Structural analysis of behavioral networks from the Internet

2008 J. Phys. A: Math. Theor. 41 224022

(http://iopscience.iop.org/1751-8121/41/22/224022)

J. Phys. A: Math. Theor. **41** (2008) 224022 (12pp) doi:10.1088/1751-8113/41/22/224022

Structural analysis of behavioral networks from the Internet

M R Meiss[1,2], **F Menczer**[1,3,4] **and A Vespignani**[3,4]

[1] Department of Computer Science, Indiana University, Bloomington, IN 47405, USA
[2] Advanced Network Management Laboratory, Indiana University, Bloomington, IN 47404, USA
[3] Department of Informatics, Indiana University, Bloomington, IN 47408, USA
[4] Complex Networks Lagrange Laboratory, ISI Foundation, 10133 Torino, Italy

E-mail: mmeiss@indiana.edu

Received 7 November 2007
Published 21 May 2008
Online at stacks.iop.org/JPhysA/41/224022

Abstract
In spite of the Internet's phenomenal growth and social impact, many aspects of the collective communication behavior of its users are largely unknown. Understanding the structure and dynamics of the *behavioral networks* that connect users with each other and with services across the Internet is key to modeling the network and designing future applications. We present a characterization of the properties of the behavioral networks generated by several million users of the Abilene (Internet2) network. Structural features of these networks offer new insights into scaling properties of network activity and ways of distinguishing particular patterns of traffic. For example, we find that the structure of the behavioral network associated with Web activity is characterized by such extreme heterogeneity as to challenge any simple attempt to model Web server traffic.

PACS numbers: 89.20.Hh, 89.75.−k

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Although the Internet was first built as an infrastructure to support other research efforts, its overwhelming success has created a complex system that has become a scientific challenge in its own right. Unsupervised and exponential growth has led to a system self-organized into a structure still only somewhat understood. Besides the physical interconnections that define Internet topology, the Internet also contains myriad *virtual* networks formed by applications as varied as the World Wide Web, e-mail and peer-to-peer (P2P) networks. Hundreds of millions of users around the world create and populate these networks. Their interactions represent a self-organizing growth process that brings about enormous, intricate and interdependent systems. The dynamics of these virtual networks are the outcome of personal interactions among users and a mixture of complex social and technical factors that define the so-called *ecology of information* [1].

The Internet thus confronts us with the scientific challenge of being the designers and explorers of the system at the same time. Much investigation has already gone into determining the physical structure of the Internet at several levels of granularity, with the goal of developing an abstract representation of Internet topology in which nodes and edges represent either routers and their physical connections, or autonomous systems (ASes) and their peering relations. Although these studies have revealed much about the physical arrangement of the Internet, they have told us relatively little about the virtual networks created by the users who now spend a significant portion of their daily lives online, carrying out a wide variety of activities in different media. Human behavior more than physical connectivity determines the structure of these networks. While network engineers try to predict future demand, they are not the ones dedicating thousands of computers to sharing the latest movies. The Internet is thus populated with a plethora of applications and communication channels lacking any formal classification, their use often unknown outside the boundaries of specific communities of users (see the discussion for the CAIDA proposed community-wide experiment at www.caida.org/projects/ditl).

In this paper we refer to these user-to-user networks, whose topology is formed by mutual use of network applications rather than the physical structure of the network, as *behavioral networks*. Understanding their properties is an essential basis for further work in modeling the structure and dynamics of Internet traffic and contributes to our overall understanding of complex and emergent systems. To this end, in section 4 we build on previous work related to the behavioral network specific to the Web [2] and present an overview of the graph properties of multiple classes of traffic. The graph structures we analyze are derived from data collected from the Abilene (Internet2) network, as described in section 3. Finally, we summarize our findings and discuss their implications for modeling and predicting network activity.

2. Background

Because the Internet lacks any privileged viewing position, researchers must rely on a variety of heuristic techniques to assemble a global perspective of the physical structure of the Internet from a collection of local views. These local views may come from passive measurement techniques, such as the analysis of Internet routing tables, or from active probes made by tools such as *traceroute*. These mapping projects [3–5] have yielded views of the Internet comprehensive enough to allow the discovery of a number of surprising properties of Internet topology, including the presence of strong heterogeneities in the degree distribution of the corresponding graphs [6–9]. Despite several controversies and a lively debate on whether heavy tail properties of the Internet graphs may be artifacts of the sampling biases associated with topology measurements [10–12], recent results suggest that while deviations from the measurements must be expected, extreme heterogeneity in degree distributions is a genuine feature of the Internet [13, 14], and the study of Internet topology has generated significant activity in the field of network modeling [15–19].

While these efforts and others have contributed greatly to our understanding of the topology of the Internet at the router and AS levels, they do not address the user-to-user interactions that make up the behavioral networks we explore in this paper. There have been efforts to characterize application behavior in various domains by analyzing access logs from servers and proxies that process client requests [20–23]. Unfortunately, these efforts are unable to give us much insight into *global* patterns of user behavior and their relation to physical connectivity.

