Perspectives on Tracing End-Hosts: A Survey Summary

Diana Zeaiter Joumblatt, Renata Teixeira, Jaideep Chandrashekar, Nina Taft

Perspectives on Tracing End-Hosts: A Survey Summary

Diana Joumblatt, Renata Teixeira
CNRS and UPMC Paris Universitas
{diana.joumblatt,renata.teixeira}@lip6.fr

Jaideep Chandrashekar, Nina Taft
Intel Labs
{jaideep.chandrashekar,nina.taft}@intel.com

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

There is an amazing paucity of data that is collected directly from users' personal computers. One key reason for this is the perception among researchers that users are unwilling to participate in such a data collection effort. To understand the range of opinions on matters that occur with end-host data tracing, we conducted a survey of 400 computer scientists. In this paper, we summarize and share our findings.

Categories and Subject Descriptors

C.2.3 [Network Operations]: Network monitoring

General Terms

Design, Measurement

Keywords

End-host data collection, Network performance diagnosis, User experience

1. INTRODUCTION

The Internet measurement community has now spent over a decade collecting Internet traffic measurements from different vantage points such as routers, servers, and gateways, and in various networking environments—ISPs, enterprise networks, cable, DSL, and wireless networks. Unfortunately, very little has been done in terms of collecting traffic data directly on end-user laptops or desktops. Having data from a population of end-host devices would enable research in numerous interesting areas, including application and network performance diagnosis [1], network management [2], profiling for security applications [3], and energy management [4]. By placing the data collection close to the user, we also open the opportunity to get the user's point of view in each of these questions. Collecting data directly on an end-host, rather than at a server or access gateway, allows researchers to obtain a much richer view of an end-host, including a broad set of applications and networking environments.

It is easy to conjecture as to why end-host measurement datasets are scarce. As researchers we imagine that users would be largely unwilling to run a measurement tool on their personal machine for two key reasons, performance and privacy. Monitoring tools can consume enough processing and storage resources for users to observe a slowdown in performance of their machine. The second reason has to do with the big brother phenomenon—the fear that someone studying their data will find something private and that this information will end up in the wrong hands. The privacy issue is vastly complicated because personal views on privacy differ across generations, cultures and countries. It is also murky because all of our privacy technologies are only partially successful. On the one hand, privacy laws do not come into play when users voluntarily install the collection software (after being made aware of the implications); in this case, consent is implicit. However, this (anticipated) fear on the part of end-users, making them unwilling to participate in such measurement efforts has discouraged many in our community from pursuing the development of such tools and subsequently the research that relies on such tools and data.

In designing such a tool, the following types of questions will surface: Would users be willing to let us collect a particular type of data? Under what conditions might they agree? How much anonymity is enough? If researchers intend to study network performance as perceived by users, then it is critical to understand the user's perception of their system's performance. In many ways, end-host network performance is a relative thing and depends upon the experience of the user. Thus such a tool should also collect feedback from the user about their perception of various network conditions. But then another question arises: how much feedback would users be willing to provide? Since we were in the midst of designing such a tool, we decided that instead of conjecturing what people would think about each of these questions, we would ask them directly.

We thus conducted a survey which was hosted on the SurveyMonkey site. Ours is a survey of 400 computer science researchers, recruited by word of mouth, at a conference (ACM SIGCOMM 2009) and via mailing lists (e.g., end2end and tccc). Our goal in doing the survey was twofold. The first intent was to help us make certain design decisions in the development of our end-host tracing tool. The second purpose was to obtain a perspective on the range of opinions for numerous subjective issues. In this paper, encouraged by requests from several survey participants, we share our findings with the networking and measurement community.

2. SURVEY PARTICIPANTS

We surveyed 400 people from 31 countries. The bulk of survey responses were collected in August and September of 2009, however the survey is still open[1]. Our survey participants came from Europe (56%), North America (19%) and

[1]You can fill out the survey at http://www-rp.lip6.fr/~zeaiter/Survey.html.

the remaining 25% were collectively from Latin America, Asia, Middle Eastern countries and Africa. Most of these people work in computer science; 43.0% have academic positions, 34% are students and 20% work in industry; the remaining 3% were not computer scientists. A summary of the participants in terms of their geographic location and occupation type is given in Table 1.

