On the feasibility of utilizing correlations between user populations for traffic inference

Kun-chan Lan
National ICT Australia Ltd
Bay 15, Australian Technology Park
Eveleigh NSW 1430, Australia
Kun-chan.Lan@nicta.com.au

John Heidemann
USC Information Sciences Institute
4676 Admiralty Way, Suite 1131
Marina Del Rey, CA 90292
johnh@isi.edu

***Abstract*— Network models today are often derived from two different methods. On one hand, detailed traffic models are generated based on traces from a single tap into the network. Alternatively, one can collect higher-level traffic-matrix data with SNMP from many routers. However, inferring flow-level details from such data is still an open research issue. Today it is infeasible to collect a fine-grained, packet-level representation of a complete, multi-router network. Even if it were economically feasible to synchronize and monitor every router in a large network, the amount of data generated would tax storage and computation resources. In this work, we propose a methodology to *infer* flow-level traffic across a network by exploiting the correlations between user populations across different networks. The contribution of this paper is twofold. First, based on traces of web traffic collected from two different sources, we observe that the user-behavior parameters of the traffic (such as user "think" time in web traffic) are correlated across time, while the application-specific parameters of the traffic (such as object size) are correlated across "similar" networks. Second, by utilizing the correlations between similar networks, we propose a methodology for inferring traffic at places where continuously taking measurements is infeasible. We evaluate the effectiveness of our methodology via simulation.**

I. Introduction

In order to get a complete picture of network-wide view of the traffic, it is necessary to integrate data collected from multiple points in a network. Recently, there is increasing interest in using link load statistics to estimate point-to-point traffic matrix via SNMP data collected from multiple routers [13], [27], [28]. However, it is still an open research problem to infer flow statistics using such coarse-grained aggregated measurements. In today's production IP network, it is typically infeasible to collect fine-grained, packet-level information at every single router in a large network due to administrative and technical issues. For example, a single direction of OC48 link can produce as much as 100GB of packet headers per hour [14]. For large ISPs with many links to monitor, such massive amounts of data would place enormous demands on storage and computation resources. To reduce the huge overhead from continuously monitoring and collecting measurements on every single network, one possible remedy is *indirect measurement*. The goal of indirect measurement is to infer traffic of one part of the network based on measurements taken from other parts of the network.

Network traffic can be correlated at different times and at different places for various reasons. It is well-known that network traffic follows a diurnal pattern [17]. People tend to work more actively and consume more network bandwidth during normal office hours. Previous studies [15], [23] showed that the organization membership has a significant impact on the degree of local sharing. In other words, members of an organization are more likely to request the same documents than users from other organizations. Such an observation suggests that users in the same organization might exhibit similar user behavior and application usage pattern.

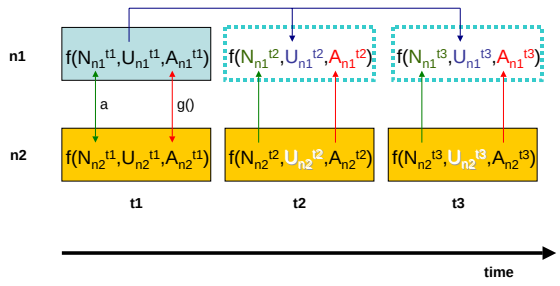


Fig. 1. Traffi c inference using the similarity between networks

In this work, we propose a methodology to infer network traffic by exploring correlations of user populations between different networks. We develop an systematic approach to model traffic on network n_1 by utilizing measurements taken from network n_2, provided networks n_1 and n_2 have *similar* user populations. Figure 1 illustrates this concept: colored squares indicate data that is collected, while data in dotted squares is inferred. Based on initial measurements at time $t1$ confirming the similarity of n_1 and n_2, we use future measurements of n_2 to predict the traffic in n_1 at times $t2$ and $t3$.

In this paper, we assume that one can model network traffic with two types of parameters: *user-behavior parameters* and *application-specific parameters*. *User-behavior parameters* characterize how users utilize the applications,

while *application-specific parameters* are the traffic parameters corresponding to the structure of one particular application. For example, *user-behavior parameters* could be the distributions of user "think" time and the number of pages requested by the users in web traffic, while *application-specific parameters* could be the distributions of object size and the number of objects in a web page. Based on measurements from two different sources, we observe that the distributions of user-behavior parameters tend to be correlated over time on the same network. Additionally, our data suggests that the distributions of application-specific parameters are likely to be correlated between two networks with *similar* user populations.

Based on the above observations, we propose the following approaches to infer traffic on network n_1 from measurements collected on network n_2, provided n_1 and n_2 have similar user populations. The first step is to collect some period of traffic on both networks n_1 and n_2. Such initial measurements are used to derive the temporal and spatial correlations between n_1 and n_2 (such as α and $g()$ in Figure 1. Note that in this paper we assume that such correlations do not vary over time. In addition, due to the limitation of our data, we only validate our methodology for its application at the time scale of hours.) and confirm the similarity between n_1 and n_2. Once the similarity between n_1 and n_2 is confirmed, we then can utilize the derived correlations to model n_1's traffic at any future time based on *only* measurements taken from n_2.