More global sources of behavioral data do exist; Internet routers themselves provide information on user-to-user communications using the abstraction of a *network flow*, which is uniquely defined by the IP address, protocol and port used by both nodes involved in a network transaction during a particular period of time. The most common form of flow data, Cisco's *NetFlow*, includes data on the volume of information exchanged but not the actual contents of any network conversation, providing a rich source of data for understanding the dynamics of user behavior on the Internet without abandoning individual privacy. Even this high-level view of activity presents technical challenges: because of the volume of data transmitted on modern networks, which can exceed half a petabyte of information daily, routers must derive flow information from a sample of actual packets, often at a rate of 1:100.

This flow-centered view of network activity has yielded substantial benefits already. For instance, inter-domain traffic has been studied on a global level by looking at data representing all traffic received by specific service providers [24]. A similar strategy has been used by the CAIDA measurement infrastructure, which allows for the construction of traffic matrices representing the traffic between pairs of ASs [25, 26]. More recently, aggregated flows have been used to detect anomalies and for time modeling of traffic [27]. Finally, increasing attention has been devoted to the evaluation of Internet performance [28] at the global level, where various projects, such as the *PingER* monitoring infrastructure [29] and RIPE [30], collect performance data among a large number of source–destination pairs. Various tools have been developed to analyze NetFlow data [31–35]. These tools, and most of the research they support, view network flow as a flat collection of records; they examine properties such as the proportion of traffic generated by particular applications or the longevity of certain classes of connection. This approach does not aid in exploring properties that relate to the graph structures of behavioral networks.

There has already been some application of complex systems analysis to application networks, mostly notably that of the Web, and these projects help to inspire the present work. The majority of Web mining studies focus on the social network built from the *link graph,* in which vertices and directed edges identify Web pages and hyperlinks, and links are seen as endorsements among pages. Data gathered in large-scale crawls [36–40] of the Web have uncovered the presence of a complex architecture with small-world properties and long-tailed distributions that characterize the structure of the graph. Examples of this complexity have included navigation patterns, community structures, congestion and other social phenomena resulting from users' behavior [39, 41–44]. Besides the Web, other overlay networks have been examined in a similar fashion, most notably email interaction and peer-to-peer networks [45–49]. Other researchers in the field have applied graph analysis to security topics, focusing on monitoring and characterization of the spread of computer viruses on the Internet [46, 50–53] and other malicious activities [54–57].

3. Constructing behavioral graphs from flow data

We now provide a technical description of the source of network flow data used in the present research. These anonymized flow data are typical of that available to a broad audience of interested researchers and do not provide access to host identities or captured packets.

The Abilene network, which is part of the Internet2 project [58], is an excellent source of flow data for behavioral studies. This high-performance TCP/IP data network spans the United States and provides high-speed connectivity to hundreds of research laboratories, colleges and universities. The backbone of the network consists of 10 Gbps fiberoptic links

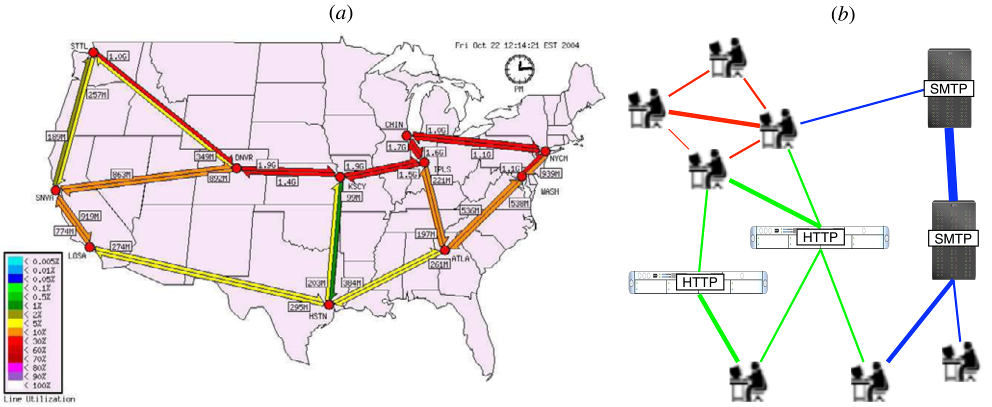

Figure 1. (*a*) Typical activity levels between core routers in the Abilene network (sustained data rates in bits per second). Source: loadrunner.uits.iu.edu/weathermaps/abilene. (*b*) Illustration of a behavioral network snapshot extracted from NetFlow data. Edge thickness represents amount of traffic. Flow records also specify traffic direction (not shown).

connecting 11 high-performance routers located in major metropolitan areas. At the time our sample was collected, the network nominally carried only academic and research traffic, with participating institutions maintaining separate connections to the commodity Internet, but it has recently expanded to include peering points with commercial ISPs. Among Abilene's users are hundreds of thousands of undergraduate students who are early adopters of new network applications. In addition, Abilene provides transit for data from dozens of international academic and research networks, serving as a major data path between Pacific Rim Nations and Europe, and giving an international character to its traffic. Finally, Abilene is never congested even during peak hours, offering a view of what users do when the network itself does not impede their behavior. Typical levels of IPv4 traffic in the Abilene network can be seen in figure 1(*a*).

