An end-host view on local traffic at home and work

Ahlem Reggani, Fabian Schneider, Renata Teixeira

An end-host view on local traffic at home and work

Ahlem Reggani[1], Fabian Schneider[2], and Renata Teixeira[1]

[1] UPMC Sorbonne Universités and CNRS, LIP6, Paris, France
[2] NEC Laboratories Europe, Heidelberg, Germany (work done at UPMC[1])

Abstract. This paper compares local and wide-area traffic from end-hosts connected to different home and work networks. We base our analysis on network and application traces collected from 47 end-hosts for at least one week. We compare traffic patterns in terms of number of connections, bytes, duration, and applications. Not surprisingly, wide-area traffic dominates local traffic for most users. Local connections are often shorter and smaller than Internet connections. Moreover, we find that name services (DNS) and network file systems are the most common local applications, whereas web surfing and P2P, which are the most popular applications in the wide-area, are not significant locally.

1 Introduction

The past couple of decades has seen many studies that characterize Internet traffic [1, 6, 7, 12]. These studies are based on packet traces collected in ISP networks, at border routers of university campuses or enterprise networks. As such, most prior studies focus on wide-area traffic. Little is known about the traffic that stays inside a network, which we call *local traffic*. The main exception is the study of traffic from one enterprise [8, 9], which shows that local traffic is different from wide-area traffic with a significant amount of name service, network file system, and backup traffic. As the authors point out their study is "an example of what modern enterprise traffic looks like" [9]. It is crucial to reappraise such analysis in other enterprises and more important in other types of edge networks. For instance, the spread of broadband Internet has caused an increase in the number of households that have a home network. Yet, there has only been limited analysis of local traffic volumes in three home networks [5], but no in depth characterization of in-home traffic patterns. The challenge of studying local traffic across multiple edge networks is to obtain measurements from *inside* multiple networks.

This paper characterizes local network traffic of multiple networks from the perspective of an end-host that connects inside an edge network. This approach is in contrast with previous work [5,9], which instruments routers in the local network. Although instrumenting routers could capture all traffic traversing the local network, it is hard to have access to routers at more than a few networks. By monitoring traffic directly at end-hosts, we can sample a larger number of networks, but we can only see the traffic from one of the hosts in the network. For smaller networks (such as home networks) a single host's traffic captures a significant fraction of those networks total traffic, whereas for larger networks (as enterprises) this fraction is less significant.

We rely on data collected at end-hosts using the HostView monitoring tool [4]. HostView records packet header traces and information about applications and user

environment. The data we study was collected from 47 users who ran HostView for more than a week each. Given that users move between different networks, this dataset contains end-host traffic from a total of 185 different networks spread over 18 different countries. Section 2 gives an overview of the HostView data. The analysis of local and wide-area traffic from HostView data is challenging, because HostView has no information of which traffic flows are local. Worse, HostView scrapes the end-host IP address from the traces to protect user's privacy, which makes the identification of local traffic more challenging. Therefore, we develop a heuristic to separate local from wide-area traffic. Section 3 describes this heuristic together with our method to categorize environments and applications in the HostView data.

Our analysis (presented in Section 4) asks some high-level questions, for instance: How does the volume of an end-host's local traffic compare to wide-area traffic? Do local and wide-area applications differ? How does traffic vary between home and work? The results show that for most users wide-area traffic dominates local traffic, but that some users have over 80 % of local traffic. Local connections are mostly shorter and smaller than wide-area connections, but sometimes they transfer a larger amount of traffic than large wide-area connections. We find that typical local applications are DNS, ssh, and network file systems (confirming previous findings [9]). Moreover, common applications at work include backup, printing, and web. Yet, these applications are rarely used at home.

2 Summary of HostView Data

In this paper, we use three of the datasets collected by the HostView tool [4]: network packet traces, application labels, and the end-host's network environment. HostView logs all this data directly at the end-host into a trace file, which is periodically uploaded to a server. A new trace is created every four hours or when a change in the network interface or the IP address is detected.

