Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets

Milan Cermak*†, Tomas Jirsik*†, Petr Velan*, Jana Komarkova*†,
Stanislav Spacek*†, Martin Drasar*, Tomas Plesnik*
* Masaryk University, Institute of Computer Science, Brno, Czech Republic
† Masaryk University, Faculty of Informatics, Brno, Czech Republic
E-mail: {cermak, jirsik, velan, komarkova, spaceks, drasar, plesnik}@ics.muni.cz

***Abstract*—Research in network traffic measurement and analysis is a long-lasting field with growing interest from both scientists and the industry. However, even after so many years, results replication, criticism, and review are still rare. We face not only a lack of research standards, but also inaccessibility of appropriate datasets that can be used for methods development and evaluation. Therefore, a lot of potentially high-quality research cannot be verified and is not adopted by the industry or the community. The aim of this paper is to overcome this controversy with a unique solution based on a combination of distinct approaches proposed by other research works. Unlike these studies, we focus on the whole issue covering all areas of data anonymization, authenticity, recency, publicity, and their usage for research provability. We believe that these challenges can be solved by utilization of semi-labeled datasets composed of real-world network traffic and annotated units with interest-related packet traces only. In this paper, we outline the basic ideas of the methodology from unit trace collection and semi-labeled dataset creation to its usage for research evaluation. We strive for this proposal to start a discussion of the approach and help to overcome some of the challenges the research faces today.**

I. Introduction

Research in the area of network traffic measurement and analysis is heavily data-driven. Good research is closely linked to the availability of network data that are realistic, current, well-documented, and publicly available ideally. Without such data and standards of their usage, researchers have limited research opportunities, cannot reliably prove their results, and tend to repeat common mistakes [1]. However, a creation of such data is, in itself, one of the challenges of the research area. Neither real-world data nor artificial data are sufficient as research input. Real-world data suffer from insufficient annotation and problems with anonymization. Artificially generated data, on the other hand, typically contain a limited set of network traffic types and do not reflect the specifics associated with real-world networks [8]. As a result, there is no generally accepted type of research data, which results in researchers' inability to justify their results, as proposed by Krishnamurthy et al. [12]. This fact is confirmed by the small number of publications focused on verification of others researchers' results, such as the work by Mehedi at al. [16], despite the fact that verification of others researchers' results is one of the pillars of science.

Although the need for research data in the area of network measurement and analysis was emphasized over the last decades, we still do not have a solution that the research community would agree on. Several widely accepted works, however, emerged during this period with the aim to bring us closer to an agreement. Well-known examples of such works are KDD [11] and DARPA [17] datasets and evaluation approaches focused on an analysis of network traffic and anomaly detection. The fact that these datasets are still widely used despite being almost twenty years old and facing criticism since their introduction [15] illustrates how much is a generally approved data source needed. In addition to these, other publicly available datasets have emerged to date based on a collection of network traffic from security competitions such as DEF CON [18], or have been generated using a simulated environment [3], [26], [28]. The opposite approach to the creation of research data is addressed by organizations like CAIDA [6], which provide real-world packet traces, but which have limited use possibilities due to anonymization and lack of annotation. Recently, top conferences put emphasis on publishing the research data along with research results [2]. Analysis by Grajeda et al. [9] shows that despite the encouragement it is not common research practice so far.

The ultimate challenge of the research in the area of network measurement and analysis is a methodology for the creation of packet trace datasets that are modifiable, extensible, reproducible, and can be publicly shared [8]. We propose to use *semi-labeled datasets* in order to achieve this goal. These datasets consist of so-called *annotated units* of network traffic, which are composed of interest-related packet traces only. These traces, with a defined process of their normalization, can be combined not only together but also with a real-world traffic in order to form semi-labeled datasets. Assuming that researchers have access to real-world traffic in their labs, the proposed approach will enable a simple creation of customized datasets. While the real-world traffic captures need to be kept private, the annotated units can be freely shared since they only contain the interest-based trace of traffic with a minimum of private information. We do not claim that semi-labeled datasets provide a universal solution to all problems related to dataset creation, usage, and sharing. However, we aim to show, that it offers more benefits than other current approaches.

In this article, we demonstrate benefits of our approach to an anomaly detection research problem. We discuss problematic areas and demonstrate how our approach helps to mitigate

some of the problems and how it compares to existing dataset creation methods. The paper serves as both an initial introduction to the concept of semi-labeled datasets, and as a discussion of relative merits of other approaches compared to our own. We are aware that there is no one-size-fits-all solution, but we believe that our approach helps to overcome many of the challenges the researchers still face today.

The rest of the paper is organized into sections according to areas related to usage of network traffic datasets. Section II defines creation process of annotated units. Section III presents a methodology of their usage. Section IV discusses possibilities of research data sharing. Section V addresses the use of the presented approach for research evaluation. Lastly, Section VI summarizes and concludes the whole approach based on semi-labeled datasets.

