EuroVis Workshop on Visual Analytics (2018)
C. Tominski and T. von Landesberger (Editors)

Towards Visual Cyber Security Analytics for the Masses

Alex Ulmer[1], Marija Schufrin[1], Hendrik Lücke-Tieke[1], Clindo Devassy Kannanayikkal[1,2] Jörn Kohlhammer[1,2]

[1]Fraunhofer IGD, Germany
[2]TU Darmstadt, Germany

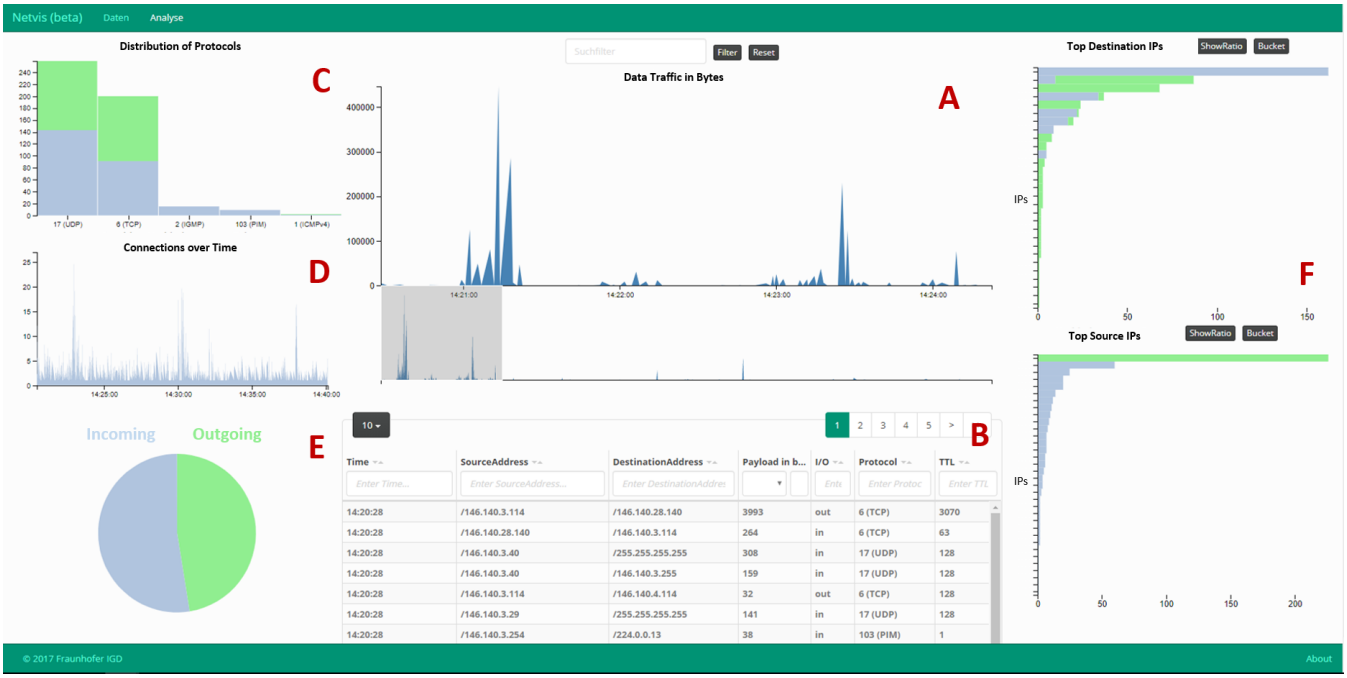

Figure 1: *Visualization of network traffic PCAP data. In (A) the user can observe the distribution of bytes over time and interactively focus on a time period. In (B) the raw data is listed for deeper investigation. (C) provides the distribution of protocols, (D) the amount of overall connections over time and (E) the relation between incoming and outgoing connections. Finally, in (F) the distribution of connections per destination and per source can be explored.*

Abstract

Understanding network activity and cyber threats is of major concern these days, for business and private users alike. As more and more online applications assist us in our daily life, there is a growing potential vulnerability to cyber crime. With this paper, we want to share our vision of cyber security analytics becoming an accessible everyday task through visual analysis tools. We describe the context of this vision and our experience with the first achievements in this direction. With our new prototype, anyone can analyze their network traffic logs and get security-relevant information out of it, a task that was too difficult and sometimes too expensive in the past. We present an open, accessible and user-friendly visual network analyzer for PCAP (packet capture) files, critically discuss our first prototype, and give an outlook to anomaly detection supported by active learning in this context.

