A Cross-Correlation Based Method for Spatial-Temporal Traffic Analysis

Jian Yuan [a] Kevin Mills [b]

[a] *Tsinghua University Beijing, China*

[b] *National Institute of Standards and Technology USA*

Abstract

Analyzing spatial-temporal characteristics of traffic in large-scale networks requires both a suitable analysis method and a means to reduce the amount of data that must be collected. Of particular interest would be techniques that reduce the amount of data needed, while simultaneously retaining the ability to monitor spatial-temporal behavior network-wide. In this paper, we propose such a method, motivated by insights about network dynamics at the macroscopic level. We define a weight vector to build up information about the influence of local behavior over the whole network. By taking advantage of increased correlations arising in large networks, this method might require only a few observation points to capture shifting network-wide patterns over time. This paper explains the principles underlying our proposed method, and describes the associated analytical process.

Key words: network traffic, timescale, cross-correlation, spatial-temporal pattern, eigenvalue, eigenvector

1 Introduction

Most extant research on network traffic analysis focuses on observing *temporal* dynamics of traffic and effects from user and protocol behavior [1–4]. In such analyses, detailed Internet Protocol (IP) packet traces on individual links reveal the characteristics of network traffic at multiple timescales, e.g., rich scaling dynamics arising over small timescales [3], and self-similarity and long-range dependence at large timescales [4]. Recently, graph wavelets have been

Email addresses: jyuan@tsinghua.edu.cn (Jian Yuan), kmills@nist.gov (Kevin Mills).

proposed for *spatial* traffic analysis with knowledge of aggregate traffic measurements over all links [5]. This method can provide a highly summarized view of traffic load throughout an entire network. Despite these advances, spatial and temporal traffic analysis still presents difficult challenges, not only because large-scale distributed networks exhibit high-dimensional traffic data, but also because current analytical methods require examination of large amounts of data, which can strain memory and computation resources in even the most advanced generation of desktop computers.

Despite these inherent difficulties, investigation of spatial-temporal dynamics in large-scale networks is an important problem because modern society grows increasingly reliant on the Internet, a network of global reach that supports many services and clients. Lacking means to predict, monitor, and adjust spatial-temporal dynamics, Internet Service Providers (ISPs) typically over-provision network capacity, which leads to under-utilized resources on average with overloaded hotspots arising from time to time. Further, the Internet appears increasingly vulnerable to attacks and failures [6,7]. These factors suggest a crucial requirement to devise and develop promising tools that can monitor network traffic in space and time to identify shifting traffic patterns. Such tools can aid in operating and engineering large-scale networks, such as the Internet. While useful network management tools might focus on either offline or online monitoring and analysis, the task of network-wide on-line monitoring presents more stringent requirements for transferring and handling traffic data in a timely fashion.

To support the development of useful network management tools, the networking research community endeavors to devise novel and accurate methods to interpret measurements, and to derive principles for extracting information from raw measurement data. For example, a recent work studies correlations between different network flows in a French scientific network, *Renater* [8]. The study defines a network flow as a packet flow transferred from a given starting router to a given destination router. Many such flows simultaneously transit a large-scale network, leading to underlying *interactions* among the flows. Unfortunately, the effects of such interactions are usually not known, and so cannot contribute to better network engineering and management. The *Renater* study uses methods from random matrix theory (RMT) to analyze cross-correlations between network flows. (RMT methods have been recently used to study correlations in financial data [9].) In essence, RMT compares a random correlation matrix—a correlation matrix constructed from mutually uncorrelated time series—against a correlation matrix for the data under investigation. Deviations between properties of the cross-correlation matrix from the investigation data and the correlations in the random data convey information about "genuine" correlations. In the case of the *Renater* study, the most remarkable deviations arise about the largest eigenvalue and its corresponding eigenvector. The largest eigenvalue is approximately a hundred times larger than the maximum eigenvalue predicted for uncorrelated time series. The largest eigenvalue appears to be associated with a strong correlation over the whole network. In addition, the eigenvector component distribution of the largest eigenvalue deviates significantly from the Gaussian distribution predicted by RMT. Further, the *Renater* study reveals that all components of the eigenvector corresponding to the largest eigenvalue are positive, which implies their collective contribution to the strong correlation. Since all network flows contribute to the eigenvector, the eigenvector can be viewed as the signature of a collective behavior for which all flows are correlated. Thus, the eigenvector might provide an important clue about macroscopic behavior of the underlying interactions. In other words, the predominant information about network dynamics at the macroscopic level can be obtained from the largest eigenvalue and its corresponding eigenvector. This insight might prove very helpful for analyzing spatial-temporal traffic patterns in large-scale networks.

In this paper, we propose a method for spatial-temporal traffic analysis using the eigenvector corresponding to the largest eigenvalue. As the macroscopic pattern emerges from all adaptive behaviors of flows in various directions, hotspots should be exposed, through their correlation information, as the joining points of significantly correlated flows. Note that the details of the components of the eigenvector of the largest eigenvalue reveal this information, with the larger components corresponding to the more correlated flows. Thus, our primary insight is to group eigenvector components corresponding to a destination routing domain (or autonomous system) together to build up information about the influence of the routing domain over the whole network. We define a weight vector for this purpose. Contrasting weights against each other in the weight vector, we not only can summarize a network-wide view of traffic load, but also locate hot spots, and even observe how spatial traffic patterns change from one time period to the next.

While our approach builds upon the *Renater* study, we must solve some special problems related to scale. The *Renater* study assumes complete information from all network connection points, which proves feasible because the *Renater* network contains only about 30 interconnected routers. Arranging for complete coverage of observations in larger networks raises issues of scale, both in gathering data from numerous measurement points and in consuming computation time and memory when analyzing data. In particular, some heavily utilized routers may fail to collect and transfer measurement data. Usually, it is impossible to monitor areas of interest without corresponding measurements from those areas. To extend our ability to monitor network-wide behavior, we exploit correlation increases arising from *collective response* of the entire network to changes in traffic. This effect has already been observed in the framework of stock correlations, where cross-correlations become more pronounced during volatile periods as compared to calm periods [9]. Indeed, higher values of the largest eigenvalue occur during periods of high market volatility, which suggests strong collective behavior accompanies high volatility. This connection should have value in our analysis because Internet traffic behavior appears to be nonstationary [10]. An increase in cross-correlation allows us to infer a shift in the spatial-temporal traffic pattern of large areas of interest outside those few areas where measurements are made. This approach could significantly reduce requirements for data, perhaps to the point where monitoring may be performed in real time.

In this paper, we use simulation results to show how our proposed technique might work in a real large-scale network. Our results derive from a simple simulation model we developed recently to study space-time characteristics of congestion in large networks, and to analyze system behavior as a coherent whole [11]. While capturing essential time details of individual packets and connections, the model accommodates spatial correlations arising from interactions among adaptive transport connections and from variations in user demands. Though simulating an abstract network, which exhibits a regular structure and homogeneous behavior, our model offers a clear-cut framework to analyze spatial-temporal traffic patterns, e.g., where will hotspots develop and how long will they persist? Coupling our new measurement and analysis technique with our existing abstract simulation model allows us to compare weight vectors at different timescales. Using this approach, we explain the timescale of interest, and show macroscopic patterns at that timescale, allowing us to observe that network-wide hotspots become more prominent as increased correlation emerges. First, we try our method assuming complete measurement data, and then we further try our method with only a few observation points. The rest of this paper is structured as four sections. Section 2 describes our adaptation of the RMT cross-correlation method. In Section 3, we present our simulation model and discuss experiment results. We remark about future work in Section 4, before concluding in Section 5.