II. Creation of Annotated Units

There is a number of factors that determine the usefulness and applicability of a newly created dataset. Both richness of content and quality of annotation depend on the process of dataset creation. Real-world network traces often require anonymization and are impossible to annotate accurately. Artificially created laboratory traffic, on the other hand, is often thin on content and may be biased. Ultimately, the best method of creating a dataset depends on the purpose of the dataset. The result is always a compromise between authenticity and accuracy of annotation. In this paper, we focus on datasets used for verification of network traffic analysis methods and propose an approach to reduce those compromises to maximize both authenticity and accuracy.

Two types of datasets can be generated from network traffic. The first type is network packet traces, which are created by capturing traffic at a certain point of the network. The second, one aggregation level above packet traces, is flow traces, which are collected as an output of a flow monitoring process. Both types are used in attack and anomaly research, but we focus only on packet traces, as they are a superset of flow traces and can be converted to flows if needed. Moreover, the content of flow traces depends on flow exporter configuration. Having the raw packet data, one can always prepare the flow traces according to their needs.

Applicable datasets used for the verification of network traffic measurement and analysis methods must address the following four challenges related to their creation and usage:

- **Anonymization** – It is difficult to anonymize traffic including application data and preserve consistency of data in a network, transport, and application layers as shown by Yurcik et al. [31].
- **Traffic annotation** – Either the traffic is captured from a live network and then annotation is inaccurate or the traffic is artificially generated with annotation and then its authenticity is compromised.
- **Capture parameters** – It is often unclear how the dataset was captured. Various settings such as network topology, network capacity, utilization, and latency might affect the dataset creation but are usually not measured or provided.
- **Dataset recency** – Datasets are created at a certain point in time, however, the composition of traffic on a network is ever changing. Therefore, each fixed dataset becomes obsolete in time [1].

To mitigate these issues, we propose to create a set of smaller datasets called *annotated units*, which can be combined with real-world background traffic into *semi-labeled datasets*. Each annotated unit is a normalized dataset containing a single event, such as a network attack. For example, a unit containing a brute-force SSH attack would consist of several complete connections of the attacker trying a different password against service on port 22. Such a unit could capture a time interval of several minutes, hours, or even days.

Creation of the annotated unit can be accomplished either by filtering the desired traffic from an existing network or by capturing a traffic from a prepared environment. An obstacle to the first approach is that it requires a capturing device and an exact knowledge of traffic parameters such as the protocol, IP addresses, and ports. Moreover, the background traffic can influence the desired traffic unpredictably by causing delays or by causing packet drops in case of too high network utilization. A prepared environment does not have these obstacles but suffers from its own problems, which are mainly related to difficulty in the environment setup. In our opinion the latter approach is superior. To alleviate its obstacles and to simplify its usage we have prepared a virtual environment to emulate the attacker-defender interactions, see Figure 1. The environment is set up using Vagrant [21] and as its input it only requires commands to be run at an attacker and defender hosts. The complete environment is built from scratch on every run, which greatly aids repeatability of the whole process. The software with an example of annotated units is available as an open-source at *https://github.com/CSIRT-MU/TraceShare*. Although the virtual environment was used to create datasets before [10], [23], we are the first to focus on automation and repeatability of the creation of individual attack units rather than the whole dataset including background traffic.

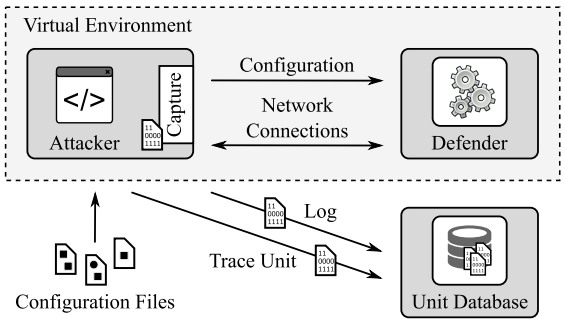

Fig. 1. Architecture of the TraceCreator virtual environment.

Following the creation, the annotated units have to be normalized and injected into the real-world background traffic to create the semi-labeled dataset. The purpose of normalization is to ensure that the annotated units can be treated uniformly in the injection process. We propose the following steps: change

MAC and IP addresses to values not usually encountered in traffic, differentiate attackers' and defenders' IP addresses, and reset timestamps of packets so that the first one starts at zero epoch time. These steps ensure uniformity of annotated units. Tools such as *tcprewrite* and *editcap* can be used for this process. We also propose the use of the PcapNg [25] format, because it is supported by the majority of traffic processing tools and offers extended timestamp precision, capture statistics, mixed link layer types, and user comments in addition to original Pcap format. The whole process of combining the annotated units with real-world background traffic is discussed in more detail in Chapter III.

We assert that the proposed approach addresses all the aforementioned issues of the current network traffic datasets. Anonymity and consistency are trivially achieved because the unit contains only controlled traffic that is not sensitive, normalization requires more attention. Annotation is accurate because the content of units is completely known. Capture parameters are both known and easily altered on demand due to the capture happening in a virtual environment with easy repeatability. Data recency is achieved by manipulating the pieces of traffic that can be either combined with newer background traffic to simulate the old attacks in new traffic conditions, or replaced by newer versions of attacks.