Network traces and application context HostView logs the first 100 bytes of each packet sent and received by the end-host with libpcap. For DNS packets, it records the whole packet to enable offline hostname to IP address mappings. In this paper, we use the connection summaries generated by previous work [3]. Each connection summary record describes both directions of a TCP or UDP connection and includes (among other fields): The source and destination IP addresses (replacing the host IP address with "0.0.0.0" to comply with French privacy laws), the source and destination port numbers, and the network protocol; The number of bytes, the number of packets, and the duration of the connection; And the name of the process executable that generated the connection.

Network environment HostView labels each trace file with information describing the network environment the end-host is connected to, including the network interface, a hash of the wireless network SSID and of the BSSID of the access point for wireless networks or a hash of the MAC address of the gateway for wired networks. It also records the ISP, the city, and the country for each trace using the MaxMind GeoIP commercial database from March 2011. When the end-host connects to a new wireless

network, HostView asks the user to specify the network type from a pre-defined list: Home, Work, Airport, Hotel, Conference meeting, Friend's home, Public place, Coffee shop or Other (with the possibility to specify). This user tag is used to classify the network the user connects to according to an environment type. Unfortunately, this tag is not available for wired connections and users sometimes skip the questionnaire. Originally, only 40 % of HostView traces had a user tag, but after applying some heuristics (which exploit the fact that users connect to the same network with both wireless and wired, for instance) previous work was able to label 78 % of the traces [3]. Still, the data includes at least one unlabeled trace per user. The next section describes our method to label most of the remaining traces with an environment type.

Dataset characteristics and biases HostView was announced in networking conferences and researcher mailing lists. Volunteer users downloaded HostView (which is available only for Mac OS and Linux) and ran it during different time intervals between November 2010 and August 2011. In this paper, we use traces from 47 users who ran HostView for at least one week; 32 of these users ran HostView for more than a month.

Because of the way HostView was advertised and its limited operating-system support, the user population is biased towards networking researchers. We acknowledge that networking researchers probably use different applications than the average user and may also work from home. It is still interesting to study examples of the differences between local and wide-area traffic. We do observe a diverse set of applications among different users and our users do use some popular applications like YouTube, Facebook and BitTorrent. Furthermore, this bias influences the types of networks we study. Importantly, "work" is often a university. Overall, we study end-hosts connected to 185 unique networks spread over 18 different countries (Italy: 25, France: 22, Germany: 21, Rest of Europe: 31, Asia: 19, US: 63, Australia: 3, and Brazil: 1); 34 distinct home networks and 38 distinct work environments (29 are universities and 9 enterprises).

Another bias comes from using data collected for a limited time period on only one single end-host in the network. It is well known that traffic characteristics can vary considerably between different networks and over time [10]. HostView can only see a small fraction of the network's traffic and there are some types of traffic that it can never observe. For example, some homes may have a media server that serves content to the TV; this type of traffic traverses the home network, but it is not originated or consumed by an end-host. Despite these shortcomings, we believe that this end-host perspective on local versus wide-area traffic offers the unique opportunity to sample traffic in a relative large number of networks. Whenever appropriate, we also contrast our findings with previous work.

3 Methodology

In this paper, we compare local and wide-area traffic in networks of different types. In addition, we are interested in the traffic application mix. We follow three steps to label HostView traces before our analysis: *(i)* Differentiation of local and wide-area traffic, *(ii)* Extension of the incomplete network type labeling, and *(iii)* Categorization of connection records into application groups.

Table 1. Examples of process names and network services to category mappings. This list is not complete and only intended to give an idea.

Category	Process name (Examples)	Application protocols
Backup	retroclient	amanda
Chat	Skype, iChat, Adium, Pidgin	ircd, SIP, msnp, snpp, xmpp
DistantControl	ssh, sshd, VNC, screen sharing	ssh(22), webmin
Email	Mail, Outlook, Thunderbird	IMAP(S), POP3(S), (S)SMTP
Personal	Media players, games, productivity	rtsp
FileTransfer	ftp, dropbox, svn, git, SW updates	ftp, rsync, svn, cvspserver
Management	traceroute, iperf, nmap, ntpd, uPNP	BOOTP, MySQL, VPN, SNMP, whois
Miscellaneous	perl, python, VirtualBox, openvpn	—
NameService	dns, nmblookup, named, nmbd, nscd	domain(53), mdns, netbios-ns
NetworkFS	smbclient, smbd, AppleFileServer	AFP, AFS, LDAP, netbios, nfs
P2P	amule, uTorrent, transmission	amule, Kazaa, BitTorrent
Printing	cupsd, lpd, HP, Lexmark	ipp, printer
Web	Firefox, Chrome, Safari, Opera, httpd, plugin-container, WebKitPluginHost	HTTP(S)

