Are We Missing Labels? A Study of the Availability of Ground-Truth in Network Security Research

Sebastian Abt
da/sec – Biometrics and Internet Security Research Group
Hochschule Darmstadt, Darmstadt, Germany
sebastian.abt@h-da.de

Harald Baier
da/sec – Biometrics and Internet Security Research Group
Hochschule Darmstadt, Darmstadt, Germany
harald.baier@h-da.de

***Abstract*—Network security is a long-lasting field of research constantly encountering new challenges. Inherently, research in this field is highly data-driven. Specifically, many approaches employ a supervised machine learning approach requiring labelled input data. While different publicly available data sets exist, labelling information is sparse. In order to understand how our community deals with this lack of labels, we perform a systematic study of network security research accepted at top IT security conferences in 2009 – 2013. Our analysis reveals that 70% of the papers reviewed rely on manually compiled data sets. Furthermore, only 10% of the studied papers release the data sets after compilation. This manifests that our community is facing a missing labelled data problem. In order to be able to address this problem, we give a definition and discuss crucial characteristics of the problem. Furthermore, we reflect and discuss roads towards overcoming this problem by establishing ground-truth and fostering data sharing.**

I. Introduction

Network security is a highly active field of research. Especially, development of effective and efficient network anomaly detection systems is constantly challenging academia and industry. For anomaly detection, the majority of contemporary research (e.g. [12], [25], [47], [57], [135]) follows a supervised machine learning or statistical approach and, consequently, requires *a-priori labelled* input data for training and evaluation. Unfortunately, such data is rare. Most public data repositories offering network traffic samples provide only anonymised data and do not contain labels. Hence, data sets available in these repositories can not be linked to other databases (e.g. blacklists) in order to derive labels. As a consequence, researchers often individually collect data sets in environments where expert knowledge of network traffic is available and, because of that, labels can be assigned automatically or semi-automatically. These environments include, but are not limited to working group, campus or industry networks. Data won in such environments typically contains sensitive information, i.e. personally identifiable information, such as IP addresses or login credentials and, unfortunately, cannot be widely shared.

In order to understand how our community handles this limitation in available data sets and which data sets our community utilises, we review in this paper 106 network security papers accepted at top IT security conferences in the years 2009 – 2013 according to the data sets used for training and evaluation. Additionally, we analyse and discuss existing publicly accessible data repositories and the data sets provided therein. Based on these analyses, we identify two main weaknesses in our community:

1) Researchers in our community tend to manually compile data sets for system design. External data sets are typically included for later evaluation. However, both data sets are typically not publicly released. We speculate about reasons for this *data sharing shortcoming* in Sect. III-D1.
2) The absence of a-priori labelled data sets combined with the previously mentioned data sharing shortcoming leads to a lack of ground-truth data. As argued in Sect. III-D2, this *missing labelled data problem* - as we are tempted to call it - affects repeatability and comparability of research.

Furthermore, we reflect the results of our analysis in context of related work in our community. Specifically, we discuss work in three complementary directions that our community may follow in order to foster data sharing and increase repeatability and comparability of research.

Our work is motivated by own experiences when performing data-driven network security experiments. Furthermore, we recognised that absence of adequate data sets and difficulties in compiling such data sets is often incidentally remarked in papers. In doing this analysis, we hope that our paper contributes to and stimulates an ongoing active discussion on availability and quality of labelled data in our community by quantifying and defining the problem we are facing. To the best of our knowledge, we are the first to perform a systematic and comparative analysis of data sets utilised in contemporary network security research.

The remainder of this paper is structured as follows: Section II gives an overview of existing public data repositories and discusses general issues and limitations of these repositories. Section III presents the results of our analysis of recent research and concludes that our community is facing a missing labelled data problem. In Section IV, we discuss and reflect possibilities to overcome this problem. Section V gives an overview and discussion of related work. Section VI summarises and concludes.

II. Analysis of Public Data Repositories

Currently, different public data repositories comprising varying network traffic traces exist. A listing of these repositories is given in Table I. Probably the most notable data repositories (simply by size) are CAIDA and PREDICT. The Cooperative Association for Internet Data Analysis (CAIDA) continuously performs Internet traffic measurements at varying scales and with varying granularity. The CAIDA repository contains public and semi-public data sets that can be freely downloaded or requested by researchers. The CAIDA repository contains, amongst others, Internet traffic statistics, as well as Internet topology data sets, backscatter traces and real-world Internet traffic captures. In the latter data sets, IP addresses are typically anonymised using Crypto-PAn [149] and packet payload is removed. Unfortunately, traffic captures provided by CAIDA are unlabelled.

The Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) is an effort to provide a distributed data repository together with centrally managed access processes. PREDICT is funded by the US Department of Homeland Security (DHS) and data sets are contributed by different data providers, one of which is CAIDA. PREDICT offers three classes of data: unrestricted data, quasi-restricted data and restricted data. Unrestricted data is available to every PREDICT user that completed the formal sign-up process. Quasi-restricted and restricted data are only accessible after completing sign-up and after request is granted by the data provider or the PREDICT application review board, respectively. Data sets indexed by PREDICT contain, amongst others, BGP routing data, DNS data, darknet and sinkhole data, Netflow data, topology data as well as packet header captures and synthetically generated data. Most data sets provided via PREDICT do not contain packet payload and many data sets contain anonymised IP addresses. PREDICT indexes 430 data sets in total, from which 30 data sets belong to class unrestricted, 284 to class quasi-restricted and 116 to class restricted. The only unrestricted packet level data source have been collected in 2003, contain anonymised IP addresses and do not contain payload. Furthermore, the traces do not contain explicit per-record class labels. However, an implicit labelling via data categories might be possible.

The other data repositories are smaller in size and mostly resulted from specific research questions. Thus, these repositories serve as good examples of how data sets can be published together with research papers. The Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD) stores data sources containing wireless network data. The DARPA Intrusion Detection Evaluation Datasets (DARPA IDEVAL) [90], [91] have been collected in 1999 and 1998 and are well known and heavily criticised [99] in our community. Despite all criticism on the data sets, both data sets are outdated and do not reflect state-of-the-art attacks seen in contemporary networks. Hence, using these data sets is not recommended for contemporary research. The Internet Traffic Archive (ITA) of the Network Research Group (NRG) at Lawrence Berkley National Laboratory (LBNL) contains anonymised traffic captures and derivatives thereof as well as tools developed for trace recording and anonymisation. The latest update contributed to the repository was in April 2008. The Monitoring and Measurement database (MOME) is a repository of tools and data of different data providers, comparable to PREDICT. The Simpleweb Traffic Traces Data Repository indexes data sets created by the Design and Analysis of Communication Systems (DACS) group of the University of Twente. The repository contains anonymised packet header traces, Netflow records, a Dropbox traffic data set as well as a labelled data set for intrusion detection. The data sets listed there seem to be single-effort data sets related to a particular study performed by the group and, hence, unfortunately do not provide continuous captures. The labelled data set has been collected using an active honeypot [129] in 2008. The UMass Trace Repository of the Laboratory for Advanced System Software of University of Massachusetts Amherst contains different network related data sets which are typically anonymised. Finally, the Waikato Internet Traffic Storage (WITS) project offers packet traces which typically have IP addresses anonymised using Crypto-PAn [149] and payload being removed.

As the above discussion of the data repositories listed in Table I shows, nearly all data sources found in these repositories show at least one of the following three characteristics that impact the sources' utility for network security research:

1) Data sets are anonymised, i.e. sequences of data are removed or modified in order to eliminate personally identifiable information (PII).
2) Data sets are static, i.e. they are compiled for a fixed period of time and may be outdated rather soon.
3) Data sets are unlabelled, i.e. records contained within the data sets are not attributed according to a-priori expert knowledge.

In the following subsections, we briefly argue why these characteristics impact the utility of data sources for network security research.

A. *Anonymisation*

Anonymisation approaches are comprehensively discussed in literature (e.g. [31], [77], [104], [148], [149]). Typically, anonymisation is applied to captures of real-world network traffic in order to remove PII from the traces. This is a necessary pre-condition for collection and publication of data in most countries and typically set by current law. Common anonymisation strategies include modification of IP addresses as well as removal of payload information. While such anonymised data may be valuable for specific measurements and statistics, it is typically of less utility to the network security research community. In fact, if anonymised data sets do not provide a-priori labels they typically render themselves useless for network anomaly detection. Specifically, modification of IP addresses in different data sources leads to data that cannot be linked across data sources in order to assign labels. Removal of payload leads to application layer attacks not being detectable.

B. *Timeliness*

Data sets collected at one specific point in time will be aged in later months or years as the attack landscapes constantly evolve. For instance, the DARPA IDEVAL data sets [90], [96], two data sets heavily utilised from 1999 to 2005, do not contain any command and control (CnC) traffic typically found in today's malware communication. Hence, these data sets are of no utility when it comes to design and evaluation of, for instance, botnet detection systems – solutions countering a highly recognised contemporary threat. Moreover, even the statistical value of a one-time data set may be highly limited, especially when the data set is heavily anonymised, as traffic patterns constantly change [45].

Data repository	URL
CAIDA	http://www.caida.org/data/overview/
CRAWDAD	http://crawdad.cs.dartmouth.edu/index.html
DARPA IDEVAL	http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/
Internet Traffic Archive	http://ita.ee.lbl.gov/
MAWI Working Group Traffic Archive	http://mawi.wide.ad.jp/mawi/
MOME	http://www.ist-mome.org/database/index.html
PREDICT	https://www.predict.org/
Simpleweb Traffic Traces Data Repository	http://www.simpleweb.org/wiki/Traces
UMass Trace Repository	http://traces.cs.umass.edu/index.php/Network/Network
WITS	http://wand.net.nz/wits/

TABLE I. DATA REPOSITORIES OFFERING NETWORK TRAFFIC TRACES.

C. Missing Labels

The absence of labels in data sets requires researchers to manually analyse and attribute data according to phenomena they try to model and detect, if a supervised approach is chosen. This has two fundamental consequences for research: *1)* Depending on the a-priori knowledge available in different research groups, outcome of manual labelling may differ among groups, even when working towards approaches having the same goal. As a consequence, ground-truth available to develop and evaluate different approaches may vary and, consequently, results are not directly comparable. *2)* If data sets are not only missing labels, but also are heavily anonymised, then a-posteriori assignment of labels is very difficult and in most cases impossible. Especially the latter phenomenon effectively diminishes a data sets utility for network security research.

As a consequence, we assume that the data sets available and described above suffering from the characteristics detailed above are currently not heavily used for system design and evaluation. Indeed, this assumption is reflected by our analysis of contemporary network security research given in the next section.

III. Analysis of Contemporary Research

This section presents the results of our review of contemporary network security research accepted at top IT security conferences in the time period 2009 – 2013. In the remainder of this section we describe our paper and conference selection strategy (Sect. III-A) as well as our analysis criteria (Sect. III-B), discuss results of our analysis (Sect. III-C), draw conclusions from these results (Sect. III-D) and reflect these conclusions with responses we received from authors (Sect. III-E).