In addition, the proposed approach provides a number of advantages. By dealing with isolated pieces of traffic it lowers the barrier of data sharing. It enables tailoring the datasets for specific environments by means of mixing them with provided background traffic. It enables simple alteration of traffic parameters in the virtual environment, thus providing variations of the same attack, e.g., in low-latency communication over a local wired network, in large distance communication, or using a poor wireless connection. Last but not least, it enables gradual updating of existing datasets, thus facilitating an ongoing validation of detection methods.

Despite the overall improvement this approach brings, there is still a number of issues that require addressing. First, the sharing of a background traffic is still a problem because of anonymity concerns. While everybody can create their own dataset using their own background traffic to test detection methods in their own environment, a direct comparison of detection methods requires a complete re-evaluation using a local dataset. The second issue is the normalization process. Any attempt to change the link and network layer information in the packet headers can lead to inconsistencies with higher level protocols. Therefore, the normalization of each annotated unit must be evaluated for its potential negative impact. New tools have to be developed to provide greater flexibility for normalization of higher level protocols. For example, when DNS traffic is involved, we need to be able to transparently change the association between DNS names and IP addresses. Although much of this work is common to anonymization as well, there is an important difference: the normalization must maintain consistency for the intended purpose of the dataset. When some application layer data are not used by the tools processing the dataset, they can be safely ignored. However, anonymization must take all the data into account to avoid privacy issues. The third issue is related to using a virtual environment. The system under attack is usually in its default mode, without any data and users. Moreover, the virtual environment provides only very limited resources. While it is not an issue for attacks such as SSH brute-force, it might affect the authenticity of annotated units for other attacks. For example, an application layer DoS would generate different responses from a production web server with many resources than from a virtual machine utilizing only a single CPU core and significantly limited amount of memory. Another problem with the controlled environment is its uniformity. Although we are able to adjust the parameters of the environment, it is difficult to predict what will the impact be, if any, on the detection methods. We believe that these issues should be subject to a further research.

III. Use of Semi-labeled Datasets

The acquisition of annotated units is the first step towards a rigorous network traffic measurement and analysis. The second step, equally essential, is a creation of a complex dataset suitable for the right use-case according to a coherent methodology. For example, a dataset used for development and evaluation of Advanced Persistent Threats detection must contain all relevant attack types instead of network scans only. In addition, the correct methodology of dataset creation must be used to unify its further processing. Examples of issues related to the use of such datasets were identified in the survey by Tavallaee et al. [24]. The authors observed that majority of surveyed papers have problems with obtaining suitable datasets and use different procedures of data processing, which complicate further research comparison. In other words, we need to know the limitations and specifics of the dataset we are planning to use for development and evaluation. An example of known specifics limiting the use of public datasets is a small, fixed range of TTL values in DARPA dataset, which makes it unsuitable for the creation of detection methods based on TTL-based network address translation [14]. This section proposes a coherent methodology for the creation and use of semi-labeled datasets for network traffic monitoring and analysis with respect to described challenges. To ease understanding of the methodology, we present development of an anomaly detection method as a model use-case.

The methodology for a creation and use of semi-labeled datasets is induced from existing implementations YD2T [28] and FLAME [5] focused on a composition of realistic synthetic datasets. The semi-labeled dataset is created by mixing annotated units with a real-world dataset (unlabeled) containing a background live traffic. First, a real-world dataset is captured via standard traffic access methods from real network and statistical characteristics of the dataset are computed. These characteristics represent for example IP address distribution, TTL distribution, capture period, distribution of packet inter-arrival times, and network link properties. Secondly, selected annotated units are modified to reflect characteristics of the real-world dataset computed in the previous step. It

is not always necessary to modify all characteristics. For example, we can keep the original IP address distribution of annotated units when developing a detection method that does not utilize IP address for detection. In this case, original IP addresses from annotated units can serve as a natural label of the inserted traffic. For such labeling purposes, IP addresses belong to a reserved range that should not be present in real-world traffic. Thirdly, a user specifies options for dataset merge, i.e. how many times is the annotated unit merged to a real-world dataset, how to allocate the unit in the background traffic (randomly vs. exact timestamp), and what labeling method should be used (labels are a part of the dataset or are stored externally). Last, annotated units and real-world dataset are merged into the semi-labeled dataset. The merge can be done for example using the *mergecap* tool [22] or the aforementioned YD2T toolkit. The adjective "semi-labeled" refers to the fact that we are able to distinguish the injected labeled traffic from the real one in the created dataset.

