A Visualization System for Multiple Heterogeneous Network Security Data and Fusion Analysis

Sheng Zhang[12], Ronghua Shi[1] and Jue Zhao[2]
[1] School of Information Science and Engineering, Central South University
Changsha, Hunan, China
[e-mail: 48209088@qq.com,shirh@csu.edu.cn]
[2] Modern Educational Technology Center, Hunan University of Commerce
Changsha, Hunan, China
[e-mail: 48209088@qq.com,yayagogo@126.com]
*Corresponding author: Sheng Zhang

Received November 16, 2015; revised February 14, 2016; accepted March 26, 2016; published June 30, 2016

Abstract

Owing to their low scalability, weak support on big data, insufficient data collaborative analysis and inadequate situational awareness, the traditional methods fail to meet the needs of the security data analysis. This paper proposes visualization methods to fuse the multi-source security data and grasp the network situation. Firstly, data sources are classified at their collection positions, with the objects of security data taken from three different layers. Secondly, the Heatmap is adopted to show host status; the Treemap is used to visualize Netflow logs; and the radial Node-link diagram is employed to express IPS logs. Finally, the Labeled Treemap is invented to make a fusion at data-level and the Time-series features are extracted to fuse data at feature-level. The comparative analyses with the prize-winning works prove this method enjoying substantial advantages for network analysts to facilitate data feature fusion, better understand network security situation with a unified, convenient and accurate mode.

Keywords: Network security visualization; multi-source data; visual fusion analysis; heatmap; treemap; radial node-link; labeled treemap ; time-series

This work was supported by National Natural Science Foundation of China (Grant No. 61402540).

http://dx.doi.org/10.3837/tiis.2016.06.019 **ISSN : 1976-7277**

1. Introduction

With the continuous development of network technology and the expansion of its applied range, the network has become an important driving force of social progress. However, network environment is worsening, and the security problems are becoming more and more serious. Although the traditional single source detection systems (such as IDS, Firewall, Netflow, etc.) have improved the security of the network to a certain extent, but due to the lack of an effective cooperation with each other, these systems are unable to monitor the security situation of the entire network. How to comprehensively analyze security states and events of various equipment, and how to rapidly grasp the network security situation are the important challenges of the modern administrators. Under the background of this need, fusion analysis of multi-source data comes into being, and quickly became a hot research topic in the field of network security.

Domestic and foreign experts and scholars have carried out a lot of work on modern network security research and achieved some results, but the study of the multi-source heterogeneous data sources is not yet mature, and needs to solve key technical problems such as framework model, data preprocessing, quantitative perception, analysis and decision system, etc. New fusion analysis methods are constantly emerging, such as: statistics, D-S theory [1], ontology [2], domain-knowledge [3],etc. As a front research field, the network security visualization brings human ability of strong pattern recognition to the network security analysis field, and has got good effects. Since the international conference on the Visualization for Cyber Security has held in 2004, a large number of visual security tools are developed, and have been playing an important role in aspects of network security safeguard and analytic decision making.

The increasing size of networks and continuous appearance of new types of attacks challenge the research on visualization for network security [4]. The first challenge is how to avoid the serious jams in information sharing and coordination treatment as a result of the increased network security equipment in the modern defense system, the fast-growing data and the more and more complex data structures. Detection, analysis, measurement, prevention and treatment of these data are great time and space consuming. One of the solutions is to reduce the processing cost with a reasonable selection of representative, related, real-time data and extraction of data features.

The second challenge is how to establish proper visualization framework. There are many visualization technologies for network security presentation, such as parallel coordinates, scatter plots, node-link, treemap, heatmap, etc. Each visual technology has its own merits and is suited for a specific analysis scenario. Different security events have different characteristics in different graphics. How to collaboratively analyze a variety of graphics, and find the hidden characteristics, relationships, patterns are the key issues.

The third challenge is how to enhance visual fusion analysis to fully grasp the security situation. Although, different visualization techniques can show different sources of data, and find anomalies in them. What they often provide is the original and lower level views. It is difficult for policymakers to make a real-time and comprehensive judgement. Through the effective combination of the data mining technology and visual technology, innovation and improvement of the existing visual model, we can show the state of the entire network and the security situation on a well-designed graph. The use of novel, practical and comprehensive

graphics to display and fuse the multi-source data network security features is a more advanced research direction.

In Section 2, we review the previous work in visualization technologies and multi-source network security systems. Section 3 addresses the appropriate selection methods of data sources. In Section 4, we choose different visualization technology for different data sources based on their characteristics. Section 5 fuses all the data mentioned above into a well-designed graph to comprehensively analyze data and grasp security situation. In Section 6, we list the advantages and disadvantages of this system. Finally, we provide conclusions in Section 7.

2. Related Work

2.1 Multidimensional visualization technologies

Parallel coordinates: it maps n attributes data to a two-dimensional space through n parallel axis, and each axis representing an attribute dimension. Parallel coordinates can intuitively express relationships between data and easy to understand. But huge amounts of data can generate a lot of overlapping lines, thus difficult to identify. The order of the parallel coordinates' axes is also the important factors that influence the relationships and trends between data.

Scatterplot matrix: scatterplot matrix is one of the most commonly used visual methods that convert the high-dimensional data into two-dimensional data. The multidimensional data are paired off and combined with an element in the matrix, which overcomes the difficulty of representing high-dimensional data onto a flat surface from a certain extent. This directly explains the relationship between any two dimensions without being affected by the size and dimension of datasets. The disadvantages are that when the dimension increases, the matrix will be constrained by the size of the screen; it can find the relationship between the two dimensions, but difficult to discover the relationship between the multiple dimensions.

Glyph: the basic idea is to use icons with visual features to express multidimensional information, and every visual feature of the icon representing one dimension of the multidimensional information. It is suitable for data with limited dimension containing special meaning, and it has good expansibility in two-dimensional space. Users can have a more accurate understanding of the meaning of these dimensions according to the icons. But the data set depicted by the glyphs is only visualized in a discrete manner. Therefore, the visualization designer should carefully review the visualization requirements before using glyphs [5].

