Large-Scale Collection and Sanitization of Network Security Data: Risks and Challenges*
(position paper)

Phillip Porras
SRI International

Vitaly Shmatikov
The University of Texas at Austin

Abstract

Over the last several years, there has been an emerging interest in the development of wide-area data collection and analysis centers to help identify, track, and formulate responses to the ever-growing number of coordinated attacks and malware infections that plague computer networks worldwide. As large-scale network threats continue to evolve in sophistication and extend to widely deployed applications, we expect that interest in collaborative security monitoring infrastructures will continue to grow, because such attacks may not be easily diagnosed from a single point in the network. The intent of this position paper is not to argue the necessity of Internet-scale security data sharing infrastructures, as there is ample research [XN05, YBU03, SY05, VFS06, Spi05] and operational examples[Sym06, DSh06, myN06, YBP04] that already make this case. Instead, we observe that these well-intended activities raise a unique set of risks and challenges.

We outline some of the most salient issues faced by global network security centers, survey proposed defense mechanisms, and pose several research challenges to the computer security community. We hope that this position paper will serve as a stimulus to spur groundbreaking new research in protection and analysis technologies that can facilitate the collaborative sharing of network security data while keeping data contributors safe and secure.

1 Introduction

Computer (in)security has become a global phenomenon. Distributed denial of service attacks, rapidly propagating viruses, self-replicating worms are a bane of computer networks worldwide, and attacks constantly grow in severity and sophistication. Following the popular success of such initiatives as DShield [DSh06] and DeepSight [Sym06], there has been a growing interest in the creation of large-scale analysis centers that collect network security information from a diverse pool of contributors and provide a real-time warning service for Internet threats, as well as a source of real data to drive new research in large-scale collaborative defense.

Availability of rich, comprehensive network security datasets collected from a broad cross-section of intrusion detection systems, firewalls, honeypots, and network sensors has the potential to cause a paradigmatic shift in computer security research. Real-time detection of zero-day attacks; large-scale picture of Internet security trends and inflection points; automatic extraction of signatures for polymorphic malware; blacklisting of attacker-controlled hosts and networks; new research on understanding malicious software, its propagation patterns and attack vectors — the possibilities are almost limitless.

It has been recognized, however, that open access to raw network security data is fraught with peril. A repository of such data becomes a single point of failure and a natural target for attackers, not to mention insider compromise. Moreover, even legitimate access to the data can be abused, and the data

*This material is based upon work supported by the Department of Defense under Contract No. H98230-05-C-1650.

contributed by well-intentioned collaborative partners can be turned against them. For example, security alerts contributed by network sensors can be used to fingerprint these sensors and to map out their locations [BFV05]. Security and audit logs may passively leak information about the contributor's vulnerabilities, as well as the data about topology of protected networks, enabled services and applications, egress filtering policies, and so on.

Protection and sanitization of network security data has received some attention in the past two years [LPS04, LPKS05, SY05, XN05], but the problem is far from being solved. The objective of this position paper is to formulate several crisp research challenges for the computer security community. We believe that these challenges will stimulate the discussion, spur design and implementation of efficient sanitization technologies that balance the utility of network security data for collaborative analysis against the need to protect contributors' privacy and security, and even lead to new paradigms for large-scale sharing of network data, including security alerts, packet traces, and so on.

Our work on formulating these challenges is motivated in part by our involvement in actual design and implementation of Internet-scale centers for privacy-preserving collaborative threat analysis. Our objective, however, is not to promote or champion specific solutions, but to raise awareness of the risks and challenges in this area, and to bring a well-informed perspective of both theoretical and pragmatic issues involved in architecting strong privacy guarantees into collaborative sharing infrastructures for network security data.

We group risks and challenges into three areas of concern, associated with, respectively, network sensors that generate the data, repositories that collect the data and make them available for analysis, and the network infrastructure which delivers the data from sensors to repositories. We pay special attention to a class of threats we refer to as *fingerprinting* attacks on network data, which have proved devastatingly effective in many contexts [BFV05, KBC05]. In a fingerprinting attack, an attacker may search for natural patterns in the data that uniquely identify a particular host (*e.g.*, clock skew [KBC05]). Alternatively, the attacker may actively influence data patterns by triggering rare rules in signature-based intrusion detection systems, employing rare port combinations, or generating certain event sequences or timing patterns that can later be recovered from the repository. (This is known in the literature as the *probe response attack* [LPS04, BFV05].)