CCS Concepts

•***Human-centered computing*** → ***Visualization systems and tools;*** *User centered design;* •***Applied computing*** → *Network forensics;*

DOI: 10.2312/eurova.20181113

1. Introduction

We believe that being aware of what is going on in one's own network will become as common and easy as using an email system. With this work, we are at the start towards this vision. We investigate how an interface for visual network analysis should be designed to address novel users with low domain expertise. To achieve this, it is important to find people that have no expertise in the field of network packet analysis but are very interested to gain access to it. Small and medium sized enterprises (SMEs) as well as public institutions and communities are representatives of this targeted user group, because they typically do not have a high budget for cyber security. Usually, general IT staff or non-IT workers are assigned to watch over the network security. These entities are increasingly affected by cyber attacks, at least in Europe, but do not have the ressources for security operation centers (SOCs) or highly specialized staff to monitor network traffic.

Recent events [ENS17] have proven again that cyber attacks cannot be fully prevented by preemptive measures. Monitoring network traffic became a complementing standard procedure for critical infrastructures to raise situation awareness and enable fast countermeasures [KA13]. Due to the large amount of monitoring data that is generated even in small networks, this task is highly demanding for security personnel. Accordingly, there are many tools and approaches to analyze log files and other sensor data, either automatically or combined with visualization. However, many techniques to process and visualize network security data [GFS*16], [Kiz17] are quite complex solutions that require considerable training and background knowledge. In the future, more cyber attacks will probably target entities with less expertise or even private households. Therefore, it is necessary to facilitate access for these users so that they can effectively inspect and analyze their own network traffic.

This effort is currently being funded by a state government in Germany, addressing the need of public communities and small companies to respond to a growing cyber security threat. Within this effort, we created a light-weight web service which accepts the common and free available PCAP (**p**acket **cap**ture) log format and lets the user interact with the data as easy as possible. The tool lets the user retrospectively analyze the network traffic to answer actual security questions. The contributions of this paper are:

1. An approach to visual cyber security analytics for general users with a description of the target group's requirements.
2. A first prototype, which is freely accessible online, with a discussion of the limitations of the initial development.
3. First steps towards an interactive active-learning approach to security analytics, avoiding the problems of anomaly detection by employing a visual analytics approach.

2. Related Work

We distinguish between two groups of related work: approaches that provide security dashboards, and approaches that support cyber analytics to visually investigate anomalies. The first group uses similar techniques to our approach and the second group has a similar general goal.

2.1. Security Dashboards

Multiple visualization dashboards were created in the past featuring a very high information density, combining advanced visualization techniques like Radviz with stream graphs and matrix views [ZLF*14, ZHZ*15] or clock views [PSZ08]. Other examples use parallel coordinate plots with dimensionality-reduced scatter plots [Leg15] to maximize the information density. In many of these approaches, the goal is to provide enhanced situational awareness. Pike et al., for instance, focus on providing a global context to the user [PSZ08]. They employ a world map displaying an overview of the current threat state. Users then drill down to individual countries to analyse the temporal patterns of threat actors over the last 24 hours. Ocelot's main goal (by Arendt et al.) is to support dynamic network management and defense [ABB*15]. One core capability of Ocelot is the ability to put systems into quarantine, i.e. dynamically reconfiguring the network to stop communication from/to quarantined systems. McKenna presented the results of a design study, the BubbleNet dashboard, to improve communication between various stakeholders about what they call "patterns" - recurring or anomal network behavior that can either be benign or malicious [MSFM16]. McKenna's central element is a world map - a design choice that makes sense for large, multinational corporations, but does not fit our target scope. All these techniques have in common that they require a substantial amount of training. We follow a user-centered design approach similar to McKenna et al. [MSM15] who showed that key questions from cyber analysts, network managers, IT directors and CEOs are very different. However, our focus is on a group typically omitted in such analyses: general users with IT experience but nearly no domain expertise. We agree with McKenna that visual representations that require significant explanations should be avoided for non-experts. In an exacerbated situation with a low budget and missing expertise, our users need a solution that is light-weight and walk-up usable.

