Issues and Etiquette Concerning Use of Shared Measurement Data

Mark Allman
ICSI
Berkeley, CA, USA
mallman@icir.org

Vern Paxson
ICSI & LBNL
Berkeley, CA, USA
vern@icir.org

ABSTRACT

In this note we discuss issues surrounding how to provide and use network measurement data made available for sharing among researchers. While previous work has focused on the technical details of enabling sharing via traffic anonymization, we focus on higher-level aspects of the process such as potential harm to the provider (e.g., by de-anonymizing a shared dataset) or interactions to strengthen subsequent research (e.g., helping to establish ground truth). We believe the community would benefit from a dialog regarding expectations and responsibilities of data providers, and the etiquette involved with using others' measurement data. To this end, we provide a set of guidelines that aim to aid the process of sharing measurement data. We present these not as specific rules, but rather a framework under which providers and users can better attain a mutual understanding about how to treat particular datasets.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General; C.2.3 [**Computer-Communication Networks**]: Network Operations; C.2.m [**Computer-Communication Networks**]: Miscellaneous

General Terms

Measurement,Experimentation

Keywords

Data Sharing,Anonymization

1. INTRODUCTION

Collecting a substantive set of network measurements generally requires both favorable circumstances and hard work. The circumstances regard having the right opportunity: administrative and legal permission (particularly for passive measurements), operational support (e.g., installation of the measurement apparatus, subsequent access to it for maintenance), and access to sufficient resources (e.g., disk space, network taps, kernel mods, smart students). The hard work spans developing and debugging associated software, calibration, monitoring the collection process for faults, organizing the resulting data, ascertaining and capturing appropriate meta-data, iterating the entire process to fix problems that arise, and interacting with the different parties that make up the favorable circumstances.

Given these considerable difficulties, there is great utility in researchers being able to share measurement datasets rather than having to independently acquire them. In addition, for Internet measurement studies in particular there is major benefit in sharing datasets in order to gain broader, more representative insight into the highly diverse nature of Internet traffic and dynamics.

Researchers have advocated sharing data for years (e.g., [4, 16, 9]). An early effort of ours, the Internet Traffic Archive [15], was established as a place where researchers could make their data available, but proved more time-consuming to maintain than we had anticipated, and we failed to keep it growing over time. More recently, several databases of released[1] (mostly passive) datasets have been established (e.g., DatCat [17] and PREDICT [3]). In addition, individual groups have begun making measurement data available in somewhat ad-hoc ways (e.g., [2, 1]). Network measurement conferences (PAM and IMC, in particular) have also been trying to encourage sharing of data through awards for high-quality new datasets released for general use. Finally, a significant degree of non-public sharing of data occurs, both within institutions and between collaborators at different institutions.

We strongly encourage the widespread sharing of measurement data. However, we are sometimes dismayed at how such data is handled. We have used and provided measurement data over the years (both publicly and privately) and we have been struck by the range of attitudes and assumptions present in the community about providing and using shared measurement data. In this note we attempt to pose a set of reasonable, high-level considerations for sharing and using measurement data. We do not attempt to offer a complete set of "ground rules". Such a list is impossible to create, as each data-sharing situation has its own unique considerations and associated threat model, requiring careful, individual evaluation. Instead, we attempt to sketch some of the high-level issues researchers should take into consideration—and sometimes explicitly specify—when providing and using shared measurement data.

We emphasize that in this note we have no aim to "lay blame," and in fact our own data sharing activities have not always adhered to the considerations for which we advocate here. Rather, our understanding of the issues has evolved with experience.

[1]We use the term "released" rather loosely. Measurements described in one of these databases have not necessarily been *publicly* released, but the researchers who collected the datasets are willing to share them with other researchers under some conditions.

IMC'07, October 24-26, 2007, San Diego, California, USA.

In the next section we present issues regarding the decision and process of releasing data for shared use. We follow this in § 3 with discussion of the use of such data by others. In § 4 we consider where releasing data meets use of data, i.e., interactions between the providers and the users. We offer final thoughts in § 5.

2. DATA RELEASE CONSIDERATIONS

Releasing data is fraught with potential problems—so much so that most institutions will not even consider it. These problems include potentially compromising the privacy of users, exposing activity that might embarrass the institution, revealing information about the site's network that could enable an attacker to more effectively mount an attack, and exposing aspects of the network's operation to possible competitors. That said, there are also significant benefits a site can gain by releasing data, in terms of furthering understanding about network activity in ways that can directly or indirectly help the institution, and garnering positive recognition within the community.

