MoVA: A Visual Analytics Tool Providing Insight in the Big Mobile Network Data

Ilias Kalamaras, Stavros Papadopoulos, Anastasios Drosou, Dimitrios Tzovaras

MoVA: A Visual Analytics tool providing insight in the Big Mobile Network Data

Ilias Kalamaras[1,2], Stavros Papadopoulos[1,2], Anastasios Drosou[2], and Dimitrios Tzovaras[2]

[1]Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, London, UK
{i.kalamaras11,s.papadopoulos11}@imperial.ac.uk
[2]Information Technologies Institute, Centre for Research and Technology Hellas, P.O. Box 361, 57001 Thermi-Thessaloniki, Greece,
{kalamar,spap,drosou,tzovaras}@iti.gr

Abstract. Mobile networks have numerous exploitable vulnerabilities that enable malicious individuals to launch Denial of Service (DoS) attacks and affect network security and performance. The efficient detection and attribution of these anomalies are of major importance to the mobile network operators, especially since there is a vast amount of information collected, which renders the problem as a Big Data problem. Previous approaches focus on either anomaly detection methods, or visualization methods separately. In addition, they utilize solely either the signaling or the Call Detail Record (CDR) activity in the network. This paper presents MoVA (Mobile network Visual Analytics), a visual analytics tool for the detection and attribution of anomalies in mobile cellular networks which combines anomaly detection and visualization, and is applied on both signaling and CDR activity in the network. In order to address the large volume of the data, the proposed application starts with an aggregated overview of the whole network and allows the operator to gradually focus on smaller sets of data, using different levels of abstraction. The proposed visualization methods are able to differentiate between different user behaviors, and enable the analyst to have an insight in the mobile network operation and easily spot the anomalous mobile devices. Hypothesis formulation and validation methods are also provided, in order to enable the analyst to formulate network security-related hypotheses, and validate or reject them based on the results of the analysis.

Keywords. Mobile network security, big data, visual analytics, information visualization, anomaly detection, hypothesis formulation and testing.

1 Introduction

This work has been partially supported by the European Commission through project FP7-ICT-317888-NEMESYS funded by the 7th framework program. The

With a steadily increasing amount of user devices, the problem of mobile network security, i.e. monitoring a mobile phone network and identifying abnormal and malicious behavior, is nowadays becoming even more challenging. The number of connected mobile devices is expected to increase even more in the next years, including diverse types of data, such as those originating from Internet of Things (IoT) devices. The vast number of mobile phone subscribers, communicating every day, results in a huge amount of signaling and billing records, containing multiple and diverse types of information. This constant flow of information from multiple sources renders the problem of mobile network security as a *Big Data* problem, posing the challenge of how to reduce the amount of information and focus on the useful aspects.

Visual analytics techniques can significantly assist in this direction. Visual analytics tools can assist the mobile network operator to have an overview of various aspects of the whole network, while allowing her/him to explore and focus on gradually smaller subsets of the data, until the desired information is reached and a decision is made.

This paper presents MoVA (Mobile network Visual Analytics), a visual analytics application for the visualization and exploration of mobile phone networks. The application has been developed for the purposes of the NEMESYS project [1], and is designed with a focus to detecting anomalies in the mobile traffic, such as Denial of Service (DoS) attacks. The application consists of various tools, which allow the operator to have an overview of the network, to view similarities in the behavior of the mobile phone users and to formulate and validate various hypotheses regarding the mobile network data. To accomplish these tasks, the tools utilize graph-based visualization techniques as well as anomaly detection algorithms, based on signaling and Call Detail Record (CDR) data of the network. The complete application allows the mobile network operator to explore the mobile network and gradually focus on specific subsets of the initial large amounts of data, in order to finally detect anomalous behavior and trace its origins. The proposed application is designed in order to handle Big Data in a time-efficient manner, starting with a large-scale overview of the network and allowing the operator to gradually focus on the desired information and acquiring useful insights. The proposed application addresses modern challenges regarding the research about Big Data, as well as practical issues posed by telecommunication and mobile network security companies.

The rest of the paper is organized as follows. Related literature is presented in Section 2. The proposed visual analytics application and the tools comprising it are described in Section 3. Experimental results for the evaluation of the application are presented in Section 4, while Section 5 concludes the paper.

---

opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission.

