

New Phenomena in Large-Scale Internet Traffic

CONTENTS

1.1	Introduction	2
1.2	Methodology	2
1.2.1	MAWI and CAIDA Internet Traffic Collection	3
1.2.2	Network Quantities from Matrices	5
1.2.3	Memory and Computation Requirements	8
1.3	Internet Traffic Modeling	10
1.3.1	Logarithmic Pooling	12
1.3.2	Modified Zipf–Mandelbrot Model	13
1.3.3	Nonlinear Model Fitting	16
1.4	Results	18
1.5	Discussion	20
1.6	Conclusions	21

Jeremy Kepner¹, Kenjiro Cho², KC Claffy³, Vijay Gadepally¹, Sarah McGuire¹, Lauren Milechin⁴, William Arcand¹, David Bestor¹, William Bergeron¹, Chansup Byun¹, Matthew Hubbell¹, Michael Houle¹, Michael Jones¹, Andrew Prout¹, Albert Reuther¹, Antonio Rosa¹, Siddharth Samsi¹, Charles Yee¹, Peter Michaleas¹

¹*MIT Lincoln Laboratory*, ²*Research Laboratory, Internet Initiative Japan, Inc.*,
³*UCSD Center for Applied Internet Data Analysis*, ⁴*MIT Dept. of Earth, Atmospheric
and Planetary Sciences*

The INTERNET is transforming our society, necessitating a quantitative understanding of Internet traffic. Our team collects and curates the largest publicly available Internet traffic data sets. An analysis of 50 billion packets using 10,000 processors in the MIT SuperCloud reveals a new phenomenon: the importance

This material is based, in part, upon the work supported by the NSF under grants DMS-1312831, CCF-1533644, and CNS-1513283, DHS cooperative agreement FA8750-18-2-0049, and ASD(R&E) under contract FA8702-15-D-0001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, DHS, or ASD(R&E).

of otherwise unseen leaf nodes and isolated links in Internet traffic. Our analysis further shows that a two-parameter modified Zipf–Mandelbrot distribution accurately describes a wide variety of source/destination statistics on moving sample windows ranging from 100,000 to 100,000,000 packets over collections that span years and continents. The measured model parameters distinguish different network streams, and the model leaf parameter strongly correlates with the fraction of the traffic in different underlying network topologies.

1.1 INTRODUCTION

Our civilization is now dependent on the Internet, necessitating a scientific understanding of this virtual universe [49, 72]. The two largest efforts to capture, curate, and share Internet packet traffic data for scientific analysis are led by our team via the Widely Integrated Distributed Environment (WIDE) project [29] and the Center for Applied Internet Data Analysis (CAIDA) [30]. These data have been used for a wide variety of research projects, resulting in hundreds of peer-reviewed publications [86], ranging from characterizing the global state of Internet traffic, to specific studies on the prevalence of peer-to-peer file sharing traffic, to testing prototype software designed to stop the spread of Internet worms.

The stochastic network structure of Internet traffic is a core property of great interest to a wide range of Internet stakeholders [72] and network scientists [11]. Of particular interest is the probability distribution $p(d)$, where d is the degree (or count) of several network quantities, such as source packets, packets over a unique source-destination pair (or link), and destination packets collected over specified time intervals. Among the earliest and most widely cited results of virtual Internet topology analysis has been the observation that $p(d) \propto 1/d^\alpha$ with a model exponent $1 < \alpha < 3$ for large values of d [10, 4, 70] fit a range of network characteristics.

Many Internet models are based on the data obtained from crawling the network from a number of starting points [94]. These webcrawls naturally sample the supernodes of the network [23], and their resulting $p(d)$ are accurately fit at large values of d by single-parameter power-law models. However, as we will show, for our streaming samples of the Internet there are other topologies that contribute significant traffic. Characterizing a network by a single power-law exponent provides one view of Internet phenomena, but more accurate and complex models are required to understand the diverse topologies seen in streaming samples of the Internet. Improving model accuracy while also increasing model complexity requires overcoming a number of challenges, including acquisition of larger, rigorously collected data sets [107, 121]; the enormous computational cost of processing large network traffic graphs [80, 7, 110]; careful filtering, binning, and normalization of the data; and fitting of nonlinear models to the data.

1.2 METHODOLOGY

This work aims to improve model accuracy through several techniques. First, for over a decade we have scientifically collected and curated the largest publicly available

Internet packet traffic data sets and this work analyzes the very largest collections in our corpora containing 49.6 billion packets (Table 1.2). Second, utilizing recent innovations in interactive supercomputing [56, 102], matrix-based graph theory [63, 57], and big data mathematics (Figure 1.2) [58], we have developed a scalable Internet traffic processing pipeline that runs efficiently on more than 10,000 processors in the MIT SuperCloud [47]. This pipeline allows us, for the first time, to process our largest traffic collections as network traffic graphs. Third, since not all packets have both source and destination Internet Protocol version 4 (IPv4) addresses, the data have been filtered so that for any chosen time window all data sets have the same number of valid IPv4 packets, denoted N_V (Figure 1.1 and Eq. 1.2). All computed probability distributions also use the same binary logarithmic binning to allow for consistent statistical comparison across data sets (Eq. 1.4)[34, 11]. Fourth, to accurately model the data over the full range of d , we employ a modified Zipf–Mandelbrot distribution [82, 87, 104]

$$p(d; \alpha, \delta) \propto 1/(d + \delta)^\alpha \quad (1.1)$$

The inclusion of a second model offset parameter δ allows the model to accurately fit small values of d , in particular $d = 1$, which has the highest observed probability in these streaming data. The modified Zipf–Mandelbrot model is a special case of the more general saturation/cutoff models used to model a variety of network phenomena (Eq. 1.6) [34, 11]. Finally, nonlinear fitting techniques are used to achieve quality fits over the entire range of d (Eq. 1.20).

Throughout this chapter, we defer to the terminology of network science. Network operators use many similar terms with significant differences in meaning. We use network topology to refer to the graph-theoretic virtual topology of sources and destinations observed communicating and not the underlying physical topology of the Internet (Table 1.1)

1.2.1 MAWI and CAIDA Internet Traffic Collection

For the analysis in following sections, the data utilized are summarized in Table 1.2 with data from Tokyo coming from the MAWI Internet Traffic Collection and the data from Chicago from the CAIDA Internet Traffic Collection.

The Tokyo data sets are publicly available packet traces provided by the WIDE project (aka the MAWI traces). The WIDE project is a research consortium in Japan established in 1988 [29]. The members of the project include network engineers, researchers, university students, and industrial partners. The focus of WIDE is on the empirical study of the large-scale Internet. WIDE operates an Internet testbed both for commercial traffic and for conducting research experiments. These data have enabled quantitative analysis of Internet traffic spanning years illustrating trends such as the emergence of residential usage, peer-to-peer networks, probe scanning, and botnets [27, 20, 45]. The Tokyo data sets are publicly available packet traces provided by the WIDE project (aka the MAWI traces). The traces are collected from a 1 Gbps academic backbone connection in Japan. The 2015 and 2017 data sets are 48-hour-long traces captured during December 2–3, 2015, and April 12–13, 2017, in JST.

Table 1.1 Network Terminology Used by Computer Network Operators and Network Scientists. Throughout this Work, the Network Science Meanings are Employed

Term	Network Operations Meaning	Network Science Meaning
Network	The <i>physical</i> links, wires, routers, switches, and endpoints used to transmit data.	Any system that can be represented as a <i>graph</i> of connections (links/edges) among entities (nodes/vertices).
Topology	The <i>layout</i> of the physical network.	The specific <i>geometries</i> of a graph and its sub-graphs.
Stream	The <i>flow of data</i> over a specific physical communication link.	A time-ordered <i>sequence of pairs</i> of entities (nodes/vertices) representing distinction in time connections (links/edges) between entities.

The IP addresses appearing in the traces are anonymized using a prefix-preserving method [43].

The MAWI repository is an ongoing collection of Internet traffic traces, captured within the WIDE backbone network (AS2500) that connects Japanese universities and research institutes to the Internet. Each trace consists of captured packets observed from within WIDE and includes the packet headers of each packet along with the captured timestamp. Anonymized versions of the traces (with anonymized IP addresses and with transport layer payload removed) are made publicly available at <http://mawi.wide.ad.jp/>.

WIDE carries a variety of traffic including academic and commercial traffic. These data have enabled quantitative analysis of Internet traffic spanning years illustrating trends such as the emergence of residential usage, peer-to-peer networks, probe scanning, and botnets [27, 28, 45]. WIDE is mostly dominated by HTTP traffic, but is influenced by global anomalies. For example, Code Red, Blaster, and Sasser are worms that disrupted Internet traffic [5]. Of these, Sasser (2005) impacted MAWI traffic the most, accounting for two-thirds of packets at its peak. Conversely, the ICMP traffic surge in 2003 and the SYN Flood in 2012 were more local in nature, each revealing attacks on targets within WIDE that lasted several months.

CAIDA collects several different data types at geographically and topologically diverse locations and makes these data available to the research community to the extent possible while preserving the privacy of individuals and organizations who donate data or network access [30, 31]. CAIDA has (and had) monitoring locations in Internet service providers (ISPs) in the USA. CAIDA's passive traces data set contains traces collected from high-speed monitors on a commercial backbone link. The data collection started in April 2008 and is ongoing. These data are useful for research on the characteristics of Internet traffic, including application breakdown

(based on TCP/IP ports), security events, geographic and topological distribution, flow volume, and duration. For an overview of all traces, see the trace statistics page [85].

Collectively, our consortium has enabled the scientific analysis of Internet traffic, resulting in hundreds of peer-reviewed publications with over 30,000 citations [86]. These include early work on Internet threats such as the Code Red worm [90] and Slammer worm [88] and how quarantining might mitigate threats [91]. Subsequent work explored various techniques, such as dispersion, for measuring Internet capacity and bandwidth [40, 41, 100]. The next major area of research provided significant results on the dispersal of the Internet via the emergence of peer-to-peer networks [54, 53], edge devices [62], and corresponding denial-of-service attacks [89], which drove the need for new ways to categorize traffic [59]. The incorporation of network science and statistical physics concepts into the analysis of the Internet produced new results on the hyperbolic geometry of complex networks [67, 66] and sustaining the Internet with hyperbolic mapping [16, 17, 18]. Likewise, a new understanding also emerged on the identification of influential spreaders in complex networks [61], the relationship of popularity versus similarity in growing networks [96], and overall network cosmology [65]. More recent work has developed new ideas for Internet classification [35] and future data centric architectures [120].

Table 1.2 Network Traffic Packet Data Sets from MAWI (Tokyo data sets) and CAIDA (Chicago data sets) Collected at Different Times and Durations over Two Years

Location	Date	Duration	Bandwidth	Packets
Tokyo	2015 Dec 02	2 days	10^9 bits/second	17.0×10^9
Tokyo	2017 Apr 12	2 days	10^9 bits/second	16.8×10^9
Chicago A	2016 Jan 21	1 hour	10^{10} bits/second	2.0×10^9
Chicago A	2016 Feb 18	1 hour	10^{10} bits/second	2.0×10^9
Chicago A	2016 Mar 17	1 hour	10^{10} bits/second	1.8×10^9
Chicago A	2016 Apr 06	1 hour	10^{10} bits/second	1.8×10^9
Chicago B	2016 Jan 21	1 hour	10^{10} bits/second	2.3×10^9
Chicago B	2016 Feb 18	1 hour	10^{10} bits/second	1.7×10^9
Chicago B	2016 Mar 17	1 hour	10^{10} bits/second	2.0×10^9
Chicago B	2016 Apr 06	1 hour	10^{10} bits/second	2.1×10^9

1.2.2 Network Quantities from Matrices

In our analysis, the network traffic packet data are reduced to origin–destination traffic matrices. These matrices can be used to compute a wide range of network statistics useful in the analysis, monitoring, and control of the Internet. Such an analysis includes the temporal fluctuations of the supernodes [107] and inferring the presence of unobserved traffic [121, 13].

To create the matrices, at a given time t , N_V consecutive valid packets are aggregated from the traffic into a sparse matrix \mathbf{A}_t , where $\mathbf{A}_t(i, j)$ is the number of valid

Table 1.3 Aggregate Network Properties

Aggregate Property	Summation Notation	Matrix Notation
Valid packets N_V	$\sum_i \sum_j \mathbf{A}_t(i, j)$	$\mathbf{1}^\top \mathbf{A}_t \mathbf{1}$
Unique links	$\sum_i \sum_j \mathbf{A}_t(i, j) _0$	$\mathbf{1}^\top \mathbf{A}_t _0 \mathbf{1}$
Unique sources	$\sum_i \sum_j \mathbf{A}_t(i, j) _0$	$\mathbf{1}^\top \mathbf{A}_t \mathbf{1}_0$
Unique destinations	$\sum_j \sum_i \mathbf{A}_t(i, j) _0$	$ \mathbf{1}^\top \mathbf{A}_t _0 \mathbf{1}$

Formulas for computing aggregates from a sparse network image \mathbf{A}_t at time t in both summation and matrix notations. $\mathbf{1}$ is a column vector of all 1's, $^\top$ is the transpose operation, and $|\cdot|_0$ is the zero-norm that sets each nonzero value of its argument to 1 [55].

packets between the source i and destination j [92]. The sum of all the entries in \mathbf{A}_t is equal to N_V

$$\sum_{i,j} \mathbf{A}_t(i, j) = N_V \quad (1.2)$$

All the network quantities depicted in Figure 1.6a can be readily computed from \mathbf{A}_t as specified in Tables 1.3 and 1.4, including the number of unique sources and destinations, along with many other network statistics [107, 121, 110].

Table 1.4 Neural Network Image Convolution Filters

Network Quantity	Summation Notation	Matrix Notation
Source packets from i	$\sum_j \mathbf{A}_t(i, j)$	$\mathbf{A}_t \mathbf{1}$
Source fan-out from i	$\sum_j \mathbf{A}_t(i, j) _0$	$ \mathbf{A}_t _0 \mathbf{1}$
Link packets from i to j	$\mathbf{A}_t(i, j)$	\mathbf{A}_t
Destination fan-in to j	$\sum_i \mathbf{A}_t(i, j) _0$	$\mathbf{1}^\top \mathbf{A}_t$
Destination packets to j	$\sum_i \mathbf{A}_t(i, j)$	$\mathbf{1}^\top \mathbf{A}_t _0$

Different network quantities are extracted from a sparse traffic image \mathbf{A}_t at time t via convolution with different filters. Formulas for the filters are given in both summation and matrix notations. $\mathbf{1}$ is a column vector of all 1's, $^\top$ is the transpose operation, and $|\cdot|_0$ is the zero-norm that sets each nonzero value of its argument to 1 [55].

Figure 1.17a depicts the major topological structures in the network traffic. Isolated links are sources and destinations that each have only one connection (Table 1.5). The first, second, third, ... supernodes are the source or destination with the first, second, third, ... most links (Table 1.6). The core of a network can be defined in a variety of ways [105, 12]. In this work, the network core conveys the con-

cept of a collection of sources and destinations that are not isolated and are multiply connected. The core is defined as the collection of sources and destinations in which every source and destination has more than one connection. The core, as computed here, does not include the first five supernodes although only the first supernode is significant, and whether or not the other supernodes are included has minimal impact on the core in these data. The core leaves are sources and destinations that have only one connection to a core source or destination (Tables 1.7 and 1.8).

Table 1.5 Properties of Isolated Links

Network Quantity	Matrix Notation
Isolated links	$\mathbf{A}_t(i_1, j_1)$
Number of isolated link sources	$\mathbf{1}^\top \mathbf{A}_t(i_1, j_1) _0 \mathbf{1}$
Number of packets traversing isolated links	$\mathbf{1}^\top \mathbf{A}_t(i_1, j_1) \mathbf{1}$
Number of unique isolated links	$\mathbf{1}^\top \mathbf{A}_t(i_1, j_1) _0 \mathbf{1}$
Number of isolated link destinations	$ \mathbf{1}^\top \mathbf{A}_t(i_1, j_1) _0 \mathbf{1}$

Different characteristics related to isolated links are extracted from a sparse traffic image \mathbf{A}_t at time t . Formulas are in matrix notation. The set of sources that send to only one destination are $i_1 = \arg(\mathbf{d}_{\text{out}} = 1)$, and the set of destinations that receive from only one destination are $j_1 = \arg(\mathbf{d}_{\text{in}} = 1)$.

Table 1.6 Properties of Supernodes

Network Quantity	Matrix Notation
Supernode source leaves	$\mathbf{A}_t(i_1, k_{\text{max}})$
Supernode destination leaves	$\mathbf{A}_t(k_{\text{max}}, j_1)$
Number of supernode leaf sources	$\mathbf{1}^\top \mathbf{A}_t(i_1, k_{\text{max}}) _0 \mathbf{1}$
Number of packets traversing supernode leaves	$\mathbf{1}^\top \mathbf{A}_t(i_1, k_{\text{max}}) + \mathbf{A}_t(k_{\text{max}}, j_1) \mathbf{1}$
Number of unique supernode leaf links	$\mathbf{1}^\top \mathbf{A}_t(i_1, k_{\text{max}}) _0 + \mathbf{A}_t(k_{\text{max}}, j_1) _0 \mathbf{1}$
Number of supernode leaf destinations	$ \mathbf{1}^\top \mathbf{A}_t(k_{\text{max}}, j_1) _0 \mathbf{1}$

Different characteristics related to supernodes are extracted from a sparse traffic image \mathbf{A}_t at time t . Formulas are in matrix notation. The identity of the first supernode is given by $k_{\text{max}} = \arg\max(\mathbf{d}_{\text{out}} + \mathbf{d}_{\text{in}})$. The leaves of a supernode are those sources and destinations whose only connection is to the supernode.

An essential step for increasing the accuracy of the statistical measures of Internet traffic is using windows with the same number of valid packets N_V . For this analysis, a valid packet is defined as TCP over IPv4, which includes more than 95% of the data in the collection and eliminates a small amount of data that use other protocols or contain anomalies. Using packet windows with the same number of valid packets

Table 1.7 Properties of Network Core

Network Quantity	Matrix Notation
Core links	$\mathbf{A}_t(i_{\text{core}}, j_{\text{core}})$
Number of core sources	$\mathbf{1}^\top \mathbf{A}_t(i_{\text{core}}, j_{\text{core}}) _0 \mathbf{1}$
Number of core packets	$\mathbf{1}^\top \mathbf{A}_t(i_{\text{core}}, j_{\text{core}}) \mathbf{1}$
Number of unique core links	$\mathbf{1}^\top \mathbf{A}_t(i_{\text{core}}, j_{\text{core}}) _0 \mathbf{1}$
Number of core destinations	$ \mathbf{1}^\top \mathbf{A}_t(i_{\text{core}}, j_{\text{core}}) _0 \mathbf{1}$

Different characteristics related to the core are extracted from a sparse traffic image \mathbf{A}_t at time t . Formulas are in matrix notation. The set of sources that send to more than one destination, excluding the supernode(s), is $i_{\text{core}} = \arg(1 < \mathbf{d}_{\text{out}} < \mathbf{d}_{\text{out}}(k_{\text{max}}))$. The set of destinations that receive from more than one source, excluding the supernode(s), is $j_{\text{core}} = \arg(1 < \mathbf{d}_{\text{in}} < \mathbf{d}_{\text{in}}(k_{\text{max}}))$.

Table 1.8 Properties of Core Leaves

Network Quantity	Matrix Notation
Core source leaves	$\mathbf{A}_t(i_1, k_{\text{core}})$
Core destination leaves	$\mathbf{A}_t(k_{\text{core}}, j_1)$
Number of core leaf sources	$\mathbf{1}^\top \mathbf{A}_t(i_1, k_{\text{core}}) _0 \mathbf{1}$
Number of core leaf packets	$\mathbf{1}^\top \mathbf{A}_t(i_1, k_{\text{core}}) + \mathbf{A}_t(k_{\text{core}}, j_1) \mathbf{1}$
Number of unique core leaf links	$\mathbf{1}^\top \mathbf{A}_t(i_1, k_{\text{core}}) _0 + \mathbf{A}_t(k_{\text{core}}, j_1) _0 \mathbf{1}$
Number of core leaf destination	$ \mathbf{1}^\top \mathbf{A}_t(k_{\text{core}}, j_1) _0 \mathbf{1}$

Different characteristics related to the core leaves are extracted from a sparse traffic image \mathbf{A}_t at time t . Formulas are in matrix notation. The core leaves are sources and destinations that have one connection to a core source or destination.

produces quantities that are consistent over a wide range from $N_V = 100,000$ to $N_V = 100,000,000$ (Figure 1.1).