Probe response and fingerprinting attacks turn the usual intrusion detection game on its head. In contrast to the standard situation, where the attacker's goal is to evade detection, here the attacker *wants* to be detected so that he can analyze the resulting report for evidence of vulnerabilities and gain better understanding of the defender's security posture. Rigorous formalization of fingerprinting attacks and development of provably secure defense mechanisms against fingerprinting are among the most important challenges identified in this paper.

We believe that techniques and methods developed for sanitizing network security data will find applications well beyond the immediate area of collaborative threat detection and analysis. For example, privacy-preserving transformation and anonymization of Internet packet traces [tcp06, PP03, PAPL06] and routing configuration data [MZX+04] have received a lot of attention in the network research community, and could potentially benefit from techniques developed for security data anonymization and anonymity-preserving data publishing protocols.

In this paper, we use the term *repository* somewhat loosely to denote both open- and restricted-access analysis centers, which collect network security data from contributors and make it available either in raw, or in sanitized form.

2 Sanitization technologies for data contributors

The first line of protection when sharing sensitive Internet security information is the contributor's selection of the elements of local security log data that may be shared with other collaborators. Traditionally, we have observed two main approaches to addressing the contributor's security and privacy concerns. The first approach is to release a bare minimum of content to the data repository (an extreme example

may be to include hashed source and destination IP addresses, source and target ports, and a rounded timestamp of the event). Unfortunately, not only does this approach significantly limit the utility of the collected data for downstream collaborative analysis, but it also fails to provide protection against fingerprinting attacks [BFV05]. The second approach is to collect security data under a strict non-disclosure agreement, with significant liability accepted by the repository operator should the data be released and used to harm the contributor. We argue that the assumption of blind trust between contributors and the data collector neither addresses the underlying privacy concerns, nor is it workable in the context of truly collaborative international grids of security sensors.

Thus far, we have used the term "network security data" rather loosely to refer to the data produced locally by the contributor to capture security-relevant operations within the contributor's network perimeter. These data can represent a diverse range of information, depending on the type of security device that produced it. In our context, this includes, but is not limited to, security logs produced by services such as firewalls, intrusion detection systems, network flow logs, and so on. The raw data produced by these sensors tend to contain fine-grained information about observed communication patterns, as well as policy decisions regarding connectivity and content analysis conclusions. In addition, each network security alert may divulge IP address information, protocol and port usage, event timing, sensor identity, and potentially even information related to payload or header contents.

Traditional objectives of large-scale network defense require high precision and accuracy from collected security data. For example, high-trust security alert repositories such as DShield [DSh06] and DeepSight [Sym06] rely on precise data from a diverse pool of contributors to identify Internet-scale security trends and provide an early warning service.

Paradoxically, if the collected data are to be made publicly available for large-scale sharing and collaborative analysis, then precision and accuracy come into conflict with security. The traditional assumption of intrusion detection — that all collected data are supplied to a trusted system administrator or an automated software program that performs analysis and assessment — is not valid anymore, because in the open-access environment, the attackers may easily gain access to the data and (mis)use it to identify their own attacks and analyze their propagation and effects. At the very least, defeating these attacks requires local sanitization of security data before they leave the organization that collected them.

A variety of techniques have been proposed for sanitizing local data before releasing them to Internet-scale analysis centers. These vary from hashing IP addresses [LPS04] to compressing them into Bloom filters [LPKS05] to "generalizing" elements of local datasets so that each element of the sanitized dataset corresponds to multiple elements of the original dataset [XN05].

All of these techniques are non-cryptographic in nature, and do not require any keys to be shared between contributors. Unfortunately, none of them provide *provable* security, especially in the face of an attacker who has access to auxiliary information. For example, IP address hashing is trivially defeated by pre-computation and dictionary attacks, especially when the set of candidate addresses is small. Similarly, if sensitive attributes are rounded up and generalized, an attacker with specific knowledge of attribute values that might have been present in the original dataset can easily infer valuable information from the presence of a generalized attribute in the sanitized dataset. Techniques based on extracting metadata such as Bloom filters [LPKS05], k-ary sketch data structures [KSZC03], and data cubes [VFS06], do not provide cryptographically strong security guarantees, either.

Research challenge 1: *Develop local sanitization methods for network security data that provide cryptographically strong security guarantees for data owners.*

Sanitization is inherently in conflict with usability, and tends to destroy usefulness of the data for subsequent analysis. For example, if IP addresses are protected by hashing, Bloom filters, or generalization, then testing equality is the only operation that can be performed on the sanitized addresses. Unfortunately, capturing *topological relationships* between addresses is necessary for many types of security analyses, *e.g.*, tracking attack vectors and propagation trends. Topological information is usually lost when addresses are sanitized. Prefix-preserving hashing [XFAM01] is one of the few sanitization techniques that preserves some topological information. To be secure against dictionary attacks, however,

hashing algorithms must be keyed.

