International Journal of Mathematical, Engineering and Management Sciences
Vol. 3, No. 1, 9–16, 2018
https://dx.doi.org/10.33889/IJMEMS.2018.3.1-002

Big Data and IT Network Data Visualization

Lidong Wang

Department of Engineering Technology
Mississippi Valley State University, USA
E-mail: lwang22@students.tntech.edu

(Received January 16, 2017; Accepted April 3, 2017)

Abstract
Visualization with graphs is popular in the data analysis of Information Technology (IT) networks or computer networks. An IT network is often modelled as a graph with hosts being nodes and traffic being flows on many edges. General visualization methods are introduced in this paper. Applications and technology progress of visualization in IT network analysis and big data in IT network visualization are presented. The challenges of visualization and Big Data analytics in IT network visualization are also discussed. Big Data analytics with High Performance Computing (HPC) techniques, especially Graphics Processing Units (GPUs) helps accelerate IT network analysis and visualization.

Keywords- Big data, Visualization, Network intrusion detection, Graphics processing units (GPUs), Data mining, Machine learning.

1. Introduction

Visualization is very important. It is often required to 1) understand the problem, generate hypotheses and define the problem; 2) identify the structure (based on user knowledge); 3) help with changing data, incomplete or incorrect data; 4) bridge the semantic gap (bring in user knowledge); and 5) steer the analysis process in dealing with massive data (local optimization) (Keim, 2014). Data can be generally represented using graphs including curves and bar charts or pie charts. Data mining supports knowledge discovery by creating models, classifying and predicting, finding associations or hidden patterns, and visualizing the data mining results using tools. For example, with visualization tools in genetic data analysis, the interactions among complex biological structures, and alignments among genomic or proteomic sequences are very efficiently represented in the forms of graphs, transformed into different types of visual displays that are easy to be understood. Thus, visual data mining and visualization have become very important in the analysis of biological data (Han and Kamber, 2006).

Visualization has been used in detecting network intrusion and cyberattacks. Network intrusion identification approaches based on graphs have been developed. In the graphs, nodes represent the agents (such as client nodes and servers) in the networks; and edges represent the communications on the network (the edges can be weighted, capturing frequency or volume). A network graph is often dynamic and can be tracked; it is assumed that the communication behaviour of a node should change when it is attacked. The score of a node is collectively computed to measure its behaviour and activity; the activity score of a node is high if it links many active nodes. Therefore, the activity vector is actually the principal eigenvector of an adjacency matrix which represent a communication graph (Akoglu et al., 2015). For big data, its volume and rate are massive. Big Data analytics has the potential of applications in many areas such as Additive Manufacturing (AM) and supply chains (Wang and Alexander, 2016a). In the methodology of Big Data analytics, most machine learning methods cannot be used in big data; there are many challenges of machine learning in Big Data analytics (Wang and Alexander, 2016b). New tools are needed, especially for big data, to correlate with external and internal data sets, interactively visualize data, and conduct deep analysis of content (Madden, 2014).

The purpose of this paper is to present advances of IT network visualization, especially Big Data analytics in IT network visualization. The organization of the paper is as follows: the second section introduces general methods in visualization; Section 3 introduces applications and technology progress in visualization; Section 4 discusses the challenges in visualization; and the final section is conclusion and future work.

2. General Methods in Visualization

Generally, data visualization components are achieved through the following phases: 1) monitoring-the characteristics of this phase are overview displays, simple displays, and flexibility; 2) analysis and diagnosis-the characteristics of this phase are filtering and interaction, exploration, multiple levels and views of data, and various data sources and correlation. Analysts often like to use multiple displays simultaneously, for example, multiple displays of each running of a visualization tool on the same data, but for various data variables or with various displaying time-spans. An additional important requirement is displaying a number of levels of data (i.e., raw packets, host information, and network sessions), and allowing users to zoom in or drill down on some items of data (Elhenawy et al., 2011).