1.2.3 Memory and Computation Requirements

Processing 50 billion Internet packets with a variety of algorithms presents numerous computational challenges. Dividing the data set into combinable units of approximately 100,000 consecutive packets made the analysis amenable to processing on a massively parallel supercomputer. The detailed architecture of the parallel processing system and its corresponding performance are described in [47]. The resulting processing pipeline was able to efficiently use over 10,000 processors on the MIT SuperCloud and was essential to this first-ever complete analysis of these data.

A key element of our analysis is the use of novel sparse matrix mathematics in

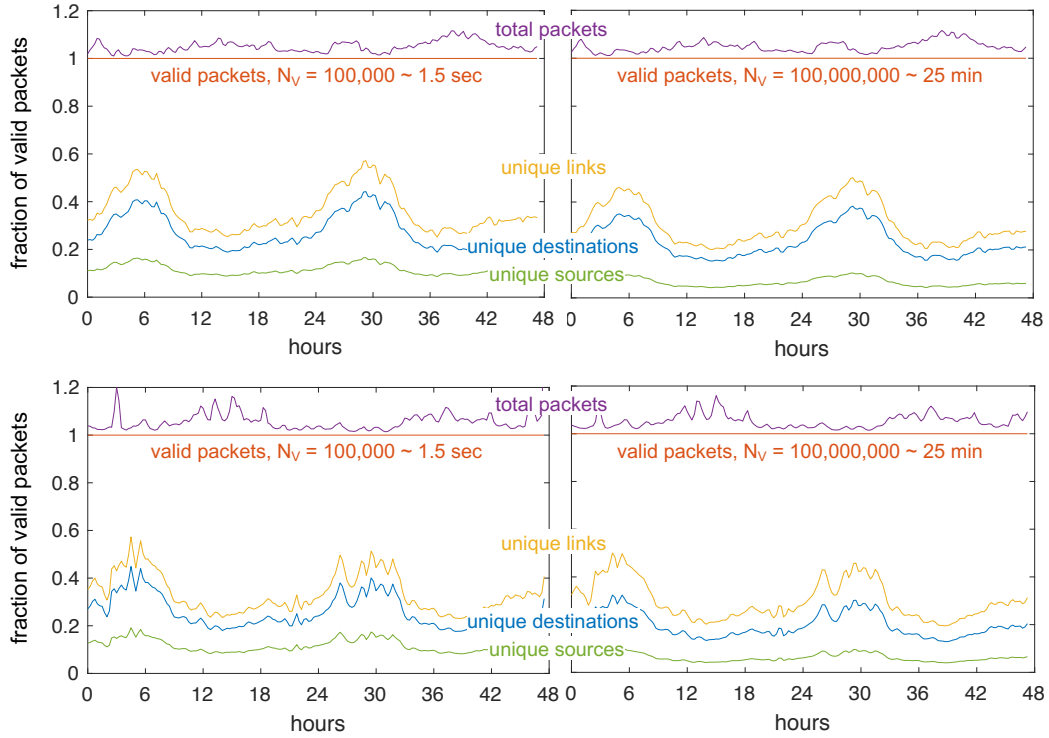


Figure 1.1 Valid packets. Analyzing packet windows with the same numbers of valid packets produces consistent fractions of unique links, unique destinations, and unique sources over a wide range of packet sizes for the Tokyo 2015 (a) and Tokyo 2017 (b) data sets. The plots show these fractions for moving packet windows of $N_V = 100,000$ packets (left) and $N_V = 100,000,000$ packets (right). The packet windows correspond to time windows of approximately 1.5 seconds and 25 minutes.

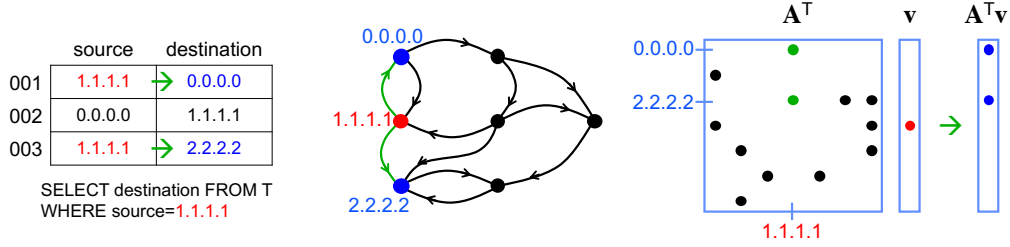


Figure 1.2 **Associative arrays.** (a) Tabular representation of raw network traffic and corresponding database query to find all records beginning with source 1.1.1.1. (b) Network graph highlighting nearest neighbors of source node 1.1.1.1. (c) Corresponding associative array representation of the network graph illustrating how the neighbors of source node 1.1.1.1 are computed with matrix vector multiplication.

concert with the MIT SuperCloud. Construction and analysis of network traffic matrices of the entire Internet address space have been considered impractical for its massive size [110]. Internet Protocol version 4 (IPv4) has 2^{32} unique addresses, but at any given collection point, only a fraction of these addresses will be observed. Exploiting this property to save memory can be accomplished by extending traditional sparse matrices so that new rows and columns can be added dynamically.

The algebra of associative arrays [58] and its corresponding implementation in the Dynamic Distributed Dimensional Data Model (D4M) software library (d4m.mit.edu) allows the row and columns of a sparse matrix to be any sortable value, in this case character string representations of the Internet addresses (Figure 1.2). Associative arrays extend sparse matrices to have database table properties with dynamically insertable and removable rows and columns that adjust as new data are added or subtracted to the matrix. Using these properties, the memory requirements of forming network traffic matrices can be reduced at the cost of increasing the required computation necessary to resort the rows and columns.

A network matrix A_t with $N_V = 100,000,000$ represented as an associative array typically requires 2 gigabytes of memory. A complete analysis of the statistics and topologies of A_t typically takes 10 minutes on a single MIT SuperCloud Intel Knights Landing processor core. Using increments of 100,000 packets means that this analysis is repeated over 500,000 times to process all 49.6 billion packets. Using 10,000 processors on the MIT SuperCloud shortens the runtime of one of these analyses to approximately 8 hours. The results presented within this chapter are products of a discovery process that required hundreds of such runs that would not have been possible without these computational resources. Fortunately, the utilization of these results by Internet stakeholders can be significantly accelerated by creating optimized embedded implementations that only compute the desired statistics and are not required to support a discovery process [76, 77].

1.3 INTERNET TRAFFIC MODELING

Quantitative measurements of the Internet [101] have provided Internet stakeholders information on the Internet since its inception. Early work has explored the early growth of the Internet [32], the distribution of packet arrival times [98], the power-law distribution of network outages [97], the self-similar behavior of traffic [69, 115, 114], formation processes of power-law networks [42, 83, 22, 113], and the topologies of Internet service providers [108]. Subsequent work has examined the technological properties of Internet topologies [73], the diameter of the Internet [70], applying rank index-based Zipf–Mandelbrot modeling to peer-to-peer traffic [104], and extending topology measurements to edge hosts [48]. More recent work looks to continued measurement of power-law phenomena [81, 60, 75], exploiting emerging topologies for optimizing network traffic [37, 68, 25], using network data to locate disruptions [46], the impact of inter-domain congestion [36], and studying the completeness of passive sources to determine how well they can observe microscopic phenomena [84].

The above sample of many years of Internet research has provided significant qualitative insights into Internet phenomenology. Single-parameter power-law fits have extensively been explored and shown to adequately fit higher-degree tails of the observations. However, more complex models are required to fit the entire range of observations. Figure 1.3a adapted from figure 8H [34] shows the number of bytes of data received in response to 2.3×10^5 HTTP (web) requests from computers at a large research laboratory and shows a strong agreement with a power law at large values, but diverges with the single-parameter model at small values. Figure 1.3b adapted from figure 9W [34] shows the distribution of 1.2×10^5 hits on web sites from AOL users and shows a strong agreement with a power law at small values, but diverges with the single-parameter model at large values. Figure 1.3c adapted from figure 9X [34] shows the distribution of 2.4×10^8 web hyperlinks and has a reasonable model agreement across the entire range, except for the smallest values. Figure 1.3d adapted from figure 4B [81] shows the distribution of visitors arriving at YouTube from referring web sites appears to be best represented by two very different power-law models with significant difference as the smallest values. Figure 1.3e adapted from figure 3A [60] shows the distribution of the number of Border Gateway Protocol updates received by the 4 monitors in 1-minute intervals and shows a strong agreement with a power law at large values, but diverges with the single-parameter model at small values. Figure 1.3f adapted from figure 21 in [75] shows the distribution of the Bitcoin network in 2011 and shows a strong agreement with a power law at small values, but diverges with the single-parameter model at large values.

The results shown in Figure 1.3 represent some of the best and most carefully executed fits to Internet data and clearly show the difficulty of fitting the entire range with a single-parameter power law. It is also worth mentioning that in each case the cumulative distribution is used, which naturally provides a smoother curve (in contrast to the differential cumulative distribution used in our analysis), but provides less detail on the underlying phenomena. Furthermore, the data in Figure 1.3 are typically isolated collections such that the error bars are not readily computable,

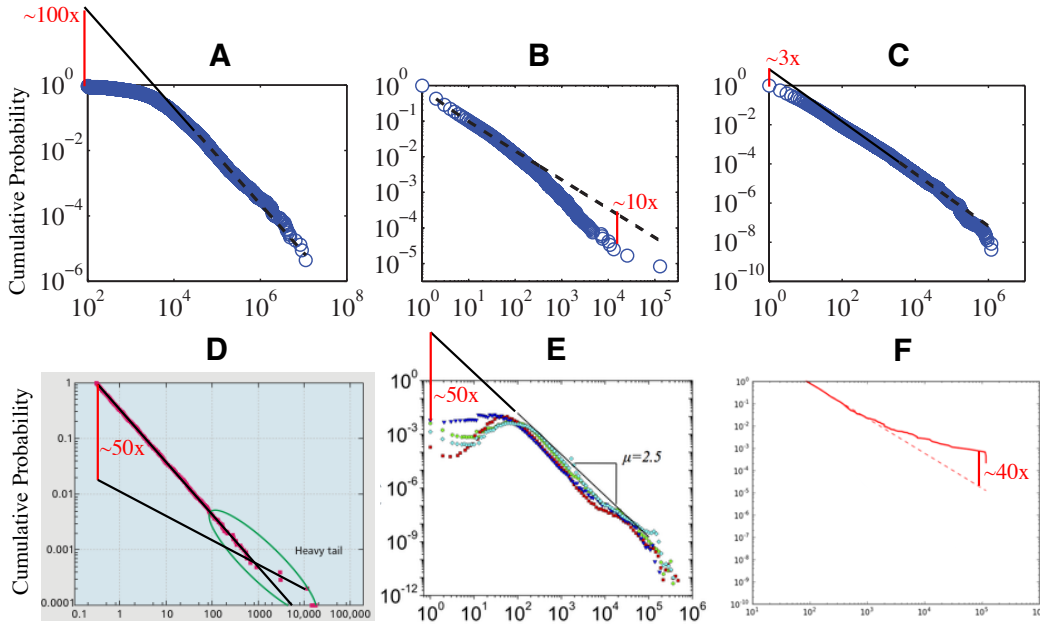


Figure 1.3 Single-parameter power-law fits of Internet data. Single-parameter fits of the cumulative distributions of Internet data have difficulty modeling the entire range. The estimated ratio between the model and the data at the model extremes is shown. (a) Figure 8H [34]. (b) Figure 9W [34]. (c) Figure 9X [34]. (d) Figure 4B [81]. (e) Figure 3A [60]. (f) Figure 21 [75].

which limits the ability to assess both the quality of the measurements and the model fits.

Regrettably, the best publicly available data about the global interconnection system that carries most of the world's communications traffic are incomplete and of unknown accuracy. There is no map of physical link locations, capacity, traffic, or interconnection arrangements. This opacity of the Internet infrastructure hinders research and development efforts to model network behavior and topology; design protocols and new architectures; and study real-world properties such as robustness, resilience, and economic sustainability. There are good reasons for the dearth of information: complexity and scale of the infrastructure; information-hiding properties of the routing system; security and commercial sensitivities; costs of storing and processing the data; and lack of incentives to gather or share data in the first place, including cost-effective ways to use it operationally. But understanding the Internet's history and present, much less its future, is impossible without realistic and representative data sets and measurement infrastructure on which to support sustained longitudinal measurements as well as new experiments. The MAWI and CAIDA data collection efforts are the largest efforts to provide the data necessary to begin to answer these questions.

1.3.1 Logarithmic Pooling

In this analysis before model fitting, the differential cumulative probabilities are calculated. For a network quantity d , the histogram of this quantity computed from \mathbf{A}_t is denoted by $n_t(d)$, with corresponding probability

$$p_t(d) = n_t(d) / \sum_d n_t(d) \quad (1.3)$$

and cumulative probability

$$P_t(d) = \sum_{i=1,d} p_t(d) \quad (1.4)$$

Because of the relatively large values of d observed due to a single supernode, 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 use the differential cumulative probability with logarithmic bins in d

$$D_t(d_i) = P_t(d_i) - P_t(d_{i-1}) \quad (1.5)$$

where $d_i = 2^i$ [34]. The corresponding mean and standard deviation of $D_t(d_i)$ over many different consecutive values of t for a given data set are denoted $D(d_i)$ and $\sigma(d_i)$. These quantities strike a balance between accuracy and detail for subsequent model fitting as demonstrated in the daily structural variations revealed in the Tokyo data (Figures 1.4 and 1.5).

Diurnal variations in supernode network traffic are well known [107]. The Tokyo packet data were collected over a period spanning two days and allow the daily variations in packet traffic to be observed. The precision and accuracy of our measurements allow these variations to be observed across a wide range of nodes. Figure 1.4 shows the fraction of source fan-outs in each of various bin ranges. The fluctuations show the network evolving between two envelopes occurring between noon and midnight that are shown in Figure 1.5.

1.3.2 Modified Zipf–Mandelbrot Model

Measurements of $D(d_i)$ can reveal many properties of network traffic, such as the number of nodes with only one connection $D(d=1)$ and the size of the supernode $d_{\max} = \arg\max(D(d) > 0)$. An effective low-parameter model allows these and many other properties to be summarized and computed efficiently. In the standard Zipf–Mandelbrot model typically used in linguistic contexts, the value d in Eq. 1.21 is a ranking with $d=1$ corresponding to the most popular value [82, 87, 104]. In our analysis, the Zipf–Mandelbrot model is modified so that d is a measured network quantity instead of a rank index (Eq. 1.21). The model exponent α has a larger impact on the model at large values of d , while the model offset δ has a larger impact on the model at small values of d and in particular at $d=1$.

The general saturation/cutoff models used to model a variety of network phenomena is denoted [34, 11]

$$p(d) \propto \frac{1}{(d + \delta)^\alpha \exp[\lambda d]} \quad (1.6)$$

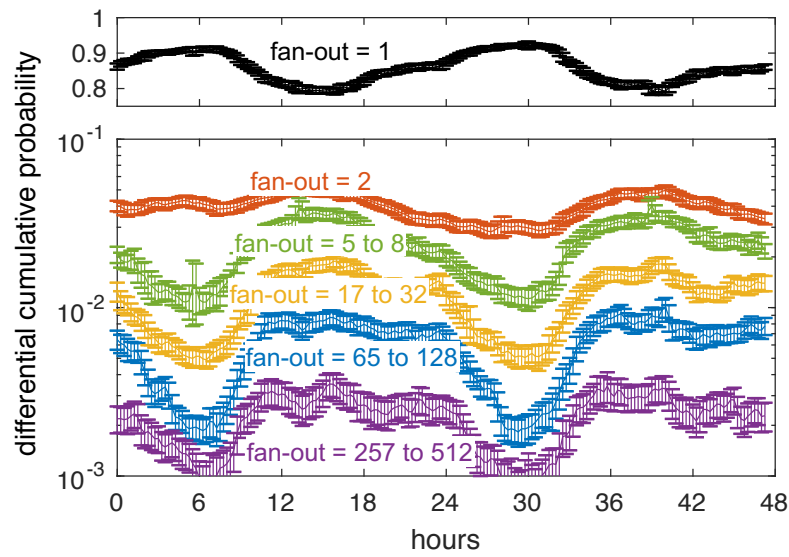


Figure 1.4 **Daily variation in Internet traffic.** The fraction of source nodes with a given range of fan-out is shown as a function of time for the Tokyo 2015 data. The $p(d = 1)$ value is plotted on a separate linear scale because of the larger magnitude relative to the other points. Each point is the mean of many neighboring points in time, and the error bars are the measured $\pm 1\sigma$. The daily variations of the distributions oscillate between extremes corresponding to approximately local noon and midnight.

