The Challenges of Effectively Anonymizing Network Data

Scott E. Coull* Fabian Monrose† Michael K. Reiter† Michael Bailey‡

*Johns Hopkins University
Baltimore, MD
coulls@cs.jhu.edu

†University of North Carolina
Chapel Hill, NC
{fabian,reiter}@cs.unc.edu

‡University of Michigan
Ann Arbor, MI
mibailey@umich.edu

1. Introduction

The availability of realistic network data plays a significant role in fostering collaboration and ensuring U.S. technical leadership in network security research. Unfortunately, a host of technical, legal, policy, and privacy issues limit the ability of operators to produce datasets for information security testing. In an effort to help overcome these limitations, several data collection efforts (e.g., `CRAWDAD` [14], `PREDICT` [34]) have been established in the past few years. The key principle used in all of these efforts to assure low-risk, high-value data is that of trace *anonymization*—the process of sanitizing data before release so that potentially sensitive information cannot be extracted.

Recently, however, the utility of these techniques in protecting host identities, user behaviors, network topologies, and security practices within enterprise networks has come under scrutiny. In short, several works have shown than unveiling sensitive data in anonymized network traces may not be as difficult as initially thought. The naïve solution to this problem is to address the specifics of these attacks as they are discovered. However, doing so fails to address the underlying problem in its entirety. While isolated advances in network data anonymization are important, without a holistic approach to the problem they will simply shift the information-encoding burden to other properties of the traces, resulting in future privacy breaches. Given the significant reliance on anonymized network traces for security research, it is clear that a more exhaustive and principled analysis of the trace anonymization problem is in order.

Luckily, the problem of anonymizing publicly released data is not new. Over the past several decades, statisticians and computer scientists have developed approaches to anonymizing various forms of microdata, which are essentially databases of attributes collected about individuals. One prominent example is census data, which collects information about the salary, marital status, and many other potentially sensitive attributes from the population of an area or country. This census microdata, much like network data, is valuable to researchers for tracking trends, and as such the anonymized microdata must provide accurate information about potentially sensitive information. At the same time, it is essential that specifics from the data cannot be linked to individuals. In response, several anonymization methods, privacy definitions, and utility metrics have been developed to ensure that researchers can use the microdata for a wide spectrum of analyses while simultaneously providing principled, concrete guarantees on the privacy of those individuals within the data.

At first glance, it would seem as though the accumulated knowledge of microdata anonymization can be directly applied to network data anonymization since the two scenarios share so much in common, including similar privacy and utility goals. Unfortunately, the inherently complex nature of network data makes direct application of these microdata methods difficult, at best. We can, however, learn from existing microdata anonymization literature and glean significant insight into how to approach the problem of network data anonymization in a principled fashion.

In this extended abstract, we compare and contrast the fields of microdata and network data anonymization to reveal the ways in which existing microdata literature may be applied to the network data anonymization problem. We further lay out several challenges that lie ahead in the development of robust network data anonymization methodologies that are grounded in the insights and lessons learned from microdata anonymization. Specifically, we examine the difficulties of clearly defining the privacy properties of network data due to its complex nature. In addition, we point out the necessity of utility measures in quantifying the extent to which anonymization may alter results obtained from analysis of the data. It is important to note that there are additional challenges that we do not address here, such as the legal and ethical issues with collecting network data. As a whole, we hope that this comparison between the fields of microdata and network data anonymization serves to fo-

cus the attention of the research community on a holistic approach to network data anonymization that enables the type of collaboration necessary to further progress in the areas of network and computer security research.

2. Microdata Anonymization

Roughly speaking, microdata can be thought of as a database with n rows and m columns, where each row in the microdata corresponds to a single entity that contributed its data. In the case of census data, for example, the rows might represent people who responded to the survey. The columns represent the attributes of those entities, such as their height or salary information. The goal of microdata anonymization is to alter the original data such that it is difficult (and quantifiably so) to infer potentially sensitive information about entities within the data while simultaneously ensuring that statistics computed on the data remain valid. As an example, average salary information for a given area should remain unchanged, but it should not be possible to infer a specific person's salary.