Node-link: it organizes the hierarchical data into a tree-like connection structure. Its nodes and edges represent data items and the relationships between them. The node usually is a small point which is difficult to contain more information. Node-link clearly and intuitively shows the relations in hierarchical data, but the gap between the edges will waste exhibition space. When the size of the data becomes very large, the edges will soon be crowded and easily generate visual confusion.

Timeline: timeline uses the time dimension for the horizontal axis, with the information displayed on the horizontal axis according to chronological order. As the time range is too long, it is difficult to fully display important details in the finite length of the axis.

Treemap: Treemap uses a series of nested rings or blocks to display hierarchical data. It can show a large amount of data within the limited space, and avoid congestion. But Treemap is

unable to display the node details, and the relationship between the nodes is not clear. In addition, Treemap is not familiar and difficult to understand for ordinary users.

The combining two or more visualization technology, using both their respective advantages, and making up for the shortcomings, has obtained a good visual effect. For example: Elena Fanea [6] combines parallel coordinates with star glyphs, the Parallel-Glyphs method, to eliminate the connecting lines overlap problem, and improve the ability of multi-dimensional data display; Michael McGuffin [7] using adjacency matrix and node-link, the combination graph, both eliminate the phenomenon of mutual crisscross and can easily find path between the nodes; IDSRadar [8] combined with radical panel, node-link and histogram, greatly extends the expression dimensions and enables more details on the screen. This study tries to employ the combination of Glyph and Treemap to improve detail-displaying ability of multi-dimensional data, and use characteristic Time-series to improve the ability of situational awareness.

2.2 Existing multi-source visualization network security systems

IDSRadar [8], a blend of intrusion detection system (IDS) and firewall system data, uses a variety of visualization techniques to help analysts identify the true anomaly pattern from a large number of false positives. The weaknesses lie in the poor scalability of the topology mapping especially when there are too many IPs to fit into the inner circle, and the single view of radial graph should be extended to multiple views to show more details.

Elvis [9] visually display various logs by importing the TCP dump packets, intrusion detection system data, and operating system SysLog and so on. Analysts can easily choose proper characteristics to analyze. The limitation is that each dataset is isolated from the others and multiple datasets cannot be combined for exploration.

Mansmann Florian [10], integrating intrusion prevention System (IPS) and Netflows, using Treemap for large-scale network traffic monitoring intend to enhance the tool to support real-time data collection and analysis via scalable data management technology

MVSec [11], by establishing the data fusion strategy for multiple heterogeneous security datasets, employs four graphics to characterize anomalies. The future work will process extremely large scale heterogeneous datasets, improve collaboration of different graphics, and provide the higher level indicator of situation awareness.

ReView [12], a tool with a special visual locality property, supports different levels of visual based querying and reasoning required for the sense making process on complex network data. Security analysts can use ReView to identify abnormal network activities and patterns resulting from attacks or stealthy malware.

Visual multi-source analysis systems, still at its novel stage, require the development of a lot of methods and technologies to support them. The main issues of modern Multi-source network security systems focus on low scalability, weak support on big data, insufficient data collaborative analysis, inadequate situational awareness, image occlusion and crowdedness, etc. This study tries to carefully select complementary data sources and fuse all data's features to form a comprehensive system for situational awareness.

3. Preparation for multi-source network security data

Modern network security systems are often integrated defense systems, which are made up of all kinds of safety equipment and produce huge amounts of heterogeneous data. These data are the basis of accurate judgment of network security situation and prediction of the future trend. Security equipments place in different positions according to their utility, and their data are

complicated and in various formats due to the historical data accumulation and new data generated in the running process. Therefore, how to choose typical, high reliable, real-time, low redundant data sources with large information capacity is the key to efficient and comprehensive analyses of network security situation.

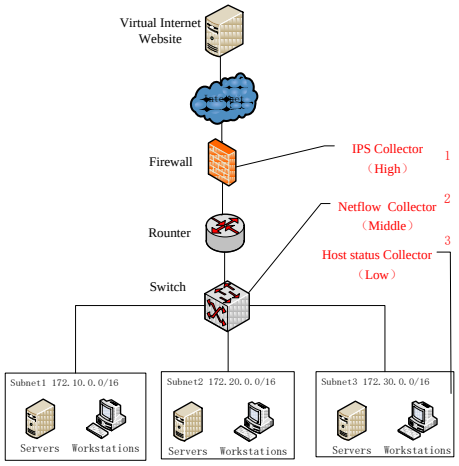


Fig. 1. Collection locations for multi-source network security data

According to these principles, we collect the logs of different security system at their installation locations. As shown in **Fig. 1**, the first kind of data is from the terminal equipment (low) such as workstations and servers' health records. They act as a mirror for operating systems, applications and the status of the host. It is very important in recording system status, monitoring system activities and maintaining system security. The second type of data is come from the network lines (middle), which record the network packages and flow information. They indicate the traffic load of network lines. The changes of the network load directly affect the change of network status. The third type of data is from the equipment, a watchdog for the input and output of network gate (high). They block attacks and reduce the internal network risks. When choosing network security dataset, we suggest giving considerations to the above three levels, so that these data sources are complementary and allow analysts to have a comprehensive perception of the real-time network changes. The main data fields are listed in **Table 1** and come from VAST Challenge 2013 competition.