2 Related work

The mobile network is comprised of millions of components and mobile devices, which exchange information at a large rate. The data collected from mobile networks are comprised of two types of information: 1) Signaling activity, and 2) CDR activity. The signaling activity is the control plane, and represents all the messages exchanged in order to activate/deactivate network services and establish connections. The CDR data represent the data plane, and contain information such as [2]: source, destination, time, duration, call type (e.g. voice, SMS, MMS etc). These two types of information contain the necessary information in order to detect anomalies in the network traffic, without using communication content and sacrificing user privacy. Signaling and CDR data are collected for each mobile subscriber, and for an extended period of time. This section presents the methods that are used for anomaly detection using these types of information.

With respect to anomaly detection using signaling data, most of the proposed approaches utilize this type of information in a aggregated format. Specifically, similarly to this paper, each signalling message is daggered over specific time periods, where the value in each time period represents the number of signals sent to the network. Using this type of information, Lee et al. [3] [4] utilized a cumulative sum (CUSUM) method for the identification of signaling attacks that the traditional detection systems cannot detect. The authors designed their method in such a way, so that it is difficult for the attackers to evade detection. They evaluated their approach on a novel DoS attack that affects the RNC and the Node-B in 3G and potentially WiMax networks. Alconzo et al. [5] proposed the use of statistical techniques applied on time-series of unidirectional feature distributions. Coluccia et al. [6] proposed two distribution-based anomaly detection methods and provided enhancements on the method introduced in [5].

With respect to anomaly detection using CDR data, Yan et al. [7] proposed SMS-Watchdog, a scheme that utilizes SMS for anomaly detection. Multiple information theoretic anomaly metrics have been used, in order to identify large deviations from the normal past behaviors. Kim et al. [8] utilized multiple statistical metrics for the identification of the mobile devices participating in an SMS-flooding attack, and achieved better results than SMS-Watchdog. Murynets et al. [9] proposed two algorithms for the identification of anomalous SMS activities on mobile networks at different levels of abstraction, including aggregate, cluster, and individual device levels.

Apart from the analytical methods for anomaly detection in mobile networks, visualization-based methods have also been proposed in the literature. Eagle et al. [10] utilized entropy measures to characterize the CDR activity of each user. Afterwards, the authors utilized 2-dimensional plots to represent the change of entropy over time. Ye et al. [11] proposed a graph-based approach, for the visualization of CDR data. The vertices of the graph represent users and the edges call communication events. Shen et al. [12] proposed mobivis, another graph-based visualization approach for the analysis of CDR data. The vertices of the graph represent ontologies (e.g. users, locations), and the edges time dependent con-

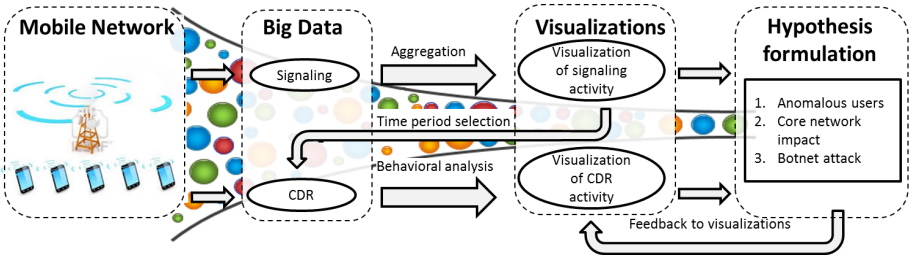

Fig. 1. The proposed data pipeline followed for the analysis of Big Data collected from mobile cellular networks.

nections between ontologies. The authors also proposed circular histograms as well as 2-dimensional histograms for visualizing the activity of each user/group over time.

There are also commercial toolkits that offer visual analytics solutions and can be used by mobile network providers. SAS VA [13] and Tableau [14] are widely-used visual analytics applications which are able to visualize any kind of multi-dimensional data, using a variety of visualization techniques. However, being generic in nature, they are not targeted to mobile network data, losing much of their effectiveness in related scenarios. Certifuge Systems, [15], another visual analytics application, is also generic in nature, but can also handle IP network data. However, it is still not focused on mobile network data.