Local vs. wide-area HostView does not collect the host IP address, so we cannot identify the local subnet based on the host IP prefix. We develop a number of heuristics to classify traffic as local or wide-area. We define *local* traffic as all the traffic exchanged between an end-user machine and a private IP address, i. e., `192.168/16`, `172.16/12`, `10/8`. We expect this classification to correctly match most local traffic at homes, as those typically connect through a NAT gateway sharing one public IP on the outside. To avoid misclassification when the ISP employs carrier-grade NAT, we develop a second heuristic that analyzes the remote IP addresses of all traffic flows classified as local. When we observe that the remote IP addresses fall in more than five different subnets, we compute the number of connections and bytes for each remote /24 to identify whether there is a "preferred subnet", i.e., a remote subnet that carries most of the traffic (>99.9 %). If there is a preferred subnet, then we leave all traffic classified as local. Otherwise, we flag the network for manual inspection. The HostView data had a total of five home networks which contacted more than five different remote subnets, four of these had a preferred subnet. We manually inspected the remaining home network and found that a large fraction of P2P traffic going to IPs in `10.*` networks. In fact, this user's home ISP is known deploy carrier-grade NAT, so we label this `10.*` traffic as wide-area and we leave the `192.*` traffic as local. For work networks, we might misclassify local traffic as wide-area when hosts connected to the local network have public IP addresses. We address this issue with a third heuristic that labels all traffic to a destination IP address that has the exact same organization name as that of the source network as local. Finally, we classify all broadcast traffic as local. We label all the remaining traffic as *wide-area*.

Extension of network environment labels As discussed in Section 2, some of the HostView traces have no network type tag (e. g., Home or Work). We manually inspect the ISP, the network interface, and the geo-location of each unlabeled trace and assign a label. For example, we label a trace annotated with *ISP: "University of California";*

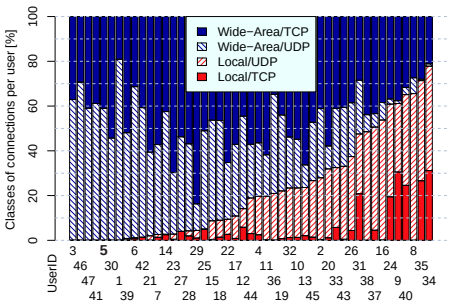

Fig. 1. Local vs. wide-area connections per user (Total number of connections per user varies between 2.5 K and 3 M.).

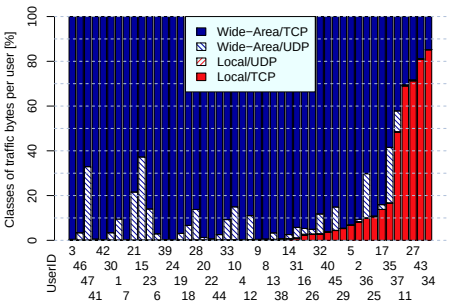

Fig. 2. Bytes transferred on local vs. wide-area connections per user (Total amount of traffic per user varies between 800 MB and 770 GB).

City: "Santa Cruz, California"; Country: "United States" as *Work*. Another example containing *ISP: "Free"; City: "Paris"; Country: "France"* is labeled *Home*. This manual classification reduced the fraction of unlabeled traces to 2 %. Some traces have no information that indicates the type of network.

Application Categorization For our analysis of popular applications we rely on a two-staged categorization process. First, we assign one of eleven application categories or "unclassified" to each connection based on the process executable name. Second, we label any connection that remains unclassified based on the application protocol as derived from the port number using the IANA mapping. We assign categories to those process names and application protocols that account for the most connections and the most volume. Table 1 lists the eleven categories and gives example process names and application protocols for each of them.

4 Results

This section first compares local and wide-area traffic in general. Then, it studies the split of local and wide-area traffic at home and at work.