Geography	Computer Science/Eng.			Non CS/E	Total
	Students	Academia	Industry		
Europe	71	114	34	6	**225**
US/Canada	31	25	14	5	**75**
Latin Am.	14	17	7	0	**38**
Asia	9	6	19	2	**36**
Australia	7	7	3	1	**18**
Middle East	2	3	0	1	**6**
Africa	1	0	1	0	**2**
Total	**135**	**172**	**78**	**15**	**400**

Table 1: Geographical location and occupation type of survey participants.

Because it is not uncommon today for individuals to have more than one computer, we asked our survey participants: "`How many computers do you use on a daily basis?`" The results are presented in Table 2. Only 13.43% of surveyed participants use a single computer on a daily basis; most people (49.25%) use two computers. In this table, we see a reasonable amount of consistency across occupation types. For example, within each occupation group, the fraction using only a single computer varied from 10% of industry participants to 16% of students. Based on a follow-up question, 86% of our users also reported that even though they used multiple computers, there was one computer they considered a primary computer.

The issue of how many computers an individual uses raises a distinction that is important to be made regarding the term *end-hosts*. This term actually refers to a device, or the coupling of a device and a user. By continuously monitoring a mobile device, one cannot claim to have a complete profile of a user, because of this modern day trend in which individuals employ numerous devices. Various monitoring methods, such as collecting data at servers, access gateways, or laptops each present successively increasing perspectives on the *user*, none of which are "complete". However, the research our community seeks to enable—such as application and network diagnosis, troubleshooting, security and energy reduction for end-hosts—is actually targeted towards diagnosing and protecting a *device*, and a user interacting with that device.

Occupation	Computer Science/Eng.			Non CS/E	Total
	Students	Academia	Industry		
1	22	21	8	3	**54**
2	70	82	39	7	**198**
3	30	48	16	2	**96**
4	8	11	6	2	**27**
5	3	5	3	1	**12**
$\geq$ 6	2	6	4	1	**13**

Table 2: "How many computers do you use on a day to day basis?"

The performance overhead of a measurement tool will vary across end-host computers, in part because of their operating systems and capabilities. We queried users as to the operating system used on their primary computer: 45% reported Windows, 31% Linux, and 24% MacOS. The trend was only slightly different for the students who reported 36%, 42% and 22% respectively. We point out that these numbers appear very different from market estimates for market shares of different operating systems in the general population[2]. This presumably reflects the bias in the community surveyed, namely computer scientists. We elected to survey computer scientists for two reasons. First, we are currently developing a tool to collect end-host measurement data that will primarily be run by users in the CS community and it seemed reasonable to target the survey accordingly. The second reason is that we expect computer scientists to be at ease in answering questions about the nature of the data being collected ("Are you comfortable with us collecting packet headers?"). Clearly the population at large would be unable to answer such questions. Moreover, in some cases we sought explicit technical feedback, such as the types of anonymization techniques people are comfortable with.

Unfortunately, not all survey participants answered all questions. In discussing the responses to particular questions, we indicate the number of respondents if fewer than all participants answered the question.

3. THE RANGE OF OPINIONS

In this section, we describe the spread of opinions on two key dimensions of end-host monitoring tools. The first dimension revolves around how privacy concerns impact a user's willingness to run such a tool on their machine. This is influenced by numerous factors including the type of data collected, the anonymization techniques used, the length of the experiment, and a personal connection (whether or not they know the people or institution running the experiment). We also explored whether some other features, such as including a *pause* button, could increase a user's willingness to participate. The second dimension concerns getting user feedback. We tried to discern to what extent users would be willing to interact with the data collection tool and provide annotations that would be useful in relating the network performance data to the users perception of the performance.