A. Paper and Conference Selection

In order to understand how our community performs data-driven analysis, design and evaluation, we analyse work accepted at highly rated IT security conferences in the years 2009 – 2013. We limit our analysis to those conferences focussing explicitly on network security or having at least a dedicated network security track. We orient our selection on conference rankings provided by Gu et al.[1], Microsoft Academic Research[2], the conference impact factor proposed by Zhou[3] as well as Google Scholar[4]. As a result, we analyse papers accepted at:

- ACM Conference on Computer and Communications Security (CCS)
- IEEE Symposium on Security and Privacy (S&P)
- International Symposium on Research in Attacks, Intrusions and Defenses (RAID)
- ISOC Network and Distributed System Security Symposium (NDSS)

We would like to underline that we neither aim at giving any particular ranking of the above listed conferences nor that we aim at discriminating any particular other conference not listed above. Specifically, we are aware of other high-quality conferences (e.g. USENIX Security, ESORICS), but at some point we had to make a cut-off in order for the analysis to be feasible. We believe that the conferences listed above constitute a representative sample of top IT security conferences applying the highest standards in peer-review and quality assurance and, thus, serve well for an analysis of contemporary network security research. Furthermore, we limit our analysis to conference papers and do not review journal articles, as we have the impression that in our community new results are preferably published via conferences and that journal articles are typically extended versions of results already published in one or more conference papers. Therefore, we believe that the bias possibly introduced to our analysis by selecting only conference papers is negligible.

In total, 793 papers have been accepted at these conferences in the time period under investigation. From these papers, we select a subset for review and analysis according to the following two criteria:

1) Papers have to focus on network security. More specifically, we do not analyse papers that focus on host-based or software security (e.g. return oriented programming, code analysis, etc.) and cryptography.
2) Papers have to utilise network traffic traces for learning or evaluation. Any paper not relying on captures of network traffic is disregarded.

As a result of this filtering, we analyse 106 papers, constituting approximately 13% of papers accepted at the conferences CCS, S&P, RAID, and NDSS within the time frame 2009 – 2013.

[1] http://faculty.cs.tamu.edu/guofei/sec_conf_stat.htm

[2] http://academic.research.microsoft.com/RankList?entitytype=3&topdomainid=2&subdomainid=2&last=5

[3] http://icsd.i2r.a-star.edu.sg/staff/jianying/conference-ranking.html

[4] http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computersecuritycryptography

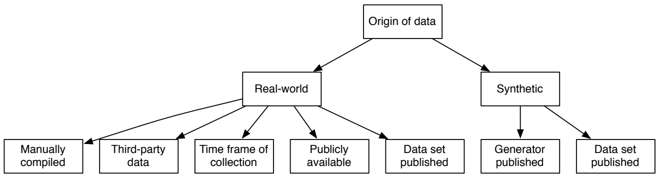

Fig. 1. Illustration and interdependence of criteria used to assess the selected papers.

The papers we analysed are listed in Table II. Again, we believe that this selection of papers and conferences constitutes a representative sample of contemporary work and, consequently, our derived results are sound.

B. Analysis Criteria

As focus of our analysis is on understanding which data sets researchers utilise in order to conduct their work, we perform our analysis according to different data set related criteria. The analysis criteria we apply are structured as illustrated in Fig. 1. As top-most criterion, we analyse whether work is based on real-world or synthetically generated network traffic. Based on the outcome of this analysis, we analyse additional result specific criteria. For real-world data, we assess whether the data set used is manually compiled, provided by third-parties or is publicly available. Furthermore, we assess whether the time span of data collection is provided in the paper and if the data set has been published after work. For synthetically generated data, we assess whether data generators or synthetically generated data sets have been published after work. These criteria are discussed in more detail as follows. Specifically, we provide and discuss a hypothesis reflecting our expectations on the outcome of each criterion. Before, however, we would like to notice that the evaluation criteria are not necessarily mutually exclusive. For instance, research work may incorporate both, synthetically generated data as well as real-world data captures, which both may either be manually compiled or third-party sponsored.

1) Origin of data: We try to understand where data sets used in papers are stemming from. More specifically, as top-most criterion, we analyse whether authors rely on *real-world* (C.1) captures of network traffic or *synthetically generated* (C.2) data. We define real-world captures as any captures that contain network traffic emitted by software that has not specially been crafted to generate traffic for the sake of the traffic itself. Along this line, we define synthetically generated data as any data set containing traffic which has been generated by a computer program that has been developed for the sake of the traffic itself. Specifically, we regard any data set created by capturing packets transmitted in a network or by running malware samples in a sandbox (e.g. [53], [88], [143]) environment as real-world captures. In contrast, we define data sets generated using simulators (e.g. [34], [137]) as synthetically generated.

Hypothesis: A prevalent impression in network security research seems to be that research results achieved using synthetically generated network traffic are less predictive of the utility of a system in real-world environments than results achieved using real-world traffic captures. We assume that this impression basically results from the difficulty in simulating Internet traffic [45]. On the other hand, Ringberg et al. [111] argue that simulation, and hence data synthesis, is a requirement for sound validation of experiments. Yet, to the best of our knowledge, no studies exist that systematically explore capabilities and limitations of synthetically generated data in network security research. As a consequence, we expect most of the work accepted at the venues above to be based on real-world evaluation.

2) Real-world data: (C.1) In our analysis, we aim at assessing where real-world data sets utilised by researchers are stemming from. As we expect the most work to rely on real-world captures, we would like to understand whether researchers leverage publicly available sources, share data amongst each other or industry, or take the expenses of manually compiling data sets. Hence, we assess papers relying on real-world data sets based on the following criteria:

a) Manually compiled: (C.1a) We define manually compiled data sets as any data set that is collected by researchers in their own premises or third-party premises. Furthermore, we define any publicly available or third-party data set lacking class labels as manually compiled, if and only if researchers manually generate class labels in order to annotate the sponsored data.

Hypothesis: Manually compiling and, especially, labelling data sets is a labour intensive process. Thus, we would expect researchers to rely on third-party or publicly available data sets wherever possible in order to be able to focus on the actual problem at hand instead of data collection. On the other hand, manually compiling data sets allows researchers to assure quality of the input data they use for system design and evaluation. Hence, outcome of research may be more predictive when manually compiled data sets are used.

b) Third-party: (C.1b) Third-party data sets are defined as any real-world data set that has not been manually compiled by the researchers itself, but has been provided by any third-party. For instance, data sets provided by network operators or other researchers are regarded as third-party.

Hypothesis: We expect research to heavily depend on especially industry-sponsored third-party data in order to evaluate own work. By evaluating own work using industry-sponsored third-party data, researchers satisfy the prevalent community belief that real-world data is essential to demonstrate real-world utility and, thus, validate contribution and impact.

c) Publicly available: (C.1c) Any data set that can potentially be publicly accessed by researchers is defined as publicly available. Specifically, we do not require the data-sponsoring entity to publish a specific data set without registration or access restriction. However, we require the data-sponsoring entity to publicly announce the availability of the data (e.g. in conference papers or on websites) and, if required, to provide a publicly accessible registration process.

Hypothesis: As mentioned in Section II, different public data repositories (e.g. CAIDA, PREDICT) exist. Unfortunately, these repositories usually offer anonymised and unlabelled data sets. As argued above, utility of such data sets for network security research may be limited as manual post-processing of the data is still required, if possible. On the other hand, using these data sets as starting ground potentially eliminates tedious data collection. Balancing these aspects, we hypothesise that researchers heavily utilise publicly available data as starting point.

d) Time frame of publication: (C.1d) For all real-world data sets we analyse whether the time frame of collection is specified in the papers. By definition, this characteristic can not be evaluated for synthetically generated data.

Hypothesis: As not only the attack and threat landscapes are constantly evolving, but also user behaviour is heavily driven by technical advancement [45], we expect this to be a commonly provided information.

3) Synthetically generated data: (C.2) The use of synthetically generated data sets allows researchers to easily craft different data sets during research. Specifically, data can be tailored to model specific aspects and to evaluate corner-cases in order to estimate the boundaries of a proposed solution. If synthetically generated data sets are used in specific work, we specifically assess whether the process of data synthesis and parameters of the underlying models are discussed.

Hypothesis: Use of synthetically generated data gives great flexibility and many degrees of freedom. On the other hand, as mentioned earlier, we recognise a common mindset in the network security research community that questions results achieved using synthetically generated data sets. More specifically, two arguments against the use of synthetically generated data commonly encountered are:

- *Incompleteness* Synthetic data is usually generated by simulation. Simulation relies on specific models of real-world. As these models hide specific aspects of reality in order to be able to terminate simulation in finite time, synthetic data inherently cannot contain the variety of subtleties found in real-world.
- *Artefacts* Synthetic data is typically generated using simulators and specific models of reality. Consequently, synthetic data shows potential to contain artefacts, such as periodicity or determinism, that may not be found in real-world.

Both characteristics of synthetically generated data may negatively impact research. As machine learning (ML) techniques, which are commonly employed for network anomaly detection, perform well in learning and recognising similarities in data, but perform worse in learning and recognising irregularities [125], ML-based approaches fed with synthetically generated data sets may be tempted to specifically learn artefacts during training. As, by definition, these artefacts are not found in real-world captures, approaches may not perform well in real-world environments. Similarly, when models are built with an incomplete representation of reality, approaches may be confronted with unknown patterns in real-world. Hence, high false alarm rates may be expected and approaches deduced from synthetic data may have little utility in real-world. Consequently, we expect the majority of studies accepted at the conferences under investigation to not rely on synthetically generated data sets without incorporating additional real-world data sets.

4) Publication of data: As mentioned earlier, we regard availability of labelled data sets as prerequisite for comparability and repeatability of experiments. Hence, we specifically analyse if data used for design and evaluation of published research is published as well. To assess this property, we analyse papers with regard to paragraphs that indicate publication of data sets or describe a processes how to access data sets used. Furthermore, we use Google search to find data sets using the title of the papers as search query. We denote *publication of real-world* data sets as criterion (C.1e). If a paper relies on synthetically generated data, we not only analyse the *publication of the synthetic data corpus* (C.2a) itself, but also for *publication of the data generator* (C.2b).

Hypothesis: We suppose that importance of data to conduct research is obvious to any active researcher. We furthermore argue that well-processed and labelled data sets are a product of every data-driven research. Hence, publication of data should be effortless after research work has been accepted for publication. On the other hand, we recognise constraints that prohibit public sharing of data. For instance, non-disclosure agreements (NDAs) may especially prohibit publication of industry-sponsored data sets due to fear of loss of customers or reputation. Additionally, data protection law may prohibit publication of data sets containing sensitive information that are vital for research (e.g. in case research focusses exactly on that part of data). Consequently, we expect researchers to publish data sets after research has been performed. In the case this is not possible, we expect researchers to discuss circumstances prohibiting data sharing.

C. Analysis Results

In this section, we present and discuss the results of our analysis of 106 network security research papers. An overview of the results of our empirical study is given in Table II. Specifically, Table II lists the papers reviewed as well as the cumulative numbers of papers categorised according to the criteria defined in Section III-B per conference and year. The last row shows the sum of papers per category for all categories and over all conferences and years. This summary given in the last row is the basis of our statistics. From our analysis we derive four key observations and some curiosities that we will discuss in the following subsections.

