Multi-Temporal Analysis and Scaling Relations of 100,000,000,000 Network Packets

Jeremy Kepner1,2,3, Chad Meiners4, Chansup Byun1, Sarah McGuire4, Timothy Davis5, William Arcand1, Jonathan Bernays4, David Bestor1, William Bergeron1, Vijay Gadepally1,2, Raul Harnasch1, Matthew Hubbell1, Micheal Houle1, Micheal Jones1, Andrew Kirby1, Anna Klein1, Lauren Milechin6, Julie Mullin1, Andrew Prout3, Albert Reuther3, Antonio Rosa1, Siddharth Sansi1, Doug Stetson4, Adam Tse4, Charles Yee1, Peter Michaleas1

1MIT Lincoln Laboratory Supercomputing Center, 2MIT Computer Science & AI Laboratory, 3MIT Mathematics Department, 4MIT Lincoln Laboratory Cyber Operations & Analysis Technology Group, 5Texas A&M, 6MIT Dept. of Earth, Atmospheric, & Planetary Sciences

Abstract—Our society has never been more dependent on computer networks. Effective utilization of networks requires a detailed understanding of the normal background behaviors of network traffic. Large-scale measurements of networks are computationally challenging. Building on prior work in interactive supercomputing and GraphBLAS hypersparse hierarchical traffic matrices, we have developed an efficient method for computing a wide variety of streaming network quantities on diverse time scales. Applying these methods to 100,000,000,000 anonymized source-destination pairs collected at a network gateway reveals many previously unobserved scaling relationships. These observations provide new insights into normal network background traffic that could be used for anomaly detection, AI feature engineering, and testing theoretical models of streaming networks.

Index Terms—Internet modeling, packet capture, streaming graphs, power-law networks, hypersparse matrices

I. INTRODUCTION

Global usage of the Internet is expected to exceed 5 billion people in 2023 [1]. Accordingly, cyberspace is a frontier as worthy of scientific study as land, sea, air, and space were during past eras of exploration [2]–[6]. Deepening our scientific understanding of cyberspace is expected to yield correspondingly equivalent societal benefits [7]–[9]. More pragmatically, effective utilization of cyberspace requires a detailed understanding of its principal characteristics: network traffic [10]. Figure 1 illustrates essential quantities found in all streaming dynamic networks. These quantities are derivable from the source and destinations in packet headers that are the foundational transactional unit of the Internet [11]. Precise measurements of these quantities on networks as vast as the Internet are computationally challenging [12]–[14]. Prior work has developed massively parallel data processing capabilities that can effectively use thousands of processors to rapidly process billions of packets in a few hours [15]–[20]. Using these capabilities, 50,000,000,000 anonymized packets from the largest publicly available collections of Internet packet data were analyzed, demonstrating the ubiquity of the Zipf-Mandelbrot distribution and the importance of the size of the packet sampling window \( N_V \). These data consisted of four 1-hour and two 2-day collections from Internet “trunk” lines, providing a high-level view of their portions of the Internet spanning years and continents [21]–[26]. These results underscore the need for processing longer time periods from additional vantage points.

Network gateways are a common view into the Internet, are widely monitored, and represent a natural observation point of network traffic. Using such a gateway we were able to collect trillions of packet headers over several years. Processing such a large volume of data requires additional computational innovations. The incorporation of GraphBLAS hypersparse hierarchical traffic matrices has enabled the processing of hundreds of billions of packets in minutes [27]–[30]. This paper presents an initial analysis of one month of anonymized source-destination pairs derived from approximately 100,000,000,000 packets. The contiguous nature of these data allows the exploration of a wide range of packet windows from \( N_V = 2^{17} \) (seconds) to...
\( N_V = 2^{27} \) (hours), providing a unique view into how network quantities depend upon time. These observations provide new insights into normal network background traffic that could be used for anomaly detection, AI feature engineering, polystore index learning, and testing theoretical models of streaming networks [31]–[33].

