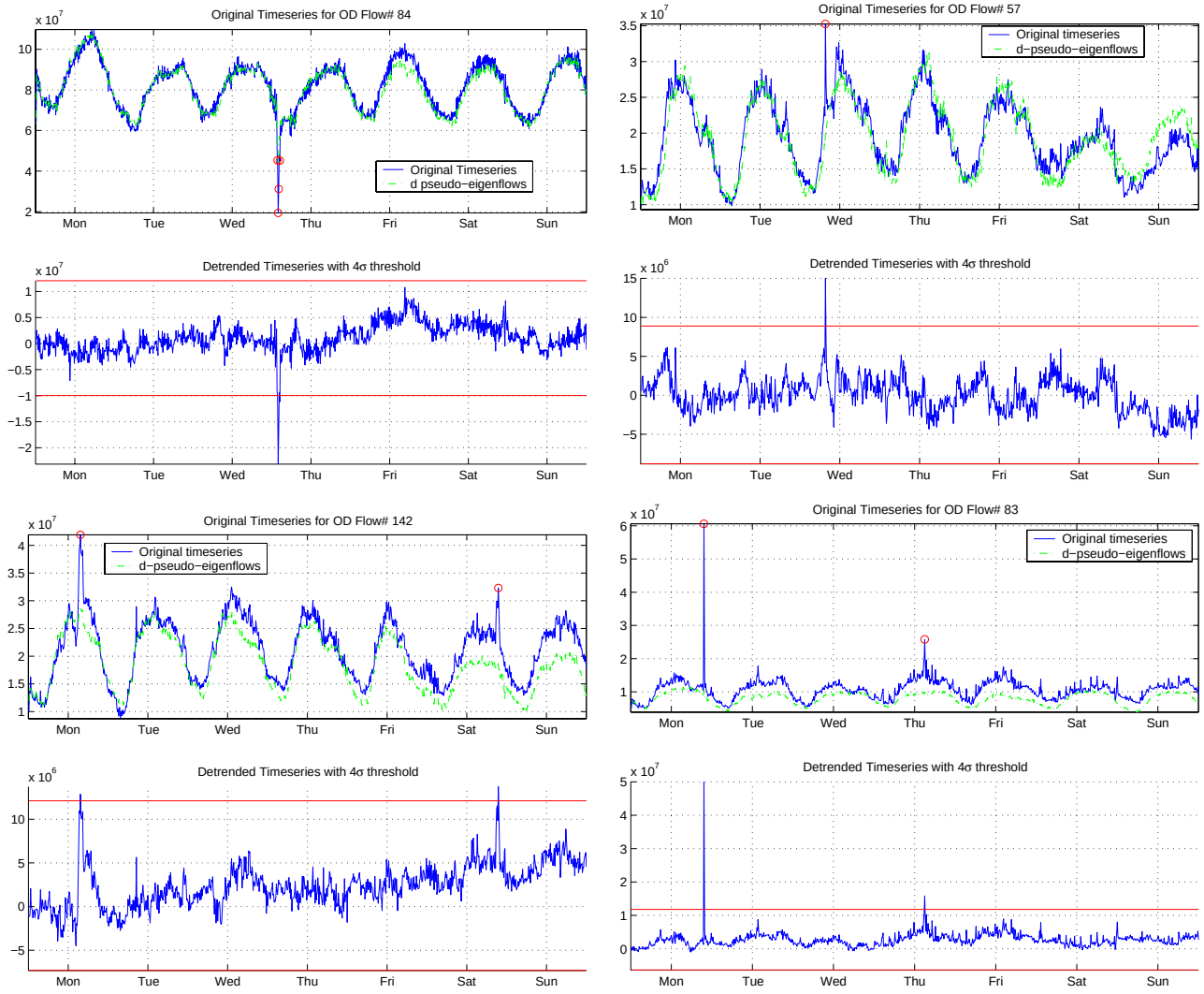


Figure 14: Examples of exploiting temporal stability to identify spikes.

To illustrate the effectiveness of this online style of detrending, we use it to identify unusual events in X_2. The approach is shown in Figure 14 on four example OD flows. The top and third rows shows the original timeseries of the OD flows taken from X_2, along with the corresponding d-pseudo-eigenflows. In the the second and bottom row, we show the same OD flow with deterministic components, as identified using the decomposition of X_1, removed. The result of removing the deterministic components appears to be a timeseries without much variation in mean, and therefore suitable for simple thresholding to identify unusual events. We adopt the arbitrary threshold of 4 standard deviations; based on this theshold, we show that unusual events (values far from the mean) can be easily identified in the original OD flow.

Taken together, these results suggest that the useful properties obtained from decomposition into eigenflows show a degree of stability from week to week that may be useful. While further

investigation is needed to determine the extent in time over which such properties are stable for any given application, we believe that the results shown here are promising.