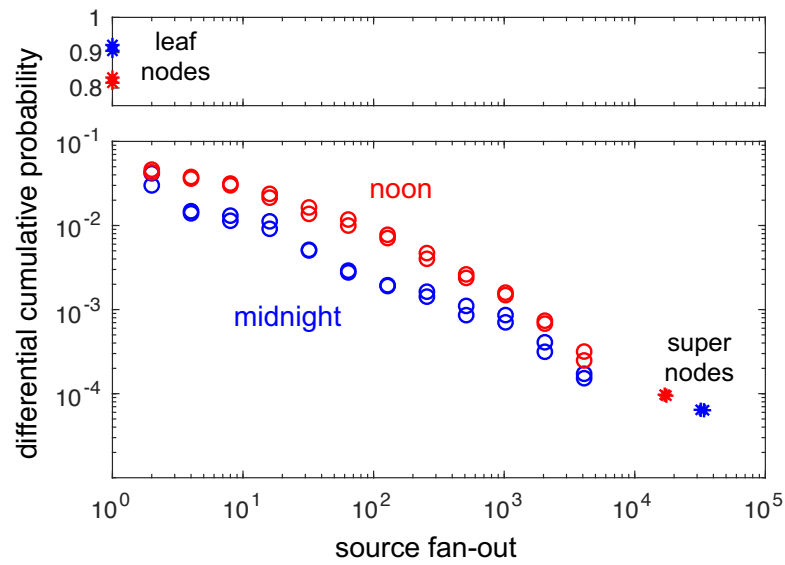


Figure 1.5 Daily limits in Internet traffic. The fraction of source nodes versus fan-out is shown for two noons and two midnights for the Tokyo 2015 data. The overlap among the noons and the midnights shows the relative day-to-day consistency in these data and shows the limits of the two extremes in daily variation. During the day, there is more traffic among nodes with intermediate fan-out. At night, the traffic is more dominated by leaf nodes and the supernode.

where δ is the low- d saturation and $1/\lambda$ is the high- d cutoff that bounds the power-law regime of the distribution. The modified Zipf–Mandelbrot is a special case of this distribution that accurately models our observations. The unnormalized modified Zipf–Mandelbrot model is denoted

$$\rho(d; \alpha, \delta) = \frac{1}{(d + \delta)^\alpha} \quad (1.7)$$

with corresponding derivative with respect to δ

$$\partial_\delta \rho(d; \alpha, \delta) = \frac{-\alpha}{(d + \delta)^{\alpha+1}} = -\alpha \rho(d; \alpha + 1, \delta) \quad (1.8)$$

The normalized model probability is given by

$$p(d; \alpha, \delta) = \frac{\rho(d; \alpha, \delta)}{\sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta)} \quad (1.9)$$

where d_{\max} is the largest value of the network quantity d . The cumulative model probability is the sum

$$P(d_i; \alpha, \delta) = \sum_{d=1}^{d_i} p(d; \alpha, \delta) \quad (1.10)$$

The corresponding differential cumulative model probability is

$$D(d_i; \alpha, \delta) = P(d_i; \alpha, \delta) - P(d_{i-1}; \alpha, \delta) \quad (1.11)$$

where $d_i = 2^i$. In terms of ρ , the differential cumulative model probability is

$$D(d_i; \alpha, \delta) = \frac{\sum_{d=d_{i-1}+1}^{d=d_i} \rho(d; \alpha, \delta)}{\sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta)} \quad (1.12)$$

The above function is closely related to the Hurwitz zeta function [38, 34, 119]

$$\zeta(\alpha, \delta_1) = \sum_{d=0}^{\infty} \rho(d; \alpha, \delta_1) \quad (1.13)$$

where $\delta_1 = \delta + 1$. The differential cumulative model probability in terms of the Hurwitz zeta function is

$$D(d_i; \alpha, \delta) = \frac{\zeta(\alpha, \delta + 3 + d_{i-1}) - \zeta(\alpha, \delta + 2 + d_i)}{\zeta(\alpha, \delta + 1) - \zeta(\alpha, \delta + 2)} \quad (1.14)$$

1.3.3 Nonlinear Model Fitting

The model exponent α has a larger impact on the model at large values of d , while the model offset δ has a larger impact on the model at small values of d and in particular at $d = 1$. A nonlinear fitting technique is used to obtain accurate model fits across the entire range of d . Initially, a set of candidate exponent values is selected, typically

$\alpha = 0.10, 0.11, \dots, 3.99, 4.00$. For each value of α , a value of δ is computed that exactly matches the model with the data at $D(1)$. Finding the value of δ corresponding to a give $D(1)$ is done using Newton's method as follows. Setting the measured value of $D(1)$ equal to the model value $D(1; \alpha, \delta)$ gives

$$D(1) = D(1; \alpha, \delta) = \frac{1}{(1 + \delta)^\alpha \sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta)} \quad (1.15)$$

Newton's method works on functions of the form $f(\delta) = 0$. Rewriting the above expression produces

$$f(\delta) = D(1)(1 + \delta)^\alpha \sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta) - 1 = 0 \quad (1.16)$$

For given value of α , δ can be computed from the following iterative equation

$$\delta \rightarrow \delta - \frac{f(\delta)}{\partial_\delta f(\delta)} \quad (1.17)$$

where the partial derivative $\partial_\delta f(\delta)$ is

$$\begin{aligned} \partial_\delta f(\delta) &= D(1) \partial_\delta [(1 + \delta)^\alpha \sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta)] \\ &= D(1) [\alpha(1 + \delta)^{\alpha-1} \sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta) + [(1 + \delta)^\alpha \sum_{d=1}^{d_{\max}} \partial_\delta \rho(d; \alpha, \delta)]] \\ &= D(1) [\alpha(1 + \delta)^{\alpha-1} \sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta) + [(1 + \delta)^\alpha \sum_{d=1}^{d_{\max}} -\alpha \rho(d; \alpha + 1, \delta)]] \\ &= \alpha D(1)(1 + \delta)^\alpha [(1 + \delta)^{-1} \sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta) - \sum_{d=1}^{d_{\max}} \rho(d; \alpha + 1, \delta)] \end{aligned} \quad (1.18)$$

Using a starting value of $\delta = 1$ and bounds of $0 < \delta < 10$, Newton's method can be iterated until the differences in successive values of δ fall below a specified error (typically 0.001), which is usually achieved in less than five iterations.

If faster evaluation is required, the sums in the above formulas can be accelerated using the integral approximations

$$\begin{aligned} \sum_{d=1}^{d_{\max}} \rho(d; \alpha, \delta) &\approx \sum_{d=1}^{d_{\text{sum}}} \rho(d; \alpha, \delta) + \int_{d_{\text{sum}}+0.5}^{d_{\max}+0.5} \rho(x; \alpha, \delta) dx \\ &= \sum_{d=1}^{d_{\text{sum}}} \rho(d; \alpha, \delta) + \frac{\rho(d_{\text{sum}} + 0.5; \alpha - 1, \delta) - \rho(d_{\max} + 0.5; \alpha - 1, \delta)}{\alpha - 1} \\ \sum_{d=1}^{d_{\max}} \rho(d; \alpha + 1, \delta) &\approx \sum_{d=1}^{d_{\text{sum}}} \rho(d; \alpha + 1, \delta) + \int_{d_{\text{sum}}+0.5}^{d_{\max}+0.5} \rho(x; \alpha + 1, \delta) dx \\ &= \sum_{d=1}^{d_{\text{sum}}} \rho(d; \alpha + 1, \delta) + \frac{\rho(d_{\text{sum}} + 0.5; \alpha, \delta) - \rho(d_{\max} + 0.5; \alpha, \delta)}{\alpha} \end{aligned} \quad (1.19)$$

where the parameter d_{sum} can be adjusted to exchange speed for accuracy. For typical values of α , δ , and d_{max} used in this work, the accuracy is approximately $1/d_{\text{sum}}$.

The best-fit α (and corresponding δ) is chosen by minimizing the $|\cdot|^{1/2}$ metric over logarithmic differences between the candidate models $D(d_i; \alpha, \delta)$ and the data

$$\operatorname{argmin}_{\alpha} \sum_{d_i} |\log(D(d_i)) - \log(D(d_i; \alpha, \delta))|^{1/2} \quad (1.20)$$

The $|\cdot|^{1/2}$ metric (or $|\cdot|_p$ -norm with $p = 1/2$) favors maximizing error sparsity over minimizing outliers [39, 24, 116, 55, 103, 21, 117]. Several authors have recently shown that it is possible to reconstruct a nearly sparse signal from fewer linear measurements than would be expected from traditional sampling theory. Furthermore, by replacing the $|\cdot|_1$ norm with the $|\cdot|_p$ with $p < 1$, reconstruction is possible with substantially fewer measurements.

Using logarithmic values more evenly weights their contribution to the model fit and more accurately reflects the number of packets used to compute each value of $D(d_i)$. Lower-accuracy data points are avoided by limiting the fitting procedure to data points where the value is greater than the standard deviation: $D(d_i) > \sigma(d_i)$.

1.4 RESULTS

Figure 1.6b shows five representative model fits out of the 350 performed on 10 data sets, 5 network quantities, and 7 valid packet windows: $N_V = 10^5, 3 \times 10^5, 10^6, 3 \times 10^6, 10^7, 3 \times 10^7, 10^8$. The model fits are valid over the entire range of d and provide parameter estimates with precisions of 0.01. In every case, the high value of $p(d=1)$ is indicative of a large contribution from a combination of supernode leaves, core leaves, and isolated links (Figure 1.17a). The breadth and accuracy of these data allow a detailed comparison of the model parameters. Figure 1.6c shows the model offset δ versus the model exponent α for all 350 fits. The different collection locations are clearly distinguishable in this model parameter space. The Tokyo collections have smaller offsets and are more tightly clustered than the Chicago collections. Chicago B has a consistently smaller source and link packet model offset than Chicago A. All the collections have source, link, and destination packet model exponents in the relatively narrow $1.5 < \alpha < 2$ range. The source fan-out and destination fan-in model exponents are in the broader $1.5 < \alpha < 2.5$ range and are consistent with the prior literature [34]. These results represent an entirely new approach to characterizing Internet traffic that allows the distributions to be projected into a low-dimensional space and enables accurate comparisons among packet collections with different locations, dates, durations, and sizes. Figure 1.6c indicates that the distributions of the different collection points occupy different parts of the modified Zipf–Mandelbrot model parameter space. Figures 1.7–1.16 show the measured and modeled differential cumulative distributions for the source fan-out, source packets, destination fan-in, destination packets, and link packets for all the collected data.

Figure 1.17b shows the average relative fractions of sources, total packets, total links, and the number of destinations in each of the five topologies for the ten data sets, and seven valid packet windows: $N_V = 10^5, 3 \times 10^5, 10^6, 3 \times 10^6, 10^7, 3 \times 10^7, 10^8$.

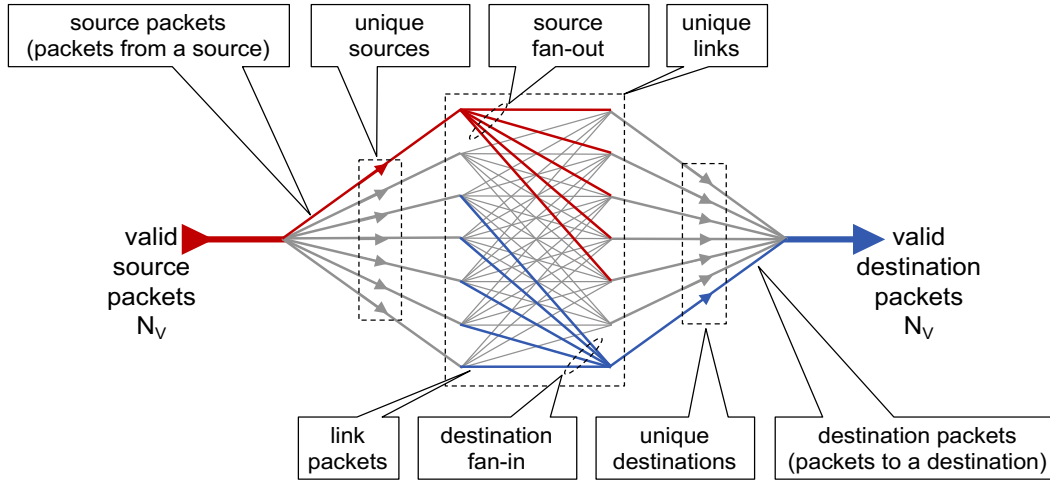
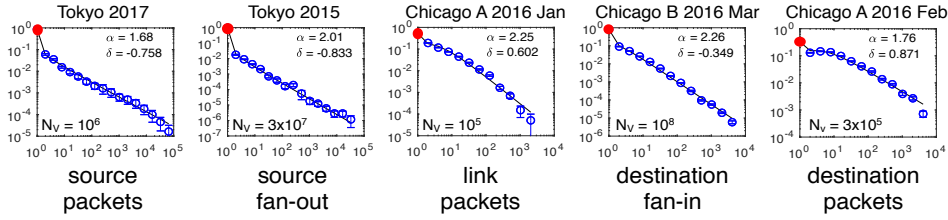
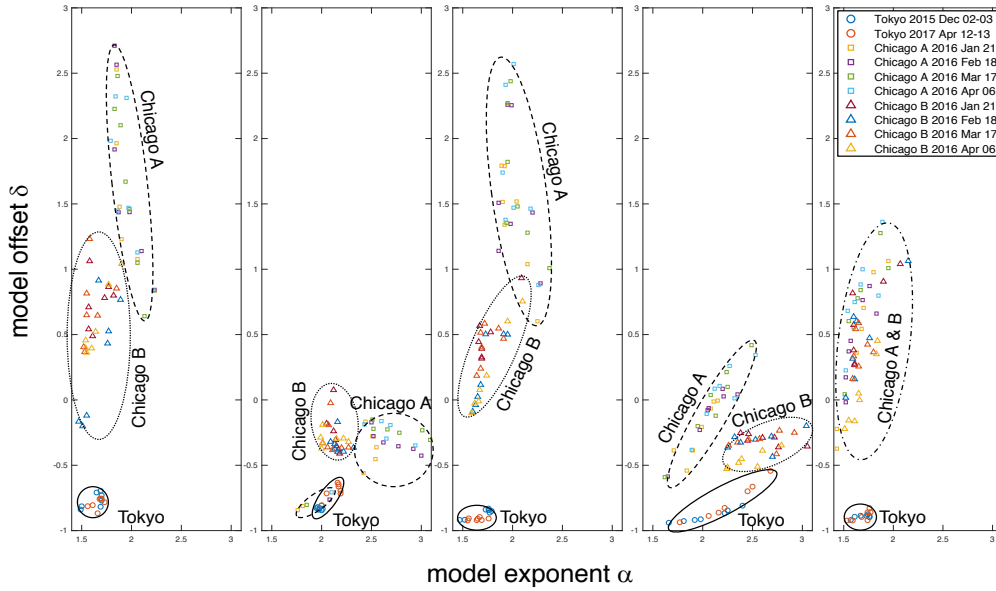
A**B**differential
cumulative
probability**C**

Figure 1.6 Streaming network traffic quantities, distributions, and model fits. (a) Internet traffic streams of N_V valid packets are divided into a variety of quantities for analysis. (b) A selection of 5 of the 350 measured differential cumulative probabilities spanning different locations, dates, and packet windows. Blue circles are measured data with $\pm 1\text{-}\sigma$ error bars. Black lines are the best-fit modified Zipf–Mandelbrot models with parameters α and δ . Red dots highlight the large contribution of leaf nodes and isolated links. (c) Model fit parameters for all 350 measured probability distributions.

The four projections in Figure 1.17b are chosen from Figures 1.18–1.21 to highlight the differences in the collection locations. The distinct regions in the various projections shown in Figure 1.17b indicate that underlying topological differences are present in the data. The Tokyo collections have much larger supernode leaf components than the Chicago collections. The Chicago collections have much larger core and core leaves components than the Tokyo collections. Chicago A consistently has fewer isolated links than Chicago B. Comparing the modified Zipf–Mandelbrot model parameters in Figure 1.6c and underlying topologies in Figure 1.17b suggests that the model parameters are a more compact way to distinguish the network traffic.

Figures 1.6c and 1.17b indicate that different collection points produce different model parameters α and δ and that these collection points also have different underlying topologies. Figure 1.22 connects the model fits and topology observations by plotting the topology fraction as a function of the model leaf parameter $1/(1 + \delta)^\alpha$ which corresponds to the relative strength of the distribution at $p(d = 1)$

$$1/(1 + \delta)^\alpha \propto p(d = 1; \alpha, \delta) \quad (1.21)$$

The correlations revealed in Figure 1.22 suggest that the model leaf parameter strongly correlates with the fraction of the traffic in different underlying network topologies and is a potentially new and beneficial way to characterize networks. Figure 1.22 indicates that the fraction of sources, links, and destinations in the core shrinks as the relative importance of the leaf parameter in the source fan-out and destination fan-in increases. In other words, more source and destination leaves mean a smaller core. Likewise, the fraction of links and total packets in the supernode leaves grows as the leaf parameter in the link packets and source packets increases. Interestingly, the fraction of sources in the core leaves and isolated links decreases as the leaf parameter in the source and destination packets increases indicating a shift of sources away from the core leaves and isolated links into supernode leaves. Thus, the modified Zipf–Mandelbrot model and its leaf parameter provide a direct connection with the network topology, underscoring the value of having accurate model fits across the entire range of values and in particular for $d = 1$.

Figures 1.23–1.27 show the fraction of the sources, links, total packets, and destinations in each of the measured topologies for all the locations as a function of the modified Zipf–Mandelbrot leaf parameter computed from the model fits of the source packets, source fan-out, link packets, destination fan-in, and destination packets taken from Figures 1.7–1.16.

1.5 DISCUSSION

Measurements of Internet traffic are useful for informing policy, identifying and preventing outages, defeating attacks, planning for future loads, and protecting the Domain Name System [33]. On a given day, millions of IPs are engaged in scanning behavior. Our improved models can aid cybersecurity analysts in determining which of these IPs are nefarious [118], the distribution of attacks in particular critical sectors [51], identifying spamming behavior [44], how to vaccinate against computer viruses

[8], obscuring web sources [52], identifying significant flow aggregates in traffic [26], and sources of rumors [95].

The results presented here have a number of potential practical applications for Internet stakeholders. The methods presented of collecting, filtering, computing, and binning the data to produce accurate measurements of a variety of network quantities are generally applicable to Internet measurements and have the potential to produce more accurate measures of these quantities. The accurate fits of the two-parameter modified Zipf–Mandelbrot distribution offer all the usual benefits of low-parameter models: measuring parameters with far less data, accurate predictions of network quantities based on a few parameters, observing changes in the underlying distribution, and using modeled distributions to detect anomalies in the data.

From a scientific perspective, improved knowledge of how Internet traffic flows can inform our understanding of how economics, topology, and demand shape the Internet over time. As with all scientific disciplines, the ability of theoreticians to develop and test theories of the Internet and network phenomena is bounded by the scale and accuracy of measured phenomena [2, 19, 109, 111]. The connections among dynamic evolution [14], network topology [92][15, 79], network robustness [71], controllability [78], community formation [99], and spreading phenomena [50] have emerged in many contexts [9, 112, 64]. Many first-principles theories for Internet and network phenomena have been proposed, such as Poisson models [98], fractional Brownian motion [115], preferential attachment [10, 4][93, 106], statistical mechanics [3], percolation [1], hyperbolic geometries [67, 66], non-global greedy routing [16, 17, 18], interacting particle systems [6], higher-order organization of complex networks from graph motifs [12], and minimum control energy [74]. All of these models require data to test them. In contrast to previous network models that have principally been based on data obtained from network crawls from a variety of start points on the network, our network traffic data are collected from observations of network streams. Both viewpoints provide important network observations. Observations of a network stream provide complementary data on network dynamics and highlight the contribution of leaves and isolated edges, which are less sampled in network crawls.

The aggregated data sets our teams have collected provide a unique window into these questions. The nonlinear fitting techniques described are a novel approach to fitting power-law data and have potential applications to power-law networks in diverse domains. The model fit parameters present new opportunities to connect the distributions to underlying theoretical models of networks. That the model fit parameters distinguish the different collection points and are reflective of different network topologies in the data at these points suggests a deeper underlying connection between the models and the network topologies.

1.6 CONCLUSIONS

Our society critically depends on the Internet for our professional, personal, and political lives. This dependence has rapidly grown much stronger than our comprehension of its underlying structure, performance limits, dynamics, and evolution. The fundamental characteristics of the Internet are perpetually challenging to research and

analyze, and we must admit we know little about what keeps the system stable. As a result, researchers and policymakers deal with a multi-trillion-dollar ecosystem essentially in the dark, and agencies charged with infrastructure protection have little situational awareness regarding global dynamics and operational threats. This paper has presented an analysis of the largest publicly available collection of Internet traffic consisting of 50 billion packets and reveals a new phenomenon: the importance of otherwise unseen leaf nodes and isolated links in Internet traffic. Our analysis further shows that a two-parameter modified Zipf–Mandelbrot distribution accurately describes a wide variety of source/destination statistics on moving sample windows ranging from 100,000 to 100,000,000 packets over collections that span years and continents. The measured model parameters distinguish different network streams, and the model leaf parameter strongly correlates with the fraction of the traffic in different underlying network topologies. These results represent a significant improvement in Internet modeling accuracy, improve our understanding of the Internet, and show the importance of stream sampling for measuring network phenomena.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the following individuals for their contributions and support: Shohei Araki, William Arcand, David Bestor, William Bergeron, Bob Bond, Paul Burkhardt, Chansup Byun, Cary Conrad, Alan Edelman, Sterling Foster, Bo Hu, Matthew Hubbell, Micheal Houle, Micheal Jones, Anne Klein, Charles Leiserson, Dave Martinez, Mimi McClure, Julie Mullen, Steve Pritchard, Andrew Prout, Albert Reuther, Antonio Rosa, Victor Roytburd, Siddharth Samsi, Koichi Suzuki, Kenji Takahashi, Michael Wright, Charles Yee, and Michitoshi Yoshida.