2 The Cross-correlation Based Method

In this section, we first discuss some important aspects associated with the *Renater* study, and then outline the cross-correlation based analysis method that we derived from the study. We describe how we represent network flow data and how we apply cross-correlation analysis to the data. Then, we explain our application of RMT (random matrix theory) to investigate cross-correlation throughout a network.

2.1 The Renater study

The French network *Renater* [1] comprises a nation-wide infrastructure to enable most French research, technological, educational, and cultural institutions to communicate with each other, and to connect to the global Internet. *Renater* has about 2 million users, supported by about 30 interconnected routers. Barthelemy and colleagues [8] studied traffic characteristics based on data collected from 26 of 30 *Renater* routers. The collected data consisted of traffic flows exchanged among routers for every sampling interval, $\tau = 5$ minutes, during a two-week period. The measured data encompass a total of $N_q = 26 \times 25 = 650$ different connections (i.e., source-destination pairs). The study considered only data for daytime traffic, which covers $L_q = 12$ sample intervals $\times$ 10 hours $\times$ 14 days = 1680 time counts, and analyzed correlation matrices using Random-Matrix Theory (RMT).

RMT describes generic behavior of different classes of systems, while deviations from its universal predictions allow the identification of system-specific properties. To apply RMT, one compares a random correlation matrix—a correlation matrix constructed from mutually uncorrelated time series—against a correlation matrix for the data under investigation. Deviations between the statistical properties of the cross-correlation matrix from investigation data and the correlations in random data convey information about "genuine" correlations. One first computes the eigenvalues λ_k (k =1, 2, ..., N_q) and the eigenvalue distribution from the investigation data, and then compares the distribution against an analytical result predicted for a corresponding random correlation matrix. Eigenvalues and eigenvectors of random matrices exhibit known statistical properties [8,9]. Particularly, in the limit $N_q \to \infty$, $L_q \to \infty$, where $Q \equiv L_q/N_q (> 1)$ is fixed, the probability density function $P_{rm}(\lambda)$of engenvalues of a random correlation matrix is given by

$$P_{rm}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \tag{1}$$

for λ within the bounds $\lambda_- \leq \lambda \leq \lambda_+$, where λ_- and λ_+ are the minimum and maximum eigenvalues, respectively given by

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}. \tag{2}$$

In the *Renater* study, the most remarkable deviation from $P_{rm}(\lambda)$ arises about the largest eigenvalue, which is found to be approximately a hundred times

[1] For more details on this network, see the web page http://www.renater.fr, which can be translated from French to English using a web-based translation service.

larger than the maximum eigenvalue predicted for uncorrelated time series. This suggests that the largest eigenvalue appears to be associated with a strong correlation in the network.

Deviation in the largest eigenvalue implies that deviation should also be displayed in the statistics of the corresponding eigenvector components. The distribution of the components $\{w_l^k; l = 1, \ldots, N_q\}$ of eigenvector $\boldsymbol{w}^k$ of a random correlation matrix should conform to a Gaussian distribution with mean zero and unit variance,

$$\rho_{rm}(w) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2}. \quad (3)$$

The *Renater* study found that, while the eigenvectors corresponding to most eigenvalues follow the predictions of RMT, the eigenvector corresponding to the largest eigenvalue λ_1 deviates significantly from the predicted Gaussian distribution. In particular, its components are nonzero and positive, which indicates correlations throughout the whole network. Since all network flows contribute to the eigenvector, the eigenvector can be viewed as the signature of a collective behavior for which all flows are correlated.

While the *Renater* study gives some inspiring results on understanding collective behavior in network flows, further studies are needed into spatial-temporal characteristics at multiple timescales and in larger networks. In addition, other issues must be considered as data is collected in larger networks. First, we may need to collect finer-grain flows in order to explore characteristics at timescales smaller than 5 minutes, and to explain temporal dynamics arising from relationships between small-scale fluctuations and long-range dependence [4]. Second, we may need to analyze collective properties over different time periods (e.g., from hour to hour) to characterize fluctuations in cross-correlation. The data used in the *Renater* study is discontinuous in the time axis, including no data for nighttime traffic. Third, to apply RMT to networks larger than *Renater*, we have to face some special problems related to scale. The *Renater* study assumes complete information from all network connection points, which proves feasible because the *Renater* network contains only about 30 interconnected routers. Arranging for complete coverage of observations in larger networks raises issues of scale, both in gathering data from numerous measurement points and in consuming computation time and memory when analyzing data. Finally, we should further identify and exploit the practical implications arising from network-wide traffic studies in order to help improve network engineering and management. A wide range of statistical techniques [19] might be explored in an effort to address these pending issues. Inspired by the *Renater* study, we derived and investigated an analysis method based on cross-correlation and deviations from RMT predictions. Next, we explain our method.

2.2 Representing network flow data

Our method requires us to represent packets flowing between distinct source-destination pairs at each sampling interval. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^T$ denote the flow vector of corresponding packet counts among all N routing domains, observed in starting domains during a given time interval in a large network. (Here ^{T} indicates transpose.) Each element of this flow vector is itself a vector defining the number of packets flowing into the corresponding domain from each of the other (starting) domains in the network. The method to obtain all flow variables in this vector is to first enumerate all the destination domains and then the starting domains by 1 to N, and group these indices by routing domain: the domains sending to the first domain in the first block, $\boldsymbol{x}_1$, and those sending to the second domain in the second block, $\boldsymbol{x}_2$, and so forth. Thus, we form $\boldsymbol{x}$ with blocks in the order $\boldsymbol{x}_1 = (x_{21}, x_{31}, \ldots, x_{N1})^T$, $\boldsymbol{x}_2 = (x_{12}, x_{32}, \ldots, x_{N2})^T$, $\boldsymbol{x}_3 = (x_{13}, x_{23}, x_{43}, \ldots, x_{N3})^T, \ldots, \boldsymbol{x}_N = (x_{1N}, x_{2N}, \ldots, x_{(N-1)N})^T$, where x_{ij} $(i \neq j)$ represents packet flow from the ith domain to the jth domain. Each flow variable x_{ij} is normalized as f_{ij} by its mean m_{ij} and standard deviation σ_{ij},

$$f_{ij} = (x_{ij} - m_{ij})/\sigma_{ij}. \tag{4}$$

Then, the normalized flow vector $\boldsymbol{f}$, corresponding to $\boldsymbol{x}$, comprises N normalized subvectors, $\boldsymbol{f}_k$ (k =1, 2, …, N), where each subvector is formed from normalized flow variables f_{ik} ($i \neq k$ and $i \leq N$). If M is the number of observed samples over the observation period of $M \times T$, then $\boldsymbol{f}$ is a $N(N-1) \times M$ matrix.

2.3 Cross-correlation analysis

Cross-correlation analysis is a tool commonly used to analyze multiple time series. We can compute the equal-time cross-correlation matrix $\boldsymbol{C}$ with elements

$$C_{(ij)(kl)} = \langle f_{ij}(t) f_{kl}(t) \rangle \,, \tag{5}$$

which measures the correlation between f_{ij} and f_{kl}, where $\langle \cdots \rangle$ denotes a time average over the period studied. The cross-correlation matrix is real and symmetric, with each element falling between –1 and 1. Positive values indicate positive correlation, while negative values indicate an inverse correlation. A zero value denotes lack of correlation.