Existing solutions that target mobile network security include tools that perform anomaly detection, but with no support from visualizations. Motive security guardian [16] was proposed by Alcatel for monitoring and analysis of signaling-related data. This method can help network providers in pinpointing malware infections accurately and take action in real time. Thanks to the power of Motive security guardian network traffic analysis, the network operators can gain actionable insights into security threats and how to minimize their impact on the network and their subscribers. NSN mobile guard [17] is a complete solution of network security proposed by Nokia. It provides anomaly detection methods that utilize signaling data, as well as an antivirus suite installed on mobile devices and protects subscribers against fraud. NSN mobile guard allows the network operators to see the security status of the devices on the network, while also enabling them to provide proactive support for infected subscribers. However, as already mentioned, both of the above tools do not provide visual analytics techniques, which would allow the operator to easily explore the available data. The proposed application aims at filling this gap in the existing applications, by utilizing visual analytics techniques, in conjunction with anomaly detection methods, in order to provide solutions for mobile network security.

3 Description of the proposed methods

As already mentioned, most of the previous work for the analysis of data collected in the mobile network, for the purposes of security, are very limited, and do not address Big Data. In addition, the majority of the aforementioned approaches focus only on signaling or CDR data, while none of these approaches combines visualization and anomaly detection into a common framework. The hereby proposed MoVA application aims at handling these limitations.

The proposed approach begins with an overview of the whole mobile network and provides different levels of abstraction, allowing the operator to gradually focus on smaller sets of data, which can be further explored and allow the detection and attribution of anomalies. A schematic representations of the proposed data pipeline is presented in Figure 1. Signaling and CDR data are collected from the mobile network. The signaling data are first visualized in an aggregated manner, which results in the selection of time periods which contain an anomaly. The CDR data of the selected time periods are afterwards used to extract multiple behavioral characteristics of the users and visualize their similarities. The network analyst can finally formulate attack related hypotheses, and validate or reject them using the results of the analysis and the visualization methods. Screenshots of MoVA are illustrated in Figure 2.

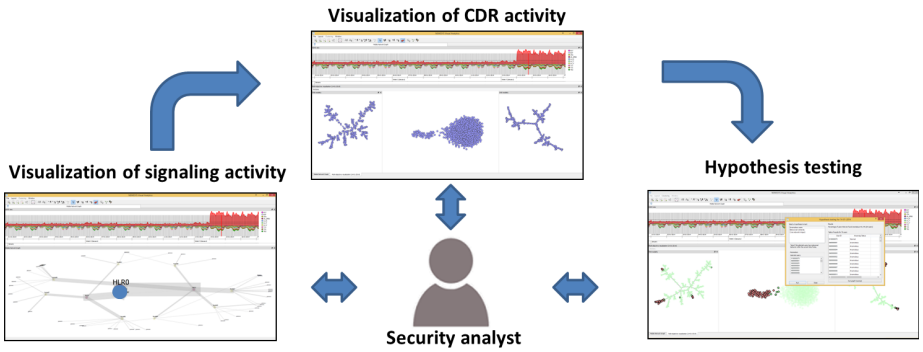

Fig. 2. Overview of the proposed visual analytics tools integrated in MoVA. An overview of the network is provided to the analyst by visualizing aggregated signaling activity. Reducing the abstraction level, CDR data of specific users are used to visualize their behavioral similarities. Finally, the analyst can formulate and test specific security-related hypotheses.

3.1 Visualization of the signaling activity in the network

The visual analytics methods presented in this section deal with the **visualization of the signaling activity** and the anomaly status over a long period of time. These methods provide an overview of the current and past network status

to the analyst, and enable the selection of appropriate subsets of data which will be analyzed in depth with the rest of the methods integrated in the system.

As noted earlier, the signaling data are Big Data, collected from millions of devices and over an extended period of time. In order to address this issue, the total signaling activity in the network is **aggregated** over short periods (e.g. per hour). This aggregation results in multiple time series, one for each type of signal in the network, while each entry of the time series represents the total amount of signals sent for the specific time period. This is represented as a matrix $S = \{s_{ij} | i \in 1 \dots K,\ j \in 1 \dots N\}$, where N is the total number of time periods, K is the total number of signals in the network, and s_{ij} is the total number of signals belonging to the i_{th} category, and sent to the network at the j_{th} time period.

Thereafter, anomaly detection methods are applied on matrix S, in order to provide an overview of the anomaly status over the different time periods. Specifically, the Local Outlier Factor (LOF) [18] is utilized on the multidimensional space defined by S, in which the position of each point are defined as $s_j = [s_{1j}, s_{2j}, ..., s_{Kj}]$. The result of the LOF analysis is an anomaly score for each time period $j \in \{1, 2, .., N\}$.

The signaling activity represented by matrix S, and the LOF scores are subsequently visualized using stacked graph representations, as shown in Figure 3(a). The lower part of this figure visualizes the normalized volume of the signaling activity for each signal, over the different time periods. The upper part shows the LOF scores for each time period. An example of anomaly is manually highlighted in this figure, concerning a sudden increase of the Location Update Total (LU_T) signal.

