J Vis (2014) 17:181–196
DOI 10.1007/s12650-014-0213-6

REGULAR PAPER

Ying Zhao · Xing Liang · Xiaoping Fan · Yiwen Wang · Mengjie Yang · Fangfang Zhou

MVSec: multi-perspective and deductive visual analytics on heterogeneous network security data

Received: 30 April 2014 / Accepted: 9 June 2014 / Published online: 25 July 2014

Abstract In this article, we present a visual analytics system, MVSec, which helps analysts understand better what information flows under network security datasets. The major contributions of this work include: (1) a data fusion strategy for multiple heterogeneous datasets by using unified event tuple and statistic tuple data structure, which compress large scale datasets and lays the foundation of cooperative visual analysis; (2) multiple coordinated views, which provide analysts with multiple visual perspectives to characterize loud events, dig out subtle events and investigate relations of events in datasets; and (3) a contextual visual analysis with deductive viewpoints, which inspires analysts to explore hypotheses and reason their deductions from visual narratives. In case studies, we demonstrate in detail how the system helps analysts draw an analytical storyline and understand network situations better in VAST Challenge 2013. Additionally, we discuss lessons learned in designing our system and participating in VAST Challenge 2013, which is helpful and applicable not only to similar network security systems but also to other domains facing visual analytics challenges.

Keywords Information visualization · Visual analytics · Network security · Vast challenge

1 Introduction

Computer network plays a very important role in information age and also suffers from all kinds of illegal access and attacks frequently. Various network security products will generate a lot of logs in monitoring and securing network infrastructure, such as NetFlow, Firewall, IPS (intrusion prevention system), and hosts health status monitor system. When network anomalies occur, clues of the anomalies will be more or less left in the corresponding logs. Consequently, understanding what information flows under these logs is concerned widely and eagerly for network security, network planning, and even counter-terrorism.

Diverse methodologies have been specially researched for analyzing network security data (Li et al. 2013; Patcha and Park 2007), such as the statistical based, the knowledge based, and the machine learning based. Although a mass of systems and algorithms have been adopted to prevent and detect network intruders automatically, humans are still crucial in the security process. As a young community

Y. Zhao · X. Liang · X. Fan · F. Zhou (✉)
School of Information Science and Engineering, Central South University, Changsha 410083, Hunan, China
E-mail: zff@csu.edu.cn

X. Fan
Laboratory of Networked Systems, Hunan University of Finance and Economics, Changsha 410205, Hunan, China

Y. Wang · M. Yang
School of Software, Central South University, Changsha 410075, Hunan, China

of network security research, visualization for network security focuses on taking advantage of the power of the human perceptual and cognitive processes by bringing robust visual tools into hands of humans in solving computer network security problems (Goodall 2007; Shiravi et al. 2012). Amounts of visual tools have played very significant roles in helping network security analysts to detect anomalies (Ren et al. 2005; Fischer et al. 2012), discover patterns (Koike et al. 2005; Mansmann et al. 2007), identify correlations (Livnat et al. 2005; Fink et al. 2005), and assess network security situation (Yin et al. 2004; Erbacher 2012).

With the increasing size of networks and continuous appearance of new types of attacks, the research on visualization for network security is facing more and more challenges (Cook et al. 2012). The first of these challenges comes from the enormous growth of network security data in both volume and complexity. The increased number of monitor systems and automated processes keep continuously logging the network status and events to the storage servers. This massively growing datasets have recently coined a new term "big data" (Manyika et al. 2011). Therefore how to conduct an effective and efficient visual analysis on the "right data" extracted from the "big data" is very challenging.

The second challenge stems from demand for collaborative visual analysis of multiple network security datasets. Security devices running at different locations of a wide network constitute a distributed audit and defense system (Bass 2000), therefore, analysts are more inclined to combine data from multiple sources for sophisticated anomaly detection and comprehensive security situation assessment. Although some visualization systems (Fischer et al. 2013; Ghidini et al. 2012) based on multiple security datasets now are springing up, how to cope with various heterogeneous data sources for more seamless collaborative analysis remains challenging.

The third challenge comes from higher need of the overall view of the whole network status. Conventional visualization systems are often limited in providing a low-level view by leaving the process of identifying abnormal events and threats in security datasets, which are obviously inefficient and inadequate for a human security analyst to make timely and informed decisions. But what is a high-level view? Network security situation awareness has been widely accepted as a high-level view for analysts on determining the overall status of all network assets based on tremendous network events. Additionally, a storyline view is another high-level view, because the narrative of what is happening is more useful than listing facts without context, and the analytic story is also an important element of effective situation awareness (Challenge 2013; Walker et al. 2013). Thus, analysts often work like detectives in analyzing the security events with reasoning and deductive method, besides the role of the decision maker. In recent research, many visualization systems (Yin et al. 2004; Erbacher 2012) have been designed to intuitively depict the security situation, yet the visualization tools which help analysts to develop hypotheses based on scattered clues and deduce findings by pulling together an analytic story are still rare. With the coordinated, staged and hidden network attacks appearing increasingly, how to introduce the visual analytics into the scenario discovering in computer network security is another very interesting challenge.

To meet these challenges, we designed a novel interactive visualization system—"MVSec". Consisted by different processing parts and four specialized visualization views, MVSec is particularly suitable for features of network security data to help analysts reach cooperative analysis on multiple heterogeneous datasets collecting from different network security devices. The main contributions of this work are the following: (1) a data fusion model of multiple heterogeneous datasets, which provides metadata with unified format and compresses "big data" into "small data" and lays the foundation of cooperative visual analysis. (2) A visual analytical system with multiple coordinated views, which provides analysts with multiple visual perspectives of security data to characterize loud events, dig subtle events and investigate relations of events. (3) A contextual visual analysis with deductive viewpoints, which helps analysts to draw an analytical storyline and understand network situation better. (4) Sharing an experience of making use of our system in VAST Challenge 2013, which is helpful and applicable to other visual analytics applications.

The remainder of this paper is structured as follows: In Sect. 2, we briefly introduce the process flow of MVSec. In Sect. 3, we explain the methods of data preprocessing. We describe the graphical user interface and the visualization components of the implemented system in Sect. 4. Additionally, related work is distributed in Sect. 1 and the front part of each view's introduction in Sect. 4. Case studies in Sect. 5 demonstrate how the system can be used to analyze multiple datasets of the Mini Challenge 3 of VAST Challenge 2013. Finally we discuss the advantages and disadvantages of the system in Sect. 6 and make conclusions in Sect. 7.

2 System overview

Figure 1 shows the overview of the process flow of MVSec. It contains four major steps: data input, data preprocessing, visual analysis and comprehensive diagnosis. The input of the system is a set of heterogeneous network security logs. The data pre-processing, which performs data checking and data extraction work for better cooperative visual analysis on multiple source data, includes two parts: the data cleaning and time synchronization part that saves the correct logs into raw database, and the data fusion part that extracts events and data statistics as metadata with unified format from raw database. In the visual analysis step, four views of MVSec are particularly designed for features of network security data to help analysts reach cooperative analysis on multiple datasets. Heat map view is automatically deployed based on network topology to monitor network traffic; event radar view is good at mining correlations of events; comparative stacked stream view aims to compare time series; and port matrix view helps analysts to discern port information patterns. Multiple interactive visualization views provide analysts with multiple visual perspectives of security data to characterize loud events, dig subtle events and investigate relations of events. Finally, with our visualization system and the detective viewpoint, analysts are able to piece out an analytic story by collecting all finding and conduct a comprehensive diagnosis of network situation.

3 Data preprocessing

Data preprocessing is the first step in the whole workflow, especially when faced with a large number of heterogeneous datasets collecting from different network security devices. In this approach, the data preprocessing mainly consists of three parts: data cleaning, time synchronization and data fusion.