We can further analyze the correlation matrix $\boldsymbol{C}$ through eigenanalysis [12]. The equation

$$\boldsymbol{C}\boldsymbol{w} = \lambda \boldsymbol{w} \tag{6}$$

defines eigenvalues and eigenvectors, where λ is a scalar, called the eigenvalue. If $\boldsymbol{C}$ is a square K-by-K matrix, e.g., $K = N(N-1)$ in the case of complete coverage, then $\boldsymbol{w}$ is the eigenvector, a nonzero K by 1 vector (a column vector). Eigenvalues and eigenvectors always come in pairs that correspond to each other. This eigenvalue problem has K real eigenvalues, some of which may repeat. An eigenvector is a special kind of vector for the matrix it is associated with, because the action of the underlying operator represented by the matrix takes a particularly simple form on the eigenvector input: namely, simple rescaling by a real number multiple. The eigenvector $\boldsymbol{w}^1$ corresponding to the largest eigenvalue λ_1 often has special significance for many applications. There are various algorithms for the computation of eigenvalues and eigenvectors [12]. Here, we exploit the MATLAB **eig** command, which uses the QR algorithm to obtain solutions [13].

2.4 Defining the weight vector

Much of the traffic flowing through the Internet must traverse multiple routing domains. Adaptive behaviors of flows in different directions play a crucial role in forming macroscopic patterns, mostly in a self-organized manner. The cross-correlation matrix contains within itself information about underlying interactions among various flows. In a study of cross-correlations in stock price changes, influence strength is defined as the sum of the cross-correlation coefficients associated with one company [14]. In that study, influence strength is used to represent the degree to which changes in a company's stock price affect the entire stock market. Similarly, we can measure the congestion level of the jth domain by summing all cross-correlation coefficients (ignoring autocorrelation) associated with the jth block, i.e., $\sum_i \sum_{k,l} C_{(ij)(kl)}$, $(i \neq j, k \neq l)$. Using this approach in our simulations yielded findings similar to those reported for stock markets [9] and for the *Renater* network [8]. That is, the majority of the properties of the correlation matrix $\boldsymbol{C}$ conformed to the results predicted by RMT [2]; thus, the correlation coefficients included substantial noise mixed with the information about macroscopic patterns. We found this to hold even when observing network traffic flows in all nodes, and to hold more strongly in cases where we observed network traffic in only a sparse number of nodes.

[2] In the *Renater* study, the eigenvalues' distribution and their spacing distribution follow approximately the predictions of RMT. And, the eigenvectors corresponding to most eigenvalues are in agreement with the results of RMT.

From this, we infer that we are more likely to find less noise (and more information) in cases that deviate from the RMT predictions. Such cases can be found by filtering the information about structural correlations through eigenanalysis.

The components of the eigenvector $\boldsymbol{w}^1$ of the largest eigenvalue λ_1 represent the corresponding flows' influences on macroscopic behavior, abstracted from the matrix $\boldsymbol{C}$ into the pair $(\lambda_1, \boldsymbol{w}^1)$. The eigenvector $\boldsymbol{w}^1$ comprises N subvectors, i.e., $\boldsymbol{w}^1 = (\boldsymbol{w}^1{}_1, \boldsymbol{w}^1{}_2, \ldots \boldsymbol{w}^1{}_N)^T$. The kth subvector, corresponding to the kth domain, is formed from components $w^1{}_{ik}$ ($i \neq k$ and $i \leq N$) representing the ith domain's contribution to the kth domain. We consider the square of each component, $(w^1{}_{ik})^2$, instead of $w^1{}_{ik}$ itself because $\sum_{i,k}(w^1_{ik})^2 = 1$ [15]. We define the weight S_k ($k = 1, 2, \ldots, N$) to be the sum of all $(w^1{}_{ik})^2$ in the kth subvector $\boldsymbol{w}^1{}_k$.

$$S_k = \sum_{i(\neq k)}^{N} (w^1_{ik})^2. \tag{7}$$

In the case of complete observations in all routing domains, S_k represents the relative strength of the contributions of the flows towards the kth routing domain. Thus, the knowledge of weight vector $\boldsymbol{S} = (S_1, S_2, \ldots, S_N)$ across varying k constitutes one summary view of network-wide traffic load.

When analyzing the spatial-temporal traffic pattern of a large-scale network, the cross-correlation matrix $\boldsymbol{C}$ can be a very large object. Usually, floating-point operations on the order of K^3 are required to find eigenvalues and eigenvectors [12]. Thus, even if such analysis yields informative results, it appears impractical to monitor the spatial-temporal pattern of large-scale networks using this method with complete coverage of observations. We exploit the property of the increased correlation in order to reduce data requirements, filling the flow vector $\boldsymbol{x}$ just with traffic measured in a few domains. This insight might allow us to infer traffic-pattern shifts in real time for large areas of interest from observations in only a few distant locations.

3 Experimental Analysis

In this section, we show some experimental results after a brief description of our simulation model. Assuming complete coverage of observations, we first discuss the timescale of interest, and also consider qualitatively the increased correlation arising at that timescale. We then demonstrate our method applied in a larger (simulated) network structure. Subsequently, we consider our method with various reductions in the number of observation points, showing how increased correlation helps to reduce the scale of measurements necessary to capture shifting network-wide traffic patterns over time. As the number of observation points decreases, there comes inevitably a level where the performance of our method degrades. We also investigate the ability of our method to reveal network-wide behavior when we divide the network into sub-areas. By focusing separately on each sub-area and performing the necessary computations in parallel, we can reduce the overall time required for analysis. Finally, we compare computation requirements among our various experiments to show that reducing data set size might permit us to support real-time monitoring and analysis.

3.1 Simulation model

Network simulation plays a key role in building an understanding of network behavior. Choosing a proper level of abstraction for a model depends very much on the objective. Studying large-scale characteristics and collective phenomena seems to require simulating networks at large scale. Appropriate models for these purposes should also include substantial detail representing protocol mechanisms across several layers (e.g., application, transport, network, and link) of functionality, yet must be restricted in space and time in order to prove computationally tractable. We propose a modeling approach that maintains the individual identity of packets to produce the full-duplex "ripple effect" at the packet level, and that can also accommodate spatial correlations in a regular network structure. The regular and homogeneous topology of our network model and accompanying routing simplifications, exhibit significant deviation from real networks. Further, our model characterizes user behavior in a highly abstract form, and depicts only the most elemental details of transport algorithms. Despite these abstractions and simplifications, our model has been tested successfully against current understanding of the timescale dynamics of network traffic, and has been used to show a significant influence of spatial span on correlation structure [11,16]. These previous experiences indicate that our simulation model, while unlikely to yield quantitative fidelity with real networks, should prove suitable as a vehicle to test our proposed analysis method.

The topology of our model comprises a variable number of interconnected domains. Figure 1, for example, shows a network of 25 domains. Each domain has two tiers: an upper tier for routers and a lower tier for hosts. Each router is attached to an equal number of sources (100 in this paper), and to a variable number of hosts ($\leq$ 500 in this paper) acting as receivers. Each source models traffic generation as an ON/OFF process, which alternates between wake and sleep periods with average durations λ_{on} and λ_{off}, and with the same shape parameter α of the Pareto distribution [11] for both ON and

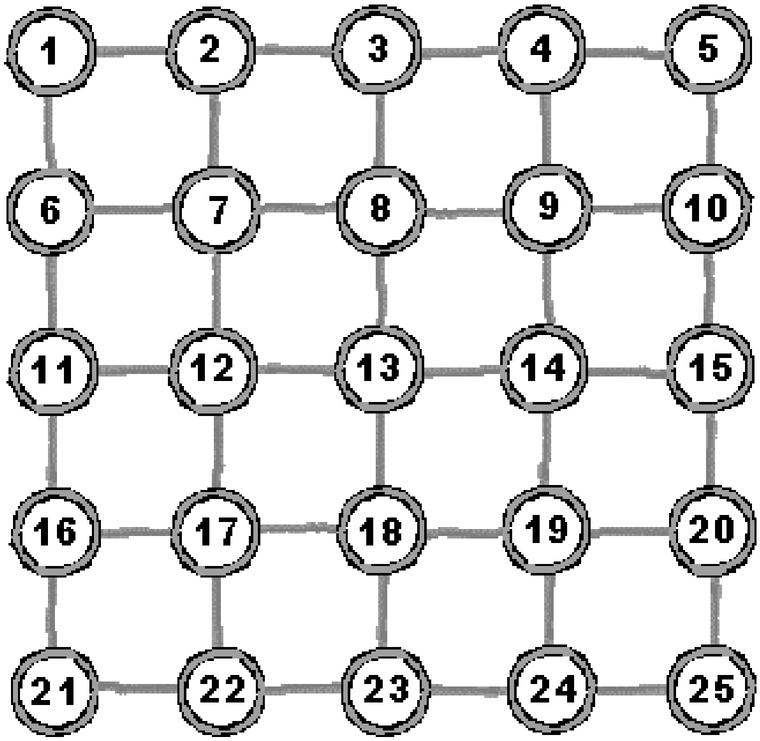

Fig. 1. The network structure with 25 routing domains.