1) Real-world data sets are preferred: As a primary result, our analysis reveals that research in the network security area preferably choses real-world captures of network traffic instead of synthetically generated data. Specifically, 88% (93 of 106 papers) of the investigated papers accepted at the conferences mentioned above used real-world captures for learning or evaluation. In contrast, only 16% (17 of 106 papers) of the papers we analysed leveraged synthetically generated data. Interestingly, 10 of the 17 papers utilising synthetic data relied on synthetic data only, i.e. did not use additional real-world data sets. Hence, only 9% (10 of 106) of all papers under investigation did not use real-world data to conduct their work. This statistic follows our initial hypothesis that we expect research to be based on real-world data. We are convinced that this figure underlines our speculation of the current mindset of our community, that research based on synthetic data does not guarantee utility in real-world.

Conf.	Year	C.1	C.2	C.1a	C.1b	C.1c	C.1d	C.1e	C.2a	C.2b	Papers
CCS	2013	5	2	4	4	2	3	0	0	0	[30], [44], [70], [101], [121], [134]
	2012	3	1	2	1	0	2	0	0	0	[24], [61], [68], [86]
	2011	4	0	4	1	0	2	0	0	0	[60], [69], [95], [138]
	2010	1	1	1	0	0	0	0	0	0	[87], [157]
	2009	3	0	3	0	1	1	1	0	0	[16], [32], [89], [133]
S&P	2013	1	1	1	1	0	0	0	1	0	[63], [85]
	2012	2	0	2	0	0	0	0	0	0	[67], [76]
	2011	3	0	2	2	1	2	0	0	0	[80], [136], [142]
	2010	8	3	7	5	4	4	0	1	0	[35], [39], [49], [75], [82], [93], [107], [122], [144]
	2009	8	1	3	2	2	6	0	1	0	[20], [29], [36], [40], [94], [118], [152]
RAID	2013	6	0	2	3	5	3	1	0	0	[13], [50], [72], [102], [109], [132], [145]
	2012	5	0	4	2	0	1	0	0	0	[18], [38], [55], [97], [155]
	2011	4	0	4	0	0	4	1	0	0	[27], [59], [100], [115], [119]
	2010	8	3	7	5	4	4	1	1	0	[19], [43], [56], [65], [84], [98], [103], [130], [131], [146]
	2009	8	1	3	2	2	6	0	1	0	[28], [42], [46], [52], [54], [83], [92], [110], [150]
NDSS	2013	9	1	8	5	1	5	0	0	0	[21], [33], [78], [81], [120], [139]–[141], [154], [156]
	2012	3	0	2	1	0	2	0	0	0	[7], [79], [151]
	2011	5	1	4	2	1	2	1	0	0	[22], [26], [41], [62], [117]
	2010	4	1	1	3	3	4	0	0	0	[105], [108], [112], [123]
	2009	3	1	1	2	1	1	0	0	1	[51], [64], [66], [128], [147]
	Σ	93	17	65	41	27	52	5	5	1	106

TABLE II. RESULTS, IN NUMBER OF PAPERS, OF THE ANALYSIS WE CONDUCTED ON 106 RESEARCH PAPERS ACCORDING TO CRITERIA DEFINED IN SECT. III-B. THE PAPERS WE ANALYSED PER CONFERENCE AND YEAR ARE LISTED IN THE RIGHTMOST COLUMN.

2) Researchers tend to manually compile data sets: From the work based on real-world captures, 44% (41 of 93) utilised third-party sponsored data sets. In contrast, 70% (65 of 93) of papers relying on real-world captures utilised manually compiled data sets for learning or evaluation, leading to the conclusion that researchers preferably compile data sets themselves. This result is especially interesting as compiling real-world data sets is a time-consuming task. On the other hand, it underlines the difficulty of obtaining real-world data sets from industry. Non-surprisingly, the most commonly referenced sources for manually compiled data sets are sandboxes or sandnets and the university or working-group network.

3) Publicly available data sets are not leveraged: Our analysis shows that publicly available data sources are leveraged by only 29% (27 of 93) of the papers we studied. This is a very interesting result contradicting our initial hypothesis. We assumed that research would heavily make use of publicly available data sets as these data sets enable rapid start of research. From the results of our analysis we conclude that the lack of class labels for publicly available data sets is an even bigger show stopper than expected. This is presumably amplified by anonymisation of public data, leading to missing sequences or sequences that cannot be linked with other data sources. Thus, post-processing publicly available data (e.g. assigning class labels) is apparently more expensive than compiling an entirely new data set.

4) Data sets are not published: One astonishing result of our survey is that the network security research community seems to be particular reserved when it comes to data publication and sharing. Our analysis of papers accepted at top IT security conferences reveals that only 5% (5 of 93) of real-world data sets used to conduct research were released after acceptance of the work. Results are slightly better for synthetically generated data sets. In 35% (6 of 17) of papers utilising synthetic data, the data set itself or the data generator was published after work has been accepted. In total, however, only for 10% (11 of 106) of the papers we analysed data has been published. This strongly contradicts our initial hypothesis that researchers usually publish their data after the corresponding work was accepted for publication. Even more surprisingly, only a negligible fraction of the papers explained why data sets could not be published.

5) Curiosities: In addition to our four observations that our analysis of contemporary network security research reveals, we found two interesting curiosities which we discuss next.

a) Unknown origin: We found that 6 of 106, i.e. almost 6%, of the papers we analysed did not reveal any information on the underlying data set. During our analysis, it was either unclear whether synthetic or real-world data had been used or where real-world data was stemming from, i.e. whether it was manually compiled, third-party sponsored or publicly available. We regard this as a very serious curiosity if we remember that the conferences under investigation are commonly regarded as top venues.

b) Reporting of time frame: Along that line, our analysis reveals that only 56% (52 of 93) of the papers relying on real-world data published the time frame during which data had been acquired. We are puzzled by this result as on one hand, reporting the period of collection is neither expensive in terms of numbers of lines to be devoted, nor in terms of time required to report. On the other hand, specifying the time frame during which data was acquired effectively helps to assess research results at a later point in time. As vulnerabilities, attacks and tools as well as human Internet behaviour are constantly changing, we regard it as a fundamental requirement of any data-driven research to reveal the time span of data acquisition. As this number is rather high, our only explanation for this observation is that publication of time frames is currently not mandated by reviewers and, hence, seems not to be a prevalent requirement for sound experiments in our community.

D. Analysis Conclusions

From the analysis results presented above, we draw two main conclusions:

1) Data sharing shortcoming: Section III-C2 shows that researchers in our community tend to manually compile their data sets for system design. External data sets are typically included for later evaluation. However, data sets are typically not publicly released together with the publication as showed in section III-C4. We are particularly astonished by this result, as any active researcher should understand our community's demand on available data sets. When speculating about this issue, we come down to two possible reasons related to a researcher's mindset. We continue to describe these reasons in a notion of stereotype researchers:

a) The restricted researcher: We regard current law of most jurisdictions as one fundamental driver of scarce data sharing amongst researchers. This holds especially true if data utilised for research is third-party sponsored or based on any real-world captures. In such cases, data privacy law usually restricts the researcher's capability of sharing data. If in such cases data sharing is possible at all, researchers are typically required to additionally process data in order to make it conform to law or contractual requirements (e.g. extensive anonymisation of data). We assume that this additional effort is typically not recognised as added value or, more specifically, that a researcher cannot predict the added value of releasing a data set, usually measured in citation count of a paper, at the time of taking the additional efforts.

b) The competing researcher: As mentioned earlier, manually compiling data sets is a very time-consuming process which may require several months or even several years of active work. For instance, if data has to be captured in real-world networks, technical details have to be discussed with operationally responsible peers, contractual (especially NDA-related) details have to be negotiated, data collectors have to be placed, and data has to be manually post-processed in order to remove noise and assign class labels. And most importantly, collection has to be conducted for a sufficiently sized period of time in order to compile a representative data set. Once finished, however, the resulting data source potentially reflects as substrate of very detailed and focussed research, contributing to the reputation of the data owner. As research is a competition on novel ideas and solutions and our community faces intensive career pressure, researchers having access to data sets with limited public accessibility have a clear competitive advantage.

Any of the two cases lead to a shortcoming of data sharing. In combination with a lack of publicly available labelled data or non-linkable publicly available data, as discussed in Section II, this shortcoming consequently leads to a problem we call the *missing labelled data problem*.

2) Missing labelled data problem: The shortcomings of public data repositories (cf. Section II) and the lack of published data sets, as confirmed by our analysis in Section III-C4, combined with the previously mentioned data sharing shortcoming leads to a lack of publicly available ground-truth data, i.e. labelled data. This absence of ground-truth data negatively affects comparability and repeatability of results and, as such, contradicts fundamental principles of science. Furthermore, as for every researcher digging into a new problem domain the expensive acquisition of data sets becomes a prerequisite for successful work, the absence of ground-truth data specifically hinders rapid research and, thus, scientific progress in our community. Hence, our community faces an intrinsic challenge. In order to be able to frame the dimensions of this challenge, we aim at explicitly defining the missing labelled data problem as follows:

Definition. The missing labelled data problem is *the problem of not having access to labelled data sets of adequate quality and utility with respect to the problem to solve at time of problem solving*.

Our definition reflects the following four dimensions that we derive from our analysis of public data repositories and research work:

a) Access to data: As mentioned earlier, one issue lies in the availability of ground-truth data. As publicly available data sets typically do not contain labelling information and only a small fraction of researchers is able or willing to publish data sets, the availability of ground-truth data is limited. However, publicly available ground-truth data sets are required in order to fulfil scientific principles, such as comparability and repeatability of research. If no common ground for analysis and evaluation is available, research results are in fact not comparable and work can not be repeated. And if research is not repeatable, new approaches can not effectively build upon previously published results. Consequently, research is self-contained within research groups and work of different groups is performed in parallel instead of being sequential.

b) Quality of data: Quality of data is commonly expected to be predictive of an approach's utility in real-world. Hence, quality of data is interchangeable with reality of data. If the data at hand is expected to be a representative sample of reality, i.e. if data is expected to be realistic, results achieved when using the data for evaluation are expected to be achieved in real-world environments as well. As consequence, quality of data serves as a measure of transferability of research results. Hence, quality is an important aspect of ground-truth data.

c) Utility of data: By intuition, we expect that ground-truth data sets can not reflect every single aspect of reality. As such, different ground-truth data sets for different problem domains, or even within one and the same problem domain, are required. In order to be able to assess impact of research conducted on a specific ground-truth data set, it is important to understand the coverage of the data set. Hence, we see a specific requirement of ground-truth data sets in the proof of utility of data for a given problem domain. Proof of utility of data for a specific problem to solve is a prerequisite for any further conclusions.

d) Timeliness: One limiting characteristic of existing publicly available data sets is that data sets are usually static, i.e. data sets are typically captured for a specific period of time and afterwards released. However, reality constantly moves and patterns change. For instance, as attacks and threats constantly evolve, network attack patterns change over time. Botnets, as one example, are a consequence of such evolution. While botnets pose a prevalent threat to our today's infrastructure, botnet CnC traffic was not present 20 years ago. Hence, data sets collected at that time are useless for research of botnet countermeasures. Consequently, one challenge and requirement of ground-truth data is its continuous development. In order to address this, next to a ground-truth data set, methodology of data collection has to be discussed in publications and tools required for data collection have to be published.

