Visual Analytics for Improving Efficiency of Network Forensics: Account Theft Investigation

Andrey Chechulin\textsuperscript{1,2}, Maxim Kolomeec\textsuperscript{1,2} and Igor Kotenko\textsuperscript{2}

\textsuperscript{1}Laboratory of Computer Security Problems, St. Petersburg Institute for Informatics and Automation (SPIIRAS), 39, 14th Liniya, Saint-Petersburg, Russia

Email: {chechulin, kolomeec, ivkote}@comsec.spb.ru

\textsuperscript{2}Saint-Petersburg ITMO University, 49, Kronverksky Prospect, St. Petersburg, Russia.

Abstract. In the paper, we propose a technique and means of visual analytics for network forensic investigation. It is assumed that experts will be able to decrease the time required for analysis and for creation of easy readable evidences, timelines and presentation for the court. Also based on an example of account theft cyber-attack investigation the technique for classification of different aspects (slices) of network traffic is proposed. The evaluation and recommendations for the technique usage are also presented.

1. Introduction
In forensics (as a branch of information security) there are as many areas as there are types of cyber attacks. It includes the hard drives analysis, RAM-dump analysis for finding malwares, and analysis of data for finding insiders and of course the analysis of network traffic. In most cases in network forensic, the investigator works with network traffic analysis. Traffic analysis is a routine task, but its analysis allows one not only to identify the cause of the security breach, but also step-by-step reconstruct a whole picture of what happened: identify the probable attacker (or at least the starting point); discover the sequence of attack actions and the list of techniques that attacker used, determine the level of his (her) knowledge and capabilities; determine attacker’s potential intentions; and develop recommendations to close the used security holes.

For traffic analysis the investigators usually use Wireshark, Network miner and other software with the graphical interface for observing network packets with extensive sorting and filtering capabilities. Traffic analysis usually is a time-consuming task - the investigator has to look through packets to discover and connect the anomalies. To expand the capabilities of investigators and to simplify their work, we propose to use visual analytics technique and means, which we propose in this paper. Visual analytics has proven itself in network attack modeling and security monitoring [1-4]. It is difficult to find a modern network security system that does not have the basic tools for visualization. These technics can be transferred to the network forensic science. Moreover, network traffic is a well-structured set and it is easy to parse and process it. The main difficulty is that there are a lot of data and the investigator has to create data slices with important information.

In this paper, we propose a visual analytics approach for network forensic. This approach includes a technique for classifying of different aspects (slices) of traffic, and a scheme for visualization models applying for these slices. This approach allows one to significantly improve the efficiency of forensic investigation, including the rapid localization of anomalies and, that is also important, the creation of easy readable graphical evidences of computer attacks for courts.

In the related papers [5-7], visual analytics is used to solve the challenges of forensics by applying one or several visualization models. The novelty of this research lies in the proposed mutual
classification technique and appropriate visualization models. Classification allows one to determine the data type and structure. The contribution of this work is the visualization technique that allows one to create data slices and visualize them with appropriate visualization model. The technique can also suggest different visualization options for one data set, depending on the case.

The rest of the paper consists of three main parts. The second section describes the visualization technique and shows the visual models that can be useful in digital forensic. In the third section, an example of a computer crime investigation which demonstrates the proposed technique for classifying different traffic slices and a scheme for the usage of different visualization models. The fourth section contains the discussion.

2. Visualization Technique

The visualization technique can be used both for visual analysis directly in the investigation of incidents, and for forming a visual representation of situation for the report. At first, the investigator must create slices of data from the traffic which may contain information about anomalies. For each slice, the investigator should define its type in order to understand which visualization model can render it. The next step is to classify the slices in order to determine what visualization models are suitable for a particular set and to limit the number of models. After choosing a model, investigator needs to bring these slices to a format suitable for visualization in accordance with the type of slice, and visualize them. Thus, the visualization technique consists of two stages: creating and classifying the analyzed data slice, selecting an appropriate model or several visualization models to bring the slice to the format required for visualization and application of the visualization model to the data.

Visualization models are used as templates for displaying different data structures. Knowing the models allows one to choose the appropriate visualization technique, depending on the situation. Models can be divided into two large groups.

The first group is numerical data models. Each value in these data is represented by a measure - either quantitative (for example, the number of events) or categorical (for example, the type of event) [8]. The second group is not numerical data models. Unlike numerical data, not numerical have one important attribute - they contain links. These links can be expressed as a relationship, nesting or dependences.