Principal Component Analysis (PCA) helps visualize data, by projecting data down to two or three dimensions which users can plot to get a better understanding of the data. In addition, the principal component vectors often indicate meaningful things about the nature of the data (Hertzmann and Fleet, 2012). Visual data mining tools including features to display classes, clusters, outliers, and associations are helpful in viewing every anomalous pattern that is identified. Graphical User Interface (GUI) associated with these tools allows security analysts to understand intrusion detection results, evaluate Intrusion Detection System (IDS) performance, and decide on future improvement for the system (Youssef and Emam, 2011).

Analysing a large amount of data generated by networks traffic in real time requires more computing power and capacity. A good option is to apply High Performance Computing (HPC) techniques, especially using Graphics Processing Units (GPUs). GPUs provide more computing power than CPUs with multiple cores. It has been developed with a high memory bandwidth, highly parallel structure, and more chip surface that are dedicated to data processing than to data caching and flow control (Barrionuevo et al., 2015). GPUs can greatly accelerate a lot of tasks. Arrays of multiple GPUs help solve large data problems. High end GPUs can provide massive parallelism and enables interactive browsing interfaces. Interactivity can create a qualitative difference (Madden, 2014).

Three main visualization techniques for network traffic are listed as follows (Marode and Chavan, 2014):

- Parallel Coordinate: A backdrop is drawn consisting of *n* parallel lines, equally spaced and typically vertical to show a set of points in an *n*-dimensional space. A point in the *n*-dimensional space is displayed as a polyline with vertices on the parallel axes.
- Bundle View: Bundle visualization technique uses rings to represent network traffic. The line dividing the circle represents a network border where network packets flowing from internal network to an external network or vice versa.
- 3-Dimentional Visualization: Network traffic is shown in a three-dimensional space.

Network visualization tools translate large amount of network traffic logs (in text based data) into visual patterns (animation), allow operators to investigate and detect anomalous internal and external network traffic. Table 1 (Marode and Chavan, 2014) shows a part of tools related to network visualization. NetFlow is a network protocol that was developed by Cisco Systems to collect IP traffic information.

Tools	**Description**
FLAVIO	Flow Loader and Virtual Information Output (FLAVIO) is a data grapher of NetFlow. It collects data from a data exporting device of NetFlow, loads the data into a MySQL database, and produces charts. Platforms: Linux/UNIX
NetFlow Monitor	It processes and evaluates NetFlow Exports from CISCO routers. It can be used for capturing, monitoring, and analysis purpose that operates in an online or offline mode. Platforms: Linux, Unix
nfdump	A series of tools for collecting and processing NetFlow data. Platforms: Linux, Unix
SiLK	A set of NetFlow tools that were developed by the CERT/NetSA Team for improving the analysis of security for big networks. Platforms: Unix, Linux
Netpy	A cross platform interactive client/server network flow visualization and analysis tool using Python, wxWindows, and C. Platforms: Linux, Unix
tcpdump	A common local traffic packet analyzer that runs under the command line and display TCP/IP traffic. It uses the libpcap library to capture packets. Platforms: Linux, Window
Wireshark	A free and open-source local packet analyzer. Packet capturing is performed with pcap library and can work in an online or offline mode. Platforms: Linux, Windows

Table 1. Some tools related to network visualization

3. Applications and Technology Progress in Visualization

As for domain-specific knowledge and visualization tools, Graphical User Interfaces (GUIs) at high levels and visualization tools are required for data mining systems. Visualization of data mining can include visual data mining and visualization of data, the mining process, and the mining results. The flexibility, variety, and quality of visualization tools can greatly affect the usability, interpretability, and attractiveness of data mining systems. Visual data mining integrates data mining and data visualization to achieve useful and implicit knowledge from huge data sets (Han and Kamber, 2006). A new and robust *K*-means clustering method with a large-scale multi-view was proposed to integrate heterogeneous data representations with a large scale. The method consistently achieves superior clustering performances. It can be easily parallelized and performed on multi-core processors for big visual data clustering (Cai et al., 2013).