2.2. Visual traffic anomaly investigation

Anomaly detection is a widely used concept in cyber security, strongly suffering from false positives, bombarding security personnel with extensive log files full of non-critical alert messages. Once a presumed anomaly has been spotted or an event has been identified to be worth investigating, low-level traffic analysis begins. While domain experts might stick with their existing tools (e.g. [Wir18]) and perform a bottom-up analysis, general IT personnel prefer a top-down approach [PHWC10]. In either case, showing the context is important. SNAPS [CW15] enables users to analyze anomalies with a radar view, which shows traffic data in a tabular manner, using cell lightness as the central indicator of relevance. While SNAPS also allows a temporal overview similar to our approach, it is strongly tailored towards domain experts who are trained to spot the relevant cues in hundreds of attributes. Goodall et al. designed TNV (Time-based Network Visualizer) to facilitate the analysis of packet-level data without the loss of context [GLRK05]. By applying Focus+Context to its main component, a matrix view across host IP addresses over time, they are able to depict an extended context. Some malware variants exhibit periodic traffic patterns, that are of interest for security personnel [HNUK16]. However, approaches that are directed at specific patterns or emphasize

the relevance of this indicator, are usually not easily understandable for general users. In essence, we see a gap in the literature of visual cyber security analytics, where effective approaches need to be highly accessible to general IT personnel.

3. Approach

Our approach adheres to the data-user-task design triangle [MA14] combined with a persona-driven approach [SMG*08], [MK11]. The main challenge is to identify how to correctly address general users in a cyber security context with the appropriate degree of complexity. Non-experts should not be overwhelmed by the amount of data and information, but should see the state of the network in sufficient detail. The study by Lee et al. [LKK17] contains further details on the use of appropriate visualizations for non-experts.

3.1. User

By targeting visual cyber security analytics for the masses we aim at a tool that is accessible for most people with a general IT experience, similar to the focus of ManyEyes [VWVH*07] that helped coin the term "InfoVis for the masses" in 2007. To support our specific domain, we also have to look more closely at our potential users. Classifying potential users by their experience with cyber-security and with visualization, we certainly find most of our target users in a group with a low expertise in cyber security and a medium visual literacy. Indeed, from our experience, only people with at least a slight interest in IT will use cyber security tools. In a company without IT experts the employee with the closest connection to IT is often given the responsibility to monitor the network. That is why SMEs and public institutions are very good representatives of such general users and the users of our approach.

Our observations were mainly made in a set of interviews with two different user groups. We interviewed IT admins from two different regional municipalities to gather hands-on feedback on their current security processes and their background in cyber security. We also interviewed two cyber analysts from a large malware analysis company and the federal office for information security in Germany. The first interviews showed that there are almost no tools available to them to reliably and accessibly assess the security state in their network. The second two interviews where necessary to determine how experts are actually working with their data sources to trace anomalies in the network. They both stated that the raw network packet data in form of PCAP files is widely used for analysis.

Based on these interviews we created personas representing two groups of users of our tool. In short, the main differences between the two personas are their motivation and goals. One persona represents the user group that is intrinsically interested in analyzing their own network to explain certain anomalies, even without a strong set of tools and experience. The other persona represents users that are rather forced to examine the company's network traffic as the most suitable employees for this task. Our approach supports both groups of novices to cyber-security monitoring.

3.2. Task and Requirement Analysis

The results of the user analysis lead to the following tasks and requirements. Although, our personas have different scenarios they have a similar goal. They want to have a retrospective overview of their network traffic to find anomalous behavior or the cause of such. Therefore, their task is to find the necessary information in large log files and interpret it so they can reasonably identify malicious events. Based on this task and the surrounding scenario, we describe the following requirements for our analytics tool:

1. The analysis should be as time-efficient as possible, thus a fast and intuitive navigation through the data is important.
2. To a novice in the field of cyber security, the tool should only display the most important information to not overwhelm, but still be useful.
3. The software has to be easily accessible through the web without time-consuming installations.

