Predicting Network Flow Behavior From Five Packets

Stefan Karpinski, John R. Gilbert, Elizabeth M. Belding

Department of Computer Science
University of California, Santa Barbara

{sgk,gilbert,ebelding}@cs.ucsb.edu

ABSTRACT

We observe that when network traffic behaviors are represented in vector spaces as relative frequency histograms of behavioral features, they exhibit low-rank linear structure. We hypothesize that this structure is due to the distribution of flow behaviors following a finite mixture model. Aside from being of theoretical interest, this hypothesis has practical consequences: it allows us to make predictions about the probabilities of future flow behaviors from a handful of a flow's initial packets. From observing five initial packets, we are able to predict the distribution of future packet sizes and inter-packet intervals with between 70% and 90% accuracy across a variety of network traces. We can predict which flow will have more packets in pairwise comparisons with between 65% and 85% accuracy. These practical applications serve dual functions. They provide highly useful tools for network management, routing decisions, and quality of service schemes. However, they also provide evidence that the hypothesized model gives a correct explanation for the observed linear structure in real network traffic.

Extended Abstract

This work begins with a particular way of representing flow behaviors as vectors. The representation is quite simple. For each feature of a flow, we represent that aspect of the flow's behavior as a *feature-frequency vector*: a vector having a dimension for each possible value of the feature and whose coordinates are the relative frequency of values. For example, the vector for the distribution of packet sizes of a flow with four 40-byte and two 145-bytes packets is

$$\text{size} = \frac{1}{4+2}(4\mathbf{e}_{40} + 2\mathbf{e}_{145}). \tag{1}$$

Different aspects of flow behavior can be represented in this way, and these representations can be combined by taking the direct sum of their representation vectors:

$$\text{flow} = \text{size} \oplus \text{ival} \oplus \text{type} \oplus \text{port} \oplus \text{pkts}. \tag{2}$$

The features here are packet size and inter-packet interval distributions, IP protocol type, source and destination port numbers, and packet count. The behavior of a feature across a collection of flows can be expressed as a matrix where each row represents a flow:

$$\text{Size} = [\ \text{size}_1\ ;\ \cdots\ ;\ \text{size}_m\ ]. \tag{3}$$

The overall behavior of the collection of flows then becomes a concatenation of these feature matrices:

$$X = [\ \text{Size Ival Type Port Pkts}\ ]. \tag{4}$$

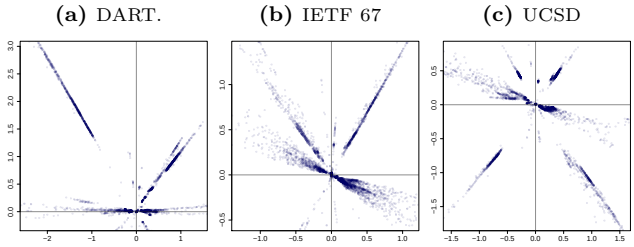

Figure 1: Scatter plots of the two most significant SVD dimensions of the feature-frequency representations of traffic samples from the six network traffic traces analyzed in our experiments.

When traffic traces are represented like this, a very curious thing happens: the resulting matrices exhibit a great deal of linear structure. Specifically, flow behaviors tend to lie near the union of a small set of low-rank subspaces. Figure 1 shows this structure visually. These scatter plots show the first two most significant dimensions of the behavior matrix after reduction via singular value decomposition (SVD) and projection onto the unit-sum hyperplane.

To explain this linear structure, we hypothesize that the behavior distribution for most flows is a mixture of a small set of "basic behaviors." Moreover, only even smaller subsets of these basic behaviors are typically combined with each other. Under these assumptions, we can express the distribution of each flow's behaviors as a finite mixture model [1]:

$$q_i(x) = \sum_{j=1}^{r} w_{ij} p_j(x). \tag{5}$$

Here q_i and p_j are probability density functions, and w_{ij} are nonnegative weights, summing to unity for each i. Equation 5 is expressed succinctly as matrix multiplication. Writing $Q_{ik} = q_i(k)$, $W_{ij} = w_{ij}$, and $P_{jk} = p_j(k)$, we have:

$$Q = WP. \tag{6}$$

The number of basic behaviors, r, is the maximum possible rank of the feature distribution matrix, Q. Moreover, we can partition the rows of P into classes such that w_{ij_1} and w_{ij_2} are both non-zero only if j_1 and j_2 are in the same class. Thus, each row of Q is associated with exactly one class, and all the points associated with a class lie in the subspace spanned by its associated rows in P.

This model explains the structures in Figure 1. Points along the same low-rank structure are in the same class. A structure is "generated" by a small set of vertices: points belonging to a structure are near the hull of its vertices. This is only one possible hypothesis that fits the data. Like any hypothesis, it must be tested. Our prediction technique, aside

Trace	Year	Type	Network
DARTMOUTH	2003	campus	Dartmouth College
IETF 60	2004	conference	IETF hotel
IETF 67	2006	conference	IETF hotel
SIGCOMM 2001	2001	conference	SIGCOMM hotel
SIGCOMM 2004	2004	conference	SIGCOMM hotel
UCSD	2007	campus	UCSD engineering

Table 1: Traffic traces used for analysis and experiments.

from providing a practical application, serves as a hypothesis test: we try to recover the matrices W and P from our noisy and imperfect observations of Q and use the recovered model to predict real flow behaviors. If the recovered model can make accurate predictions, this provides evidence that our model and hypothesis approximate reality.