Human-Computer-Interaction (HCI) patterns of large data with an unstructured form are visual-oriented; therefore, they are displayed in a multimodal and heterogeneous format including graphics, flow charts, text, code snippets, and are often screen shots disparate (Gaffar et al., 2014). Meta-visualization aims at providing additional information about the data, the setup, or the current analysis session. Two kinds of meta-visualizations can help to establish orientation: 1) a history graph; 2) an explicit, abstract representation of, for instance, the data sets and their inter-dependencies (Streit, 2011).

Visualization tools have been utilized in the analysis of telecommunication data. Tools for clustering, outlier visualization, linkage visualization, association visualization, and OLAP (On-Line Analytical Processing) visualization are very helpful in analysing telecommunication data (Han and Kamber, 2006). Topology-preserving compressed graphs were proposed, which shows promising results in reducing visual complexity of large networks, and consequently provide a time-efficient solution for big graph anomaly analysis. Through on-site live demonstration and videos, the compressed graphs have many potential applications (e.g., security and network management) in different scenarios of diversified networks (Liao et al., 2012). It is easy for analysts to lose the sight of the "big picture" due to focusing on low-level details when they conduct packet-level analysis for intrusion detection. Time-based Network Traffic Visualizer (TNV), an information visualization tool, was developed to avoid the loss of context and augment of existing tools for intrusion detection (Goodall et al., 2005).

Most intrusion detection methods with visualization are anomaly-based detection methods and visualize audit data rather than alerts themselves. The host-based visualization method for intrusion detection is to learn normal states of commends or programs that are achieved by users and compare audit data with profiles for visualization. So, they detect an intrusion when an attack differs from graph characteristics with normal states, and extract diagnostic features of attacks for embodying anomaly detection. However, these methods do not visualize alerts themselves, but visualize audit data. Therefore, these are useful for detecting attacks that emit much traffic such as distributed denial-of-service attack (DDoS) or worm. These do not offer clear characteristics for attacks that emit little traffic (Elhenawy et al., 2011). Some attacks that are detectable using visualization methods are listed as follows (Marode and Chavan, 2014):

- Host scan attack (Ping): An attacker tried to send traffic to too many hosts on a network by changing the IP address of the destination.
- Port scan attack: Port scanning is a process of identifying listening ports of a target system. An attacker checks if the target system is on, then looks for open ports, and finds weak points to attack.
- Denial of Service (DoS) attack: It attempts to make a network resource or machine unavailable to its intended users through sending large amount of traffic to the specified host.
- Distributed Denial of Service (DDoS) attack: Many compromised systems (bot) attack a single target, therefore resulting in denial of service for the users of the target system.
- Backscatter: An attacker tried to do both host scan and port scan.

A system was proposed to support the analysis of the logs of an Intrusion Detection System (IDS), which pivots huge sets of NetFlows visually. Especially, two kinds of visual representation of the flow data were compared. The first one is the TreeMap visualization of local network hosts which are linked through hierarchical edge bundles with external hosts; the second one is a graph representation using a force-directed layout for visualizing the structure of host communication patterns. The proposed visual analytics system monitors large-scale network traffic. The system obtains NetFlows, stores them in a database, and helps querying with suitable index structures. Tools permit loading events of intrusion detection, combining with preset queries on NetFlows as a starting point of a search for malicious activities on the network. After suspect hosts are selected, analysts can match them to traffic on the network under surveillance. The analysis process can be assisted by the following two kinds of visualization: 1) the visualization of a geometrically clustered flow which displays the internal network as a TreeMap surrounded by external hosts; 2) the graph-based visualization emphasizing the structural properties of communication flows between internal and external hosts. NetFlow views and host details show more details in a tabular form which is augmented with graphical properties, e.g. colour (Mansmann et al., 2009). Fig. 1 (Mansmann et al., 2009) illustrates a graphical view of emphasizing structural properties of the connectivity between hosts, e.g. groups of interconnected hosts.