3.3. Data

Although packet captures offer various network data attributes, in our first iteration we decided to provide only the most important, yet generally understandable information to get a basic overview of the network. While it is likely that the provided information will not provide deeper insights, the chosen information selection is appropriate for a first level of situation awareness [End11]:

1. Amount of traffic over time: One of the most frequently used warning signals for suspicious traffic is a notably high amount of data transferred at one specific point of time. This kind of anomaly can signal a DDoS attack or other malicious activity.
2. Connections over time: Partially high amounts of connections can indicate port scans or other malicious activity.
3. Recorded destination addresses and their activity: The information about how often a destination address is targeted can reveal the most relevant player and machines in the network. Frequently targeted unknown addresses are suspicious.
4. Recorded source addresses and their activity: Similar to destination addresses, frequently occurring unknown source addresses should be further investigated.
5. Used protocols: Detecting protocols not typically used (e.g. SCTP or IGMP) by the own enterprise can be a symptom of an undesirable communication.
6. Incoming and outgoing traffic: Extraordinary relations that do not reflect the expected communication behaviour within the network can be a signal for unwanted network activities. Furthermore, these attributes are interesting to filter by in subsequent analysis steps.
7. Raw data: To enable the maximum possible trust and traceability, the raw data should be accessible in detail.

All of these information sources are especially useful if there is the possibility to observe them in combination.

3.4. Visual Prototype

3.4.1. Exploration Module

There are six linked visualizations in our prototype (see Figure 1), which are connected via brushing & linking. For choosing the visual mappings for the data we used insights from various publications [LKK17], [SK16]. Following Shneiderman's mantra [Shn03], the user can select the time span of interest from the overview in

the center and thereby filter the observed data, using animated updates to highlight the changes. Additionally, a search bar enables the user to filter specific host names or IP addresses. The whole interface decodes the incoming and outgoing connections by using two different colors (green for outgoing and blue for incoming). This enables the user to more quickly determine whether irregular activity is caused by outside sources or by internal hosts.

While we would like to provide more sophisticated visualizations in our interface, the experience in this project has shown that the general user has difficulties grasping such advanced visualizations. Because our approach targets users with low expertise in cyber security and information visualization, we decided to first provide the most common visualization techniques.

3.4.2. Analytics Module

This is also true for the advanced analytical capabilities of the current web-based prototype that we are developing with Python 3.6, JavaScript and elasticsearch [ela18] to support the user in finding anomalies. We want to guide the user's exploration of automatically detected irregularities. Based on a semi-supervised machine learning technique the system learns how normal traffic looks like. Then, the algorithm predicts, which part of the traffic is anomalous by calculating a confidence value. The interface of the prediction system is seen in Figure 2. The user can again select a time frame and drill down to the respective netflows, which are rated by the algorithm to show anomalies with high and low confidence on the one side, and normal traffic on the other. Then, she can select these flows and see, which IPs and protocols were most frequently used.

By selecting protocols or IPs the data can be further filtered to specific data that is for example related to an attack. Finally, these netflows are displayed in a raw data table with the ability to change the classification. The user's feedback re-trains a Gaussian mixture model and a k-means clustering to adapt to the specific user's network context. The initial training of the models is only done with positive data, meaning without malicious traffic. Our first experiments showed that this is better than training on positive and negative data, which is in line with the results of [SPNS16]. In this way the tool is easier to deploy in networks, as it only needs regular traffic to learn its initial model. Further requirements have to be met for this functionality to be seamlessly integrated to the web service, and is a task for future work.

4. Discussion and Future Work

In this paper, we have described the current state of the design process concerning our web-based tool for retrospective examination of network packet data. To date our first prototype has already been reviewed in hallway testing and cognitive walk-throughs and we collected some feedback from public institutions and SMEs. In the following, we critically reflect on the response to the current state of our prototype. On the positive side, the tool and its purpose has been well accepted and appreciated. The feedback revealed that the tool has the ability to increase the efficiency in the analysis of network data. The high amount of information hidden in the network data is broken down to the most important characteristics. The tool enables the user to easily navigate through the data and find critical spots. Furthermore, the accessibility of the tool through a web

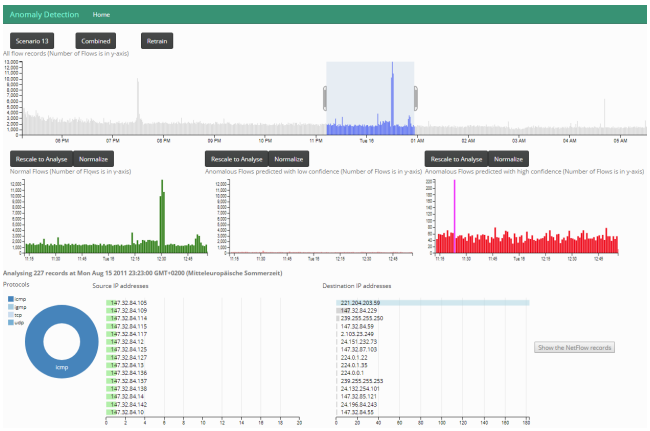

Figure 2: *Analytics module for anomaly detection in network data. (A) shows the total netflows over time, while (B) depicts the prediction of normal traffic (left), anomalous traffic with low confidence (center), and anomalous traffic with high confidence (right). (C) provides visual filters for IPs and protocols.*