In general, the goal of a data release is to minimize the potential problems while at the same time trying to maximize the research value of the data, *at least for some context* (see below, and also [13] for our exploration with colleagues of such issues in the context of a particular data release). To aid with this process, researchers have developed a number of anonymization techniques to scrub data for release [10, 18, 5, 14, 13]. While useful, these techniques do not—and cannot—provide guaranteed protection against information leakage. Therefore, as discussed in more detail in [13], ultimately the choice about what to release, how to obscure the data, and to whom to release the data, are *policy decisions*. Furthermore, different policies apply to different situations. For example, a university's network operators might consider professors, their students, and the public, to constitute three different threat models, deserving three different datasets anonymized in three different ways. Another example concerns providing data with only a narrowly defined task in mind. In this case, the provider might well discard the more detailed elements of each data record (e.g., a NetFlow record or a packet trace), rendering the data generally useless outside the context of the given study.

The first (obvious) guideline for data release is for providers to carefully understand the threat model for each situation and to use this understanding to frame the anonymization policy applied to the data. A caveat is that no matter how careful a provider is, they need to understand that *they are releasing more information than they think*. Network measurement is often about inferring subtle information from observed traffic, and the techniques for doing so evolve continually. Therefore, as researchers contrive better inference mechanisms, previously "safe" data can become vulnerable to some forms of information leakage. A key counterpoint here, however, is that *as data ages it often becomes less sensitive*, because the network has changed, IP addresses no longer reflect current users or remote servers, meta-data that links one type of activity to another has disappeared, apparent vulnerabilities have since been addressed, and so on.

The second guideline is to enumerate an *explicit* Acceptable Use policy for the data. While it may seem obvious what constitutes "scholarly use" of measurement data, the sensitivity can in fact significantly depend on where data was collected. Therefore, the best course of action is for the provider to explicitly state the bounds of what sort of analysis they wish to allow or disallow researchers to pursue with their data. For example (illustrative, not exhaustive):

- The user must not attempt to de-anonymize the shared measurement data.
- The user may use the data only for assessing the following characteristics of traffic: transfer length in bytes and duration.
- The user may use the data to develop new techniques for finding subverted hosts that are part of botnets.
- The shared data is not a general resource; each use of the data should be undertaken only after gaining explicit permission from the provider.
- Users noticing failures in the anonymization process are asked to please inform the data provider.

Clearly a simple statement disallowing some activity will not stop it. However, if the users of the data violate the explicit terms then they run the risk of receiving no more data from the given provider (and, possibly others with whom the provider shares their experiences). Further, our hope is that Acceptable Use policies can be cited in papers—and therefore guide reviewers, program committees and editors as they evaluate work that may violate the given use constraints. If this were to be the case, then researchers who do not adhere to the guidelines would run the risk of enduring censure when attempting to publish their work.

A third guideline for providers is to be explicit about the interactions they are willing to have with the users. Analyzing measurement data often leads to questions about the environment, collection strategy, filtering artifacts, additional activity, etc. Often, more details about the traffic garnered from a raw, non-anonymized version of the dataset can shed light on these questions. Data providers should be explicit about what sorts of assistance they can and will provide to data users when these sorts of questions crop up.

A final guideline, related to the above, is that providers should explicate what sort of raw data and meta-data pertaining to the shared data that the provider intends to retain, and for how long. For instance, consider a provider who uses a packet trace to produce a log of TCP connection summaries that they then anonymize and share. It may be quite useful for users of the summaries to understand how long (or even if) the provider retains the raw packet traces such that the user can ask appropriate questions when chasing down puzzles in the data. Further, it can be quite helpful for users to understand the degree to which the provider keeps ancillary data about hosts, servers, networks, etc., that can shed light on aspects of the shared data (e.g., a snapshot of the network topology at the time a dataset was collected).

Lastly, providers should consider explicitly stating what notification they desire regarding use of the data or its appearance in a publication, and the desired form of acknowledgment that such publications should include.