Annotated units contain only precisely defined and completely annotated traffic captured from the virtual environment or a real network. Hence, they capture the essence of network traffic and are suitable for initial comprehension of analyzed network traffic and ground truth establishment. The presence of ground truth enables us to create and evaluate various analyses and detection methods rigorously. These two properties are the most valuable assets of these annotated units. Although the annotated units serve well to network traffic comprehension, use of annotated units without real-world data prevents understanding of the network traffic in a broader concept. Therefore, annotated units should be used in combination with real-world data. The result is represented by the semi-labeled dataset, which is a balanced trade-off between the need for a dataset representing real-world traffic and the need for labeled datasets. This dataset has all the main characteristics of real-world traffic while still being able to keep the partial comprehension. This methodology can not provide an absolute ground truth though, as it can not guarantee an absence of unlabeled data of interest in the whole dataset. Nevertheless, we suggest a coherent approach to an evaluation facing this issue in Chapter V. The semi-labeled datasets are suitable for evaluation and improvement of network traffic analysis methods in a semi-controlled environment closely resembling the real-world conditions. The creation process (mixing annotated units with real-world data) ensures sufficient variability and provides a wide range of different environments for the evaluation.

Figure 2 demonstrates the use of the semi-labeled dataset on the example of a development of network threat detection inspired by PDSA methodology [13]. Annotated units containing threat-related traffic are used for initial comprehension of the threat mechanism. Based on the comprehension, a detection method is derived and a prototype is implemented. Optionally, we can combine threat-related units with other annotated units to evaluate the prototype against another, still annotated, traffic. Once 100 % recall and precision are achieved, the detection method works properly in a controlled environment and is ready to be tested using a semi-labeled dataset.

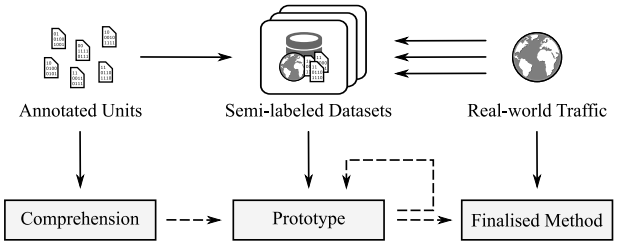

Fig. 2. Usage workflow of annotated units and semi-labeled datasets.

The semi-labeled dataset is created by merging the previously used annotated units with a real-world traffic. The detection method should still be able to detect the threats contained in merged annotated units. Nevertheless, a false positive (FP) detection may occur when real-world part of the semi-labeled dataset is analyzed. All detected events from the real-world part should be inspected manually. If the inspection shows true positive (TP), a new annotated unit can be created from associated traffic. In case of FP detection, a detection algorithm is enhanced to provide the correct result. The evaluation process is iterative. Each time a method is enhanced, it is re-evaluated using the original semi-labeled dataset. The re-evaluation continues until FP rate reaches the desired level. After that, a new semi-labeled dataset can be created by merging annotated units from the previous iteration with a new real-world dataset. The optimization process is repeated with the new semi-labeled dataset. In this fashion, several semi-labeled datasets can be tested until a desired robustness of the detection method is achieved.

The proposed methodology can be demonstrated more precisely on an SSH brute-force attack detection as follows. First, a network traffic pattern identifying an SSH attack is derived from an annotated unit and a prototype matching the identified traffic pattern is implemented and evaluated against the annotated units. Secondly, the semi-labeled dataset is used. If a new, unlabeled SSH attack is detected, the attack is inspected manually. It is possible that FP detection occurs as a common traffic pattern that fits a detection pattern, e.g. automated RDP authentication resembles SSH attack in terms of flow count and timing. Based on the result of the inspection, SSH attack pattern matching options are revised in case of FP, or a label is added to the detected flows in case of TP detection. This process is repeated until the pattern is able to detect the SSH attack with a specific accuracy and precision.

Using the combination of annotated units and a semi-labeled dataset allows us to develop and evaluate a detection method in the near-real-world conditions which results in a high chance of a successful method deployment. We also recommend gathering a feedback from users of the detection method even during the real-world deployment. The deployment can discover a special network traffic that could disrupt the detection. The anomalous network traffic samples can be used for further improvement of the method and should be collected for further analysis, development, or sharing.

IV. Dataset Sharing

Information sharing and cooperation is a key component of efficient research. While hackers and other cyber-attackers are aware of this and have already adopted these habits, it is still a challenge for a network related research community [4]. The community cooperates well thanks to a number of top conferences, but in the case of data sharing there is clearly room for improvement [9]. Some of these top conferences have recently begun to address this issue and encourage authors to share their data [2]. Aside from influencing the efficiency of development, datasets sharing is necessary for research repeatability and for the reproduction of results. Even so, most researchers that use their own datasets decide not to make them public. Previous research shows that custom datasets are made public only in 4 % of all cases [9]. However, if there already exists a public dataset, it is reused in 50 % of all cases. There is clearly noticeable demand for publicly available datasets.

There are several platforms for sharing datasets currently available. Digital Corpora [7] is a platform that stores disk images, memory dumps, and network packet captures. Pcapr [19] is a platform that shares a large amount of network traffic captures in the Pcap format. Similarly, Network Forensics and Network Security Monitoring sharing platform [20] provides another collection of packet traces with a malware network traffic and other captured events. Unfortunately, each of these platforms suffers from common problems.