OFF processes. This traffic-generation process mimics the heavy-tailed distribution of transferred file sizes observed from empirical measurements on the Internet [2]. When a source initiates a connection (ON period), a destination routing domain (differing from the source domain) is chosen randomly and uniformly. Our model generates and routes individual messages, called packets. To store and forward packets, which travel a constant, shortest path between a source-destination pair for each flow, routers maintain a queue of limited length (160 packets/router here), where arriving packets are stored until they can be processed: first-in, first-out. For convenience, in this paper we assume that every discrete simulation time-step is 1 millisecond. If a source is in an ON period, the source can create one packet every millisecond, subject to the control of TCP (Transmission-Control Protocol) constraints, and forward it to the buffer of its directly attached router. However, each router can forward multiple packets (10 here) during one millisecond. This simulates the difference between access links and backbone links in a hierarchically structured network.

With our model, we can simulate spatial and temporal traffic dynamics through high user variability ($\alpha = 1.5$, $\lambda_{on} = 50$ and $\lambda_{off} = 3000$), and through adaptive transport (TCP) connections. We first model a network with $N = 25$ domains (Figure 1). Note that there is no structural bottleneck in our model because routing assumes a periodic boundary condition, which allows the edges of our grid topology to form a closed structure [11]. Given homogeneous variation

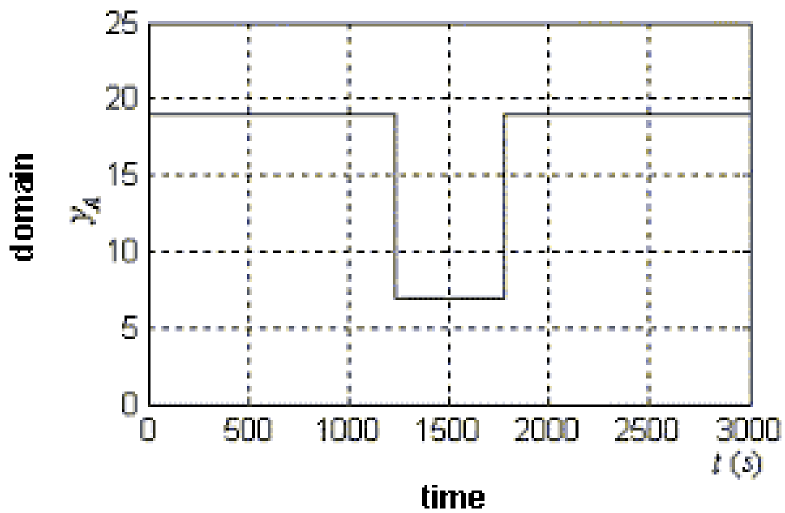

Fig. 2. The most congested domain changes over time.

of traffic demand in space and time, heavily congested subnets are induced only infrequently. To deliberately induce congestion, we let one domain have an additional two percent probability for selection as the destination domain. This is a natural way to change the network-wide traffic demand at longer timescale. We measure the most congested domain, i.e., the domain serving as destination for the greatest number of connections. Figure 2 shows the address, y_A $(= 1, 2, \ldots, 25)$, of the most congested domain changing over time. During the first period, the 19^{th} domain is the most congested. Then, at $t = 800$ s, the 7^{th} domain is selected as a new location to induce the next congestion, but the second period of congestion actually starts from $t = 1232.9$ s. At $t = 1600$ s, the 19^{th} domain is again selected as the hotspot, but the third period of congestion arises 542.2 s into the second period. This congestion-induction technique offers an easily interpreted framework to analyze spatial-temporal pattern shifts driven by varying traffic demand.

3.2 Timescale of interest

When focusing on network-wide behavior, the timescale of interest should not be fine-grained. The microscopic fluctuations observed at shorter timescales usually reflect local details, while the driving force of traffic demand seems to vary over much longer timescales. The timescale of interest in our experiments appears at a middle range, similar to the concept of a critical timescale beyond which traffic fluctuation is supposed to exhibit greater influence [17]. At this middle timescale, macroscopic behavior forms a connecting link between microscopic fluctuations and the longer-range driving force of variations in traffic demand. This expected coherence emerges as a result of adaptive behaviors of flows in different directions, but continues to shift its spatial-temporal pattern

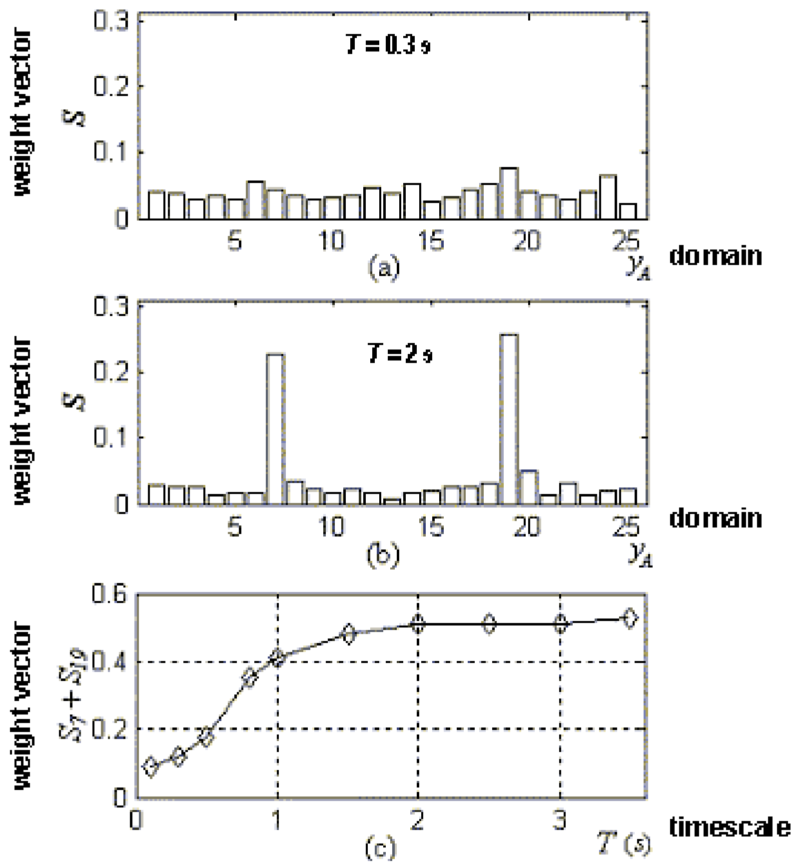

Fig. 3. Two weight vectors at $T = 0.3$ s (a) and $T = 2$ s (b), and (c) the sum of S_{19} and S_7 changing at different timescales.

under the force of traffic demand.

In our simulation, we observe at granularity of 100 ms (i.e., each 100 model time steps) every fine-grain flow between all domain pairs, filling the flow vector $\boldsymbol{x}$ with 600 variables (24 destination domains for each of the 25 source domains). Such complete coverage of observation allows us to analyze cross-correlations of all flows aggregated at various time granularities, denoted by T.