Key management is a notoriously difficult problem in a massively distributed setting, such as an Internet-scale analysis center with thousands of contributors. To enable cross-contributor comparison, the keys must be shared across all contributors, which means that security of the scheme is as weak as security of the weakest machine on which keys are stored. It is not clear whether there exist scalable solutions to this problem, although introduction of small, tamper-proof, special-purpose hardware devices to which cryptographic operations could be outsourced locally may solve the problem in some scenarios. (The authors are grateful to Tadayoshi Kohno for pointing this out.)

To enable Internet-scale analysis centers to track propagation and topology of Internet threats, we believe it will be necessary to effectively *virtualize* the contributors' IP address space, so that reported data capture all topological relationships without revealing the actual IP addresses contained in the original data.

Research challenge 2: *Develop techniques for IP address virtualization that preserve topological relationships between IP addresses without revealing the contributors' true addresses.*

Sanitization techniques can and should exploit the difference between the objectives and incentives of attackers and honest contributors. The goal of legitimate Internet-scale collaborative analysis is to discover *global* trends and inflection points, while the goal of the adversary is to pinpoint vulnerabilities within a specific *local* system or network. Therefore, we would like each contributor to release locally collected data *if and only if* other collaborative partners have observed the same or similar events. Ideally, this should be achieved with minimal involvement of a global coordinating authority.

Research challenge 3: *Design efficient distributed protocols and similarity metrics for network security data to ensure that each contributor only reveals the data if a threshold number of other participants are ready to reveal similar data.*

3 Sanitization technologies for data repositories

Network security data stored inside global repositories are arguably *more* vulnerable than those stored at the contributors' local sites. In open-access repositories, the attacker may browse and analyze the data at will, looking for evidence of his own and other attacks and actively discovering vulnerabilities such as obsolete intrusion detection systems, old versions of network services, and so on. Even restricted-access repositories are vulnerable, because they are bound to attract malicious attention and become the single point of failure of the system. Finally, it is difficult to prevent malicious insiders with legitimate access to the data from abusing their access privileges.

For the purposes of protecting network security data within the repository, we may as well assume that the repository is completely controlled by the attacker. Defense methods that are robust even under this assumption will also defeat more restricted attacks.

3.1 Understanding and defeating fingerprinting attacks

Fingerprinting attacks effectively enable the attacker to recognize the "signature" of a particular site within the data related to that site. This attack can be passive (*i.e.*, the attacker simply observes the site's unique natural characteristics, such as clock skew [KBC05]), or active (*i.e.*, the attacker probes the system and actively induces a particular attack signature with the goal of eventually recognizing the victim's response in the dataset reported by multiple sites). More generally, the objective of the fingerprinting attack is to uncover the identity of an object within a sanitized dataset by associating the object's attributes to actions that the adversary has knowledge of or control over.

Effectiveness of fingerprinting and probe response attacks has been shown for TCP traces [KBC05] as well as the data reported by network security sensors [BFV05]. We expect that fingerprinting attacks based on unique event sequences, patterns of intrusion alert production, triggering of rare intrusion

detection rules, and so on will prove devastating for naive data collection and sanitization schemes. To the best of our knowledge, there have been no attempts to rigorously define fingerprinting attacks as a class, nor to design sanitization schemes that are provably secure against fingerprinting.

Research challenge 4: *Rigorously formalize fingerprinting attacks and design sanitization schemes for network security data that are provably secure against fingerprinting.*

3.2 Privacy-preserving data mining and analysis

Restricting access to repositories containing network security data may provide some protection for data contributors *if and only if* (a) the repository itself is trusted (which may or may not be realistic, depending on the deployment scenario), and (b) the repository manager takes active measures to protect the data contained within, while making them available in some form for collaborative analysis.

Privacy-preserving data mining has been a subject of very intensive research, so we limit our attention to a few common approaches, focusing in particular on the data mining and learning tasks that are most relevant in the context of collaborative analysis of Internet threats.