Issues of network data sharing platforms may be categorized into two groups. The first group contains issues that relate to the transformation the datasets must go through prior to being shared. This transformation encompasses anonymization and normalization. The second group contains issues that involve the dataset sharing platform itself. For example, what requirements should the platform fulfill to accommodate the needs of as many researchers as possible while remaining user-friendly and manageable for an extended period of time. Some of these issues, when left unresolved, have already led to a demise of several sharing platforms in the past. The following list outlines these challenges and presents possible solutions so the platform can be successfully adopted by the community.

- **Anonymization** – A process of anonymization is the main obstacle to a publication of datasets. Assisted anonymization of uploaded datasets should be one of the key features of a central dataset sharing platform. Such anonymization alleviates part of the workload the anonymization demands from the dataset authors. This procedure can be applied to all uploaded datasets and significantly contributes to the credibility of the platform.
- **Data heterogeneity** – Network data are typically gathered by various methods in countless different environments. A certain level of disorder is sure to develop in such a collection and leads to difficult navigation and complicated searching for a certain dataset. Central sharing platform should have clearly defined types of datasets it collects. Then it is possible to establish a standard representation of data and adhere to it where possible. An example is a definition of a normalization process that will be invoked every time a new dataset is being uploaded [30].
- **Platform sustainability** – Providing an up-to-date content is a critical part of the sharing platform and a necessity for an outgoing research [30]. Several projects of dataset sharing platform early ended in the past after only a few years of operation. An example is Digital Corpora, updated for the last time in 2014. When preparing such a project, it is necessary to include personal and financial sustainability. Another solution is a creation of the platform as an open community hub and thus lower the requirements for the involvement of platform managers.
- **Initial content** – Both community and centrally managed sharing platforms must contain a sufficient number of datasets when launched. These datasets should already be up-to-date and usable in current research. Otherwise, the researcher will have no incentive to visit the platform and the project will probably end with little success.

To tackle the above-mentioned challenges, we are currently developing a web-based centralized dataset sharing platform that will store annotated units of network traffic data. The sharing platform will be built upon basic functions of uploading, searching, downloading, and mixing of annotated units. Other functions will provide user account management and inter-user communication. In our platform, the user will have access to a wide range of annotated units uploaded by us or the community itself. A good example of such sharing platform is OpenML [27] focused on other areas of research. The user interface of the OpenML platform is illustrated in Figure 3. This platform enables the sharing of datasets for training machine learning algorithms and provides several community functions that are suitable to our needs as well.

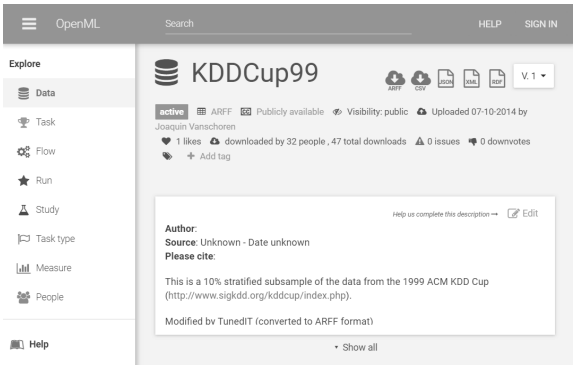

Fig. 3. User interface of the OpenML sharing platform [27].

We plan to assist the users with uploading of annotated units and to ensure that datasets are normalized and anonymized by a unified procedure described in more detail in Section II. In addition to the capture itself, a relevant annotation process will be provided. Aside from plaintext annotation, the unit can be given tags based on the content and intended use. The platform user will also be able to edit this information after the unit is shared and other users can leave comments

to any shared annotated unit. Because the users will have a simple, direct way to contact the author of any annotated unit, the annotation is likely to be kept up-to-date and accurate. The assistance provided by our platform will simplify the process of normalization while providing a unified form of the output unit. That should keep the units organized and easy to search for. Searching for an annotated unit containing a specific event will be based on the unit's identifier – its name, annotation and assigned tags. A tag may denote the captured protocol or application (SSH), identify the operating system of communicating devices (Linux) or a name of a captured anomaly, etc. By combining these tags, any user should be able to find an annotated unit suitable to his/her specific needs. The user will be able to download an annotated unit that is being shared on the sharing platform. Besides downloading a single unit, the platform will allow users to combine various annotated units to ease follow-up creation of semi-labeled dataset, as described in Section III. It is worth mentioning that we plan to enable sharing of the combination of the annotated units through the platform as well.

The proposed dataset sharing platform is intended as a community hub, where the community is provided with tools for simple network data sharing. The unified procedures for normalization with anonymization included will narrow the heterogeneity of the shared data and, at the same time, build trust in the sharing platform. Furthermore, the community-based approach will help to alleviate some workload from the hub managers and ensure project sustainability with regular updates. The initial content will be provided by us and cooperating research teams before the sharing platform will be publicly accessible. We have a large network with deployed monitoring infrastructure at our disposal, which enables us to provide units with various network events captured. Moreover, we have a large virtual research environment KYPO [29] to create artificial unit traces and regenerate existing ones with updated features. We believe that the dataset sharing platform based on our design will make dataset sharing more common and help the reproducibility of results in future research.