Specifically, the attributes of the microdata are divided into three categories: (*i*) identifiers, (*ii*) key attributes (i.e., quasi-identifiers), and (*iii*) sensitive attributes. Identifiers are attributes that trivially identify the row, such as name or social security number. Key attributes can be used to make inferences on the identity of the row from auxiliary sources of information. Though these key attributes do not directly identify the row, unique attribute values can be used to link rows in the anonymized microdata with other databases that do have identifying information. For instance, if a row in the microdata had a unique combination of height, weight, and age key attributes, then the adversary could use these attributes to look up the row's identity in a secondary database that includes the height, weight, age, and name. Finally, sensitive attributes are those that are not available from other data sources, and which the adversary would like to link to specific identities. To achieve the goal of anonymization, the data publisher removes identifiers, and applies one or more anonymization methods to alter the relationship between the key attributes and sensitive attributes to ensure that such inferences are unlikely. The resultant sanitized microdata can then be measured to quantify its level of privacy and utility.

2.1. Anonymization Methods

Several techniques are used by data publishers to anonymize microdata for publication. Truncation methods remove or reorganize records in the microdata to hide the relationship between the key attributes and sensitive attributes. These methods include removing rows, removing attributes, suppression of key attribute values in specific rows, or generalization (i.e., recoding) where several key attributes are combined into a single equivalence class (e.g., $25 \leq age \leq 35$) [39]. Additionally, several methods based on perturbation of the sensitive attributes exist. Some examples of perturbation include swapping the values of sensitive attributes among different rows [15], sampling the data, or adding noise to the values [4, 17].

In addition to the truncation and perturbation-based methods, two methods have have been proposed which do not directly sanitize the microdata, but instead provide the underlying statistics of the data in alternate ways. The first of which, synthetic data generation [35, 28], attempts to model the original data and generate completely new microdata from that statistical model. Since this new data is generated from a model, the resultant microdata has no connection to real individuals and at the same time the specific statistical properties captured by the model are guaranteed to be preserved. The second method stores the data on a secure remote server, where the data user can access it only through a query interface [3, 40, 19]. Thus, the user only gets the answer to specific queries, and the query interface ensures that no queries are answered if they are harmful to privacy.

2.2. Measuring Privacy

Obviously, naïvely applying anonymization methods to the data is not enough to guarantee privacy. In fact, inappropriate application of anonymization methods may provide several avenues of information leakage. For instance, a recent study by Narayanan and Shmatikov [30] showed that an anonymized dataset of movie recommendations released by NetFlix fails to meet the accepted privacy definitions for microdata, which results in re-identification of several users in the data. To prevent such information leakage, it is necessary to concretely measure the privacy of the resultant anonymized data. As the extensive literature in microdata privacy measures indicates, however, developing privacy definitions that encapsulate all areas of information leakage is not as straightforward as one might hope.

A common microdata privacy definition, known as k-anonymity, was proposed by Samarati and Sweeney [39]. The definition quantifies the difficulty of an adversary in determining which row in the microdata belongs to a given identity by requiring that every row must look like at least $k-1$ other rows with respect to their key attributes. In effect, this creates equivalence classes of key attributes where the adversary would have a $1/k$ chance of identifying the correct row using the key attributes. Chawla *et al.* [10] provide a similar notion of anonymity that applies to microdata containing numerical, rather than categorical, data types.

The notion of k-anonymity provides a necessary, but not sufficient, condition for privacy since without it a row can be trivially identified by the uniqueness of its key attributes. Further restrictions are necessary, however, when we want to prevent the inference of sensitive attributes and not just which rows belong to a given identity. It may be possi-

ble, for example, to have an equivalence class that meets the k-anonymity definition, and yet has only one or a small number of distinct sensitive values. Thus, any individual that falls into such a class will have their sensitive attributes revealed. Machanavajjhala *et al.* [26] proposed ℓ-diversity to strengthen the k-anonymity property by requiring that each class have at least ℓ distinct sensitive values. Truta and Vinay [43] concurrently developed p-sensitive k-anonymity to provide the same requirement.

The ℓ-diversity property was further strengthened by Li *et al.* [25] since it may still be possible to leak information (in an information theoretic sense) about the sensitive attributes for an individual if the distribution of sensitive values in that individual's equivalence class is significantly different than those of the population. Essentially, the distribution within the equivalence class gives the adversary a more refined distribution of potential sensitive values for an individual than the adversary would have access to without the anonymized microdata. The t-closeness property [25] requires that the distribution of sensitive values in all equivalence classes be within a distance t of the population distribution across all rows in the microdata. This property ensures that the data publisher has greater control over the amount of information the adversary can gain about sensitive values of the individuals in the equivalence classes, thought small values of t clearly have a deleterious effect on the utility of the data.