3. DATA USE CONSIDERATIONS

While care is clearly necessary (and exercised) when releasing data, care is also required when *using* others' data for scholarly purposes. First and foremost, users should understand that releasing data is difficult (at best) for the provider. For example, we recently publicly released packet traces recorded on internal networks at Lawrence Berkeley National Laboratory [2]. The effort to design and implement suitable anonymization policies (described in [13]), as well as to obtain approval to release traces given the policies, took months[2] of effort on the part of several individuals. From discussions with colleagues, we believe this experience is indicative of that of others who have released network measurement data.

[2]Not including the effort to collect and study the data, as described in [12].

Given such costs, clearly data providers will strongly desire that users of shared measurement data should do nothing to hinder the ability of the providers from releasing more data (to that user or others) in the future, or anything that would have a broader chilling effect on the community's ability to release data. Organizations have an immense list of reasons to say "No" to providing data to the research community. Organizations that say "Yes" do so with the hopes of helping the community and the understanding of networks, as well as gaining positive recognition within the community. The community should therefore endeavor to treat the data as responsibly as possible. At a minimum, data users should scrupulously follow Acceptable Use policies that accompany the data. In addition, if use drifts into any sort of "grey area," the user should consult the data provider, as discussed below.

We now explore four topics in more detail.

3.1 Reporting

Reporting findings obtained from using others' data can sometimes be tricky. Ultimately, researchers strive to learn something *new* about the network, protocols, services or hosts present in a dataset. Very often, these results reflect characteristics somewhat particular to the network from which the shared data comes. The presentation of this new information can possibly lead to an undesirable impact on the data provider.

For instance, consider a study on a new mechanism for detecting hosts scanning a network. It may well be useful for a researcher to use data such as that provided in [13] to attempt to reverse engineer the mechanism the site used to detect scanners, such that the researcher can then deduce a plausible set of scans that passed through the site's filter—and then whether these scans would be caught by the proposed new mechanism. Success here (finding scans previously uncaught) casts the researcher's innovation in a favorable light; but, by publicizing the approach the site currently uses to detect, can undermine the site's security posture.

We offer two coping strategies for researchers to use in this regard:

- **Aggregate.** When reporting results, often one can aggregate information to reduce its sensitivity. For example, in the above example one could characterize the general nature of the scans that made it through the site's firewall over the course of a day (say), instead of providing estimates for the fine-grained thresholds on the detection algorithm.

- **Further Anonymization.** When discussing particular artifacts in a dataset, researchers can go beyond simply reusing the anonymization applied by the data provider and instead *re-anonymize* the data they report, such as referring to hosts in abstract terms ("host A") rather than identifying specific hosts present in the dataset.

3.2 Purpose-Provided Data

Publicly released datasets clearly provide a major benefit to the community in terms of allowing broad access to measurement data. However, it behooves us to also consider data shared more informally during collaborations; among friendly researchers to help each other out; or with students for a particular project. Often in these cases the provider is more lax on anonymizing the data because they have significant trust in the researcher. For instance, Blanton developed *tcpurify* [5] as a quick way to obscure IP addresses from students who needed to work with packet header traces for a specific project [6]. In this context, the students could be generally trusted, and access to the packet traces carefully controlled, but the operators still felt much more comfortable with not providing the students with the direct data, due to important concerns regarding both privacy and accidental revelations that might turn up during analysis of the datasets.

In such contexts, in addition to allocating less effort to anonymization, the data provider likely allocates less effort to developing Acceptable Use statements (as suggested above). We offer a few general guidelines for this situation:

- A researcher *must not* further re-distribute non-public data a provider has shared with the researcher.

- Closely related to the last point, researchers should exercise great care when storing non-public measurement data, to *ensure* that the data remains inaccessible to anyone outside the given project.

- When sharing non-public data, researchers should explicitly inform providers as to who will have access to the data. E.g., a professor should identify the students who will work on the given project, rather than simply assuming the provider understands that students will naturally be given access to the data.

- Researchers should employ the data only for the project / analysis for which it was provided. Researchers *must not* treat privately shared data as a general resource that can be analyzed at will without the explicit consent of the provider.[3]

  Note: In general it makes sense for researchers to keep data they have collected to address concerns with their analysis, or to follow up on questions generated by peer review or subsequent publication. However, it also makes sense to delete someone else's data—provided for a specific purpose—as soon as it is no longer needed, such that the data does not fall into the wrong hands, or lead to the temptation for reuse in another context. The tension between these competing goals is fairly fundamental, and therefore we encourage providers and users to explicitly address data retention as part of an Acceptable Use policy.