It’s important to pay attention to the order of classification [9]. A tree is always a planar graph. The planar graph is always semi-structured, and so on from the list. The reverse way of the list is different. For example, only some of the planar graphs are trees, and so on. This indicates an important feature: visualization models for the relevant structure can always visualize structures that are higher in the list. At the same time, they cannot visualize the structures that are below.

Having determined the structure of the data, it is easy enough to choose the appropriate visualization model. Models of visualization can display a slice of only a specific format. For example, it is impossible to represent the topology of a computer network using a linear chart. Models of visualization can be: Bar Chart, Parallel Coordinates, Scatter Plot, Circle Packing, TreeMaps, Voronoi Maps, Chord Diagram and Triangle Coordinates. Let’s consider their features.

Bar Chart [10] allows one to compare the values. Usually, one axis represents categories, and the second axis represents quantitative data. Bar Chart visualizes numerical data [11]. Parallel coordinates are well suited for multidimensional data [12]. Each vertical line is the axis of one of the metrics. It does not have to be quantitative. The axis can be made categorical for some relationships [13]. Triangle coordinates [14] allows one to visualize three values expressed in percentages. Scatter Plot [10] allows us to represent the objects distribution. Circle Packing [15,16] is a simple visualization of trees. Object’s metrics are displayed using circle’s size or color. Unlike the Circle Packing, TreeMaps [17] do not have unused space. Like the Circle Packing, TreeMaps [18] show the connections by nesting, but express values of various metrics using the size of the rectangles. Voronoi Maps [19] are suitable for visualization of planar data. Each cell is represented by an object, and the tiny edges between cells are links between objects. Thick lines represent separators. Using the size and color of the cells, one can display metrics. The Chord diagram [9] is a good example of semi-structured data. It is suitable for visualizing objects in which one type of connection is expressed by a hierarchy, and the second is unstructured.
3. Illustration of the Proposed Approach Application

This section describes the legend of the attack (security breach story), the description of the investigation using Wireshark (basic solving) and the description of the investigation using visual analytics (visual analysis solving).

There was a room with 4 persons - Patrick, Lisa, John and Harry. It is becoming known that John received several spam messages from Lisa. Lisa said that she hasn’t sent these messages. The system administrator provided on computer network configuration and recorded computer network traffic in the PCAP format.

To solve the test, we had made data slices from pcap-traffic, which can contain traces of the attack. The first slice was formed for routing analysis. It was made from ARP packets and contained 2 fields: Source MAC and Source IP. The second slice was configured to analyze the network activity of each host. It was made on the basis of all packets and contained 4 fields: Source MAC, Packet Type, Packet Length, Packet Timestamp. The third file contained time windows with traffic. For each window, the percentage of packet types in the network was recorded. Let's look at the visualization of these slices.

For the first slice, we choose the model of parallel coordinates (Figure 1). The left column represents the IP-address, and the right – MAC-address. If there is a line between the elements, it means that in the slice there is information about the packets with such a pair of IP and MAC addresses. Obvious feature that one can see is MAC with eight IP addresses (right column, second from the top, light-blue rectangle on Figure 1). This host is a router that broadcasts broadcast-packets, so these seven lines are a kind of data noise in the analyzed slice. If one does not consider the data from the router, then all MAC addresses have one IP, except one. Host with MAC address 00:0c:29: bf: 22:13 (the right column, the first from bottom, the pink rectangle on Figure 1) has three IP-addresses. This behavior clearly stands out against the background of the behavior of other network hosts.

Let's look at the second slice. In it, we calculate the number of packets of the same type for each of the devices and represent it using the basic Bar Chart. For all types of packets except ARP (Figure 2) anomalies were not detected. When ARP-packets are displayed, the host with MAC-address 00:0c:29: bf: 22:13 is allocated (the first from the bottom bar on Figure 2), which sent ARP packets more than the router (the router is the second from the top bar on Figure 2).

This ARP-anomaly can be detected in the third slice. For its visualization, we used the model of triangular coordinates (Figure 3). The slice was aggregated to three values: percentage of ARP traffic, TCP traffic and all other traffic. Every few seconds of observation are represented by the individual red dot. Most of the observations are on the lower edge (they did not have any ARP traffic). When ARP traffic appears on the network, the dots are slightly climbing. These two dots on the left edge are time intervals in traffic dump that have increased number of ARP packets (20% and 30% of ARP packets in one time window). We can check the time intervals of these dots and look at them in Wireshark in details.