While k-anonymity and t-closeness provide controls over the information disclosed by the key attributes and sensitive attributes, respectively, there are still other avenues of information leakage which the adversary can take advantage of. Zhang *et al.* [45] recently showed that it is possible to reverse the anonymization of a dataset if the adversary has knowledge of the anonymization method used (e.g., generalization). The key observation is that anonymization proceeds deterministically from anonymizations with the best utility (e.g., minimal equivalence class sizes) to those with worse utility, and will stop at the first anonymization that meets the privacy definition. Zhang *et al.* suggest the notion of p-safe, p-optimal anonymization, where anonymized microdata produced to meet privacy definition p (e.g., k-anonymity) is considered safe if it has more than one potential original microdata that could have produced it.

An alternative approach to these uncertainty, or indistinguishability, definitions is provided by the notion of differential privacy [16]. Differential privacy is primarily applied to interactive query systems where users interact with the data via a secure query interface. The notion of differential privacy states that the probability of a privacy breach occurring for a person is similar whether or not that person's information is contained in the data. The primary difference between differential privacy and the uncertainty-based definitions is that differential privacy is unable to quantify exactly what sensitive information could be leaked by the data, and instead focuses on the slightly more general guarantee that no additional harm will be done by adding a record.

2.3. Measuring Utility

The primary motivation for publishing anonymized microdata is to provide some utility, such as the ability to calculate statistics on the attributes, to researchers who make use of the data. Clearly, the data would be useless if the privacy definitions above are achieved at the expense of utility. As a result, several utility measures have been developed to provide researchers with metrics that allow them to gauge the confidence they should have in the results gained by analysis of the anonymized data. Most utility measures for microdata focus on *specific* utilities that are meant to be preserved. The obvious problem is that in doing so one can only anticipate a limited set of utilities and therefore can not offer guidance about other uses of the data.

Recently, some global utility measures have been proposed to try and quantify a wide range of utilities in a single metric [44, 20]. These global measures, however, can be difficult to interpret and often times do not strongly predict the available utilities. Specifically, these measures are loosely correlated with the extent to which utility is preserved, but they are unable to communicate to the researcher the exact way in which a particular utility is affected by the anonymization. For instance, Karr *et al.* 's use of the Kullback-Leibler divergence [20] between the anonymized and original data provides a broad notion of the similarity of the two distributions of attribute values, but that value has no direct connection to the changes to specific utilities.

3. Network Data Anonymization

Network data can be viewed in much the same way as microdata; containing n rows each representing a single packet (or summarized network flow) and m columns representing the fields in the packet. Unlike microdata, which generally contains only categorical or numerical data, network data contains a variety of data types that make application of well-known anonymization methods difficult, if not impossible. Some fields in network data, like IP addresses, have a highly complex hierarchical ordering structure that often needs to be preserved after anonymization. Moreover, the relationship among different fields in network data is semantically rich, which means that the values taken by certain fields is dependent on their context with respect to other values within the data – both within the same row and within other rows – and these dependencies must be maintained in order for the data to be semantically meaningful.

The goals of network data anonymization are also superficially similar in nature to those of microdata insofar as they are focused on preventing the disclosure of sensitive information about certain entities present within the

data. However, these goals are far more nebulous in the network data case since this sensitive information cannot be defined as a single field, nor can it be quantified for just a single row. Network data publishers are concerned with the privacy of workstations on the network and their users, which can be associated with multiple rows (e.g., packets) within the data. The sensitive information about these entities is often encoded in complex relationships among multiple fields across several different rows, such as a user's web browsing patterns or computer virus activity. Unfortunately, these goals remain ill-defined even in the most recent work in this area, which necessarily limits the efficacy of the anonymization procedures.

3.1. Anonymization Methods

Currently, the anonymization of network data is performed by applying one of a limited number of techniques, many of which are shared with microdata, to fields in the data chosen by the data publisher and defined in an anonymization policy language [33, 42]. The most widely used of these techniques are truncation, randomization, quantization, and pseudonymization. Truncation and randomization effectively destroy the semantics of the field they are applied to, but are helpful when dealing with fields that are likely to contain highly sensitive data. One example is the payload of packets, which might contain usernames and passwords and are removed from the data as standard practice. Quantization techniques, such as limiting the precision of time stamps, are applied to reduce the information gained about the identity of the workstations from timing attacks [21]. Perhaps the most widely used technique, pseudonymization, replaces IP addresses found in the data with linkable, prefix-preserving pseudonyms [32, 18]. These pseudonyms preserve the hierarchical relationships found in the prefixes of the original addresses. The underlying goal is to enable the analysis of packets generated from hosts, or whole prefixes, without providing the actual IPs.