The contribution of this paper is twofold. First, based data collected from two different sources, we observe that the user-behavior parameters of the traffic (such as user "think" time in web traffic) are correlated across time on the same network, while the application-specific parameters of the traffic (such as object size) are correlated across similar networks (Section III). By utilizing the correlations between similar networks, we then propose a methodology to infer traffic at places where continuously taking measurements is not feasible. We also evaluate the effectiveness of approach via simulations (Section IV).

II. Related work

Our work builds on prior work in traffic inference, traffic correlation and bursty traffic.

A. Traffic inference

Several research efforts tried to infer traffic based on indirect measurements. They can be basically categorized into two different directions. One direction is to infer internal network behavior based on results of active probing from end points. Another form of indirect measurements focuses on using link load statistics to estimate point-to-point traffic matrix.

The first type of traffic inference aims to to characterize internal network behavior based on end-to-end performance measurements, including MINI, IEPM, AMP, RIPE, Surveyor [3] and TReno [12] etc. They utilize either unicast probes (via tools such as *traceroute* and variants of *ping*) or multicast probes, and correlate end-to-end traffic measurements collected at different monitoring machines across the Internet to infer statistical properties of the network such loss, delay and topology.

Another form of indirect measurements, known as traffic matrix estimation, emphasizes on estimating individual flow characteristics based on aggregated traffic measurements [4], [5], [20]. The idea is to infer traffic matrices (the set of traffic between all pairs of sources and destinations) based on link bytes counts which are readily available through SNMP that is provided by most of the commercial routers.

Although our work also relies on utilizing passive measurements to infer traffic, our problem domain is different from previous work in traffic matrix estimation. While the goal of traffic matrix estimation is to infer flow-level details from coarse-grained aggregated measurements collected from multiple routers, our work focuses on exploring the temporal and spatial correlations across "similar" networks Our goal is to avoid the overhead from *continuously* monitoring *all* networks by projecting traffic from measurements taken at a few, similar networks. We expect our work can complement previous traffic inference infrastructures.

B. Traffic correlation

Prior work has shown that network traffic is often correlated at different times and at different places. For example, web caching work has shown that web access is more similar between clients in the same organization than between random clients. They would likely access the same set of documents and each client tends to browse back and forth within a short period of time. Web caching utilizes these temporal and spatial correlations by caching documents which are very likely being requested again in the future to save the network bandwidth and lower access latency for the clients [21].

Another example that shows that traffic can be correlated temporarily is the diurnal pattern. It is well-known that network traffic follows a daily pattern. People tend to work more actively and consume more network bandwidth during normal office hours

Finally, traffic can be correlated due to the sharing of common resources. For example, when several TCP flows compete for bandwidth in a common gateway, it has been observed experimentally that it will result in similar bandwidth oscillation behavior due to the synchronization of window increase/decrease cycles [25]. Traffic to multiple clients can become synchronized as they wait for service from a busy server [6].

Our work explores another form of traffic correlation: the temporal and spatial correlations between two similar networks. We utilize such correlations to infer traffic at places where continuously taking measurement is infeasible.

C. Bursty traffic

Previous studies of Internet traffic have shown that a very small percentage of flows consume most of the network bandwidth [7], [16], [26], Sarvotham et al. [18] show that traffic bursts typically arise from just a few high-volume connections, which is caused by large file transmissions over high bandwidth links (a more recent study [7] also confirmed

Trace	RES	UNI
Date	Dec 2002	Oct 2003
Duration (hr)	168	168
Total Packets	79.7M	218M
Bytes	2.6G	8.4G
TCP Packets	58M (72%)	197M (90 %)
Bytes	1.8G (69%)	6.6G (78%)
UDP Packets	21.6M (27%)	13M (6 %)
Bytes	0.8G (30 %)	1.5GB (18 %)
HTTP Packets	32M (40 %)	101M (46 %)
Bytes	1.2G (46 %)	4.2G (50 %)

TABLE I
SUMMARY OF TRACES

User behavior

- The number of pages requested per user
- Page inter-arrival time (i.e. user "think" time)

Page

- Number of objects within each page
- Object inter-arrival time

Object

- Size of object

Fig. 2. Structural model of web traffi c

their observation). In our work, we observe that, at a lower level of traffic aggregation, traffic distributions between two similar networks tend to be correlated in the body but vary significantly in the tail. Furthermore, we show that the variations in the tail might be contributed by such bursty connections.

III. TRAFFIC CORRELATIONS BETWEEN SIMILAR NETWORKS

In this section, we consider web traffic as an example to demonstrate the traffic correlation between different networks. We characterize web traffic as a three-level application level model as shown in Figure 2. For web traffic, the *user-behavior parameters* include number of page per user and user think time, while the *application-specific parameters* consist of number of objects per page, object inter-arrival and object size. We derive the distributions of user-behavior and application-specific parameters of web traffic using an existing tool RAMP [10]. Based on data from two different sources, we first observe that the distributions of user-behavior parameters of web traffic are correlated over time. Next, we show that distributions of application-specific parameters of web traffic can be correlated between similar networks. Finally, we discuss the effect of "dominant" flows on the tail of the distributions.