Non-interactive data mining. In non-interactive data mining, the dataset is sanitized and then released to the users, who access it locally in any way they want. Sanitization may involve statistical randomization of the data [AS00, EGS03, CDM+05], which enables users to compute certain statistical properties of the original dataset while preserving privacy of individual data entries. An alternative is provably secure database obfuscation [NS05], which restricts the types of *queries* that users can feasibly evaluate on the sanitized dataset. The latter can be viewed as a form of uncircumventable access control that does *not* rely on tamper-proof enforcement software or hardware. The class of access control policies that can be enforced in this way, however, is relatively small.

There has been very little research to date on rigorously defining which functions must be efficiently computable (and to what degree of precision) on the sanitized datasets in order to enable common forms of collaborative security analysis. For example, it is clear that classifying network traffic, extracting malware signatures and tracking propagation of attacks through the Internet are among the most critical tasks of collaborative analysis. Yet, to the best of our knowledge, there has been no research on adapting privacy-preserving data mining algorithms to support evaluation of functions that are relevant for these tasks.

Research challenge 5: *Design and implement efficient privacy-preserving data mining algorithms that enable traffic classification, signature extraction, and propagation analysis on sanitized data without revealing the values of individual dataset entries.*

Interactive data mining. Privacy-preserving data mining can also take place as an *interactive* protocol, where, instead of releasing a sanitized dataset to users, the repository accepts queries from users and, to ensure that no sensitive information is revealed, either audits queries [KPR03, KMN05], or randomizes its responses [DN03, BDMN05].

For interactive data mining, it is also necessary to investigate which functions other than simple statistical calculations and learning algorithms such ID3 and Perceptron must be supported in order to enable collaborative analysis, detection, and tracking of Internet threats.

Other methods for sanitizing network security data include random alert sampling and/or suppression of rare data attributes. For example, the repository may only reveal records that have some information in common if the total number of such records exceeds some threshold. (The threshold may be randomized to prevent flushing attacks — see section 3.3). Finally, the repository may add synthetic or artificial records to the datasets in order to introduce uncertainty into the attacker's analysis of the data.

None of these method provide cryptographically strong security guarantees. The objective, however, is worthwhile: to introduce uncertainty to the attacker's observations of the data repository and make it

difficult to determine whether the absence of a particular fingerprint is due to the lack of collaborative detection, sampling percentage, selective or threshold-based event filters, or the repository's distribution policy, all of which can be controlled and dynamically adjusted by the repository manager and/or data contributors.

The absence of cryptographically strong security should not discourage attempts to quantitatively measure the degree of protection accorded by various sanitization technologies. The metrics should focus on the attacker workfactor need to stage a successful attack, *e.g.*, the number of probes that must be launched before the response can be recognized in the reported data, number of addresses scanned, number of packets generated, and so on.

Research challenge 6: *Develop quantitative metrics for estimating attacker workfactor for different data sanitization and protection technologies.*

3.3 Preventing data poisoning and enforcing accountability

Protecting sources of data and identities of data contributors (discussed further in section 4) is inherently in conflict with preserving utility of the collected data. A completely anonymous system can be abused in many ways. For example, the attacker may stage a blending attack [SDS02] by submitting a large number of fake records that contain some information in common with some record of interest that may or may not be contained in the dataset. The attacker's hope is that these records will be released together (*e.g.*, because the number of records sharing this information exceeds the threshold) and he will then be able to recognize the target record in the released set. Attackers may also stage a denial of service attack by flooding the repository with spam and fake records, poisoning the dataset and rendering it unusable.

Combining anonymity with accountability is a difficult task. In the context of network security data collection, one possible solution involves a registration phase, during which each contributor is issued an *anonymous credential* or a cryptographic key that enables him to compute a digital group signature on his messages.

Anonymous credentials (*e.g.*, [CL01]), group signatures and group authorization mechanisms have been a subject of very intensive research in cryptography — see bibliographies in [Lip06, Wan06]. The particular flavor of group credentials that is relevant in our context should enable the repository to verify that a given contributor is an authorized member of the group (*i.e.*, he successfully passed through the registration protocol at some point in the past), yet his identity remains anonymous (up to the entire set of group members). Obviously, availability of revocation mechanisms is extremely important, in case one of the contributors is compromised.

Research challenge 7: *Investigate applications of anonymous credential schemes for spam and abuse prevention in large-scale network security data collection schemes.*

4 Technologies for anonymous data delivery

We expect that many of the voluntary contributors to global data analysis centers will be interested in protecting their identities even from the centers themselves. There are several reasons for this: (i) probe response and fingerprinting attacks become more difficult if the source of the data is hidden; (ii) anonymous data are more secure against insider attacks; (iii) even direct compromise of data repositories will not necessarily enable attackers to link data records with their creators.