Table 1. Multi-source Network Security Data

Source name	Main Data Fields	Location
Host status logs	DateTime\priority\operation\messageCode\protocol\srcIp\destIp\srcPort\destPort\destServic\direction\flags\command	Terminal equipment (low)
Netflow	ParsedDate\dateTimeStr\ipLayerProtocol\SrcIP\DestIP\Port\DestPort\durationSeconds\SrcPayloadBytes\DestPayloadBytes\SrcTotalBytes\nDestTotalBytes\SrcPacketCount\DestPacketCount\recordForceOut	Network lines (middle)
IPS system	ParsedDate\Hostname\Servicename\Currenttime\statusVal\Bbcontent\Receivedfrom\diskUsagePercent\pageFileUsagePercent\numProcs\loadAveragePercent\physicalMemoryUsagePercent\connMade	Web portal (high)

Data pre-process is an important step in visualization, especially when faced with a large number of heterogeneous data collecting from different network security devices. In this study, the data pre-process mainly consists of three parts: data cleaning, data integration and data transformation:

There is difficulty in ascertaining valid data due to data redundancy and the different schema definition of data coming from different sources. To avoid this, first of all, we remove the unrelated fields of the source data tables, filter invalid records, discretize the numeric fields, and process time synchronization of data sources. This is our data cleaning. In the step of data integration, we combine data from numerous heterogeneous devices into a coherent process that can be used to evaluate the overall situation of cyberspace. Thus, we extract security events and data statistics from multi-source logs as metadata. At last, data transformation converts data into a form suitable for mining. We construct new properties according to the needs of analysis to help understand the characteristics of the data, and process data standardization for needs of visualization to make the image reasonable and beautiful.

4. Visualization of multi-source data

4.1 Host status visualization

For visualization, Heatmap is by far the most popular graph, which compacts large amounts of information into a small space to bring out coherent patterns in the data [13]. It can be applied to medical services [14, 15], traffic [16], social network [17], etc. Moderately large data matrices (several thousand rows/columns) can be displayed effectively on a high resolution color monitor and even larger matrices can be handled in print or in megapixel displays [18]. Brewer [19] divides color maps into three classes: qualitative, sequential, and diverging. In this system the red and blue sequential mode is used for the Heatmap. Each host's health value is indicated by color saturation, which uses a blue-red gradient, as shown in **Fig. 2**. Deep blue indicates host's health value is 0, and deep red is 100, when the host's health value is too high or too low, it indicates abnormal. Meanwhile, the temporal characteristic is reflected on the X-axis horizontally, and hosts which have undergone some important or significant changes are demonstrated vertically using TOP N method (select N recorders at the front with column sorting). From the hotspot's distribution in **Fig. 2**, the analyst can visually detect 9 issues, which are located in bluer or redder sections. The problems in Phase 1 (IPS not installed) are more often than those in Phase 2 (IPS installed). The installation of IPS is real and effective to protect critical internal hosts, but it is not a panacea umbrella, the entire network system security requires more protection measures.

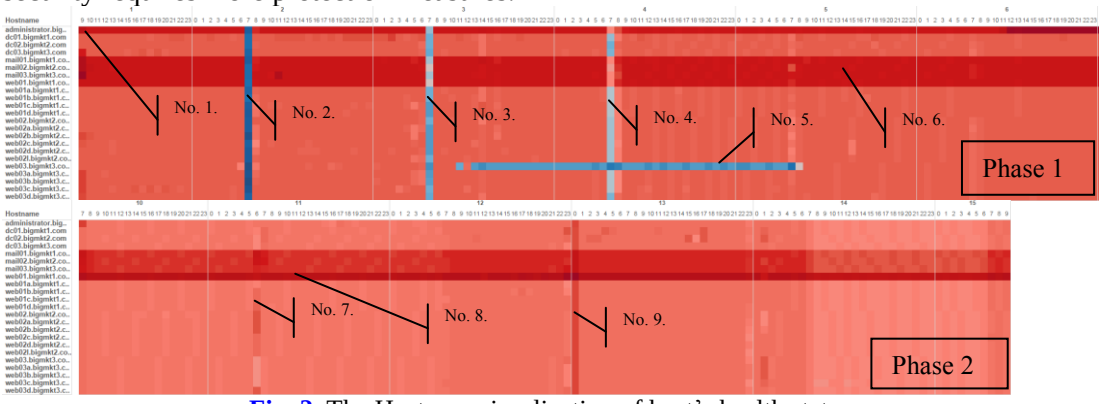
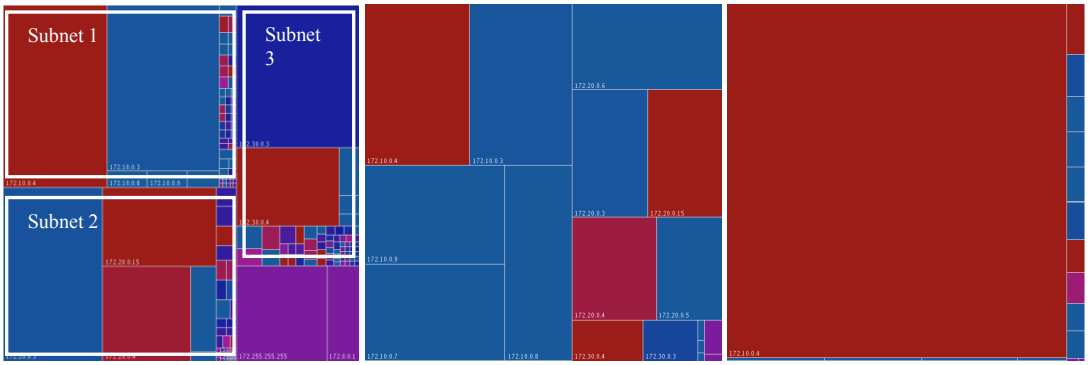


Fig. 2. The Heatmap visualization of host's health status

4.2 Netflow visualization

Treemap is introduced by Johnson in 1991, which is suitable for hierarchical data. Since its birth, Treemap is widely welcomed and deeply studied for many years. Now it has already changed from the original simple form to the better practical visualization methods, such as: Columnar Treemap [20], the Hilbert/Moore Treemap [21], Edge Equalized Treemaps [22], Spiral Tree Layout [23], Rectangular Tree [24]. In essence, traditional Treemap, a kind of mapping to present hierarchical data [25], effectively uses screen space and reduces the image occlusive, which helps to identify data hierarchy and node size in an easy way.