We first calculate the weight vector $\boldsymbol{S}$ with M data points ($M = 200$ in this paper), which span a first period ($M/2$ points) and a second period ($M/2$ points). We calculate two weight vectors at the aggregated levels $T = 0.3$ s and $T = 2$ s, shown respectively in Figure 3(a) and 3(b). The weight vector with $T = 2$ s shows two prominent weights at the 7^{th} and 19^{th} domains (S_7 and S_{19}), revealing the network-wide pattern of congestion arising in these two domains. However, the pattern does not appear when $T = 0.3$ s. To clarify the role of timescale here, we further show in Figure 3(c) the sum of S_7 and S_{19} at different aggregated levels. We find that the sum of S_7 and S_{19} gradually

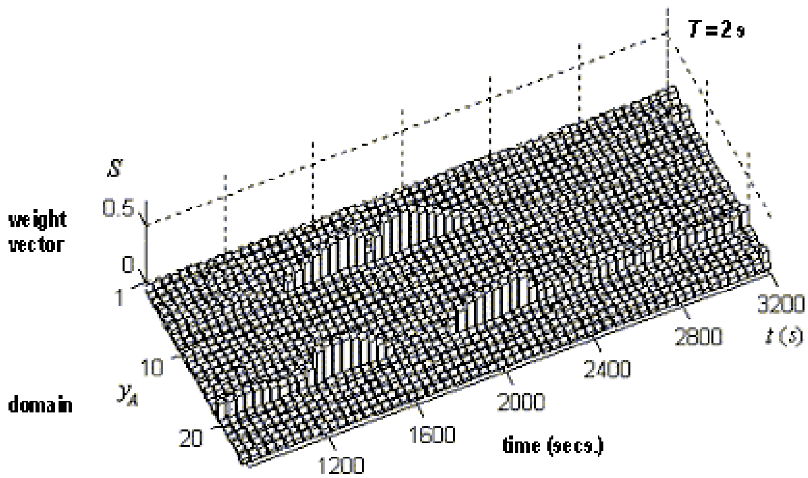

Fig. 4. The spatial-temporal pattern evolving with $T = 2$ s.

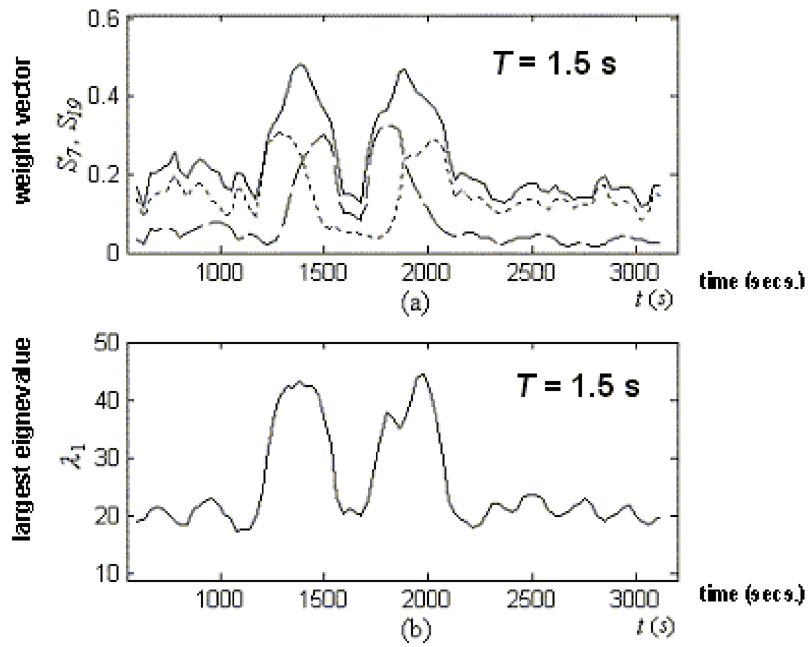

Fig. 5. (a) S_7 (dashed line), S_{19} (dotted line), and the sum of S_7 and S_{19} (solid line), and (b) the largest eigenvalue λ_1 with $T = 1.5$ s.

increases as T increases, but levels off from about $T = 2$ s. To show how the spatial traffic pattern changes, we calculate the weight vector $\boldsymbol{S}$ using M data points within a moving time window MT from one time period to the next. Figure 4 shows the weight vector $\boldsymbol{S}$ evolving with $T = 2$ s and with the time window MT ($= 200 \times 2$ s $= 400$ s) sliding ahead every 40 s. The time axis indicates the end of the moving time window. This technique provides a useful way to observe network-wide congestion patterns shifting over time.

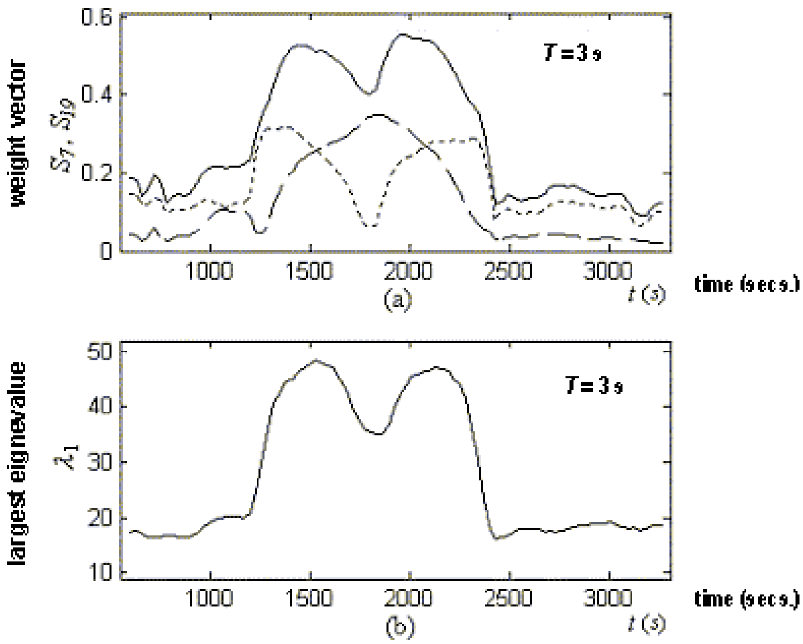

Fig. 6. (a) S_7 (dashed line), S_{19} (dotted line), and the sum of S_7 and S_{19} (solid line), and (b) the largest eigenvalue λ_1 with $T = 3$ s.

3.3 Increased correlation

Figure 5(a) shows S_7 (dashed line), S_{19} (dotted line), and the sum of S_7 and S_{19} (solid line), which are calculated with $T = 1.5$ s and with the time window MT (= 200 × 1.5 s = 300 s) sliding ahead every 30 s. Figure 5(b) shows the corresponding λ_1. While S_7 and S_{19} are distinguishable in three periods, both become enhanced during periods of pattern shifting. The sum of S_7 and S_{19}, and the largest eigenvalue λ_1 undulate in the same way, and reach higher values during periods of pattern shifting than during calm periods. The increased correlation in the simulation data suggests a *collective response* over the entire network to changes in traffic demand. During transient periods, flows in different directions have to adapt their behaviors to the changing impulse of the driving force, and continue to react to each other until they reach collectively a new coherent pattern. With the measurement and analysis method, as outlined above, applied at the appropriate timescale, as cross-correlations become more pronounced, traffic patterns over the whole system become more visible.