A. Traces

The datasets used in our study are from two sources. One was collected from two subnets of a large research institute. There are several research divisions under this institute. One of these two subnets is used by the networking division while the other carries traffic from the AI division. The users of these two subnets are mainly researchers. The number of users (we consider an unique IP address as an "user") for both subnets are about the same (around 60). For the rest of the paper, we refer this institute as RES and these two subnets as RES-a and RES-b. The second data set was collected from two subnets of a large university (we refer this university as UNI and the two subnets as UNI-a and UNI-b). Each subnet serves a different computing center whose users are mainly students from CS and EE departments. The number of users in UNI-a and UNI-b are also very close (around 150). Both traces captured all inbound and outbound traffic but only TCP/IP header information was recorded. Both RES and UNI traces were collected during a seven-day period. The details of traces are given in Table I. Note that a significant portion of UDP traffic in RES trace is contributed by NFS (Network File System) traffic. The traffic volume of UNI traces is about three times of that of RES's.

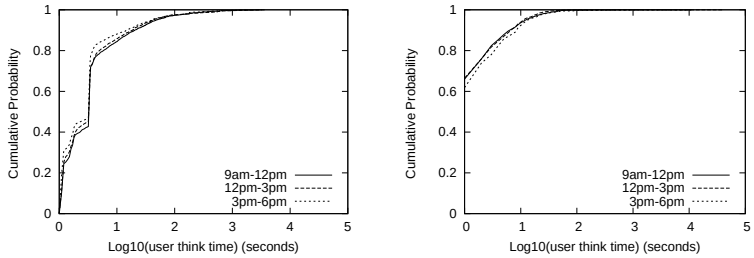
B. Temporal correlations of the traffic on the same network

Previous studies have shown that the same group of users tend to exhibit certain navigation behavior patterns when they surf on the web [2], [24]. Our data suggests that the distributions of *user-behavior parameters* derived from measurements on the same network tend to be correlated over time. Specifically, we examine the distributions of user-think time and the number of pages per user in web traffic. We find that both distributions are strongly correlated across time for both RES and UNI data. As shown in Figure 3[1], the distributions of user think time are strongly correlated across three 3-hour periods for both RES-a and UNI-a data. Note that although the distributions of user-behavior parameters are strong correlated across time on the same network, they vary significantly between different group of users as shown in Figure 3.

C. Spatial correlations of the traffic between similar networks

Prior work [15], [23] has shown that members of the same organization are likely to request similar documents from the web. For example, researchers rely strongly on search engines such as Google to search for existing literatures, and students tend to spend significant amount of time on browsing class web pages and related web sites. Based on our traces, we observe that the distributions of *application-specific parameters* tend to be correlated between two "similar" networks. Specifically, we examine the distributions of object size, the number of objects per page and object inter-arrival time in web traffic. We find that all three distributions show strong spatial correlation in RES and UNI data. For brevity, we only show the results for the distributions of object size here. As shown in Figure 4, although there is some variation in the tails, the distributions of object size are strongly correlated in the body of the distribution between two similar subnets for a 3-hour period of traffic, Note that, although the distributions of *application-specific parameters* between RES-a and RES-b are similar, their aggregated traffic can still vary significantly, as shown in Figure 5.

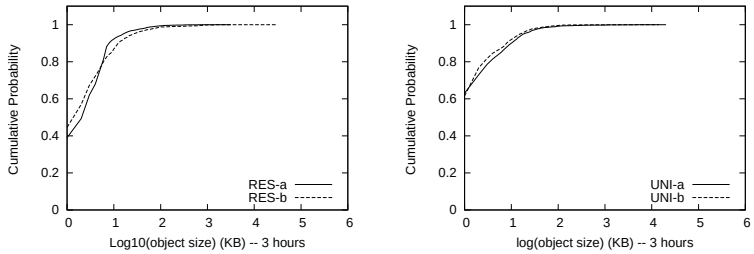
[1]Due to the space limitation, we do not show the same plots for the distribution of number of pages per user here, although the results are similar



(a) Comparison of user think time distribution at RES-a across time

(b) Comparison of user think time distribution at UNI-a across time

Fig. 3. Comparison of distribution of user think time at the same subnet across time