Netflow data, a kind of hierarchical dataset, its different properties are displayed by the location, size, and color in the Treemap. Here, we use node's location to represent the host position in a subnet. The hosts in the same subnet will be assigned to a same square area, as shown in **Fig. 3(a)**. The size and color of nodes can represent any two dimensions of the Netflow, and we can choose the dimensions, such as the number of source (destination) ports, number of source (destination) IP, number of source (destination) packets, source (destination) total bytes, etc., to observer the changes. In the **Fig. 3**, the size represents the number of destination packets and the color represents the destination total bytes. The bigger the size is, the greater the number of packets is. The redder the color is, the greater the value of total bytes is. Whether network status is normal or not depends on the Treemap distribution. If the graph (using Squarified layout algorithm [26]) is too concentrated or too fragmented, the network has a high probability of abnormal events. **Fig. 3(a)** is a graph of the normal network status, while **Fig. 3(b)** and **3(c)** are not. The normal network is represented as a left rectangular chunk accompanied by many small patches on the right or down sides in each subnet. This is because the target network consists of servers and clients. Servers are heavily loaded, and clients are lightly loaded. According to the Squarified layout algorithm, servers occupy the left larger space of each subnet, while other clients are shown in small patches which are tightly against the large area. In **Fig. 3(b)**, the evenly distributed rectangular blocks indicate that most of the host states are suffering relatively uniform attacks from outside, which conforms to the characteristics of port scanning. **Fig. 3(c)** is displayed as one or few large blocks. This is because few hosts are under massive network traffic attacks, a typical characteristic of Dos attack. However, some patterns will be a little similar, such as P2P and port scan (they establish a large number of connections) and DOS and centralized FTP access (they will carry a huge traffic flow). A more accurate distinction needs analysis of visual fusion.



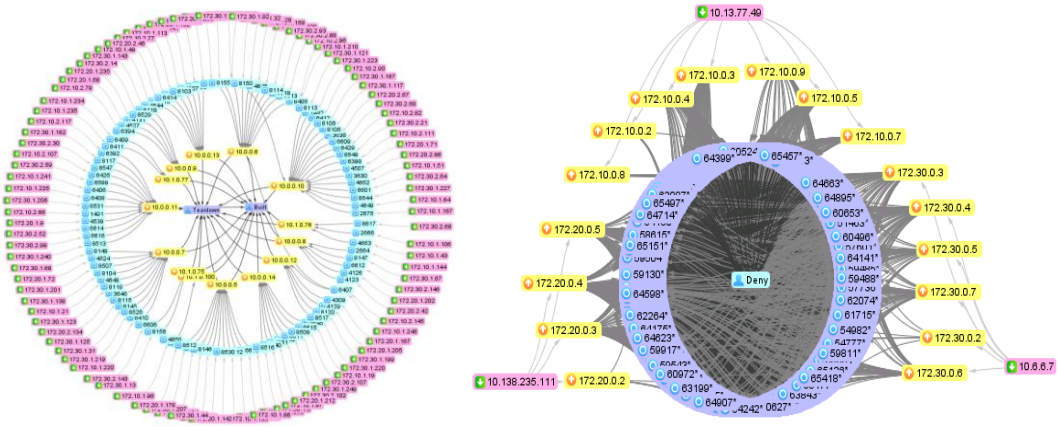
(a) normal network (b) port scan (c) Dos attack

Fig. 3. The Treemap visualization of Netflow

4.3 IPS visualization

Node-link graph, a connection structure that organizes hierarchical data into a tree, in which nodes and edges represent the dimensions and the relationship between them. It can be applied in network forensics [27], wireless sensor network configuration [28], host activity supervision [29] and so on. Thus, this study tries to improve node-link technology and the radial layout is adapted to the overall structure. Firstly, considering the different types of data hierarchy, different types of nodes are sequentially arranged in radiation ring of the circle. So that data nodes congestion is avoided and the crosses of edges are reduced. For easy distinction, different colors are used for different nodes. Edge bundling technology is employed for edges merged into the same node. Bezier curves [30] are used to bring these edges together, which can further avoid crossing and form convergence naturally, as shown in **Fig. 4**.

At the same time, by changing the data mapping sequence, we can directly build up, bring forth and analyze the network use cases. The left to right arrangement of data mapping model is transformed into outside to inside layout of radial node-link graph. Such as **Fig. 4(a)**, like the devil's eye, uses the "source IP-> source port -> target IP->action" visual mapping mode. It shows the internal hosts (red) access external servers (yellow) and uses a fixed port (blue) to share an Internet connection. **Fig. 4(b)**, like a frog's eye, which employs the "source IP -> target IP -> destination port -> action" mapping mode, reveals that external hosts (red) are probing into a large number of ports (purple) of internal hosts (yellow).



(a) client access (devil's eyes) (b) port scan (Frog's eye)

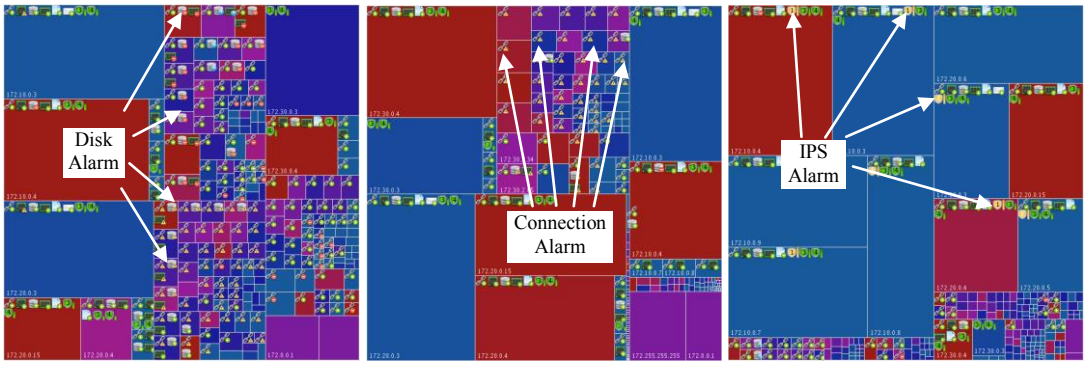
Fig. 4. The radial Node-link visualization of IPS

5. Visual fusion

5.1 Visual fusion and analysis at data-level

Data-level fusion processes raw data and provides the fusion of the local information [31]. It maintains the original information as much as possible. However, it brings about the problem of fusing large quantity of data in limited space. In a small area, there is insufficient space to express fusion data. Its advantage is that the information loss is small, and fusion accuracy is high. Its disadvantage is that data value is large, processing cost is high, and real-time process is weak.

