On the Lack of Typical Behavior in the Global Web Traffic Network

Mark Meiss
mmeiss@indiana.edu
Dept. of Computer Science
and Adv. Network Mgmt. Lab

Filippo Menczer
fil@indiana.edu
School of Informatics and
Dept. of Computer Science

Alessandro Vespignani
alexv@indiana.edu
School of Informatics

Indiana University
Bloomington, IN 47405

ABSTRACT

We offer the first large-scale analysis of Web traffic based on network flow data. Using data collected on the Internet2 network, we constructed a weighted bipartite client-server host graph containing more than 18×10^6 vertices and 68×10^6 edges valued by relative traffic flows. When considered as a traffic map of the World-Wide Web, the generated graph provides valuable information on the statistical patterns that characterize the global information flow on the Web. Statistical analysis shows that client-server connections and traffic flows exhibit heavy-tailed probability distributions lacking any typical scale. In particular, the absence of an intrinsic average in some of the distributions implies the absence of a prototypical scale appropriate for server design, Web-centric network design, or traffic modeling. The inspection of the amount of traffic handled by clients and servers and their number of connections highlights non-trivial correlations between information flow and patterns of connectivity as well as the presence of anomalous statistical patterns related to the behavior of users on the Web. The results presented here may impact considerably the modeling, scalability analysis, and behavioral study of Web applications.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Network topology*; C.2.2 [**Computer-Communication Networks**]: Network Protocols—*HTTP*; C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network management, Network monitoring, Public networks*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks, Performance evaluation (efficiency), WWW*

General Terms

Measurement, Networks

Keywords

Web usage, traffic statistics, network flows, degree, strength, scale-free networks, power laws

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-046-9/05/0005.

1. INTRODUCTION

In recent years the research community has devoted a large effort to the understanding and characterization of the World-Wide Web, in the attempt to provide a theoretical and practical understanding of what has been termed the *ecology of information* [11]. A fundamental step in this direction is represented by experiments aimed at studying the graph structure of the Web, in which vertices and directed edges identify Web pages and hyperlinks, respectively. These studies are based on crawlers that explore the connectivity of the Web by following the links on each discovered page, thus reconstructing the topological properties of the visited graph. In particular, data gathered in large-scale crawls [3, 7, 14, 1, 15] have indicated the presence of a rich and complex architecture underlying the structure of the Web graph.

Foremost among these properties have been the in-degree and out-degree distributions of the network, where *in-degree* represents the number of Web pages that link to a particular page, and *out-degree* represents the number of links that a page contains to other pages. These distributions turn out to have heavy tails, showing close fits to power-law distributions over several orders of magnitude. This last feature is the signature of complex topological properties with statistical fluctuations that extend over many length scales; they are not exclusive to the Web and can be found in a wide range of network structures spanning various domains such as ecology, biology, and social and technological systems [2, 4, 9, 17].

While there is much to be learned from the link structure of the Web—it forms the basis of Brin and Page's PageRank algorithm [6]—it does not tell us about what people actually do when browsing the Web. Indeed, it has been recognized that the complexity of the Web encompasses not only its topology but also the dynamics of information. Examples of this complexity are navigation patterns, community structures, congestion, and other social phenomena resulting from users' behavior [12, 13, 1, 17]. In order to gain this kind of insight, a variety of usage data must be studied, such as server logs, hit counts, and router statistics. These sources of information tell us about the *behavioral* network of the World-Wide Web, in which the nodes correspond to individual hosts on the Internet and the edges correspond to actual HTTP transfers among these hosts. Network managers and capacity planners are accustomed to this view of the Web;

numerous tools exist for analyzing server logs, determining trends in the quantity of HTTP traffic on a network, and so forth. These tools offer high-level insight into such aspects of user behavior as what pages are most popular, what referring sites are most common, what percentage of traffic in a local network is devoted to Web traffic, etc.

The aforementioned studies, however, do not consider the statistical properties of the Web from the perspective of an international transit network that serves as a conduit rather than an endpoint for Web traffic. Such a study amounts to a global large-scale investigation of the behavioral network itself and represents the focus of the present work. In particular, we discuss the properties of a very large sample of Web-related network flow data taken from the Internet2 (Abilene) network. The data collected allow us to construct a client-server interaction network whose weighted connections characterize the traffic flows. To our knowledge, this is the first large-scale weighted graph representation of the Web interaction network and its traffic. A valued bipartite graph is used to represent the network mathematically, and a thorough analysis is performed to uncover the statistical laws characterizing the patterns of traffic. We find that client-server interaction patterns show marked scale-free properties, with statistical distributions of traffic properties varying over a wide range of length scales. In some cases, scaling over eight orders of magnitude is observed in conjunction with a surprisingly slow decay of the distribution tails, indicating the presence of unbounded fluctuations in all measures characterizing the behavior of Web traffic. These fluctuations in traffic and connectivity patterns suggest several questions related to the planning and modeling of Web traffic. Finally, we provide evidence that large-scale analysis of traffic might result in a useful tool for the statistical detection of anomalous patterns related to malicious and exploratory use of the Web.

2. COLLECTION OF FLOW DATA

In order to gather data on the global Web traffic we performed a set of passive measurements on the Abilene (Internet2) network. The Abilene network is an TCP/IP data network that provides high-speed Internet connectivity to research laboratories, colleges, and universities throughout the United States.[1] The backbone of the network consists of 10-Gbps fiberoptic links connecting eleven high-performance routers located in major metropolitan areas such as Los Angeles, Chicago, and New York City. Individual institutions connect to Abilene either directly or through large regional connectors. Abilene carries only academic and research traffic; participants must maintain their own separate connections to the commodity Internet. As of this writing, over 200 universities and corporate research laboratories within the United States connect to the Abilene network. In addition, Abilene also provides transit for data from dozens of international academic and research networks, particularly between Pacific Rim nations and Europe. The traffic thus includes not only data to and from hosts on Abilene, but also international data routed across Abilene. While the network provides native support for newer protocols such as IPv6, the great majority of all hosts on the Abilene network use IPv4, just as in the commodity Internet.