Figure 3(b) illustrates the visualization method that is utilized for presenting the aggregated signaling activity of a subset of users in the network. This subeset represents the top k mobile devices with the highest activity. The specific time period is selected using the stacked graphs representation shown in Figure 3(a). As shown in Figure 3(b), the components of the network and the mobile devices connected to them are positioned in different layers, namely, 1) UE (User Equipment): the actual mobile devices, 2) NodeB: the NodeBs that provides the mobile devices with wireless connectivity, 3) RNC: the RNCs (Radio Network Controller)which are responsible for controlling the NodeBs that are connected to them, and 4) Core Network: the componets of the core network.

The visualization of Figure 3(b) shows the distribution of the signaling messages to the different network components. This allows the mobile network operator to identify the most active users, as well as their activity distribution to the network. More details on the method presented in this section for the visualization of the signaling activity in the network can be found in [19].

3.2 Visualization of common user behaviors

As a second stage of analysis, after **focusing on a specific time period**, the Multi-objective **behavior visualization** tool is provided to the operator, which visualizes groups of users with common behavior, utilizing multiple channels of

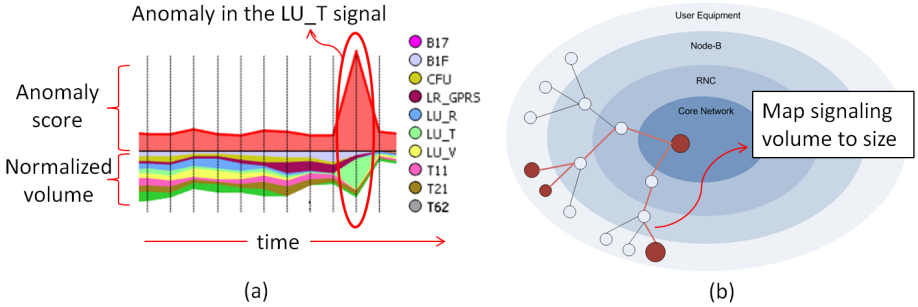

Fig. 3. Visualization of the signaling activity in the network. (a) Stacked graphs visualization of the anomaly scores and signaling volume. (b) A scheme of the proposed layered layout of the Mobile Network Graph.

CDR information. During mobile phone usage, multiple and diverse types of information are stored, such as the time of a call, the identity of the user starting the call and the recipient etc. Each of these attributes can be used to extract behavioral descriptors from the users, encoding certain characteristics of their behavior.

The descriptors are extracted by considering each attribute separately and constructing histograms of their values for each user, for the time period in consideration. These histograms are indicators of the behaviors of users so that a user with normal behavior has a histogram with a much different form than a user with abnormal behavior. Two types of descriptors are used, namely the *Time Histogram Descriptor (THD)*, which is formed from the hours in a day that a user communicates with other users, and the *Recipient Histogram Descriptor (RHD)*, which is formed from the IDs of the recipients of the user's communications.

The multiple descriptors extracted from a user can be used to visualize the similarities between users, with respect to the multiple recorded attributes. Although each descriptor could be used separately to visualize a specific behavioral aspect, considering all descriptors simultaneously allows the operator to view multiple aspects at once, and to have a more complete insight into the data, such that could not be gained by using each descriptor separately. For the purposes of the mobile network behavior visualization, the multi-objective visualization of [20] has been used. Each descriptor is used to calculate unimodal distances between users. Subsequently, a distance graph is formed, where each user is considered as a vertex and there are edges between each pair of users, weighted by the distance between them. The Minimum Spanning Tree of the graph is extracted and is used to define a force-directed-based objective function [20]. The minimization of this objective leads to a placement of the vertices on the two-dimensional plane, such that the connected vertices are kept close to each other.

Although large graph analysis is generally a computationally expensive task, the proposed approach deals with it effectively, in two ways: First, the overall

top-down architecture of the approach is meant to use less complex methods, such as the stacked plots and the Mobile Network Graph on large amounts of data, and then to focus on smaller datasets, so that the more complex multi-objective method is run with few computational resources required. Second, by using the Minimum Spanning Tree, instead of the full graph or a neighborhood graph, only a few number of edges are kept, specifically equal to the number of vertices minus one. The few number of edges renders the minimization of the force-directed potential function computationally efficient.