E. What the authors say

In order to gain more insight into the problem and to reflect our conclusions, we considered surveying a sample of the authors whose paper we reviewed during our analysis. Specifically, we were interested in surveying why authors deliberately decided to release data sets and why not. However, we were particularly unsure how to survey those authors that not released data. Especially, we expected those answers, if received at all, to refer to sensitivity of information collected in a specific restricted-access context and, particular, to NDAs and legal requirements. Indeed, we got similar answers in prior work when we performed experiments that we wanted to compare. Unfortunately, we would not have been able to assess these answers. Particularly, we assume that we would not have been able to judge whether the answering author is of restricted researcher or of competitive researcher stereotype, which would have given interesting insight into our community. We are currently unsure on how to frame such survey best and leave that part for future work.

Nevertheless, we decided to reach out by email to all authors of those papers that published data sets. We presented the authors a brief summary of key results of our analysis and asked to briefly explain why they chose to publicly release the data set. Actually, two authors responded as follows:

- *Overall, we thought that while many malware repositories are available, there was a real need for (largely) labeled malware datasets. Hopefully, other groups can use it to evaluate their malware classification techniques.*
- *We shared the data because: 1) There aren't enough security datasets available so security research is not very repeatable. We felt that this gap needed to be bridged. 2) The privacy laws in [...] were relaxed so it was easier to share the datasets after anonymization.*

Interestingly, the responses exactly stress that our community, while having various publicly accessible data repositories, is missing *labelled* data sets and that, without such data sets, research is not *repeatable*. Both aspects are in line with our argumentation and analysis results. Hence, even if the sample size is small, this result fully supports our conclusions. Also, we find it particular interesting that one of the respondents explicitly mentions the causality between privacy law and ability to release anonymised data. While an analysis of this causality, and especially implications thereof, is not covered in this paper, we regard it as highly interesting research question for future work.

IV. Overcoming the Problem

From Sect. III we conclude that our community inherently faces the missing labelled data problem. If data sets for research are unavailable, experiments can not be repeated and results or claims can not be verified. Additionally, future work can not be evaluated using the same data set. Hence, different results achieved in work with similar objectives are not comparable. While we recognise this as an inherent property of our domain, it fundamentally contradicts principles of science. Consequently, our community has to develop solutions in order to not loose credibility over time. In the remainder of this section, we present and discuss three complementary approaches as a step towards this direction.

The approaches we discuss here have as well been proposed in different work by others (cf. discussion of related work in Section V). However, from our analysis we conclude that our community has not significantly changed since. We can only speculate why this is the case, but we believe that it is due to the intrinsic difficulty of the problem we describe in this paper.

Additionally, we would like to note that we are aware that the problem we discuss in this paper and the possible approaches towards overcoming it are not necessarily unique to network security research or computer science in general. However, we regard ourselves as experts in network security only and hence hesitate to generalise from our observations in this community. Nevertheless, we are convinced that the missing labelled data problem is especially prevalent in network security research as, from our experience, people are increasingly becoming aware of sensitivity of network data; which is a good development demonstrating some success in our field on one hand, on the other hand making sound data-driven network experiments even harder.

A. Establishing ground-truth

In order to address the lack of ground-truth, research has to focus not only on solving prominent problems, but also on generating common ground-truth data sets. That is, research has to accept missing labelled data sets as a problem of itself. For the conferences under investigation in this paper, compilation of ground-truth data had not been listed in the latest calls for papers. From this observation we conclude that working towards this direction is not heavily recognised in our community and, consequently, less attractive for researchers. From a scientific and, specifically, methodological point of view, however, that kind of research is of great value to the community. Hence, work on ground-truth should become one central topic of interest for relevant IT security conferences in order to stimulate research.

Work towards compilation of ground-truth can comprise the following aspects:

1) Real-world captures: Capturing, post-processing and publishing real-world traffic is a challenging effort. In order to assure that traffic is not biased by behaviour of a specific user group (e.g. behaviour of IT security specialists being connected to the working group network), traffic has usually to be collected on more central points in the network [125]. This essentially requires much communication with operationally responsible peers within the own or within other organisations until formal requirements are met and technical issues can be tackled.

Probably most challenging in that direction is finding an anonymisation tradeoff between legal or contractual requirements and utility of data sets. As shown in Section II, heavily anonymised data sets which are not labelled are available.

Our analysis results in section III-C3, however, show us that such data sets can hardly be leveraged by our community as the data can not easily be linked to other sources and, hence, labels can not be easily derived. Finding appropriate anonymisation techniques that satisfy both, the requirements of our community and those of the data sponsoring party is challenging. Especially, data providers' trust in such techniques may be undermined by publications demonstrating effective data leakage due to attacks on the anonymisation technique [17].

Additionally, such approaches should ideally work towards a continuous data capturing platform in order to be able to continuously track changes of network traffic patterns and to be able to access a representative sample at every point in time. As discussed in Section II, this is a fundamental issue of most data sets that have been crafted for one specific research project. Moreover, being able to provide a constant stream of labelled data would effectively stop overstudy of data sets or publication of irrelevant results on outdated data sets, as seen in case of the DARPA IDEVAL data sets.

By definition, a continuous data capturing system deployed at representative sites in real-world would theoretically address all dimensions of the missing labelled data problem given in Section III-D2 and, consequently, would solve the problem. However, we are aware that it is a long road towards this direction, if possible at all. Nevertheless, research towards this direction is valuable and should especially focus on *methodology*. The more we can learn on *how* to securely design such systems, *how* to technically bridge the gap between anonymisation and utility and *how* to sociologically solve privacy concerns, the faster we can proceed. Literally, at the time of writing, we are convinced that the emphasis is on *'how'* to sensibly capture and provide data, i.e. methodology, and not on the data capture itself.

2) Synthesis software: As mentioned above, utility of synthetically generated data is often challenged in our community. Consequently, our analysis in Section III-C1 shows that only 16% of papers under review utilised synthetically generated data. However, to the best of our knowledge, it is yet unproven that synthetically generated data cannot be used to draw valid conclusions. We are convinced that it is possible to build efficient and effective anomaly detection systems that perform well in real-world. Hence, developing a data synthesis toolchain and assessing utility of data generated using these tools reveals as important future work. In fact, Ringberg et al. [111] and Sonchak et al. [126] argue that synthetically generated data is indeed a requirement for performing repeatable network security experiments. In any case, if in an ideal world a data synthesis tool can be generated that is capable of generating traffic samples of high quality and utility, the missing labelled data problem can effectively be solved for the domain addressed by this tool.

However, the challenge of assessing the quality of synthetically generated data remains. One straightforward approach is to perform statistical tests. If synthetically generated data equals real-world captures with regard to statistical distribution of key aspects of the problem domain, probability is high that (statistical) learners being trained on synthetically generated data sets work well on real-world data sets as well. As an alternative, existing learners published in the problem domain of interest (e.g. classifiers) can be used to assess quality of synthetically generated data. If learners showing high performance on real-world data are capable of detecting events in synthetically generated data and vice versa, we can conclude that the synthetically generated data reflects our current understanding of reality. Obviously, however, the disadvantage of these two approaches is the dependency on real-world data, leading to a recursive problem. The advantage, on the other hand, is that those having access to real-world data would be able to derive synthetically generated data sets that can freely be shared without restrictions, increasing the availability of data in our community.

Nonetheless, we are aware and specifically want to highlight that utilising synthetically generated data sets is just the next best approach compared to real-world data. However, we are convinced that publicly accessible synthetically generated data sets and synthesis toolchains can not only greatly increase comparability and repeatability of network security research, but also foster our understanding of network traffic patterns. In any case, we would like to remind and encourage our community to study capabilities and limitations of synthetically generated data as well as methodology of data synthesis. If, after intensive research, our community comes to the conclusion that we can not establish protocols that support effective and efficient sharing of real-world data, we have to live with the second best approach and accept synthetically generated data as ground-truth.

3) Labelling public data: As our analysis and discussion of data repositories in Section II shows, different publicly available data sources exist. However, class labels describing specific characteristics of the data records are usually missing. This correlates with our observation in Section III-C3 that publicly available data sources are rarely utilised in network security research. Specifically, to the best of our knowledge, the only contemporary data sets providing labelled data records with emphasis on intrusion detection evaluation are provided by Sperotto et al. [129] and Song et al. [127]. However, these data sets have both been collected utilising active honeypots.

One valuable approach in compiling a common ground-truth, thus, would be to focus on exactly filling this gap of missing labels. Hence, researchers should focus on generating and publishing class labels for already existing data sources as this would unleash the full utility of already ongoing data collection efforts. Additionally, doing so would release from the burden associated with manual collection of data and allows for rapid advancement and supports comparability of research.

Furthermore, we would like to note that labelling publicly available data sets also comes without costs when data sets are utilised for research anyway. As shown in Section III-C3, 29% of the papers we reviewed utilised publicly available data sets. If labels would have been released afterwards, these studies would have contributed to solve the missing labelled data problem our community is facing. At that time, we can only speculate about the reasons not to release such labels and come to the conclusion that the lack of labelled data and the contribution the authors could have made to the community is virtually not present to the authors as our community as a whole has not fully internalised the problem. Again, from this we conclude that formalising and discussing the missing labelled data problem, as we do throughout this paper and particularly in Section III-D2, is essential in order to overcome it.

B. Indexing data

In order to solve the missing labelled data problem, fulfilling the requirement of access to data, as described in Sect. III-D2, is essential. One step towards that direction is the establishment of a common data sharing platform which can be used to uniquely index data sets, comparable to the digital object identifier (DOI) system. Such data indexing serves two goods:

1) *Referencing of data.* By assigning unique identifiers to data sets published in a data sharing platform, data sets can uniquely be referenced. Hence, data sets can easily be integrated in literature and it will be trivial to look up specific characteristics of data sets.
2) *Availability of data.* Alongside with indexing, data sets should be reliably stored in the data sharing platform. Thus, data sets will be available and accessible for long time spans, making not only research more comparable and transparent, but also supports other research areas, such as the systematical analysis of evolution in our community.

From the data repositories listed in Section II, especially CAIDA, PREDICT and MOME aim at providing such a platform. Both data repositories, PREDICT and MOME, list various data sets (and, for MOME, even tools) of different data providers while CAIDA basically provides access to own data of affiliated institutes and universities. However, there's a significant overlap between the data repositories. Especially, PREDICT lists a significant proportion of data sets also available in the CAIDA repository. The problem we see here is that neither of these repositories aims at developing a unique and standardised naming scheme. Even worse, neither procedures to access restricted data, nor naming of data classes and data sets is identical in all cases. This variability is confusing, especially to new researchers in our community, and should be removed by defining and agreeing on a community standard in data set naming, attributing and indexing.

We are aware that having a data indexing platform is worthless if we are missing appropriate data to index. Hence, we regard establishment of such a data indexing platform as complementary to establishment of ground-truth, as discussed in Section IV-A. On the other hand, as described previously, we already have a significant amount of (unlabelled) data sets available in our industry that would highly benefit from being uniquely indexed by and accessible via such a platform.