- If a researcher wishes to employ data for another task, they should *seek permission* from the provider. First and foremost, the provider may not wish their data to be used for the new purpose the researcher has in mind. In fact, the provider might already be engaged in research on the new topic themselves, for which being "scooped" by someone using their own data would prove highly frustrating.

  In addition, the provider may have transformed the shared data in a way that can (sometimes invisibly) render the researcher's results incorrect. For instance, consider a packet trace for which the provider removed large bulk transfers corresponding to backup traffic, because these consumed a great amount of space in the trace yet had little bearing on the analysis for which the provider originally made the data available. However, if that data were subsequently studied for network utilization, it would show the network much less loaded than actually was the case.

  Finally, the sensitivity "threat model" the provider had in mind when originally providing the trace may differ in the context of the new form of analysis, for which the researcher using the data has a strong obligation to honor the provider's concerns.

[3] In our experience, reusing data can prove a significant temptation, due to the general difficulty of obtaining rich, apt datasets.

3.3 De-anonymizing Data

More than any other activity, efforts to de-anonymize shared measurement data have the potential to cause serious problems for future data release *across the community*. Careless de-anonymization efforts can violate privacy, increase a site's exposure to security problems, or potentially embarrass a data provider. In addition, careless reporting on such activities from the research community itself—the very people the data provider is trying to help—can profoundly change the threat model applied to future data release. Thus, it is important to appreciate that such activities can have a chilling effect across the community, rendering potential providers not even involved with a de-anonymized dataset quite reluctant to release data in the future.

Therefore, first and foremost we emphasize that *de-anonymization of measurement data should not be undertaken as sport*.

That said, there are scholarly reasons to attempt to de-anonymize measurement data. If a researcher can illustrate how to leverage a modicum of information to untangle an anonymization scheme, and doing so points to better anonymization techniques, then such an effort can comprise a significant benefit for the community. However, researchers wishing to engage in this sort of analysis must proceed carefully. First, they should undertake such an activity only with the consent of the data provider—either because such activity is part of the normal Acceptable Use policy provided with the data, or per a specific arrangement with the data provider. Second, reporting on such an investigation should refrain from openly publishing the specific, de-anonymized data.

The data provider often holds the *ground truth*, such that they can inform researchers whether their de-anonymization schemes succeeded.[4] Therefore, as a part of the scientific process, an investigator attempting to de-anonymize data for scholarly purposes should try to verify their results with the data provider. A natural hesitation to approaching providers in this way can arise because of a perceived conflict-of-interest: the data provider may simply indicate that the researcher did not properly de-anonymize the data—regardless of whether the researcher was accurate or not—in the hopes that the researcher will thus not publicize their efforts, and hence keep unrevealed any information problematic for the provider. On the other hand, the data provider has a vested interest in understanding flaws in their anonymization scheme, that they might fix the problems before releasing more data. In addition, if the provider and researcher follow the general advice in this note, then there will already be an understanding of what the researcher is doing, and therefore likely a working relationship such that a reflexive "nope, not right" reaction from the provider becomes less likely.

We also note that reports of de-anonymization techniques should not directly un-mask details of a dataset (e.g., IP addresses). Significantly better is to describe the process and note that the provider has verified that it indeed recovers sensitive data. (Of course, providing a fix for the problem in terms of a more secure way to perform the anonymization is also quite useful.) We encourage reviewers, program committees and editors to require authors to follow this path, rather than publishing sensitive details of datasets.

To avoid the thorny issues of dealing with and reporting on others' data, a different approach for researchers studying attacks on anonymization is to focus on the anonymization *techniques* rather than the anonymized data. That is, the researchers can re-apply the anonymization used by a particular data provider, but to data that they themselves capture. They then assess the possible attacks against the new data. Such an approach can completely factor out a provider's sensitive data from the investigation. On the other hand, collecting data from a variety of sources inevitably yields different artifacts. Therefore, without using the provider's data, the researcher may not get as full a picture of the strength of the provider's anonymization techniques.

3.3.1 Case Study

As a concrete example, [7] reports on an investigation into de-anonymizing several recently released datasets. The authors of the study do this with an eye towards enhancing the community's understanding of anonymization techniques and where they break down. The study provides a useful illustration of several of the items discussed above.