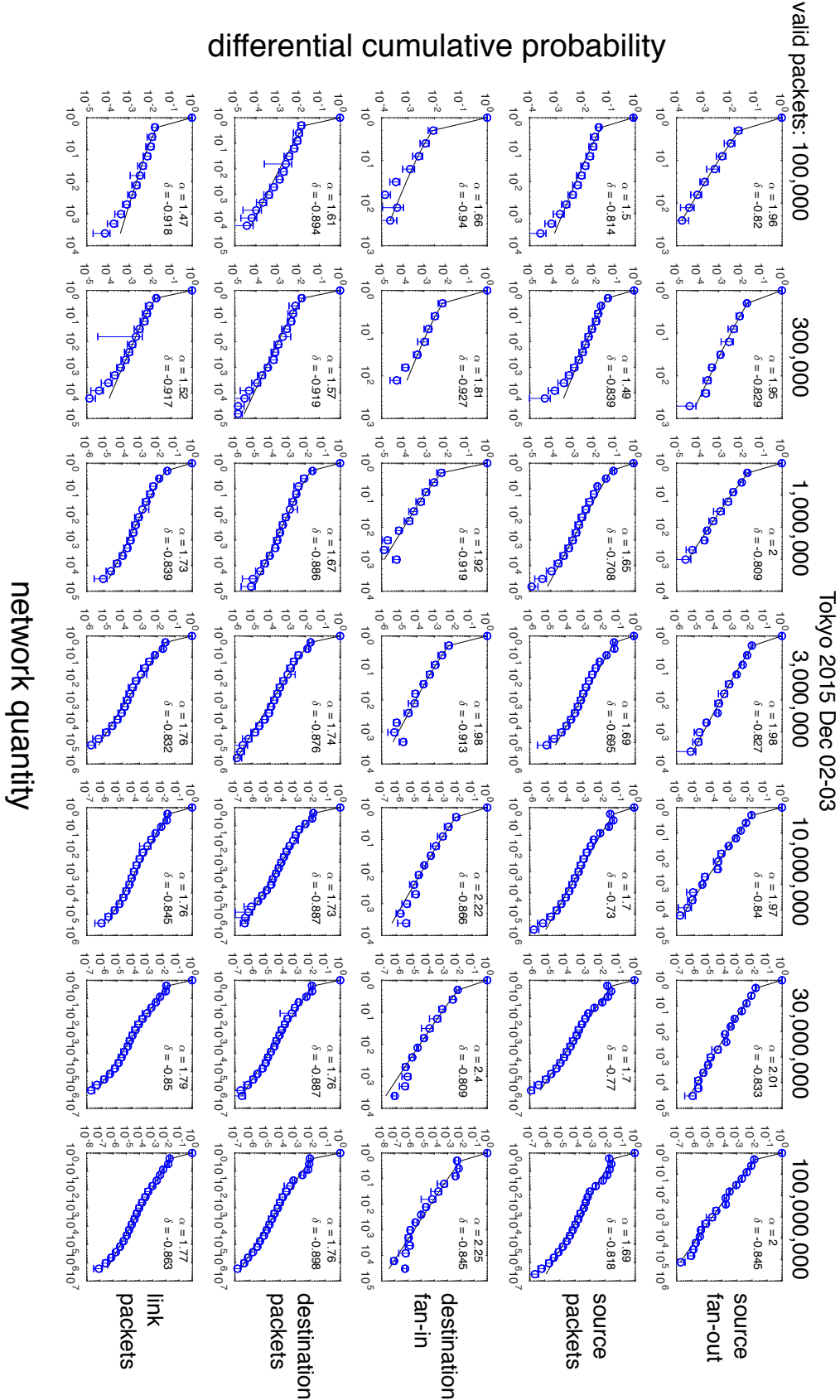


Figure 1.7 Measured differential cumulative probabilities. Blue circles with ± 1 - σ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Tokyo 2015 Dec 02.

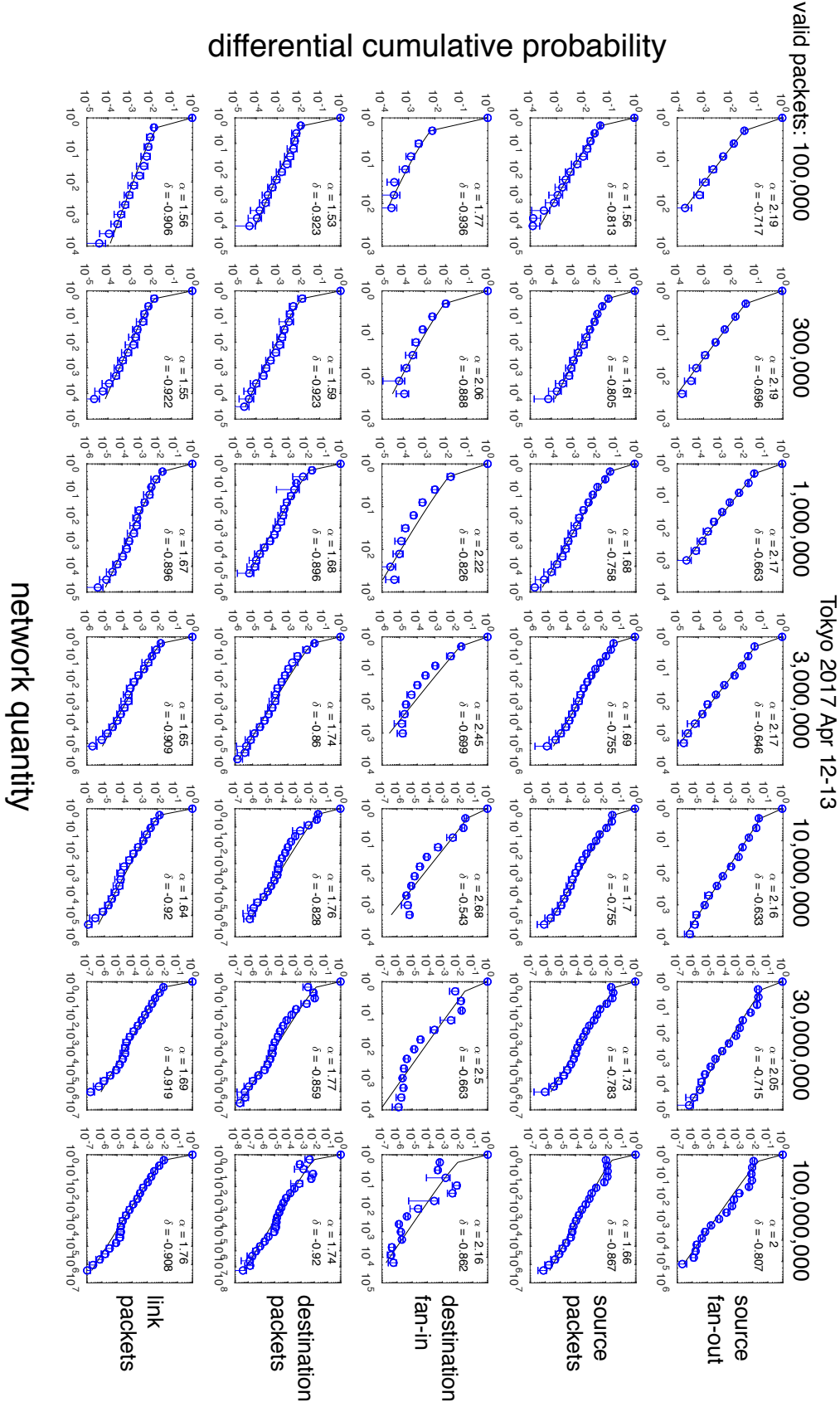


Figure 1.8 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Tokyo 2017 Apr 12.

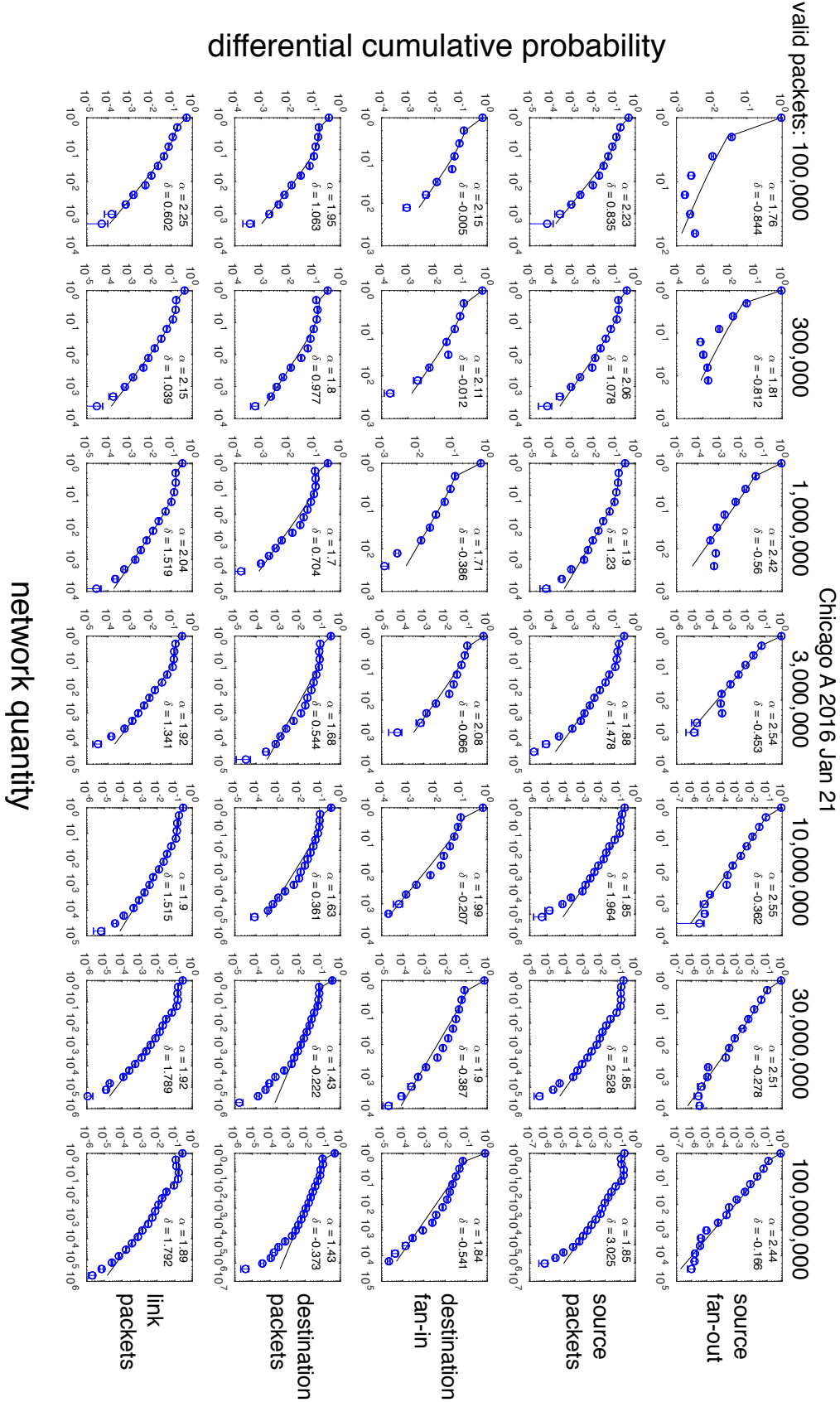


Figure 1.9 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago A 2016 Jan 21.

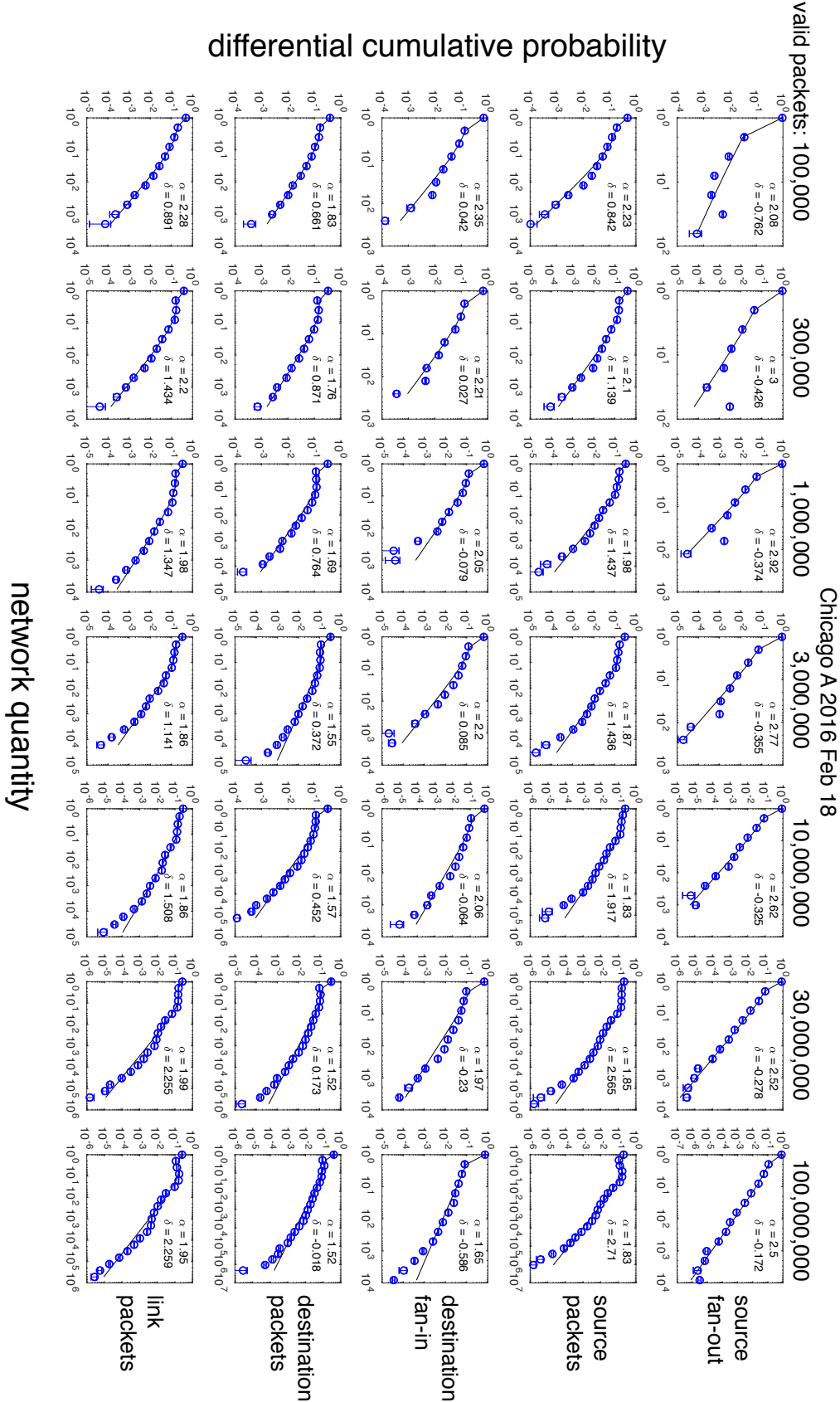


Figure 1.10 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago A 2016 Feb 18.

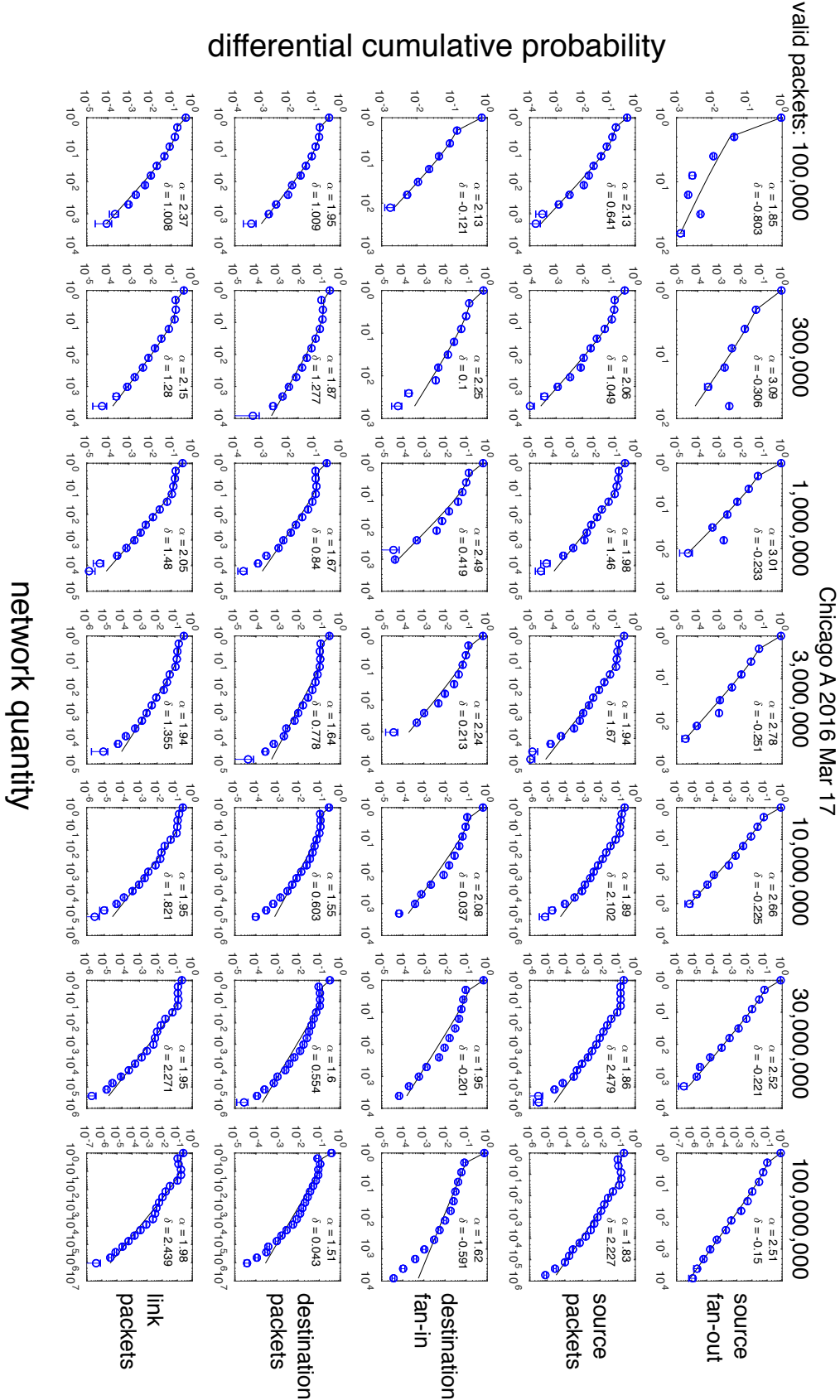


Figure 1.11 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago A 2016 Mar 17.

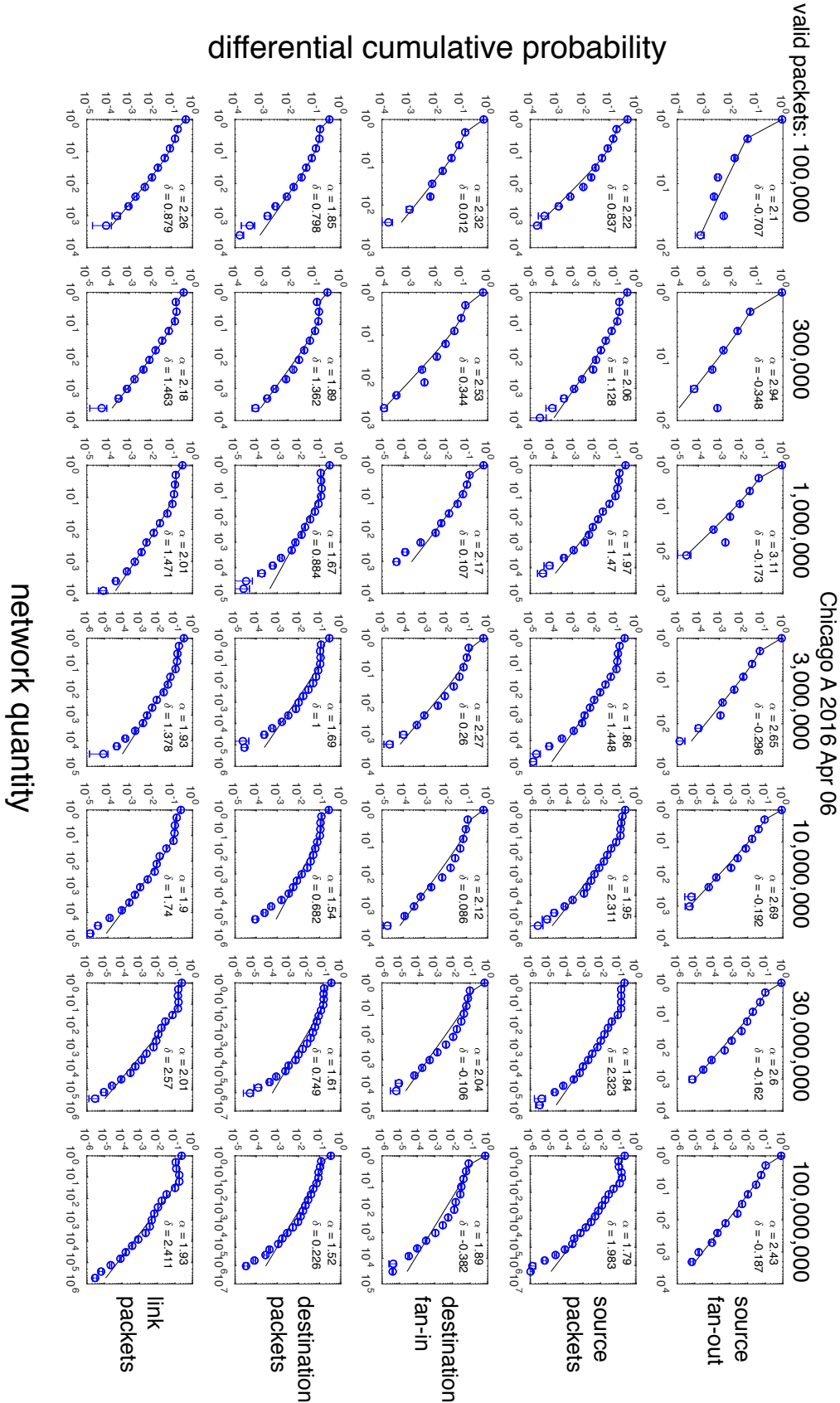


Figure 1.12 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago A 2016 Apr 06.