(a) Comparison of object size distribution between RES-a and RES-b

(b) Comparison of object size distribution between UNI-a and UNI-b

Fig. 4. Comparison of distribution of object size across similar subnets

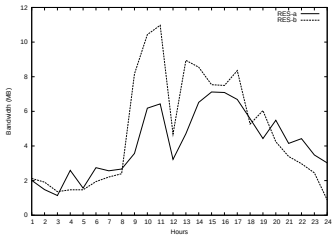


Fig. 5. Aggregated traffi c of RES-a and RES-b across a day

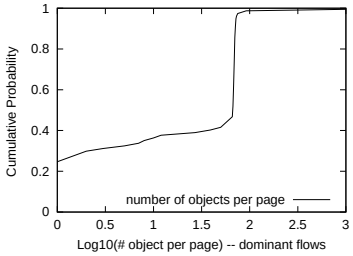


Fig. 6. Distribution of number of objects per page for the dominant flows

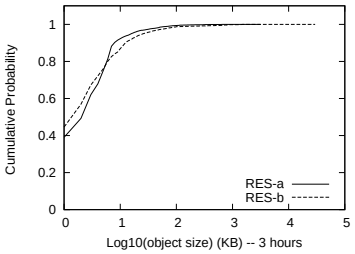
Furthermore, the spatial correlation is stronger in UNI data than in RES data in both the body and the tail of the distribution. We hypothesize that such an observation is due to different levels of traffic aggregation, which is discussed in next section.

D. Effect of the level of traffic aggregation

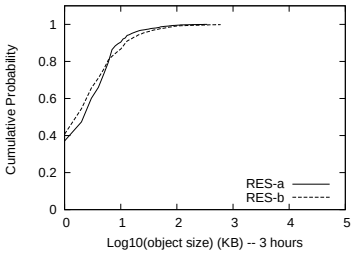
Previous studies have shown that a very small percentage of flows consume most of the network bandwidth [1], [8], [19]. We hypothesize that the variation in the tails, as described in Section III-C, might be due to such "dominant" flows. In addition, we hypothesize that the effect of such flows on the tail might be less significant when traffic is highly aggregated.

To understand the effect of "dominant" flows, we first look at only flows with a size larger than 1MB. We find that the distribution of number of objects per page for these flows is almost bi-modal, as shown Figure 6. Around 25% of such connections contain only one single big object. We then remove these domain flows from the trace and compare the distributions of object size before and after removing such connections. As shown in Figure 7, the difference in the tail become less significant once such connections are removed. Note that, as shown in Figure 7, the probability of small objects (with a size less than 1KB) also decreases once the dominant flows are filtered out.

We hypothesize that the effect of such "dominant" flows can be reduced by increasing the level of traffic aggregation. One can model the level of aggregation (G) as the product of the amount of traffic generated by the sources (S) and the length of measurement period (T). That is, $G = S \times T$. Furthermore, S can be described as a function of the number of users and the amount of traffic generated by each user. Such a model implies that one might expect that the effect of dominant flows is less significant when traffic data is collected either from a larger

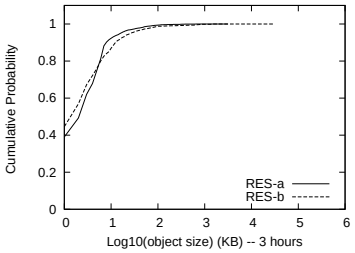


(a) Object size distribution of the original traffic

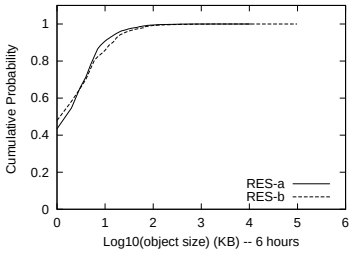


(b) Object size distribution after filtering out the dominant flows

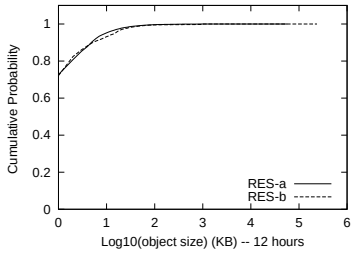
Fig. 7. Effect of dominant flows on the distributions of object size



(a) Comparison of object size distribution between RES-a and RES-b for a 3-hour interval



(b) Comparison of object size distribution between RES-a and RES-b for a 6-hour interval



(c) Comparison of object size distribution between RES-a and RES-b for a 12-hour interval

Fig. 8. The effect of traffic aggregation over time

user population or over a longer period of time.

To verify this hypothesis, we first look look at the effect of the size of user population. As described previously, the number of users in UNI data is about 2.5 times of that in RES data. Hence, one might expect that there is less variation in the tail in UNI data in RES data, provided that each user generates similar amount of traffic. As shown in Figure 4, there is a stronger match in the tail in UNI data than in RES data. We next look at the distributions of object size between RES-a and RES-b for different length of measurement period. As shown in Figure 8, we find that the differences in the tail become less significant as the sampling period becomes longer. Additionally, the probability of having more smaller objects increases as a result of a longer sampling period.

While our study is mainly based datasets obtained from university subnets, the above observation suggests the possibility of applying our methodology to larger networks such as POPs of an ISP. Although it is not obvious that two different POPs will have "similar" user populations, it might be possible that two different POPs with large aggregations of different user populations could still exhibit "similar" traffic statistics.