To support broad participation in data collection efforts, the data delivery infrastructure must provide (perhaps optionally) *anonymous message delivery* mechanisms for transmitting the data from contributors to data repositories. Obviously, standard Internet protocols reveal source IP addresses and thus do not provide much anonymity even against passive eavesdroppers. Therefore, we envision anonymous data delivery technologies that will "piggyback" on existing mix-based anonymity networks.

Mix networks are a practical way to enable *unlinkable* communications on public networks. A *mix*, first proposed by Chaum [Cha81], can be thought of as a server that accepts incoming connections and forwards them in such a way that an eavesdropper cannot easily determine which outgoing connection corresponds to which incoming connection. To protect message sources even when some of the mixes in the network are compromised, messages are typically routed through a mix chain.

Since real-time detection of Internet threats is one of the envisioned applications of global analysis centers, we are especially interested in *low-latency* mix networks such as JAP [BFK00] and Tor [DMS04]. Unfortunately, low-latency networks tend to be extremely vulnerable to *traffic analysis* based on correlating packet stream characteristics and/or message dispatch and arrival times [Tim97, Tim99, LRWW04]. An attacker who controls both the data repository and public network links in the vicinity of the sensor generating the data can easily collect traffic observations required for traffic analysis and completely de-anonymize messages received at the repository.

Research challenge 8: *Research new models and implementations for anonymous data delivery networks that provide low- or mid-latency guarantees as well as resistance to traffic analysis attacks even when connection endpoints are directly observable by the attacker.*

Of course, anonymity must go hand in hand with *accountability*. As described in section 3.3, group authorization and anonymous credential mechanisms may need to be deployed to prevent attackers from dumping garbage data into the network. Another promising direction involves reputation systems, which must be combined with identity protection [DMS03].

Research challenge 9: *Investigate applications of reputation systems for ensuring quality of network security data collected anonymously from a broad pool of contributors.*

5 Conclusions

In recent years, as Internet attacks increased in scale, frequency, and severity, there has been a growing interest in creating global analysis centers that would gather network security data from a wide variety of network sensors, use it for real-time collaborative analysis to detect inflection points and global security trends, identify propagation patterns and attack vectors of malware, and make the data available for network security researchers.

Successful deployment of global analysis centers will require resolving a number of fundamental tradeoffs between increased global network security, privacy of data contributors, potential for malicious abuse of the reported data, liability of data repositories, usefulness of the data for network security research, and practical efficiency. This position paper outlines several specific research challenges in this area. They vary from rigorous formalization of fingerprinting attacks to better understanding of traffic analysis attacks which de-anonymize the data contributed to global analysis centers. We hope that our challenges will become part of the research program for computer scientists working in this area. It is unlikely that global Internet defense will succeed without solving them.

References

[AS00] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. ACM SIGMOD International Conference on Management of Data*, pages 439–450, 2000.

[BDMN05] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *Proc. 24th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 128–138, 2005.

[BFK00] O. Berthold, H. Federrath, and S. Köpsell. Web MIXes: a system for anonymous and unobservable Internet access. In *Proc. Workshop on Design Issues in Anonymity and Unobservability*, pages 115–129, 2000.

[BFV05] J. Bethencourt, J. Franklin, and M. Vernon. Mapping Internet sensors with probe response attacks. In *Proc. 14th USENIX Security Symposium*, pages 193–208, 2005.

[CDM+05] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Towards privacy in public databases. In *Proc. 2nd Theory of Cryptography Conference (TCC)*, pages 363–385, 2005.

[Cha81] D. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–88, 1981.

[CL01] J. Camenisch and A. Lysyanskaya. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In *Proc. EUROCRYPT*, pages 93–118, 2001.

[DMS03] R. Dingledine, N. Mathewson, and P. Syverson. Reputation in P2P anonymity systems. In *Proc. Workshop on Economics of Peer-to-Peer Systems*, 2003.

[DMS04] R. Dingledine, N. Mathewson, and P. Syverson. Tor: the second-generation onion router. In *Proc. 13th USENIX Security Symposium*, pages 303–320, 2004.

[DN03] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proc. 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 202–210, 2003.

[DSh06] DShield. `http://www.dshield.org`, 2006.

[EGS03] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy-preserving data mining. In *Proc. 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 211–222, 2003.

[KBC05] T. Kohno, A. Broido, and K. Claffy. Remote physical device fingerprinting. In *Proc. IEEE Symposium on Security and Privacy*, pages 211–225, 2005.