3.1 Privacy and Willingness

A significant hurdle for any data collection designed to run on the end-host is the wariness users feel about data being exported from their computers. It is intuitive that a user's comfort with data being exported depends directly on which type of data is gathered. We asked users how they felt about six types of data: CPU LOAD, MOUSE CLICKS, ACTIVE PROCESSES, PACKET HEADERS, CONTENT TYPE AND URLS. For each data type, we asked participants to rate them as either *comfortable*, *uncomfortable* or *deal breaker*. The label *uncomfortable* was intended to capture those who were hesitant and indecisive. By *deal breaker* we meant to find out who would refuse to install the tool if they knew it collected that particular data type. Our goal was not to ascertain the underlying causes for users having a specific point of view, but merely to find the line dividing between what is acceptable and what isn't for the majority of users.

The survey responses are summarized in Table 3. The datum are listed here in an order that we supposed intuitively would be most to least *comfortable*. The numbers in Table 3 are logical in that respondents view exporting data

[2]Up-to-date statistics of operating system's market share are available from http://marketshare.hitslink.com.

Data Type	Comfortable	Uncomfortable	Deal breaker
CPU load	**87.61%**	7.85%	4.53%
Mouse clicks	**55.29%**	28.10%	16.62%
List of active programs/processes	35.35%	**37.76%**	26.89%
Content type (port 80 connections)	26.28%	**37.16%**	36.56%
Packet Headers	26.89%	**38.07%**	35.05%
URLs	9.67%	22.66%	**67.67%**

Table 3: "How you feel about an end-host collection tool that collects each of these data types: are you completely comfortable, uncomfortable, or is any particular data item a deal breaker (i.e. you wouldn't install the tool)?"

with *higher* information content *more* negatively. We see that most of the respondents, 87.61%, don't mind CPU load being exported. More than half, 55.29% are also comfortable with mouse clicks being exported. For datum such as processes, packet headers and content type, we find that roughly 1/4 to 1/3 or participants find this acceptable. It is interesting how the numbers for URLs differ dramatically from all other types of datum. Exactly 2/3 of our respondents considered this a deal breaker. Numbers like this indicate that when building an end-host tracing tool, URLs should either be excluded altogether, or else need to be anonymized with a high degree of confidence.

There were two results that surprised us. We were surprised that only 55% were comfortable with mouse click logging as we expected this to be much higher. It seems to us that mouse clicks have very little potential to reveal private or personal information, as these are typically logged merely to indicate user presence. We suspect that there may have been a confusion in the survey itself that led to these numbers. What we had meant by MOUSE CLICKS was basically a binary signal, essentially timestamps of when the mouse was engaged. Perhaps the users misinterpreted the question as somehow gathering more information (such as what was clicked) than what we intended. Similarly, we were surprised by the resistance to recording content type of HTTP connections. By CONTENT TYPE we meant that we would record whether the data transferred in the HTTP connections was of type *text, audio, image* or *video*. We would not have expected this simple annotation to be threatening, and thus here again, we suspect that the survey participant read more into this question than we intended. This points to a lesson learned. Researchers in the Internet measurement community can benefit by working with others– HCI researchers for one– who have experience doing surveys and who can anticipate how questions may be perceived by individuals.

We wondered whether there were any differences between students, academics and those in industry, in terms of their comfort level with a given data type being collected. To look at this we break down the responses by occupation. Figure 1 shows those who selected "comfortable" for a given data type for each occupation group. Similarly, in Figure 2 we break down the responses, by occupation, of those who selected *deal-breaker* for a given data type. We see that academics are the least comfortable among the three groups with data logging, especially when it comes to things like processes, packet headers and URLs. We found it interesting, and somewhat surprising that there appears to be a trend between "academics" and "industry" people, in which industry folks seem to be roughly 10% more comfortable with data logging. Industry professionals are less likely than academics to consider a particular data type as a "deal-breaker" (Fig. 2) and are more likely to be comfortable with data logging (across all data types) (Fig. 1) than those in academia. Academics would probably argue that this is because they know best what can go wrong! On the other hand, it could also be cultural. During discussions with some of our industrial participants, they indicated that people who work for large corporations are already used to the fact that their IT departments do a great deal of logging of their devices to ensure their usage complies with company policy. Similarly, the fact that students are generally less resistant than academics could be a cultural effect based upon generational differences; it is well known that the younger generation cares less about privacy than others. Because our survey was limited to 400 participants, we are not drawing major conclusions on this front. Instead, we believe we have enough evidence to put forth a hypothesis that there exists a trend among computer scientists in which academics are more resistant to end-host monitoring than industry and student researchers.