IV. Traffic Inference

The observation in Section III suggests a possibility to infer traffic by utilizing the "similarity" between two networks. In this section, we first define and quantify the "similarity" between two networks. By exploring traffic correlations between similar networks, we next propose a methodology to infer traffic at places where continuously taking measurement is infeasible.

A. Tests for similarity

Intuitively we can describe two networks as "similar" if they have user populations with similar characteristics. For example, traffic generated by the CS department and by the EE department in an university could be similar because students from both departments might have similar application usage patterns. Traffic generated by the finance division and by the accounting division in a big corporate could be similar because users of these two divisions might share some common tasks and applications. In other words, we consider two networks to be *similar* as long as there exist correlations in the application-level view of the traffic.

While this intuitive definition of similarity is appealing, a more formal procedure is required to determine similarity for traffic inference. We propose the following procedure to test the similarity. First, we derive distributions of user-behavior and application-specific parameters of the traffic (for example, distributions of user "think" time and object size in web traffic) from measurements. We then compare the traffic statistics of two networks qualitatively and quantitatively to determine if they are similar. By qualitatively, we visually inspect the CDF plots of the derived distributions between two networks. By quantitatively, we first normalize the derived distributions.

We then perform statistical tests to determine if two distributions are significantly different in mean, variance and shape. Specifically we utilize Student's t-Test to test mean, the F-Test to evaluate variance and Kolmogorov-Smirnov Goodness of Fit Test to test shape [11]. We consider two distributions are *strictly similar* if they pass all three tests at 99% confidence level.

The *similarity tests* mentioned above are very strict. One can imagine some networks with similar user populations might not be able to pass all three tests. Furthermore, in some situation one might consider that the difference in one particular metric (such as mean) is more important. By relaxing the above testing procedure, we define a simple "similarity" function *s* as a linear combination of differences in mean, variance and shape as the following:

$$s = w_1 \times m + w_2 \times v + w_3 \times D$$

where

$$m = |\mu(N1) - \mu(N2)|/MAX(\mu(N1), \mu(N2))$$
$$v = |\sigma(N1) - \sigma(N2)|/MAX(\sigma(N1), \sigma(N2))$$

μ is the mean and σ is variance of the data, and D is the Kolmogorov-Smirnov D value (Kolmogorov-Smirnov D value is the largest absolute difference between the cumulative distributions of two sets of data). N1 and N2 are the data samples taken from two different networks, and w_1, w_2 and w_3 are positive user-defined weights that allow user to prioritize different metrics (i.e. mean or variance or shape). The above definitions also imply that the range of possible values for m, v and D can vary from 0 to 1. Intuitively, two networks are more "similar" if the computed s is closer to 0.

The definition of "similarity" we describe above is based on the differences in first-order statistics. An area of future work would be to explore more sophisticated definitions of similarity that also consider higher-order statistics of the traffic, such as scaling properties.

B. Methodology

Our approach is based on the source-level modeling approach [9], [22]. At its simplest form, we assume that network traffic (T) can be modeled as a function of three sets of parameters: number of user (N), user behavior (U) and application-specific (A) parameters. In other words, $T = f(N, U, A)$. For example, the user-behavior parameter could be the distribution of user "think" time while the application-specific parameter could be the distribution of object size in web traffic. For the rest of discussion, we assume that U consists of a set of parameters u_1, u_2, u_3 etc., while A consists of a set of parameters a_1, a_2, a_3 etc..

In the previous section, we observe that the distributions of user behavior is correlated over time on the same network. The reasoning for this observation could be that a user might follow similar patterns in using an application (e.g. similar web browsing patterns). By utilizing such an observation, one can model distributions of user behavior at time t_2 based on measurements taken at a previous time t_1 on the same network n_1 (i.e. ${U_{n_1}}^{t_1} \approx {U_{n_1}}^{t_2}$). Additionally, we observe that distributions of application-specific parameters between two similar networks are likely to be correlated when traffic is highly aggregated (but only correlated in the body of the distributions at a lower level of traffic aggregations). The intuition behind such an observation could be that similar user populations tend to result in similar application usage patterns, such as the downloading of similar web pages. That is,
if (**traffic aggregation** == ***high***)

$$\begin{aligned} A(t_1, n_1) &= A_{body}(t_1, n_1) + A_{tail}(t_1, n_1) \\ &\approx A(t_1, n_2) \\ &= A_{body}(t_1, n_2) + A_{tail}(t_1, n_2) \end{aligned}$$

else

$$A_{body}(t_1, n_1) \approx A_{body}(t_1, n_2)$$

provided n_1 and n_2 are two similar networks.

By utilizing the above observations, we propose the following procedure to infer traffic on network n_1 using measurements obtained from network n_2, provided n_1 and n_2 have similar user population. The first step is to collect some period of traffic (say, at time t_0) on both networks n_1 and n_2. Such initial measurements are used to derive the three sets of traffic parameters (i.e. N, U and A) of n_1 and n_2. We then compare the derived statistics to determine if there is any spatial correlation between n_1 and n_2 (i.e. decide if n_1 and n_2 are "similar") by employing similarity tests described in Section IV-A.