C. Incentivising the researcher

Probably the most important, and even most challenging step towards overcoming the missing labelled data problem is incentivising the researcher. As derived in Section III-D1, we consider two stereotype researchers describing the intrinsic motivation and mindset found in our community. While we have no formal proof for these stereotypes to correctly reflect all individuals within our community, we nevertheless believe that it broadly characterises the majority of researchers. When discussing these stereotypes with colleagues, they invariably were able to agree.

The stereotype *restricted researcher*, as discussed in Section III-D1, may be willing to publish his data sets but may be restricted due to outer constraints, the *competing researcher* in contrary may be able to share, but is unwilling to do so. Hence, the restricted researcher may be intrinsically motivated, while the competing researcher may not. One approach to motivate both researchers even stronger is to incentivise publication of data, i.e. to extrinsically motivate researchers until publication becomes a matter of course. One way of achieving this would be to mandatorily demand release or specification of at least one data set or, if synthetically generated data has been used, parameters required to generate data for validation of research alongside with paper submission for all top-ranked publication platforms. This proceeding effectively enforces comparability and repeatability of research and, hence, essential principles of scientific work. Furthermore, it enables the community to incorporate insight from previous work into new work much stronger, even across different research groups, and, thus, allows us to systematically and sequentially solve problems instead of working in parallel, as discussed in Section III-D2.

Ideally, in case of release of new data sets, data sets should be submitted to data sharing and indexing platforms (cf. Section IV-B) and linked to the paper under submission. We believe that such requirement would initiate reconsideration of paper design, especially of data sets used for evaluation of work. In the long term, this proceeding effectively eliminates the missing labelled data problem we are facing thus far. As the time of writing, however, we are not aware of any conference or journal in network security research mandating researchers to specify one publicly accessible reference for repetition of experiments and comparison of results or otherwise incentivises researchers to publish data. On a step towards this direction, ACM Internet Measurement Conference (IMC) is offering a dedicated award for papers contributing novel data sets. Similarly, USENIX Symposium on Networked Systems Design and Implementation (NSDI) offers a community award for the best paper publishing its data and/or code. We propose to adopt similar awards for key network security venues.

We are aware that forcing researchers to specify a publicly accessible data source in order to repeat research and compare results in conjunction with an accepted paper would severely affect our community. Nevertheless, we are convinced that this proceeding is effective in overcoming the issues arising from the missing labelled data problem ware are facing today. Specifically, we would like to note that we do not insist in mandating researchers to publicly release private/restricted data sets in general. As mentioned throughout this paper, we are well aware and understand constraints that prohibit such data release. However, we propose to mandate researchers to give reference to one publicly accessible data set that, in addition to the private/restricted data set, has been used to evaluate the system proposed in a research paper in order to give fellows the possibility to repeat experiments and compare results. Such mandate simply causes the researcher to take the burden of additional efforts of labelling publicly available data sets, crafting a synthetically generated data set or specifying parameters used by an established data generator to synthesise data sets. We believe that this burden is feasible and especially is outweighed by the long-term benefit to the community. Moreover, we predict that the burden of such mandate monotonically decreases when time elapses, as, after a while, a significant amount of reference data would be publicly accessible by definition. In a significantly lowered version of the above, major conferences could introduce special *data sponsoring papers* sessions. To these sessions, the above requirements should be applied and only papers fulfilling the criteria mentioned above should be considered for acceptance. Hence, such sessions would specifically incentivise papers that focus on contributing data sets and data collection methodology and would explicitly raise broad awareness for the problem, which, as described in Section IV-A, seems currently not to be the case.

On a different location in the continuum, data sharing can practically be incentivised by the data providers. Specifically, we propose data providers to release data sets together with a well-crafted usage codex which especially enforces that labelling information is fed back to the data providers and linked to the data set. When analysing the data repositories mentioned in Section II, we find that data providers typically restrict usage of data (e.g. data sets may not be used to perform research targeting at breaking anonymisation strategies employed) and require researchers to link to a specific paper describing the data collection process. Furthermore, some data providers regularly ask researchers for published work utilising data sets found in their data repositories in order to have that work listed on the data provider's websites (e.g. CAIDA, PREDICT). However, for the repositories we analysed, we did not find any data usage codex requiring researchers to submit information that enrich the publicly accessible data sets. In prior work [11], we propose such codex our community should adhere to. Citing one rule of that codex, we ask that *researchers should publish the results they achieved when utilising a specific set of data. Specifically, the results should be re-submitted to the data providing organisation and should be linked, together with a reference to the research work, online together with the data set [11].* Requiring such codex effectively contributes to establishing ground-truth by labelling public data as proposed in Section IV-A3.

V. Related Work

To the best of our knowledge, no similar comparative study of data sets utilised in network security research has been conducted so far. As data is highly relevant to our community, we therefore believe that the epistemological work we present here is justified. Comparable to our work in spirit is a comparative analysis of malware samples utilised in malware research by Rossow et al. [113]. In this study, 36 academic publications on malware analysis in the time frame 2006 – 2011 have been analysed. Amongst others, the paper identifies shortcomings in transparency, realism and methodology for a significant amount of analysed publications. While this paper analyses malware data sets used, whereas we focus on network traffic captures, the results of [113] are comparable to our results and indicate that our community is facing issues in performing scientific sound experiments.

Recent work supporting our line of argumentation and conclusions in earlier sections is presented in [17], [48], [58], [73], [106], [116], [126], [158]: Ethics and issues of sharing measurement data have been discussed by Allman and Paxson [17]. Specifically, [17] provides considerations for data providers as well as data receivers. However, the usage codex proposed in [17] gives no recommendation for returning supplementary information to the data providers. We especially regard this as an easy and effective way of data sharing. A discussion and classification of different data available to and required by our community is given by Heidemann and Papdopoulos in [58]. Back in 2009, the authors formulated our community's requirement on annotations and metadata as future research topic. Our work underlines this requirement and quantifies the demand and degree of data sharing. Also, our analysis demonstrates that our community has not significantly evolved with regard to data sharing and availability of labelled data within the last 5 years. This is also underlined by work of Sonchack et al. [126] and Ringberg et al. [111]. In [126], the authors discuss the need of labelled data for evaluation of large scale collaborative intrusion detection systems in order to perform repeatable experiments. To bridge this data gap, Sonchack et al. propose a data synthesis approach called parametrised trace scaling, which aims at expanding small real-world traffic samples to generate large and realistic data sets. In [111], the authors argue that synthetically generated data is required in order gain experimental control and to be able to repeatedly evaluate intrusion detection systems. Specifically, the authors argue that synthetic data should be used for training and evaluation and, afterwards, systems should be verified in real-world. However, the use of static data sets for intrusion detection system evaluation - especially if data is synthetically generated - is also challenged in our community. In [99], McHugh intensively criticises methodology and results of the DARPA IDEVAL data sets [90], [91]. In fact, for these data sets we have seen how research has been tuned to the data sets, data sets have been overstudied and systematic deficiencies of data sets can render research results useless. Nevertheless, this approach has heavily stimulated research activity in our community and contributed a lot to the evolution of our community. For future work, we have to incorporate lessons learned from the DARPA approach and especially have to make sure that labelled data sets are continuously compiled, as proposed in Section IV-A. If we achieve to continuously compile realistic data sets, the necessity of relying on old and overstudied data sets vanishes and technical program committee members have a profound argument to withdraw work that is tailored to data sets or based on arbitrarily old data. In [73], Kenneally and Claffy discuss privacy issues in data sharing and propose a privacy-sensitive sharing (PS2) framework. Specifically, the authors emphasise the need of network traffic data for empiric studies and attribute especially industry hesitation to the challenge in balancing advantages and disadvantages of data sharing due to different legal regimes and flawed technology models. With the PS2 framework, [73] provides a viable guideline and demonstrates its utility in the CAIDA use case. Further approaches describing data acquisition and PII-removal methodologies and challenges are described in [48], [106], [116], [158]. One of the most prominent and heavily used IP address anonymisation schemes called Crypto-PAn is described in [149]. This scheme proposed by Xu et al. is prefix-preserving, i.e. if two anonymised IP addresses $j' = f(j)$ and $k' = f(k)$ coincide with the first n bits, then the first n bits of the original IP addresses j and k are equal, too. Crypto-PAn is based on cryptographic hash functions. Indeed, the author demonstrates that the scheme is cryptographically strong.

VI. Summary and Conclusion

Research in the area of network security is heavily data-driven. Especially, in our daily work we observed that our community heavily relies on the availability of a-priori labelled data. As we experienced in own work, such data is hard to find. Indeed, an analysis of data repositories we performed shows that publicly accessible labelled data sets is a rare good. For inherently empirical studies, this observation is quite idiosyncratic. On one hand, data is a necessity in order to perform empirical studies and to be able to publish results. Given the number of empirical studies our community publishes per year, we conclude that such data indeed exists. On the other hand, finding publicly accessible labelled data sets is nearly impossible. From that observation, we hypothesise that our community does not share data sets. In order to be able to accept or reject this hypothesis, we perform a systematic study of 106 network security research papers accepted at CCS, S&P, RAID and NDSS conferences in the years 2009 – 2013. As a result of our analysis, we find that the majority, i.e. 70%, of the papers we review relies on manually compiled data sets. Furthermore, our analysis reveals that only a very small fraction, i.e. 10%, of the papers we analyse release data sets or data generators after compilation. We also notice that a significant amount of work we review, i.e. 44%, tend to utilise external data sets from industry, which are not released either. Interestingly, a surprisingly small fraction of the investigated papers, i.e. 29%, utilise existing public sources containing network traffic. From that analysis, we have to accept our hypothesis and conclude that our community is facing a missing labelled data problem. To the best of our knowledge, we are the first to quantify this severe issue of our community by empirical analysis of contemporary research. In order to be able to frame the problem our community is facing, we derive a definition of the missing labelled data problem and discuss its crucial dimensions. Furthermore, we propose different research areas and challenges towards establishment of ground-truth and propose to establish a common data sharing and indexing platform. Furthermore, we propose how to incentivise researchers to publish and share data.

While we are aware that some of our proposals are rigorous and would severely affect methodology in our community, we deliberately chose to bluntly formulate them. We are convinced that these proposals have the potential to contribute to and stimulate an active discussion in our community. We are aware that similar issues exist in other academic sciences. As computer science is a particular young discipline, we propose to learn from more matured disciplines. Specifically, we are aware that approaches to overcome the missing labelled data problem comparable to those raised by us are currently established in other disciplines of science. For instance, the Nature journal[5] requires authors to share and submit their complementary data. Similar requirements can be found for publishing in Science[6] and the Oxford Journal of Heredity[7]. From that observation, we conclude that an elaborate discussion of this phenomenon in our community is satisfied. Especially, we recognise and want to point out that absence of data, on which empirical analysis is based on, contradicts basic principles of science. Specifically, unavailability of data hinders repeatability of research and comparability of results. While we are well aware and understand constraints that limit general availability of data, as a community we nevertheless have to take care of maintaining a scientific approach in order to not loose credibility over time. Especially, notwithstanding healthy competition and career pressure, we have to make sure that the competing researcher mindset we discussed in this paper does not become prevalent. We are tempted to claim that this is as important to our community as solving the everyday security issues we're facing.

Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under grant number 03FH005PB2 (INSAIN) and has been supported by CASED. We would like to thank Vern Paxson for his inspiration to coin the problem missing *labelled* data problem instead of missing data problem.

References

[1] *30th IEEE Symposium on Security and Privacy (S&P 2009), 17-20 May 2009, Oakland, California, USA*. IEEE Computer Society, 2009.

[2] *Proceedings of the Network and Distributed System Security Symposium, NDSS 2009, San Diego, California, USA, 8th February - 11th February 2009*. The Internet Society, 2009.

[3] *31st IEEE Symposium on Security and Privacy, S&P 2010, 16-19 May 2010, Berleley/Oakland, California, USA*. IEEE Computer Society, 2010.

[4] *Proceedings of the Network and Distributed System Security Symposium, NDSS 2010, San Diego, California, USA, 28th February - 3rd March 2010*. The Internet Society, 2010.

[5] *32nd IEEE Symposium on Security and Privacy, S&P 2011, 22-25 May 2011, Berkeley, California, USA*. IEEE Computer Society, 2011.

[6] *Proceedings of the Network and Distributed System Security Symposium, NDSS 2011, San Diego, California, USA, 6th February - 9th February 2011*. The Internet Society, 2011.

[7] *19th Annual Network and Distributed System Security Symposium, NDSS 2012, San Diego, California, USA, February 5-8, 2012*. The Internet Society, 2012.

[8] *IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA*. IEEE Computer Society, 2012.

[9] *2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013*. IEEE Computer Society, 2013.

[10] *20th Annual Network and Distributed System Security Symposium, NDSS 2013, San Diego, California, USA, February 24-27, 2013*. The Internet Society, 2013.

[11] S. Abt and H. Baier, "A darknet-driven approach to compilation of hostile network traffic samples," in *Proceedings of Sicherheit in vernetzten Systemen: 20. DFN-Workshop*, C. Paulsen, Ed. DFN-Cert Services GmbH, 2013.

[12] S. Abt, C. Dietz, H. Baier, and S. Petrović, "Passive remote source nat detection using behavior statistics derived from netflow," in *Emerging Management Mechanisms for the Future Internet*. Springer, 2013, pp. 148–159.

[13] M. Akiyama, T. Yagi, K. Aoki, T. Hariu, and Y. Kadobayashi, "Active credential leakage for observing web-based attack cycle," in *RAID*, ser. Lecture Notes in Computer Science, S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds., vol. 8145. Springer, 2013, pp. 223–243.

[5] http://www.nature.com/authors/policies/availability.html

[6] http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml

[7] http://www.oxfordjournals.org/our_journals/jhered/for_authors/msprep_submission.html

[14] E. Al-Shaer, S. Jha, and A. D. Keromytis, Eds., *Proceedings of the 2009 ACM Conference on Computer and Communications Security, CCS 2009, Chicago, Illinois, USA, November 9-13, 2009*. ACM, 2009.

[15] E. Al-Shaer, A. D. Keromytis, and V. Shmatikov, Eds., *Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS 2010, Chicago, Illinois, USA, October 4-8, 2010*. ACM, 2010.

[16] M. Q. Ali, H. Khan, A. Sajjad, and S. A. Khayam, "On achieving good operating points on an roc plane using stochastic anomaly score prediction," in *ACM Conference on Computer and Communications Security*, E. Al-Shaer, S. Jha, and A. D. Keromytis, Eds. ACM, 2009, pp. 314–323.

[17] M. Allman and V. Paxson, "Issues and etiquette concerning use of shared measurement data," in *Internet Measurement Conference*, C. Dovrolis and M. Roughan, Eds. ACM, 2007, pp. 135–140.

[18] B. Amann, R. Sommer, A. Sharma, and S. Hall, "A lone wolf no more: Supporting network intrusion detection with real-time intelligence," in *RAID*, ser. Lecture Notes in Computer Science, D. Balzarotti, S. J. Stolfo, and M. Cova, Eds., vol. 7462. Springer, 2012, pp. 314–333.

[19] M. Antonakakis, D. Dagon, X. Luo, R. Perdisci, W. Lee, and J. Bellmor, "A centralized monitoring infrastructure for improving dns security," in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 18–37.

[20] M. Backes, B. Köpf, and A. Rybalchenko, "Automatic discovery and quantification of information leaks," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2009, pp. 141–153.

[21] G. Bai, J. Lei, G. Meng, S. S. Venkatraman, P. Saxena, J. Sun, Y. Liu, and J. S. Dong, "Authscan: Automatic extraction of web authentication protocols from implementations," in *NDSS*. The Internet Society, 2013.

[22] M. Balduzzi, C. T. Gimenez, D. Balzarotti, and E. Kirda, "Automated discovery of parameter pollution vulnerabilities in web applications," in *NDSS*. The Internet Society, 2011.

[23] D. Balzarotti, S. J. Stolfo, and M. Cova, Eds., *Research in Attacks, Intrusions, and Defenses - 15th International Symposium, RAID 2012, Amsterdam, The Netherlands, September 12-14, 2012. Proceedings*, ser. Lecture Notes in Computer Science, vol. 7462. Springer, 2012.

[24] A. Bianchi, Y. Shoshitaishvili, C. Kruegel, and G. Vigna, "Blacksheep: detecting compromised hosts in homogeneous crowds," in *ACM Conference on Computer and Communications Security*, T. Yu, G. Danezis, and V. D. Gligor, Eds. ACM, 2012, pp. 341–352.

[25] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, and C. Kruegel, "Disclosure: detecting botnet command and control servers through large-scale netflow analysis," in *Proceedings of the 28th Annual Computer Security Applications Conference*. ACM, 2012, pp. 129–138.

[26] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "Exposure: Finding malicious domains using passive dns analysis," in *NDSS*. The Internet Society, 2011.

[27] N. Boggs, S. Hiremagalore, A. Stavrou, and S. J. Stolfo, "Cross-domain collaborative anomaly detection: So far yet so close," in *RAID*, ser. Lecture Notes in Computer Science, R. Sommer, D. Balzarotti, and G. Maier, Eds., vol. 6961. Springer, 2011, pp. 142–160.

[28] D. Bolzoni, S. Etalle, and P. H. Hartel, "Panacea: Automating attack classification for anomaly-based network intrusion detection systems," in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 1–20.

[29] K. Borders and A. Prakash, "Quantifying information leaks in outbound web traffic," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2009, pp. 129–140.

[30] K. Borgolte, C. Kruegel, and G. Vigna, "Delta: automatic identification of unknown web-based infection campaigns," in *ACM Conference on Computer and Communications Security*, A.-R. Sadeghi, V. D. Gligor, and M. Yung, Eds. ACM, 2013, pp. 109–120.

[31] 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 *Privacy Enhancing Technologies*. Springer, 2006, pp. 179–196.

[32] J. Caballero, P. Poosankam, C. Kreibich, and D. X. Song, "Dispatcher: enabling active botnet infiltration using automatic protocol reverse-engineering," in *ACM Conference on Computer and Communications Security*, E. Al-Shaer, S. Jha, and A. D. Keromytis, Eds. ACM, 2009, pp. 621–634.

[33] D. Canali and D. Balzarotti, "Behind the scenes of online attacks: an analysis of exploitation behaviors on the web," in *NDSS*. The Internet Society, 2013.

[34] G. Carneiro, "Ns-3: Network simulator 3," in *UTM Lab Meeting April*, vol. 20, 2010.

[35] H. Chan and A. Perrig, "Round-efficient broadcast authentication protocols for fixed topology classes," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 257–272.

[36] S. Chen, Z. Mao, Y.-M. Wang, and M. Zhang, "Pretty-bad-proxy: An overlooked adversary in browsers' https deployments," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2009, pp. 347–359.

[37] Y. Chen, G. Danezis, and V. Shmatikov, Eds., *Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011, Chicago, Illinois, USA, October 17-21, 2011*. ACM, 2011.

[38] J. Chu, Z. Ge, R. Huber, P. Ji, J. Yates, and Y.-C. Yu, "Alert-id: Analyze logs of the network element in real time for intrusion detection," in *RAID*, ser. Lecture Notes in Computer Science, D. Balzarotti, S. J. Stolfo, and M. Cova, Eds., vol. 7462. Springer, 2012, pp. 294–313.

[39] P. M. Comparetti, G. Salvaneschi, E. Kirda, C. Kolbitsch, C. Kruegel, and S. Zanero, "Identifying dormant functionality in malware programs," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 61–76.

[40] P. M. Comparetti, G. Wondracek, C. Krügel, and E. Kirda, "Prospex: Protocol specification extraction," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2009, pp. 110–125.

[41] S. E. Coull, F. Monrose, and M. Bailey, "On measuring the similarity of network hosts: Pitfalls, new metrics, and empirical analyses," in *NDSS*. The Internet Society, 2011.

[42] G. F. Cretu-Ciocarlie, A. Stavrou, M. E. Locasto, and S. J. Stolfo, "Adaptive anomaly detection via self-calibration and dynamic updating," in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 41–60.

[43] J. Cucurull, M. Asplund, and S. Nadjm-Tehrani, "Anomaly detection and mitigation for disaster area networks," in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 339–359.

[44] V. Dave, S. Guha, and Y. Zhang, "Viceroi: catching click-spam in search ad networks," in *ACM Conference on Computer and Communications Security*, A.-R. Sadeghi, V. D. Gligor, and M. Yung, Eds. ACM, 2013, pp. 765–776.

[45] S. Floyd and V. Paxson, "Difficulties in simulating the internet," *IEEE/ACM Trans. Netw.*, vol. 9, no. 4, pp. 392–403, 2001.

[46] J. François, H. J. Abdelnur, R. State, and O. Festor, "Automated behavioral fingerprinting," in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 182–201.

[47] J. François, S. Wang, W. Bronzi, T. Engel *et al.*, "Botcloud: Detecting botnets using mapreduce," in *Information Forensics and Security (WIFS), 2011 IEEE International Workshop on*. IEEE, 2011, pp. 1–6.

[48] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl, "You are what you say: Privacy risks of public mentions," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. New York, NY, USA: ACM, 2006, pp. 565–572.

[49] M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan, "Synthesizing near-optimal malware specifications from suspicious behaviors," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 45–60.

[50] J. Fritz, C. Leita, and M. Polychronakis, "Server-side code injection attacks: A historical perspective," in *RAID*, ser. Lecture Notes in Computer Science, S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds., vol. 8145. Springer, 2013, pp. 41–61.

[51] C. Gates, "Coordinated scan detection," in *NDSS*. The Internet Society, 2009.

[52] F. Giroire, J. Chandrashekar, N. Taft, E. M. Schooler, and D. Papagiannaki, “Exploiting temporal persistence to detect covert botnet channels,” in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 326–345.

[53] C. Gorecki, F. C. Freiling, M. Kührer, and T. Holz, “Trumanbox: improving dynamic malware analysis by emulating the internet,” in *Stabilization, Safety, and Security of Distributed Systems*. Springer, 2011, pp. 208–222.

[54] K. Griffin, S. Schneider, X. Hu, and T. cker Chiueh, “Automatic generation of string signatures for malware detection,” in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 101–120.