Since there are multiple descriptors extracted from the users, multiple notions of distance are considered and multiple objective functions are defined. The simultaneous minimization of all the objective functions is normally infeasible; however, such problems are solved by multi-objective optimization methods, which result in a set of solutions, instead of one. The set of solutions, namely the Pareto set, contains those solutions which cannot be further improved without sacrificing one of the objectives. These solutions represent different optimal trade-offs among the multiple objectives and can provide insight into various aspects of the mobile network data and the user behaviors. More details can be found in [20].

3.3 Hypothesis formulation and validation

Moving in even less abstract levels, MoVA allows the operator to **formulate network security-related hypotheses** for specific limited sets of data, and attempt to verify or reject them. The operator formulates a hypothesis by selecting a subset of the available CDR records, either manually or through one of the visualization tools. The hypothesis data are provided as input to the various visual analytics tools of the proposed system. Depending on the type of question submitted by the analyst, appropriate anomaly detection or visualization methods are employed in order to provide answers. The produced answers, either quantitative or visual, are presented to the analyst. Using this information as a guideline, the analyst can further explore the dataset by focusing on specific aspects of the data, or altering the question. The operator can submit one of the following three supported hypotheses.

Anomalous users hypothesis: The anomalous users hypothesis is testing if a subset of users exhibits abnormal behavior within a given time frame. Initially, the analyst selects a time frame of interest and then selects a subset of users which he/she desires to check if they have abnormal behavior or not. Then, the activity graphs are created for each user and provided as input to an anomaly detection algorithm, in order to identify if the selected users exhibit abnormal behavior. These activity graphs are directed graphs, in which the vertices represent users, and edges the communication between them. The weights of the edges represent the number of communication events between two users. Multiple features are then extracted from the graph created from the neighbors of each user, and are used for the supervised classification (using random forests [21]) of each user into two classes, normal and anomalous. The extracted features are: 1) number of edges, 2) Edge weight entropy, 3) Graph entropy, 4)

Maximum LOF value of the edge weights, 5) Ratio of outwards towards inwards edge weights, 6) Average Outward/Inward Edge Weight, and 7) Number of Outward/Inward Edges with a specific weight. The result of anomaly detection is a classification of the selected users as anomalous or normal, which is utilized to calculate the probability that the original hypothesis is true.

Core network impact hypothesis: The core network impact hypothesis tests if a set of selected users exhibit abnormal behavior simultaneously, indicating a DoS attack which has a negative impact to the core network. The data available for the users are split into time windows of duration determined by the operator and anomaly detection is performed for each of the time frames. This results in a series of classification results, which are finally used to test the original hypothesis.

Botnet attack hypothesis: In this hypothesis, the analyst suspects that the detected attacks against the network were launched by a botnet, installed on multiple mobile devices. In order to validate or reject this hypothesis, the system searches in the past activity of the selected users for common patterns, which indicate 1) the botnet command and control interfaces, and 2) the infection methods that they use to spread and infect additional devices. To achieve this, the past CDR activity of each user is used to construct a symbolic time series, in which each symbol represents a specific communication event (e.g. Call towards user-X, or SMS from user-Y). Multiple symbolic time series are created, one for each user. These time series are provided as input to a Frequent Episode Mining algorithm (FEM) [22], for the identification of common behaviors. The results are filtered based on a list of events which are known to be normal. The patterns that remain correspond to either the command and control interfaces, or the infection methods, or both.

It should be noted that the whole application can be used either with stored historical network data, in an off-line manner, as well as with streaming data, arriving at each time instance, in real time, in an on-line manner. The arriving signals can be initially visualized with the stacked graphs visualization approach, described in Section 3.1, where the new signals are appended to the right part of the plots. By selecting a time instance of the new data, the Mobile Network Graph can be visualized too. If a small time window around the arriving data is considered and selected by the operator, e.g. one day, behavioral features can be extracted from the CDR records and be used within the multi-objective behavior visualization. Using different durations for the time window, features can be extracted that correspond to smaller or larger time periods, thus allowing the operator to overview various temporal characteristics of the data and cope with the arriving data velocity. The various supported hypotheses can also be verified using this time window around the arriving data, for further analysis.