The outline of the rest of the paper is as follows. First, the network quantities and their distributions are defined in terms of traffic matrices. Second, multi-temporal analysis of the network quantities and their distributions are defined in networks [31]–[33].

The network quantities depicted in Figure 1 are computable from origin-destination matrices that are widely used to represent network traffic [34]–[37]. It is common to filter the packets down to a valid set for any particular analysis. Such filters may limit particular sources, destinations, protocols, and time windows. To reduce statistical fluctuations, the streaming data should be partitioned so that for any chosen time window all data sets have the same number of valid packets [20]. At a given time \( t \), \( N_V \) consecutive valid packets are aggregated from the traffic into a sparse matrix \( A_t \), where \( A_t(i, j) \) is the number of valid packets between the source \( i \) and destination \( j \). The sum of all the entries in \( A_t \) is equal to \( N_V \)

\[
\sum_{i,j} A_t(i, j) = N_V \tag{1}
\]

All the network quantities depicted in Figure 1 can be readily computed from \( A_t \) using the formulas listed in Table I.

Each network quantity will produce a distribution of values whose magnitude is often called the degree \( d \). The corresponding histogram of the network quantity is denoted by \( n_d \). The largest observed value in the distribution is denoted \( d_{\text{max}} \). The normalization factor of the distribution is given by

\[
\sum_d n_d(d) \tag{2}
\]

with corresponding probability

\[
p_d = n_d(d) / \sum_d n_d(d) \tag{3}
\]

and cumulative probability

\[
P_d = \sum_{i=1,d} p_d \tag{4}
\]

Because of the relatively large values of \( d \) observed, the measured probability at large \( d \) often exhibits large fluctuations. However, the cumulative probability lacks sufficient detail to see variations around specific values of \( d \), so it is typical to pool the differential cumulative probability with logarithmic bins in \( d \)

\[
D_d = P_d(d_i) - P_d(d_{i-1}) \tag{5}
\]

where \( d_i = 2^i \) [39]. All computed probability distributions use the same binary logarithmic binning to allow for consistent statistical comparison across data sets [39], [40]. The corresponding mean and standard deviation of \( D_d \) over many different consecutive values of \( t \) for a given data set are denoted \( D_t(d_i) \) and \( \sigma(d_i) \). Figure 2 provides an example distribution of external → internal source packets using a packet window of \( N_V = 2^{17} \). The means and standard deviations are computed using 1024 consecutive packet windows. The resulting distribution exhibits the power-law frequently observed in network data [41]–[47].

III. MULTI-TEMPORAL ANALYSIS

Network traffic is dynamic and exhibits varying behavior on a wide range of time scales. A given packet window size \( N_V \) will be sensitive to phenomena on its corresponding timescale. Determining how network quantities scale with \( N_V \) provides insight into the temporal behavior of network traffic. Efficient computation of network quantities on multiple time scales can be achieved by hierarchically aggregating data in different time windows [20]. Figure 3 illustrates a binary aggregation of different streaming traffic matrices. Computing each quantity at each hierarchy level eliminates redundant computations that would be performed if each packet window was computed separately. Hierarchy also ensures that most computations are performed on smaller matrices residing in faster memory. Correlations among the matrices mean that adding two matrices each with \( N_V \) entries results in a matrix with fewer than \( 2N_V \) entries, reducing the relative number of operations as the matrices grow.