Several properties of the Abilene network make it an ideal environment for studying network traffic. As a wide-area transit network that includes both domestic and international traffic, it offers a global view of the Internet unavailable in many smaller networks. It also has a heterogeneous user base that includes hundreds of thousands of high-spirited undergraduates as well as researchers and college faculty. Finally, even during peak hours, the Abilene network is never congested, which offers a view of what users do when the network itself does not impede their behavior. Typical traffic levels in the network can be seen in Figure 1.[2]

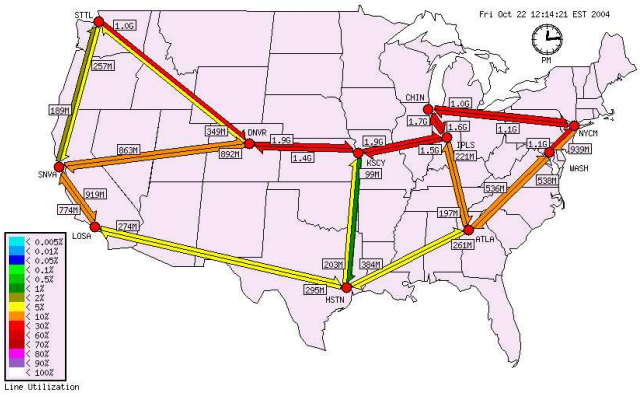

Figure 1: Typical activity levels between core routers in the Abilene network. The numbers refer to sustained data rates measured in bits per second.

Our passive measurement strategy is based on information about the traffic handled by the network that comes in the form of *flow records* generated by the core routers and sent over the network to management systems. Each flow record contains information about a single network flow, which is defined as one or more packets sent from a particular source host and port, to a particular destination host and port, using a particular protocol, over some time interval. In the case of most Web traffic, either the source or the destination port will be 80, the assigned port number for HTTP. The routers do not have capacity for generating full information on every flow in the network; instead, they must sample the data, which is done periodically at a rate of approximately one in a hundred packets. Note that even though TCP connections involve bidirectional traffic, network flows as defined above are only unidirectional. Thus, every TCP connection potentially generates *two* flows: one from the client to the server, and another from the server to the client. Sampling implies that we may see only one of these flows, or possibly neither.

The network flow records contain a variety of information, which is described in detail in Cisco's documentation on the "netflow-v5" format.[3] The relevant fields for the analysis at hand are the source and destination IP addresses, the source and destination ports, and the total number of bytes in the flow. In conformance with the privacy policies of Internet2, we do not examine the actual source and destination IP addresses found in the flow records; instead, they are replaced with index values that maintain their identity only over the

[1] http://abilene.internet2.edu/

[2] http://loadrunner.uits.iu.edu/weathermaps/abilene/

[3] http://www.cisco.com/univercd/cc/td/doc/product/rtrmgmt/nfc/nfc_3_0/nfc_ug/nfcform.htm

course of a single day. This index is stored only in system memory and is discarded at the end of the day. The index values are unique across the entire set of core routers.

Even with the routers sampling one packet in a hundred, the total amount of flow data collected is substantial. On a typical weekday, the Abilene routers produce between 700 and 800 million of these flow records, of which around 40 percent involve HTTP (i.e., TCP connections on port 80). At 48 bytes per flow record, this means that a full day of flow data consumes about 35 GB of disk space and arrives at a mean rate of 3.4 Mbps.

The analysis in this paper is based on a full 24-hour day of network flow data captured and saved to disk starting at midnight Eastern Standard Time (UTC-5) on September 30, 2004. During this interval, we collected information on approximately 742 million flows between almost 30 million individual hosts. Of these flows, about 319 million (43%) involved TCP connections with an endpoint on port 80, which we take to be indicative of Web traffic.

3. A WEIGHTED GRAPH REPRESENTATION OF WEB TRAFFIC

The data collected have been used to construct a graph formed by considering hosts involved in Web flows as vertices (nodes) and the directed aggregate traffic between pairs of hosts as edges. The resulting graph contains 18.5 million nodes and 68.1 million edges. This view of the data, however, makes no distinction between Web servers and clients. For the sake of consistency in statistical analysis, it is thus more appropriate to partition the graph into subsets. One subset $C = \{i_1, i_2, \cdots, i_{N_C}\}$ identifies with each vertex i a host who has acted as client, and the second subset $S = \{j_1, j_2, \cdots, j_{N_S}\}$ consists of hosts that have acted as servers. The data clearly indicate that a number of hosts act both as server and clients; these hosts are therefore represented in both sets. Each flow record contributes a directed edge representing a server-to-client (j, i) or client-to-server (i, j) connection, depending on whether the source or destination uses port 80. To each edge we assign a *weight* indicating the aggregate amount of data associated with that pair of hosts in the network. Each weight w_{ij} is the total amount of sampled data sent from a particular client to a particular server over the course of the day. Similarly, the weights w_{ji} give the total amount of data sent from a particular server to a particular client. This graph representation results in a weighted bipartite digraph, since we have two disjoint subsets C and S such that each directed edge connects only a vertex from C to one from S, or vice-versa [8]. In Figure 2 we show a pictorial representation of the obtained bipartite digraph.

From the analysis of this weighted graph we can derive a number of characteristic quantities and basic statistical distributions [5, 16]:

- The number of servers $n_S(i)$ with which each client communicates in server-to-client and client-to-server connections. We measure in like manner the number of clients $n_C(j)$ handled by each server.
- The in-degree k_{in} of a vertex i (j) is the number of directed edges with i (j) as their terminal vertex. The in-degree of server vertices is therefore the number of clients by each server in client-to-server connections. Similarly, the in-degree of client vertices is the number of servers sending HTTP data to each client.
- The out-degree k_{out} of a vertex i (j) is the number of directed edges with i (j) as their initial vertex. The out-degree of server vertices is again an estimate of the number of clients handled by each server in server-to-client flows. Similarly, the out-degree of client vertices is a measure of the number of servers contacted by each client. It is worth remarking that while k_{in} and k_{out} may differ on a single vertex, the analysis of the in- and out-degree in the two subsets S and C provides the same statistical information concerning the number of clients per server and servers per client, respectively.
- The client out-strength is defined as
$$s_{out}(i) = \sum_j w_{ij},$$
and represents the total number of bytes sent from each client i to servers, which will consist largely of requests and posted form data. Similarly, the client in-strength
$$s_{in}(i) = \sum_j w_{ji}$$
is the amount of data that each client receives from servers (i.e., client downloads).
- The server out-strength is defined as
$$s_{out}(j) = \sum_i w_{ji},$$
and represents the total number of bytes sent from each server j to its clients. The out-strength therefore amounts to the volume of downloads handled by each server. Analogously, server in-strength
$$s_{in}(j) = \sum_i w_{ij}$$
is the amount of incoming data (requests and uploads) handled by each server.