V. Research Evaluation

In order to catch up the objective performance of the evaluated method, the datasets must be used for evaluation in a consistent manner. Be it an objective method comparison for benchmarking, evaluation of method on a given network, or evaluation done for parameter estimation for higher level methods, the problems and suggested solutions for dataset usage outlined in this section apply to all of them. In this section we discuss the methodology of using semi-labeled datasets for evaluation and benchmarking. We outline the problems and their possible solutions and discuss advantages and disadvantages of respective approaches.

There are several works discussing the challenges related to evaluation methodology. McHugh [15] extensively covers deficiencies found in 1998 and 1999 DARPA intrusion detection systems evaluation. Apart from the critique of the datasets themselves, he points out the problematic estimation of the sample of analysis and its distortion of results of an evaluation. He also debates the fitness of ROC (Receiver Operating Curve) for the purpose of expressing the evaluation results. Gharib et al. [8] present a formula to estimate quality of datasets for evaluation purposes and allow comparison. The formula is based on 11 desired dataset characteristics: knowledge of full dataset underlying network configuration, complete traffic capture, labeling, complete interaction patterns, wide range of available protocols, attack diversity, anonymity, heterogeneity of data sources (packet captures, logs, ...), rich feature set, and complete metadata. Tavallaee et al. [24] survey the evaluation of anomaly detection systems performed by researchers. They focus on reliability and validity of the evaluation methodology. The reliability indicates the repeatability and robustness of the evaluation, the validity indicates how much does the result of the evaluation experiment attribute to the proposed approach. In their survey, most of the recent works are found lacking. They observe that metrics based on false/true positive/negative are most often used for performance evaluation.

Problems associated with evaluation using dataset can be, in essence, divided into two categories: *qualitative* and *quantitative*. The qualitative aspect of evaluation using datasets relates to the properties of the dataset itself – in order to achieve objective evaluation, the dataset must contain realistic, diverse data, that accurately reflect actual traffic. The quantitative aspect contains the actual process of evaluation of the test results. The evaluation process must give an objective metric of the method efficiency. The most commonly used approach is confusion matrix composed of *true positive* (TP), *false positive* (FP), *true negative* (TN), and *false negative* (FN) values. The confusion matrix is also used to compute another derived metrics commonly used for the quantitative evaluation, such as *false positive/negative rate*, *precision*, *accuracy*, and *recall*.

There are essentially three types of datasets: real-world datasets, artificial datasets, and mixed (semi-labeled) datasets. Each type has its own advantages and disadvantages in terms of use for evaluation. We will discuss each type in detail and show there is a trade-off between the qualitative and quantitative aspect of evaluation for each type of dataset.

Datasets based on captured real-world traffic are best in terms of unbiased representation of real traffic. However, the evaluation of results is complicated. In order to compute the metrics, the dataset needs to be labeled. That has proven to be an extremely demanding task and there are only a few labeled datasets available, most of them outdated. Furthermore, such datasets cannot be easily modified for evaluation of a method for a specific network or specific attack type. Since it is not very likely that a general approach for labeling captured traffic will be developed anytime soon, the datasets based solely on captured traffic are not suitable for the purpose of objective method evaluation.

On the other hand, the fully synthesized artificial datasets are extremely well suited for quantitative evaluation since the nature of the data in the dataset is fully known and the dataset is well annotated. However, creating a dataset that accurately imitates traffic in a real network and fulfills the qualitative

demands is challenging. The traffic in the network is generated by an enormous variety of applications and the captured traffic is often malformed due to errors adding another layer of complexity. A solution of the problems with fully synthesized datasets is not available at the moment. The synthesized traffic is too out-of-the-book to challenge the evaluated methods.

The proposed semi-labeled datasets consisting of a mix of real-world traffic and annotated units balance the qualitative and qualitative aspects of the evaluation. The annotated unit consists of attacks while the real-world traffic provides a background for those attacks. The semi-labeled datasets can be created easily and can be tailored to a specific network. They are not completely realistic, but the capture provides varied traffic, which accounts for a wide range of applications and network errors. However, the background traffic can also contain malicious traffic, which makes the quantitative evaluation inaccurate. The problem is illustrated in Figure 4. The dark gray area corresponds to malicious traffic from annotated units, the rest is unlabeled. The light gray area corresponds to data that were identified by the evaluated method as malicious. Those identified data can be, with some effort, also labeled since it should be a fraction of the original data (the gray area is known). The true negative area is out of concern for us at the moment, as the evaluated method does not identify these data as malign and they are truly not malign. The inaccuracy of evaluation comes from the white part of the false negative region. This unknown volume of data comes from the real-world dataset and is not identified by the method. Its volume must be accounted for. Otherwise, all the measures of false/true or positive/negative rates are not accurate. To address this issue, we plan to consider a probabilistic-based evaluation approach. We aim to experimentally estimate the probability of occurrence of an attack in real-world traffic. Based on this information, we will be able to derive the probability of occurrence of FN during the evaluation and asses the method with a certain degree of confidence.