In this study, due to the limitation on Treemap expression of the data source attributes(only position, size and color), Glyph is used to cover the drawbacks of Treemap, which makes data level fusion possible. Within Treemap rectangular space, icons are placed to represent memory, CPU, hard disk, page file, network connection of the host status, the alarm type and the quantity of the IPS. This kind of visualization technology is named Labeled Treemap. The green "+" icon at the bottom right corner of host status indicates normality; the yellow "!" indicates a warning; the red "-" indicates a problem; the blue "?" indicates an inability to receive status information. The color of IPS alerts shield expresses the severity of the warning. The green indicates no harm or slight harm. Yellow indicates secondary hazards and red indicates critical danger. Indicators on the right side of the shield express the number of alarms in 5 levels. When plenty of yellow or red icons appear, administrators should pay high attention. As shown in **Fig. 5 (a)**, the appearance of a large number of red or yellow disk alarms indicates that virus replicates itself furiously, resulting in a fast decline of disk space. The numerous yellow connection errors in **Fig. 5(b)** indicate that the network connection is not smooth. The yellow shields in **Fig. 5(c)** indicate that a large number of connections from outside are rejected and the internal network is being attacked.



(a) disk alarm (b) connection alarm (c) IPS alarm

Fig. 5. Data-level fusion on Labeled Treemap

5.2 Visual fusion and analysis at feature-level

Feature-level fusion extracts feature information and compresses the observable data. It reduces the demand of bandwidth, improves the anti-interference ability and prediction ability for the overall network. But it may lose some useful information and lead to the problem of reduced precision. Feature level fusion requires effective complementary features extracted from single mode data, and combines these characteristics and time-scale to organic combination, which converts single-mode data to united multi-modal data. Although Labeled Treemap can represent status of the network at a certain time, but the status is not clear at a glance due to large data to be shown. Moreover, it can't show the trend of the network change on the time dimension. In order to make up for the drawback of Labeled Treemap in grasping the trend of the future, we need to understand multi-source network security situation with other technologies. Time-series graph plays an important role in interpreting data, examining the major causes, and forecasting future circumstances [32] [33]. So this study uses the Time-series to integrate multi-source features.

Network security data from different sources have different attributes and characteristics. The choice of the method for feature fusion is very important. Netflow data can provide multi-level, multi-angle, multi-granularity traffic information in real-time with a full range of

network connections and network event information, etc. It has the characteristics of reflecting rapid change, being real-time, providing large amount of information, and enjoying high randomness. To eliminate redundancy and uncertainty, the method of information entropy is selected in this study. If the data are concentrated in one point, that is, all the data have the same value, the information entropy is 0. On the contrary, if the data are widely distributed, the information entropy will be great.

The status of the host can be indicated through memory, CPU, hard disk, page file, network connection, etc. Each indicator is in a reasonable range to show that host is in a normal state. When some indicators are beyond the scope, an abnormal event is suggested. We use comprehensive weighted method to obtain the host state from the indicators mentioned above. We assign a weighted factor to each attribute. The bigger the weighted factor is, the more important is for the attribute to decide the host status.

IPS system can monitor network or network equipment's transmission behavior, which can immediately interrupt, adjust or isolate some abnormal or dangerous network transmission behavior. One of the biggest challenges with IPS is the vast number of false positives. So two possible solutions help address this problem. The first solution is to install the IPS on the firewall as the combination of IPS with firewall helps reduce false positives; in the second solution, we adopt the method of statistics to summarize alarms in a certain time. Based on black lists, certain values known as false positives will be filtered out, and the number of true rejected connections will be counted.

Table 2. Feature fusion method

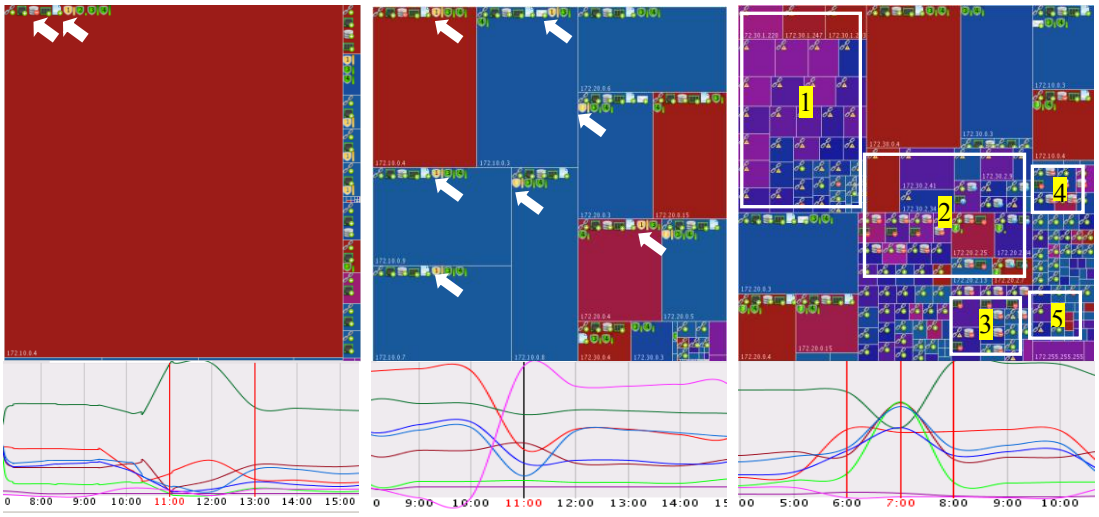
No	Source	Feature Name	Color	Fusion Method
1	Netflow	The source address（SrcIP） The destination address（DestIP） The source port（SrcPort） The destination port（DestPort） The source flow bytes per packet（SrcBpp） The destination flow bytes per packet（DestBpp）	Red Dark red Green Fresh green Blue Light blue	Information entropy
2	Host Health	Host status（HostStatus）	Light purple	Comprehensive weighted method
3	IPS	The number of rejected connections（IPSDeny）	Dark purple	Statistic