Current technology does not allow the collection of flow data for every network conversation on Abilene. Each of the core routers sample their traffic load at a rate of about 1:100, using them to generate flow information in Cisco's 'netflow-v5' format[5], which is then sent to an analysis system at our university. According to Internet2 privacy policy, this system removes the source and destination IP addresses of each flow, replacing them with index values that maintain their identity only over the course of a single day. This anonymized flow information is saved for analysis. On a typical day, the analysis system records around 700 million flow records; a full day of data consumes over 30 GB of disk space and arrives at a mean rate of 3.1 Mbps.

We construct different varieties of graph structures depending on the way in which we aggregate the flow data. Each record describes the transmission of some quantity of data from some host and port to some other host and port, without identifying explicitly which host acts as a client and which acts as a server. We are thus able to use either hosts or ports as nodes in the graph, and we can accumulate weights over repeated edges throughout some time interval. Figure 2 illustrates different types of graphs we derive from flow data, which we refer to as *behavioral, functional* and *application* networks.

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

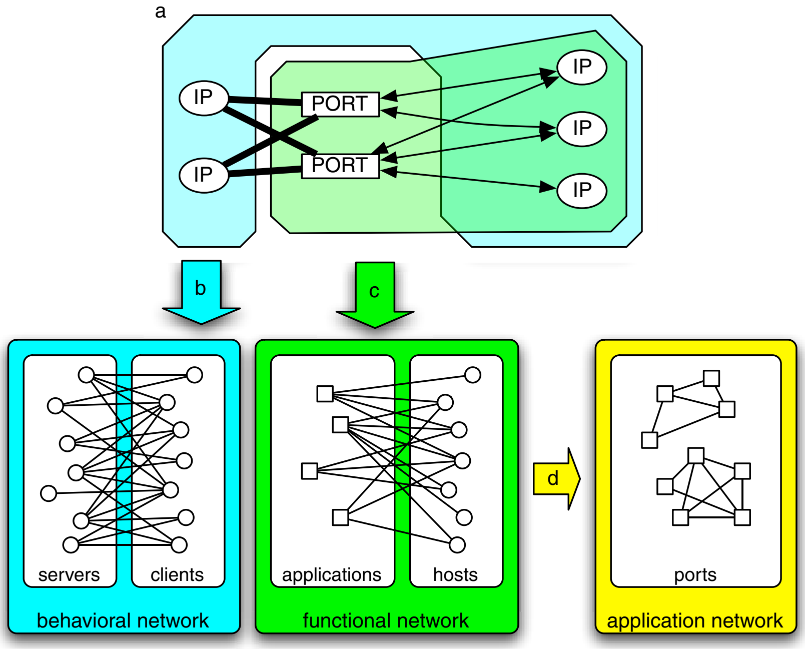

Figure 2. The construction of behavioral, functional and application networks from flow data. (*a*) Each raw flow record describes how many bytes are exchanged between two hosts using the indicated ports on the sending and receiving hosts. Flows are aggregated over a day so that we have the total amount of data exchanged by two hosts for each pair of ports. (*b*) By aggregating flow data for each IP address across application ports, or focusing on individual standard ports, we can build networks of hosts (clients and servers) that describe how users are connected with each other and with services across the Internet. (*c*) By disregarding servers, and retaining ports as entities, we can build a functional network describing the relationships between user hosts and applications. (*d*) By focusing on strength correlations among ports, we can cluster and identify applications.

The first step in deriving the behavioral network associated with an application or group of applications (figure 1(*b*)) involves recovering the roles of clients and servers, since their behavior is quite different for most applications. We do this by examining the total number of flows that reference a particular port because clients use ephemeral port numbers and servers' port numbers must be known in advance, in any particular record, the server will almost always be the system with the more frequently used port number. We can thus partition the set of all hosts into a subset $C = \{i_1, i_2, \ldots, i_{N_C}\}$ of systems that act as clients and a subset $S = \{j_1, j_2, \ldots, j_{N_S}\}$ of system that act as servers. Some computers on the Internet, especially those involved in P2P networks, act as both clients and servers and are assigned to both sets. Using the sets C and S, we construct a *behavioral* graph in which the nodes represent individual hosts and edges represent the directed transmission of data between a pair of hosts, aggregated over the course of a day. Each weight w_{ij} represents the total amount of sampled data sent from client i to server j over the course of a day, and w_{ji} represents the amount of data sent from server j back to client i. We thus have a bipartite digraph between clients and servers, weighted by aggregate volumes of traffic, as shown in figure 2(*b*). This representation of the behavioral networks of Internet2 hosts is the basis for the analysis in the following section.

There are other ways to project the flow data to obtain different network views of Internet traffic. By retaining port information one can build a *functional network* among port numbers and client IP addresses, capturing the variety of activities in which each particular user engages (figure 2(*c*)). Each weight in the network represents the extent to which a host on the network has made use of a particular TCP port. Since each port generally corresponds to a specific application, this functional network can be used to characterize applications by their profiles, i.e., by the amounts of traffic exchanged by each participating host. One can then study the associations among applications by comparing these host profiles, using the basic intuition that correlated use of two applications by users is evidence that they have a similar purpose, similar to the way in which co-citation can serve as a measure of relatedness for academic papers. This process also allows us to construct *application networks* (figure 2(*d*)) having ports as nodes and weighted edges representing the usage correlations (similarity) among ports. The use of application networks to predict the function of *unknown* ports based on their observed profiles is discussed elsewhere [59]; we focus here on the analysis of behavioral networks.