If one look at the pcap-file with time packets (Figure 4), when these two points were fixed, traces of ARP-spoofing will be detected.

Let's go back again to the slice number 2 to consider who related to MAC-address 00:0c:29: bf: 22:13. Let's represent each user in the form of separate Scatter Plot in which the Y-scale is the packet length, the X-scale is the timeline of the packet, color is the type of packet, point - packet. In this graph (Figure 5), one can see some features of user activity. For example, consecutive vertical lines may indicate the download of files within a single session. Consecutive horizontal lines in which packets of the same length appear at regular intervals may indicate services. But on this plot, the obvious anomaly is that Patrick (the fourth from the bottom Scatter Plot) left the network during the attack. The attacker himself 00:0c:29: bf: 22:13 (the first bottom Scatter Plot) was active only at the time when the Patrick was inactive (highlighted in red rectangles).

Consider this discover in a simpler form. Figure 6 shows Scatter Plots without packet lengths in the form of horizontal columns. On it, the sequence of actions is more obvious: the Patrick leaves the network -> the attacker appears -> the attacker leaves the network -> the Patrick appears.
Figure 1. Parallel Coordinates represents IP-MAC pairs anomaly around host 00:0c: bf: 22:13.

Figure 2. Bar Chart represents ARP traffic anomaly around host 00:0c: bf: 22:13.
Figure 3. Triangle Coordinates represents for ARP traffic.

Figure 4. Representation of the ARP-spoofing attack using Wireshark.
**Figure 5.** Scatter Plot of network traffic

**Figure 6.** Scatter Plot of network traffic without packet length

### 4. Discussion

Visual analytics allows solving two challenges at once. First, it is used as an analysis tool, which the investigator applies in his work. In this case, the investigator must first create slices of data which may contain information about anomalies. Then, he leads these slices to a format suitable for visualization, depending on the type of slice: if it is numerical data, the investigator can use tables or relational
databases; if this is not numerical data, the investigator can use JSON files or NoSQL databases. Then he uses a specific visualization model. The investigator can use different visualization models for one set, depending on the situation.

It is likely that on one visualization model the anomalies will not look obvious, but on other - not (compare Figures 5 and 6). Secondly, visualization is used as a proof, which is perfectly suited to the report, allowing one to explain the phenomena not by numbers, but in the form of a graphical image, which is much easier to understand.

Often decisions are made by people who are not experts in information security. It will be difficult to explain them basic technical aspects, especially when it is necessary to keep a lot of data in mind. Visualization simplifies this process.

Obviously, visual analytics tools cannot replace the work with traffic in Wireshark, but they do not pretend to do this. The goal of visual analytics is not to replace the work of investigators with automatic algorithms and the drawing of visual stories. The goal of visual analytics is to give powerful tools to investigators that will save time and simplify the work with data.

The presented example shows how visual analysis effectively allows one to detect an anomaly and localize an anomaly part of the traffic. However, not every anomaly in the data is a security breach. For example, in Figure 1, multiple IP-MAC connections (the right column, second on top, light-blue host) are not a security breaches, but just an ARP broadcast of packets from the router that were forgotten to be excluded from the data set (an intentional oversight on our part).

Determine a security breach can only be done with the help of Wireshark, but the investigator should still explore the detected phenomenon at a lower level - at the packet level. But visualization allows one to localize these phenomena and suggest areas in the traffic that one should look firstly. So, for example, Figure 3 demonstrates narrowing of the search area to just two points. It is therefore necessary to view just a few seconds of ARP-traffic that related to these 2 points to make sure that there was ARP spoofing on the network.

At the same time, visualization allows one to show the deviation from the template. Image in Figure 3 allows one to convince a person who does not understand computer networks that the found phenomenon is different from the trends and typical behavior of the computer network. To prove the presence of an attack using bright visual representation becomes easier.

In addition, the images obtained as a result of the analysis can be combined into one visual history. Step by step, one can designate the sequence of attacks in a graphical form. Together with the usual evidence based on the analysis of traffic packets, the visual representation of the events story will help to bring the evidence together, creating solid base for suspicion and will serve as an excellent addition to the investigation report.

5. Conclusion
In this paper, we examined how one can use visual analysis in network forensics.

The approach to the use of techniques and means of visual analytics for network forensics was proposed, a technique for classifying different sets (slices) of traffic and a scheme for using various models of visualization for these slices was demonstrated.