So far, the questions were posed assuming the data would be collected without any level of anonymization performed on the data. To the specific individuals that rated a particular datum as a *deal breaker*, we posed a subsequent question which was: "`Would you reconsider and run the tool if the source IP address and all local machine identifiers were anonymized before the data is exported?`" Of the 269 respondents answering this question, 55% answered in the affirmative. Using a free form text box, we asked our participants to indicate what kind of anonymization technique would make them comfortable using such a tool. We received 40 responses. Typical examples include "CryptoPAN", "Zero Knowledge", "aggregation to meta data (i.e. means, variances, etc.)" and "deletion of IP address".

This free form feedback yielded other interesting suggestions that do not relate explicitly to anonymization techniques, but instead point to other mechanisms that could increase someone's willingness to participate in a data collection effort that collects the type of data above. First, many people said that making the code **open-source**, allowing them access to the source, would increase their likelihood of participation. Second, others said that they would feel more comfortable participating if they knew personally the people or institution(s) carrying out both the data collection and the subsequent storage of the data. This is a matter of **personal trust**, one that is established outside the technical features of a tool. Third, some said they would participate if they could **see the data** before it was exported and were able to delete sections of the trace.

We suspected that one cause of resistance to installing a monitoring tool is because users do not like the idea of a tool running *all* the time. We imagined that users might like to turn off a monitoring tool, for short periods of time, when they do something that they perceive could generate sensitive information in the logged data, for example entering credentials on email websites that use plain text passwords. This concern could be addressed by

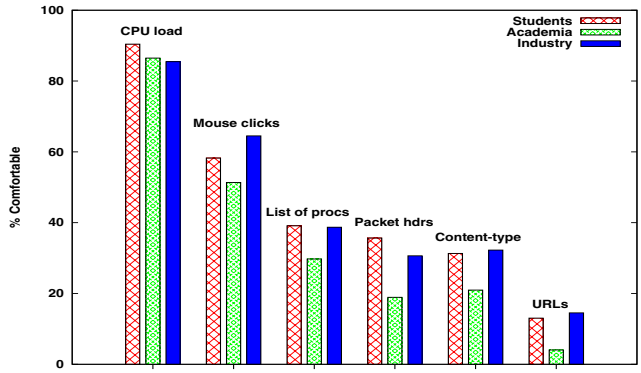

Figure 1: Percentage of users per occupation that feel comfortable collecting each of the data types.

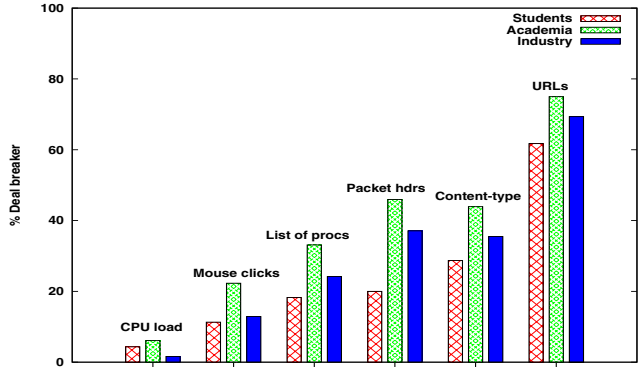

Figure 2: Percentage of users per occupation that answered "deal breaker" to the collection of each of the data types.

providing a mechanism for users to pause or suspend the data collection and logging for a brief time when a sensitive activity is carried out (similar to what is available in the Nano tool [1]). To understand if such a feature would allay users' fears, we posed the following question: "`If privacy is your over-riding concern in not wanting to run this tool on your system, would you reconsider if we implemented a` *pause* `mechanism?`" Of the 338 who answered this question, 65% did so affirmatively.