| Aggregate Property | Summation Notation | Matrix Notation |
|---------------------|-------------------|----------------|
| Valid packets \( N_V \) | \( \sum_t \sum_j A_t(i, j) \) 1^T A_t \| 0 | \( \sum_t 1^T A_t \) |
| Unique links | \( \sum_t |A_t(i, j)|_0 \) 1^T A_t \| 0 | \( \sum_t |A_t|_0 \) |
| Link packets from \( i \) to \( j \) | \( A_t(i, j) \) | \( A_t \) |
| Max link packets \( (d_{\text{max}}) \) | \( \max_i A_t(i, j) \) \( \max(A_t) \) | \( \max(A_t) \) |
| Unique sources | \( \sum_i |A_t(i, j)|_0 \) 1^T A_t \| 0 | \( \sum_i |A_t|_0 \) |
| Packets from source \( i \) | \( A_t(i, j) \) \( A_t \) | \( A_t \) |
| Max source packets \( (d_{\text{max}}) \) | \( \max_j A_t(i, j) \) max(A_t) | \( \max(A_t) \) |
| Source fan-out from \( i \) | \( |A_t(i, j)|_0 \) | \( |A_t|_0 \) |
| Max source fan-out \( (d_{\text{max}}) \) | \( \max_j |A_t(i, j)|_0 \) max(|A_t|_0) | \( \max(|A_t|_0) \) |
| Unique destinations | \( \sum_j |A_t(i, j)|_0 \) 1^T A_t \| 0 | \( \sum_j 1^T A_t \) |
| Destination packets to \( j \) | \( A_t(i, j) \) 1^T A_t \| 0 | \( 1^T A_t \) |
| Max destination packets \( (d_{\text{max}}) \) | \( \max_i A_t(i, j) \) max(1^T A_t) | \( \max(1^T A_t) \) |
| Destination fan-in to \( j \) | \( |A_t(i, j)|_0 \) 1 \( A_t \) | \( 1^T A_t \) |
| Max destination fan-in \( (d_{\text{max}}) \) | \( \max_j |A_t(i, j)|_0 \) max(1^T A_t) | \( \max(1^T A_t) \) |
Fig. 2. External → internal source packet degree distribution. Differential cumulative probability (normalized histogram) of the number (degree) of source packets from each source using logarithmic bins $d_i = 2^i$. Circles represent the averages of 1024 packet windows each with $N_V = 2^{17}$. Error bars represent one standard deviation. Sources sending out a single packet are denoted “leaf nodes”. The source with the largest number of packets $d_{\text{max}}$ is referred to as the “supernode”.

One of the important capabilities of the SuiteSparse GraphBLAS library is support for hypersparse matrices where the number of nonzero entries is significantly less than either dimensions of the matrix. If the packet source and destination identifiers are drawn from a large numeric range, such as those used in the Internet protocol, then a hypersparse representation of $A_t$ eliminates the need to keep track of additional indices and can significantly accelerate the computations [30].

Network gateways are a common view into the Internet, are widely monitored, and represent a natural observation point on network traffic. Using such a gateway we were able to collect trillions of packets over several years. The traffic matrix of a gateway can be partitioned into four quadrants (see Figure 4). These quadrants represent different flows between nodes internal and external to the gateway. A gateway will typically see the external → internal and internal → external traffic.

IV. RESULTS

Several trillion packet headers have been collected at a gateway over several years. This work focuses on analyzing one month of anonymized source-destination pairs derived from approximately 100,000,000,000 packet headers. The

Fig. 4. Gateway network traffic matrix. The traffic matrix of a gateway can be divided into quadrants separating internal and external traffic. The gateway will observe the external → internal traffic (upper left) and the internal → external traffic (lower right).

Fig. 5. (top) Unique external → internal source fraction. Average total number of unique sources in a packet window of width $N_V$ measured at each time over a month. (bottom) Normalized external → internal unique source fraction. Scaling (top) data by $(N_V/2^{17})^{2/5}$ aligns the different packet windows, indicating that the number of uniques sources is proportional to $N_V^{-2/5} = N_V^{2/5}$ (see Table II).
TABLE II
APPROXIMATE SCALING RELATIONS.