Once the similarity between n_1 and n_2 is confirmed, one can then derive the spatial correlations between n_1 and n_2. For example, the similarity between n_1 and n_2 might suggest that the number of users in network n_1 is a function of the number of users in network n_2 at any given time (i.e. $N_{n_1} = \alpha \times N_{n_2}$, where α is a function of time or some constant). By utilizing traces collected at time t_0, one can compute this scaling factor α. Similarly, there might exist some functions g_1, g_2, etc. so that a_{1n_1}=$g_1(a_{1n_2})$, a_{2n_1}=$g_2(a_{2n_2})$.... These functions (i.e. g_1, g_2, etc) can also be identified by comparing the measurements taken at time t_0. At its simplest form, such a function can be as simple as $g(x) = y$.

Once the spatial correlations between n_1 and n_2 are derived from the initial measurements, we can infer traffic of n_1 at any future time using only the measurements from n_2. Specifically, we can infer the number of active users of n_1 since $N_{n_1} = \alpha \times N_{n_2}$. We can infer the distributions of application-specific parameters of n_1 based on traces collected on n_2 because
A_{n_1}=$(a_{1n_1}, a_{2n_1}, ...)$=$(g_1(a_{1n_2}), g_2(a_{2n_2}), ...)$=$g(A_{n_2})$.
Finally, since the user-behavior parameters are correlated over time on the same network (i.e. ${u_1}_{n_1}^{t_1} = {u_1}_{n_1}^{t_2}$ and ${u_2}_{n_1}^{t_1} = {u_2}_{n_1}^{t_2}$), the user-behavior parameters of n_1 traffic at any given time can be inferred based on the statistics derived previously at t_0. In other words,
${T_{n_1}}^t$=$f({N_{n_1}}^t, {U_{n_1}}^t, {A_{n_1}}^t) \approx f(\alpha \times {N_{n_2}}^t, {U_{n_1}}^{t_0}, g({A_{n_2}}^t))$.
Note that here we assume that the spatial correlation functions do not vary over time, which might not be true for some networks.

C. Evaluation

To evaluate our methodology, we utilize two 2-hour traces (one from 10am to 12am and the other from 3pm to 5pm) from RES-a and RES-b respectively. We refer these traces as RES-a-am, RES-a-pm, RES-b-am and RES-b-pm for the rest of this section. The goal is to infer the traffic of RES-a-pm based on measurements of RES-a-am, RES-b-am and RES-b-pm. We evaluate our results by comparing the RES-a-pm trace with simulation results generated using our methodology.

We first employ the similarity tests described in Section IV-A to confirm the similarity between RES-a-am and RES-b-am. Here we only show results for the distributions of object size. The results for other distributions are similar. We use student's t Test, F Test and Kolmogorov-Smirnov Goodness of Fit Test to test the differences of mean, variance and shape between RES-a-am and RES-b-am. The first two tests passed, but KS-test fails at 99% confidence level (corresponding to a critical value of 0.00874). Additionally, we also compute the m, v and D values as described in in Section IV-A. The m, v and D values between RES-a-am and RES-b-am are 0.01, 0.009 and 0.011 respectively.

Next, we derive the spatial correlations between RES-a and RES-b by comparing RES-a-am with RES-b-am data [2]. We then utilize the derived statistics to project a model of RES-a-pm based on RES-b-pm trace. Finally, we input the projected RES-a-pm model into the ns-2 simulator and compare the generated synthetic traffic with real RES-a-pm traffic.

To evaluate the results of our model, we compare the CDF plots of some first-order statistics (such as flow size and flow duration) and the wavelet scaling plots [10] between the RES-a-pm trace and the simulation results. As shown in Figure 9, all plots match closely.

Furthermore, to evaluate the effectiveness of the similarity test described in Section IV-A, we took another two 2-hour traces during the same period from another subnet of RES. This subnet serves mainly the users from the business office of RES. We refer this subnet as RES-x and the corresponding traces as RES-x-am and RES-x-pm for the rest of this section. Intuitively, one might expect that the user population of RES-x is "different" from RES-a's and RES-b's.

We first perform similarity tests on RES-x-am and RES-b-am. As expected, all three tests fail. Additionally, the computed m, v and D values are 0.16, 0.9 and 0.049, which are significantly larger than the results from the comparison of RES-a-am and RES-b-am. Nevertheless, to understand that if it is still possible to infer RES-x traffic from RES-b, we construct a model of RES-x-pm which is projected from RES-b-pm trace. As shown in Figure 10, the projected model of RES-x-pm has significant deviations from the real traffic in the flow statistics, and higher energy in the wavelet scaling plot.

[2]We set the scaling factor α to 1 and choose a simple linear function $g(x) = y$ to model the correlation function since that the number of users and distributions of application-specific parameters between RES-a-am and RES-b-am are very similar

V. Future work and Conclusion

The similarity tests described in Section IV-A are based on the comparison of some first-order statistics of the traffic between two networks. Another possible direction to evaluate the level of similarity is to compare the higher order statistics of the traffic. We plan to study this issue as future work.