From training data we recover estimates, W^* and P^*, of the factors in Equation 6. To detect the low-rank linear structures, we use Ma *et al.*'s algorithm for segmenting multivariate data into subspaces using lossy data coding and compression [2]. Then we determine the hull points of each linear structure using nonnegative matrix factorization (NMF) [3, 4]: if Q_c is a sub-matrix of rows in the same structure class, we want to find nonnegative matrices, W_c and P_c, such that $Q_c \approx W_c P_c$. Our reconstructed P^* is a vertical concatenation of these P_c matrices, while W^* is a row-permutation of the direct sum of W_c matrices. We use Kim and Park's alternating non-negative least squares algorithm [4] for rapid initial convergence, but refine the result using Lee and Seung's Euclidean algorithm [3]. Good prediction performance requires special initialization of the NMF algorithms, using new techniques that we lack room to detail here.

To predict flow behavior, we separate flow features into those observed and those to be predicted:

$$X_\mathrm{o} = [\ \mathrm{Size}_\mathrm{init}\ \mathrm{Ival}_\mathrm{init}\ \mathrm{Type}\ \mathrm{Port}\ ], \quad (7)$$

$$X_\mathrm{p} = [\ \mathrm{Size}_\mathrm{rest}\ \mathrm{Ival}_\mathrm{rest}\ \mathrm{Pkts}\ ]. \quad (8)$$

$\mathrm{Size}_\mathrm{init}$ is the packet size matrix for the first five packets, while $\mathrm{Size}_\mathrm{rest}$ is the matrix for the remainder of the packets, and similarly for inter-packet intervals. From an observation matrix, X_o, and the recovered model parameters, P^*, we can make predictions about X_p. Let $P^* = [P^*_\mathrm{o}\ P^*_\mathrm{p}]$ be the recovered model parameter matrix with separated observable and predictable features. From an observation matrix for test data, X_o, we estimate the matrix of weights by minimizing the squared Frobenius error:

$$W^* = \mathrm{argmin}_W \|X_\mathrm{o} - W P^*_\mathrm{o}\|^2_\mathrm{frob} \quad (9)$$

with the constraint that W be nonnegative. We can estimate the underlying feature distributions for the flows:

$$Q^* = W^* P^*. \quad (10)$$

The "predictable" portion, Q^*_p, contains predictions of packet size distribution, inter-packet interval distribution and distribution of packet counts for each flow. To evaluate the quality of these predictions, we compare the distributions in Q^*_p to the matrix, X_p, of actual test flow behaviors.

For our experiments, we use randomly sampled traffic from six network traces. The traces are freely available from the CRAWDAD trace repository [5]. Details of the traces are shown in Table 1. We randomly sampled 5000 flows from each trace for training and another 5000 flows for testing. The results are shown in Figure 2. The left panel shows accuracy rates for predicting packet size and inter-packet interval distributions. Since no previous work attempts to either model individual flow behavior or predict flow behavior from initial observations, we compare our prediction technique to two simple and obvious approaches: predicting the already observed behavior of each flow, and predicting the average behavior of the training trace. Accuracy is computed by comparing the distribution of Kolmogorov-Smirnov (K-S) test p-values to an empirical ideal distribution of K-S p-values, taking the maximum deviation from the ideal as the error rate. The right-top panel shows the accuracy rate for predicting which flow will have more packets between random pairs of flows. Choosing randomly gives 50% accuracy, which we exceed significantly on all traces.

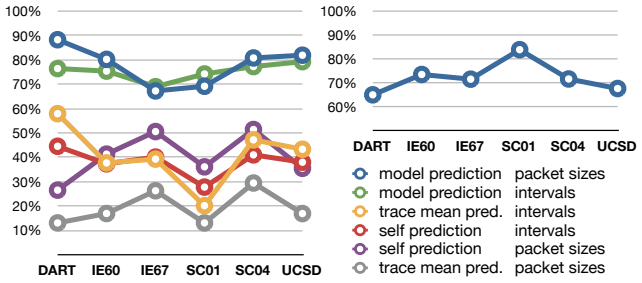

Figure 2: Accuracy rates of various methods of predicting flow behavior from five initial packets across six data sets.

Our method does not yield perfect predictions, but the nondeterministic nature of flow behavior implies that it is impossible to achieve perfect prediction. Moreover, we do not know what the inherent upper limit on prediction quality is. No prior work has provided detailed statistical models of individual flow behaviors, or attempted to predict individual flow behavior from initial packets. The fact that this technique can accurately predict flow behavior from so few initial packet observations is evidence that our hypothesized mixture model for flow behavior has merit. With improvements in the algorithms used to recover the model parameters, we are confident that even better prediction accuracy can be achieved. Furthermore, the same model can be applied to traffic classification from flow behavior, and to generation of realistic synthetic network traffic from collections of trace data. These applications, however useful, are merely pleasant side effects of the real breakthrough of this work: a realistic, detailed statistical model for individual flow behaviors across whole networks.

REFERENCES

[1] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley and Sons, New York NY, USA, 2000.

[2] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9), September 2007.

[3] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing*, 13, 2001.

[4] H. Kim and H. Park. Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method. *SIAM Journal in Matrix Analysis and Applications*, 30(2), May 2008.

[5] J. Yeo, D. Kotz, and T. Henderson. CRAWDAD: a community resource for archiving wireless data at Dartmouth. *SIGCOMM Computer Communication Review*, 36(2), April 2006.