4 Experimental results

MoVA has been evaluated with synthetic data, in a scenario of Denial-of-Service (DoS) attack, caused by SMS-sending malware. In this section, the dataset used

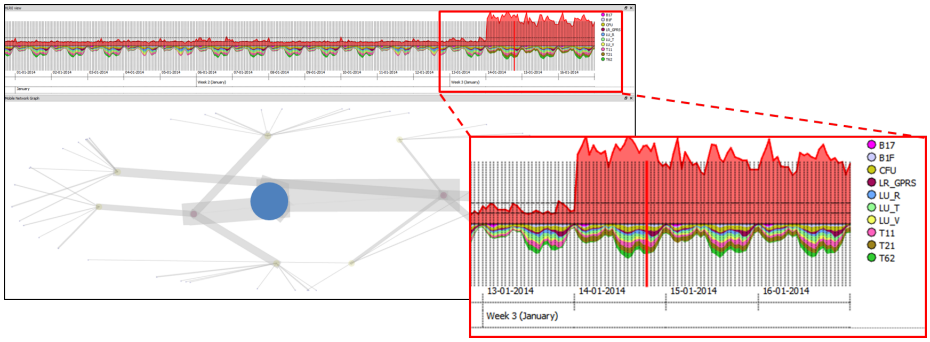

Fig. 4. The visualization of the signaling activity. The anomaly score on 14-Jan-2014 has suddenly increased when compared to the previous days. In addition, the volume of the T21 signaling message (i.e. the number of SMS received by the network subscribers) has also increased. The graph representation of the signaling activity is shown on the lower part, and shows that activity has uniform distribution with respect to the network topology, and thus, the attacking devices are not geographically constrained.

is described, followed by the evaluation of each of the visual analytics tools and methods presented in Section 3.

4.1 GEDIS dataset

GEDIS Studio [23] is an online tool for generating Call Detail Record (CDR) data. For the purposes of evaluating the proposed MoVA application, CDR data corresponding to the SMS flood scenario of [8] have been produced. Six months (25 weeks) of data have been generated, for 4800 mobile phone users. The users are split into two groups. The first group consists of 4500 normal users with an SMS sending rate of 8 messages per day, while the rest 300 users are considered to be infected by SMS flooding malware, sending SMS messages with a rate of 64 messages per day, when the malware is active. The attacks are considered to occur from day 106 to 126 and from day 148 to 154, i.e. 4 weeks in total.

4.2 Results

Figure 4 presents the visualization of the signaling activity. As shown in the stacked graphs representation, the anomaly score on 14-Jan-2014 has suddenly increased by a large factor, when compared to the previous days. Along with this increase, there is also an increase in the T21 signaling message. This signaling message represents the number of SMS received by the network subscribers. The graph representation of the signaling activity on 14-Jan-2014 for a specific hour is shown on the lower part of Figure 4. The signaling activity has uniform distribution with respect to the network topology, and thus, the attacking devices are not geographically constrained.

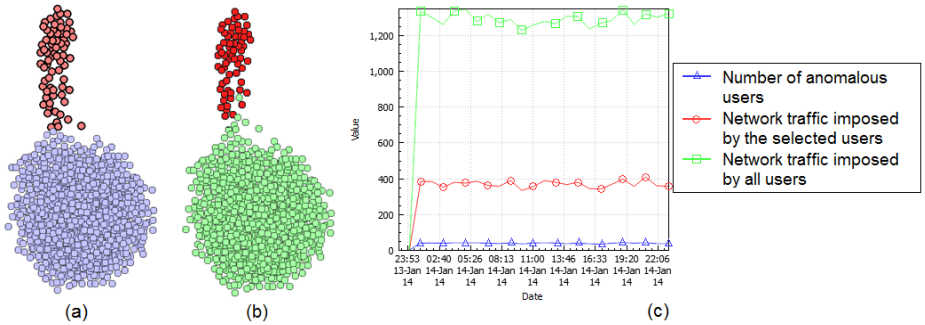

Fig. 5. The results of the hypothesis testing tool. (a) The result of the multi-objective visualization method. The red/dark color denotes the users selected by the operator. (b) The results of the anomalous users hypothesis. The red/dark color denotes the users identified as anomalous and the green/light color the normal users. (c) The results of the core network impact hypothesis.

Selecting the 14-Jan-2014 as the time period for more detailed analysis, the analyst filters the CDR records and proceeds with the visualization of the CDR activity, using the multi-objective visualization tool. The THD and RHD descriptors were extracted and used as the multiple modalities. Figure 5(a) depicts the results of the visualization. This visualization corresponds to a solution from the center of the Pareto front. A small group of users appear to separate from the rest of the users, indicating that they have a much different, probably abnormal, behavior. The red/dark color in Figure 5(a) denotes the users selected by the operator for further analysis. By selecting this group, the operator wishes to test various hypotheses regarding these users.