Local vs. Internet: Connection and Bytes Figures 1 and 2 show the fraction of local (two bottom bars) and wide-area (two top bars) traffic for each user (UserIDs are the same across figures for comparison). For each user, we separate UDP (shaded bars) from TCP (solid bars) traffic. We consider the composition of traffic by number of connections (Figure 1) and bytes (Figure 2).

Take the example of the rightmost user in Figure 1, UserID *34*, 77 % (46 % UDP and 31 % TCP) of this user's connections are local. The remaining traffic is directed to the Internet (0 % UDP and 23 % TCP). In general, we observe that Internet traffic dominates both in number of connections and bytes, although this dominance is much more pronounced for bytes. In total, we classify 780 GB as local and 3 TB as wide-area traffic. Furthermore, we see that UDP dominates local connections for almost 80 % of the users. The absence of shaded bars in Figure 2 clearly shows that almost all bytes are transferred in TCP connections (>89 %).

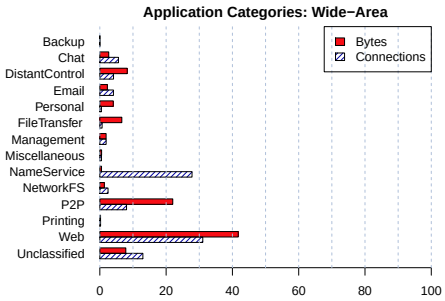

Fig. 3. Application mix for wide-area traffic.

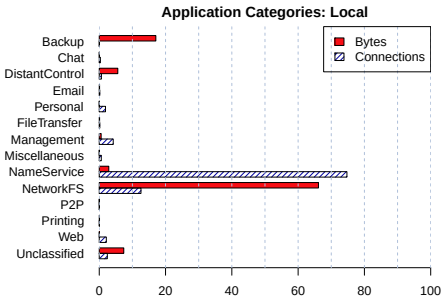

Fig. 4. Application mix for local traffic.

We observe that the four rightmost users in Figure 2 transfer more bytes locally than in the wide-area. As we discuss in the next section, most of this traffic corresponds to network file system, so these users could be playing music or watching videos from a local network storage. In Figure 2, more than half of the users exchange almost all traffic with hosts in the wide-area (corroborating previous findings [5]). In the rare cases these users do exchange traffic with hosts in the local network, they mainly perform file transfers.

Local vs. Internet: Application Mix We now study how local and wide-area applications differ. Figures 3 and 4 show the application mix in terms of connections (shaded bars) and data bytes (solid bars). These figures use the application categorization method described in Section 3, which leaves no more than 12 % of connections and 7 % of bytes *unclassified*.

Figure 3 shows the application mix for wide-area traffic. We see that the proportion of bytes per application class agrees with results from previous studies [6,7]. Web traffic and P2P are the top applications. In addition, we see some file transfers and distant control traffic (ssh and VNC). When we classify in terms of number of connections, the mix changes and name services take the second place behind Web. Chat and Email are also more prevalent in terms of connections than bytes.

Figure 4 shows that name services (e. g., DNS) dominates local traffic in terms of connections, whereas backup and network file systems (e. g., AFP and SMB) in terms of bytes. A previous study of enterprise traffic [9] also found that network file system and name service dominate local traffic, but their study found considerably more local email and web traffic than what we find. A significant part of our data is of home traffic, which may explain this difference. We now split the traffic into home and work.

Traffic at Home and Work Our analysis so far has mixed traffic from multiple network environments, including home, work, airports, coffee shops, or hotels. Based on our extended environment labels (see Section 3) we investigate the differences not only between local and wide-area traffic, but also across different types of network environments. Figure 5 shows the distribution of traffic and users over the different environments. Note that a single user can visit multiple environments. After applying our heuristics the ‘Other’ category, which includes instances when users labeled the environment as other and when our heuristic could not label the environment, only accounts

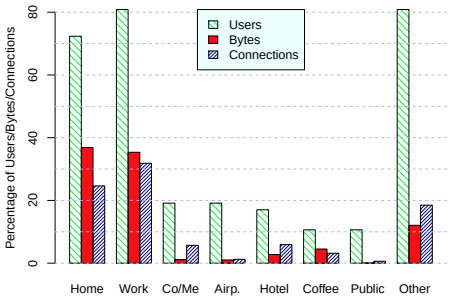

Fig. 5. Percentage of Users, volume, and connections by environment.