|                      | Internal → External (Average) | Internal → External (Standard Deviation) | External → Internal (Average) | External → Internal (Standard Deviation) |
|----------------------|-------------------------------|------------------------------------------|-----------------------------|------------------------------------------|
| Unique links         | 2 x $N_V^{1/2}$              | 14 x $N_V^{1/4}$                         | 11 x $N_V^{1/2}$           |                                          |
| Max link packets ($d_{\text{max}}$) | 0.03 x $N_V^{-1}$           | 0.03 x $N_V^{-1}$                         | 0.1 x $N_V^{-1}$           | 0.1 x $N_V^{-1}$                    |
| Link packets norm ($\Sigma_p(d)$) | 2 x $N_V^{1/2}$              | 10 x $N_V^{1/4}$                         | 0.01 x $N_V^{1/6}$        | 11 x $N_V^{1/2}$                     |
| Unique sources       | $5 x N_V^{-1/5}$            | 8 x $N_V^{-1}$                            | 12 x $N_V^{-3/5}$         | 0.2 x $N_V^{-1/3}$                    |
| Max source packets ($d_{\text{max}}$) | 0.15 x $N_V^{-1}$           | 0.03 x $N_V^{-1}$                         | 0.1 x $N_V^{-1}$           | 0.1 x $N_V^{-1}$                    |
| Source packets norm ($\Sigma_d(d)$) | $6 x N_V^{-1/5}$             | 8 x $N_V^{-1}$                            | 1 x $N_V^{-3/5}$          | 8 x $N_V^{-1/3}$                     |
| Max source fan-out ($d_{\text{max}}$) | $3 x N_V^{-3/7}$             | 12 x $N_V^{-5/6}$                        | 0.03 x $N_V^{-5/6}$       | 11 x $N_V^{-1/2}$                    |
| Source fan-out norm ($\Sigma_f(d)$) | $5 x N_V^{-1/5}$             | $9 x N_V^{-1}$                            | $1 x N_V^{-3/5}$          | $10 x N_V^{-1/3}$                    |
| Unique destinations  | 2 x $N_V^{1/2}$              | 7 x $N_V^{1/4}$                           |                              |                                          |
| Max destination packets ($d_{\text{max}}$) | 0.04 x $N_V^{-1}$           | 0.03 x $N_V^{-1}$                         | 0.3 x $N_V^{-1}$          | 2 x $N_V^{3/4}$                     |
| Destination packets norm ($\Sigma_d(d)$) | 2 x $N_V^{1/2}$             | 12 x $N_V^{1/5}$                         | 3 x $N_V^{3/5}$          | 35 x $N_V^{2/5}$                     |
| Max destination fan-in ($d_{\text{max}}$) | 0.03 x $N_V^{-2/5}$         | 0.05 x $N_V^{-1/3}$                      | 4 x $N_V^{-2/7}$         | 9 x $N_V^{-1/3}$                    |
| Destination fan-in norm ($\Sigma_f(d)$) | 2 x $N_V^{1/2}$             | 12 x $N_V^{1/5}$                         | 3 x $N_V^{3/5}$          | 35 x $N_V^{2/5}$                     |

contiguous nature of these data allows the exploration of a wide range of packet windows from $N_V = 2^{17}$ (seconds) to $N_V = 2^{27}$ (hours), providing a unique view into how network quantities depend upon time. Figure 5 (top) shows the average total number of unique sources in a packet window of width $N_V$ measured at each time over a month normalized by $N_V$. Figure 5 (bottom) is the result of scaling the data by $(N_V/2^{17})^{2/5}$ to align the different packet windows, indicating that the number of unique sources is proportional to $N_V^{-1/2} = N_V^{3/5}$. Performing a similar analysis across many network quantities produced the scaling relations shown in Table II. These results reveal a strong dependence on these quantities as function of the packet window size $N_V$. To our knowledge, these scaling relations have not been previously reported and represent a new view on the background behavior of network traffic.

V. SIMPLE TOPOLOGY ANALYSIS

Power-law network data are a result of complex network topologies that are an ongoing area of investigation [48], [49]. Bounds on the underlying network topologies can be derived by analyzing several simple topologies. Figure 6 shows a schematic of networks and gateway traffic matrices for four simple topologies whose $N_V$ scaling behavior can be readily computed. Isolated links are source-destination pairs that have only one packet. Single link implies all traffic flows between one internal and one external node. An internal supernode has all traffic occurring between a single internal node and many external nodes. An external supernode has all traffic occurring between a single external node and many internal nodes. The resulting scaling relationships are shown in Table III. In all cases the simple traffic topologies produce relationships that scale as $N_V^0$ (constant) or $N_V^{1/2}$ (linear). The observed relationships in Table II are between these two extremes consistent with observed power-law topologies. In particular, the observed maximum link, source, and destination packets are consistent with the single link model, suggesting that a fraction of the traffic is on a single link perhaps combining an internal supernode and an external supernode.