In an effort to maintain as much of the original data as possible, data publishers apply these methods to as few fields as possible; normally, just the IP addresses, time stamps, and payloads. In fact, fields within the network data are typically anonymized only when they are shown to leak information via published attacks. As a result, the unaltered fields of the data provide significant information that can be used as key attributes to link objects in the data to their real identities. This reactionary anonymization policy has lead to the discovery of several attacks which use the unaltered features of the data to re-identify workstations and their behaviors [37, 5, 6, 12], and identify web pages that the users visit [22, 11].

3.2. Measuring Privacy

Given the reactionary nature of network data anonymization, it comes as no surprise that network data does not have well-defined privacy measures, due in part to the difficulty in clearly defining the privacy properties desired by data publishers. To date, there have been a few attempts to quantify the uncertainty that the adversary has in identifying which pseudonyms or values in the data belong to which real world workstations. For instance, Ribeiro *et al.* [37] derive fingerprints, such as the port numbers used, for each IP address in both the anonymized and original data, and compare the two sets of fingerprints to determine the equivalence classes for each IP address. Those workstations with highly unique fingerprints are considered to be privacy risks for the data publisher. Coull *et al.* [13] also examines the similarity between the anonymized and original data, but examines a broader range of distributions of values found in the data. In doing so, they quantify the privacy of workstations in terms of the number of other workstations with similar value distributions, and also discover those fields in the data that negatively affect privacy. Kounine and Bezzi [23] perform a similar analysis with respect to the privacy of individual values after they have been anonymized rather than workstation privacy as a whole. The problem, of course, is that each of these techniques focus exclusively on workstation or individual field privacy, and yet network data can contain several different types of entities whose privacy is equally important.

3.3. Measuring Utility

The idea of quantifying the utility of network data is only just beginning to gain traction in the network data anonymization community, though the complex nature of the data makes such measures as important, if not more so, as those proposed in microdata anonymization. One such utility measure was recently proposed by Lakkaraju and Slagell [24], and compares the performance of a well-known intrusion detection system on the anonymized and unanonymized data. Another measure was proposed by Burkhart *et al.* [8] and applies anomaly detection methodologies to the anonymized data to quantify the way in which it affects its performance. Both methods closely resemble those of Brickell and Shmatikov [7] that apply machine learning tasks to microdata to determine the degradation in accuracy. In addition, the global utility measure of Woo *et al.* [44] can also be adapted to network data due to its use of standard statistical classification techniques. As with microdata, the use of highly specific measures, such as evaluations under specific anomaly detection methodologies or intrusion detection systems, leads to results that may not be applicable in a more general context. Similarly, global measures still remain difficult to interpret due to their disconnection from concrete utilities, and may in fact be even

more difficult to apply effectively to network data because of its inherently complex and interdependent nature.

4. The Challenges Ahead

Clearly, the problem of anonymizing microdata has received significant attention over the past three decades, and that attention has served to develop several methodologies for providing private and useful microdata to researchers. It is equally clear that network data anonymization is only just beginning to mature as an area of active research, and it can benefit from the substantial body of work generated by microdata anonymization research due to the similarities between the two areas. That said, microdata and network data have a number of non-trivial differences that make direct application of well-known microdata anonymization concepts meaningless. In this section, we outline three broad challenges that lie ahead in the development of effective methods for anonymizing network data.

4.1. What are we protecting?

Before we can begin developing effective anonymization methods for network data, we must first have a clear understanding of exactly what it is we hope to protect. For microdata, this question is easily answered because there is a natural one-to-one correspondence between the rows in the data and the entities being protected. With network data, however, this connection is not as clear. Publishers of network data are interested in protecting the privacy of a number of entities: the network users, the network's security procedures, and the hosts that operate on the network. What makes it difficult to clearly define these entities is the fact that network data is inherently multifaceted. A single record in the network data may actually affect the privacy of many entities of varying types. Moreover, the privacy of those entities is not contingent on only a single row in the data, but on many rows that define their behavior over time. These issues naturally raise questions about how we define each of the entities for which the data publisher is interested in providing privacy.