[KMN05] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing. In *Proc. 24th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 118–127, 2005.

[KPR03] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Auditing Boolean attributes. *J. Comput. Syst. Sci.*, 66(1):244–253, 2003.

[KSZC03] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch-based change detection: methods, evaluation, and applications. In *Proc. 3rd ACM SIGCOMM Conference on Internet Measurement*, pages 235–247, 2003.

[Lip06] H. Lipmaa. Group signature schemes. `http://www.cs.ut.ee/~lipmaa/crypto/link/signature/group.php`, 2006.

[LPKS05] M. Locasto, J. Parekh, A. Keromytis, and S. Stolfo. Towards collaborative security and P2P intrusion detection. In *Proc. IEEE Information Assurance Workshop*, pages 333–339, 2005.

[LPS04] P. Lincoln, P. Porras, and V. Shmatikov. Privacy-preserving sharing and correlation of security alerts. In *Proc. 13th USENIX Security Symposium*, pages 239–254, 2004.

[LRWW04] B. Levine, M. Reiter, C. Wang, and M. Wright. Timing attacks in low-latency mix systems. In *Proc. 8th International Conference on Financial Cryptography*, pages 251–265, 2004.

[myN06] myNetWatchman. http://www.mynetwatchman.com, 2006.

[MZX+04] D. Maltz, J. Zhan, G. Xie, H. Zhang, G. Hjálmtýsson, A. Greenberg, and J. Rexford. Structure preserving anonymization of router configuration data. In *Proc. 4th ACM SIGCOMM Conference on Internet Measurement*, pages 239–244, 2004.

[NS05] A. Narayanan and V. Shmatikov. Obfuscated databases and group privacy. In *Proc. 12th ACM Conference on Computer and Communications Security (CCS)*, pages 102–111, 2005.

[PAPL06] R. Pang, M. Allman, V. Paxson, and J. Lee. The devil and packet trace anonymization. *ACM SIGCOMM Computer Communication Review*, 36(1):29–38, 2006.

[PP03] R. Pang and V. Paxson. A high-level programming environment for packet trace anonymization and transformation. In *Proc. ACM SIGCOMM '03*, pages 339–351, 2003.

[SDS02] A. Serjantov, R. Dingledine, and P. Syverson. From a trickle to flood: active attacks on several mix types. In *Proc. 5th International Workshop on Information Hiding*, pages 36–52, 2002.

[Spi05] L. Spitzner. Know your enemy: Honeynets. http://project.honeynet.org/papers/honeynet, 2005.

[SY05] A. Slagell and W. Yurcik. Sharing computer network logs for security and privacy: a motivation for new methodologies of anonymization. In *Proc. SECOVAL: The Workshop on the Value of Security through Collaboration*, 2005.

[Sym06] Symantec. DeepSight threat management system. http://tms.symantec.com, 2006.

[tcp06] tcpdpriv. Program for eliminating confidential information from traces. http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html, 2006.

[Tim97] B. Timmerman. A security model for dynamic adaptable traffic masking. In *Proc. New Security Paradigms Workshop*, pages 107–116, 1997.

[Tim99] B. Timmerman. Secure adaptive traffic masking. In *Proc. New Security Paradigms Workshop*, pages 13–24, 1999.

[VFS06] A. Valdes, M. Fong, and K. Skinner. Data cube indexing of large-scale Infosec repositories. In *Proc. Australian Computer Emergency Response Team Conference*, May 2006.

[Wan06] G. Wang. Bibliography on group-oriented signatures. http://www.i2r.a-star.edu.sg/icsd/staff/guilin/bible/group-oriented.htm%, 2006.

[XFAM01] J. Xu, J. Fan, M. Ammar, and S. Moon. On the design and performance of prefix-preserving IP traffic trace anonymization. In *Proc. 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 263–266, 2001.

[XN05] D. Xu and P. Ning. Privacy-preserving alert correlation: a concept hierarchy based approach. In *Proc. 21st Annual Computer Security Applications Conference (ACSAC)*, pages 537–546, 2005.

[YBP04] V. Yegneswaran, P. Barford, and D. Plonka. On the design and use of Internet sinks for network abuse monitoring. In *Proc. Recent Advances in Intrusion Detection: 7th International Symposium (RAID)*, pages 146–165, 2004.

[YBU03] V. Yegneswaran, P. Barford, and J. Ullrich. Internet intrusions: global characteristics and prevalence. In *Proc. ACM SIGMETRICS '03*, pages 138–147, 2003.