The study reports apparent mappings between the IP addresses in anonymized datasets and the real IP addresses, as inferred by the authors' techniques. However, the authors did not approach the data providers regarding this attempt, for fear of creating a conflict of interest [11], illustrating the uncertain community culture regarding how to undertake such studies.[5] The down-side to the study not including such interactions is that the authors were unable to compare their results with ground truth—and thus, in fact, ended up incorrectly de-anonymizing nearly all of the IP address mappings reported in the paper for our LBNL dataset [2], to the detriment of assessing the underlying scientific issues. On the other hand, the authors' de-anonymization scheme clearly has significant merit, since in addition they employed the same anonymization techniques as used on the LBNL traces *on their own packet traces* (in line with the approach for which we advocate above), for which their de-anonymization techniques worked well. Taken together, the above two points also nicely illustrate why a breadth of data can be highly beneficial when analyzing measurement data.

The final point we draw from this example is that the concerns expressed in this section are *not theoretical*. Exposing a purported mapping of anonymized IP addresses to real IP addresses risks making further release of data from LBNL more difficult. While the general form of the techniques presented in [7] is well known, and was in fact taken into account by LBNL [13], the threat model used at LBNL was of a malicious attacker—*not* the research community—scouring the data for information. At a minimum, the actions of the *research community* will need to be explained and defended to LBNL's decision makers (likely the CIO) before additional data gains approval for release.

As noted in § 1, our understanding of the issues with data release and use has evolved over time. In this case, we failed to accompany our data with a discussion of expectations regarding use of the data, and the terms of our commitment to work with researchers studying it. We believe the community's understanding of these issues is also evolving. Therefore, we should aim to understand the different perspectives of the involved researchers, and from this work towards a common data-sharing culture for the community.

3.4 Notification and Acknowledgment

Earlier we discussed how data providers should consider explicitly stating what sort of notification they would like when a researcher uses their data or when it later appears in a publication, and what sort of acknowledgment any such publications should include. Naturally, data users should honor these requests. In addition, if data users are uncertain regarding the provider's desired

[4] This is not always true, as some data providers (partially) anonymize using keys that they subsequently discard [8].

[5] Furthermore, the authors of [7] perceived a *double* conflict-of-interest in this case, because the data providers were also the authors of the anonymization techniques.

notification policies, they should attempt to contact the provider to learn them. In the absence of explicit guidance, best is to assume that the provider desires notification and an acknowledgment in publications, including the location of the data, if publicly available.

The above points may seem somewhat obvious, but we note them here to frame an additional facet of such notification/credit of which data users are often unaware: some data providers find it highly beneficial—either internal to their organization, or when interacting with their funders—to tabulate the uses that researchers have made of the released data. Since data release entails significant work for the provider in gaining the institute's approval, obtaining funding to support an altruistic activity, and anonymizing the data it behooves the community to give providers the necessary "ammo" for presenting a case that such release has broad benefit, and results in positive public recognition for both the institute and those responsible for the data release.[6]

4. INTERACTIONS

In the previous two sections we have discussed what data providers should furnish to users (§ 2) and what responsibilities the users of the data have to the provider (§ 3). In this section, we explore issues regarding subsequent interactions between data providers and data users. Analyzing measurement data is often a messy process, whereby additional context can often shed much light on the observed phenomena. Unfortunately, often when using someone else's data, the amount of additional context is in short supply. In addition, anonymizing data—nearly always a requirement when sharing—tends to introduce additional blind spots into the analysis process, which can leave researchers using shared data with lingering questions.

We advocate that researchers should ask data providers explicit questions when such situations arise, rather than making independent assumptions or assertions about the data. When doing so, researchers should temper their questions to the data providers to only those that are vital to their analysis. Of course, data providers may or may not be able to answer the questions for a variety of reasons (e.g., lack of time/energy, lack of additional context, privacy/competitive concerns, etc.). Further, seemingly mundane questions can result in a large amount of analysis to find an acceptable answer. While we do not wish to put providers on the hook for answering every question that comes their way, we suggest that providers make reasonable efforts to answer reasonable questions about data sets they release, to help foster an effective culture for sharing measurement data.

As discussed in § 3.3, one natural place where puzzles arise concerns working on de-anonymizing data. As sketched above, this activity should only be conducted under mutual agreement between the provider and the researcher. As part of this agreement, the parties should discuss information the provider will convey when checking the de-anonymized data against ground truth.