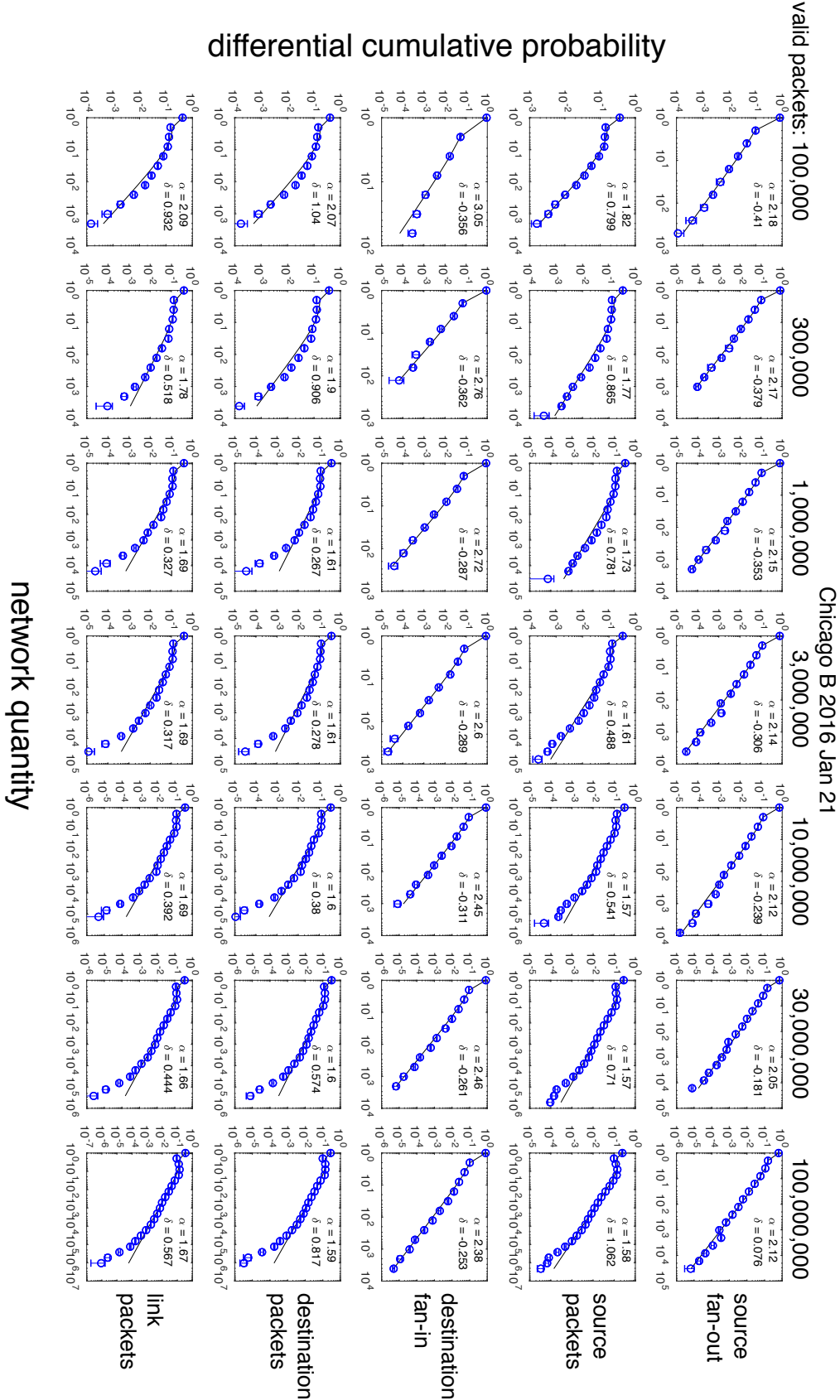


Figure 1.13 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago B 2016 Jan 21.

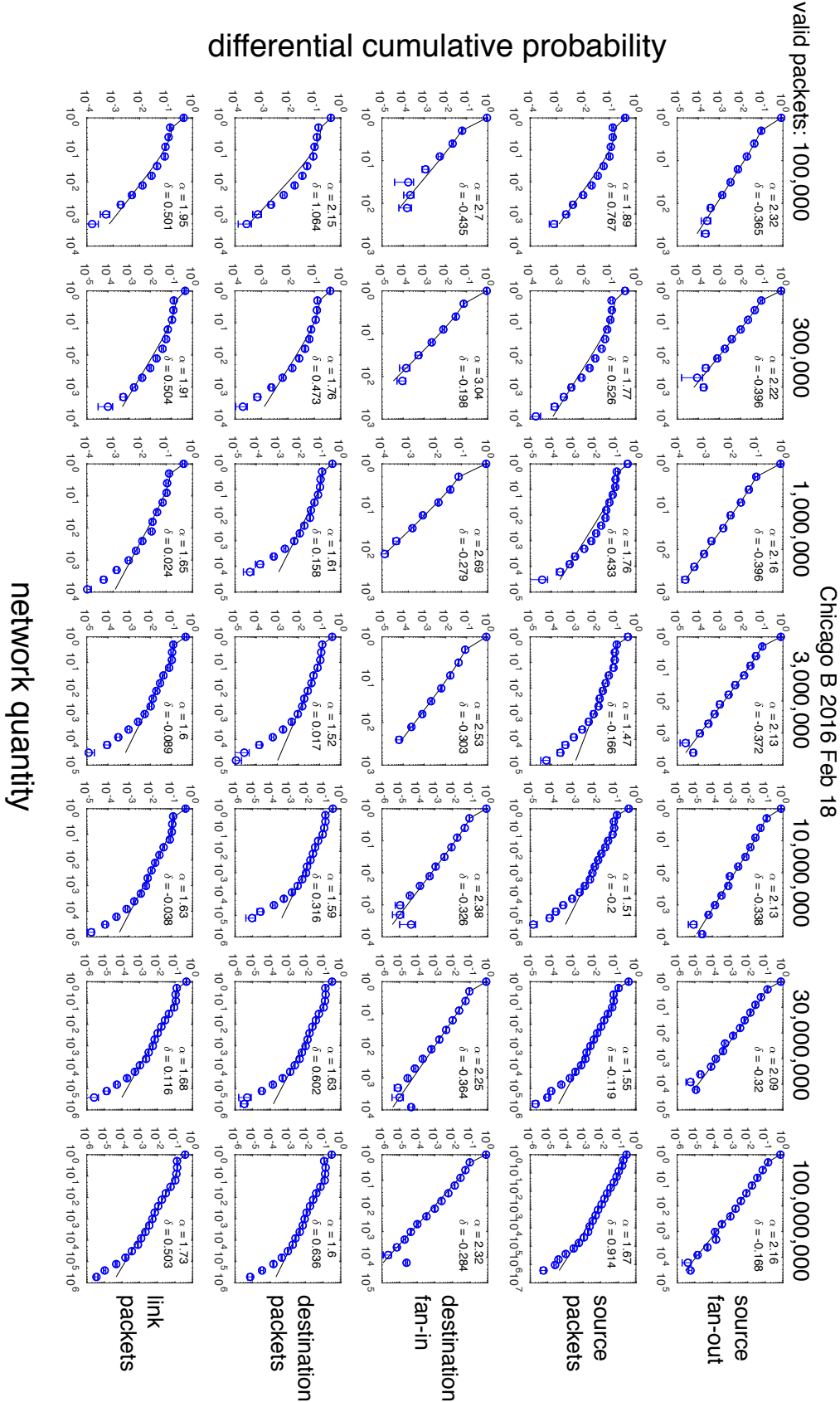


Figure 1.14 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago B 2016 Feb 18.

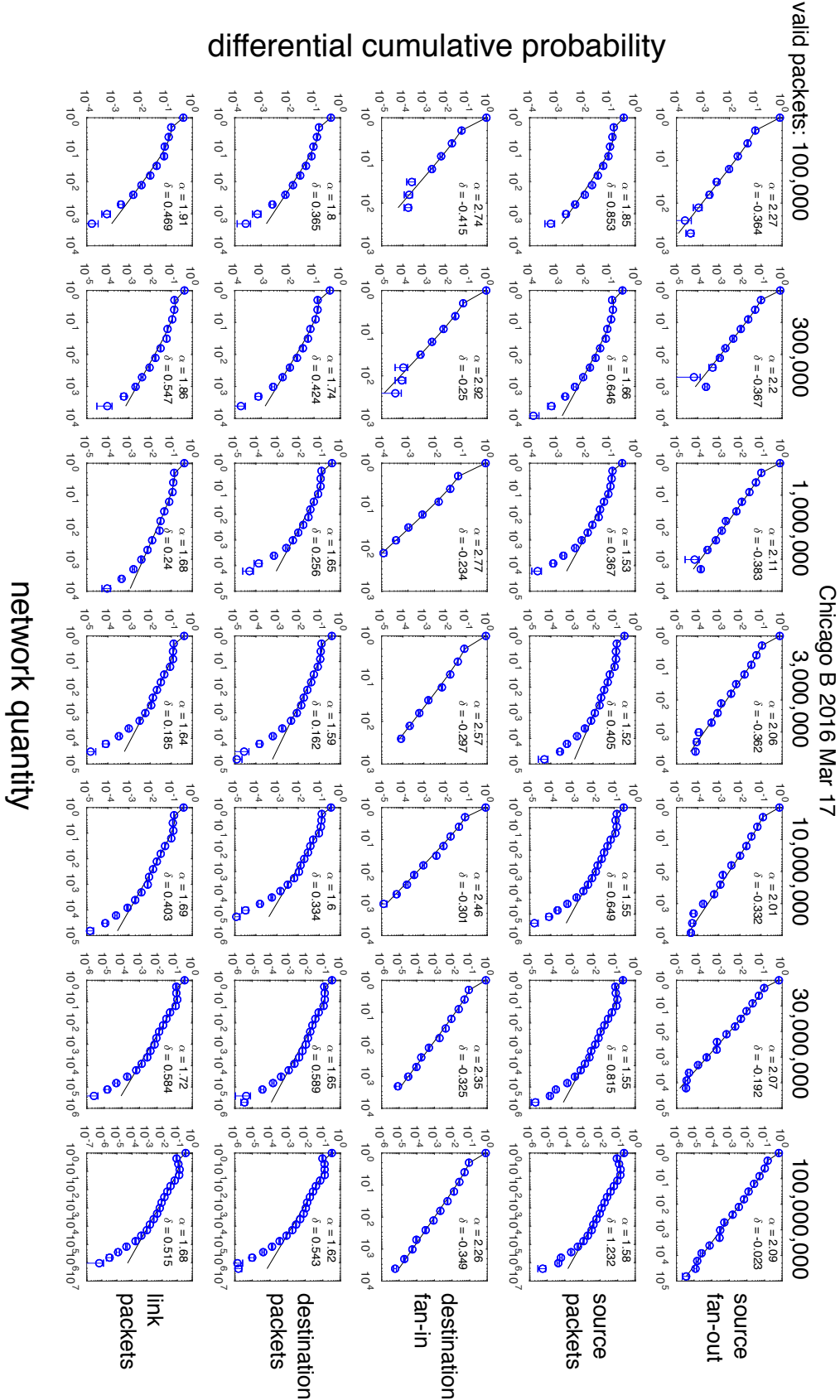


Figure 1.15 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago B 2016 Mar 17.

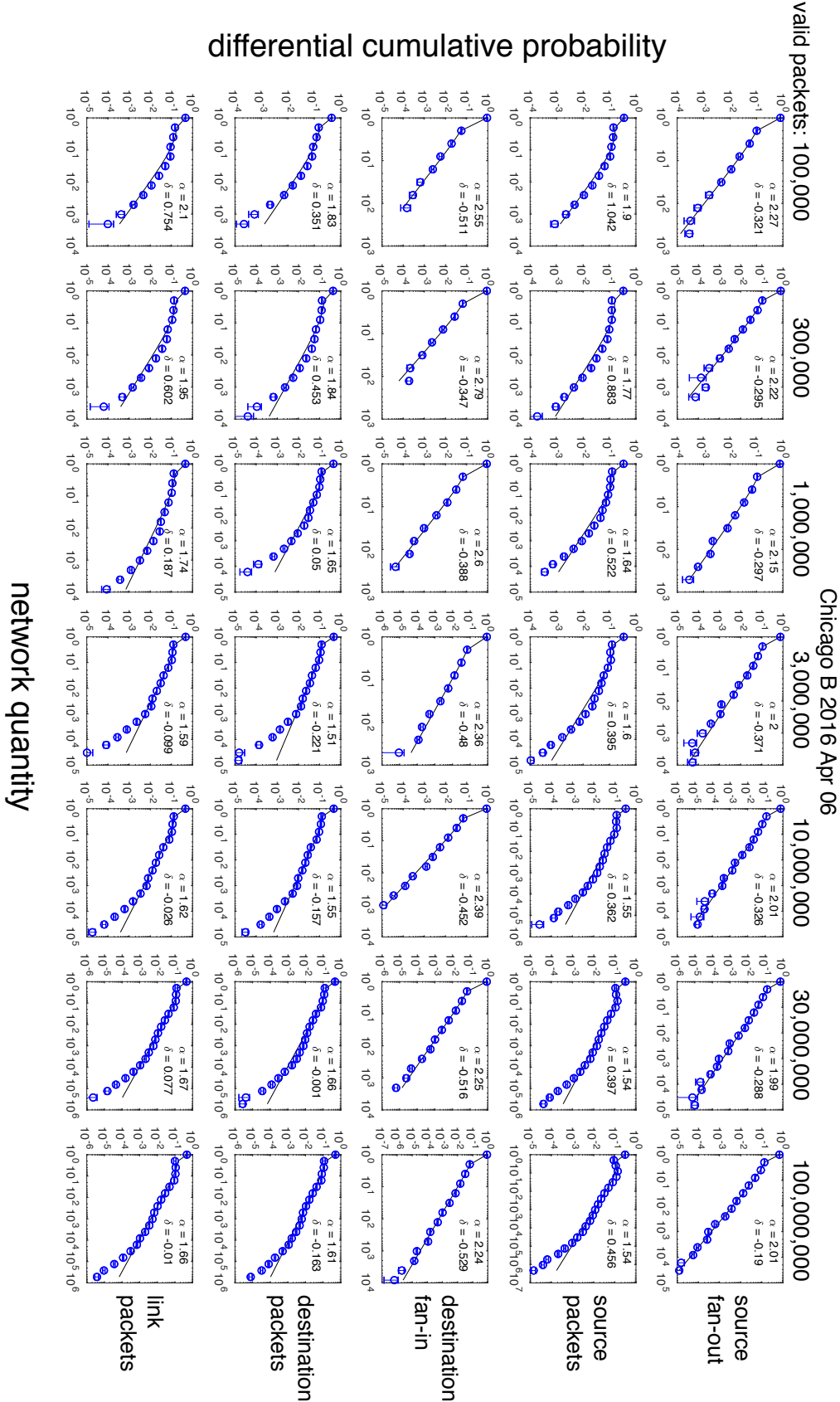


Figure 1.16 Measured differential cumulative probabilities. Blue circles with $\pm 1\text{-}\sigma$ error bars, along with their best-fit modified Zipf-Mandelbrot models (black line) and parameters α and δ performed for Chicago B 2016 Apr 06.

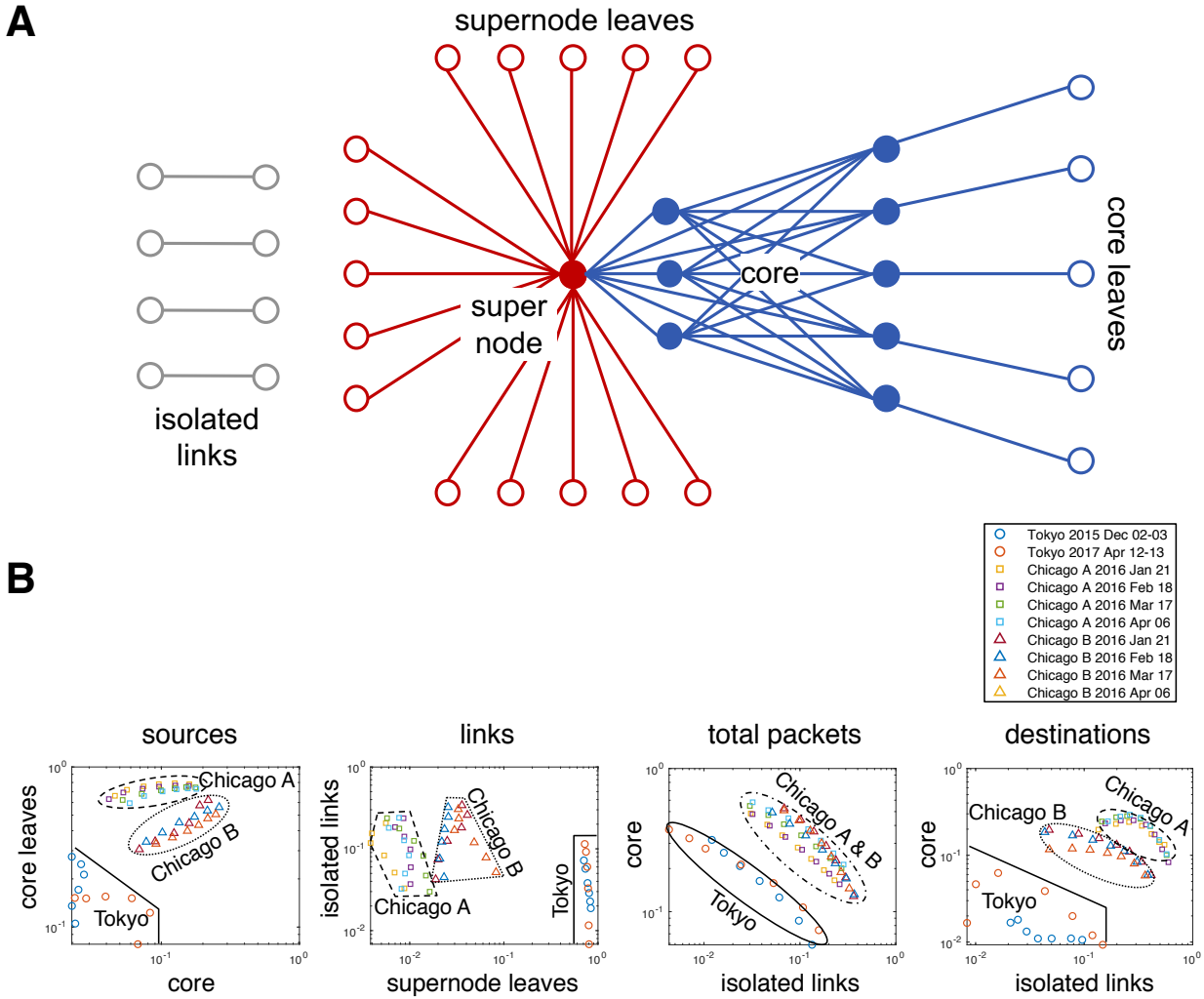


Figure 1.17 Distribution of traffic among network topologies. (a) Internet traffic forms networks consisting of a variety of topologies: isolated links, supernode leaves connected to a supernode, and densely connected core(s) with corresponding core leaves. (b) A selection of four projections showing the fraction of data in various underlying topologies. Horizontal and vertical axes are the corresponding fraction of the sources, links, total packets, and destinations that are in various topologies for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$). These data reveal the differences in the network traffic topologies in the data collected in Tokyo (dominated by supernode leaves), Chicago A (dominated by core leaves), and Chicago B (between Tokyo and Chicago A).