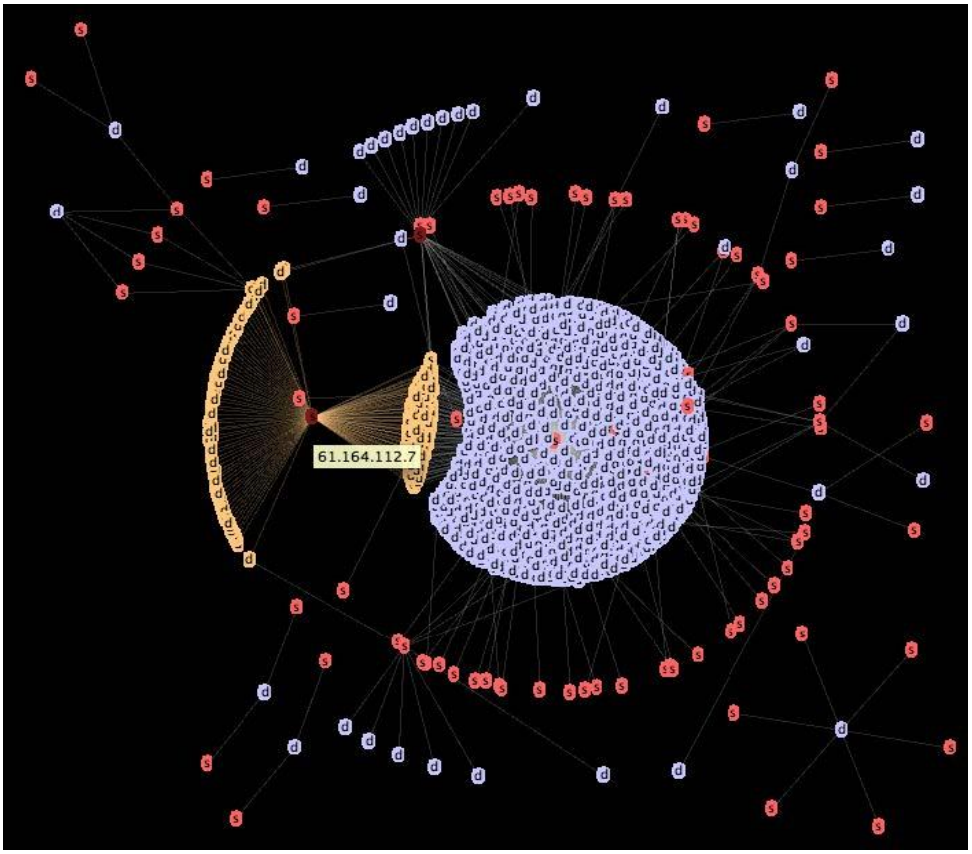

Fig. 1. Graphical visualization displaying communication flows between destination hosts (blue) and source (red)

Big Data technology helps visualize multi-dimensional data including structured data and unstructured data (Cao, 2014). The unit-circle algorithm was developed to provide unit-circle representations to both intrusion and regular traffic. This can reduce problems in visualizing big data through mapping huge numbers of data points to a unit-circle (Suthaharan, 2014). The VSOutlier system was developed to support interactive exploration of outliers in big stream data. VSOutlier not only helps detect many types of outliers for efficient and innovative outlier identification strategies, but also presents many interactive interfaces to explore outliers in real time. It has been demonstrated that the VSOutlier system is able to let analysts identify, understand, and respond to the situations of interest more efficiently in near real-time even when it is used for a large volume of streams (Cao et al., 2014).

4. Challenges in Visualization

Data usually needs to be visualized so that an analyst can see insights. When Hadoop is used, integrated visualization is difficult because it is required to build middleware to get the data out of

Hadoop and put it into the visualization layer (Datameer, 2013). Human's eyes are only good at seeing numeric data with two or three dimensions; therefore, data visualization methods are weak in identifying outliers in a dataset with high dimensionality or with many categorical attributes. Powerful visualization tools are needed to view any detected anomalous patterns. This kind of tools should have the functions of viewing outliers, clusters, and associations (Han and Kamber, 2006).