In the case where researchers ultimately must make assumptions about the data because they cannot get answers from the provider, they should explicitly note these assumptions when reporting on the data analysis. In addition, they should frame their efforts to work with the data provider.

In turn, peer reviewers should expect communication on key points of the analysis between providers and researchers, and resist cases where researchers have seemingly not made efforts to validate their assumptions with data providers.[7] We stress, however, that we advocate applying such a standard only for *key* analysis points; not for minor or tangential aspects of the data analysis.

[6]For instance, we would advocate that preparing and releasing a broadly useful dataset be considered a valuable scholarly activity when the researchers involved are evaluated.

5. SUMMARY

Our goal for this note is to help evolve the community's understanding about the care required when releasing measurement data, and the sensitivity of others using such data. We advocate that data providers be explicit in terms of a dataset's acceptable use, and researchers thoughtful in the reporting of potentially sensitive information gleaned from others' data. Data providers should also convey what interactions they desire or will accommodate, and researchers should comply with such interactions.

In general, measurement is a painfully laborious undertaking, and therefore there is great benefit in leveraging others' efforts in the form of shared measurement data. But providing such data is not without its own significant labors. Thus, it behooves the research community to foster a culture to support such sharing as best we can.

6. ACKNOWLEDGMENTS

We thank Paul Barford, Fabian Monrose and Mike Reiter for their discussion of the issues presented in this note, and for feedback on a draft version. This work was supported in DHS Award HSHQPA4X03322 and NSF Awards ITR/ANI-0205519 and NSF-0433702. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD). http://crawdad.cs.dartmouth.edu/.

[2] Enterprise tracing project. http://www.icir.org/enterprise-tracing/.

[3] Protected Repository for the Defense of Infrastructure against CyberThreats. http://www.predict.org/.

[4] M. Allman, E. Blanton, and W. Eddy. A Scalable System for Sharing Internet Measurements. In *Passive and Active Measurement Workshop*, Mar. 2002.

[5] E. Blanton. tcpurify, May 2004. http://irg.cs.ohiou.edu/~eblanton/tcpurify/.

[6] E. Blanton. Personal communication, Apr. 2007.

[7] S. Coull, C. Wright, F. Monrose, M. Collins, and M. Reiter. Playing Devil's Advocate: Inferring Sensitive Information from Anonymized Network Traces . In *Proceedings of the Network and Distributed System Security Symposium*, Feb. 2007.

[8] J. Heidemann. Personal communication, Apr. 2007.

[9] k. claffy, M. Crovella, T. Friedman, C. Shannon, and N. Spring. Community-Oriented Network Measurement Infrastructure (CONMI) Workshop Report. *ACM Computer Communication Review*, 36(2):41–48, Apr. 2006.

[10] G. Minshall. tcpdpriv, Aug. 1997. http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html.

[11] F. Monrose and M. Reiter. Personal communication, Apr. 2007.

[12] R. Pang, M. Allman, M. Bennett, J. Lee, V. Paxson, and B. Tierney. A First Look at Modern Enterprise Traffic. In

[7]Researchers can indicate communication by explaining the outcome and citing "personal communication" with the data providers. More speculatively, program chairs and editors may wish to contact data providers themselves with specific questions about the dataset.

ACM SIGCOMM/USENIX Internet Measurement Conference, Oct. 2005.

[13] R. Pang, M. Allman, V. Paxson, and J. Lee. The Devil and Packet Trace Anonymization. *ACM Computer Communication Review*, 36(1), Jan. 2006.

[14] R. Pang and V. Paxson. A High-Level Programming Environment for Packet Trace Anonymization and Transformation. In *ACM SIGCOMM*, Aug. 2003.

[15] V. Paxson. Internet Traffic Archive. http://ita.ee.lbl.gov/.

[16] V. Paxson. Strategies for Sound Internet Measurement. In *ACM SIGCOMM Internet Measurement Conference*, Oct. 2004.

[17] C. Shannon, D. Moore, K. Keys, M. Fomenkov, B. Huffaker, and kc claffy. The Internet Measurement Data Catalog. *ACM Computer Communication Review*, 35(5), Oct. 2005.

[18] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon. Prefix-Preserving IP Address Anonymization: Measurement-Based Security Evaluation and a New Cryptography-Based Scheme. In *Proc. of the 10th IEEE International Conference on Network Protocols*, 2002.