[55] D. Hadziosmanovic, L. Simionato, D. Bolzoni, E. Zambon, and S. Etalle, “N-gram against the machine: On the feasibility of the n-gram network analysis for binary protocols,” in *RAID*, ser. Lecture Notes in Computer Science, D. Balzarotti, S. J. Stolfo, and M. Cova, Eds., vol. 7462. Springer, 2012, pp. 354–373.

[56] I. U. Haq, S. Ali, H. Khan, and S. A. Khayam, “What is the impact of p2p traffic on anomaly detection?” in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 1–17.

[57] W. He, G. Hu, and Y. Zhou, “Large-scale ip network behavior anomaly detection and identification using substructure-based approach and multivariate time series mining,” *Telecommunication Systems*, vol. 50, no. 1, pp. 1–13, 2012.

[58] J. Heidemann and C. Papdopoulos, “Uses and challenges for network datasets,” in *Conference For Homeland Security, 2009. CATCH ’09. Cybersecurity Applications Technology*, March 2009, pp. 73–82.

[59] M. Heiderich, T. Frosch, and T. Holz, “Iceshield: Detection and mitigation of malicious websites with a frozen dom,” in *RAID*, ser. Lecture Notes in Computer Science, R. Sommer, D. Balzarotti, and G. Maier, Eds., vol. 6961. Springer, 2011, pp. 281–300.

[60] M. Heiderich, T. Frosch, M. Jensen, and T. Holz, “Crouching tiger - hidden payload: security risks of scalable vectors graphics,” in *ACM Conference on Computer and Communications Security*, Y. Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp. 239–250.

[61] C.-Y. Hong, F. Yu, and Y. Xie, “Populated ip addresses: classification and applications,” in *ACM Conference on Computer and Communications Security*, T. Yu, G. Danezis, and V. D. Gligor, Eds. ACM, 2012, pp. 329–340.

[62] A. Houmansadr and N. Borisov, “Swirl: A scalable watermark to detect correlated network flows,” in *NDSS*. The Internet Society, 2011.

[63] A. Houmansadr, C. Brubaker, and V. Shmatikov, “The parrot is dead: Observing unobservable network communications,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2013, pp. 65–79.

[64] A. Houmansadr, N. Kiyavash, and N. Borisov, “Rainbow: A robust and invisible non-blind watermark for network flows,” in *NDSS*. The Internet Society, 2009.

[65] C.-H. Hsu, C.-Y. Huang, and K.-T. Chen, “Fast-flux bot detection in real time,” in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 464–483.

[66] X. Hu, M. Knysz, and K. G. Shin, “Rb-seeker: Auto-detection of redirection botnets,” in *NDSS*. The Internet Society, 2009.

[67] L. Invernizzi and P. M. Comparetti, “Evilseed: A guided approach to finding malicious web pages,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2012, pp. 428–442.

[68] M. A. Jamshed, J. Lee, S. Moon, I. Yun, D. Kim, S. Lee, Y. Yi, and K. Park, “Kargus: a highly-scalable software-based intrusion detection system,” in *ACM Conference on Computer and Communications Security*, T. Yu, G. Danezis, and V. D. Gligor, Eds. ACM, 2012, pp. 317–328.

[69] J. Jang, D. Brumley, and S. Venkataraman, “Bitshred: feature hashing malware for scalable triage and semantic analysis,” in *ACM Conference on Computer and Communications Security*, Y. Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp. 309–320.

[70] M. Javed and V. Paxson, “Detecting stealthy, distributed ssh brute-forcing,” in *ACM Conference on Computer and Communications Security*, A.-R. Sadeghi, V. D. Gligor, and M. Yung, Eds. ACM, 2013, pp. 85–96.

[71] S. Jha, R. Sommer, and C. Kreibich, Eds., *Recent Advances in Intrusion Detection, 13th International Symposium, RAID 2010, Ottawa, Ontario, Canada, September 15-17, 2010. Proceedings*, ser. Lecture Notes in Computer Science, vol. 6307. Springer, 2010.

[72] N. Jiang, Y. Jin, A. Skudlark, and Z.-L. Zhang, “Understanding sms spam in a large cellular network: Characteristics, strategies and defenses,” in *RAID*, ser. Lecture Notes in Computer Science, S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds., vol. 8145. Springer, 2013, pp. 328–347.

[73] E. Kenneally and K. Claffy, “Dialing privacy and utility: A proposed data-sharing framework to advance internet research,” *Security Privacy, IEEE*, vol. 8, no. 4, pp. 31–39, July 2010.

[74] E. Kirda, S. Jha, and D. Balzarotti, Eds., *Recent Advances in Intrusion Detection, 12th International Symposium, RAID 2009, Saint-Malo, France, September 23-25, 2009. Proceedings*, ser. Lecture Notes in Computer Science, vol. 5758. Springer, 2009.

[75] C. Kolbitsch, T. Holz, C. Kruegel, and E. Kirda, “Inspector gadget: Automated extraction of proprietary gadgets from malware binaries,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 29–44.

[76] C. Kolbitsch, B. Livshits, B. G. Zorn, and C. Seifert, “Rozzle: De-cloaking internet malware,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2012, pp. 443–457.

[77] D. Koukis, S. Antonatos, D. Antoniades, E. P. Markatos, and P. Trimintzios, “A generic anonymization framework for network traffic,” in *Communications, 2006. ICC’06. IEEE International Conference on*, vol. 5. IEEE, 2006, pp. 2302–2309.

[78] T. Lauinger, M. Szydlowski, K. Onarlioglu, G. Wondracek, E. Kirda, and C. Krügel, “Clickonomics: Determining the effect of anti-piracy measures for one-click hosting,” in *NDSS*. The Internet Society, 2013.

[79] S. Lee and J. Kim, “Warningbird: Detecting suspicious urls in twitter stream,” in *NDSS*. The Internet Society, 2012.

[80] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage, “Click trajectories: End-to-end analysis of the spam value chain,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2011, pp. 431–446.

[81] C. Lever, M. Antonakakis, B. Reaves, P. Traynor, and W. Lee, “The core of the matter: Analyzing malicious traffic in cellular carriers,” in *NDSS*. The Internet Society, 2013.

[82] A. B. Lewko, A. Sahai, and B. Waters, “Revocation systems with very small private keys,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 273–285.

[83] P. Li, D. Gao, and M. K. Reiter, “Automatically adapting a trained anomaly detector to software patches,” in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 142–160.

[84] P. Li, L. Liu, D. Gao, and M. K. Reiter, “On challenges in evaluating malware clustering,” in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 238–255.

[85] Z. Li, S. A. Alrwais, Y. Xie, F. Yu, and X. Wang, “Finding the linchpins of the dark web: a study on topologically dedicated hosts on malicious web infrastructures,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2013, pp. 112–126.

[86] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang, “Knowing your enemy: understanding and detecting malicious web advertising,” in *ACM Conference on Computer and Communications Security*, T. Yu, G. Danezis, and V. D. Gligor, Eds. ACM, 2012, pp. 674–686.

[87] T. Limmer and F. Dressler, “Dialog-based payload aggregation for intrusion detection,” in *ACM Conference on Computer and Communications Security*, E. Al-Shaer, A. D. Keromytis, and V. Shmatikov, Eds. ACM, 2010, pp. 708–710.

[88] M. Lindorfer, C. Kolbitsch, and P. M. Comparetti, “Detecting environment-sensitive malware,” in *Recent Advances in Intrusion Detection*. Springer, 2011, pp. 338–357.

[89] Z. Ling, J. Luo, W. Yu, X. Fu, D. Xuan, and W. Jia, “A new cell counter based attack against tor,” in *ACM Conference on Computer and*

Communications Security, E. Al-Shaer, S. Jha, and A. D. Keromytis, Eds. ACM, 2009, pp. 578–589.

[90] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 darpa off-line intrusion detection evaluation," *Computer networks*, vol. 34, no. 4, pp. 579–595, 2000.

[91] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham *et al.*, "Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation," in *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings*, vol. 2. IEEE, 2000, pp. 12–26.

[92] L. Liu, G. Yan, X. Zhang, and S. Chen, "Virusmeter: Preventing your cellphone from spies," in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 244–264.

[93] Y. Liu, P. Ning, and H. Dai, "Authenticating primary users' signals in cognitive radio networks via integrated cryptographic and wireless link signatures," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 286–301.

[94] M. T. Louw and V. N. Venkatakrishnan, "Blueprint: Robust prevention of cross-site scripting attacks for existing browsers," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2009, pp. 331–346.

[95] L. Lu, R. Perdisci, and W. Lee, "Surf: detecting and measuring search poisoning," in *ACM Conference on Computer and Communications Security*, Y. Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp. 467–476.

[96] M. V. Mahoney and P. K. Chan, "An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection," in *Recent Advances in Intrusion Detection*. Springer, 2003, pp. 220–237.

[97] S. Marchal, J. François, R. State, and T. Engel, "Proactive discovery of phishing related domain names," in *RAID*, ser. Lecture Notes in Computer Science, D. Balzarotti, S. J. Stolfo, and M. Cova, Eds., vol. 7462. Springer, 2012, pp. 190–209.

[98] S. Mathew, M. Petropoulos, H. Q. Ngo, and S. J. Upadhyaya, "A data-centric approach to insider attack detection in database systems," in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 382–401.

[99] J. McHugh, "Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 262–294, Nov. 2000. [Online]. Available: http://doi.acm.org/10.1145/382912.382923

[100] S. A. Mehdi, J. Khalid, and S. A. Khayam, "Revisiting traffic anomaly detection using software defined networking," in *RAID*, ser. Lecture Notes in Computer Science, R. Sommer, D. Balzarotti, and G. Maier, Eds., vol. 6961. Springer, 2011, pp. 161–180.

[101] Y. Nadji, M. Antonakakis, R. Perdisci, D. Dagon, and W. Lee, "Beheading hydras: performing effective botnet takedowns," in *ACM Conference on Computer and Communications Security*, A.-R. Sadeghi, V. D. Gligor, and M. Yung, Eds. ACM, 2013, pp. 121–132.

[102] Y. Nadji, M. Antonakakis, R. Perdisci, and W. Lee, "Connected colors: Unveiling the structure of criminal networks," in *RAID*, ser. Lecture Notes in Computer Science, S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds., vol. 8145. Springer, 2013, pp. 390–410.

[103] A. J. Oliner, A. V. Kulkarni, and A. Aiken, "Community epidemic detection using time-correlated anomalies," in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 360–381.

[104] R. Pang, M. Allman, V. Paxson, and J. Lee, "The devil and packet trace anonymization," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 29–38, 2006.

[105] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G. M. Voelker, V. Paxson, N. Weaver, and S. Savage, "Botnet judo: Fighting spam with itself," in *NDSS*. The Internet Society, 2010.

[106] P. Porras and V. Shmatikov, "Large-scale collection and sanitization of network security data: Risks and challenges," in *Proceedings of the 2006 Workshop on New Security Paradigms*, ser. NSPW '06. New York, NY, USA: ACM, 2007, pp. 57–64.

[107] Z. Qian, Z. M. Mao, Y. Xie, and F. Yu, "Investigation of triangular spamming: A stealthy and efficient spamming technique," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 207–222.

[108] ——, "On network-level clusters for spam detection," in *NDSS*. The Internet Society, 2010.