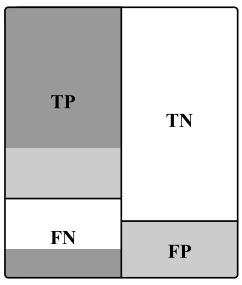

Fig. 4. Confusion matrix in semi-labeled datasets.

There are several ways to improve the precision of the quantitative evaluation based on semi-labelled datasets. Either a condition on the mixing process can be formed in order to bound the evaluation result's error, or the unspecified part of the false negative can be disregarded as it is likely to be small, not affecting the evaluation results. In the mixing process, either positive data can be added by mixing more attacks into the traffic or the number of negative samples can be increased by adding benign traffic.

However, adding either positive or negative annotated units can skew the traffic composition and therefore influence the evaluation of results as well. Adding positive units decreases the influence of the unknown volume of false negatives, but it increases the a priori probability that a sample is positive. On the other hand, adding negative units can artificially decrease false positive rate because they would probably be examples of completely regular traffic. The most likely scenario is that there will be several real-world traffic samples that are frequently used as the background traffic in semi-labeled datasets used for evaluation. Each use would lead to an examination of the positive results not included in mixed annotated units. The more the traffic samples are used, the fewer samples fall into the unknown area of false negative results. The effective reduction of the unknown area of false negative results requires an extensive sharing of the positive results not included in mixed datasets among the research community.

VI. Summary and Conclusion

Research validation and verification are fundamental principles of good scientific work. In terms of research in the area of network traffic measurement and analysis, however, these principles pose a great challenge. The research heavily depends not only on the correct processes of data usage but also on the availability of network traffic datasets that meet the common requirements and are publicly available. Without these datasets, we will never be able to reliably repeat, validate, and analyze research results. To overcome this challenge, we proposed the semi-labeled datasets approach based on sharing of small annotated units of network traffic that adopts our lessons learned from other research works.

The main idea of the discussed approach is based on annotated units of network traffic that can be synthetically generated, or derived from real-world traffic. These units typically contain only a minimum of personal data, so they can be shared and, thanks to the restrictions on the inclusion of interest-related traffic only, be easily annotated. Annotated units can be generated with a variety of traffic types, protocols, or network attacks. They can be easily normalized and combined with each other or with a real-world traffic to create so-called semi-labeled datasets. The semi-labeled dataset is represented by a combination of private traffic capture of non-annotated real-world network traffic and an annotated baseline that can be publicly shared. We are currently developing an open sharing platform that will allow the community to provide such research data. We have also outlined the basic methodology of using semi-labeled datasets to evaluate research results and allow mutual comparison of different analysis methods. We strive to cover all areas related to the issue of research provability and believe that our approach provides a comprehensive solution to challenges of research provability and data sharing. All relevant source codes and additional documentation of the proposed methodology are publicly available at *https://github.com/CSIRT-MU/TraceShare*.

This article is the first introduction of the concept of semi-labeled datasets in the area of network traffic sharing and we are aware that there are many challenges that need to be addressed in further research. Our goal is not to deal with all identified problems at this point, but to present a general solution in order to start a discussion of its usability. We hope that the follow-up discussions will help us to move forward to a solution that will be accepted by the research community, help us to establish better research conditions, and make research more accessible to other researchers and the industry as well.

ACKNOWLEDGEMENT

This research was supported by the Security Research Programme of the Czech Republic 2015-2020 (BV III/1 – VS) granted by the Ministry of the Interior of the Czech Republic under No. VI20162019014 – Simulation, detection, and mitigation of cyber threats endangering critical infrastructure.

REFERENCES

[1] S. Abt and H. Baier, "Are We Missing Labels? A Study of the Availability of Ground-Truth in Network Security Research," in *2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*. IEEE, sep 2014, pp. 40–55.

[2] Annual Computer Security Applications Conference. (2017) Call for Submissions. Accessed: 2018-02-20. [Online]. Available: https://www.acsac.org/2017/cfp/

[3] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Towards generating real-life datasets for network intrusion detection," *International Journal of Network Security*, vol. 17, no. 6, pp. 683–701, 2015.

[4] S. Bratus, "What Hackers Learn that the Rest of Us Don't: Notes on Hacker Curriculum," *IEEE Security & Privacy*, vol. 5, no. 4, 2007.

[5] D. Brauckhoff, A. Wagner, and M. Martin, "FLAME: a Flow-Level Anomaly Modeling Engine," in *Proceedings of the Conference on Cyber Security Experimentation and Test (CEST)*. San Jose, CA: USENIX Association, 2008, p. 6.

[6] CAIDA. (2018) The CAIDA Anonymized Internet Traces Dataset 2008 – Ongoing. Accessed: 2018-05-11. [Online]. Available: http://www.caida.org/data/passive/passive_dataset.xml

[7] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt, "Bringing science to digital forensics with standardized forensic corpora," *Digital Investigation*, vol. 6, no. SUPPL., pp. S2–S11, sep 2009.