With that said, for some types of entities the answer to this question is relatively straightforward. When considering the privacy of hosts on the network, for example, these host entities can be defined by assuming that the IP addresses in the network data consistently and uniquely identify a host. Even so, the relatively simple entity definition of hosts is not without its caveats, such as the possibility that multiple hosts may use the same IP. More complex entities, like users or web pages, are more difficult to define without significant auxiliary information (e.g., audit logs). Using those auxiliary data sources to mark the entities associated with each record in the data is one potential avenue for defining the entities of interest in the network data.

4.2. What is sensitive?

Network data has a wide variety of information encoded within it. One need only consider some of its uses in network research to appreciate its scope: e.g., measurements of network traffic characteristics, testing new networking methodologies and tools, and studying emerging phenomena. As we move forward, we must decide which of these pieces of information encoded within the network data should be considered to be sensitive. Again, the relatively simple structure of microdata allows for an elegant definition of sensitive information – any attribute in the data that is likely to be unavailable from an external information source should be labeled as sensitive. The sensitivity of attributes are often easily intuited from knowledge of the underlying data. Unfortunately, such intuitive definitions are simply not applicable to network data.

The very same information-rich properties that make network data so useful to the research community also lead to two significant challenges in defining which pieces of information might be considered sensitive. First, potentially sensitive information encoded within the network data is not restricted to a single column in the data. In fact, the relationships between the columns and across several records often indicate the most sensitive of information. For instance, the distribution of ports used by a host in combination with other fields may indicate that the host is infected by a virus, whereas the distribution of ports alone would not. Similar arguments could be made for whether a user visited an illicit web site, or if the network is using a particular security system. Second, many of the fields present within network data contain a combination of both publicly known and private values. As an example, the distribution of ports used by a host may indicate the services it offers, both publicly and privately within the local network. These scenarios are particularly troublesome since the known values within the column of port numbers can act as key attributes, while the unknown values act as sensitive attributes that the adversary may seek to infer.

Many of the attacks that have been discovered for anonymized network data take advantage of these issues in subverting the privacy of the data. Host profiling attacks [37, 5, 6, 12], for instance, use some of the ports and IP pseudonyms in the data as key attributes to link the hosts to their real identities, and then use the remaining ports to infer the hosts hidden services. Rather than attempt to adapt the static notions of key and sensitive attributes to multifaceted network data, current approaches to measuring privacy of network data (e.g., [37, 13, 23]) instead focus on the uniqueness of a piece of data as an indicator for sensitivity. The underlying assumption is that a sufficiently unique behavior encoded within the data is likely to be unavailable from other data sources.

4.3. Defining Utility for Network Data

An area of considerable interest for both microdata and network data anonymization is the development of metrics that measure the utility of the data after it has been anonymized. These metrics are especially important in the case of network data, where the inherent difficulties of defining sensitivity and entities within the data may lead to essentially useless data. For instance, if we follow the definition that sensitive information in network data is any piece of information that is sufficiently unique, then it is easy to imagine a scenario in which the network data contains only homogenous behavior. This type of homogenous data would be useless to researchers who are interested in investigating anomalous incidents or who want to get an accurate estimation of traffic characteristics. In these types of scenarios, it is imperative that researchers have access to utility metrics with respect to certain properties of the data so that they, and those that review their work, can adequately gauge its appropriateness to the task at hand.

Specific utility measures may provide an adequate short term solution to the problem. In general, a utility measure can be derived by comparing the results of a particular utility on the anonymized data to those of the unanonymized data. The problem, of course, lies in predicting the utilities that will be used. One simple way to address this concern is for the data publisher to publish a set of metrics for standard utilities on the data, and allow researchers to request additional utility measures as necessary. However, this type of arrangement is a significant burden on data publishers and researchers, since data publishers would need to run various analyses on the data and researchers would be unable to perform exploratory analyses in a timely fashion. A slightly different approach might be to adapt the concept of a remote verification server, such as the one proposed by Reiter *et al.* [36], to allow researchers to automatically compare their results from the anonymized data with those from the original data with respect to a specific utility.