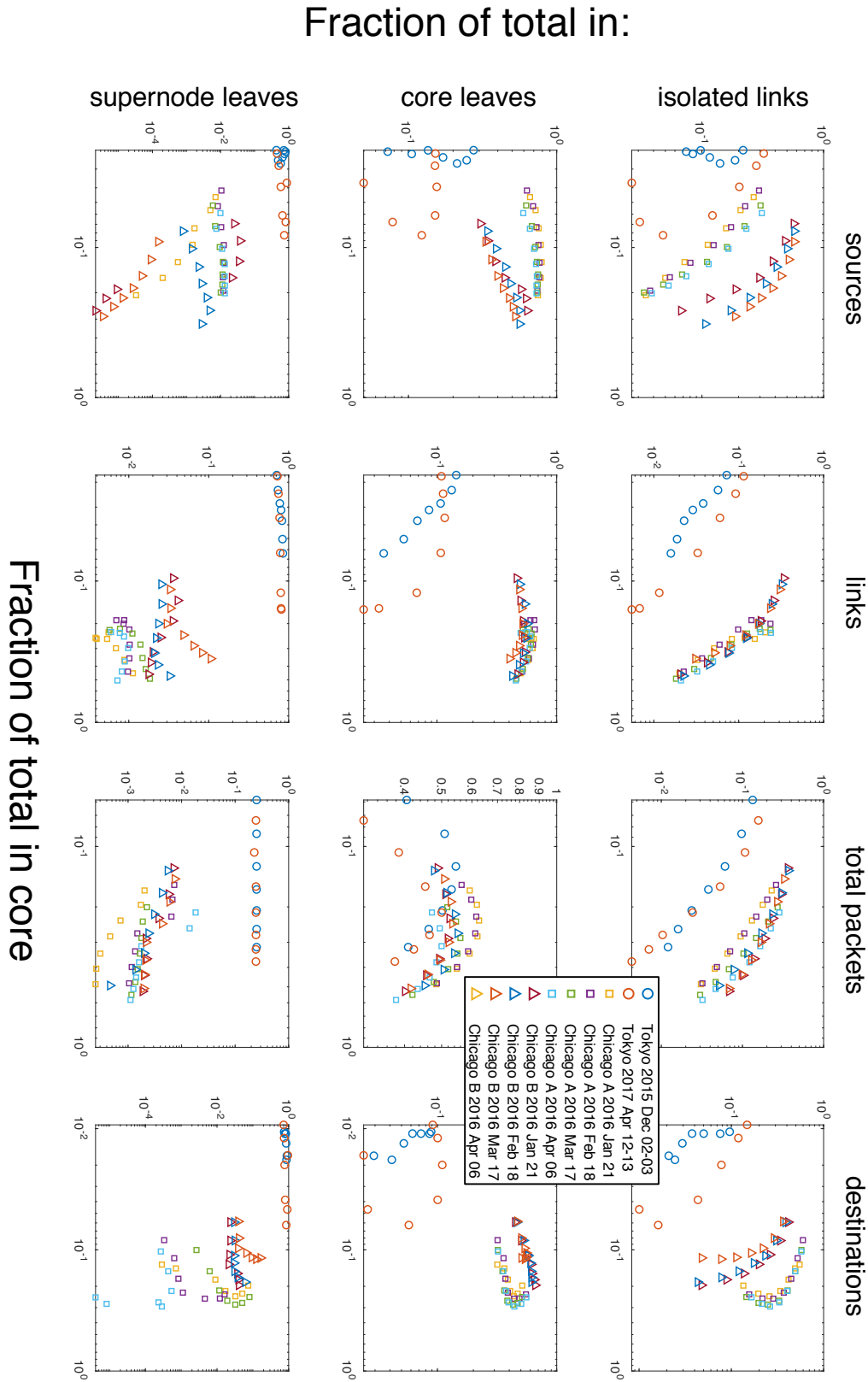


Figure 1.18 **Fraction of data in different network topologies.** Fraction of data in isolated links, core leaves, and supernode leaves versus the fraction of data in the core for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

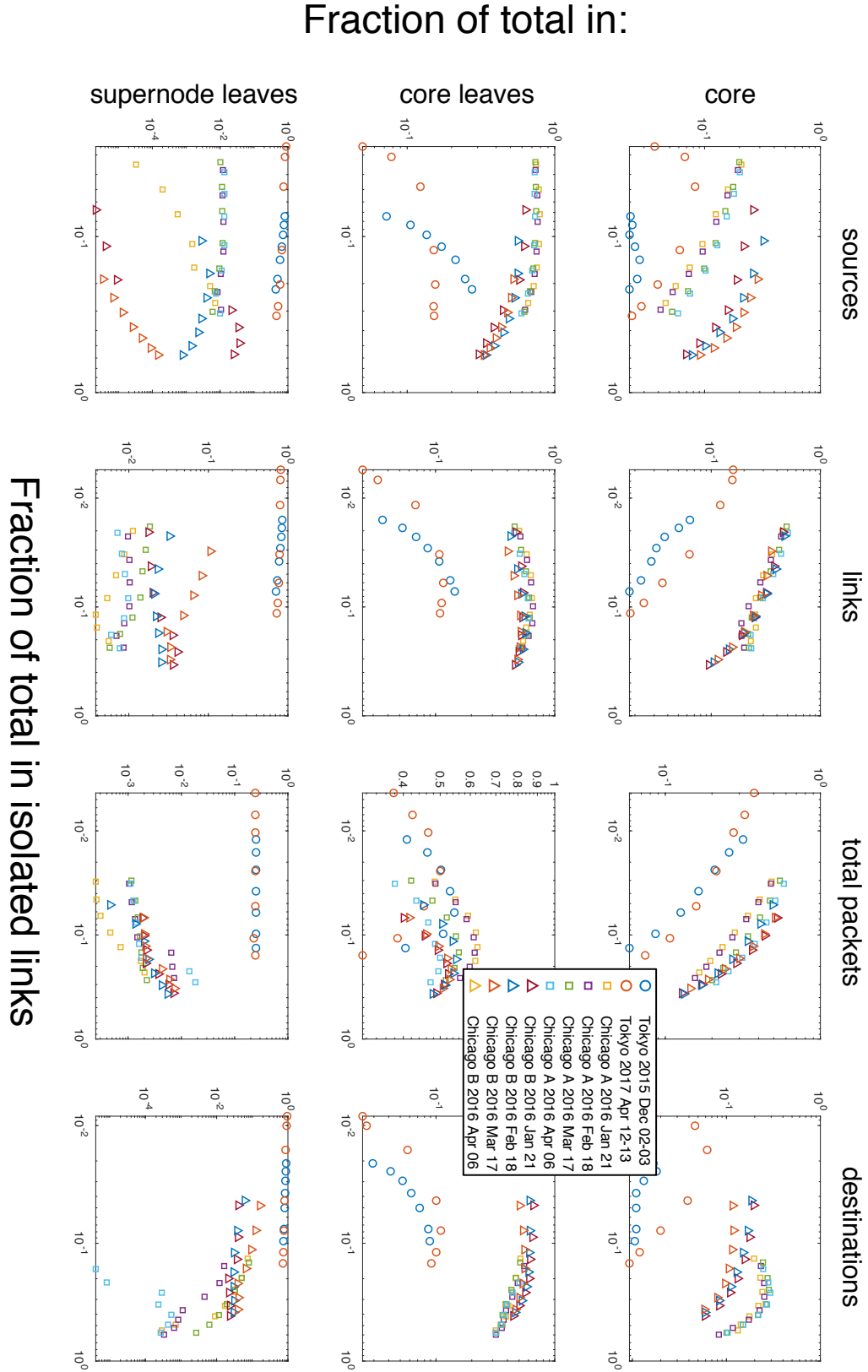


Figure 1.19 **Fraction of data in different network topologies.** Fraction of data in the core, core leaves, and supernode leaves versus the fraction of data in isolated links for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

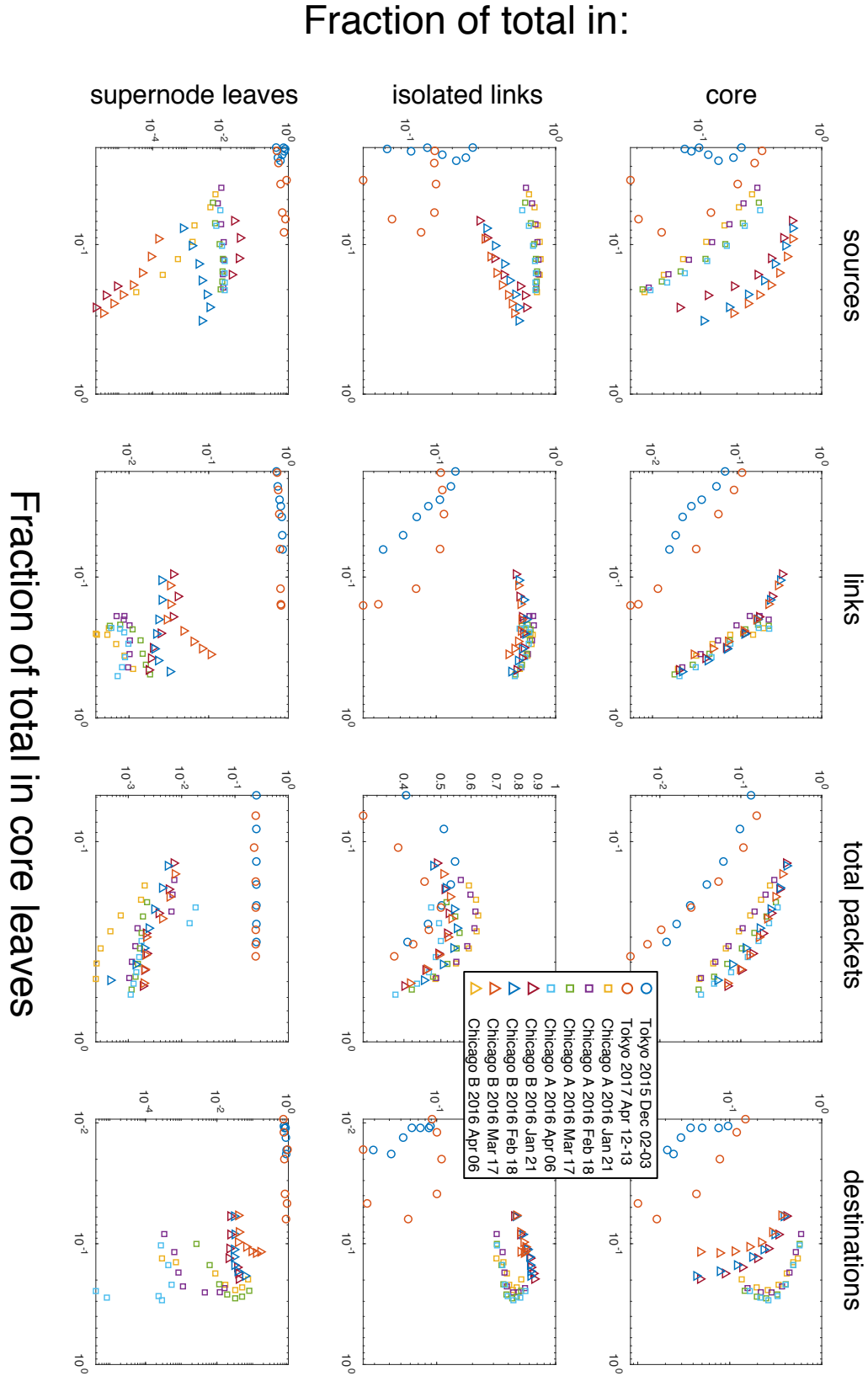


Figure 1.20 **Fraction of data in different network topologies.** Fraction of data in the core, isolated links, and supernode leaves versus the fraction of data in core leaves for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

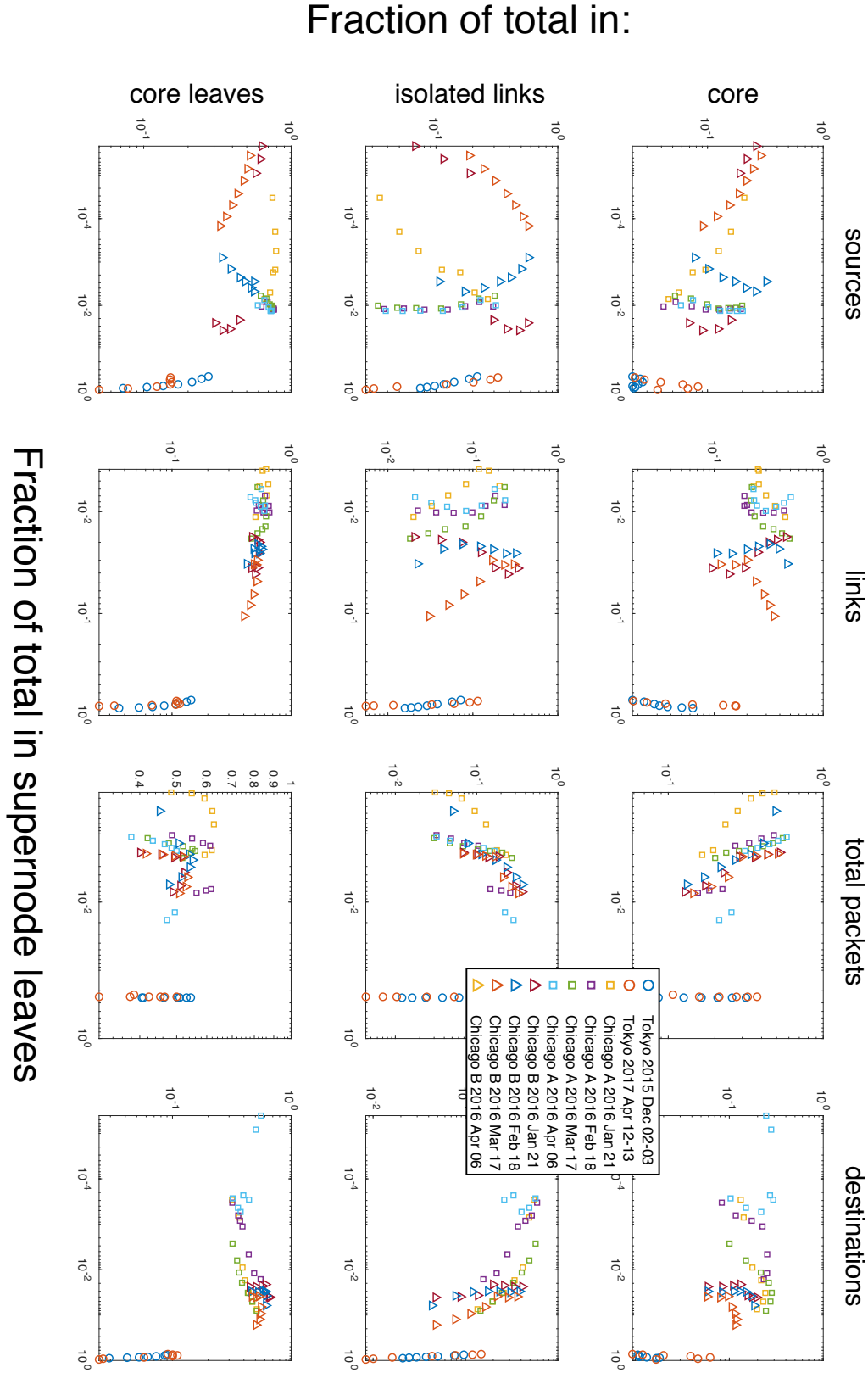


Figure 1.21 **Fraction of data in different network topologies.** Fraction of data in the core, isolated links, and core leaves versus the fraction of data in supernode leaves for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

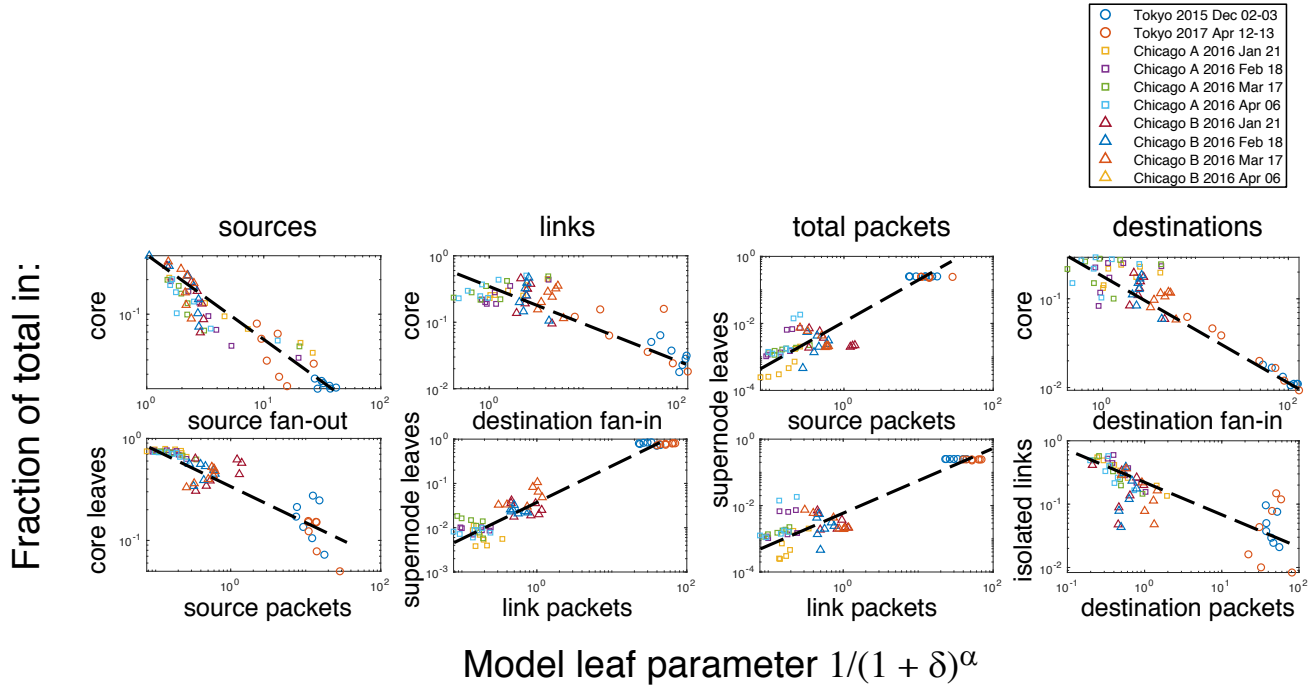


Figure 1.22 Topology versus model leaf parameter. Network topology is highly correlated with the modified Zipf–Mandelbrot model leaf parameter $1/(1 + \delta)^\alpha$. A selection of eight projections showing the fraction of data in various underlying topologies. Vertical axis is the corresponding fraction of the sources, links, total packets, and destinations that are in various topologies. Horizontal axis is the value of the model parameter taken from either the source packet, source fan-out, link packet, destination fan-in, or destination packet fits. Data points are for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