5.3 Fusion analysis

DDos controls a lot of computers as a platform to attack one or more target. Thus, the power of the attacks is exponentially increased. About 11:00 a.m. - 12:00 p.m. on April 11, 2013, Time-series has a drastic fluctuation, as shown in **Fig. 6(a)**. SrcPort (green) rises high, which means the source port number is soaring. SrcIP (red) shows a peak, which means an increase of attackers. But DestIP (dark red) and DestPort (fresh green) go down, which means the attack target and port are highly concentrated. In the Labeled Treemap, a large red rectangle occupies 90% of its space, which means the attack flow is quite centralized. Host 172.10.0.4 shows disk problems and IPS shows a large number of yellow alerts at the same time. Analysis result is that DDos caused great influence on the internal host 172.10.0.4, but the installation

of the IPS system succeeded in preventing the vast majority of DDos attacks on other internal hosts, making these attacks not very successful.

Switching to the figures before fusion, we can find more attack details. As shown in issue No.8 of **Fig. 2**, the host web01.bigmkt1.com(172.10.0.4) has been presented deep red, that is to say the host's load is heavy and resources are exhausted. At the same time, in the lower right corner of **Fig. 7(a)** appears a more severe alarm, level 4 ASA-4-106023, which indicates ACL blocked outside access to the real inside address. **Fig. 7(a)** helps make more explicit that the affected scope is focused on the lower right hosts (yellow).

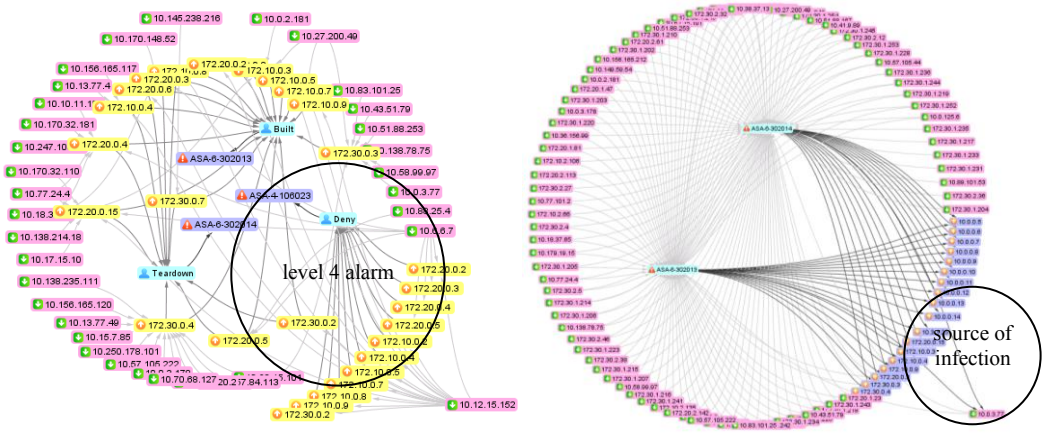


(a) DDos Attack (b) port Scan (c) virus infection

Fig. 6. Fused diagrams of Multi-source Network Security Data

Port scan is a method that the hacker likes. Hackers can detect weakness in the target by using it. Generally, it falls into horizontal scan (aiming at the same port of hosts for the entire network), vertical scan (aiming at all ports of a host), and mix of both (multiple ports of the entire network hosts). Port scan usually paves the way for other types of attack. About 11:00 a.m. on April 12, 2013, Time-series has a drastic fluctuation, as shown in **Fig. 6(b)**. The DestPort (fresh green) sinks to the bottom, which means that the same port of many internal hosts have been accessed by external parties. SrcBpp (blue) and DestBpp (light blue) show a fall, which means a large number of the same bytes package have been sent. The IPSDeny (light purple) shows a peak, which means a lot of connections are prohibited. Looking at the Treemap, image distribution is relatively uniform. And many yellow warning shields (IPSDeny) are emerged in rectangle, which almost cover the 172.10.X, 172.20.X, 172.30.X. The attack type should be port scan.

Switching to the **Fig. 4(b)**, more details are found. Three external hosts (red), 10.13.77.49, 10.138.235.111 and 10.6.6.7, are probing into a large number of ports (purple) of internal hosts (yellow). This is not a good start, which indicates the next round attacks against the network are about to begin.



(a) DDos Attack (b) virus infection(two rushing waterfalls)

Fig. 7. Pre-fusion cyber threat diagrams

Viruses (Trojans) are the No. 1 enemy of the modern Internet, which seriously hurt network security and undermine the stability of the network. As shown in **Fig. 6(c)**, 6: 00-8: 00 am on April 11, 2013 all of the curves exhibit strong fluctuations, an indication of rampant network activities. In the Labeled Treemap, a large number of red flag symbol "-" and yellow "!" appeared, and the problems focused on memory, hard disk and network connections (white box 1 to 5). It is estimated the hosts in the network are infected by viruses.

Issue No.7 in **Fig. 2** before the fusion shows some hosts arranged in a column become redder than ever. These hosts are infected because of a sudden increase in resource consumption. **Fig. 7(b)** like two rushing waterfalls indicates that the host 10.0.3.77 acted as SrcIP and DestIP produces a lot of warnings: ASA-6-302013 and ASA-6-302014 (abnormally build and teardown TCP connections). Through these connections, the 10.0.3.77 sends a lot of control packets to the internal hosts, causing the IPS to constantly trigger level 6 alarms (Botnet infection). At last we found the source of infection.

6. Comparison and evaluation:

6.1 Prize-winning works of the VAST Challenge 2013

In order to evaluate the practicability and validity of the visual fusion method, we conducted a comparative analysis with the prize-winning works of the VAST Challenge 2013. They all use the same data sources but select different visual technology to represent security events.