4. Behavioral network analysis

We now present the results of our analysis of behavioral networks based on 24 h of Abilene flow data gathered starting at midnight EST on 14 April 2005. This was a typical day, with no known major outages or disruptions of service. Our findings are consistent with those of earlier studies [2]. In the course of this day, the flow collector received over 600 million flows involving almost 15 million hosts. Of these flows, 258 million (41.3%) were Web-related and 82 million (13.1%) were associated with known P2P applications. The remaining 285 million (45.6%) flows describe all other traffic, which includes network performance tests, pings, e-mail, interactive logins and a wide variety of miscellaneous and unidentified applications (see figure 3(*a*)).

Of the total number of hosts served, 5.82 million were observed behaving as clients and nearly twice as many, 11.1 million, behaving as servers. Such a high proportion of servers to clients is a symptom of scanning traffic on the network: rogue clients routinely search for vulnerable servers. We find that the opposite is the case for Web and P2P applications. When we examine Web flows in isolation, we find 3.97 million hosts behaving as clients and 0.68 million (less than one-fifth as many) behaving as servers. Similarly, for P2P traffic, there were 0.71 million clients and only 0.14 million servers. The remaining traffic shows 2.48 million clients and 10.6 million servers. The behavioral graph that includes all hosts and applications contains 131 million edges. If we examine subgraphs related only to particular classes of application, we find that the Web graph contains 50.1 million edges (38.0% as many as the full graph), the P2P graph contains 7.89 million edges (6.0%) and remaining TCP traffic contains 54.9 million edges (41.6%) (see figure 3(*b*)).

For each category of traffic (Web, P2P and the remainder), we also examine the degree of overlap between C and S, which we represent with the quantity:

$$O = (|C| \cap |S|)/(|C| \cup |S|). \tag{1}$$

When $O = 0$, no host acts as both client and server; when $O = 1$, every host does so. We would expect O to be lower for traditional client–server applications than modern P2P applications, and indeed, we find that $O = 0.14$ overall and 0.097 for P2P traffic, but only 0.013 for Web traffic. This is a strong indication that hosting content for the Web is much less of a participant sport than sharing personal files, in that relatively few users run both Web clients and servers.

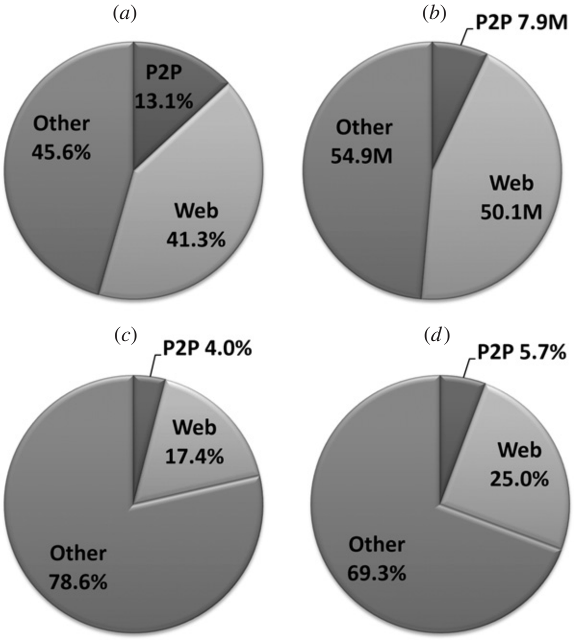

Figure 3. (*a*) Proportion of collected flows generated by each category of traffic. (*b*) Relative sizes of bipartite graphs for each category of traffic. (*c*) Proportion of traffic volume consumed by each category of traffic, with *iperf* test traffic included and (*d*) with *iperf* traffic omitted.

The total volume of traffic recorded is approximately 1.85 trillion bytes, with a mean of 124 kB per host. However, because of the sampling involved in constructing the flow data, the true amount of traffic will be about 100 times greater than this value. 17.4% of total traffic is Web-related, with a mean of 81 kB of data for clients and 471 kB for servers. 4.0% of traffic is P2P, with a mean of 105 kB of data for clients and 515 kB for servers. The remaining traffic comprises 78.6% of the total, with a mean of 586 kB per client and 137 kB per server. This is influenced by a large volume of *iperf* test traffic generated by the Abilene network operations center to monitor performance; when these data are removed, the mean client data drop to 222 kB, and the proportion of remaining traffic drops to 69.3% of the total (see figures 3(*c*), (*d*)).

The statistics just described provide little insight into the actual behavior of the community or the role a typical user plays in the network. We thus turn our attention to the *structure* of the behavioral networks corresponding to these three categories: the Web, P2P applications and everything else.