We examined: how to classify data slices; by what means, depending on the type of slice, it can be visualized; showed an example of using visual analytics to investigate an attack by comparing traffic analysis in Wireshark and visual analysis.

The possibilities of using visual analytics in forensic science need to be developed further. Future work will focus on simplifying data preprocessing mechanisms for visualization, creating a coherent methodology for visual historical analysis, and extending the benefits of visual analysis to areas of digital forensics other than forensics.

6. Acknowledgments
This research was partially supported by grants of RFBR (projects No. 16-29-09482 and 18-07-01488), by the budget (project No. AAAA-A16-116033110102-5), and by Government of the Russian Federation, Grant 08-08.
7. References

[1] Kotenko I 2003 Active Vulnerability Assessment of Computer Networks by Simulation of Complex Remote Attacks Proceedings of the International Conference on Computer Networks and Mobile Computing. ICCNMC-03 (IEEE Computer Society) pp 40-47

[2] Kotenko I 2005 Agent-Based Modeling and Simulation of Cyber-Warfare between Malefactors and Security Agents in Internet Proceedings of the 19th European Simulation Multiconference “Simulation in wider Europe” (ESM’05) pp 533-543

[3] Kotenko I, Ulanov A 2006 Simulation of Internet DDoS Attacks and Defense Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 4176 (Springer) pp 327-342

[4] Kotenko I, Stepashkin M, Doynikova E 2011 Security Analysis of Information Systems taking into account Social Engineering Attacks. Proceedings of the 19th Euromicro International Conference on Parallel, Distributed and network-based Processing (PDP 2011) (IEEE Computer Society) pp 611-618

[5] Promrit N, Mingkhwan A, Simcharoen S, Namvong N 2011 Multi-Dimensional Visualization for Network Forensic Analysis 7th International Conference on Networked Computing pp 68-73

[6] Krasser S, Conti G, Grizzard J, Gribschaw J, Owen H 2005 Real-time and forensic network data analysis using animated and coordinated visualization Proceedings from the Sixth Annual IEEE Systems, Man and Cybernetics (SMC) Information Assurance Workshop pp 42-49

[7] Erbacher R, Christiansen K, Sundberg A 2006 Visual network forensic techniques and processes 1st Annual Symposium on Information Assurance: Intrusion Detection and Prevention. p 72-80

[8] JPMorgan Chase & Co. 2013 Report of JPMorgan Chase & Co. Management Task Force Regarding 2012 CIO Losses [Internet] [cited 9 December 2017] Available from: http://files.shareholder.com/downloads/ONE/2272984969x0x628656/4cb574a0-0bf5-4728-9582-625e4519b5ab/Task_Force_Report.pdf

[9] Kolomeec M, Gonzalez-Granadillo G, Doynikova E, Chechulin A, Kotenko I, Debar H 2017 Choosing Models for Security Metrics Visualization Lecture Notes in Computer Science pp 75-87

[10] Marty R 2009 Applied security visualization. NJ: Addison-Wesley. Upper Saddle River.

[11] Jacobs J, Rudis B 2014 Data-driven security John Wiley & Sons, Inc. Indianapolis, Indiana.

[12] Tricaud S, Saadé P 2009 Applied parallel coordinates for logs and network traffic analysis Journal in Computer Virology vol 6 pp 1-29

[13] Choi H, Lee H 2005 PCAV: Internet Attack Visualization on Parallel Coordinates Information and Communications Security pp 454-466

[14] Whitaker R 2010 Applying information visualization to computer security applications. [Master’s thesis]. Utah State University

[15] Arendt D, Best D, Burtner R, Lyn Paul C 2016 CyberPetri at CDX 2016: Real-time network situation awareness 2016 IEEE Symposium on Visualization for Cyber Security (VizSec) pp 1-4

[16] National Security Agency Information Assurance Directorate Cyber defense exercise (CDX), February 2016 [Internet] [cited 9 October 2017] Available from: https://www.iad.gov/iad/programs/cyber-defense-exercise/

[17] Bederson B, Shneiderman B, Wattenberg M 2002 Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies ACM Transactions on Graphics vol 21(4) pp 833-854

[18] Kotenko I, Novikova E 2013 VisSecAnalyzer: A Visual Analytics Tool for Network Security Assessment Security Engineering and Intelligence Informatics pp 345-360

[19] Kolomeets M, Chechulin A, Kotenko I 2016 Visualization Model for Monitoring of ComputerNetworks Security Based on the Analogue of Voronoi Diagrams Lecture Notes in Computer Science pp 141-157