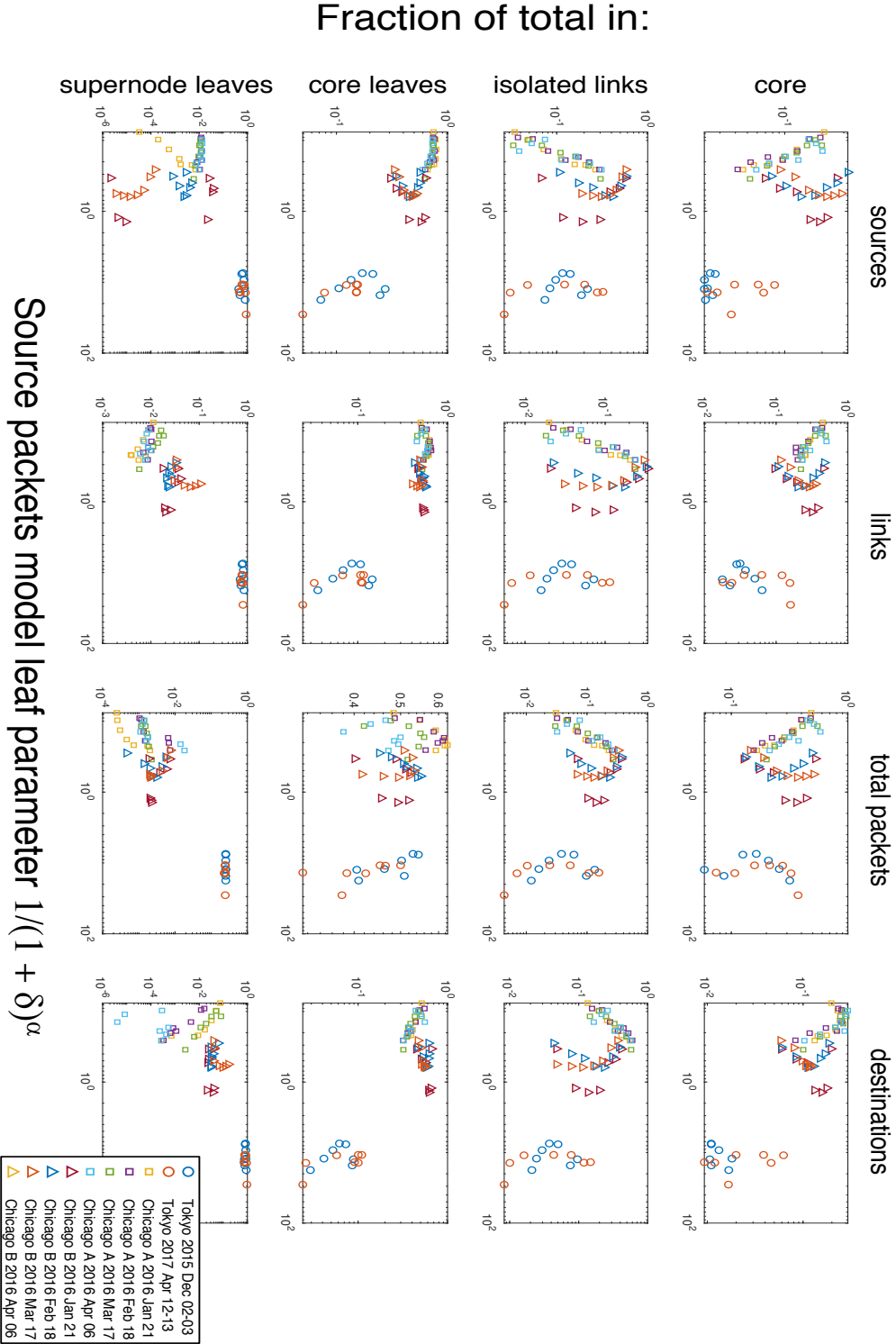


Figure 1.23 **Topology versus modified Zipf–Mandelbrot model leaf parameter.** Fraction of data in the core, isolated links, core leaves, and supernode leaves versus the model leaf parameter computed from the source packet modified Zipf–Mandelbrot model fits for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

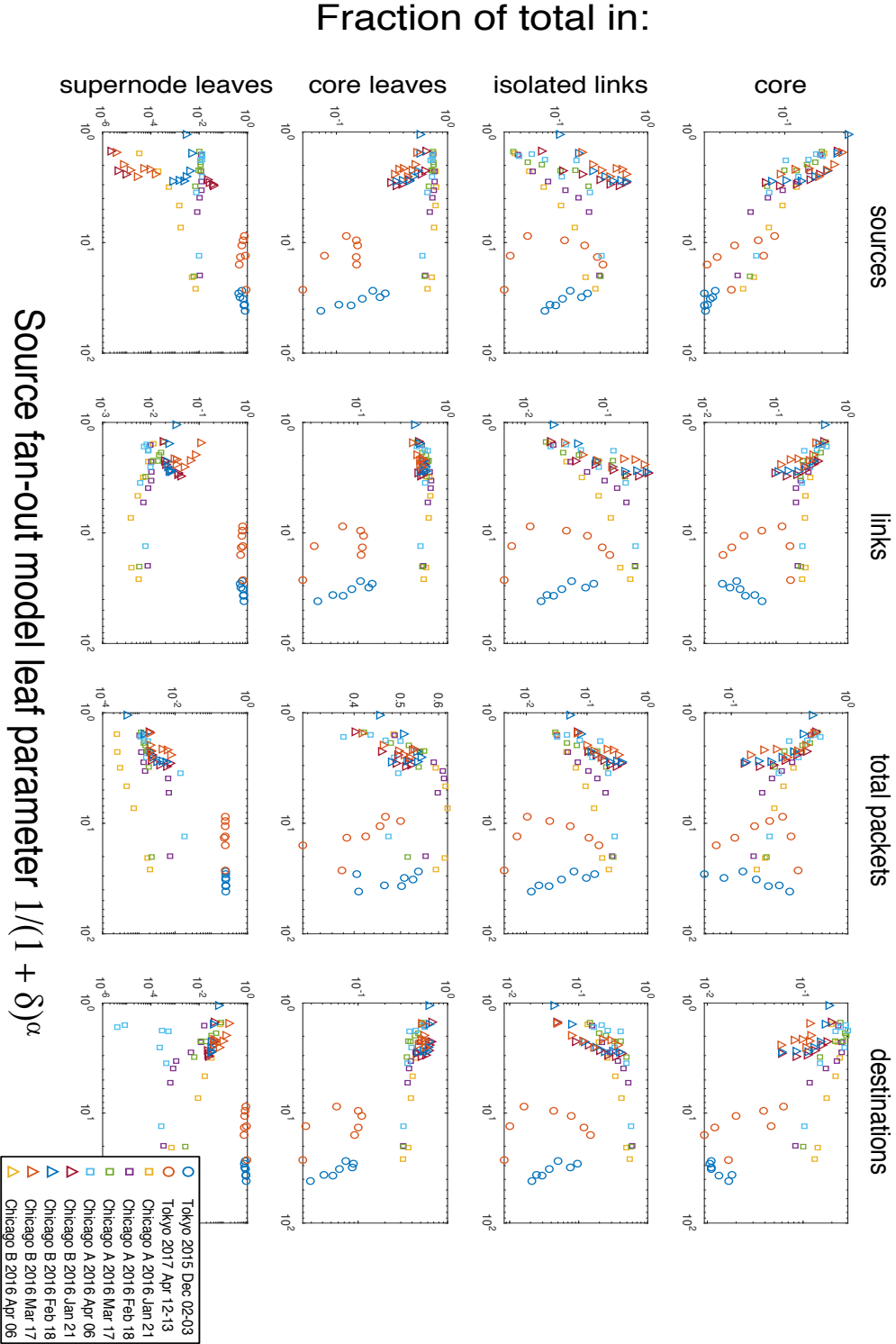


Figure 1.24 **Topology versus modified Zipf–Mandelbrot model leaf parameter.** Fraction of data in the core, isolated links, core leaves, and supernode leaves versus the model leaf parameter computed from the source fan-out modified Zipf–Mandelbrot model fits for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

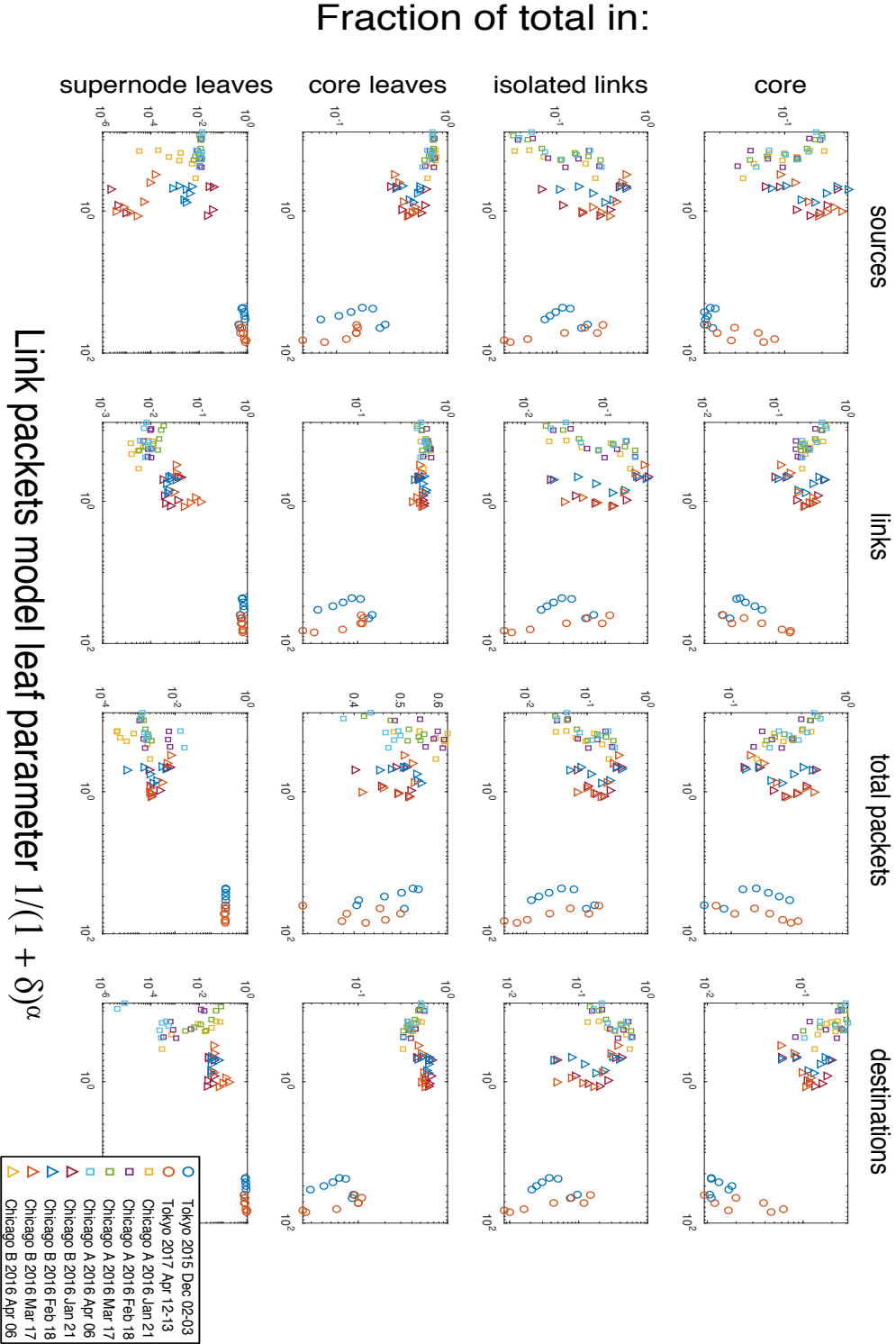


Figure 1.25 **Topology versus modified Zipf–Mandelbrot model leaf parameter.** Fraction of data in the core, isolated links, core leaves, and supernode leaves versus the model leaf parameter computed from the link packets modified Zipf–Mandelbrot model fits for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

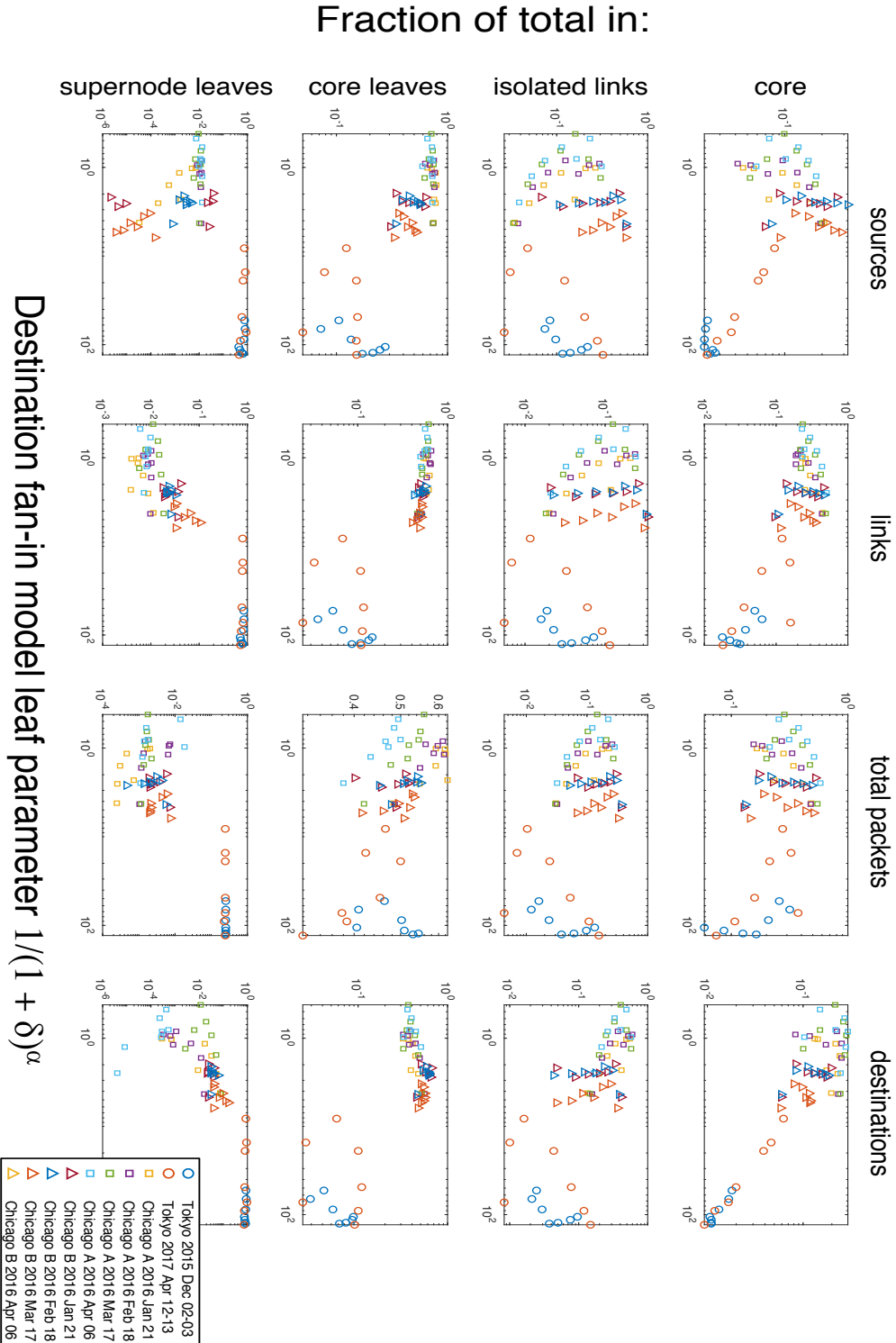


Figure 1.26 **Topology versus modified Zipf–Mandelbrot model leaf parameter.** Fraction of data in the core, isolated links, core leaves, and supernode leaves versus the model leaf parameter computed from the destination fan-out modified Zipf–Mandelbrot model fits for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

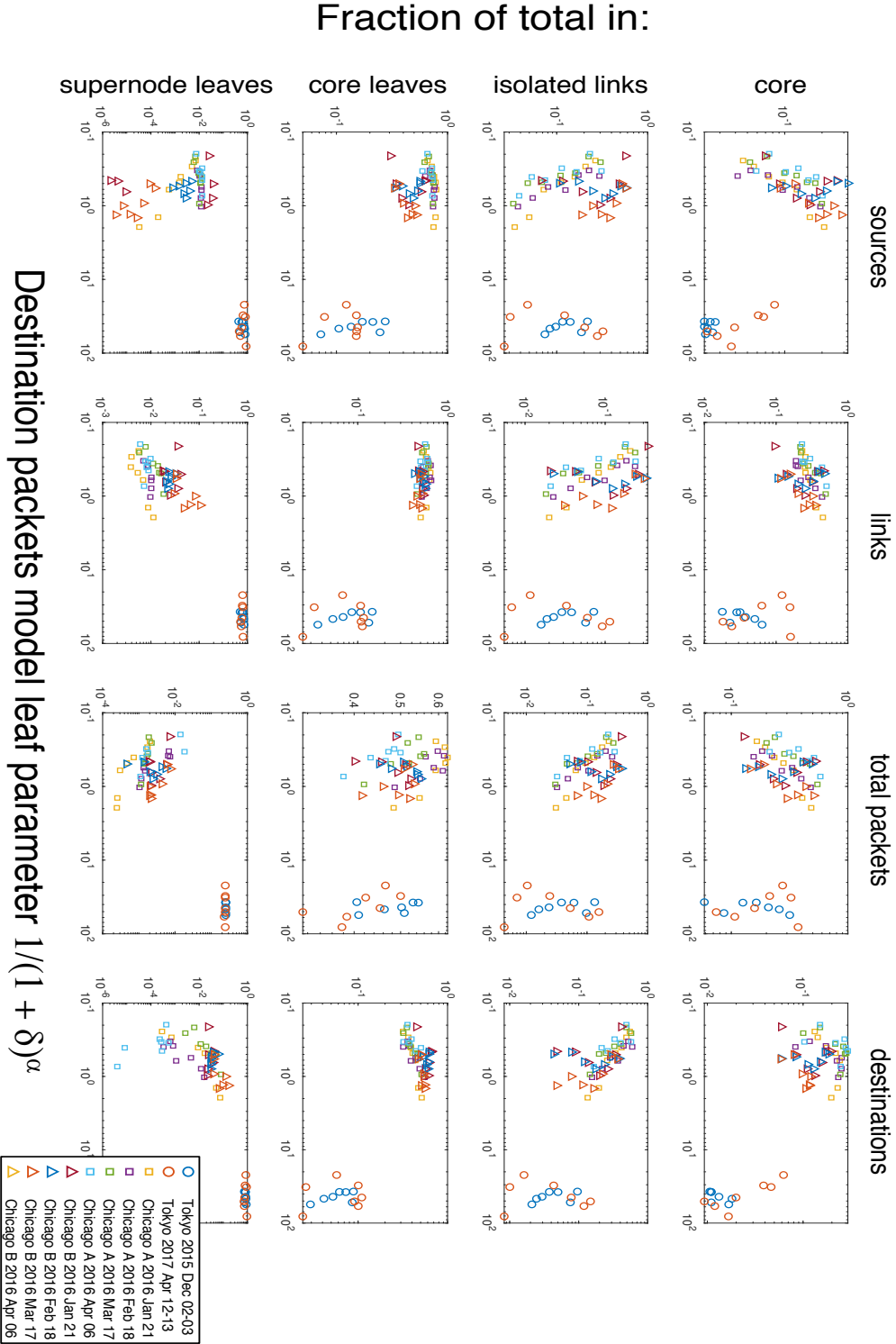


Figure 1.27 **Topology versus modified Zipf–Mandelbrot model leaf parameter.** Fraction of data in the core, isolated links, core leaves, and supernode leaves versus the model leaf parameter computed from the destination packets modified Zipf–Mandelbrot model fits for each location, time, and seven packet windows ($N_V = 10^5, \dots, 10^8$).