We begin by considering the distributions of degree and strength for the nodes in the behavioral network. Given a node N with i initial edges and j terminal edges, we define the degree as $d_N = i + j$ and the strength as

$$s_N = \sum_{k=1}^{i} w_{N,N_k} + \sum_{k=1}^{j} w_{N_k,N}, \tag{2}$$

where $w_{a,b}$ denotes the weight of the edge between nodes a and b. In other words, the degree of a node in the behavioral network reflects the total number of users with which it has exchanged

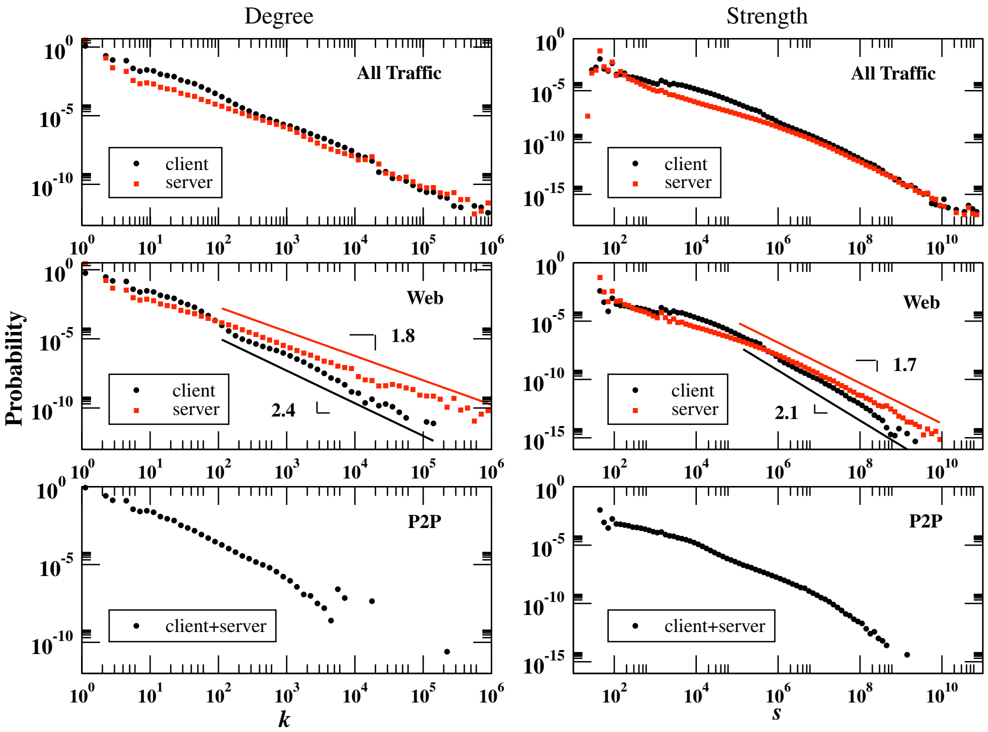

Figure 4. Probability distributions for degree (left) and strength (right) in the Internet2 behavioral network, shown for all data (top), the Web (middle) and peer-to-peer applications (bottom). The data are grouped into logarithmically-sized histogram bins normalized by the width of the bin and size of the distribution, so that we are estimating a probability density function. The annotated lines in the Web plots show statistically significant best-fit power-law approximations to the actual data, with $R^2 \geqslant 0.995$.

data, and the strength reflects the total amount of data it has exchanged. Because they share the same labels, the behavioral networks of different applications are directly comparable.

Because both the degree and strength distributions reflect the individual decisions made by a large population, it might seem plausible for their form to be roughly normal. This turns out to be far from the case, however, as shown in figure 4. All of the degree and strength distributions shown have extremely long tails, some spanning almost ten orders of magnitude. As an example, the mean strength of a client is approximately 318 kB, but the standard deviation of the distribution is 72.6 MB; the level of statistical fluctuation is over two orders of magnitude larger than the mean value. Indeed, the distributions are so skewed that in the case of the behavioral networks for the Web and for all traffic, we are able to approximate both the degree and strength distributions with a power-law function $P(n) \sim n^{-\gamma}$ over several orders of magnitude.

Let us examine the slopes of these power-law approximations. When $2 < \gamma < 3$, the second moment $\langle n^2 \rangle = \int n^2 P(n)\, \mathrm{d}n$ diverges; the average value $\langle n \rangle$ is no longer typical, and we lack any characteristic mean for the system, giving us behavior often described as 'scale-free.' We have an appreciable probability of finding a client that has contacted any arbitrary number of servers or downloaded any arbitrary amount of data. The averages thus seem to be of no value in predicting user behavior.