One might hypothesize that system-wide visibility depends on choosing an appropriate timescale. For example, observe the system at a coarser timescale of $T = 3$ s, as shown in Figure 6. We show S_7 (dashed line), S_{19} (dotted line), and the sum of S_7 and S_{19} (solid line) in Figure 6(a), and the corresponding λ_1 in Figure 6(b). As T increases, doubling from Figure 5 to Figure 6, we find that two transient processes seem to converge gradually, and that the second period (seen in Figure 5) becomes indistinct (in Figure 6), as if a hotspot appears in the 7^{th} domain for some time. When T is above 4 s (not shown), congestion in the 19^{th} domain never appears to diminish.

3.4 Sparse observation posts

While our proposed data analysis method provides substantial visibility into network-wide behavior at the critical timescale, it appears impractical to collect fine-grain traces for every source-destination pair over a large network. Even if complete observations could be arranged, challenges remain: such as, obtaining reliable data transfer to the analysis point and implementing processing power sufficient to analyze the data within a meaningful time. In particular, some heavily utilized routers may fail to collect and transfer data, but often happen to be the parts of interest to monitor (due to their congested nature). Given these real constraints, it would be appealing to reduce the amount of data to transfer and process, while retaining our ability to monitor network-wide behavior.

It could prove feasible to design sample-based techniques suitable to identify network-wide patterns that remain invariant for a long time. When traffic demands vary over a large dynamic space-time range, these same techniques might fail to detect more quickly changing patterns. However, by exploiting the increased correlation arising during volatile periods, we might be able to use a sample-based version of our proposed method to identify shifting network-wide congestion patterns. In the following, we provide some preliminary results regarding this idea.

Figure 7 shows a larger simulated network with 81 domains and L $(= 16)$ observation points (shaded). For each source, we use the following traffic-generation parameters: $\alpha = 1.5$, $\lambda_{on} = 50$ and $\lambda_{off} = 5000$. We record traffic flowing out from each observation point to all other domains with $T = 2.1$ s, and we fill the flow vector $\boldsymbol{x}$ with $L \times (N-1)$ $(= 16 \times 80 = 1280)$ variables, representing a substantial reduction from the 6480 variables that would be needed for complete monitoring. We select a total of four domains as hotspots, and increase congestion in two of the domains in each of two different time periods. Figure 8(a) shows how the most congested domains, y_A $(= 1, 2, \ldots, 81)$, change over time. In the first period (up to about 1830 s), we arrange for the 21^{st} and 61^{st} domains to be most congested. In the second period (after 1830 s), we arrange for the 25^{th} and 57^{th} domains to be most congested. We then calculate the weight vector $\boldsymbol{S}$ with 200 data points spanning the two periods. In Figure 8(b), the weight vector shows four prominent weights at the 21^{st}, 25^{th}, 57^{th} and 61^{st} domains (S_{21}, S_{25}, S_{57} and S_{61}), and thus reveals the network-wide pattern that we stimulated. From this, we infer that

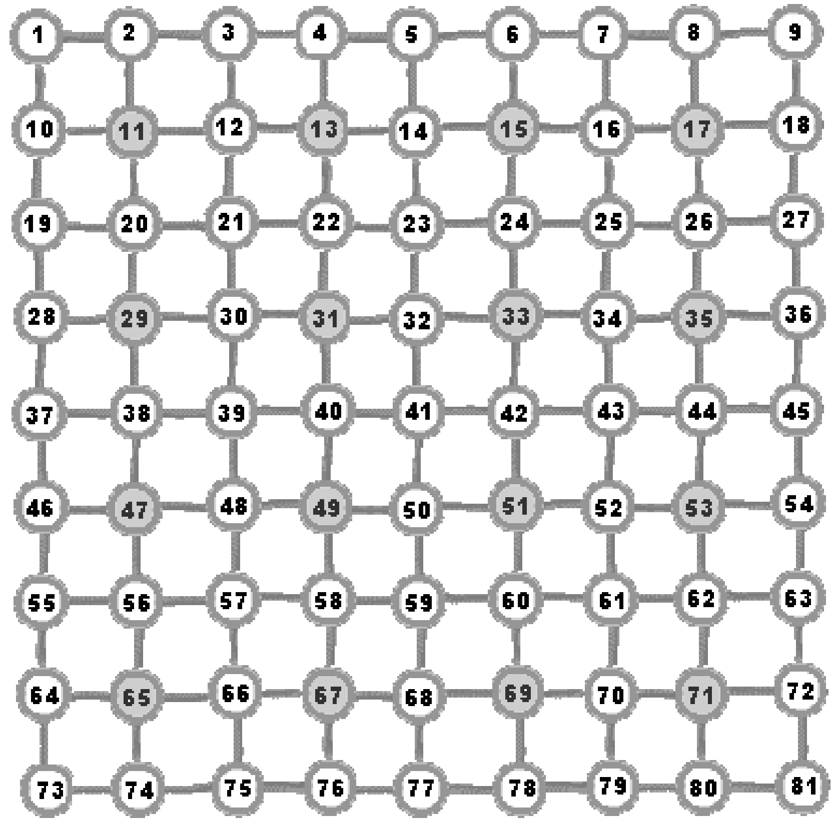

Fig. 7. The larger network with 81 domains and 16 selected observation points (gray).

such patterns can be detected even without complete observations. Note also, that this technique managed to find the congestion pattern without sampling packets flowing out of the congested domains.

What if we further reduce the number of sample points? Select $L = 8$ as the number of observation points (i.e., the 13^{th}, 15^{th}, 29^{th}, 35^{th}, 47^{th}, 53^{rd}, 67^{th}, and 69^{th} domains here). Thus, the flow vector $\boldsymbol{x}$ has $L \times (N-1) = 8 \times 80 = 640$ variables. The related weight vector $\boldsymbol{S}$, calculated with 200 data points spanning two periods, is shown in Figure 8(c), which is almost the same as Figure 8(b). Next, with the observed data from only these eight sample points, we calculate the weight vector $\boldsymbol{S}$ using M data points within a moving time window MT from one time period to the next. Figure 9 shows the weight vector $\boldsymbol{S}$ evolving with T = 2.1 s and the time window MT (= 200 × 2.1 s = 420 s) sliding ahead every 42 s. With a few observation points visibility into time-varying network congestion appears indistinguishable during non-transient periods; however, we find that the effect of transient periods is very helpful for capturing the network-wide pattern shifting over time.

Can the proposed method succeed with still further reduction in the number of sample points? We finally select $L = 4$ as the number of observation points (i.e.,

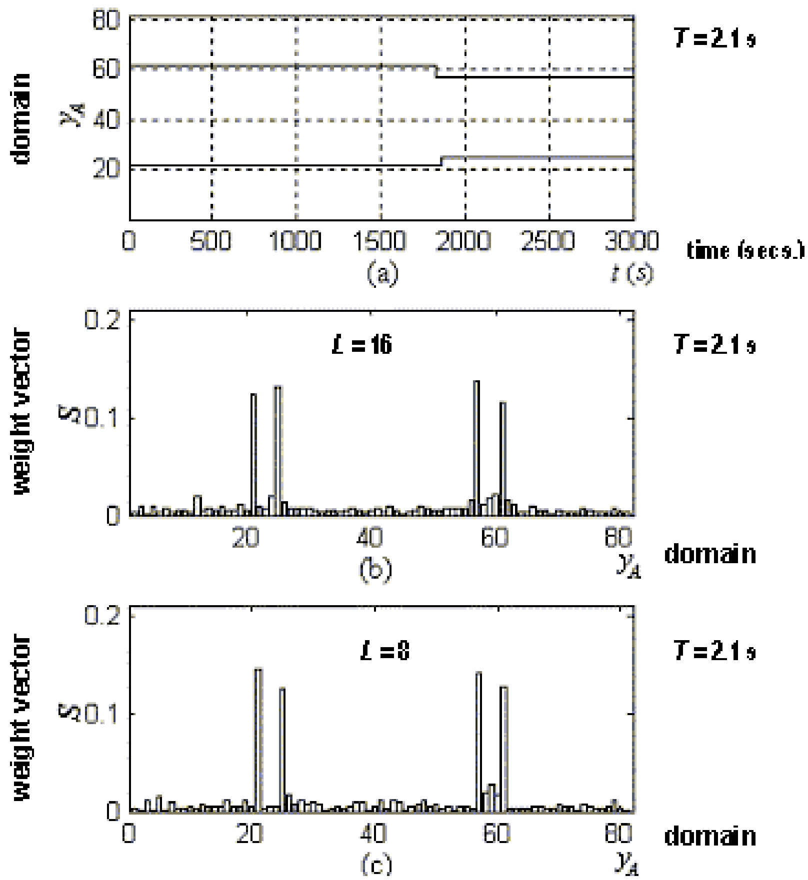

Fig. 8. (a) The most congested domains changing over time, and two weight vectors with $L = 16$ (b) and $L = 8$ (c).