Data cleaning performs the check work to ensure that collected data is accurate when raw data from related network appliances is being loaded into the database or files suitable for visual analysis. Meanwhile, the inconsistency of time format is another common problem existing in different network security logs. Even for devices in the same network, they may be scattered across different time zones in a large-scale network around the world. Thus, the time synchronization is one of the important issues of data preprocessing, including establishing the unified time reference, time format and time precision.

This visualization system input would be the data from numerous heterogeneous network security appliances, such as Packet Sniffer, NetFlow, Firewall and IDS (Intrusion Detection System). The heterogeneous devices generate non-uniform logs, and different logs reflect different facets of the network security. Thus, a significant challenge is to combine data from numerous heterogeneous devices into a coherent process that can be used to evaluate the overall situation of cyberspace. In the visualization system, the core issue of data fusion is to extract metadata with unified format for carrying out cooperative visual analysis on multiple data sources. Thus, we extract security events and data statistics as metadata from multi-source logs.

Definition 1 TupleEvent: In this paper, each network security event is defined as the TupleEvent with seven tuples: TupleEvent (Time, EventType, Priority, SourceIP, DestinationIP, SourcePort, DestinationPort).

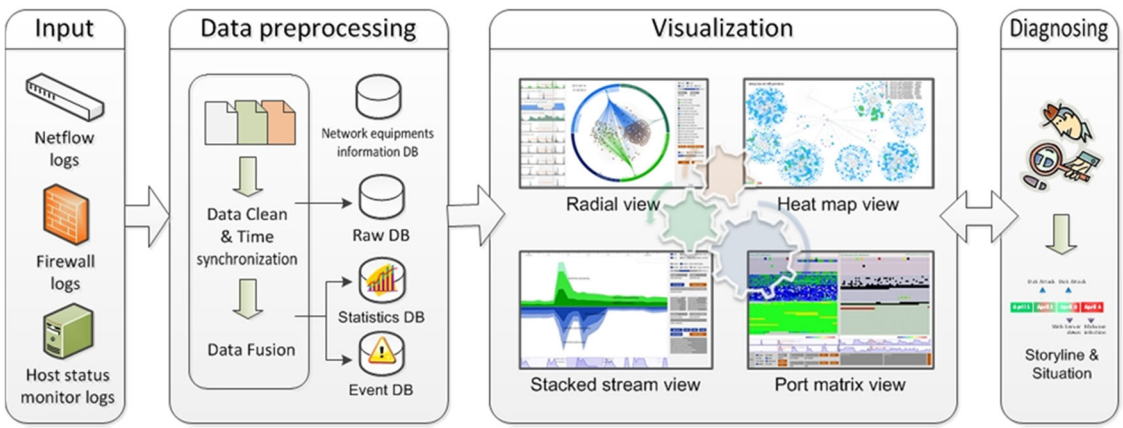

Fig. 1 The workflow of MVSec

Network security events are very suitable for fusion of information because of their uniform data structure which is abstracted from network security data. There are many sources of TupleEvent. Raw logs are the first source. For example, an alert record in IDS logs can be directly used as a TupleEvent. Another source of network security events may be the activities that exceeded preset threshold or broke rules. For instance, we can set the threshold of flow bytes in NetFlow logs to identify flow flooding event. Moreover, TupleEvent can also come from data mining or machine learning systems based on expertise knowledge.

Definition 2 TupleStatistic: In this paper, each statistical unit of network security data is defined as the TupleStatistic with eight tuples: TupleStatistic (TimeBegin, TimeEnd, TimeInterval, StatisticType, StatisticScope, StatisticParam, StatisticValue, MeasureUnit).

The statistics can be easily abstracted into one uniform data structure and is suitable for the fusion of information. Common statistics in analyzing network security data includes the number of connections, flow counts, flow bytes. Statistics can also be computed from the whole network, from the subnet, from the single hosts, from the single port, or from the single prototype.

4 Visualization

In this section, we will introduce design details of four visualization views in MVSec.

4.1 Heat map view: presenting network traffic based on network logic topology

Network logic topology helps analysts locate hot areas and abnormal activities efficiently in visual analysis process. Many visualization techniques have been developed to implement the IP level topology of enterprise network based on the IP address prefix rule, such as grid graph used in NVisionIP (Lakkaraju et al. 2004), square technique used in Quadtree (Teoh et al. 2002) and Treemap used by Florian Mansmann (Mansmann et al. 2007). A common drawback of these systems is not intuitive enough to express the logic network topology. We are inspired by dandelion to design a hierarchical layout of network logic topology based on IP address prefix rule and force-directed algorithm.

The Big Marketing enterprise of which the network architecture is shown in Fig. 2a is taken to introduce the layout strategy. Organizationally, Big Marketing consists of three different branches within each of them possessing around 400 workstations and one group of servers. Shown in Fig. 2b, seeds of dandelion spread out from the center of flower stalk, of which each seed looks like a lollipop to some extent, while many fluffy antennas extend out from the seed. We referred this natural hierarchical two-layer structure as a metaphor to network layout design. First, IP address is divided into two parts (subnet and host). For instance, a network IP 172.10.1.15 can be divided into 172.10.1 (subnet) and 15 (host). Each seed of the dandelion will be regarded as a subnet and antennas on this seed represent hosts in this subnet. As all seeds directly link to the flower stalk, we can define the stalk as the only entrance to the intranet of enterprise's network, which means the stalk may be a core switcher or router showed as the red nodes in the center of the global layout. During deployment, we will first deploy the subnet nodes as the first layer by forced-layout algorithm. Then hosts in each subnet will be deployed, respectively according to another force-layout algorithm, during which, in case of the mess between nodes of different subnets, nodes in the same subnet have larger attraction than repulsion while nodes among different subnets have larger repulsion than attraction. Users are allowed to adjust the position of subnet nodes during deployment. Figure 2c shows the compact layout used in radial graph including the whole hosts and servers in the network of Big Marketing. Furthermore, incompact layout owns a larger space including the external hosts appeared in the logs and internal mapping servers, shown in Fig. 2d, which is used in heat map view.

Heat map view based on network topology, shown in Fig. 2e, performs well in monitoring activities of the hosts in real-time or historical data within colors from cool to warm tune representing the growth of active extent of activities. Analysts are capable of directly finding vulnerable hosts as well as analyzing their correlations on spatial and temporal dimensions.

4.2 Event radar view: exploring correlations of network security events

In the field of network security visualization, there are several tools using radial graph to analyze correlations of events, such as VisAlert (Livnat et al. 2005), IDSRadar (Zhao et al. 2013) and AlertWheel (Dumas

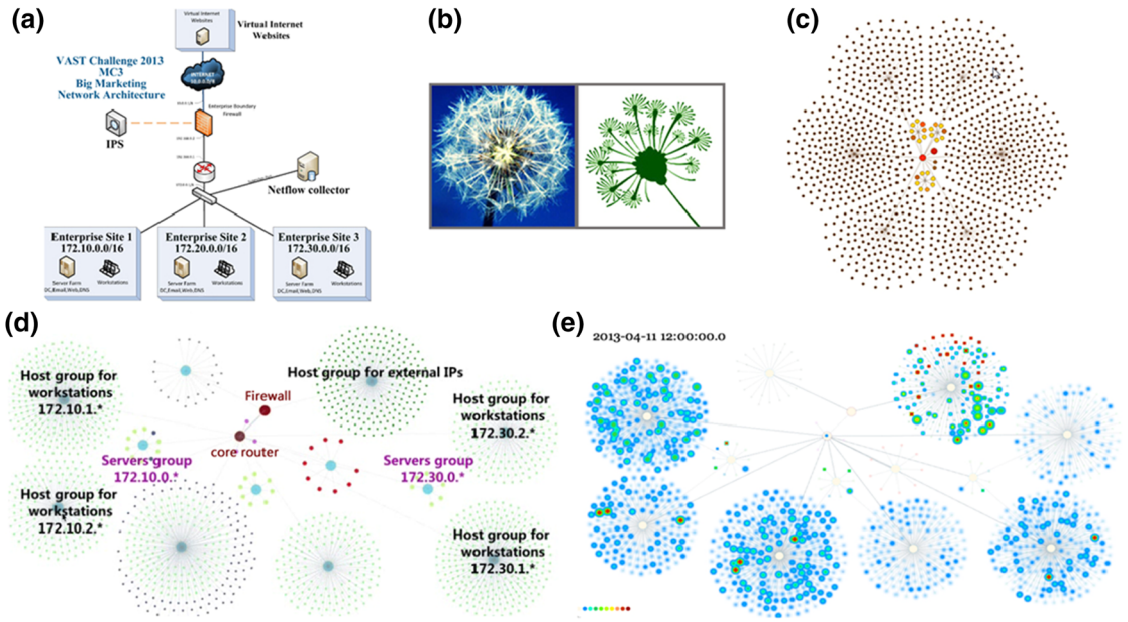

Fig. 2 Heat map view. **a** Network architecture of Big Marketing. **b** Diagram of the dandelion. **c** The compact layout of Big Marketing for radial graph view. **d** The incompact layout of Big Marketing for heat map view. **e** The network traffic visualization in heat map view