5. Conclusion

The uncertainties that currently exist about the efficacy of network data anonymization, from both technical and policy perspectives, leave the research community in a vulnerable position. Even as the field marches forward, it does so with little understanding of the implications of publishing anonymized network data on the privacy of the networks being monitored and the utility to researchers. Without that understanding, data publishers are left to wonder what fields must be anonymized to avoid legal fallout, while researchers question the confidence of results gained from the data. However, the extensive work done on microdata anonymity provides the network research community with several useful insights about how to effectively apply anonymization to published data. At the same time, this prior wisdom cannot be applied directly without first overcoming several challenges, including the development of appropriate privacy and utility definitions for the more complex case of network data. Addressing these challenges is essential, in our view, to ensure the continued, yet responsible, availability of network trace data to support security research.

Acknowledgements

This work was supported in part by the U.S. Department of Homeland Security Science & Technology Directorate under Contract No. FA8750-08-2-0147.

References

[1] M. Allman and V. Paxson. Issues and Etiquette Concerning Use of Shared Measurement Data. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, page To appear, 2007.

[2] R. J. Bayardo and R. Agrawal. Data Privacy through Optimal k-Anonymization. In *Proceedings of the 21st International Conference on Data Engineering*, pages 217–228, 2005.

[3] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical Privacy: The SuLQ Framework. In *Proceedings of the 24th ACM Symposium on Principles of Database Systems*, pages 128–138, 2005.

[4] R. Brand. Microdata Protection through Noise Addition. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 97–116, 2002.

[5] T. Brekne and A. Årnes. Circumventing IP-Address Pseudonymization. In *Proceedings of the* 3^{rd} *IASTED International Conference on Communications and Computer Networks*, pages 43–48, October 2005.

[6] T. Brekne, A. Årnes, and A. Øslebø. Anonymization of IP Traffic Monitoring Data – Attacks on Two Prefix-preserving Anonymization Schemes and Some Proposed Remedies. In *Proceedings of the Workshop on Privacy Enhancing Technologies*, pages 179–196, May 2005.

[7] J. Brickell and V. Shmatikov. The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 70–78, 2008.

[8] M. Burkhart, D. Brauckhoff, and M. May. On the Utility of Anonymized Flow Traces for Anomaly Detection. In *Proceedings of the 19th ITC Specialist Seminar on Network Usage and Traffic*, October 2008.

[9] A. J. Burstein. Conducting Cybersecurity Research Legally and Ethically. In *USENIX Workshop on Large-scale Exploits and Emergent Threats*, April 2008.

[10] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in Public Databases. In *Proceedings of the* 2^{nd} *Annual Theory of Cryptography Conference*, pages 363–385, 2005.

[11] S. Coull, M. Collins, C. Wright, F. Monrose, and M. K. Reiter. On Web Browsing Privacy in Anonymized NetFlows. In *Proceedings of the* 16^{th} *USENIX Security Symposium*, pages 339–352, August 2007.

[12] S. Coull, C. Wright, F. Monrose, M. Collins, and M. K. Reiter. Playing Devil's Advocate: Inferring Sensitive Information from Anonymized Network Traces. In *Proceedings of*

the 14^{th} *Annual Network and Distributed System Security Symposium*, pages 35–47, February 2007.

[13] S. E. Coull, C. V. Wright, A. D. Keromytis, F. Monrose, and M. K. Reiter. Taming the Devil: Techniques for Evaluating Anonymized Network Data. In *Proceedings of the* 15^{th} *Network and Distributed Systems Security Symposium*, pages 125–135, 2008.

[14] CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. `http://crawdad.cs.dartmouth.edu`.

[15] T. Dalenius and S. P. Reiss. Data-swapping: A Technique for Disclosure Limitation. *Journal of Statistical Planning and Inference*, 6:73–85, 1982.

[16] C. Dwork. Differential Privacy. In *Proceedings of the* 33^{rd} *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 1–12, 2006.

[17] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our Data, Ourselves: Privacy via Distributed Noise Generation. In *Proceedings of Advances in Cryptology–EUROCRYPT*, pages 486–503, 2006.

[18] J. Fan, J. Xu, M. Ammar, and S. Moon. Prefix-preserving IP Address Anonymization: Measurement-based Security Evaluation and a New Cryptography-based Scheme. *Computer Networks*, 46(2):263–272, October 2004.

[19] S. Gomatam, A. F. Karr, J. P. Reiter, and A. P. Sanil. Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers. *Statistical Science*, 20(2):163, 2005.

[20] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *The American Statistician*, 60(3):224–232, 2006.

[21] T. Kohno, A. Broido, and K. Claffy. Remote Physical Device Fingerprinting. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 93–108, May 2005.