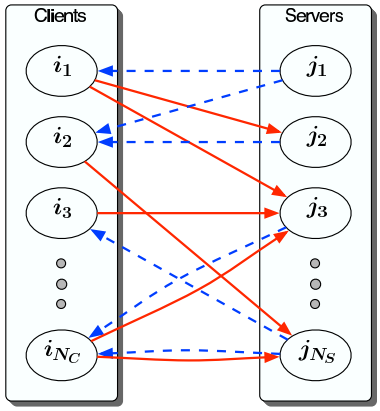

Figure 2: Bipartite digraph representation of the client-server traffic flows. To each directed edge is associated a weight w_{ij} or w_{ji} representing the total amount of data sent from one host to the other.

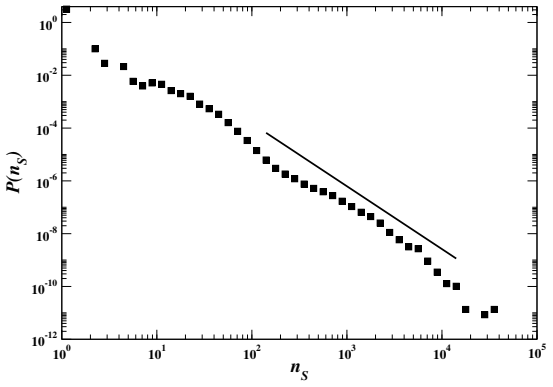

Figure 3: Probability distribution for any given client to contact n_S servers. As a visual guide (solid line) we report the power-law behavior with slope (exponent) -2.4.

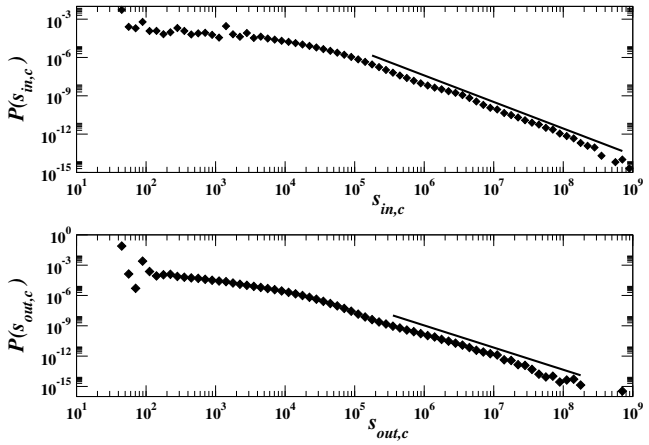

Figure 4: Probability distribution of the total incoming data (in-strength) and total outgoing data (out-strength) of clients. The solid lines illustrate power-law behavior with slopes of -2.2 and -2.1 for the upper and lower graphs, respectively.

In the following section, we will analyze the statistical properties of the graph by looking at clients and servers separately. It is worth remarking, however, that the number of client-to-server connections and server-to-client connections are not identical. In particular, the server-to-client connections involve 2.38 million clients, 148,000 servers, and 17,500 vertices that play both roles. Client-to-server connections are more numerous and involve 18.1 million vertices, of which 17.7 million are clients, 363,000 are servers, and 29,000 are both.[4]

4. DATA ANALYSIS

In this section, we discuss our findings for the properties of the behavioral Web network from two different points of view: the client and the server. In the client section, we are concerned with properties that have implications for the design and modeling of Web browsers, crawlers, and other user agents. In the server section, we are concerned with properties that affect the design and modeling of both Web servers and the networks built to support them.

Unless otherwise noted, our technique for analyzing a distribution involves grouping the data into logarithmically-sized histogram bins normalized by the width of the bin and the size of the distribution so that we are estimating a probability density function. The results are then plotted on a log-log scale with the bin centers on the x-axis and the estimated probability on the y-axis.

4.1 Web Clients

4.1.1 Traffic heterogeneity

The first quantity of interest is the number of servers contacted by each client. This can be extracted from the in-degree and out-degree of client vertices in our bipartite graph. In Figure 3 we report the probability distribution $P(n_S)$ describing the likelihood that any given client contacts n_S servers. For our sample, we get a mean of $\langle n_S \rangle = 3.52$ servers per client and a standard deviation of $\sigma(n_S) = 35.1$. Strikingly, the standard deviation that indicates the level of statistical fluctuation is an order of magnitude larger than the mean value. This is due to the heavy-tailed and skewed probability distribution that matches a power law $P(n_S) \sim n_S^{-\gamma}$, with exponent $\gamma \simeq 2.4 \pm 0.2$, on a range of values spanning several orders of magnitude. For such a distribution, the second moment $\langle n_S^2 \rangle = \int n_S^2 P(n_S) dn$ eventually diverges; the standard deviation is not an intrinsic value of the distribution and is only bounded by the size of the statistical sample. It is clear that in such a case, the average value $\langle n \rangle$ is no longer a typical value, and we lack any characteristic length in the system: this is the so-called "scale-free" behavior. In particular, we have an appreciable probability of finding clients that handle a disproportionate number of servers.

As a confirmation of the scale-free behavior of client connections, we studied also the probability distributions $P(k_{out,c})$ and $P(k_{in,c})$ that any given client has out-degree $k_{out,c}$ and in-degree $k_{in,c}$, respectively. These two distributions refer to number of servers that contacted the client or were contacted by the client. Statistically, it is reasonable to expect the same scaling behavior as obtained for the distribution of total number of servers per client. Indeed, this analysis again recovered power-law behavior with exponents $\gamma_{in} \simeq \gamma_{out} \simeq 2.4$, in agreement with the scaling of $P(n_S)$.