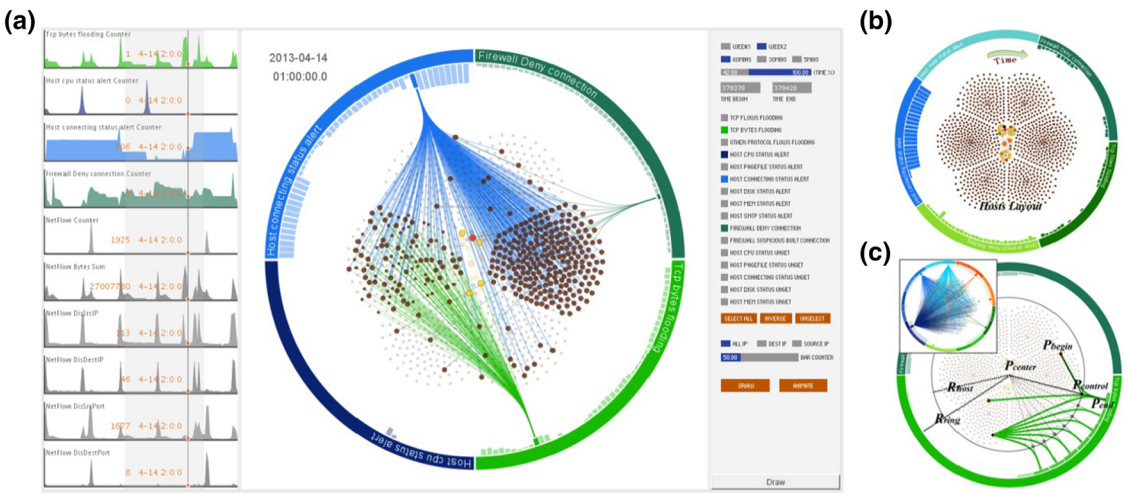

Fig. 3 Event radar view. **a** The interface of event radar view. **b** The representation of three core attributes of event: what, when and where. **c** *Line* bundling in radial view

et al. 2012). However, some issues within them still need to be improved. The first issue is the lack of multiple views; the second one is the demand of a better host layout based on network topology; the last problem is visual clutter issue caused by messy straight lines appearing in the same time. Event radar view, shown in Fig. 3a, takes a more comprehensive and vivid perspective at how to make sense of the abundant amount of security events and how to analyze their correlations.

As the main part of the interface, the radar view provides an overview of events and their inherent associations. A timeline group view is shown on the left of the interface, representing several statistical time series, such as network traffic and the number of distinct active IPs over time, which is very helpful to provide the overall network state information while performing event analysis. Additionally, the interactive control panel on the right of the interface provides network security analysts with the

function of setting and selecting parameters. From the definition, a TupleEvent must contain the three most fundamental elements, namely: What, When, and Where. Corresponding to these three core attributes, the radial graph is designed with three parts: network security event types, histograms over time and intranet hosts (servers and workstations), shown in Fig. 3b. In the center of radar, the colored nodes arranged by hierarchical force-directed layout algorithm are servers and workstations of a corresponding corporate network, which is composed of more than one thousand workstations and about twenty servers. The colored arcs evenly divided on the ring are known as the network security event types which are selected on the control panel. The histogram inside the ring is drawn clockwise along. The height of histogram in a particular color arc represents the amount of this event type happened or observed in a sampling interval which can be minutes or hours tuned by the security analysts. A line links the three properties of an event from the top of histograms, implying what types of and when the event happened, to a specific host in internal network layout, implying where the events happened. The color of the line is the same with the color of the event type it links. The width of the line and the size of the highlighted host are related to the dangerous level which is set by network security analysts in advance.

When the distribution of the hosts related to selected events at that time is very wide, the straight lines comes from the histograms will fan out, shown on the top left corner of Fig. 3c. To alleviate this visual mess, we designed a line binding strategy. This binding strategy is to utilize Bezier curves to bind the straight lines within the same event type and the same time interval. Shown in Fig. 3c, the start point of curve is the host location point, the end point of curve is the head of corresponded bar of histogram, and the two middle control points overlap each other on the one crossover point of the inner circle and the line linking from the circle center to the corresponded bar of histogram.

By providing multiple interactions, users are able to drill down correlations of events on the perspectives of What, When and Where as well as discover the dangerous behavior patterns by bringing into the analysis of statistics. In real-time monitoring, our system will display the times that selected event types happened and the correlated hosts in these events. The timeline group will also walk forward as the time goes by. When the arc is unable to bear more histograms, the old histograms will be substituted by new histograms. In history data analysis, after loading the selected data, analysts are allowed to filter data by simply clicking on any of the hosts, servers, event types and bars of histograms, then all the related hosts and lines will be highlighted. If analysts want to analyze correlations of events, multi-selection is provided.

4.3 Comparative stacked stream view: investigating implicit information within multiple time series

Some traditional tools utilized to analyze time series like line graph and bar graph were no longer able to meet the higher demand, so stacked graph tools, like Tstat (Finamore et al. 2011), were particularly designed to assist our analysts in better observing the changes of multiple time series. Usually, multiple time series in one stacked graph belong to the same type, but comparisons among different types of time series are also very important in the analysis of network security data, which is just out of the capacity of basic stacked graph. For example, if there are three time series about the number of connections from three particular workstations and another three time series about flow bytes of them over time, comparing two groups of data can lead to more implicit information. Inspired by ThemeRiver (Havre et al. 2002) and stacked area graph (Plonka 2000), we designed the comparative stacked stream view which is able to meet analysts' higher need of comparing different types of time series. Shown in Fig. 4, two groups of time series are stacked upside and downside, respectively, along the central line of the painting area, which reaches the goal of analyzing the differences and tendency both among the same type of time series and between two different types of time series.

The interface of our time series tool is made up of three parts. Shown in Fig. 4, the central part is the stacked stream graph, the right panel is configuration panel, and the bottom is a single timeline. Timeline highlights the current time period used in the stacked stream graph, meanwhile, it provides analysts with another comparative dimension. To simplify users' some repetitive operations of selecting time series, we offered users some frequently-used configuration files to directly get the desired outcomes and users are also able to output their configurations as a new configuration file. Two groups of time series are, respectively drew in green tone and blue tone within different color depth in one group of time series coding different themes. Other interactions include the option of displaying theme labels and values, the option of zooming in or zooming out, the option of painting in straight line or curve and so on.