7 Related Work

Although, to our knowledge, dimensional analysis using PCA has not been previously applied to network traffic measurements, it is a well-established tool for analyzing dimensionality and structure in other disciplines. Areas where it has been successfully employed in this way include face recognition [12], brain imaging [28], meteorology [21] and fluid dynamics [14].

Modeling traffic timeseries on a single link has attracted considerable research. Examples of recent studies that characterize timeseries of link traffic in backbone networks over long timescales are [22, 23].

In contrast, there is little prior work on OD flows, despite their engineering importance. Directly measuring OD flows requires additional and intensive monitoring on many routers, a task that demands considerable resources for high speed networks. Recently, however, network operators and researchers have started to use sampling schemes to measure OD flows [9]. It is the recent availability of such data that makes a study like ours now possible.

Two measurement studies that depart from the link-level traffic characterization and examine inter-PoP flows in a commercial Tier-1 backbone instead are [2, 9]. The authors of [2] observed many different types of OD flows, which behaved differently depending on link speed, type of relationship (peer or customer) and popularity. The implication is that it is difficult to devise a single model (or even a family of models) that characterizes a general PoP to PoP level flow.

Although there is little work that is closely related to ours, the work we report here has implications for a number of related networking problems. Our principal results (low dimensionality of OD flows, and differences in OD flow characteristics based on rate) can inform other work in a number of contexts. Here we briefly contrast our proposed approach with existing methods for a few candidate problems.

Traffic Matrix Estimation: The traffic matrix estimation problem, as originally formulated in [29], is an ill-posed linear inverse problem of the form $\mathbf{y} = \mathbf{A}\mathbf{x}$, where one seeks to estimate $\mathbf{x}$, the vector of OD flows, given $\mathbf{y}$, the vector of link traffic, and the routing matrix A (as defined in Section 2.1). The central difficulty of this problem stems from the fact that the apparent dimensionalilty of $\mathbf{x}$ is much larger than that of $\mathbf{y}$. Most of the methods proposed to date (e.g., [29, 4, 26, 17, 31, 32]) estimate $\mathbf{x}$ over hour-long stationary periods, when OD pairs are presumed to be independent. Our work demonstrates that on the timescales of days, which is the timescale of interest for many applications of traffic engineering, the effective dimensionality of OD flows is much smaller. In such scenarios therefore, the traffic matrix estimation problem may be more tractable and yield to direct solution methods.

Anomaly detection in timeseries: Anomalies in OD flow timeseries are difficult to identify without manual inspection. Simple thresholding schemes cannot be applied because the timeseries are nonstationary. A number of change detection methods have been proposed that rely on wavelet denoising techniques [1] and deviations from forecasted behavior [3, 13] to identify outliers. An

alternative approach is to detrend the flow timeseries using its d-eigenflows and then perform simple threshold tests on the resulting timeseries. The elements of this approach were briefly examined in Section 6 (Figure 14).

Traffic Forecasting: The state of the art in traffic forecasting for IP networks relies on forecasting models built on predictable trends of traffic, which are inturn isolated using wavelets [19]. An alternative approach to a wavelet-based isolation of trends in an OD flow is to simply use its d-eigenflows. Having done so, we can build forecasting models for the d-eigenflows and forecast the traffic for the entire set of OD flows. An advantage of such a PCA-based approach is that it allows simultaneous examination and forecasting of the entire ensemble of OD flow timeseries.

Traffic Engineering: The finding that large OD flows are mainly periodic and small OD flows are predominantly noise has been observed by others anecdotally [31]. Using PCA, we can systematically evaluate this effect with a fair amount of precision. Such an understanding of the structure of collections of OD flows has use in modeling and simulation studies.

An investigation of these problems constitutes our ongoing work.

8 Conclusions

In this paper, we have analyzed the structure of complete sets of Origin-Destination flow timeseries from two different networks: the European Sprint backbone network and the Abilene Internet2 backbone.

The first question we asked was whether complete sets of OD flows can be captured with low dimensional representations. Prior work suggested that because OD flows number on the order of hundreds in medium-sized networks and because each OD flow serves a different customer population, they are complicated structures to collectively model. Using Principal Component Analysis, we found that the hundreds of OD flows from both networks can be accurately described in time using 5-10 independent dimensions.

This surprising low dimensionality motivated us to ask a second question: how best can we understand the ways in which an ensemble of OD flows are similar and the ways in which they differ. We found that by examining the eigenflows, which are the common patterns of variation underlying OD flows, we could develop considerable understanding of the structure of OD flows. We found that the set of OD flows shows three features: deterministic trends, spikes and noise. Furthermore, the largest OD flows most strongly exhibit deterministic trends and the smallest OD flows are dominated by noise. Thus using PCA, we were able to quantitatively decompose the structure of each OD flow into its constitutent features.