An important measure of client behavior is the total amount of data sent $s_{out,c}$ and received $s_{in,c}$ when interacting with servers. Indeed, the performance evaluations of client applications depend on the typical workloads measured. Yet in this case, we find that the distributions $P(s_{out,c})$ and $P(s_{in,c})$ are extremely broad, with a range of values spanning nine orders of magnitude. In Figure 4, we report both distributions. The heavy-tail behavior is well approximated by a power-law behavior over three to four orders of magnitude where the distribution assume the forms $P(s_{out,c}) \sim s_{out,c}^{-\alpha_{out}}$ and $P(s_{in,c}) \sim s_{in,c}^{-\alpha_{in}}$ with $\alpha_{in} = 2.1 \pm 0.1$ and $\alpha_{out} = 2.2 \pm 0.1$. It is worth mentioning that because of the sampling done on the data flows, the actual traffic values are $s_{in,c} \times 10^2$ and $s_{out,c} \times 10^2$. This multiplicative factor affects neither the performed analysis nor the shape of the

[4]The relatively small number of hosts operating in both client and server roles suggests that the practice of using port 80 to mask peer-to-peer applications is uncommon.

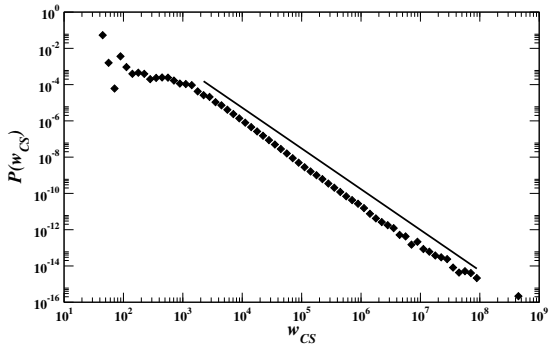

Figure 5: Probability distribution of the client-to-server connection traffic. The solid line has slope -2.3.

distribution. For the in-strength we find a mean value of $\langle s_{in,c} \rangle = 9.28 \times 10^4$ and a standard deviation of $\sigma(s_{in,c}) = 2.05 \times 10^6$. Analogously, we find $\langle s_{out,c} \rangle = 1.11 \times 10^3$ and a standard deviation of $\sigma(s_{out,c}) = 1.44 \times 10^5$. In both cases the standard deviation is two orders of magnitude larger than the average value, showing the lack of any characteristic strength value and the massive heterogeneity in the amount of data handled by client hosts. Also striking is the evidence for similar behavior in both ingoing and outgoing traffic, since the nature of the Web as a broadcast medium led us to expect much greater asymmetry between incoming and outgoing data.

Finally, we can consider the probability distribution of the values w_{ij}. These weights represent the aggregate flow between specific client-server pairs and allow us to study the probability $P(w_{CS})$ that any given connection carries traffic w_{CS}. Figure 5 shows that we have also in this case a heavy-tailed distribution with a best fit to a power-law distribution $P(w_{CS}) \sim w_{CS}^{-\delta}$ with $\delta \approx 2.4 \pm 0.1$. Each weight represents the total amount of data sent from a particular client to a particular server over the course of a full day. The variability of the distribution thus provides evidence for scale-free traffic heterogeneity even at the level of single connections.

4.1.2 Behavioral patterns

In order to provide more insight on behavioral patterns on the Web, let us inspect the correlation between the traffic and the number of connections handled by clients. Intuitively, the strength $s_{out,c}$ ($s_{in,c}$) behaves as an increasing function of the client out-degree $k_{out,c}$ ($k_{in,c}$). The power-law character of the distribution $P(s_{out,c})$ might be considered less surprising from this perspective, as it could arise directly from the power-law behavior of the server-per-client distribution. However, further inspection of the strength behavior uncovers the peculiar nature of the traffic distribution. In Figure 6 (top) we report the behavior of the average in-strength $\langle s_{in,c}(k_{in,c}) \rangle$ for clients with in-degree $k_{in,c}$. Analogously, in Figure 7 (top) we report the behavior of $\langle s_{out,c}(k_{out,c}) \rangle$. In both cases, we find that the strength is increasing as a power law of the degree, yielding the relations

$$\langle s_{in,c}(k_{in,c}) \rangle \sim k_{in,c}^{-\beta_{in}}$$

and

$$\langle s_{out,c}(k_{in,c}) \rangle \sim k_{out,c}^{-\beta_{out}}.$$

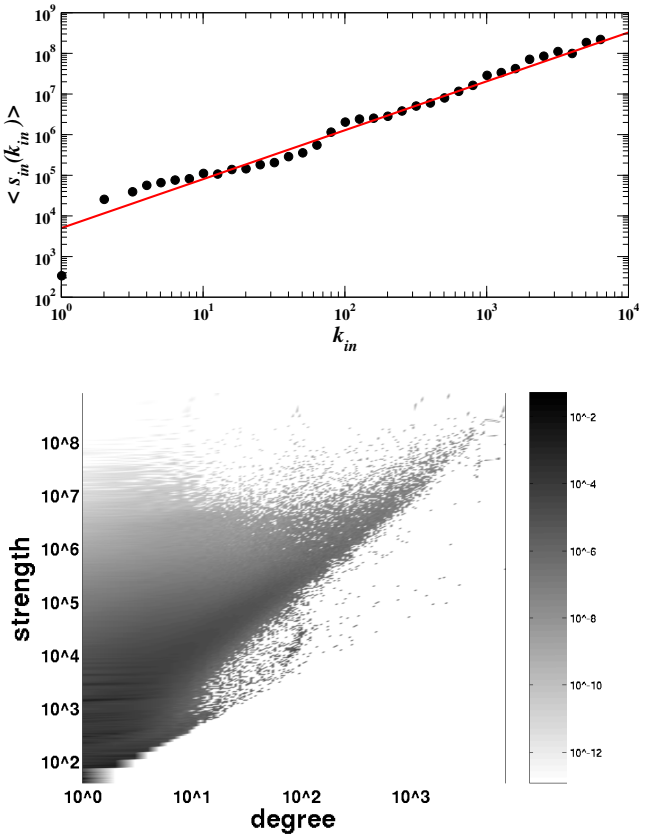

Figure 6: Behavior of the incoming traffic (strength) s_{in} as a function of the number of server-to-client connections (degree) k_{in}. Top: behavior of the average in-strength $\langle s_{in}(k_{in}) \rangle$ as a function of the in-degree. The behavior is linear on a double logarithmic scale and is well approximated by power-law behavior with slope $\beta_{in} \simeq 1.2$. Bottom: frequency of clients with given in-strength for each value of the in-degree. In this and the following distribution maps, tones represent frequencies of strength values, normalized within each degree bin, on a log scale.