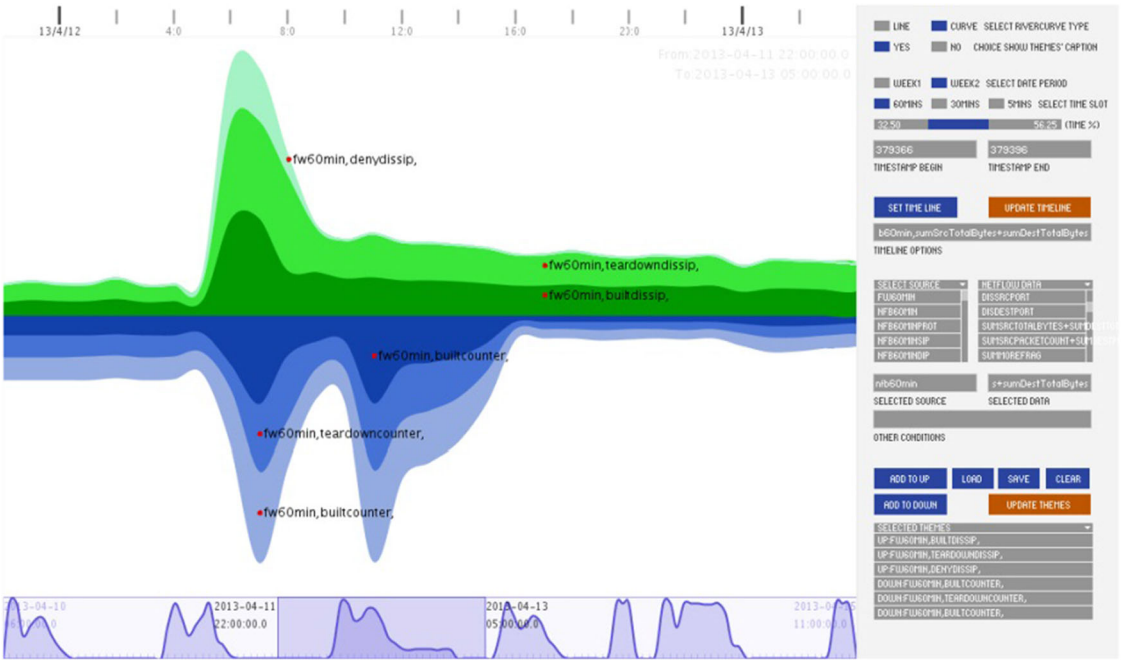

Fig. 4 Comparative stacked stream view

4.4 Port matrix view: depicting the characteristics of port activities

In computer network, the port-based analysis helps analysts to track activities on the specific applications and tie attacks to vulnerable services. Many visualization tools have been implemented in this area, such as PortVis (McPherson et al. 2004) and NetBytes (Taylor et al. 2007), but how to better deal with 65,536 port numbers remains a challenge. In our approach, we conduct a user-interest design that over sixty thousands ports are divided into four groups:

- The user customized ports: They are the most interested ports selected by analysts. What analysts most concerned about is the ports associated with essential services or applications in their perspectives.
- The well-known ports: The port numbers from 0 to 1,023 are the well-known ports, and they are used by system processes that provide widely used types of network services.
- The registered ports: They are occupied by some famous applications and the range of port numbers is from 1,024 to 49,151.
- The uncommon ports: All other ports that are not declared in previous groups.

Port matrix view provides a port-based overview of network activities, which has an improved representation for depicting characteristics of port activities in the entire intranet or a particular host. Shown in Fig. 5, each port is represented as a grid while ports in each group range from bottom to top and from left to right. For the uncommon ports, 100 continuous ports are placed into one grid cell because of a large number of ports classified into this group. The interface includes three parts: port matrix, timeline group and control panel. The main part is the port matrix view with the source port matrix on the left and the destination port matrix on the right. The activity of a port is color-coded from cool tune to warm tune to indicate how active it is. For instance, cool tune indicates low network traffic like black and blue; on the contrary, warm tune indicates high network traffic like orange and red. If analysts do not specify an IP address on the control panel, the port matrix view would depict a port-based traffic situation for the entire intranet in selected time period. The filter operation of clicking on a port cell is very effective for discovering associations between source ports and destination ports. Figure 5 shows that the port 80 is highlighted in the destination port matrix as a high level traffic; moreover, a large number of un-well-known ports in the source port matrix are related with it. The second part of the interface is the timeline group which provides users with four time series to grasp trends or observe statistical values during the current analysis period in historical data or in real time monitoring. The default settings of the timeline group include four time series on the entire intranet: the number of network flows, the sum of bytes, the number of distinct active destination IPs and

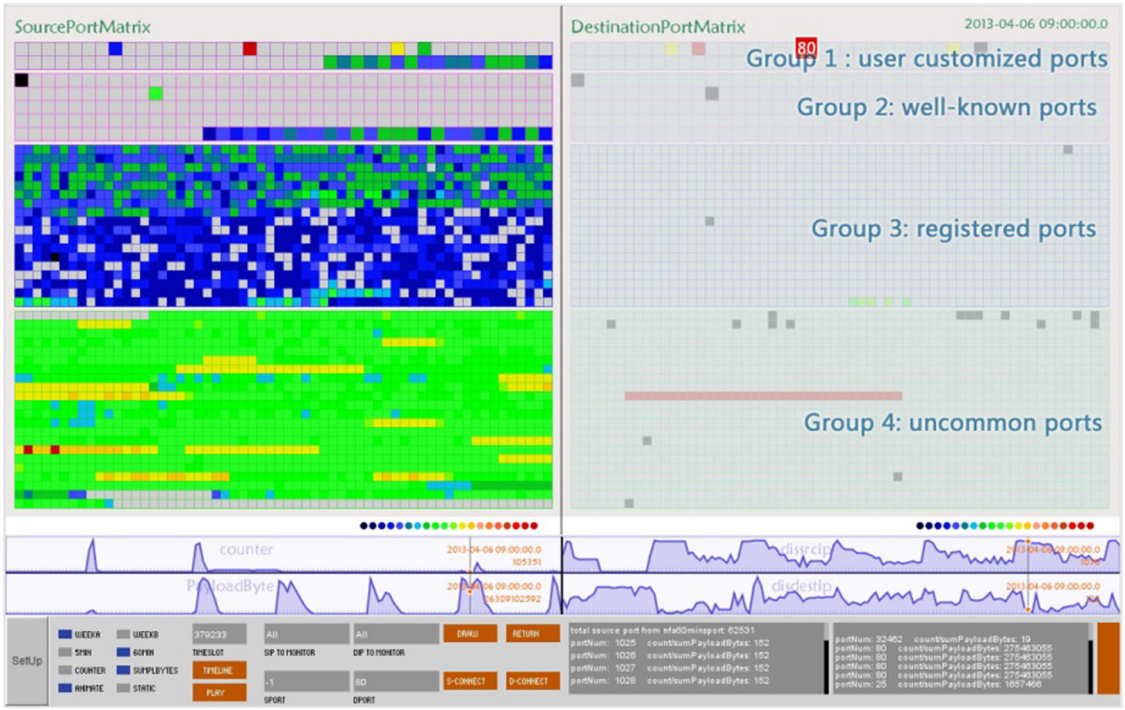

Fig. 5 Port matrix view

source IPs over time. In addition, the interactive control panel at the bottom of the interface provides users with the function of setting and selecting parameters.

5 Case studies

In this section, we will demonstrate two interesting cases to illustrate how our visualization system works, how it helps users cooperatively confirm anomalies from multiple perspectives, and how it catches small clues in loud events to deduce new findings and make the scenario analysis in context.

We utilize the MC3 (Mini Challenge 3 of VAST Challenge 2013) (VAST Challenge Homepage [EB/OL]. 2013) as case studies. The MC3 was the culmination of a three-year evolution of cyber network visual analytics contests (Cook et al. 2012; Challenge 2013; Grinstein et al. 2011). It represented an increase in complexity from previous years in several important ways. First, the security data of a hypothetical "Big Market" corporation had a longer time span (2 weeks) and a bigger volume (nearly 10G) than the previous contests in the series, meanwhile it contained three types of security logs—NetFlow, IPS and Big Brother (health and status metrics) logs. Second, the embedded story included over 30 ground truth events, and participants were required to not only detect anomalies but also analyze events in context for revealing the underlying story and providing situation awareness of its large computer network.