VI. CONCLUSIONS AND FUTURE WORK

We have developed an efficient method for computing a wide variety of streaming network quantities on diverse time scales. Applying these methods to 100,000,000,000 anonymized source-destination pairs collected at a network gateway reveals many previously unobserved scaling relationships as a function of the packet window $N_V$. These observations provide new insights into normal network background traffic. While the details of these scaling relationships are likely to be specific to the gateway, the number and variety of relationships suggests that scaling relationships could be observed at other network vantage points. The observed relationships provide simple, low-dimensional models of the network traffic that could be used for anomaly detection. These relationships can also assist with feature engineering for the development of AI based traffic categorization. Deeper theoretical models can also be constrained and tested, and the observed relationships can provide a useful target for theoreticians. Future work will expand this analysis to the full several trillion packet data set and other data sets collected at different vantage points, apply these results to anomaly detection problems, and test theoretical models of dynamic streaming networks.

ACKNOWLEDGMENTS

The authors wish to acknowledge the following individuals for their contributions and support: Bob Bond, K Claffy, Cary Conrad, David Clark, Alan Edelman, Jeff Gottschalk, Tim Kraska, Charles Leiserson, Dave Martinez, Mimi McClure, Steve Rejto, Marc Zissman.
Fig. 6. Simple network traffic topologies and traffic matrices. Isolated links are source-destination pairs that have only one packet. Single link implies all traffic flows between one internal and one external node. Internal supernode has all traffic occurring between a single internal node and many external nodes. External supernode has all traffic occurring between a single external node and many internal nodes.

| TABLE III | SIMPLE TOPOLOGY SCALING RELATIONS |
|-----------|----------------------------------|
| For balanced traffic with equal numbers of packets flowing in each direction, the simple traffic topologies produce relationships that scale as $N^V_0$ (constant) or $N^V_1$ (linear). |

| Isolated Links | Internal External | Internal External | Internal External | Internal External | Internal External | External External |
|----------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Unique links   | 0.5 x N^1_1       | 0.5 x N^1_2       | 1.0 x N^0_1       | 1.0 x N^0_2       | 0.5 x N^1_3       | 0.5 x N^1_4 |
| Max link packets ($d_{max}$) | 1.0 x N^2_3       | 1.0 x N^2_4       | 0.5 x N^1_5       | 0.5 x N^1_6       | 1.0 x N^0_7       | 1.0 x N^0_8 |
| Link packets norm ($L_{p}(d)$) | 0.5 x N^1_9       | 0.5 x N^1_10      | 1.0 x N^0_11      | 1.0 x N^0_12      | 0.5 x N^1_13      | 0.5 x N^1_14 |
| Unique sources | 0.5 x N^1_16      | 0.5 x N^1_17      | 1.0 x N^0_18      | 1.0 x N^0_19      | 0.5 x N^1_20      | 0.5 x N^1_21 |
| Max source packets ($d_{max}$) | 1.0 x N^2_22      | 1.0 x N^2_23      | 0.5 x N^1_24      | 0.5 x N^1_25      | 1.0 x N^0_26      | 1.0 x N^0_27 |
| Source packets norm ($L_{s}(d)$) | 0.5 x N^1_28      | 0.5 x N^1_29      | 1.0 x N^0_30      | 1.0 x N^0_31      | 0.5 x N^1_32      | 0.5 x N^1_33 |
| Max source fan-out ($d_{max}$) | 1.0 x N^2_34      | 1.0 x N^2_35      | 0.5 x N^1_36      | 0.5 x N^1_37      | 1.0 x N^0_38      | 1.0 x N^0_39 |
| Source fan-out norm ($L_{s}(d)$) | 0.5 x N^1_40      | 0.5 x N^1_41      | 1.0 x N^0_42      | 1.0 x N^0_43      | 0.5 x N^1_44      | 0.5 x N^1_45 |
| Unique destinations | 0.5 x N^1_46      | 0.5 x N^1_47      | 1.0 x N^0_48      | 1.0 x N^0_49      | 0.5 x N^1_50      | 0.5 x N^1_51 |
| Max destination packets ($d_{max}$) | 1.0 x N^2_52      | 1.0 x N^2_53      | 0.5 x N^1_54      | 0.5 x N^1_55      | 1.0 x N^0_56      | 1.0 x N^0_57 |
| Destination packets norm ($L_{d}(d)$) | 0.5 x N^1_58      | 0.5 x N^1_59      | 1.0 x N^0_60      | 1.0 x N^0_61      | 0.5 x N^1_62      | 0.5 x N^1_63 |
| Max destination fan-in ($d_{max}$) | 1.0 x N^2_64      | 1.0 x N^2_65      | 0.5 x N^1_66      | 0.5 x N^1_67      | 1.0 x N^0_68      | 1.0 x N^0_69 |
| Destination fan-in norm ($L_{d}(d)$) | 0.5 x N^1_70      | 0.5 x N^1_71      | 1.0 x N^0_72      | 1.0 x N^0_73      | 0.5 x N^1_74      | 0.5 x N^1_75 |