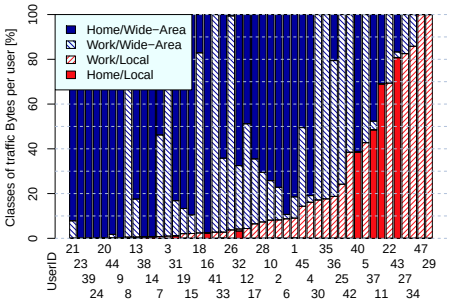

Fig. 6. Bytes transferred at home vs. work and traffic target per user.

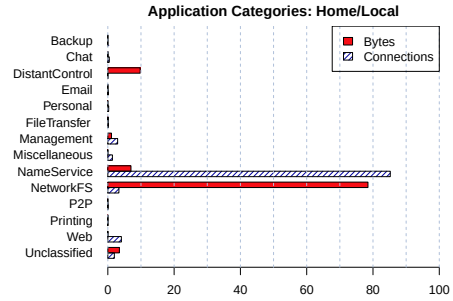

Fig. 7. Application Mix for Home/Local traffic.

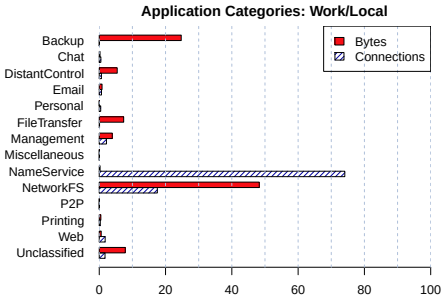

Fig. 8. Application Mix for Work/Local traffic.

for 12 % of the bytes and 18 % of the connections. We see that users (light shaded bars) are primarily at home or work, thus we select these two environments for further study. These environments include 56 % of the connections (heavy shaded bars) and 72 % of the bytes (solid bars). Moreover, our analysis of local traffic in different environments (not shown) shows that the fraction of local traffic in all environments but home and work is marginal ($<$1.25 %).

Figure 6 shows the number of bytes sent and received per user for all four combinations: home/wide-area, work/wide-area, work/local, and home/local. As expected, we see a similar split between local (bottom) and wide-area (top) traffic The differences between Figure 6 and Figure 2 happen because here we only include traffic from home and work. The majority of users has more local traffic at work. Only four users have a significant fraction of local traffic at home.

Application Mix at Home and Work Now that we established a basic understanding of how traffic differs between home and work as well as local and wide-area, we investigate the application mix in each of these cases. The analysis of wide-area traffic at work (omitted for conciseness) shows almost no P2P traffic, but a considerable fraction of file transfers and distant control traffic. These results are consistent with previous findings by Pang et al. [9].

We study the application mix of local traffic at home in Figure 7 and at work in Figure 8. Local traffic at work includes file transfers and backup traffic, which are not present in home traffic. Different from Pang et al. [9], we see little local email or web

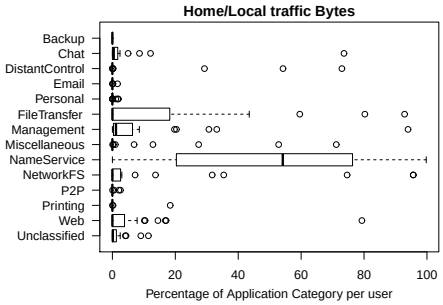

Fig. 9. Boxplot of application mix per user for Home/Local traffic.

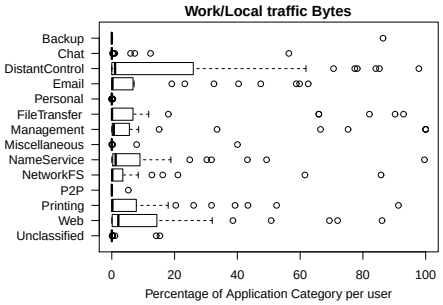

Fig. 10. Boxplot of application mix per user for Work/Local traffic.

traffic at work. Indeed, it turns out that email traffic of most HostView users is wide-area. A possible explanation is that they are typically mobile and hence rely less on local infrastructure.