interface allows the user to upload own network data, which turned out to be a valuable aspect of the tool. In our investigation of tools with a similar goals and techniques, we observed that one of the widespread barriers is the uncomfortable access and the modest availability of such tools. This is a high initial hurdle for novice users. Therefore, we strongly recommend and lead by example to provide an open web access for cyber analytics tools.

Nevertheless, some shortcomings have become apparent as well. To provide an easy understanding of the visualizations for the whole target group, we were forced to choose very simple charts in our first prototype. While the feedback concerning the visualizations varied from too trivial to too complex, the table with the raw network data was often seen as overwhelming. The degree between seeming trivial and losing the attention of the user is very narrow, if there is less prior knowledge. Therefore, our intent for the next steps is a careful calibration of the visual complexity closely tied to a feedback-loop with the user. Further, we have been confronted with the inability to combine and reuse previously uploaded log files and will therefore add a data management feature. This also corresponds to our analytics module, where multiple log files are needed to train the detection algorithm.

Our visual prototype is currently provided to local SMEs and public communities, deployed in their local networks for confidentiality reasons. However, the tool can also be accessed online [Fra18] and used with any PCAP file. Our future work will focus on extending this service to support larger amounts of data and making the visualizations customizable as proposed by Méndez et al. [MHN17]. Finally, we will improve our semi-supervised anomaly detection by optimizing the active learning features and integrate it into the current web service.

5. Acknowledgments

This work is funded by the Hessian Ministry of the Interior and Sports (HMdIS) within the "Round Table Cybersecurity@Hessen".

References

[ABB*15] ARENDT D. L., BURTNER R., BEST D. M., BOS N. D., GERSH J. R., PIATKO C. D., PAUL C. L.: Ocelot: user-centered design of a decision support visualization for network quarantine. In *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)* (Oct. 2015), pp. 1–8. 00014. doi:10.1109/VIZSEC.2015.7312763. 2

[CW15] CAPPERS B. C. M., WIJK J. J. V.: SNAPS: Semantic network traffic analysis through projection and selection. In *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)* (Oct. 2015), pp. 1–8. 00008. doi:10.1109/VIZSEC.2015.7312768. 2

[ela18] ELASTIC: The elasticsearch database, 2018. URL: https://www.elastic.co/products/elasticsearch. 4

[End11] ENDSLEY M. R.: *Designing for Situation Awareness: An Approach to User-Centered Design, Second Edition*, 2nd ed. CRC Press, Inc., Boca Raton, FL, USA, 2011. 3

[ENS17] ENSIA: European union agency for network and infromation security - threat landscape report, 2017. URL: https://www.enisa.europa.eu/publications/enisa-threat-landscape-report-2017. 2

[Fra18] FRAUNHOFER IGD: Netvis webpage. http://netvis.igd.fraunhofer.de/, 2018. Accessed: 2018-02-28. 4

[GFS*16] GUIMARAES V. T., FREITAS C. M. D. S., SADRE R., TAROUCO L. M. R., GRANVILLE L. Z.: A survey on information visualization for network and service management. *IEEE Communications Surveys & Tutorials 18*, 1 (2016), 285–323. 2

[GLRK05] GOODALL J. R., LUTTERS W. G., RHEINGANS P., KOMLODI A.: Preserving the Big Picture: Visual Network Traffic Analysis with TN. In *IEEE Workshop on Visualization for Computer Security 2005 (VizSEC 05)(VIZSEC)* (2005), p. 6. 00000. URL: doi.ieeecomputersociety.org/10.1109/VIZSEC.2005.17, doi:10.1109/VIZSEC.2005.17. 2

[HNUK16] HUYNH N. A., NG W. K., ULMER A., KOHLHAMMER J.: Uncovering periodic network signals of cyber attacks. In *2016 IEEE Symposium on Visualization for Cyber Security (VizSec)* (Oct. 2016), IEEE Computer Society, pp. 1–8. 00001. doi:10.1109/VIZSEC.2016.7739581. 2