the 31^{st}, 33^{rd}, 49^{th}, and 51^{st} domains here). The flow vector $\boldsymbol{x}$ has $L \times (N-1) = 4 \times 80 = 320$ variables. The weight vector $\boldsymbol{S}$, again calculated with 200 data points spanning two periods, is shown in Figure 10(a). Here, the performance of the method appears to degrade. While Figure 10(a) reveals the network-wide pattern to some extent, it also exhibits differences with Figure 8(b) and Figure 8(c). We attribute these differences to local effects being amplified in the weight vector, but not appearing in the global pattern of Figure 8(b) and Figure 8(c). For example, S_{12} is very prominent in Figure 10(a), but not in Figure 8. This occurs because traffic from our four sampling domains to the 12^{th} domain appears jammed because the routing algorithm in our model [11] forwards packets through the congested 21^{st} domain. Despite degraded performance, the weight plot in Figure 10(a), though derived from only four sample points, is still helpful for inferring the network-wide pattern.

Can we derive further insight by decomposing the network into parts with regard to the data analysis? We divide the network into three parts (i.e., $1^{st} \sim 27^{th}$, $28^{th} \sim 54^{th}$, and $55^{th} \sim 81^{st}$), and analyze each separately. Since all hotspots exist in the first and third parts, Figure 10(b) and 10(c) show respectively their weight vectors, each of which is calculated with the flow

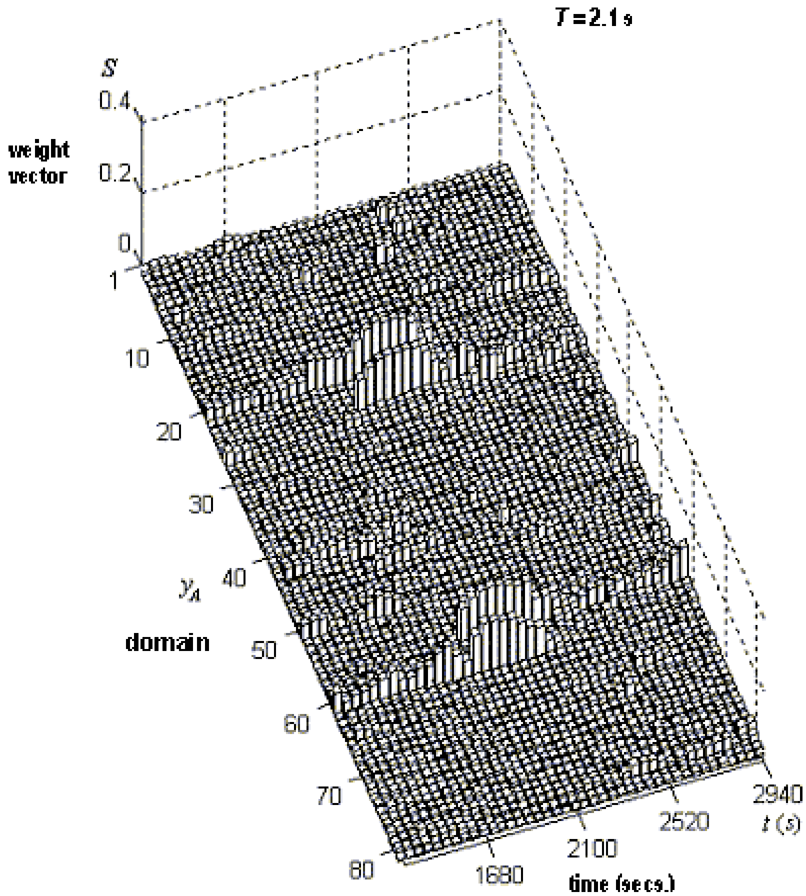

Fig. 9. The spatial-temporal pattern observed with T =2.1 s at eight observation posts.

vector of 4 × 27 = 108 variables. Notice that the weights of the domains are enhanced in these local maps. Figure 10(d) shows distinctly S_{57} (dotted line) and S_{61} (solid line) within the third part, which change with the time window MT (= 200 × 2.1 s = 420 s) moving ahead every 21 s (recall Figure 9).

Our experiments suggest that we can gain network-wide knowledge of changing congestion patterns with substantially reduced data sets, but what effect does this reduced data have on computation requirements? Might we perform data analysis to support real-time monitoring? To produce Figure 10(c) requires just 0.06 s for computing the correlation matrix, all eigenvalues and eigenvectors with MATLAB on a 1 GHz computer. Our other analyses required more computation: 1.10 s for Figure 10(a), 9.98 s for Figure 8(c), and 82.92 s for Figure 8(b).