If the goal of collecting the end-host data is to build profiles of how the user interacts with their networked device, then the end-host tracing tools need to be run for a period of time that is long enough for the profile to stabilize. Earlier research illustrated that for some end-host profiles [5], or whitelists [3] this can take approximately two weeks. This means the tool would need to run for at least a month to collect equal amounts of training and testing data. On the other hand, users may only be willing to run the tool for a limited period of time, especially if the tool itself impacts the performance of the machine. To gauge people's perspectives on this, we asked our participants how long they would be willing to run a data collection tool. The results are in Table 4. We see that roughly 55% of those surveyed would be willing to run the tool for one month or longer.

3.2 Interactive Feedback

A user's perception of quality is a personal and subjective measure. Nevertheless, it is important to understand if the end-host data is meant to inform research on performance diagnosis and management. This is because the user's perspective often suggests the target performance level. Such

Occupation	Experiment Duration				
	Never	One week	One month	Three months	As long as necessary
Students	11.32%	27.36%	23.58%	8.49%	29.25%
Academia	23.74%	27.34%	19.42%	4.32%	25.18%
Industry	14.29%	19.64%	25.00%	12.50%	28.57%
Total	**17.3%**	**26.6%**	**22.1%**	**7.4%**	**26.6%**

Table 4: "Knowing what data is being logged and the level of interaction required, how long would you be willing to run the tool on your own computer?" (This question was answered by 306 participants.)

Occupation	Click		Record Incident	
	Yes	No	Yes	No
Students	86	22	70	38
Academia	112	29	67	74
Industry	46	11	35	22
Total	**244**	**62**	**172**	**134**

Table 5: "Would you use an *I am annoyed* button? If you answered yes, would you be willing to record the incident/experience that prompted you to push this button?"

research frequently needs to correlate a user's perception of network quality and performance to quantitatively measured parameters of network quality. However obtaining good quality user feedback exposes an inherent tension in the design of the monitoring tool. On one hand, we would want a tool to be as unobtrusive as possible, yet on the other hand, we need the user's feedback for many important research activities.

In order to obtain user feedback, Human Computing Interaction (HCI) researchers frequently employ two mechanisms called, *participant triggered feedback* and *experience sampling* [6]. We believe these techniques could be adopted into our community's efforts in developing end-host tracing tools. We also used our survey to understand to what extent people would employ or respond to these mechanisms. For participant triggered feedback, we considered including an I'M ANNOYED button in our tool. Such a button would exist as a small icon on the bottom of a screen, and the user is free to click on it when they are annoyed or unhappy with network performance. This feedback can thus be as simple as clicking a button once. To obtain more information about why the user is annoyed, an optional text box can be provided for the user to describe the associated application and/or provide more context. We asked our survey participants "`Would you use an "I am annoyed" button: a small icon in your tray that you can optionally click on when annoyed at the performance of an application?`". 80% of our respondents said 'yes' to this question. We subsequently asked "`If you answered yes, would you be willing to record the incident/experience that prompted you to push the "I am annoyed" button?`". Here, 57% said 'yes'. These results are shown in more detail in Table 5. This generally positive response indicates that it is worth providing a feature such as an I'M ANNOYED button in an end-host tracing tool. We found that more participants were willing to fill out a text box, and provide information (e.g. symptoms) about a poor performance incident, then we expected: 65% of students, 61% of industry professionals, and 48% of people in academia said they would do so. (Again we see a slight trend of academics being most resistant to various forms of participation.)

The Experience Sampling Method (ESM) originates from the field of psychology [7] and has been widely adopted within the HCI community [6]. The idea behind ESM is to prompt the user a few times a day to respond to a small set of questions pertaining to their experience immediately prior to the moment that the questions are presented. A typical questionnaire might include 2-3 multiple choice questions, such as asking the user to rate the quality of their network connection on a standard Likert scale from 1 to 5 (widely used scale in survey research). In addition, an optional free form text box allows the user to enter any information they like about their experience. The former provides quantitative data that can be used for statistical summaries, while the latter provides qualitative data. Users can be automatically prompted for feedback either at regular intervals, at random, or based upon traffic load. Triggering questionnaires at low, medium and high traffic levels allows researchers to calibrate users and check for consistency. We asked our participants how often they would be willing to answer a short questionnaire. 45% of our respondents replied that they would be willing to answer 2 to 3 times per day or more (i.e. this number includes those who would respond more often than 2-3 times/day). Amazingly enough, 6% said they would be willing to provide feedback every hour! We conclude that including an ESM questionnaire in an end-host tracing experiment is worth doing because if roughly half of the participants actually complete the questionnaire regularly, then this could provide a substantial amount of rich data for researchers (assuming a reasonably sized set of participants).