Ease of access to relevant data for visualization and performing comprehensive visualization are a challenge (Glisic et al., 2014), especially in the following situations: 1) data are registered both in a dynamic and static mode (Sigurdardottir et al., 2012); 2) various sensing methods are mixed in the same application, such as wired fiber optics (Li and Wu, 2008; Ravet et al., 2009), wired or wireless electrical sensors (Gangone et al., 2011; Lynch, 2007); and 3) heterogeneous data are captured, such as acceleration, image, temperature, and strain, etc. (Gangone et al., 2011).

Visualization has proven its value for the interactive analysis of homogeneous data. However, in order for analysing enormous amounts of heterogeneous and interrelated data sets, visualization alone is often not sufficient. The heterogeneous sets of complex data comprise major challenges confronting the field of visual analytics: different sources, various formats, different types, and various levels of scale. Due to the multidisciplinary nature, data is again available in various forms (free text, images, and statistical tables, etc.), in various representations (tabular, tag clouds, and visualizations, etc.), and at multiple levels of details. All of them need to be integrated into one seamless and interactive analysis process to fulfil efficient collaboration (Streit, 2011; Attunity, 2012).

There are two main computational challenges in large-scale data clustering (Cai et al., 2013): 1) how to integrate the features of heterogeneous data to improve the performance of data categorization? 2) how to reduce the computational cost of clustering algorithms for large-scale applications? Anomaly analysis and visualization of large graphs are still a challenge due to the non-linear increase of complexity and highly dynamic nature of large networks. Large-scale graph analysis and visualization is an important research topic in Big Data (Liao et al., 2012).

There are two major challenges regarding large-scale communications networks: 1) numerous nodes makes it very difficult to monitor them individually. In addition, the behaviour of the nodes is often dependent on each other; therefore, monitoring the nodes in an isolated manner will bypass their correlations; 2) numerous edges makes it very difficult to analyse the highly dynamic time-series of communications volume in tandem (Akoglu et al., 2015). The characteristics of big data make it a challenge to visualize big data. The available visualization methods such as dimensionality reduction and data projection can only present an abstract view of data. However, the abstract view does not present complete geometric representations of the data (Suthaharan, 2014).

5. Conclusion and Future Work

Data visualization holds promise as an approach to helping achieve data integration. The capability of providing multiple views of the same or related data is significant for achieving multiple views and levels of data. Network-based visualization method for intrusion detection shows the source address, destination address, port number, and the network's packets using visual graphs. Network intrusion detection methods based on graphs are useful approaches to the visualization of network intrusion. This kind of visualization presents the dynamically growing and changing nature of networks.

Analysing anomalies in large-scale networks is important but challenging due to the non-linear dynamics and complexity when the graph size increases. Interaction with complex and heterogeneous data is also a challenge. Large volumes of heterogeneous data bring a lot of technical challenges, which is a barrier to Big Data analytics. Big Data analytics with HPC, especially GPUs are very useful in accelerating scientific computing, network analysis, and network visualization. Future research topics can be Big Data analytics for streaming data, big data with complex structures, and big data with uncertainty.

Reference

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, *29*(3), 626-688.

Attunity (2012). Enabling big data analytics by replicating oracle data to Greenplum, database trends and applications, September 12.

Barrionuevo, M., Lopresti, M., Miranda, N., & Piccoli, M. F. (2015). Solving a big-data problem with GPU: the network traffic analysis. *Journal of Computer Science & Technology*, *15*(1), 30-39.

Cai, X., Nie, F., & Huang, H. (2013). Multi-View K-means clustering on big data. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (*IJCAI*), 2598- 2604.

Cao, L., Wang, Q., & Rundensteiner, E. A. (2014). Interactive outlier exploration in big data streams. *Proceedings of the VLDB Endowment*, *7*(13), 1621-1624.

Cao, N. (2014). *Big data analytics lecture: data visualization*. Columbia University.