[KA13] KOTT A., ARNOLD C.: The promises and challenges of continuous monitoring and risk scoring. *IEEE Security & Privacy 11*, 1 (2013), 90–93. 2

[Kiz17] KIZZA J. M.: *Guide to computer network security*. Springer, 2017. 2

[Leg15] LEGG P. A.: Visualizing the insider threat: challenges and tools for identifying malicious user activity. In *2015 IEEE Symposium on Visualization for Cyber Security, VizSec 2015, Chicago, IL, USA, October 25, 2015* (2015), Harrison L., Prigent N., Engle S., Best D. M., (Eds.), IEEE Computer Society, pp. 1–7. 00013. doi:10.1109/VIZSEC.2015.7312772. 2

[LKK17] LEE S., KIM S.-H., KWON B. C.: Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 551–560. 3

[MA14] MIKSCH S., AIGNER W.: A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics 38* (2014), 286–290. 3

[MHN17] MÉNDEZ G. G., HINRICHS U., NACENTA M. A.: Bottom-up vs. top-down: Trade-offs in efficiency, understanding, freedom and creativity with infovis tools. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 841–852. 4

[MK11] MIASKIEWICZ T., KOZAR K. A.: Personas and user-centered design: How can personas benefit product design processes? *Design Studies 32*, 5 (2011), 417–430. 3

[MSFM16] MCKENNA S., STAHELI D., FULCHER C., MEYER M. D.: BubbleNet: A Cyber Security Dashboard for Visualizing Patterns. *Computer Graphics Forum 35*, 3 (Jan. 2016), 281–290. 00001. doi:10.1111/cgf.12904. 2

[MSM15] MCKENNA S., STAHELI D., MEYER M.: Unlocking user-centered design methods for building cyber security visualizations. In *IEEE Symp. on Visualization for Cyber Security (VizSec)* (2015), IEEE, pp. 1–8. 2

[PHWC10] PARK J., HAO M. C., WONG P. C., CHEN C. (Eds.):. *Cognitive task analysis of network analysts and managers for network situational awareness* (Jan. 2010), vol. 7530. 00000. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.845488, doi:10.1117/12.845488. 2

[PSZ08] PIKE W. A., SCHERRER C., ZABRISKIE S.: Putting Security in Context: Visual Correlation of Network Activity with Real-World Information. In *VizSEC 2007* (2008), Mathematics and Visualization, Springer, Berlin, Heidelberg, pp. 203–220. 00012. URL: https://link.springer.com/chapter/10.1007/978-3-540-78243-8_14, doi:DOI:10.1007/978-3-540-78243-8_14. 2

[Shn03] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*. Elsevier, 2003, pp. 364–371. 3

[SK16] SKAU D., KOSARA R.: Arcs, angles, or areas: individual data encodings in pie and donut charts. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 121–130. 3

[SMG*08] STOLL J., MCCOLGIN D., GREGORY M., CROW V., EDWARDS W. K.: Adapting personas for use in security visualization design. In *VizSEC 2007*. Springer, 2008, pp. 39–52. 3

[SPNS16] SAKAI T., PLESSIS M. C. D., NIU G., SUGIYAMA M.: Semi-supervised classification based on classification from positive and unlabeled data. *arXiv preprint arXiv:1605.06955* (2016). 4

[VWVH*07] VIEGAS F. B., WATTENBERG M., VAN HAM F., KRISS J., MCKEON M.: Manyeyes: a site for visualization at internet scale. *IEEE transactions on visualization and computer graphics 13*, 6 (2007). 3

[Wir18] WIRESHARK FOUNDATION: Wireshark homepage. https://www.wireshark.org/, 2018. Accessed: 2018-02-28. 2

[ZHZ*15] ZHOU F., HUANG W., ZHAO Y., SHI Y., LIANG X., FAN X.: Entvis: A visual analytic tool for entropy-based network traffic anomaly detection. *IEEE computer graphics and applications 35*, 6 (2015), 42–50. 2

[ZLF*14] ZHAO Y., LIANG X., FAN X., WANG Y., YANG M., ZHOU F.: Mvsec: multi-perspective and deductive visual analytics on heterogeneous network security data. *Journal of Visualization 17*, 3 (2014), 181–196. 2