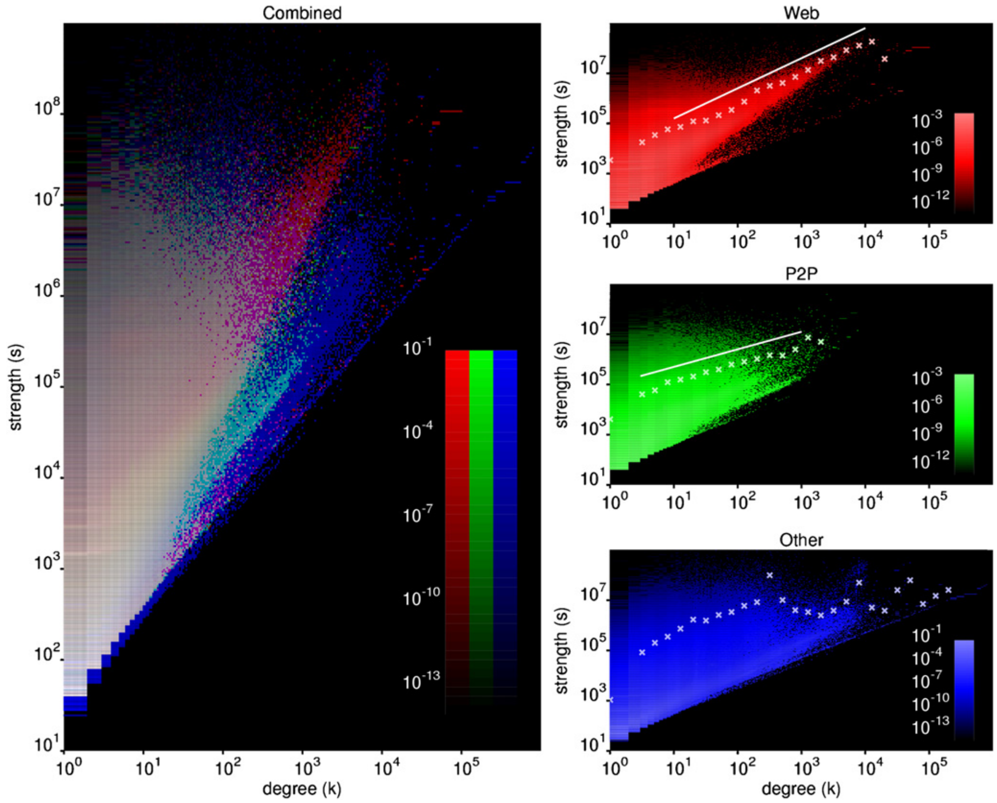

Figure 5. Behavior of strength (total data) s as a function of degree (number of hosts contacted) k. This distribution map shows the relationship of degree versus strength for Web, P2P and other traffic, both combined (left) and as separate plots (right). The tones represent the frequencies of strength values, normalized within each degree bin, on a logarithmic scale. The points in the maps to the right show the mean strength for each degree bin, and the lines in the Web and P2P plots show best-fit power-law approximations to the actual data. The behavior for these two traffic categories is roughly linear on a double-logarithmic scale, with $R^2 \geqslant 0.999$ over the full data set. A color version of this plot, available online, splits the application classes into the red, green, and blue color planes for more direct comparison.

When $\gamma < 2$, as is the case for both the degree and strength distributions of Web servers, we have an even more dramatic situation where the first moment $\langle n \rangle = \int n P(n)\,\mathrm{d}n$ diverges and is bounded only by the size of the sample. Neither the mean number of connections nor the mean amount of data transmitted are intrinsic to the system. This extreme heterogeneity presents a challenge for any attempt to model Web traffic: there is no typical Web server in terms of expected traffic load.

In the case of P2P networks, we do observe heavy-tailed distributions, but there appears to be a definite exponential cutoff, after which the probability function decays more quickly than a power-law fit would predict. We conjecture this may be due to the limited computing and network capacity of the computers participating in peer-to-peer networks, which are usually the personal computers of individual users. If such is the case, we can expect the tail to lengthen over time as the processing power and bandwidth available to users continue to increase.

The scaling relationship between degree and strength, which describes the relation between the number of hosts contacted to the amount of data exchanged, is also of considerable interest in understanding user behavior. Because of the power-law nature of distributions of degree and strength considered separately, it is unsurprising for strength to increase as a function of degree, again following a power law, as shown in figure 5.

Because basic power-law behavior $\langle s(k)\rangle \sim k^{\beta}$ can be expected of this interaction, it is the value of the exponent β that is of critical interest. In the case of server behavior we find a linear or sublinear relationship ($\beta \leqslant 1$), but in the case of Web clients we see evidence of a superlinear relationship $\beta = 1.2 \pm 0.1$. This implies that the amount of data exchanged with each Web server tends to increase as a user contacts more servers: the more sites surfed, the more data are received from each of these sites. Such a nonlinear growth mechanism must be accounted for in generative models of Web traffic and may be the basis of techniques to disambiguate the behavior of individual Web surfers and large-scale crawlers.

The relationship between the in- and out-distributions of strength may also facilitate the discovery of unrestricted proxy servers that are the launch points for a wide variety of security attacks on the Internet. Most of the traffic associated with a proxy for an application will be repeated: requests from a client to the proxy are retransmitted from the proxy to a server, and the server response is likewise retransmitted by the proxy back to the client. We would thus expect a proxy to exhibit an unusually high level of symmetry in its role in a behavioral network; not only would it function as both client and server, but also its in-strength and out-strength would be nearly equal.