Another difference is the lack of backup traffic at home, which may reflect users' preference to backup directly at external disks when at home, instead of over the network. The backup traffic at work is mainly from a single user, who is responsible for almost all the bytes of backup traffic in Figure 8. We do also observe some file transfer traffic locally at work. Most of that is transmit (file transfer client for Mac OS) and FTP, but some is Dropbox (a cloud storage/synchronization service). Given it is a cloud service (cloud = wide-area) we did not expect to find Dropbox locally. It turns out that Dropbox is using a direct connection for synchronization across devices in the same LAN. Dropbox constitutes half of the file transfers in our local home traces.

As single users can have a distorting impact on the overall traffic composition, we now calculate the application mix per user. Figures 9 and 10 show boxplots[3] of the application mix per user in terms of bytes. Each row shows the distribution of the individual contribution of the corresponding application category across all users. We find that although network file system traffic dominates local traffic, most users have less than 10 % of traffic in this category both at home and at work. Reversely, although name service represents a small percent of the total number of bytes in Figure 7, the median across all users is over 50 %. We find similar effects for file transfers at home. At work, contrary to Figure 8, we do see web, email, and printing usage.

Connection size and duration We end our analysis with a study of the characteristics of local and wide-area connections both at home and work. We show the complimentary cumulative distribution of the number of bytes per connection in Figure 11 and connection durations in Figure 12. For example, the 'work/local' point at x = 10kB in Figure 11 indicates that only 1 % (y-axis) of all the connections are larger that 10kB.

In terms of bytes, we observe in general larger (further to the right) connections for wide-area traffic. Local connections are typically small, but the largest local connections exceed the size and duration of wide-area connections. This observation confirms one

[3] The box (line inside the box) shows the quartiles (median); whiskers show nearest values not beyond a standard span from the quartiles; points beyond (outliers) are drawn individually.

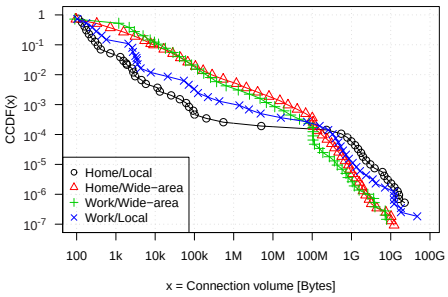

Fig. 11. CCDF of connection volumes.

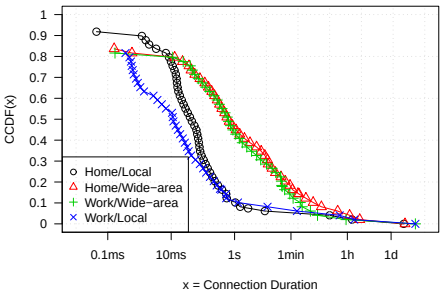

Fig. 12. CCDF of conn. durations (log-linear)

previous study showing that home traffic sometimes have short spikes [5]. Although the connection durations in Figure 12 are limited by the 4 hour trace file cutoff, most connections are shorter than this limit. We also see the local connections (circles and crosses) are up to two orders of magnitude shorter than wide-area connections.

5 Related work

Wide-area traffic measured from *inside* the network has been analyzed from different angles over the past decades [1, 6, 7, 12]. These measurements, however, cannot capture local traffic in networks at the edge. Our study analyzes local traffic and how it compares with wide-area traffic with data collected directly at end-hosts using HostView [4]. Other studies have collected and analyzed similar end-host data in the past [2, 11]. In particular, Giroire et al. [2] has compared network traffic from end-hosts across three network environments (inside the company, VPN to company, and outside the company). Different from ours, their study has not characterized local network traffic in depth and although it measured laptops of a larger number of users than HostView measured, they are all employees of a single enterprise.

Most similar to our work are the studies of one enterprise network [8, 9] and of three home networks [5]. These prior studies instrument the local network to collect packet traces and can hence observe most local and wide-area traffic. Our study measures one (or at most a couple) of end-host in each network and hence cannot have such a complete view of each of the studied networks, but it can sample a larger number of networks. The home network study focuses mainly on network performance, not on traffic characterization. Their few traffic characterization results show that wide-area traffic dominates local traffic in the three homes, but that there are some, rare spikes of local traffic. The analysis in the enterprise study [9] is most similar to ours and we contrasted their findings with ours throughout this paper. Given that Internet traffic can vary significantly among sites and over time [10], our study contributes to show the diversity of traffic patterns in different network environments.