Table 3. Prize-winning works of the VAST Challenge 2013

No	Authors(Award)	Visualization Technology	Flaws
1	Ying Zhao, et al. (Outstanding Comprehensive Solution)	Stacked stream graph to statistically analyze Netflow; radial graph to guide targets and directions; matrix graph to analyze port usage.	Inadequate fusion analyses as the network problems are scattered in different figures.
2	Siming Chen, et al. (Outstanding Situation Awareness)	Radial graph to represent targets; Time-series graph to represent flow feature changes; parallel coordinates to fuse multi-source network security data.	Easy confusion of targets for directions under complex attack environ ment in radial graphs.

3	Chen Zhong, et al. (Noteworthy Collaborative Analysis Strategy)	Timeline-Heatmap to highlight the change of individual subjects in time; histogram to statistically analyze.	Inability for Heatmap to show the multiple dimensions of data at the same time.
4	Fabian Fischer, et al. (Intriguing Visualization)	Stacked stream graph to observe the overall situation; Treemap to represent the distribution characteristics of Netflow; Node-link to show the connection changes of the overall network.	Relatively coarse data processing, and weak ability in identifying anomalous events; mainly suitable for wide-screen display.

6.2 Advantages

The main advantage of this system lies in its visual fusion of multi-source network security data at the data-level and feature-level which can reduce the administrators' pressure of cognition and response and make visualized analysis more unified, convenient and accurate.

Compared with the running systems, such as IDSRadar, HoNe, NFlowVis, PortVis, BGPlay, etc., they focus only on one or a few kinds of data sources, and their detected problems are limited. These systems fail to control the entire network in terms of security situation. The data collected in our solution are from three levels of network architecture (security portal, network lines, terminal hosts), and hence are both widely representative and complementary. This kind of solution has formed a comparative complete multisource analysis system.

Compared with methods, which represent inner network by Dot Matrix, this system uses Treemap, a hierarchy structure, to manage large and super large networks so that problems such as inadequate display space, crowded graphs and image occlusive can be avoided. At the same time, The Glyph is imported to expand the expression dimension of Treemap and enhance the ability of the visual expression and data fusion.

Eight primary dimensions of heterogeneous network security data are extracted to show the network situation. According to the characteristics of different data sources, the corresponding feature extraction algorithm—information entropy, comprehensive weighted method, information statistics are used to draw time-series diagram so that analysts can grasp the network situation more intuitively and analyze attack mode more efficiently.

6.3 Limitations

We install the system in the campus network and try to find its shortcomings through this practical application.

Efficiency: Our fusion analysis depends on a large number of aggregated data preprocessing, such as the huge amounts of data access and feature extraction algorithm. Some administrators reported network attacks with logs explosion, such as DDos and virus infection, which lead to slow response and poor interactivity. In the following work, we plan to use distributed processing framework for big data to improve our response for large-scale data and real-time analysis.

Usability: Some stealthy attacks cannot be identified. For example, when a malware is residing in host machine and communicating with a remote server via HTTP protocol in a normal volume of data traffic, our method will possibly fail to detect the network anomaly. So administrators suggested that the system should provide functions for free choices of data sources, browsing and editing for raw data; a high level fusion need to be strengthened. We have fused image at data-level and feature-level. But the higher level: decision-level is what

we are aiming at. Some experts suggest a combination of machine's discovery with humans' discovery to enhance the ability of decide-making.

Expressiveness: Multiple visualization techniques are utilized in our solution, but they are all in two-dimensional space. In the future, we will try to use 3D plots to add more dimensions of data to an interface and make fusion analysis more intuitive and effective.

7. Conclusion

In this article, we presented a novel visual analytic system for network security data. Our system establishes three single-source views and one fuse view for different data source: a Heatmap view to present the variation of host status, a Treemap view to display the behavior pattern of Netflow, and a Node-link view to depict the characteristics of IPS. At last, the fuse view, using labeled Treemap and characterized Time-series, is constructed to display a variety of data on a graph and grasp security situation. Experimental results show that the fuse analysis method has a substantial advantage for network analysts to understand network security situation, identify anomalies, discover attack pattern and remove the false positives, etc.

References

[1] Y. Wei, Y. Lian, and D. Feng, "A Network Security Situational Awareness Model Based on Information Fusion," *Journal of Computer Research and Development*, vol. 46, no. 3, pp. 353-362, March, 2009.

[2] C. Si, H. Zhang, Y. Wang, and J. Liu, "Network Security Situation Elements Fusion Method Based on Ontology," in *Proc. of 7th International Symposium on Computational Intelligence and Design (ISCID)*, pp. 272-275, December 13-14, 2014. Article (CrossRef Link).

[3] H. Zhang, D. Yao, and N. Ramakrishnan, "Detection of Stealthy Malware Activities with Traffic Causality and Scalable Triggering Relation Discovery," in *Proc. of 9th ACM Symposium on Information, Computer and Communications Security (ASIACCS'14)*, June 4-6, 2014. Article (CrossRef Link).

[4] K. Cook, G. Grinstein, M. Whiting, M. Cooper, P. Havig, K. Liggett, B. Nebesh, and C. L. Paul, "VAST Challenge 2012: Visual Analytics for Big Data," in *Proc. of 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 251-255, October 14-19, 2012. Article (CrossRef Link).

[5] T. Ropinski, S. Oeltze, and B. Preim, "Survey of glyph-based visualization techniques for spatial multivariate medical data," *Computers & Graphics*, vol. 35, no. 2, pp. 392-401, April, 2011. Article (CrossRef Link).

[6] E. Fanea, S. Carpendale, and T. Isenberg, "An Interactive 3D Integration of Parallel Coordinates and Star Glyphs," in *Proc. of 2005 IEEE Symposium on Information Visualization* pp. 20-20, Octobet 23-25, 2005. Article (CrossRef Link).