Our initial results indicate that our methodology performs well for inferring traffic at the time scale of hours for small networks such as university subnets. Due to the limitation of our data, we are not able to evaluate the applicability of our approach for larger time scales (such as month or year) and for larger networks. Additionally, due to the nature of our traces, we employ a simple *correlation function* in our simulation study for projecting traffic. It might be non-trivial to compute such a function for other networks. To further validate our approach, we plan to collect more traces from other places. We are particularly interested in looking at places where traffic is highly aggregated such as the POPs of a large ISP. Specifically, we'd like to investigate if the aggregation of different user populations could still introduce similarity in the traffic.

Obtaining a network-wide view of traffic requires data collection from multiple points of the network. Unfortunately, it is economically and technically infeasible to continuously collect packet-level information at all routers in a large network. In this work, we propose a methodology to infer network traffic by exploring the correlations of user populations between different networks. The main contributions of this paper are the following: first, based on traces of web traffic collected from two different sources, we observe that the user-behavior parameters of the traffic (such as user "think" time in web traffic) are correlated across time, while the application-specific parameters of the traffic (such as object size) are correlated across "similar" networks. Our data also suggests that, at a lower level of traffic aggregation, the distributions of traffic between two similar networks tend to be correlated in the body but vary significantly in the tail. Furthermore, we show that the variations in the tail might be due to bursty connections. Second, by utilizing the correlations between similar networks, we present a methodology for inferring traffic at places where continuously taking measurements is infeasible. We then evaluate the effectiveness of our methodology via simulations.

References

[1] Supratik Bhattacharyya, Christophe Diot, Jorjeta Jetcheva, and Nina Taft. Pop-level and access-link-level traffic dynamics in a tier-1 POP. In *Proceeding of ACM SIGCOMM Internet Measurement Workshop 2001*, pages 39–54, San Francisco Bay Area, November 2001.

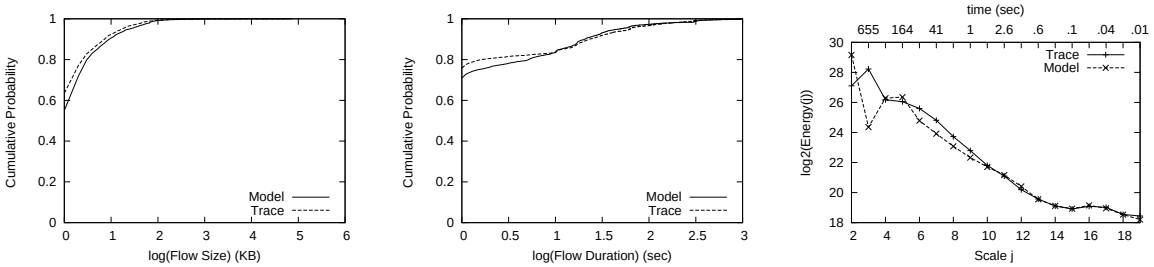
[2] Jose Borges and Mark Levene. Data mining of user navigation patterns. In *Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99)*, pages 31–36, San Diego, CA, August 1999.

[3] CAIDA. Internet measurement infrastructure. http://www.caida.org/analysis/performance/measinfra/.

[4] J. Cao, D. Davis, S. Wiel, and B. Yu. Time-varying network tomography : Router link data. *The Journal of American Statistics Association*, 95(452):1063–1075, February 2000.

[5] J. Cao, Scott Vander Wiel, Bin Yu, and Zhengyuan Zhu. A scalable method for estimating network traffic matrices. *Bell Labs Tech. Report*, 2000.

[6] Kun chan Lan and John Heidemann. Multi-scale validation of structural models of RealAudio traffic. Technical Report ISI-TR-544, USC/Information Sciences Institute, September 2001.

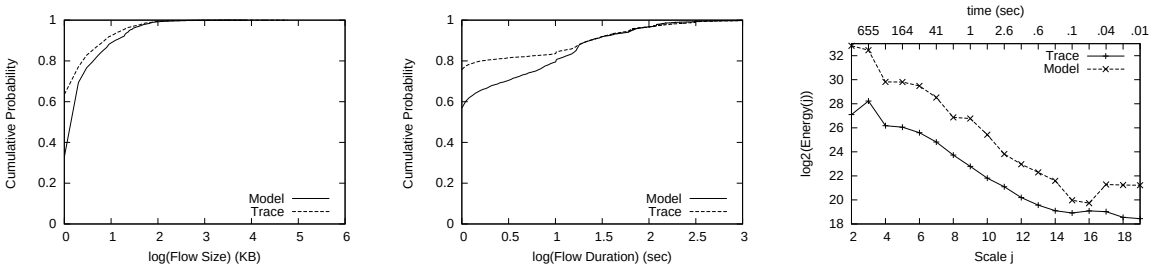


(a) Comparison of flow size distribution

(b) Comparison of flow duration distribution

(c) Comparison of wavelet scaling plot between trace and model

Fig. 9. Comparison between trace and model for networks with similar user population



(a) Comparison of flow size distribution

(b) Comparison of flow duration distribution

(c) Comparison of wavelet scaling plot between trace and model

Fig. 10. Comparison between trace and model for networks with different user population

[7] Kun chan Lan and John Heidemann. On the correlation of internet flow characteristics. *USC/ISI Technical Report ISI-TR-574*, 2003.