In the data-preprocessing part, we first cleaned these three kinds of datasets, synchronized them over time and then loaded them into MySQL database. In the fusion based on statistics, we made large numbers of statistical computations in order to acquire times series with three kinds of time granularities—60, 30 and 5 min. In the fusion based on events, we utilized the network traffic analysis based on threshold judging and the high-risk ports activities extraction in the NetFlow and IPS logs to customize abnormal events. Besides the customized anomalies, the host status alerts from Big Brother logs and the connection deny alerts from IPS logs were directly marked as abnormal events. Next, let us enter the visual analysis part.

5.1 The DoS attacks and server crash in the first week

The analysis begins with monitoring the global network traffic in chronological order. The heat map view is very easy for us to notice any singularly active host in global network. At 5:14, 2nd April, shown in Fig. 6a,

we find that many internal web servers in the heat map, especially 172.30.0.4, turn into red in this moment which means they were very active compared to other nodes in the network. At the same time, some of the external hosts turn into red as well which means they might be involved in this activity. For more details, we switch to the port matrix view to look for the port information during this period in NetFlow logs. When selecting the destination port 80, we find that the port matrix view, shown in Fig. 6b, displays an obvious attack pattern that over 60,000 source ports marked as dark blue, light blue and green in the left matrix accessed port 80 which is marked as red in the right matrix. Furthermore, shown in Fig. 6c, after these dense connections, connection problems exploded in Big Brother logs. So far, we have grasped three perspectives of the unusual events around 5:14, 2nd April. As a phased conclusion, the web servers, especially 172.30.0.4, were very likely to receive the DDoS attack from many external hosts, and the overload of those web servers made themselves fail to correctly response the requests submitted by internal workstations.

In some ways, the DDoS attack is a typical kind of loud events easy to be detected, however we can extend our findings by following some valuable clues that we find from our multi-faceted visual analysis on this loud attack. In order to verify the situation of 172.30.0.4, we begin the statistical analysis on 172.30.0.4. To check its detailed records count in NetFlow logs, we draw the records count of 172.30.0.4 as source IP and destination IP, respectively on the stacked stream view, shown in Fig. 7a. Spikes of two streams tell us that the connection to 172.30.0.4 increased suddenly twice showing that it was attacked twice. After that, an apparent blank period from 13:00 April 3 to 6:00 April 5 reminds us that 172.30.0.4 could not be accessed during that time and it also did not install any connection to others. Therefore, we suspect that after twice of dense attacks, web server 172.30.0.4, went down for 2 days. Additionally, the radial view during that period, shown in Fig. 7b, also tells us that web server was involved in unavailable memory status alert, unavailable disk status alert and unavailable pagefile status alert, which again proves that 172.30.0.4 went down. Even though quiet events like the crash of web server we found here is hard to discern, they do play more important roles for analysts than loud events in explaining how serious the DDoS attack was.

Another type of loud events is port scan which can be clearly recognized from port matrix view. Most of port scan events in the first week have similar patterns shown in Fig. 6b, but the port scan at 11:15 6th April showed a different pattern. Shown in Fig. 8a, almost all destination ports were involved in this port scan while the highly risky ports like FTP (File Transfer Protocol) port 21 and SMTP (Simple Mail Transfer Protocol) port 25 in the destination port matrix were more attractive at this moment. We think that the unusual activities on port 21 and 25 are more valuable clues leading further detections, so we drill down for more information about these ports in the stacked stream view. Shown in Fig. 8b, the network traffic on port 20 and 21 are drawn at the upper stream, and the network traffic on the internal mail servers (172.10.0.3, 172.20.0.3 and 172.30.0.3 as destination IPs) are stacked at the lower stream, and the overall network traffic is drawn in the timeline at the bottom of interface. Two spikes around 11:00 6th April illustrate that the possible outbreaks of spam and data exfiltration were hidden under this port scan event, thus we strongly suspect that the crash of web server was utilized by sly attackers to plant malware into the enterprise intranet. Minor changing of the overall network traffic on April 6 enlightens us little in discovering those two subtle events around 11:00 April 6, but the tiny clues in a distinct port scan lead us to find them out from another perspective.

5.2 The IPS defense and botnet infection in the second week

What kind of role was played by IPS which appeared in the second week, and what were the differences in two different weeks. The visual deductive analysis begins with the overall perspective of the whole week by

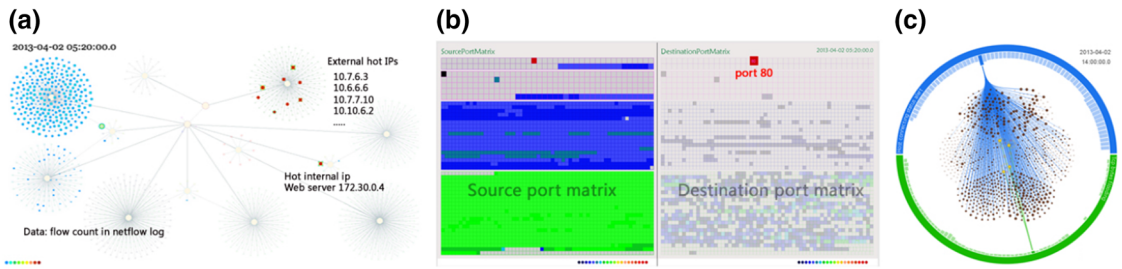

Fig. 6 The DDoS attack in the morning of April 2. **a** Hot external IPs and internal web servers shown in the heat map. **b** Port scan pattern shown in the port matrix. **c** The explosion of connection problems shown in the radar view

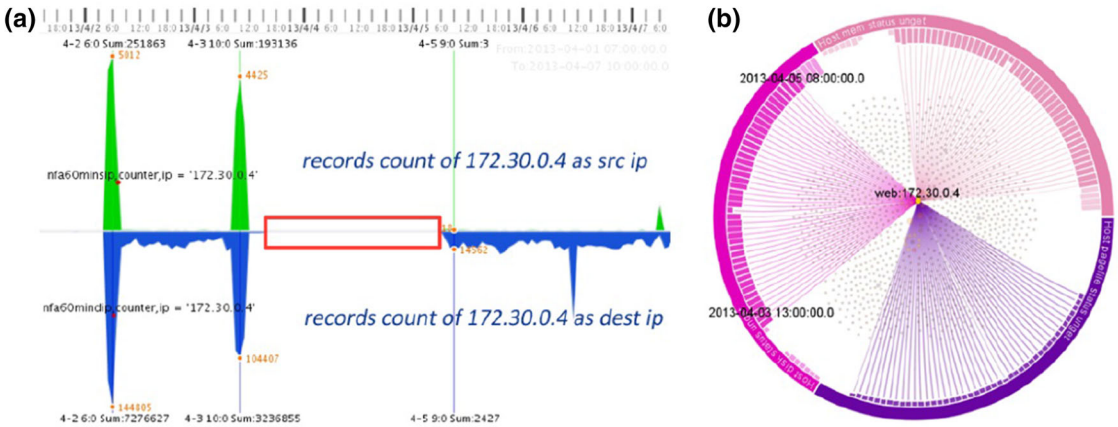

Fig. 7 The crash of internal web server 172.30.0.4. **a** The peaks and blank period of network traffic of 172.30.0.4 shown in stacked stream view. **b** The health alerts of 172.30.0.4 shown in radar