Our colleagues in the HCI community pointed out to us an important issue when using ESM questionnaires, namely that it is always hard to write user feedback questions that are unambiguous. They highly recommend a pilot study of the questionnaire to uncover ambiguities in how volunteers interpret the questions. They also suggested writing feedback questionnaires in the native language where the measurement collection is carried out, to make it as easy as possible for the user to complete the questionnaire.

4. CONCLUSION

In this survey, we learned that the range of opinions on the topic of privacy and willingness is broad. By this we mean that there are plenty of people in each camp, those who are willing, those who won't, and interestingly there are many people who are in an intermediate category of being unsure or hesitant. This means that it is worth either adding into the tool particular features (such as a pause button, or a particular type of anonymization technique), or removing some features (such as the collection of a particular data type) in the hopes of winning over the hesitant crowd to participation. When researchers plan to do data collection on end-hosts, they need to ask a larger number of people than they expect to participate, since only a fraction will actually do so; our data indicates that for this community asking roughly three to four times the number of desired participants is a good target. Overall we find the numbers in this survey encouraging in that they indicate that those willing to participate are not a small set of people. The issue of incentives also plays an important role in addressing the general willingness of people to participate in end-host tracing; however we do not address this issue herein because our survey did not explore this aspect of our participants views.

It is not common practice in the Internet measurement community of researchers to conduct user surveys. We found this experience to be very useful in that it helped us to make explicit design decisions in an end-host tracing tool we are developing[3]. In doing this survey, we learned a fair amount about the methodology of conducting user surveys, such as the importance of carefully worded questions, developing follow-on questions, and understanding the many ways in which a simple question can be interpreted. We believe that conducting user surveys is a technique that could prove useful to many research efforts in our community, especially those interested in research that influences user perception of performance.

ACKNOWLEDGEMENTS

The authors would like to thank Sunny Consolvo, Tye Rattenbury, Jaeyeon Jung and Allison Woodruff for their input on participant feedback and experience sampling. We also thank Nick Feamster for his valuable comments and all participants of the survey. This work was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) no. 223850 and the ANR project C'Mon.

5. REFERENCES

[1] M. Tariq, M. Motiwala, N. Feamster, and M. Ammar, "Detecting Network Neutrality Violations with Causal Inference," in *ACM CoNEXT*, 2009.

[2] T. Karagiannis, R. Mortier, and A. Rowstron, "Network exception handlers: Host-network control in enterprise networks," *ACM SIGCOMM*, 2008.

[3] F. Giroire, J. Chandrashekar, N. Taft, E. Schooler, and K. Papagiannaki, "Exploiting Temporal Persistence to Detect Covert Botnet Channels," in *in Proc. of Recent Advances in Intrusion Detection (RAID)*, 2009.

[4] S. Nedevschi, J. Chandrashekar, J. Liu, B. Nordman, S. Ratnasamy, and N. Taft, "Skilled in the Art of Being Idle: Reducing Energy Wasted in Networked Systems," in *In Proc. of Networked Systems Design and Implementation (NSDI)*, April 2009.

[5] T. Karagiannis, K. Papagiannaki, N. Taft, and M. Faloutsos, "Profiling the End Host," in *Passive and Active Measurement Workshop (PAM)*, April 2007.

[6] S. Consolvo and M. Walker, "Using the Experience Sampling Method to Evaluate Ubicomp Applications," *IEEE Pervasive Computing Magazine*, vol. 2, no. 2, 2003.

[7] M. Csikszentmihalyi and R. Larson, "Validity and Reliability of the Experience-Sampling Method," *Journal of Nervous and Mental Disease*, no. 175, pp. 526–536, 1987.

[3] http://cmon.lip6.fr/EMD.