[8] WenJia Fang and Larry Peterson. Inter-AS traffic patterns and their implications. In *Proceeding of IEEE GLOBECOM 99*, pages 1859–1868, Rio de Janeiro, Brazil, 1999.

[9] Sally Floyd and Vern Paxson. Difficulties in simulating the Internet. *ACM/IEEE Transactions on Networking*, 9(4):392–403, February 2001.

[10] Kunchan Lan and John Heidemann. Rapid model parameterization from traffic measurement. *ACM Transactions on Modeling and Computer Simulation*, 12(3):201–229, July 2002.

[11] Massey and F. J. Jr. The kolmogorov-smirnov test of goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, March 1951.

[12] M. Mathis and J. Mahdavi. Diagnosing internet congestion with a transport layer performance tool. In *Proceedings of INET '96*, Montreal, June 1996.

[13] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic matrix estimation: existing techniques and new directions. In *ACM SIGCOMM*, pages 161–174, Pittsburgh, Pennsylvania, USA, August 2002. ACM.

[14] M. Thorup N.G. Duffield, C. Lund. Charging from sampled network usage. In *Proceeding of ACM SIGCOMM Internet Measurement Workshop 2001*, San Francisco Bay Area, November 2001.

[15] Venkata N. Padmanabhan and Lili Qui. The content and access dynamics of a busy web site: findings and implicatins. In *ACM SIGCOMM*, pages 111–123, Stockholm, Sweden, August 2000. ACM.

[16] D. Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian, and C. Diot. A pragmatic definition of elephants in internet backbone traffic. In *Proceeding of ACM SIGCOMM Internet Measurement Workshop 2002*, pages 175–176, Marseille, France, November 2002.

[17] Vern Paxson. Empirically derived analytic models of wide-area TCP connections. *IEEE/ACM Transactions on Networking*, 2(4):316–336, 1994.

[18] Shriram Sarvotham, Rudolf Riedi, and Richard Baraniuk. Connection-level analysis and modeling of network traffic. In *Proceeding of ACM SIGCOMM Internet Measurement Workshop 2001*, San Francisco Bay Area, November 2001.

[19] Anees Shaikh, Jennifer Rexford, and Kang Shin. Load-sensitive routing of long-lived IP flows. In *Proceedings of the ACM SIGCOMM*, pages 215–226. ACM, August 1999.

[20] Y. Vardi. Network tomography : Estimating source-destination traffic intensities from link data. *The Journal of American Statistics Association*, 91(433):365–377, March 1996.

[21] Jia Wang. A survey of web caching schemes for the internet. *ACM Computer Communication Review*, 29(x):xxx, October 1999.

[22] W. Willinger, V. Paxson, and M. Taqqu. Self-similarity and heavy-tails: Structural modeling of network traffic. In *Self-Similarity and Heavy-Tails: Structural Modeling of Network Traffic, in A Practical Guide To Heavy Tails: Statistical Techniques and Applications, R.J. Adler, R.E. Feldman and M.S. Taqqu, editors. ISBN 0-8176-3951-9. Birkh auser, Boston, 1998.*, 1998.

[23] Alec Wolman, Geoffrey M. Voelker, Nitin Sharma, Neal Cardwell, Molly Brown, Tashana Landray, Denise Pinnel, Anna R. Karlin, and Henry M. Levy. Organization-based analysis of web-object sharing and caching. In *USENIX Symposium on Internet Technologies and Systems*. USENIX, 1999.

[24] Osmar R. Zaiane, Man Xin, and Jiawei Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. In *Proceedings of the Advances in Digital Libraries (ADL'98)*, pages 19–29, Santa Barbara, CA, April 1998.

[25] L. Zhang and D. Clark. Oscillating behavior of network traffic: a case study simulation. *Internetworking: Research and Experience*, 1(2):101–12, 1990.

[26] Yin Zhang and Lili Qiu. Understanding the end-to-end performance impact of RED in a heterogeneous environment. *Cornell CS Technical Report TR2000-1802*, 2000.

[27] Yin Zhang, Matthew Roughan, Nick Duffield, and Albert Greeberg. Fast accurate computation of large-scale ip traffic matrices from link loads. In *Proceedings of the ACM SIGMETRICS*, pages 206–217, San Diego, CA, USA, May 2003. ACM.

[28] Yin Zhang, Matthew Roughan, Carsten Lund, and David Donoho. An information-theoretic approach to traffic matrix estimation. In *ACM SIGCOMM*, pages 301–312, Karlsruhe, Germany, August 2003. ACM.