A best linear fit of the obtained curves gives $\beta_{in} = 1.2 \pm 0.1$ and $\beta_{out} = 1.2 \pm 0.1$. These two exponents can be considered as a signature of non-trivial correlation between the number of connections and the traffic handled by clients. Indeed, one would expect a linear correlation behavior $\beta = 1$. The super-linear behavior found for clients is therefore a hint of non-linear mechanisms at work in the growth in the amount of downloaded and uploaded data associated with each additional connection handled by a client. These results may prove extremely relevant for the design of scalable client applications.

Interestingly, the exponents β_{in} and β_{out} allow us to relate the degree and strength distribution exponents through a simple scaling argument. For the sake of simplicity, let us consider all variables as continuous and plug in the scaling behavior $s(k) \sim k^{\beta}$ in the strength distribution $P(s)ds \sim s^{-\alpha}$, where the various subscript indexes are implicitly considered. We obtain $P(k)dk \sim k^{-\beta(\alpha-1)-1}dk$, which by def-

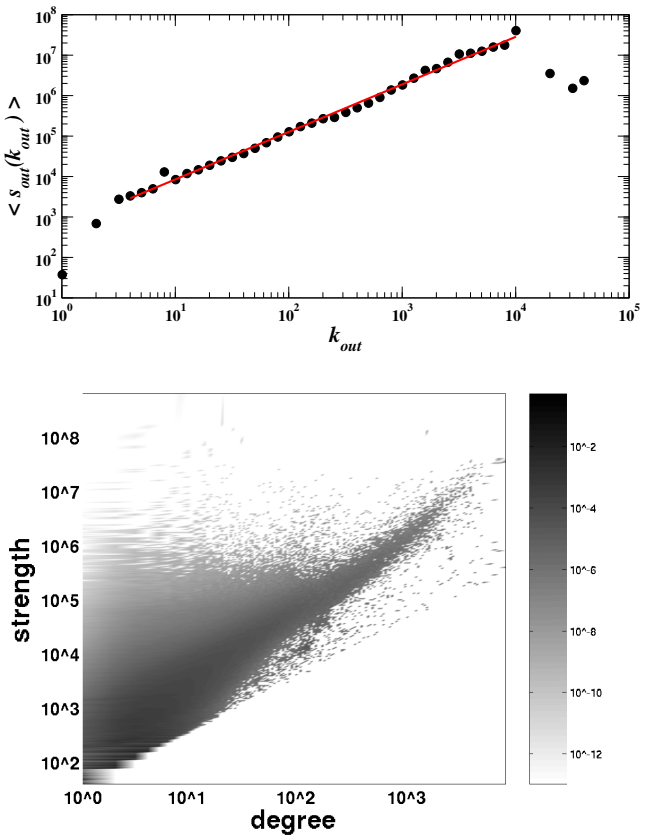

Figure 7: Behavior of the outgoing traffic (strength) s_{out} as a function of the number of client-to-server connections (degree) k_{out}. Top: behavior of the average out-strength $\langle s_{out}(k_{out})\rangle$ as a function of the out-degree. The behavior is linear on a double logarithmic scale and is well approximated by a power-law behavior with slope $\beta_{out} \simeq 1.2$. Bottom: frequency of clients with given out-strength for each value of the out-degree.

inition has to match the behavior $P(k)dk \sim k^{-\gamma}dk$. The comparison of the two scaling behaviors readily provides an equality among exponents that in our case yields two relations:

$$\beta_{out} = \frac{\gamma_{out} - 1}{\alpha_{out} - 1} \quad \text{and} \quad \beta_{in} = \frac{\gamma_{in} - 1}{\alpha_{in} - 1}, \tag{1}$$

obtained by considering the *out* and *in* distributions, respectively. The values obtained empirically for the exponents satisfy the above scaling relations within the error bars for the data, and support a consistent scale-free picture for properties of client behavior.

While the study of traffic as a function of the number of connections ($\langle s_{in,c}(k_{in,c})\rangle$ and $\langle s_{out,c}(k_{out,c})\rangle$) provides relevant information on the scaling of traffic, it is clear that examining the mean value washes out the presence of anomalies and outliers in the population. In particular, for fixed values of $k_{in,c}$ and $k_{out,c}$ we observe a wide variation in the strength associated with different clients. In order to study the spread of strength values with respect to the number of connections, we show in Figures 6 and 7 (bottom) a two-dimensional color plot indicating for each number of

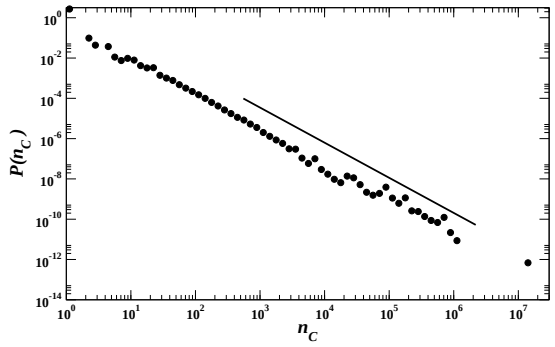

Figure 8: Distribution of the number of clients n_C handled by each server. The solid line has slope -1.8.

connections, the frequency of clients with a given strength. While the plots confirm the accumulation of clients along the power-law distribution studied previously, we also observe the presence of a large spread of strength values for the same number of connections. In addition, we notice the clear presence of outliers falling two to three orders of magnitude below the expected value. These points correspond to clients with a very large number of server connections and a disproportionally small amount of traffic handled. These anomalies might indicate malicious activity or massive scanning strategies. In particular, for outgoing client data, points corresponding to a very large number of client-to-server connections ($k_{out,c}$) and a very low volume of outgoing traffic ($s_{out,c}$) would imply some sort of scanning activity, for example, a project trying to estimate the density of Web servers in assigned IP space.