6 Summary

This paper presented a comparison of local traffic in different network environments from the perspective of end-hosts. The advantage of using end-hosts as vantage points

is that we study traffic collected from over one hundred different edge networks. Our results showed that there is a large diversity in importance of local traffic relative to wide-area traffic, but that in general wide-area traffic dominates. In some networks (like airports and coffee-shops), we rarely see any local traffic, the only local traffic is DNS. At home and work, we do observe a non-negligible fraction of local traffic. Most local traffic is composed by short connections, but sometimes local connections transfer an extremely large number of bytes. Besides DNS, the most typical local applications are network file system and backup, but the composition of local traffic depends on the user and the network. The drawback of measuring local traffic from end-hosts is that we can only see a small fraction of each network's traffic. In the future, we plan to collect data directly from home gateways to measure all traffic from a single home over a longer period of time. In fact, home users are already deploying home gateways modified to perform measurements. We are working with the developers of Bismark (http://projectbismark.net/) to collect passive traffic measurements as well.

Acknowledgments

We thank D. Joumblatt and O. Goga for their help with the HostView data. This work was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) no. 258378 (FIGARO) and carried out at LINCS (www.lincs.fr).

References

1. CÁCERES, R., DANZIG, P. B., JAMIN, S., AND MITZEL, D. J. Characteristics of wide-area TCP/IP conversations. In *Proc. ACM SIGCOMM* (1991), pp. 101–112.
2. GIROIRE, F., CHANDRASHEKAR, J., IANNACCONE, G., PAPAGIANNAKI, K., SCHOOLER, E. M., AND TAFT, N. The cubicle vs. the coffee shop: behavioral modes in enterprise end-users. In *Proc. PAM* (2008), pp. 202–211.
3. JOUMBLATT, D., GOGA, O., TEIXEIRA, R., CHANDRASHEKAR, J., AND TAFT, N. Characterizing end-host application performance across multiple networking environments. In *Proc. INFOCOM (Mini-Conference)* (2012).
4. JOUMBLATT, D., TEIXEIRA, R., CHANDRASHEKAR, J., AND TAFT, N. Hostview: annotating end-host performance measurements with user feedback. *SIGMETRICS Perform. Eval. Rev. 38* (January 2011), 43–48.
5. KARAGIANNIS, T., CHRISTOS, G., AND KEY, P. Homemaestro: Distributed monitoring and diagnosis of performance anomalies in home networks, Oct. 2008. Tech. Rep. MSR.
6. LABOVITZ, C., IEKEL-JOHNSON, S., MCPHERSON, D., OBERHEIDE, J., AND JAHANIAN, F. Internet inter-domain traffic. In *Proc. ACM SIGCOMM* (2010), pp. 75–86.
7. MAIER, G., FELDMANN, A., PAXSON, V., AND ALLMAN, M. On dominant characteristics of residential broadband internet traffic. In *Proc. ACM IMC* (2009), pp. 90–102.
8. NECHAEV, B., ALLMAN, M., PAXSON, V., AND GURTOV, A. A preliminary analysis of tcp performance in an enterprise network. In *Proc. INM/. WREN'10* (2010), pp. 7–7.
9. PANG, R., ALLMAN, M., BENNETT, M., LEE, J., PAXSON, V., AND TIERNEY, B. A first look at modern enterprise traffic. In *Proc. ACM IMC* (2005).
10. PAXSON, V. Empirically-derived analytic models of wide- area tcp connections. *IEEE/ACM Transactions on Networking 2* (August 1994).
11. SAIKAT, G., CHANDRASHEKAR, J., TAFT, N., AND PAPAGIANNAKI, K. How healthy are today's enterprise networks? In *Proc. IMC* (2008), pp. 145–150.

12. THOMPSON, K., MILLER, G., AND WILDER, R. Wide-area internet traffic patterns and characteristics. *Network, IEEE 11*, 6 (Nov/Dec 1997), 10 –23.