REFERENCES

[1] “Cisco Visual Networking Index: Forecast and Trends, 20182023,” https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.
[2] V. Bush, “Science The Endless Frontier.” Report to the President of the United States, 1945.
[3] K. Claffy, “Internet tomography,” Nature, Web Matter, 1999.
[4] R. Prasad, C. Dovrolis, M. Murray, and K. Claffy, “Bandwidth estimation: metrics, measurement techniques, and tools,” IEEE network, vol. 17, no. 6, pp. 27–35, 2003.
[5] H. Kim, K. C. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, “Internet traffic classification demystified: myths, caveats, and the best practices,” in Proceedings of the 2008 ACM CoNEXT conference, p. 11, ACM, 2008.
[6] M. Boguna, D. Krioukov, and K. C. Claffy, “Navigability of complex networks,” Nature Physics, vol. 5, no. 1, p. 74, 2009.
[7] K. Claffy, “Measuring the internet,” IEEE Internet Computing, vol. 4, no. 1, pp. 73–75, 2000.
[8] B. Li, J. Springer, G. Bebb, and M. H. Gunes, “A survey of network flow applications,” Journal of Network and Computer Applications, vol. 36, no. 2, pp. 567–581, 2013.
[9] M. Rabinovich and M. Allman, “Measuring the internet,” IEEE Internet Computing, vol. 20, no. 4, pp. 6–8, 2016.
[10] k. Claffy and D. Clark, “Workshop on internet economics (wie 2019) report,” SIGCOMM Comput. Commun. Rev., vol. 50, p. 5359, May 2020.
[11] D. Huang, A. Chowdhary, and S. Pisharody, Software-Defined networking and security: from theory to practice. CRC Press, 2018.
[12] A. Lumsdaine, D. Gregor, B. Hendrickson, and J. Berry, “Challenges in parallel graph processing,” Parallel Processing Letters, vol. 17, no. 01, pp. 5–20, 2007.
[13] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” SIAM review, vol. 51, no. 3, pp. 455–500, 2009.
[14] M. Hilbert and P. Lópeiz, “The world’s technological capacity to store, communicate, and compute information,” Science, p. 1200970, 2011.
[15] J. Kepner, Parallel MATLAB for Multicore and Multinode Computers. SIAM, 2009.
[16] J. Kepner and J. Gilbert, Graph algorithms in the language of linear algebra. SIAM, 2011.