While further study and analysis are needed to fully exploit the possibilities of the analysis presented here, it appears that the study of large-scale traffic data might result in a viable tool to detect anomalies and behavioral patterns at the global level that may pass unnoticed through more conventional intrusion detection systems.

4.2 Web Servers

Having looked at the behavior of Web clients, and having found that there is no such thing as a *typical* client, we turn to the behavior of Web servers. This analysis is perhaps the more interesting of the two, since information on the demands placed on a Web server is instrumental to both server and network design, as well as modeling Web traffic. We will follow the same general course as we did for client behavior.

4.2.1 Traffic heterogeneity

In the case of Web servers, we again start by analyzing the number of clients n_C that each server handles. For our sample, we get an average of 142 clients per server and a standard deviation of 2.34×10^4. In Figure 8, we report the probability distribution $P(n_C)$ indicating the likelihood that a server handles n_C clients. It is not surprising that we again find a heavy-tailed distribution with power-law behavior, as shown by the very large standard deviation. What is more surprising is that the power-law behavior shows a fit to $P(n_C) \sim n_C^{-\gamma}$ with $\gamma \approx 1.8 \pm 0.1$. It appears that $\gamma < 2$ quite definitively, presenting us with a new scenario. It is known that the first moment of a power-law distribution

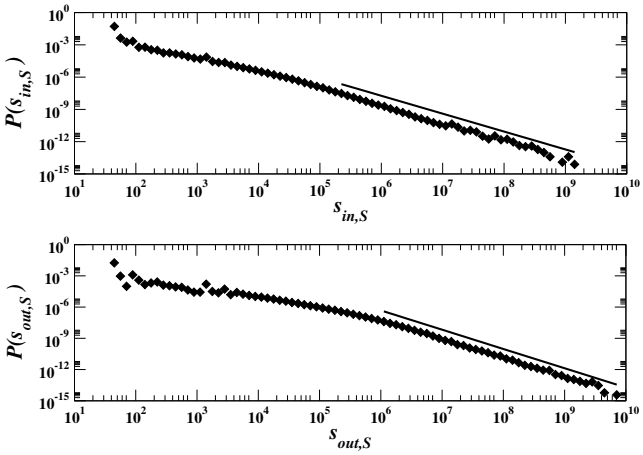

Figure 9: Probability distribution of the total incoming data (in-strength) and total outgoing data (out-strength) of servers. The solid lines refer to power-law behavior with slope -1.7 and -1.8 for the upper and lower graph, respectively.

with exponent $\gamma \leq 2$ diverges, which means that the average value of the distribution is not statistically defined. This is more peculiar than the case of power-law distributions with $2 < \gamma \leq 3$, such as the servers-per-client data reported in Figure 3, where the average is defined but the variance is not. In the case of clients-per-server data, the average degree is bounded only by the finite size of the Web. In such a situation, even the global mean number of connections is no longer a physical intrinsic quantity. This extreme heterogeneity is not usually found in technological networks [17]; it provides no indication of a global average quantity and, most importantly, no hint as to the scale to which it is most appropriate to target the design of a general-purpose Web server. As in the case of the client analysis, we also studied the probability distributions $P(k_{out,c})$ and $P(k_{in,c})$ that describe the chance that any given server has out-degree $k_{out,c}$ and in-degree $k_{in,c}$, respectively. The analysis recovers a power-law behavior with exponents $\gamma_{in} \simeq \gamma_{out} \simeq 1.8$, consistent with the behavior of $P(n_C)$.

The server in-strength consists of the amount of data that each server has received from its clients; this is the sum total of requests, form postings, and so forth. As was the case with the corresponding client graphs, the sampling done at the routers causes these values to be rescaled by two orders of magnitude with respect to the actual values, without affecting the shape of the distribution itself. We find a mean value of 8.42×10^4 and a standard deviation of 5.41×10^6. Analogously, the out-strength of servers indicates the amount of data sent to Web clients by each server; this is probably the quantity of greatest importance when designing Web server software. Bearing in mind the underestimate caused by sampling, we find a mean of 1.35×10^6 and standard deviation of 3.91×10^7. As shown in Figure 9, both distributions yet again match power-law behavior $P(s_{out,S}) \sim s_{out,S}^{-\alpha_{out}}$ and $P(s_{in,S}) \sim s_{in,S}^{-\alpha_{in}}$, this time with $\alpha_{in} = 1.71 \pm 0.1$ and $\alpha_{out} = 1.8 \pm 0.1$. In this case we are in presence of another harsh reality: there is no typical amount of incoming data that a Web server can expect to handle over the course of a day. This fact in itself is not surprising, given the broad range of popularity among Web sites. What is surprising is that both the first and second moments of the mean diverge. Analogously, there is no typical amount of outgoing traffic for a Web server. From the standpoint of an arbitrary Web server, Web traffic has no characteristic scale at all!

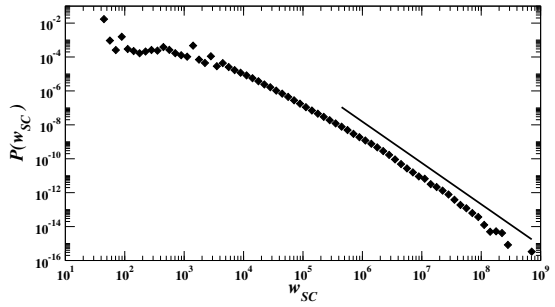

Figure 10: Distribution of aggregate server-to-client traffic. The solid line has slope -2.2.

Finally, we consider the distribution of weights w_{ji} representing the amount of data sent from a particular server to a particular client over the course of the day. The probability distribution showing the chance that the traffic on a given server-to-client connection has a value w_{SC} is no exception, and in Figure 10 we show the power-law behavior $P(w_{SC}) \sim w_{SC}^{-\delta}$, with $\delta = 2.2 \pm 0.1$.