Bibliography

- [1] Dimitris Achlioptas, Raissa M. D'souza, and Joel Spencer. Explosive percolation in random networks. *Science*, 323(5920):1453–1455, 2009.
- [2] Lada A. Adamic and Bernardo A. Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115, 2000.
- [3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.
- [4] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130, 1999.
- [5] Mark Allman, Vern Paxson, and Jeff Terrell. A brief history of scanning. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement*, pages 77–82. ACM, 2007.
- [6] Chris G. Antonopoulos and Yilun Shang. Opinion formation in multiplex networks with general initial distributions. *Scientific Reports*, 8(1):2852, 2018.
- [7] David A. Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner. *Graph Partitioning and Graph Clustering*, volume 588. American Mathematical Society, Ann Arbor, 2013.
- [8] Justin Balthrop, Stephanie Forrest, Mark E.J. Newman, and Matthew M. Williamson. Technological networks and the spread of computer viruses. *Science*, 304(5670):527–529, 2004.
- [9] Albert-László Barabási. Scale-free networks: A decade and beyond. *Science*, 325(5939):412–413, 2009.
- [10] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [11] Albert-László Barabási et al. *Network Science*. Cambridge University Press, Cambridge, 2016.
- [12] Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.
- [13] Vineet Bharti, Pankaj Kankar, Lokesh Setia, Gonca Gürsun, Anukool Lakhina, and Mark Crovella. Inferring invisible traffic. In *Proceedings of the 6th International Conference*, page 22. ACM, 2010.

- [14] Ginestra Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.
- [15] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I. Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.
- [16] Marián Boguná and Dmitri Krioukov. Navigating ultrasmall worlds in ultra-short time. *Physical Review Letters*, 102(5):058701, 2009.
- [17] Marian Boguna, Dmitri Krioukov, and Kimberly C. Claffy. Navigability of complex networks. *Nature Physics*, 5(1):74, 2009.
- [18] Marián Boguná, Fragkiskos Papadopoulos, and Dmitri Krioukov. Sustaining the internet with hyperbolic mapping. *Nature Communications*, 1:62, 2010.
- [19] Tom Bohman. Emergence of connectivity in networks. *Evolution*, 11:13, 2009.
- [20] Pierre Borgnat, Guillaume Dewaele, Kensuke Fukuda, Patrice Abry, and Kenjiro Cho. Seven years and one day: Sketching the evolution of internet traffic. In *INFOCOM 2009, IEEE*, pages 711–719. IEEE, 2009.
- [21] Maria Brbic and Ivica Kopriva. ℓ_0 -motivated low-rank sparse subspace clustering. *IEEE Transactions on Cybernetics*, 50:1–15, 2018.
- [22] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.
- [23] Jing Cao, Yu Jin, Aiyu Chen, Tian Bu, and Z-L Zhang. Identifying high cardinality internet hosts. In *INFOCOM 2009, IEEE*, pages 810–818. IEEE, 2009.
- [24] Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.
- [25] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. Are we one hop away from a better internet? In *Proceedings of the 2015 Internet Measurement Conference*, pages 523–529. ACM, 2015.
- [26] Kenjiro Cho. Recursive lattice search: Hierarchical heavy hitters revisited. In *Proceedings of the 2017 Internet Measurement Conference*, pages 283–289. ACM, 2017.
- [27] Kenjiro Cho, Kensuke Fukuda, Hiroshi Esaki, and Akira Kato. The impact and implications of the growth in residential user-to-user traffic. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 207–218. ACM, 2006.

- [28] Kenjiro Cho, Kensuke Fukuda, Hiroshi Esaki, and Akira Kato. Observing slow crustal movement in residential user traffic. In *Proceedings of the 2008 ACM CoNEXT Conference*, page 12. ACM, 2008.
- [29] Kenjiro Cho, Koushirou Mitsuya, and Akira Kato. Traffic data repository at the wide project. In *Proceedings of USENIX 2000 Annual Technical Conference: FREENIX Track*, pages 263–270, 2000.
- [30] Kimberly C. Claffy. Internet tomography. *Nature, Web Matter*, 1999.
- [31] Kimberly C. Claffy. Measuring the internet. *IEEE Internet Computing*, 4(1):73–75, 2000.
- [32] Kimberly C. Claffy, Hans-Werner Braun, and George C. Polyzos. Tracking long-term growth of the nsfnet. *Communications of the ACM*, 37(8):34–45, 1994.
- [33] David Clark et al. The 9th workshop on active internet measurements (aims-9) report. *ACM SIGCOMM Computer Communication Review*, 47(5):35–38, 2017.
- [34] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
- [35] Alberto Dainotti, Antonio Pescapè, and Kimberly C. Claffy. Issues and future directions in traffic classification. *IEEE Network*, 26(1), 2012.
- [36] Amogh Dhamdhere, David D. Clark, Alexander Gamero-Garrido, Matthew Luckie, Ricky K.P. Mok, Gautam Akiwate, Kabir Gogia, Vaibhav Bajpai, Alex C. Snoeren, and Kimberly C. Claffy. Inferring persistent interdomain congestion. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 1–15. ACM, 2018.
- [37] Amogh Dhamdhere and Constantine Dovrolis. The internet is flat: modeling the transition from a transit hierarchy to a peering mesh. In *Proceedings of the 6th International Conference*, page 21. ACM, 2010.
- [38] *NIST Digital Library of Mathematical Functions, Release 1.0.20 of 2018-09-15*. <http://dlmf.nist.gov/25.11>. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.
- [39] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
- [40] Constantinos Dovrolis, Parameswaran Ramanathan, and David Moore. What do packet dispersion techniques measure? In *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 905–914. IEEE, 2001.

- [41] Constantinos Dovrolis, Parameswaran Ramanathan, and David Moore. Packet-dispersion techniques and a capacity-estimation methodology. *IEEE/ACM Transactions on Networking*, 12(6):963–977, 2004.
- [42] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29-4, pages 251–262. ACM, 1999.
- [43] Jinliang Fan, Jun Xu, Mostafa H. Ammar, and Sue B. Moon. Prefix-preserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme. *Computer Networks*, 46(2):253–272, 2004.
- [44] Osvaldo Fonseca, Elvertton Fazzion, Italo Cunha, Pedro Henrique Bragioni Las-Casas, Dorgival Guedes, Wagner Meira, Cristine Hoepers, Klaus Steding-Jessen, and Marcelo H.P. Chaves. Measuring, characterizing, and avoiding spam traffic costs. *IEEE Internet Computing*, 20(4):16–24, 2016.
- [45] Romain Fontugne, Patrice Abry, Kensuke Fukuda, Darryl Veitch, Kenjiro Cho, Pierre Borgnat, and Herwig Wendt. Scaling in internet traffic: A 14 year and 3 day longitudinal study, with multiscale analyses and random projections. *IEEE/ACM Transactions on Networking (TON)*, 25(4):2152–2165, 2017.
- [46] Romain Fontugne, Cristel Pelsser, Emile Aben, and Randy Bush. Pinpointing delay and forwarding anomalies using large-scale traceroute measurements. In *Proceedings of the 2017 Internet Measurement Conference*, pages 15–28. ACM, 2017.
- [47] Vijay Gadepally, Jeremy Kepner, Lauren Milechin, William Arcand, David Bestor, Bill Bergeron, Chansup Byun, Matthew Hubbell, Micheal Houle, Micheal Jones, et al. Hyperscaling internet graph analysis with d4m on the mit supercloud. *IEEE High Performance Extreme Computing Conference (HPEC)*, 2018.
- [48] John Heidemann, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, Genevieve Bartlett, and Joseph Bannister. Census and survey of the visible internet. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, pages 169–182. ACM, 2008.
- [49] Martin Hilbert and Priscila López. The world’s technological capacity to store, communicate, and compute information. *Science*, 332:60, 2011. doi:10.1126/science.1200970.
- [50] Petter Holme. Modern temporal network theory: A colloquium. *The European Physical Journal B*, 88(9):234, 2015.
- [51] Martin Husák, Nataliia Neshenko, Morteza Safaei Pour, Elias Bou-Harb, and Pavel Celeda. Assessing internet-wide cyber situational awareness of critical sectors. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, ARES 2018, pages 29:1–29:6, New York, 2018. ACM.

- [52] Mobin Javed, Cormac Herley, Marcus Peinado, and Vern Paxson. Measurement and analysis of traffic exchange services. In *Proceedings of the 2015 Internet Measurement Conference*, pages 1–12. ACM, 2015.
- [53] Thomas Karagiannis, Andre Broido, Nevil Brownlee, Kimberly C. Claffy, and Michalis Faloutsos. Is p2p dying or just hiding? In *IEEE Global Telecommunications Conference, 2004, GLOBECOM'04.*, volume 3, pages 1532–1538. IEEE, 2004.
- [54] Thomas Karagiannis, Andre Broido, Michalis Faloutsos, et al. Transport layer identification of p2p traffic. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, pages 121–134. ACM, 2004.
- [55] Juha Karvanen and Andrzej Cichocki. Measuring sparseness of noisy signals. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 125–130, 2003.
- [56] Jeremy Kepner. *Parallel MATLAB for Multicore and Multinode Computers*. SIAM, 2009.
- [57] Jeremy Kepner and John Gilbert. *Graph Algorithms in the Language of Linear Algebra*. SIAM, Philadelphia, PA, 2011.
- [58] Jeremy Kepner and Hayden Jananthan. *Mathematics of Big Data: Spreadsheets, Databases, Matrices, and Graphs*. MIT Press, Cambridge, MA, 2018.
- [59] Hyunchul Kim, Kimberly C. Claffy, Marina Fomenkov, Dhiman Barman, Michalis Faloutsos, and KiYoung Lee. Internet traffic classification demystified: Myths, caveats, and the best practices. In *Proceedings of the 2008 ACM CoNEXT Conference*, page 11. ACM, 2008.
- [60] Maksim Kitsak, Ahmed Elmokashfi, Shlomo Havlin, and Dmitri Krioukov. Long-range correlations and memory in the dynamics of internet interdomain routing. *PloS one*, 10(11):e0141481, 2015.
- [61] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888, 2010.
- [62] Tadayoshi Kohno, Andre Broido, and Kimberly C. Claffy. Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing*, 2(2):93–108, 2005.
- [63] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
- [64] Christopher J. Koliba, Jack W. Meek, Asim Zia, and Russell W. Mills. *Governance Networks in Public Administration and Public Policy*. Routledge, Abingdon, 2018.

- [65] Dmitri Krioukov, Maksim Kitsak, Robert S. Sinkovits, David Rideout, David Meyer, and Marián Boguñá. Network cosmology. *Scientific Reports*, 2:793, 2012.
- [66] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.
- [67] Dmitri Krioukov, Fragkiskos Papadopoulos, Amin Vahdat, and Marián Boguñá. Curvature and temperature of complex networks. *Physical Review E*, 80(3):035101, 2009.
- [68] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. Internet inter-domain traffic. *ACM SIGCOMM Computer Communication Review*, 41(4):75–86, 2011.
- [69] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking (TON)*, 2(1):1–15, 1994.
- [70] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 177–187. ACM, 2005.
- [71] Aming Li, Sean P. Cornelius, Y.-Y. Liu, Long Wang, and A.-L. Barabási. The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046, 2017.
- [72] Bingdong Li, Jeff Springer, George Bebis, and Mehmet Hadi Gunes. A survey of network flow applications. *Journal of Network and Computer Applications*, 36(2):567–581, 2013.
- [73] Lun Li, David Alderson, Walter Willinger, and John Doyle. A first-principles approach to understanding the internet’s router-level topology. In *ACM SIGCOMM Computer Communication Review*, volume 34, pages 3–14. ACM, 2004.
- [74] Gustav Lindmark and Claudio Altafini. Minimum energy control for complex networks. *Scientific Reports*, 8(1):3188, 2018.
- [75] Matthias Lischke and Benjamin Fabian. Analyzing the bitcoin network: The first four years. *Future Internet*, 8(1):7, 2016.
- [76] Alex X. Liu, Chad R. Meiners, and Eric Torng. Tcam razor: A systematic approach towards minimizing packet classifiers in tcams. *IEEE/ACM Transactions on Networking (TON)*, 18(2):490–500, 2010.
- [77] Alex X. Liu, Chad R. Meiners, and Eric Torng. Packet classification using binary content addressable memory. *IEEE/ACM Transactions on Networking*, 24(3):1295–1307, 2016.

- [78] Yang-Yu Liu and Albert-László Barabási. Control principles of complex systems. *Reviews of Modern Physics*, 88(3):035006, 2016.
- [79] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016.
- [80] Andrew Lumsdaine, Douglas Gregor, Bruce Hendrickson, and Jonathan Berry. Challenges in parallel graph processing. *Parallel Processing Letters*, 17(01):5–20, 2007.
- [81] Aniket Mahanti, Niklas Carlsson, Anirban Mahanti, Martin Arlitt, and Carey Williamson. A tale of the tails: Power-laws in internet measurements. *IEEE Network*, 27(1):59–64, 2013.
- [82] Benoit Mandelbrot. An informational theory of the statistical structure of language. *Communication Theory*, 84:486–502, 1953.
- [83] Alberto Medina, Ibrahim Matta, and John Byers. On the origin of power laws in internet topologies. *ACM SIGCOMM Computer Communication review*, 30(2):18–28, 2000.
- [84] Jelena Mirkovic, Genevieve Bartlett, John Heidemann, Hao Shi, and Xiyue Deng. Do you see me now? sparsity in passive observations of address liveness. In *Network Traffic Measurement and Analysis Conference (TMA)*, 2017, pages 1–9. IEEE, 2017.
- [85] http://www.caida.org/data/passive/trace_stats/.
- [86] <http://www.caida.org/data/publications/>.
- [87] Marcelo A. Montemurro. Beyond the zipf–mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3-4):567–578, 2001.
- [88] David Moore, Vern Paxson, Stefan Savage, Colleen Shannon, Stuart Staniford, and Nicholas Weaver. Inside the slammer worm. *IEEE Security & Privacy*, 99(4):33–39, 2003.
- [89] David Moore, Colleen Shannon, Douglas J. Brown, Geoffrey M. Voelker, and Stefan Savage. Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems (TOCS)*, 24(2):115–139, 2006.
- [90] David Moore, Colleen Shannon, et al. Code-red: A case study on the spread and victims of an internet worm. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement*, pages 273–284. ACM, 2002.
- [91] David Moore, Colleen Shannon, Geoffrey M. Voelker, and Stefan Savage. Internet quarantine: Requirements for containing self-propagating code. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 3, pages 1901–1910. IEEE, 2003.

- [92] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.
- [93] Mark E.J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.
- [94] Christopher Olston, Marc Najork, et al. Web crawling. *Foundations and Trends® in Information Retrieval*, 4(3):175–246, 2010.
- [95] Robert Paluch, Xiaoyan Lu, Krzysztof Suchecki, Bolesław K. Szymański, and Janusz A. Hołyst. Fast and accurate detection of spread source in large complex networks. *Scientific Reports*, 8(1):2508, 2018.
- [96] Fragkiskos Papadopoulos, Maksim Kitsak, M. Ángeles Serrano, Marián Boguñá, and Dmitri Krioukov. Popularity versus similarity in growing networks. *Nature*, 489(7417):537, 2012.
- [97] Vern Paxson. End-to-end routing behavior in the internet. *ACM SIGCOMM Computer Communication Review*, 26(4):25–38, 1996.
- [98] Vern Paxson and Sally Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking (ToN)*, 3(3):226–244, 1995.
- [99] Matjaz Perc, Jillian J. Jordan, David G. Rand, Zhen Wang, Stefano Boccaletti, and Attila Szolnoki. Statistical physics of human cooperation. *Physics Reports*, 687:1–51, 2017.
- [100] Ravi Prasad, Constantinos Dovrolis, Margaret Murray, and Kimberly C. Claffy. Bandwidth estimation: metrics, measurement techniques, and tools. *IEEE Network*, 17(6):27–35, 2003.
- [101] Michael Rabinovich and Mark Allman. Measuring the internet. *IEEE Internet Computing*, 20(4):6–8, 2016.
- [102] Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, et al. Interactive supercomputing on 40,000 cores for machine learning and data analysis. *IEEE High Performance Extreme Computing Conference (HPEC)*, 2018.
- [103] Naoki Saito, Brons M. Larson, and Bertrand Bénichou. Sparsity vs. statistical independence from a best-basis viewpoint. In *Wavelet Applications in Signal and Image Processing VIII*, volume 4119, pages 474–487. International Society for Optics and Photonics, 2000.
- [104] Osama Saleh and Mohamed Hefeeda. Modeling and caching of peer-to-peer traffic. In *Proceedings of the 2006 14th IEEE International Conference on Network Protocols, ICNP’06*, pages 249–258. IEEE, 2006.

- [105] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.
- [106] Paul Sheridan and Taku Onodera. A preferential attachment paradox: How preferential attachment combines with growth to produce networks with log-normal in-degree distributions. *Scientific Reports*, 8(1):2811, 2018.
- [107] Augustin Soule, Antonio Nucci, Rene Cruz, Emilio Leonardi, and Nina Taft. How to identify and estimate the largest traffic matrix elements in a dynamic environment. In *ACM SIGMETRICS Performance Evaluation Review*, volume 32, pages 73–84. ACM, 2004.
- [108] Neil Spring, Ratul Mahajan, and David Wetherall. Measuring isp topologies with rocketfuel. *ACM SIGCOMM Computer Communication Review*, 32(4):133–145, 2002.
- [109] Michael P.H. Stumpf and Mason A. Porter. Critical truths about power laws. *Science*, 335(6069):665–666, 2012.
- [110] Paul Tune, Matthew Roughan, H. Haddadi, and O. Bonaventure. Internet traffic matrices: A primer. *Recent Advances in Networking*, 1:1–56, 2013.
- [111] Yogesh Virkar and Aaron Clauset. Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, pages 89–119, 2014.
- [112] Zhen Wang, Chris T. Bauch, Samit Bhattacharyya, Alberto d’Onofrio, Piero Manfredi, Matjaz Perc, Nicola Perra, Marcel Salathé, and Dawei Zhao. Statistical physics of vaccination. *Physics Reports*, 664:1–113, 2016.
- [113] Walter Willinger, David Alderson, and John C. Doyle. Mathematics and the internet: A source of enormous confusion and great potential. *Notices of the American Mathematical Society*, 56(5):586–599, 2009.
- [114] Walter Willinger, Ramesh Govindan, Sugih Jamin, Vern Paxson, and Scott Shenker. Scaling phenomena in the internet: Critically examining criticality. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2573–2580, 2002.
- [115] Walter Willinger, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson. Self-similarity through high-variability: Statistical analysis of ethernet lan traffic at the source level. *IEEE/ACM Transactions on Networking (ToN)*, 5(1):71–86, 1997.
- [116] Zongben Xu, Xiangyu Chang, Fengmin Xu, and Hai Zhang. $l_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1013–1027, 2012.
- [117] Hongbo Dong Yaghoub Rahimi, Chao Wang and Yifei Lous. A scale invariant approach for sparse signal recovery. *arXiv preprint arXiv:1812.08852*, 2018.

- [118] Shui Yu, Guofeng Zhao, Wanchun Dou, and Simon James. Predicted packet padding for anonymous web browsing against traffic analysis attacks. *IEEE Transactions on Information Forensics and Security*, 7(4):1381–1393, 2012.
- [119] Xuecheng Yu and Tianguang Chu. Link prediction from partial observation in scale-free networks. In *Chinese Intelligent Systems Conference*, pages 199–205. Springer, 2017.
- [120] Lixia Zhang, Alexander Afanasyev, Jeffrey Burke, Van Jacobson, Patrick Crowley, Christos Papadopoulos, Lan Wang, Beichuan Zhang, et al. Named data networking. *ACM SIGCOMM Computer Communication Review*, 44(3):66–73, 2014.
- [121] Yin Zhang, Matthew Roughan, Carsten Lund, and David L. Donoho. Estimating point-to-point and point-to-multipoint traffic matrices: An information-theoretic approach. *IEEE/ACM Transactions on Networking (TON)*, 13(5):947–960, 2005.