Testing the anomalous users hypothesis is performed by running the anomaly detection method described in Section 3.3. The results of the anomaly detection are illustrated in Figure 5(b), where the red/dark color denotes the users detected as anomalous. It is apparent that most of the selected users of the small cluster are indeed identified as anomalous by the anomaly detection algorithm. The probability that the hypothesis is valid, i.e. that the selected users exhibit anomalous behavior, is 87.6%.

The core network impact hypothesis can be used by the operator to test if the selected users exhibit a behavior that has an impact to the core network, rather than to individual users. Figure 5(c) contains the results of testing the network impact hypothesis. From the leftmost figure, it is apparent that the number of anomalous users is maintained throughout the whole day, which is a first indication that the attack aims at the network malfunction. Furthermore, the network traffic of the selected users is close to 50% of the total traffic of all users, causing a significant negative impact to the core network.

Finally, the same users were utilized in order to run the botnet attack hypothesis. The past 14 days of CDR activity were utilized for the detection of common behaviors. The FEM algorithm identified multiple patterns which con-

Table 1. Comparison to commercial mobile network security applications.

Name	Generic	IP network security	Mobile network security	Techniques
SAS VA [13]	✓	-	-	Line charts, pie charts, colored tables.
Tableau [14]	✓	-	-	Line charts, histograms, pie charts, colored tables, maps.
Centrifuge Systems [15]	✓	✓	-	Histograms, graphs, pie charts, colored tables.
Motive Security Guardian [16]	-	-	✓	Anomaly detection techniques.
NSN Mobile Guard [17]	-	-	✓	Anomaly detection techniques.
MoVA	✓	✓	✓	Stacked plots, graphs, k-partite graphs, multi-objective visualization, anomaly detection algorithms, hypothesis testing.

tained the normal event $\langle saf.gr \rangle$. Removal of this event results in one dominant pattern, namely, $\langle mal.gr, 10.10.10.1 \rangle$. This pattern shows that all the anomalous users visited the webpage *mal.gr*, from which they were infected by the botnet. Afterwards, the botnet communicated with the specific IP address *10.10.10.1* for the command and control interface.

Compared to existing commercial solutions available, the proposed MoVA application has several advantages. Table 1 summarizes the differences between existing solutions and the proposed one. Widely used visualization applications, such as SAS VA [13] and Tableau [14], although they provide a variety of visualization techniques, are generic in nature. They can be used with any type of multi-dimensional data and they are not targeted to mobile network security. Certifuge Systems [15], is further able to handle data related to IP network security, but it still lacks support for mobile network security data. On the other hand, applications that are targeted to mobile network security, such as Motive Security Guardian [16] and NSN Mobile Guard [17], do not provide visualization tools that would allow the mobile network operator to have an insightful overview of the network operation. The proposed MoVA application is able to fill these gaps, by providing visualizations targeted at mobile network security, while still being able to be easily extended in order to handle IP network data, as well as any generic type of data.

Furthermore, existing mobile network-oriented applications utilize one kind of data, either signaling or CDR. Utilizing both signaling and CDR data for the analysis and visualizations of the proposed method has the advantage of allowing the implementation of the proposed top-down approach. The large amounts of constantly arriving signaling data are used in the initial stages of the visual analysis, utilizing more straightforward techniques, such as the stacked plots and

the Mobile Network Graph, which present an overview of the large available data. Then, the more elaborate techniques, such as the multi-objective visualization and hypothesis testing are performed on smaller sets of data, utilizing the CDR records available for the users. This allows the application to handle big data, by allowing both a large-scale overview of the data and smaller-scale analysis and visualizations.

Finally, compared to methods that only utilize anomaly detection algorithms, the use of visual analytics, in the proposed application, has the advantage of data exploration. Visual analytics, by definition, involves the human operator in the visualization process, by allowing him/her to select data subsets, fine-tune parameters, compare different visualizations, etc. Combining the fast analytic processes of the automatic anomaly detection and analysis tools with the human perception, through visualization and interaction, leads to an effective means to explore the available data, to allow the operator to view different aspects of them and, ultimately, to gain more insight and arrive at a more accurate decision. Applications such as Motive Security Guardian [16] and NSN Mobile Guard [17], although able to detect malware activity, do not provide the mobile network operator with visualizations that would allow him/her to instantly view the infected users, the origins of an attack, etc.