4.2.2 Behavioral patterns

Following the same lines as in the client analysis, we plot the mean in-strength of servers of a given in-degree as a function of the in-degree. This allows us to explore the relationship between the number of clients handled by a Web server and the total amount of data received from those clients. Figure 11 (top) shows the obtained behavior, yielding the relation

$$\langle s_{in,S}(k_{in,S}) \rangle \sim k_{in,S}^{-\beta_{in}}$$

with $\beta_{in} = 0.9 \pm 0.1$.

Figure 12 (top) shows that the relation between out-strength and out-degree, i.e., traffic sent to clients as a function of the number of clients, exhibits the same behavior with exponent $\beta_{out} = 0.9 \pm 0.1$. In this case, servers appear to behave differently than do clients. The super-linear increase in traffic with increasing degree is not found; on the contrary, the data are compatible with a linear or slightly sub-linear behavior. This suggests that Web servers have a more conventional coupling between data traffic and number of connections, signalling a more predictable behavior in this respect. As in the case of clients, the server scaling exponents β for incoming and outgoing connections are predictable from the exponents of the strength and degree distributions α and γ, in agreement with Equations 1.

Finally, we study the spread of strength values with respect to the number of connections. In Figures 11 and 12 (bottom) we map the frequency of servers with given in-strength and out-strength as a function of the number of client-to-server (in-degree) and server-to-client (out-degree) connections, respectively. As with clients, we observe a large variability around the expected average behavior at all values of connectivity. The presence of outliers is less notable than in the client case, in agreement with the passive role of

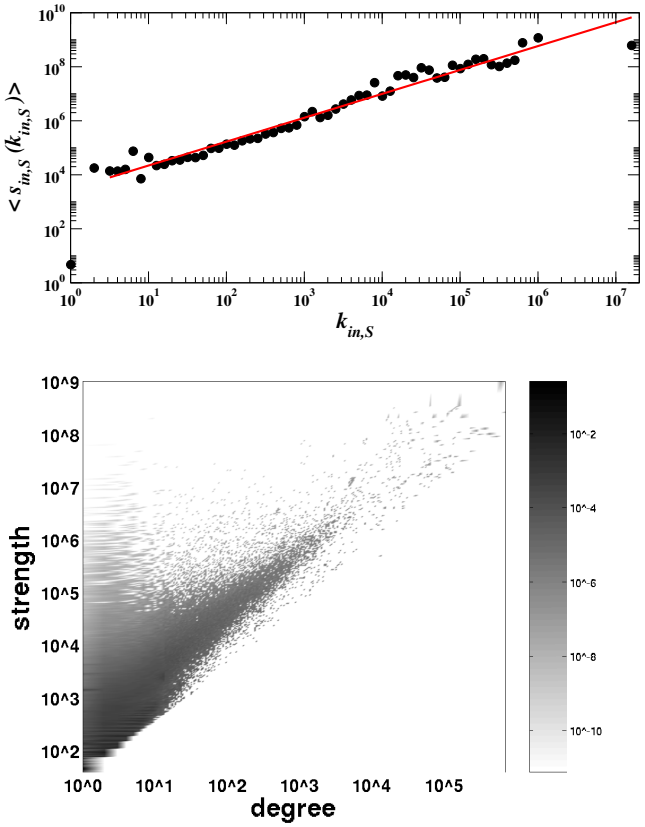

Figure 11: Behavior of the incoming traffic (strength) s_{in} as a function of the number of client-to-server connections (degree) k_{in}. Top: behavior of the average in-strength $\langle s_{in}(k_{in})\rangle$ as a function of the in-degree. The behavior is linear on a double logarithmic scale and is well approximated by power-law behavior with slope $\beta_{in} \simeq 0.9$. Bottom: frequency of servers with a given in-strength for each value of the in-degree.

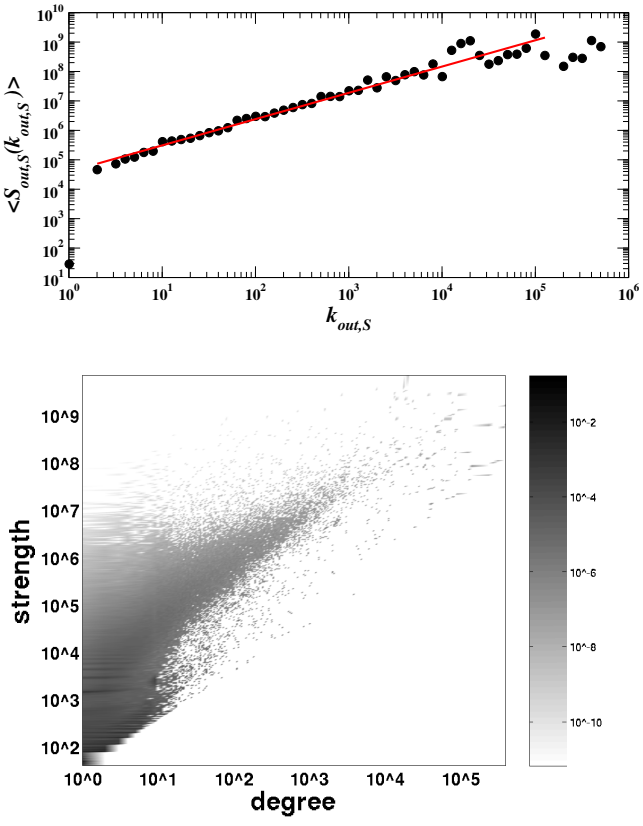

Figure 12: Behavior of the outgoing traffic (strength) s_{out} as a function of the number of server-to-client connections (degree) k_{out}. Top: behavior of the average out-strength $\langle s_{out}(k_{out})\rangle$ as a function of the out-degree. The behavior is linear on a double logarithmic scale and is well approximated by power law behavior with slope $\beta_{out} \simeq 0.9$. Bottom: frequency of servers with a given out-strength for each value of the out-degree.

servers. The presence of anomalies or outliers corresponding to servers with very low in-strength and very high in-degree could indicate the presence of distributed DoS attacks.