[109] M. Z. Rafique and J. Caballero, "Firma: Malware clustering and network signature generation with mixed network behaviors," in *RAID*, ser. Lecture Notes in Computer Science, S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds., vol. 8145. Springer, 2013, pp. 144–163.

[110] M. Rehák, E. Staab, V. Fusenig, M. Pechoucek, M. Grill, J. Stiborek, K. Bartos, and T. Engel, "Runtime monitoring and dynamic reconfiguration for intrusion detection systems," in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 61–80.

[111] H. Ringberg, M. Roughan, and J. Rexford, "The need for simulation in evaluating anomaly detectors," *Computer Communication Review*, vol. 38, no. 1, pp. 55–59, 2008.

[112] W. K. Robertson, F. Maggi, C. Kruegel, and G. Vigna, "Effective anomaly detection with scarce training data," in *NDSS*. The Internet Society, 2010.

[113] C. Rossow, C. Dietrich, C. Grier, C. Kreibich, V. Paxson, N. Pohlmann, H. Bos, and M. van Steen, "Prudent practices for designing malware experiments: Status quo and outlook," in *Security and Privacy (SP), 2012 IEEE Symposium on*, 2012, pp. 65–79.

[114] A.-R. Sadeghi, V. D. Gligor, and M. Yung, Eds., *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013*. ACM, 2013.

[115] M. B. Salem and S. J. Stolfo, "Modeling user search behavior for masquerade detection," in *RAID*, ser. Lecture Notes in Computer Science, R. Sommer, D. Balzarotti, and G. Maier, Eds., vol. 6961. Springer, 2011, pp. 181–200.

[116] B. Sangster, T. O'Connor, T. Cook, R. Fanelli, E. Dean, W. J. Adams, C. Morrell, and G. Conti, "Toward instrumenting network warfare competitions to generate labeled datasets," in *Proc. of the 2nd Workshop on Cyber Security Experimentation and Test (CSET'09)*, 2009.

[117] M. Schuchard, A. Mohaisen, D. F. Kune, N. Hopper, Y. Kim, and E. Y. Vasserman, "Losing control of the internet: Using the data plane to attack the control plane," in *NDSS*. The Internet Society, 2011.

[118] M. I. Sharif, A. Lanzi, J. T. Giffin, and W. Lee, "Automatic reverse engineering of malware emulators," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2009, pp. 94–109.

[119] S. Shin, R. Lin, and G. Gu, "Cross-analysis of botnet victims: New insights and implications," in *RAID*, ser. Lecture Notes in Computer Science, R. Sommer, D. Balzarotti, and G. Maier, Eds., vol. 6961. Springer, 2011, pp. 242–261.

[120] S. Shin, P. A. Porras, V. Yegneswaran, M. W. Fong, G. Gu, and M. Tyson, "Fresco: Modular composable security services for software-defined networks," in *NDSS*. The Internet Society, 2013.

[121] S. Shin, V. Yegneswaran, P. A. Porras, and G. Gu, "Avant-guard: scalable and vigilant switch flow management in software-defined networks," in *ACM Conference on Computer and Communications Security*, A.-R. Sadeghi, V. D. Gligor, and M. Yung, Eds. ACM, 2013, pp. 413–424.

[122] K. Singh, A. Moshchuk, H. J. Wang, and W. Lee, "On the incoherencies in web browser access control policies," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 463–478.

[123] S. Sinha, M. Bailey, and F. Jahanian, "Improving spam blacklisting through dynamic thresholding and speculative aggregation," in *NDSS*. The Internet Society, 2010.

[124] R. Sommer, D. Balzarotti, and G. Maier, Eds., *Recent Advances in Intrusion Detection - 14th International Symposium, RAID 2011, Menlo Park, CA, USA, September 20-21, 2011. Proceedings*, ser. Lecture Notes in Computer Science, vol. 6961. Springer, 2011.

[125] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010, pp. 305–316.

[126] J. Sonchack, A. J. Aviv, and J. M. Smith, "Bridging the data gap: Data related challenges in evaluating large scale collaborative security systems," in *Presented as part of the 6th Workshop on*

Cyber Security Experimentation and Test. Berkeley, CA: USENIX, 2013. [Online]. Available: https://www.usenix.org/conference/cset13/workshop-program/presentation/Sonchack

[127] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, “Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation,” in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, ser. BADGERS ’11. New York, NY, USA: ACM, 2011, pp. 29–36. [Online]. Available: http://doi.acm.org/10.1145/1978672.1978676

[128] Y. Song, A. D. Keromytis, and S. J. Stolfo, “Spectrogram: A mixture-of-markov-chains model for anomaly detection in web traffic,” in *NDSS*. The Internet Society, 2009.

[129] A. Sperotto, R. Sadre, F. Vliet, and A. Pras, “A labeled data set for flow-based intrusion detection,” in *Proceedings of the 9th IEEE International Workshop on IP Operations and Management*, ser. IPOM ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 39–50. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04968-2_4

[130] A. Srivastava and J. T. Giffin, “Automatic discovery of parasitic malware,” in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 97–117.

[131] S. Stafford and J. Li, “Behavior-based worm detectors compared,” in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 38–57.

[132] S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds., *Research in Attacks, Intrusions, and Defenses - 16th International Symposium, RAID 2013, Rodney Bay, St. Lucia, October 23-25, 2013. Proceedings*, ser. Lecture Notes in Computer Science, vol. 8145. Springer, 2013.

[133] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. A. Kemmerer, C. Kruegel, and G. Vigna, “Your botnet is my botnet: analysis of a botnet takeover,” in *ACM Conference on Computer and Communications Security*, E. Al-Shaer, S. Jha, and A. D. Keromytis, Eds. ACM, 2009, pp. 635–647.

[134] G. Stringhini, C. Kruegel, and G. Vigna, “Shady paths: leveraging surfing crowds to detect malicious web pages,” in *ACM Conference on Computer and Communications Security*, A.-R. Sadeghi, V. D. Gligor, and M. Yung, Eds. ACM, 2013, pp. 133–144.

[135] F. Tegeler, X. Fu, G. Vigna, and C. Kruegel, “Botfinder: finding bots in network traffic without deep packet inspection,” in *Proceedings of the 8th international conference on Emerging networking experiments and technologies*, ser. CoNEXT ’12, ACM. New York, NY, USA: ACM, 2012, pp. 349–360. [Online]. Available: http://doi.acm.org/10.1145/2413176.2413217

[136] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, “Design and evaluation of a real-time url spam filtering service,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2011, pp. 447–462.

[137] A. Varga *et al.*, “The omnet++ discrete event simulation system,” in *Proceedings of the European Simulation Multiconference (ESM’2001)*, vol. 9. sn, 2001, p. 185.

[138] G. Vasiliadis, M. Polychronakis, and S. Ioannidis, “Midea: a multi-parallel intrusion detection architecture,” in *ACM Conference on Computer and Communications Security*, Y. Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp. 297–308.

[139] S. Venkataraman, D. Brumley, S. Sen, and O. Spatscheck, “Automatically inferring the evolution of malicious activity on the internet,” in *NDSS*. The Internet Society, 2013.

[140] D. Y. Wang, S. Savage, and G. M. Voelker, “Juice: A longitudinal study of an seo botnet,” in *NDSS*. The Internet Society, 2013.

[141] A. M. White, S. Krishnan, M. Bailey, F. Monrose, and P. A. Porras, “Clear and present data: Opaque traffic and its security implications for the future,” in *NDSS*. The Internet Society, 2013.

[142] A. M. White, A. R. Matthews, K. Z. Snow, and F. Monrose, “Phonotactic reconstruction of encrypted voip conversations: Hookt on foniks,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2011, pp. 3–18.

[143] C. Willems, T. Holz, and F. Freiling, “Toward automated dynamic malware analysis using cwsandbox,” *Security & Privacy, IEEE*, vol. 5, no. 2, pp. 32–39, 2007.

[144] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, “A practical attack to de-anonymize social network users,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2010, pp. 223–238.

[145] C. Wressnegger, F. Boldewin, and K. Rieck, “Deobfuscating embedded malware using probable-plaintext attacks,” in *RAID*, ser. Lecture Notes in Computer Science, S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds., vol. 8145. Springer, 2013, pp. 164–183.

[146] C. V. Wright, C. Connelly, T. Braje, J. C. Rabek, L. M. Rossey, and R. K. Cunningham, “Generating client workloads and high-fidelity network traffic for controllable, repeatable experiments in computer security,” in *RAID*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds., vol. 6307. Springer, 2010, pp. 218–237.

[147] C. V. Wright, S. E. Coull, and F. Monrose, “Traffic morphing: An efficient defense against statistical traffic analysis,” in *NDSS*. The Internet Society, 2009.

[148] J. Xu, J. Fan, M. Ammar, and S. B. Moon, “On the design and performance of prefix-preserving ip traffic trace anonymization,” in *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*. ACM, 2001, pp. 263–266.

[149] 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 *Network Protocols, 2002. Proceedings. 10th IEEE International Conference on*. IEEE, 2002, pp. 280–289.

[150] G. Yan, S. Eidenbenz, and E. Galli, “Sms-watchdog: Profiling social behaviors of sms users for anomaly detection,” in *RAID*, ser. Lecture Notes in Computer Science, E. Kirda, S. Jha, and D. Balzarotti, Eds., vol. 5758. Springer, 2009, pp. 202–223.

[151] T.-F. Yen, Y. Xie, F. Yu, R. P. Yu, and M. Abadi, “Host fingerprinting and tracking on the web: Privacy and security implications,” in *NDSS*. The Internet Society, 2012.

[152] H. Yu, C. Shi, M. Kaminsky, P. B. Gibbons, and F. Xiao, “Dsybil: Optimal sybil-resistance for recommendation systems,” in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2009, pp. 283–298.

[153] T. Yu, G. Danezis, and V. D. Gligor, Eds., *the ACM Conference on Computer and Communications Security, CCS’12, Raleigh, NC, USA, October 16-18, 2012*. ACM, 2012.

[154] J. Zhang and G. Gu, “Neighborwatcher: A content-agnostic comment spam inference system,” in *NDSS*. The Internet Society, 2013.

[155] J. Zhang, C. Yang, Z. Xu, and G. Gu, “Poisonamplifier: A guided approach of discovering compromised websites through reversing search poisoning attacks,” in *RAID*, ser. Lecture Notes in Computer Science, D. Balzarotti, S. J. Stolfo, and M. Cova, Eds., vol. 7462. Springer, 2012, pp. 230–253.

[156] J. Zhang, Y. Xie, F. Yu, D. Soukal, and W. Lee, “Intention and origination: An inside look at large-scale bot queries,” in *NDSS*. The Internet Society, 2013.

[157] K. Zhang, Z. Li, R. Wang, X. Wang, and S. Chen, “Sidebuster: automated detection and quantification of side-channel leaks in web application development,” in *ACM Conference on Computer and Communications Security*, E. Al-Shaer, A. D. Keromytis, and V. Shmatikov, Eds. ACM, 2010, pp. 595–606.

[158] M. Zimmer, ““but the data is already public”: On the ethics of research in facebook,” *Ethics and Information Technology*, vol. 12, no. 4, pp. 313–325, 2010.