5 Conclusions

In this paper, a visual analytics application for the monitoring of large mobile phone networks, focused on network security and aiming at detecting traffic anomalies, namely MoVA, has been proposed. MoVA provides the mobile network operator with several tools allowing her/him to have an overview of the whole network and gradually focus on less abstract data that are candidates to be abnormal, thus gaining useful insights in the network operation. For an overview of the whole network, the Mobile Network Graph and stacked graphs visualizations are provided, which present a spatio-temporal view of the network. Focusing on a specific time period, the behavioral similarities of a set of users is analyzed through the multi-objective visualization tool, which aggregates information from many traffic attributes, in order to visually separate normal from abnormal behaviors. Finally, the operator can focus on a set of candidate anomalous users and use the hypothesis testing tool to verify specific hypotheses regarding the anomalous behavior of the users, its impact to the core network, and its root causes. Experimental evaluation of MoVA with a CDR dataset verifies the applicability of the proposed tools in visualizing large network traffic data and facilitating the operator. Future work includes adjusting and evaluating the proposed approach in various mobile network-related use cases.

References

1. "NEMESYS project." http://www.nemesys-project.eu/nemesys/, 2015.
2. J. K. Petersen, *The telecommunications illustrated dictionary.* CRC Press, 2002.

3. P. P. C. Lee, T. Bu, and T. Woo, "On the detection of signaling DoS attacks on 3G wireless networks," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pp. 1289–1297, IEEE, 2007.
4. P. P. C. Lee, T. Bu, and T. Woo, "On the detection of signaling DoS attacks on 3G/WiMax wireless networks," *Computer Networks*, vol. 53, no. 15, pp. 2601–2616, 2009.
5. A. D'Alconzo, A. Coluccia, F. Ricciato, and P. Romirer-Maierhofer, "A distribution-based approach to anomaly detection and application to 3G mobile traffic," in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, pp. 1–8, IEEE, 2009.
6. A. Coluccia, A. DAlconzo, and F. Ricciato, "Distribution-based anomaly detection in network traffic," in *Data Traffic Monitoring and Analysis*, pp. 202–216, Springer, 2013.
7. G. Yan, S. Eidenbenz, and E. Galli, "Sms-watchdog: Profiling social behaviors of sms users for anomaly detection," in *Recent Advances in Intrusion Detection*, pp. 202–223, Springer, 2009.
8. E. K. Kim, P. McDaniel, and T. La Porta, "A detection mechanism for SMS flooding attacks in cellular networks," in *Security and Privacy in Communication Networks*, pp. 76–93, Springer, 2013.
9. I. Murynets and R. P. Jover, "Anomaly detection in cellular Machine-to-Machine communications," in *Communications (ICC), 2013 IEEE International Conference on*, pp. 2138–2143, IEEE, 2013.
10. N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.
11. Q. Ye, T. Zhu, D. Hu, B. Wu, N. Du, and B. Wang, "Cell phone mini challenge award: Social network accuracyexploring temporal communication in mobile call graphs," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, IEEE, 2008.
12. Z. Shen and K.-L. Ma, "Mobivis: A visualization system for exploring mobile data," in *Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific*, pp. 175–182, IEEE, 2008.
13. "SAS Visual Analytics." http://www.sas.com, 2015.
14. "Tableau Analytics." http://www.tableau.com/, 2015.
15. "Certifuge Systems." http://centrifugesystems.com/, 2015.
16. "Alcatel Security-guardian." http://www.alcatel-lucent.com/solutions/security-guardian, 2015.
17. "NSN mobile guard." http://nsn.com/sites/default/files/document/nsn_mobile_guard_executive_summary.pdf, 2015.
18. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *ACM Sigmod Record*, vol. 29, pp. 93–104, ACM, 2000.
19. S. Papadopoulos, V. Mavroudis, A. Drosou, and D. Tzovaras, "Visual Analytics for Enhancing Supervised Attack Attribution in Mobile Networks," in *Information Sciences and Systems 2014*, pp. 193–203, Springer, 2014.
20. I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-Objective Optimization for Multimodal Visualization," *Multimedia, IEEE Transactions on*, vol. 16, no. 5, pp. 1460–1472, 2014.
21. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
22. B. Ding, D. Lo, J. Han, and S.-C. Khoo, "Efficient mining of closed repetitive gapped subsequences from a sequence database," in *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pp. 1024–1035, IEEE, 2009.
23. GenieLog, "GEDIS Studio online," 2014.