[8] A. Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "An Evaluation Framework for Intrusion Detection Dataset," in *2016 International Conference on Information Science and Security (ICISS)*, no. Cic. IEEE, dec 2016, pp. 1–6.

[9] C. Grajeda, F. Breitinger, and I. Baggili, "Availability of datasets for digital forensics – And what is missing," *Digital Investigation*, vol. 22, pp. S94–S105, aug 2017.

[10] J. Guerra and C. Catania, "Improving the Generation of Labeled Network Traffic Datasets Through Machine Learning Techniques," in *XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).*, 2017.

[11] S. Hettich and S. D. Bay, "The UCI KDD Archive," Irvine, CA: University of California, Department of Information and Computer Science, Tech. Rep., 1999, accessed: 2018-03-10. [Online]. Available: http://kdd.ics.uci.edu

[12] B. Krishnamurthy, W. Willinger, P. Gill, and M. Arlitt, "A Socratic method for validation of measurement-based networking research," *Computer Communications*, vol. 34, no. 1, pp. 43–53, jan 2011.

[13] G. J. Langley, R. D. Moen, K. M. Nolan, T. W. Nolan, C. L. Norman, and L. P. Provost, *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*, ser. Wiley Desktop Editions. Wiley, 2009.

[14] M. V. Mahoney and P. K. Chan, "An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection," in *In Proceedings of the Sixth International Symposium on Recent Advances in Intrusion Detection*, vol. 2820. Springer, Berlin, Heidelberg, 2003, pp. 220–237.

[15] J. McHugh, "Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory," *ACM Transactions on Information and System Security*, vol. 3, no. 4, pp. 262–294, nov 2000.

[16] M. Mehedi, A. Pritom, C. Li, B. Chu, and X. Niu, "A Study on Log Analysis Approaches Using Sandia Dataset," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, 2017.

[17] MIT Lincoln Laboratory. (2017) DARPA'98 and DARPA'99 datasets. Accessed: 2018-03-10. [Online]. Available: https://www.ll.mit.edu/ideval/docs/index.html

[18] J. Moss. (2018) DEF CON Media Server. Accessed: 2018-03-10. [Online]. Available: https://media.defcon.org/

[19] Mu Dynamics. (2018) Pcapr. Accessed: 2018-02-20. [Online]. Available: https://pcapr.net

[20] NETRESEC. (2018) Network Forensics and Network Security Monitoring. Accessed: 2018-02-20. [Online]. Available: https://www.netresec.com

[21] M. Peacock, *Creating Development Environments with Vagrant*. Packt Publishing, 2013.

[22] S. Renfro and B. Guyton, *mergecap*, accessed: 2018-03-14. [Online]. Available: https://www.wireshark.org/docs/man-pages/mergecap.html

[23] M. Ring, S. Wunderlich, D. Grüdl, D. Landes, and A. Hotho, "Flow-based benchmark data sets for intrusion detection," in *Proceedings of the 16th European Conference on Cyber Warfare and Security*, 2017, pp. 361–369.

[24] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 5, pp. 516–524, sep 2010.

[25] M. Tuexen, F. Risso, J. Bongert, G. Combs, and G. Harris. (2018) PCAP Next Generation (pcapng) Capture File Format. Accessed: 2018-03-11. [Online]. Available: https://github.com/pcapng/pcapng

[26] M. J. M. Turcotte, A. D. Kent, and C. Hash, "Unified Host and Network Data Set," *CoRR*, vol. abs/1708.07518, 2017. [Online]. Available: http://arxiv.org/abs/1708.07518

[27] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "OpenML: Networked Science in Machine Learning," *SIGKDD Explorations*, vol. 15, no. 2, pp. 49–60, 2013. [Online]. Available: https://www.openml.org/

[28] E. Vasilomanolakis, C. G. Cordero, N. Milanov, and M. Muhlhauser, "Towards the creation of synthetic, yet realistic, intrusion detection datasets," in *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*. IEEE, apr 2016, pp. 1209–1214.

[29] J. Vykopal, R. Ošlejšek, P. Čeleda, M. Vizváry, and D. Tovarňák, "KYPO Cyber Range: Design and Use Cases," in *Proceedings of the 12th International Conference on Software Technologies - Volume 1: ICSOFT*. Madrid, Spain: SciTePress, 2017, pp. 310–321.

[30] Y. Yannikos, L. Graner, M. Steinebach, and C. Winter, "Data Corpora for Digital Forensics Education and Research," in *Advances in Digital Forensics X*, G. Peterson and S. Shenoi, Eds., vol. 433. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 309–325.

[31] W. Yurcik, C. Woolam, G. Hellings, L. Khan, and B. Thuraisingham, "Toward Trusted Sharing of Network Packet Traces Using Anonymization : Single-Field Privacy / Analysis Tradeoffs," *Design*, 2007. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/0710/0710.3979.pdf