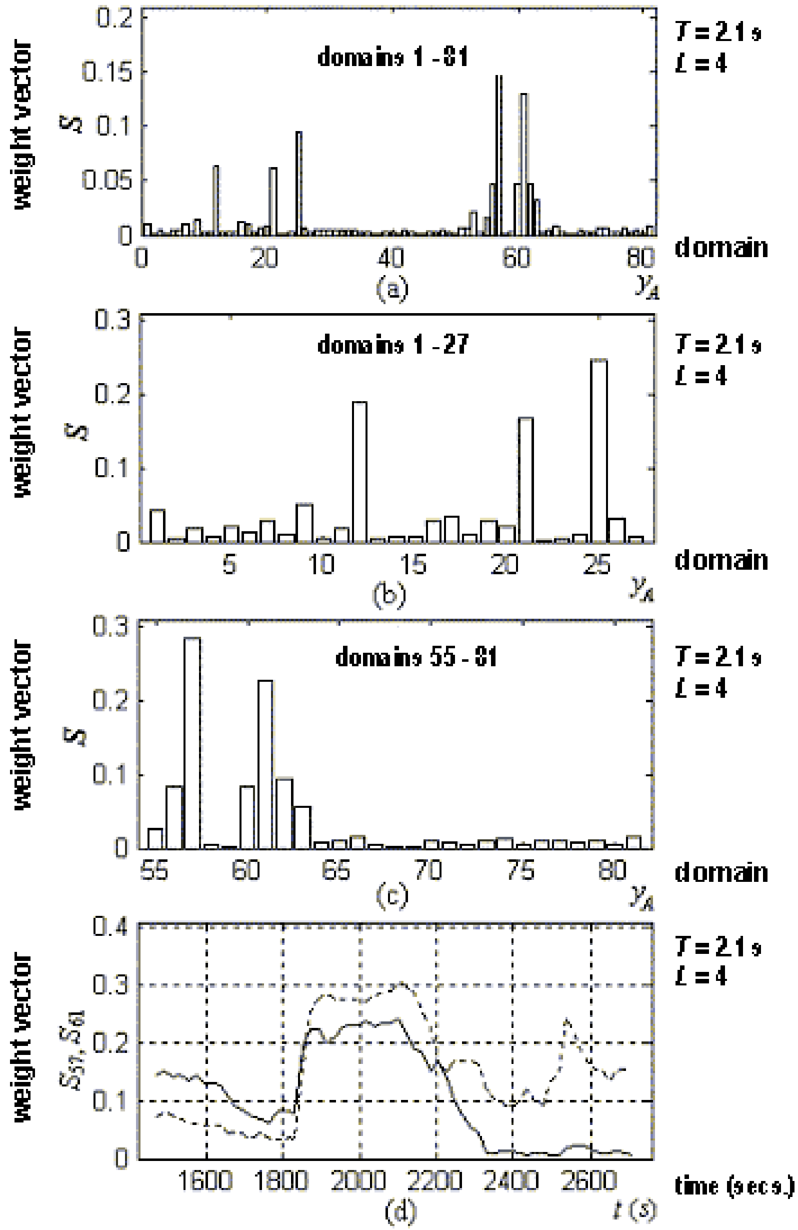

Fig. 10. (a) a weight vector with L =4, (b) (c) two weight vectors for the first (1^{st} $\sim 27^{th}$) and third parts ($55^{th} \sim 81^{st}$), and (d) S_{57} (dotted line) and S_{61} (solid line) of the third part.

4 Future work

The preliminary results we presented here encourage further investigation. We can conceive future work along two dimensions: scientific and engineering. Along the scientific dimension, we plan to investigate the applicability of our proposed cross-correlation based method to observe complex phenomena using a more realistic simulation model of a large-scale network. Such investigation should further test the utility of our analysis method. We also need to understand differences between our simulation results and the cross-correlations reported in the Renater study, which appear much stronger than those we found from our simulation model. For example, the largest eigenvalue reported in the Renater study is at least four times larger than the eigenvalues estimated from our simulations. This might imply that actual traffic demand varies much more violently than the simulated “square wave” used in our experiments. If this implication proves valid, then our proposed method might be quite well suited for use in network operation and engineering. We can further investigate this question as suitable data becomes available from Internet measurements. At the same time, we could consider insights provided by other statistical techniques [19], such as process control methods for multivariate correlated data.

Along the engineering dimension, we plan to investigate a range of applications for our proposed analysis method. For example, using a more realistic network simulation, we plan to explore the ability of cross-correlation based analysis to reveal the macroscopic effect of distributed denial of service (DDoS) attacks. Can our method reveal the dynamics of various attack types, such as constant rate, increasing rate, natural-network-like-congestion, subgroup, and pulsing attacks? Can our method distinguish the existence of multiple attack targets and the location of attack sources? Simulation might also allow us to examine the utility of our method to guide real-time traffic engineering in response to shifting network demands. If such simulation experiments yield encouraging results, then we could use real Internet data, once available, to test the applicability of our analysis method to a large operational network. Simulation could also help us evaluate appropriate values for various parameters, e.g., the number of data points to collect (M), the time granularity to observe (T), and the number (L) and location of observation points, associated with our proposed analysis method. After we understand better theoretical parameters to use, we can consider practical engineering methods associated with deployment and application. Though we can imagine data recorded, possibly by NetFlow [18], and transmitted frequently to a collection server, significant practical questions remain. For example, could such data collection induce measurement artifact into the cross-correlation eigenvalue depiction of the network? How many observation points can be deployed in the Internet, and where? Should a central site manage data collection and analysis, or could decentralized sites collaborate to exchange subsets of data collected and analyzed independently? Further, can network-wide traffic monitoring be deployed as a real-time service to support scientific research, to aid traffic engineering, to inform end users about network conditions, and to provide early warning of possible DDoS attacks? If such a service proves feasible, then how can network-wide shifts in traffic patterns be used to trigger more detailed monitoring activities, for example, to verify that a hotspot really exists or that a DDoS attack is underway?

5 Conclusions

Operating and engineering large-scale networks could benefit from development of promising tools to monitor network-wide traffic in space and time. In this paper, we investigated spatial-temporal traffic analysis using a cross-correlation method, based on the eigenvector of the largest eigenvalue. To illustrate the method, and reveal its promise, we reported simulation results from some experiments using a rather simple network model. Through a defined weight vector, we could identify macroscopic traffic patterns within the simulated network at the critical timescale, which allowed us to observe the more prominent weights of congested domains as increased correlation arises. We evaluated our method with various reductions in the number of observation points, and suggested that we could still capture the network-wide pattern shifting over time. We identified some degradation in the performance of our proposed method as the number of sample points passed below a threshold; however, we also suggested that we could compensate for this degradation somewhat by dividing the network into sub-areas, and then focusing on each smaller area separately. Our experiments suggest a possibility to observe network-wide shifts in congestion patterns with substantially reduced data sets and lower computation requirements, which might enable data analysis in support of real-time monitoring.

References

[1] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling", in *Proc. ACM SIGCOMM '94*, pp. 257-268, 1994.

[2] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes", in *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, May 1996.

[3] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control", in *Proc. ACM SIGCOMM '99*, pp. 301-313, 1999.

[4] A. Feldmann, A. C. Gilbert, W. Willinger and T. G. Kurtz, "The changing nature of network traffic: Scaling phenomena", *ACM SIGCOMM Computer Communication Review 28* (2) (1998) 5-29.

[5] M. Crovella and E. Kolaczyk, "Graph Wavelets for Spatial Traffic Analysis", in *Proceedings of IEEE Infocom 2003,*San Francisco, CA, USA, April 2003.

[6] R. Mahajan, S. Bellovin, S. Floyd, J. Ioannidis, V. Paxson, and S. Shenker, Controlling high bandwidth aggregates in the network, Technical Report, AT&T Center for Internet Research at ICSI, July 2001.

[7] P. Barford and D. Plonka, “Characteristics of network traffic flow anomalies”, in *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, San Francisco, CA, USA, November 2001.

[8] M. Barthelemy, B. Gondran, and E. Guichard, “Large-scale cross-correlations in Internet traffic”, *Physical Review E* 66 (2002) 056110.

[9] V. Plerou, *et al.*, “Random matrix approach to cross correlations in financial data”, *Physical Review E* 65 (2002) 066126.

[10] K. Thompson, G. J. Miller, and R. Wilder, “Wide-Area Internet Traffic Patterns and Characteristics”, *IEEE Network* 11 (6) (1997) 10-23.

[11] J. Yuan and K. Mills, “Simulating the Time-Scale Dynamics of Network Traffic Using Homogeneous Modeling”, submitted to the *Journal of Research of the National Institute of Standards and Technology*, 2004.

[12] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[13] MATLAB Ueser’s Guide, The MathWorks, Inc., Natick, MA, USA, 1998.

[14] H. J. Kim, Y. Lee, B. Kahng, and I. Kim, “Weighted scale-free network in financial correlations”, *Journal of the Physical Society of Japan* 71 (9) (2002) 2133-2136.

[15] K. I. Goh, B. Kahng, and D. Kim, “Spectra and eigenvectors of scale-free networks”, *Physical Review E* 64 (2001) 051903.

[16] J. Yuan, K. Mills, “Exploring Collective Dynamics in Communication Networks”, *Journal of Research of the National Institute of Standards and Technology* 107 (2) (2002) 179-191.

[17] M. Grossglauser and D. Tse, “A time-scale decomposition approach to measurement-based admission control”, in: *Proceedings of IEEE Infocom ’99*, pp. 1539-1547, New York, NY, March 1999.

[18] Cisco NetFlow. http://www.cisco.com/warp/public/732/netflow/index.html.

[19] J. O. Ramsay and B. W. Silverman, Functional Data Analysis, Springer, New York, 310 pages, 1997.