[22] D. Koukis, S. Antonatos, and K. Anagnostakis. On the Privacy Risks of Publishing Anonymized IP Network Traces. In *Proceedings of Communications and Multimedia Security*, pages 22–32, October 2006.

[23] A. Kounine and M. Bezzi. Assessing Disclosure Risk in Anonymized Datasets. In *Proceedings of FloCon*, 2008.

[24] K. Lakkaraju and A. Slagell. Evaluating the Utility of Anonymized Network Traces for Intrusion Detection. In *Proceedings of the 4th Annual Conference on Security and Privacy in Communication Networks*, September 2008.

[25] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proceedings of the* 23^{rd} *IEEE International Conference on Data Engineering*, pages 106–115, April 2007.

[26] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-Diversity: Privacy Beyond k-Anonymity. In *Proceedings of the* 22^{nd} *IEEE International Conference on Data Engineering*, pages 24–35, April 2006.

[27] J. M. Mateo-Sanz, J. Domingo-Ferrer, and F. Sebé. Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata. *Data Mining and Knowledge Discovery*, 11(2):181–193, 2005.

[28] J. P. Mateo-Sanz, A. Martinez-Balleste, and J. Domingo-Ferrer. Fast Generation of Accurate Synthetic Microdata. In *Proceedings of the International Workshop on Privacy in Statistical Databases*, pages 298–306, 2004.

[29] J. Mirkovic. Privacy-Safe Network Trace Sharing via Secure Queries. In *Proceedings of the 1st ACM Workshop on Network Data Anonymization*, October 2008.

[30] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[31] P. Ohm, D. Sicker, and D. Grunwald. Legal Issues Surrounding Monitoring During Network Research (Invited Paper). In *ACM SIGCOMM/USENIX Internet Measurement Conference*, San Deigo, October 2007.

[32] R. Pang, M. Allman, V. Paxson, and J. Lee. The Devil and Packet Trace Anonymization. *ACM Computer Communication Review*, 36(1):29–38, January 2006.

[33] R. Pang and V. Paxson. A High-Level Environment for Packet Trace Anonymization and Transformation. In *Proceedings of the ACM Special Interest Group in Communications (SIGCOM) Conference*, pages 339–351, August 2003.

[34] PREDICT: Protected Repository for the Defense of Infrastructure Against Cyber Threats. `http://www.predict.org`.

[35] T. E. Raghunathan, J. P. Reiter, and D. B. Rubin. Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, 19:1–16, 2003.

[36] J. P. Reiter, A. Oganian, and A. F. Karr. Verification Servers: Enabling Analysts to Assess the Quality of Inferences from Public Use Data. *Computational Statistics and Data Analysis*, forthcoming.

[37] B. Ribeiro, W. Chen, G. Miklau, and D. Towsley. Analyzing Privacy in Enterprise Packet Trace Anonymization. In *Proceedings of the* 15^{th} *Network and Distributed Systems Security Symposium, to appear*, 2008.

[38] D. B. Rubin. Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-Imputed Microdata. *Journal of Official Statistics*, 9:461–468, 1993.

[39] P. Samarati and L. Sweeney. Protecting Privacy When Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[40] B. Schouten and M. Cigrang. Remote Access Systems for Statistical Analysis of Microdata. *Statistics and Computing*, 13(4):381–389, 2003.

[41] C. Shannon, D. Moore, and K. Keys. The Internet Measurement Data Catalog. *ACM SIGCOMM Computer Communications Review*, 35(5):97–100, Oct. 2005. See `http://imdc.datcat.org/browse`.

[42] A. Slagell, K. Lakkaraju, and K. Luo. FLAIM: A Multilevel Anonymization Framework for Computer and Network Logs. In *Proceedings of the* 20^{th} *USENIX Large Installation System Administration Conference*, pages 63–77, 2006.

[43] T. M. Truta and B. Vinay. Privacy Protection: p-Sensitive k-Anonymity Property. In *Proceedings of the 2nd International Workshop on Privacy Data Management*, page 94, 2006.

[44] M. Woo, J. P. Reiter, A. Oganian, and A. F. Karr. Global Measures of Data Utility in Microdata Masked for Disclosure Limitation. *Journal of Privacy and Confidentiality*, forthcoming.

[45] L. Zhang, S. Jajodia, and A. Brodsky. Information Disclosure Under Realistic Assumptions: Privacy versus Optimality. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*, pages 573–583, 2007.