5. DISCUSSION AND FUTURE WORK

Several features of our data set merit further note. We speculate that the larger size of the client-to-server graph is a consequence of ongoing denial-of-service attacks and a possible sampling bias in the routers toward smaller flows. Though we are in contact with the router vendor to determine more exact details of flow sampling, we have no definitive answer as of yet. If this bias does exist, it should not affect the results of our analysis, as the effects we describe persist over many orders of magnitude.

Though all of the results presented here derive from a single day's worth of traffic data, analysis of other days yields nearly identical results. In the absence of any known catastrophic network events, we would expect this to be the case simply because of the extreme size of our data set. We know of no other study of Web traffic which has examined this quantity of flow information, although there has been large-scale analysis of Web traffic using Akamai server logs and client packet traces [10].

It is natural to wonder whether the properties of these distributions are changing over time. Others have reported Web traffic as using a smaller percentage of total Abilene traffic in 2001 [18], thus we suspect that traffic patterns are indeed changing. We plan on repeating our analysis with the intention of identifying long-term trends in global Web traffic behavior.

Our study provides general evidence that Web traffic is characterized by scale-free statistical distributions that signal a large heterogeneity of behaviors and the impossibility of relying on typical quantities or characteristic properties. In Table 1 we summarize the results obtained for the various distributions analyzed in the present work. All of them are heavy-tailed and exhibit standard deviations two to three orders of magnitude larger than the mean values indicated by the distributions. The presence of these overwhelming fluctuations is underscored in the case of Web servers by traffic distributions that follow power-law behavior with diverging first moments; here, no intrinsic average quantity can be inferred from the statistical distribution. This fea-

Table 1: Summary of statistical properties of the variable characterizing Web traffic data. For each variable x we report the average value $\langle x \rangle$ and standard deviation σ. It is possible to appreciate how in all cases the standard deviation is one to two orders of magnitude larger than the mean. The exponent characterizing the power-law behavior is evaluated in each case by a best fit procedure of the distribution tail.

variable x	$\langle x \rangle$	σ	exponent
n_S	3.52×10^0	3.51×10^1	2.4 ± 0.2
$s_{in,c}$	9.28×10^4	2.05×10^6	2.1 ± 0.1
$s_{out,c}$	1.11×10^3	1.44×10^5	2.2 ± 0.1
n_C	1.42×10^2	2.34×10^4	1.8 ± 0.1
$s_{in,S}$	8.42×10^4	5.41×10^6	1.7 ± 0.1
$s_{out,S}$	1.35×10^6	3.91×10^7	1.8 ± 0.1

ture has major consequences for Web server and network design. The fact that there is no typical amount of traffic faced by a Web server means that no one server design can be appropriate for all sites and the scalability of servers over time will be uncertain.

Client traffic also has diverging fluctuations, but not a diverging first moment. While this suggests a more regular behavior, clients exhibit a super-linear growth of traffic handled as a function of their numbers of connections. This behavioral pattern of client users points to a difficulty in designing scalable client applications as well. We are developing techniques to classify traffic according to the type of client: browsers, crawlers, scanners, etc. The statistical behaviors of these applications are expected to be distinguishable at the level of flow data, and this may provide additional insight into the design of scalable clients.

At the theoretical level, models of Web traffic must account for the fact that the traffic distributions follow power laws so strongly that the familiar parameters of mean and standard deviation are useless in characterizing the underlying system. In addition, the non-linear coupling between traffic and connections calls for models able to produce non-linear behavioral patterns and non-trivial correlations. An open issue is then how to relate the scale-free properties of the link structure of the Web with the Web traffic itself, where neither Web clients nor servers operate according to conventional distributions.

The fact that Web traffic follows distributions even more scale-free than the link structure of the Web also has implications for search engine design. The PageRank algorithm at the heart of Google deals solely with link structure, but large-scale analysis of data in a transit network offers additional information on how often users actually click on those links. Using traffic data to weigh the edges in the link network—thus turning PageRank into "ClickRank"—may improve substantially the ordering of search results. We intend to pursue this line of investigation in future research.

6. ACKNOWLEDGMENTS

The authors would like to thank the "Networks and Agents Network" in the IU School of Informatics as well as Katy Börner for useful discussions, the Advanced Network Management Laboratory at IU for support and infrastructure, and Internet2 for its generous policies on the use of anonymized network flow data. We are also grateful to three anonymous reviewers for helpful comments.

7. REFERENCES

[1] L. A. Adamic and B. A. Huberman. The Web's hidden order. *Communications of the ACM*, 44(9):55–60, 2001.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[3] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the World-Wide Web. *Nature*, 401:130–131, 1999.

[4] L. Amaral, A. Scala, M. Barthélemy, and H. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA*, 97:11149, 2000.

[5] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA*, 101:3747, 2004.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, 33(1-6):309–320, 2000.

[8] G. Chartrand and L. Lesniak. *Graphs and Digraphs*. Chapman & Hall/CRC, 1996.

[9] S. Dorogovtsev and J. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, UK, 2003.

[10] A. Feldmann, N. Kammenhuber, O. Maennel, B. Maggs, R. D. Prisco, and R. Sundaram. A methodology for estimating interdomain web traffic demand. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 322–335. ACM Press, 2004.

[11] B. Huberman. *The Laws of the Web*. MIT Press, 2001.

[12] B. Huberman and R. Lukose. Social dilemmas and internet congestion. *Science*, 277:535, 1997.

[13] B. Huberman, P. Pirolli, J. Pitkow, and R. Lukose. Strong regularities in World Wide Web surfing. *Science*, 280(5360):95–97, 1998.

[14] S. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the Web graph. In *Proc. 41st Annual IEEE Symposium on Foundations of Computer Science*, pages 57–65, Silver Spring, MD, 2000. IEEE Computer Society Press.

[15] L. Laura, S. Leonardi, S. Millozzi, U. Meyer, and J. Sibeyn. Algorithms and experiments for the Webgraph. In *Proc. European Symposium on Algorithms*, 2003.

[16] M. Newman. Analysis of weighted networks. Technical report, http://arxiv.org/abs/cond-mat/0407503, 2004.

[17] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet*. Cambridge University Press, Cambridge, UK, 2004.

[18] S. Shalunov and B. Teitelbaum. Internet2 TCP use and performance. Technical report, Internet2 Technical Report, 2001.