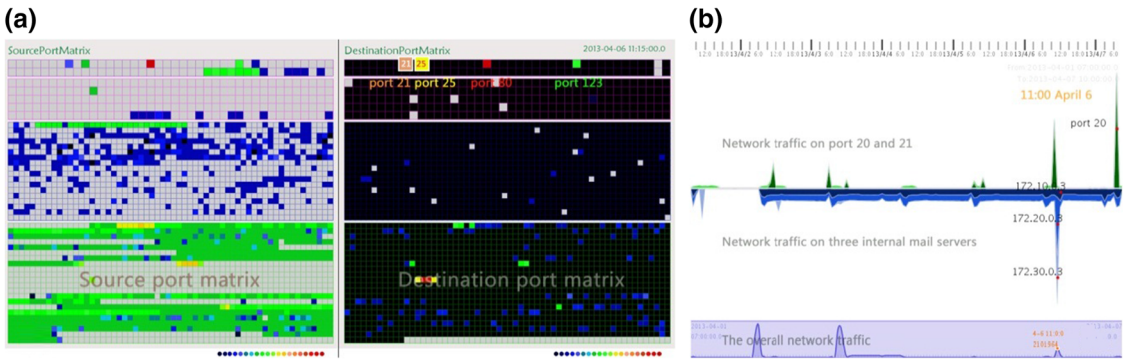

Fig. 8 The port scan and abnormal activities on port 21 and 25 around 11:15 April 6. **a** A particular port scan pattern shown in port matrix. **b** Two spikes shown in stacked stream view illustrate the possible outbreaks of spam and data exfiltration

comparing multiple time series. Shown in Fig. 9a, in the upper stream area, three stacked streams represent the number of deny, tear down and built operations records in IPS logs colored with the light green, the dark green and the darkest green, respectively. Below the central axis, the blue stream represents the overall TCP (Transmission Control Protocol) traffic in NetFlow logs. Figure 9a shows that there were large numbers of deny records in the whole week and TCP traffic suddenly increased twice around 11th April 11:00 AM and 14th April 15:00. In addition, it's easy to find that the activities on Port 21 and 25 significantly were reduced in the second week. These clues deliver the idea that even though the IPS blocked the known threats in the first week for many times, it still failed to defend the attack twice and the illegal external hosts were successful to break into the intranet and generated massive TCP traffic. Figure 9b, c show the further analysis of the peaks of network traffic on the morning of April 11 from other perspectives. A large number of internal hosts were involved in the explosion of network traffic, and a dozen of external IPs, such as 10.12.14.15, 10.6.6.7, and 10.78.100.150, made use of over 60,000 different source ports to raise DDoS attack to port 80 of web servers 172.10.0.4, 172.20.0.4, 172.30.0.4 and 172.20.0.15. So far, a preliminary analysis of the situation of the whole network in the second week is given to depict the role of IPS and the pattern of the obvious DDoS attack.

Were there any other clues behind the loud DDoS and IPS deny events? Looking from the heat map view, shown in Fig. 9d, it is easy to overlook that eight internal hosts were a little more active than other internal hosts when the network gradually cooled down after rush hour at noon on April 11. When utilizing the radial view to analyze alerts of these eight hosts, shown in Fig. 10a, it is noticeable that IPS deny alerts in IPS logs kept emerging for a long time and TCP flooding alerts in NetFlow logs also emerged for many times, which intensifies our doubts on the these eight hosts. Looking into the detailed records of these eight

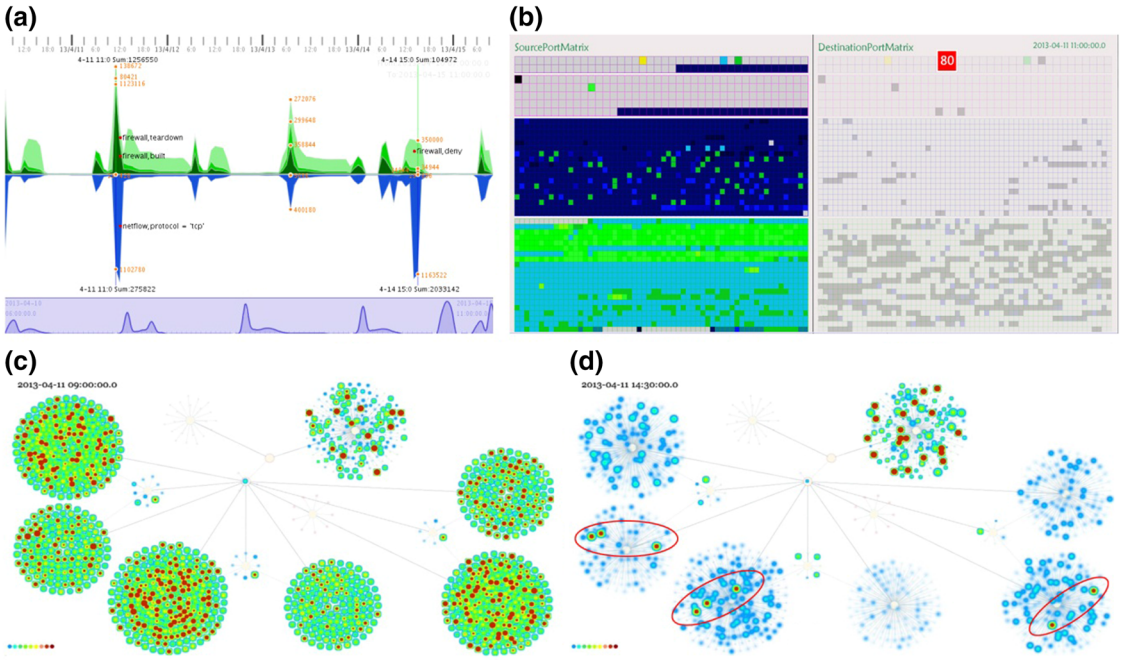

Fig. 9 The DDoS attack around 11:00 April 11. **a** A comparison between IPS operation and TCP traffic. **b** The pattern of port activities of DDoS attack. **c** The TCP traffic explosion on the morning of April 11. **d** Eight suspicious internal hosts in the heat map view

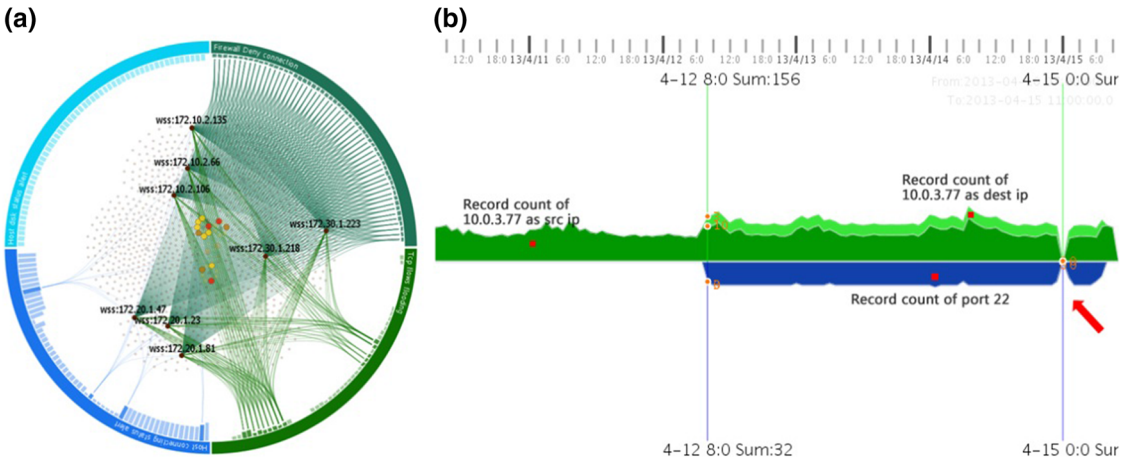

Fig. 10 The botnet infection. **a** The connection deny alerts of eight internal hosts. **b** The connections between 10.0.3.77 and eight internal hosts on SSH port 22

suspicious hosts in database, we find that these eight suspicious hosts periodically connected to external host 10.0.3.77 through SSH (Secure Shell Protocol) port 22. Hence, we turn to the stacked stream view to look for details about the external IP 10.0.3.77 and port 22 in NetFlow logs. Shown in Fig. 10b, both of activities on port 22 and connections to IP 10.0.3.77 in the whole network began around 8 o'clock on April 12, which means that port 22 was only utilized to communicate between 10.0.3.77 and these eight hosts. Therefore, these phenomena successfully verify our assumption that non-IRC botnet is very likely to be infected soundlessly in the enterprise intranet.