Our last objective was to examine the extent to which the structure of OD flows unearthed by PCA varies over time. We found using the results of PCA of a previous week to decompose the structure of OD flows in the current week introduced very little error. Thus, the low-dimensional coordinate space formed by PCA shows some evidence of stability over time.

9 Acknowledgements

We are grateful to Rick Summerhill, Mark Fullmer (Internet 2), Matthew Davy (Indiana University) for helping us collect and understand the flow measurements from Abilene. We also thank Bjorn Carlsson, Jeff Loughridge (SprintLink) and Richard Gass (Sprint ATL) for instrumenting and collecting the Sprint NetFlow measurements. Finally, we thank Supratik Bhattacharyya (Sprint ATL) and Kavé Salamatian (LIP 6) for helpful discussions.

References

[1] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Internet Measurement Workshop*, 2002.

[2] S. Bhattacharyya, C. Diot, J. Jetcheva, and N. Taft. Pop-Level and Access-Link-Level Traffic Dynamics in a Tier-1 POP. In *Internet Measurement Workshop*, 2001.

[3] J. Brutlag. Aberrant behavior detection in timeseries for network monitoring. In *USENIX LISA*, 2000.

[4] J. Cao, D. Davis, S. V. Weil, and B. Yu. Time-Varying Network Tomography. *J. of the American Statistical Association*, 2000.

[5] NetFlow. At `www.cisco.com/warp/public/732/Tech/netflow/`.

[6] M. Crovella and E. Kolaczyk. Graph Wavelets for Spatial Traffic Analysis. In *IEEE INFOCOM*, 2003.

[7] D. Donoho. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. In *American Math. Society. Available at: `www-stat.stanford.edu/~donoho/Lectures/AMS2000/`*, 2000.

[8] N. Duffield, C. Lund, and M. Thorup. Estimating Flow Distributions from Sampled Flow Statistics. In *ACM SIGCOMM*, 2003.

[9] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True. Deriving traffic demands for operational IP networks: Methodology and experience. In *IEEE/ACM Transactions on Neworking*, 2001.

[10] N. Hohn and D. Veitch. Inverting Sampled Traffic. In *Internet Measurement Conference*, 2003.

[11] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.*, pages 417–441, 1933.

[12] M. Kirby and L. Sirovich. Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1990.

[13] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch-based Change Detection: Methods, Evaluation, and Applications. In *Internet Measurement Conference*, 2003.

[14] L. Sirovich and K. S. Ball and L. R. Keefe. Plane Waves and Structures in Turbulent Channel Flow. *Phys. Fluids. A*, page 2217:2226, 1990.

[15] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. Analysis of OD Flows (Raw Data). Technical Report BUCS-2003-021, Boston University, 2003.

[16] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *Transactions on Networking*, pages 1–15, Feburary 1994.

[17] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic Matrix Estimation: Existing Techniques and New Directions. In *ACM SIGCOMM*, 2002.

[18] K. Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian, and C. Diot. A Pragmatic Definition of Elephants in Internet Backbone Traffic. In *Internet Measurement Workshop*, 2002.

[19] K. Papagiannaki, N. Taft, Z. Zhang, and C. Diot. Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models. In *IEEE INFOCOM*, 2003.

[20] V. Paxson and S. Floyd. Wide Area Traffic: The Failure of Poisson Modeling. *Transactions on Networking*, pages 236–244, June 1995.

[21] R. W. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.

[22] M. Roughan and J. Gottlieb. Large scale measurement and modeling of backbone internet traffic. In *SPIE ITCom*, 2002.

[23] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, and Y. Zhang. Experience in measuring backbone traffic variability: Models, metrics, measurements and meaning. In *International Teletraffic Conference (ITC-18)*, 2003.

[24] S. Sarvotham, R. Riedi, and R. Baraniuk. Network Traffic Analysis and Modeling at the Connection Level. In *Internet Measurement Workshop*, 2001.

[25] G. Strang. *Linear Algebra and its Applications*. Thomson Learning, 1988.

[26] C. Tebaldi and M. West. Bayesian Inference of Network Traffic Using Link Data. *J. of the American Statistical Association*, pages 557–573, June 1998.

[27] Traffic Sampling. At `www.juniper.net/techpubs/software/junos/junos60/swconfig60-policy/html/sampling-overview.html`.

[28] D. T'so, R. D. Frostig, E. E. Lieke, and A. Grinvald. Functional Organization of primate visual cortex revealed by high resolution optical imaging. *Science*, page 417:420, 1990.

[29] Y. Vardi. Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *J. of the American Statistical Association*, pages 365–377, 1996.

[30] V. Yegneswaran, P. Barford, and J. Ullrich. Internet Intrusions: Global Characteristics and Prevalence. In *ACM SIGMETRICS*, 2003.

[31] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Loads. In *ACM SIGMETRICS*, 2003.

[32] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An Information-Theoretic Approach to Traffic Matrix Estimation. In *ACM SIGCOMM*, 2003.