[7] J. M. Michael, and S. Zhao, "Hybrid Visualization for Tree and Network Structures," *Communications for the CCF*, vol. 7, no. 4, pp. 8-13, April, 2011.

[8] Y. Zhao, F. Zhou, X. Fan, X. Liang, and Y. Liu, "IDSRadar: a real-time visualization framework for IDS alerts," *Science China Information Sciences*, vol. 56, no. 8, pp. 1-12, June, 2013. Article (CrossRef Link).

[9] C. Humphries, N. Prigent, C. Bidan, and F. Majorczyk, "ELVIS: Extensible Log VISualization," in *Proc. of 10th Workshop on Visualization for Cyber Security*, pp. 9-16, October 23-28, 2016. Article (CrossRef Link).

[10] F. Mansmann, F. Fischer, D. A. Keim, and S. C. North, "Visual support for analyzing network traffic and intrusion detection events using TreeMap and graph representations," in *Proc. of 2009 Symposium on Computer Human Interaction for the Management of Information Technology*, pp. 3, November 7-8, 2009. Article (CrossRef Link).
[11] Y. Zhao, X. Liang, X. Fan, Y. Wang, M. Yang, and F. Zhou, "MVSec: multi-perspective and deductive visual analytics on heterogeneous network security data," *Journal of Visualization*, vol. 17, no. 3, pp. 181-196, August, 2014. Article (CrossRef Link).
[12] H. Zhang, M. Sun, D. Yao, and C. North, "Visualizing Traffic Causality for Analyzing Network Anomalies," in *Proc. of 2015 International Workshop on Security and Privacy Analytics (IWSPA'15)*, March 02-04, 2015. Article (CrossRef Link).
[13] M. Andrade, "Heatmap," http://en.wikipedia.org/, 2008.
[14]C. G. A. Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61-70, September, 2012. Article (CrossRef Link).
[15] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, and D. Sonkin, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603-607, March, 2012. Article (CrossRef Link).
[16] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743-761, April, 2012. Article (CrossRef Link).
[17] R. Gove, N. Gramsky, R. Kirby, E. Sefer, A. Sopan, C. Dunne, B. Shneiderman, and M. Taieb-Maimon, "NetVisia: Heat map & matrix visualization of dynamic social network statistics & content," in *Proc. of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*, pp. 19-26, October 9-11, 2011. Article (CrossRef Link).
[18] L. Wilkinson, and M. Friendly, "The history of the cluster heat map," *The American Statistician*, vol. 63, no. 2, January, 2012. Article (CrossRef Link).
[19] C. A. Brewer, *Designing Better Maps: A Guide for GIS Users by Cynthia A. Brewer*: ESRI Press Redlands, Calif, 2005.
[20] J. Armitage, "Method and system for generating a columnar tree map," *Google Patents*, 2011.
[21] S. Tak, and A. Cockburn, "Enhanced spatial stability with hilbert and moore treemaps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 1, pp. 141-148, January, 2013. Article (CrossRef Link).
[22] A. Kobayashi, K. Misue, and J. Tanaka, "Edge Equalized Treemaps," in *Proc. of 2012 16th International Conference on Information Visualisation*, pp. 7-12, July 11-13, 2012. Article (CrossRef Link).
[23] I. Jusufi, A. Kerren, V. Aleksakhin, and F. Schreiber, "Visualization of mappings between the gene ontology and cluster trees," in *Proc. of 2012 SPIE 8294 Visualization and Data Analysis*, pp. 82940N-12, January 24, 2012. Article (CrossRef Link).
[24] H. Song, X. Cai, and Y. Fu, "Rectangular tree browser: A navigation and visualization tool for large hierarchies," *Journal of Information & Computational Science*, vol. 8, no. 2, pp. 354-361, February, 2011.
[25] X. Zhang, and X. Yuan, "Treemap Visualization," *Journal of Computer-Aided Design & Computer Graphics*, vol. 24, no. 9, pp. 1113-1124, December ,2012. Article (CrossRef Link).
[26] M. Bruls, K. Huizing, and J. J. V. Wijk, *Squarified Treemaps*: Springer Vienna, 2000.
[27] Z. TIAN, W. JIANG, and Y. LI, "A Transductive Scheme Based Inference Techniques for Network Forensic Analysis," *China Communications*, vol. 12, no. 2, pp. 167-176, February, 2015. Article (CrossRef Link).
[28] Q. He, F. Chen, S. Cai, J. Hao, and Z. Liu, "An efficient range-free localization algorithm for wireless sensor networks," *Science China Technological Sciences*, vol. 54, no. 5, pp. 1053-1060, March, 2011. Article (CrossRef Link).
[29] F. Mansman, L. Meier, and D. A. Keim, "Visualization of host behavior for network security," in *Proc. of 4th International Workshop on Visualization for Cyber Security*, pp. 187–202, October 29, 2007.

[30] J. H. Gallier, *Curves and surfaces in geometric modeling: theory and algorithms*: Morgan Kaufmann, 2000.
[31] K. CHEN, Z. ZHANG, and J. LONG, "Multisource Information Fusion:Key Issues，Research Progress and New Trends," *Computer Science*, vol. 40, no. 8, pp. 6-13, October, 2013. Article (CrossRef Link).
[32] M. Krstajic, E. Bertini, and D. A. Keim, "Cloudlines: Compact display of event episodes in multiple time-series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2432-2439, December, 2011. Article (CrossRef Link).
[33] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu, "RankExplorer: Visualization of ranking changes in large time series data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2669-2678, December, 2012. Article (CrossRef Link).

Sheng Zhang a Ph.D. candidate in information science and engineering of Central South University, is presently a software system analyst at Hunan University of Commerce, and a member of China Computer Federation. His research interests are network and information security, computer-aided learning, and network application.

Ronghua Shi is presently a Professor and the Vice Dean of the School of Information Science and Engineering of Central South University. His research interests include information security, quantum cryptography and network security.

Jue Zhao is presently an associate professor at the Hunan University of Commerce. She graduated from Hunan University with M.A. degree in 2008. Her research interests include computer-supported cooperative learning, e-commerce and network application.