How serious impact was brought to the enterprise intranet by the infected botnet? The interruption of connections on port 22 in Fig. 10b gives us a tiny clue of possible network downtime, which is proved by the disappearance of network traffic during 23:49 April 14 to 1:43 April 15, shown in Fig. 11a, b. By the

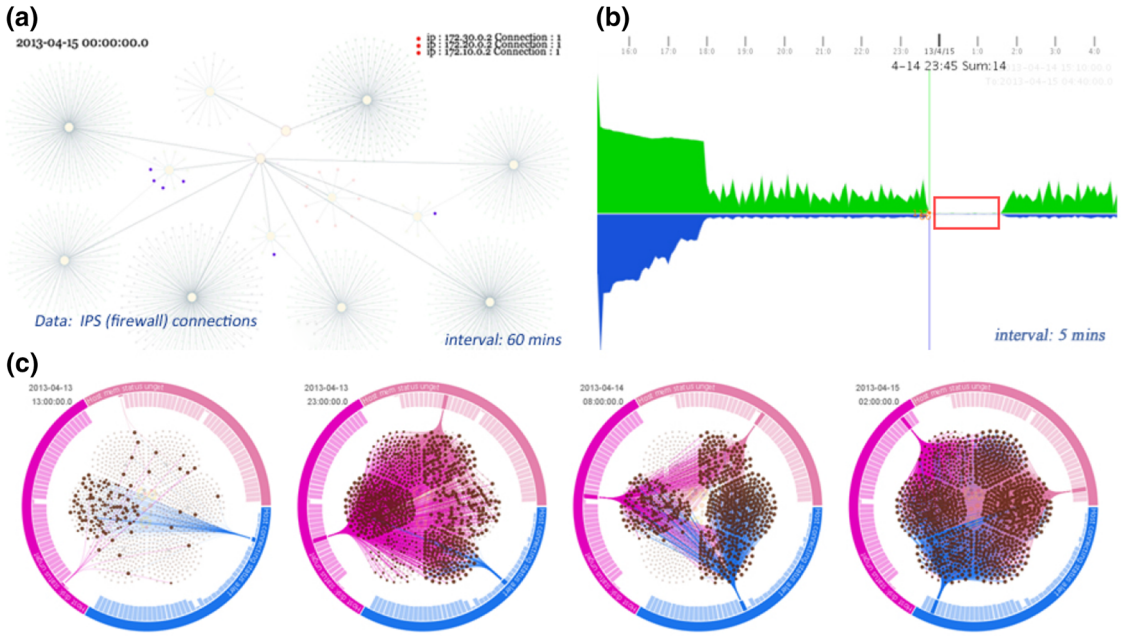

Fig. 11 The network downtime around 0:0 April 15. **a** The connections in IPS logs. **b** The network traffic in NetFlow logs. **c** The host health alerts increased from April 13 to April 15 in Big Brother logs

radial view, shown in Fig. 11c, when we select four periods of the second week and draw three types of host status alerts, we will find more and more these alerts emerged in the intranet, which represents that the condition of the whole intranet got worse and worse after the infection of botnet.

As a conclusion of two case studies, we'd like to deduce the whole storyline and assess the situation in 2 weeks. Shown in Fig. 12, the colored time line represents our evaluation of network situation based on all events in 2 week, and the important loud events and subtle events are arranged along the time line. The network infrastructures of the Big Marking enterprise had been continued undermined by hackers or competitors for 2 week. In the first week, web servers and mail servers suffered a lot from the repeated DDoS attacks which were raised by many external hosts from April 2 to April 3, and then the crash of web server 172.30.0.4 as victim of DDoS attacks was very likely to be utilized by sly attackers to plant malware into the enterprise intranet from April 4 to April 5. At the end of the first week, the data exfiltration and the dissemination of spam, as the evidences of the destructive activities of malware, were found. In the second week, IPS was deployed to block those known attacks after maintenance of network from April 8 to April 10; therefore the attackers had to adapt their attack strategy. The multiple DDoS attacks on April 13 and April 14 are highly suspicious to related with the botnet infected in the morning on April 12, and the network situation was getting worse and worse at the end of the second work that the final downtime in the morning on April 14.

6 Discussion

The development tools used to implement MVSec include: Processing, Eclipse, MySQL. In this section, we will discuss lessons learned in designing and developing the MVSec system, and share our experience of visual analytical process for the VAST Challenge 2013 contest datasets, which would be helpful and applicable to other visual analytics applications.

6.1 Advantages of the system

Because of its clear analytical goal and superior data quality, the Mini Challenge 3 of VAST Challenge 2013 performed a perfect platform to evaluate MVSec's efficiency. Compared to other contestants, the advantages

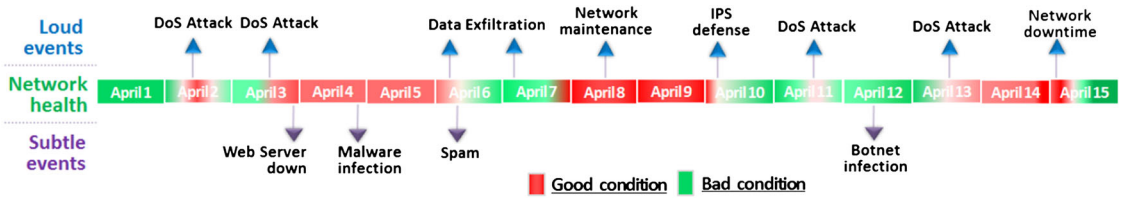

Fig. 12 The storyline and network situation evolution of the Mini Challenge 3 of VAST Challenge 2013

of our solution can be concluded in three aspects: the highest accuracy of anomaly detection, the discovery of some subtle events and the story-telling in a context.

The fusion of multiple heterogeneous datasets lays the solid foundation of the entire system, which provides metadata with unified format and compresses "big data" into "small data". As the strong adaptive faculty of metadata for its unified data structure, metadata can be called by any visualization view and hence these views are able to analyze data collaboratively from different sources. On the other side, dataset from the same source can be visualized on different views, which makes it possible to be linked with other data from different perspectives and thereby avoids information isolation and increases the available clues in analysis on events.

Multiple coordinated views are the main strategy for cooperative anomalies detection. The multiple perspectives brought by them do not only help users differentiate between events which require normal maintenance and events which are critical threats, but also differentiate between events that represent infrastructure failures, malicious activity against the network, or other aberrant behavior, and further help users to understand connections between events and their locations within the computer network. The loud events always show up as massive changes in colors or in other visual attributes in our visualization design, making them very hard to be missed in our four specialized views.

Contextual analysis deduced by visual perceptions in visualization views is a novel approach for visual analysis, which provides analysts with more tiny clues and the detective viewpoints to trace back the development of events step by step and even unfold the whole story. In this process, we strive to find all possible hidden clues in every identified attack by using multiple views so that some important subtle events are able to be pulled out, such as web server crash, spam explosion, and botnet infection. Furthermore, it helps analysts to draw an analytic storyline and make a better understand of network situation.

6.2 Problems of the system

We critically assess the system in four aspects: scalability, usability, collaboration and situation awareness.

6.2.1 Scalability

Even though the two-level layout of network logic topology is responsible for a mid-size network like the Big Marketing, it seems that the heat map and radial graph clusters would become too large and unwieldy if considering a network of 1 million hosts. Additionally, G-level data provided in MC3 can't be considered seriously as truly big data, and the system efficiency is good when running on the datasets after preprocessing, but it is not efficient enough to meet the operation on the raw datasets. Consequently, more research on the extreme large scale network and dataset is necessary.

6.2.2 Usability

Usability refers to the ease with which users can execute the desired tasks in our application. One usability issue about our system is that ordinary users may don't know how to get start, because it is only suitable for expert users who can acclimate to the variety of settings and interactions. Another usability issue is that a web-based graphical user interface should be provided to users for the easily access from any workstation.