5. Conclusions

We believe the results presented here constitute one of the first efforts to understand the behavior of individual Internet users through analysis of the networks implicitly formed by their actions. The properties uncovered by this view of individual behavior will be essential for agent-based modeling of user populations, which has implications for Internet epidemiology, network design and capacity planning. In particular, the pervasive presence of distributions with extremely long and heavy tails implies that user behavior rarely follows normal distributions, but is so diverse as to defy characterization with a mean value. Superlinear behavior in Web clients demands that any behavioral models be able to account for non-trivial coupling between degree and strength. Furthermore, the differences observed between the Web and P2P application groups imply that behavioral analysis can yield statistical signatures for different types of applications, allowing network managers to identify applications being run covertly on non-standard ports. Current network security products commonly employ rate-based thresholds to detect traffic anomalies [34]. However, our results show that ‘normal’ traffic would cause many false alarms irrespective of the threshold used. The analysis of behavioral networks may offer more effective methods of detecting malicious or otherwise anomalous behavior on the Internet.

The analytical framework we have described offers a practical way of understanding the actions of individual Internet users through the behavioral and application networks they generate. Because we avoid any reliance on captured packets or non-anonymized flow data, a much wider audience of researchers can test these techniques for themselves than has been the case with previous studies. None of the processing steps we describe require extensive computing resources; a single high-end workstation can perform all of the analysis described in section 4 in less than half an hour. The analysis of application networks is quadratic in the number of applications considered, but even it can be performed in a fraction of the data collection time window.

Acknowledgments

The authors would like to thank the Advanced Network Management Laboratory at Indiana University for support and infrastructure, as well as Internet2 for its generous policies on the use of anonymized network flow data. This work is funded in part by NSF awards 0348940 and 0513650 to FM and AV respectively, and by the Indiana University School of Informatics.

References

[1] Pirolli P and Card S 1997 The evolutionary ecology of information foraging *Technical Report* UIR-R97-01, Xerox PARC

[2] Meiss M, Menczer F and Vespignani A 2005 On the lack of typical behavior in the global Web traffic network *Proc. 14th Int. World Wide Web Conf.* pp 510–8

[3] The National Laboratory for Applied Network Research (NLANR) http://moat.nlanr.net

[4] The Cooperative Association for Internet Data Analysis (CAIDA) http://www.caida.org/home

[5] Dimes Project http://www.netdimes.org

[6] Faloutsos M, Faloutsos P and Faloutsos C 1999 On power-law relationships of the internet topology *SIGCOMM* pp 251–62

[7] Broido A and Claffy K C 2001 Internet topology: connectivity of IP graphs *San Diego Proc. SPIE Int. Symp. on Convergence of IT and Communication*

[8] Chen Q, Chang H, Govindan R, Jamin S, Shenker S J and Willinger W 2002 The origin of power laws in Internet topologies revisited *Proc. IEEE Infocom*

[9] Pastor-Satorras R, Vázquez A and Vespignani A 2001 Dynamical and correlation properties of the Internet *Phys. Rev. Lett.* **87** 258701

[10] Chang H, Govindan R, Jamin S, Shenker S J and Willinger W 2004 Towards capturing representative AS-level Internet topologies *Comput. Netw. J.* **44** 737–55

[11] Lakhina A, Byers J, Crovella M and Xie P 2003 Sampling biases in IP topology measurements *INFOCOM*

[12] Achlioptas D, Clauset A, Kempe D and Moore C 2005 On the bias of traceroute sampling *STOC*

[13] Dall'Asta L, Alvarez-Hamelin I, Barrat A, Vázquez A and Vespignani A 2005 Exploring networks with traceroute-like probes: theory and simulations *Theoretical Computer Science, Special Issue on Complex Networks*

[14] Siganos G, Faloutsos M, Faloutsos P and Faloutsos C 2003 Power-laws and the AS-level Internet topology *Trans. Netw.* **11** 514–24

[15] Medina A and Matta I 2000 BRITE: A flexible generator of Internet topologies *Technical Report* BU-CS-TR-2000-005, Boston University

[16] Jin C, Chen Q and Jamin S 2000 INET: internet topology generators *Technical Report* CSE-TR-433-00, EECS Department, University of Michigan

[17] Yook S-H, Jeong H and Barabási A-L 2002 Modeling the Internet's large-scale topology *PNAS* **99** 13382–6

[18] Pastor-Satorras R and Vespignani A 2004 *Evolution and Structure of the Internet* (Cambridge: Cambridge University Press)

[19] Fabrikant A, Koutsoupias E and Papadimitriou C H 2002 Heuristically optimized trade-offs: a new paradigm for power laws in the Internet *ICALP*

[20] Cherkasova L and Gupta M 2004 Analysis of enterprise media server workloads: access patterns, locality, content evolution and rates of change *IEEE/ACM J. Trans. Netw. (ToN)* **12** 781–94

[21] Acharya S, Smith B and Parnes P 1998 Characterizing user access to videos on the world wide web *SPIE/ACM Conference on Multimedia Computing and Networking (MMCN)*

[22] Arlitt M and Williamson C 1996 Web server workload characterization: the search for invariants *ACM SIGMETRICS*

[23] Harel N, Vellanki V, Chervenak A and Abowd G 1999 Workload of a media-enhanced classroom server *IEEE Workshop on Workload Characterization*

[24] Uhlig S and Bonaventure O 2001 The macroscopic behavior of Internet traffic: a comparative study *Technical Report* Infonet-TR-2001-10, University of Namur