Datameer, Inc. (2013). The guide to big data analytics, White Paper, 1-39.

Elhenawy, I., Riad, A. E. D., Hassan, A., & Awadallah, N. (2011). Visualization techniques for intrusion detection- a survey. *International Journal of Computer Science and Engineering Survey*, *2*(3),107-119

Gaffar, A., Darwish, E. M., & Tridane, A. (2014). Structuring heterogeneous big data for scalability and accuracy. *International Journal of Digital Information and Wireless Communications*, *4*(1), 10-23.

Gangone, M. V., Whelan, M. J., & Janoyan, K. D. (2011). Wireless monitoring of a multi-span bridge superstructure for diagnostic load testing and system identification. *Computer-Aided Civil and Infrastructure Engineering*, *26*(7), 560–579.

Glisic, B., Yarnold, M. T., Moon, F. L., & Aktan, A. E. (2014). Advanced visualization and accessibility to heterogeneous monitoring data. *Computer-Aided Civil and Infrastructure Engineering*, *29*(5), 382-398.

Goodall, J. R., Lutters, W. G., Rheingans, P., & Komlodi, A. (2005). Preserving the big picture: Visual network traffic analysis with TNV. In *IEEE Workshop on Visualization for Computer Security*, (*VizSEC 05*). 47-54.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques, second edition*. Morgan Kaufmann Publishers.

Hertzmann, A., & Fleet, D. (2012). *Machine learning and data mining lecture notes*. University of Toronto Publisher, February 6.

Keim, D. A. (2014). Exploring big data using visual analytics. *EDBT/ICDT Workshops*.

Li, S., & Wu, Z. (2008). A model-free method for damage locating and quantifying in beam-like structure based on dynamic distributed strain measurements. *Computer-Aided Civil and Infrastructure Engineering*, *23*(5), 404–413.

Liao, Q., Shi, L., & Sun, X. (2012). Anomaly analysis and visualization through compressed graphs. *IEEE LDAV Poster Session*.

Lynch, J. P. (2007). An overview of wireless structural health monitoring for civil structures. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *365*(1851), 345-372.

Madden, S. (2014). Tackling the challenges of big data: visualizing twitter. *Workshop*, Massachusetts Institute of Technology.

Mansmann, F., Fischer, F., Keim, D. A., & North, S. C. (2009, November). Visual support for analyzing network traffic and intrusion detection events using TreeMap and graph representations. In *Proceedings of the Symposium on Computer Human Interaction for the Management of Information Technology* (pp. 19-28). ACM.

Marode, M. D. K., & Chavan, R. K. (2014). Survey of network traffic visualization techniques. *International Journal of Computer Technology & Applications*, *5*(3), 876-883.

Ravet, F., Briffod, F., Glisic, B., Nikles, M., & Inaudi, D. (2009). Submillimeter crack detection with Brillouin-based fiber-optic sensors, *IEEE Sensors Journal*, *9*(11), 1391-1396.

Sigurdardottir, D. H., Afonso, J. P. S., Hubbell, D. L. K., & Glisic, B. (2012). Streicker bridge: a two-year monitoring overview, bridge maintenance, safety, management, resilience and sustainability-*Proceedings of the Sixth International Conference on Bridge Maintenance, Safety and Management, Stresa*, Italy, July 8–12, pp.790–797.

Streit, M. (2011). Guided visual analysis of heterogeneous data. Dissertation. Graz University of Technology, Graz, Austria, February.

Suthaharan, S. (2014). Big data classification: problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, *41*(4), 70-73.

Wang, L., & Alexander, C. A. (2016a). Additive manufacturing and big data. *International Journal of Mathematical, Engineering and Management Sciences*, *1*(3), 107–121.

Wang, L., & Alexander, C. A. (2016b). Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences*, *1*(2), 52–61.

Youssef, A., & Emam, A. (2011). Network intrusion detection using data mining and network behaviour analysis. *International Journal of Computer Science & Information Technology*, *3*(6), 87-98.