6.2.3 Collaboration

Despite the multiple separated views working in coordination lend much to the analysis, shown in Fig. 13, they bring into users onerous actions to move between different visualization views. Currently, the timeline

Fig. 13 Using system with four tied displays in the control room

is the only way to coordinate with multiple views, so how to reach the efficient collaboration is an urgent problem. One solution is to develop a single integrated picture for the large-screen display, and another one is to evolve a distributed and co-located collaborative visual analytics (Grinstein et al. 2011).

6.2.4 Situation awareness

Almost all views of system are focusing on understanding and characterizing particular events, but administrators may prefer to have an overview picture about the current performance of the network, about the state of information flows, and about the state of network infrastructure at a functional level. A higher level indicator of situation awareness is needed to show the state of the entire network, even though one strength of our system is to help analysts to understand situation better by analyzing related sets of events in context.

7 Conclusions and future work

In this article, we presented MVSec, a visual analytics system for computer network security. Our system establishes on the data fusion from multiple heterogeneous security datasets, and the visualization is composed of four view: a heat map view to present network traffic on hierarchical layout of network logic topology, an event radar view to explore correlations of security events, a comparative stacked stream view to investigate implicit information within multiple time series, and a port matrix view to depict the characteristics of port activities. The combination of four visualization parts are quite effective in providing analysts with diverse aspects of data to characterize anomalies, investigate event correlations and conduct a deductive visual analysis in context. In the process of solving the Mini Challenge 3 of VAST Challenge 2013, the system played a significant role in helping analysts draw an analytic storyline and make a better understand of network situation. Another part of our contributions is that the experiences we shared in visualization design and actual experiments in MC3. We hope that it is useful not only to similar network security systems but also to other domains with visual analysis challenges.

In the future, we plan to continue research in the better collaboration way of four views, continue working on providing the higher level indicator of situation awareness and web-based interface for security analysts, and further investigate the better scalability and usability to implement a more robust of visual analytics application for computer network security.

A demo video can be found at http://www.youtube.com/watch?v=XehlHdDrNMk.

Acknowledgments This work was supported by National Natural Science Foundation of China (Grant No. 61103108), Hunan Provincial Science and Technology Program (Grant Nos. 2012RS4049), Hunan Provincial Natural Science Foundation of China (Grant No. 12JJ3062), and Postdoc Research Funding in Central South University. The authors would also like to thank the data providers, IEEE VAST Challenge.

References

Bass T (2000) Intrusion detection systems and multisensor data fusion[J]. Commun ACM 43(4):99–105

Cook K, Grinstein G, Whiting M et al (2012) VAST challenge 2012: visual analytics for big data[C]. In: Proceeding of the 2012 IEEE conference on visual analytics science and technology (VAST). IEEE, New York, pp 251–255

Dumas M, Robert JM, McGuffin MJ (2012) Alertwheel: radial bipartite graph visualization applied to intrusion detection system alerts[J]. Netw IEEE 26(6):12–18

Erbacher RF (2012) Visualization design for immediate high-level situational assessment[C]. In: Proceedings of the ninth international symposium on visualization for cyber security. ACM, New York, pp 17–24

Finamore A, Mellia M, Meo M et al (2011) Experiences of internet traffic monitoring with tstat[J]. Netw IEEE 25(3):8–14

Fink GA, Muessig P, North C (2005) Visual correlation of host processes and network traffic[C]. In: IEEE workshop on visualization for computer security, 2005 (VizSEC 05). IEEE, New York, pp 11–19

Fischer F, Fuchs J, Vervier P A et al (2012) VisTracer: a visual analytics tool to investigate routing anomalies in traceroutes[C]. In: Proceedings of the ninth international symposium on visualization for cyber security. ACM, New York, pp 80–87

Fischer F, Fuchs J, Mansmann F et al (2013) BANKSAFE: visual analytics for big data in large-scale computer networks[J]. Inform Vis

Ghidini G, Das S K, Gupta V (2012) Fuseviz: a framework for web-based data fusion and visualization in smart environments[C]. In: Proceeding of the 2012 IEEE ninth international conference on Mobile Adhoc and Sensor Systems (MASS). IEEE, New York, pp 468–472

Goodall JR (2008) Introduction to visualization for computer security[M]. In: VizSEC 2007. Springer, Berlin, pp 1–17

Grinstein G, Cook K, Havig P et al (2011) VAST 2011 challenge: cyber security and epidemic[J]. IEEE VAST 2011:299–301

Havre S, Hetzler E, Whitney P et al (2002) Themeriver: visualizing thematic changes in large document collections[J]. IEEE Trans Vis Comput Graph 8(1):9–20

Koike H, Ohno K, Koizumi K (2005) Visualizing cyber attacks using IP matrix[C]. In: IEEE workshop on visualization for computer security, 2005 (VizSEC 05). IEEE, New York, pp 91–98

Lakkaraju K, Yurcik W, Lee AJ (2004) NVisionIP: netflow visualizations of system state for security situational awareness[C]. In: Proceedings of the 2004 ACM workshop on visualization and data mining for computer security. ACM, New York, pp 65–72

Li B, Springer J, Bebis G et al (2013) A survey of network flow applications[J]. J Netw Comput Appl 36(2):567–581

Livnat Y, Agutter J, Moon S et al (2005) Visual correlation for situational awareness[C]. In: IEEE symposium on information visualization, 2005. INFOVIS 2005. IEEE, New York, pp 95–102

Mansmann F, Keim DA, North SC et al (2007a) Visual analysis of network traffic for resource planning, interactive monitoring, and interpretation of security threats[J]. IEEE Trans Vis Comput Graph 13(6):1105–1112

Mansmann F, Keim DA, North SC et al (2007b) Visual analysis of network traffic for resource planning, interactive monitoring, and interpretation of security threats[J]. IEEE Trans Vis Comput Graph 13(6):1105–1112

Manyika J, Chui M, Brown B et al (2011) Big data: the next frontier for innovation, competition, and productivity[J]

McPherson J, Ma KL, Krystosk P et al (2004) Portvis: a tool for port-based detection of security events[C]. In: Proceedings of the 2004 ACM workshop on visualization and data mining for computer security. ACM, New York, pp 73–81

Patcha A, Park JM (2007) An overview of anomaly detection techniques: existing solutions and latest technological trends[J]. Comput Netw 51(12):3448–3470

Plonka D (2000) FlowScan: a network traffic flow reporting and visualization tool[C]. In: LISA, pp 305–317

Ren P, Gao Y, Li Z et al (2005) IDGraphs: intrusion detection and analysis using histographs[C]. In: IEEE Workshop on visualization for computer security, 2005. (VizSEC 05). IEEE, New York, pp 39–46

Shiravi H, Shiravi A, Ghorbani AA (2012) A survey of visualization systems for network security[J]. IEEE Trans Vis Comput Graph 18(8):1313–1329

Taylor T, Brooks S, McHugh J (2008) NetBytes viewer: an entity-based netflow visualization utility for identifying intrusive behavior[M]. In: VizSEC 2007. Springer, Berlin, pp 101–114

Teoh ST, Ma KL, Wu SF et al (2002) Case study: interactive visualization for internet security[C]. In: Proceedings of the conference on Visualization'02. IEEE Computer Society, pp 505–508

VAST Challenge 2013 (2013) Situation awareness and prospective analysis[C]. In: IEEE conference on visual analytics science and technology (VAST). IEEE, New York

Walker R, ap Cenydd L, Pop S et al (2013) Storyboarding for visual analytics[J]. Inform Vis

Yin X, Yurcik W, Treaster M et al (2004) VisFlowConnect: netflow visualizations of link relationships for security situational awareness[C]. In: Proceedings of the 2004 ACM workshop on visualization and data mining for computer security. ACM, New York, pp 26–34

Zhao Y, Zhou FF, Fan XP et al (2013) IDSRadar: a real-time visualization framework for IDS alerts[J]. Sci China Inform Sci 1–12

VAST Challenge Homepage [EB/OL]. http://www.vacommunity.org/VAST+Challenge+2013