[25] Claffy KC 1999 Internet measurement and data analysis: topology, workload, performance and routing statistics *NAE '99 Workshop* CAIDA

[26] Huffaker B, Fomenkov M, Moore D, Nemeth E and Claffy K C 2000 Measurements of the Internet topology in the Asia-Pacific region *INET '00 (Yokohama, Japan, 18–21 July 2000)* (Washington, DC: The Internet Society)

[27] Lakhina A, Papagiannaki K, Crovella M, Diot C, Kolaczyk E D and Taft N 2004 Structural analysis of network traffic flows *ACM SIGMETRICS* pp 61–72

[28] Carbone L, Coccetti F, Dini P, Percacci R and Vespignani A 2003 The spectrum of Internet performance *Proc. Passive and Active Measurement (PAM2003)*

[29] SLAC Internet end-to-end performance monitoring http://www-iepm.slac.stanford.edu

[30] Réseaux IP Européens, Ripe network coordination centre http://www.ripe.net

[31] Fullmer M Flow-tools information http://www.splintered.net/sw/flow-tools

[32] CAIDA.org. cflowd: traffic flow analysis tool http://www.caida.org/tools/measurement/cflowd

[33] CAIDA.org Flowscan—network traffic flow visualization and reporting tool http://www.caida.org/tools/utilities/flowscan

[34] Arbor Networks Peakflow http://www.arbor.net/products_platform.php

[35] Estan C, Savage S and Varghese G 2003 Automatically inferring patterns of resource consumption in network traffic *Proc. ACM SIGCOMM Conf.*

[36] Barabási A-L and Albert R 1999 Emergence of scaling in random networks *Science* **286** 509–12

[37] Broder A, Kumar S R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A and Wiener J 2000 Graph structure in the Web *Comput. Netw.* **33** 309–20

[38] Kumar S R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A and Upfal E 2000 Stochastic models for the Web graph *Proc. 41st Annual IEEE Symp. Foundations of Computer Science (Silver Spring, MD)* (Los Alamitos, CA: IEEE Computer Society Press) pp 57–65

[39] Adamic L A and Huberman B A 2001 The Web's hidden order *Commun. ACM* **44** 55–60

[40] Laura L, Leonardi S, Millozzi S, Meyer U and Sibeyn J F 2003 Algorithms and experiments for the Webgraph *European Symposium on Algorithms*

[41] Huberman B A and Lukose R 1997 Social dilemmas and Internet congestion *Science* **277** 535

[42] Huberman B A, Pirolli P L T, Pitkow J E and Lukose R M 1998 Strong regularities in World Wide Web surfing *Science* **280** 95–7

[43] Menczer F 2002 Growing and navigating the small world Web by local content *Proc. Natl Acad. Sci. USA* **99** 14014–9

[44] Menczer F 2004 The evolution of document networks *Proc. Natl Acad. Sci. USA* **101** 5261–5

[45] Ebel H, Mielsch L-I and Bornholdt S 2002 Scale-free topology of e-mail networks *Phys. Rev.* E **66** 035103

[46] Newman M E J, Forrest S and Balthrop J 2002 E-mail networks and the spread of computer viruses *Phys. Rev.* E **66** 035101

[47] Saroiu S, Krishna Gummadi P and Gribble S D 2002 A measurement study of peer-to-peer file sharing systems *Proc. Multimedia Computing and Networking 2002 (MMCN '02) (San Jose, CA)*

[48] Ripeanu M, Foster I and Iamnitchi A 2002 Mapping the gnutella network: properties of large-scale peer-to-peer systems and implications for system design *IEEE Internet Comput.* **6** 50–7

[49] Li C and Chen C 2007 Gnutella: topology dynamics on phase space *Preprint* cs/0702022

[50] Moore D, Shannon C and Brown J 2002 Code-Red: a case study on the spread and victims of an Internet worm *Proc. 2nd Internet Measurement Workshop*

[51] Staniford S, Paxson V and Weaver N 2002 How to own the Internet in your spare time? *Proc. 11th USENIX Security Symposium (Security '02)*

[52] Pastor-Satorras R and Vespignani A 2001 Epidemic spreading in scale-free networks *Phys. Rev. Lett.* **86** 3200–3

[53] Forrest S, Hofmeyr S and Somayaji A 1997 Computer immunology *Commun. ACM* **40** 88–96

[54] Moore D, Voelker G and Savage S 2001 Inferring internet denial of service activity *Proc. 2001 USENIX Security Symp.*

[55] Garetto M, Gong W and Towsley D 2003 Modeling malware spreading dynamics *Proc. INFOCOM 2003, 22nd Annual Joint Conf. IEEE Computeer and Communications Societies*

[56] Zou Cliff, Towsley Don and Gong Weibo 2004 Email worm modeling and defense *Proc. 13th International Conf. Computer Communications and Networks (ICCCN'04)*

[57] Singh S, Estan C, Varghese G and Savage S 2004 Automated worm fingerprinting *Proc. ACM/USENIX Symp. Operating System Design and Implementation*

[58] Internet2 project http://abilene.internet2.edu

[59] Meiss M, Menczer F and Vespignani A 2007 A framework for analysis of anonymized network flow data *Proc. NSF Symp. Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (NGDM'07)*