Investigating Network Traffic Through Compressed Graph Visualization

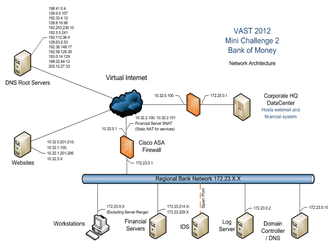
IEEE Conference on Visual Analytics Science and Technology (VAST) Challenge Workshop, Seattle, Washington, October 14, 2012

Award: Good Adaptation of Graph Analysis Techniques

Lei Shi (Institute of Software, Chinese Academy of Sciences), **Qi Liao** (Central Michigan University) and **Chunxin Yang** (Northwestern Polytechnical University)

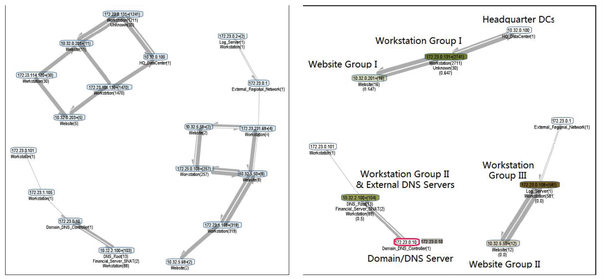
VAST 2012 MC2 Scenario

The Bank of Money regional office network has been reconfigured to support the addition of call-center activity. The network configuration is illustrated in the below figure. This Bank of Money regional office has been configured with approximately 4000 workstations and approximately 1000 servers. The office operates 24 hours a day. Some of the financial transactions are performed on financial servers inside the regional bank office's network. Other financial transactions travel to the corporate headquarters datacenter. The Bank of Money regional office uses web-based mail which is housed in the corporate headquarters datacenter.



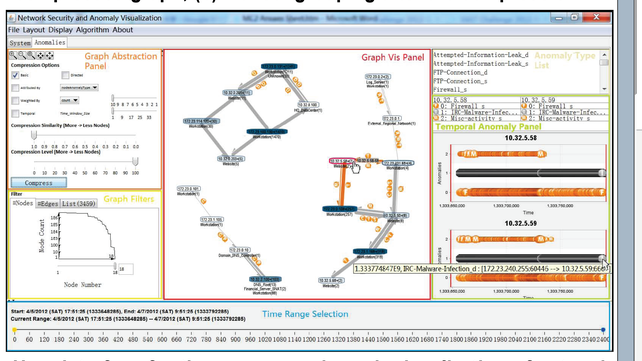
Bank of Money (BoM) Regional Headquarters Network"

We propose a novel visual analytic technique called compressed graphs for effectively analysis of anomalies in large-scale dynamic network traffic graphs. Compressed graphs can significantly reduce the size of original large graphs by a factor of 10 while reducing most node clutter and edge crossing. Unlike traditional clustering technique, compress graphs retain all the connectivity of the original graphs.



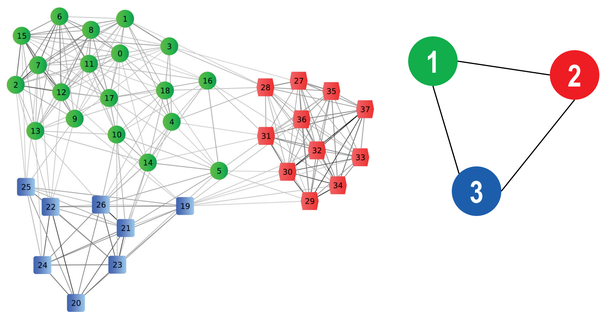
(a) (b)

40-hour BoM network traffic from the firewall and IDS logs: (a) compressed graph; (b) manual grouping after the compression.

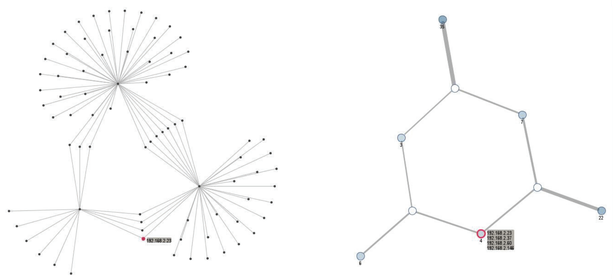


User interface for the compressed graph visualization of network security traffic. Left: controllers for the compression operation. Middle: compressed graph visualization. Right: anomaly panels showing different types of security alerts.

Compressed Graphs

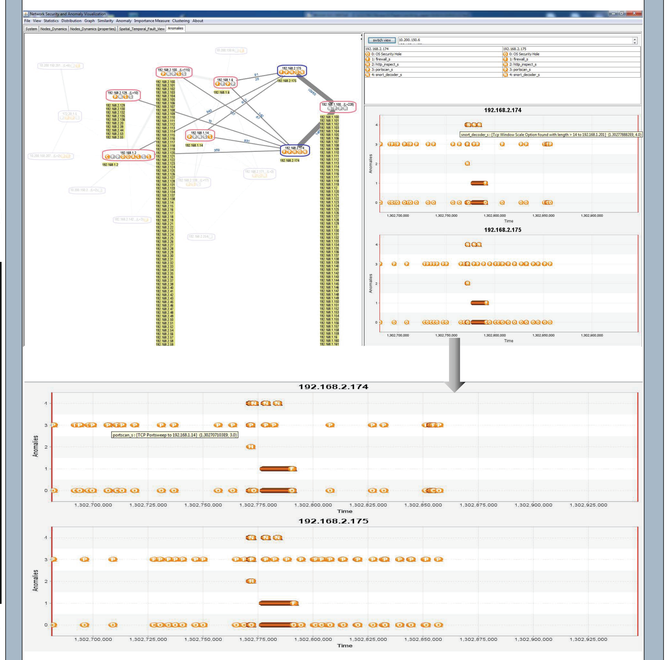


Graph clustering (or communities) may lose important topological information (e.g., edges within a community) during large graph analysis.



Original graph Compressed Graph

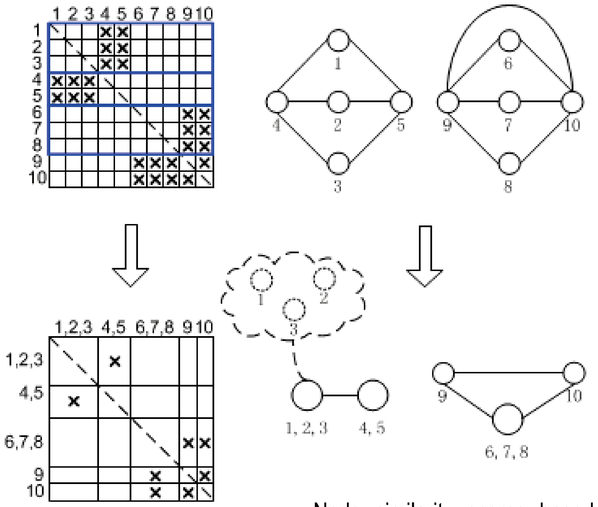
Illustration of concept of topology-preserving compressed graph on a basic subgraph.



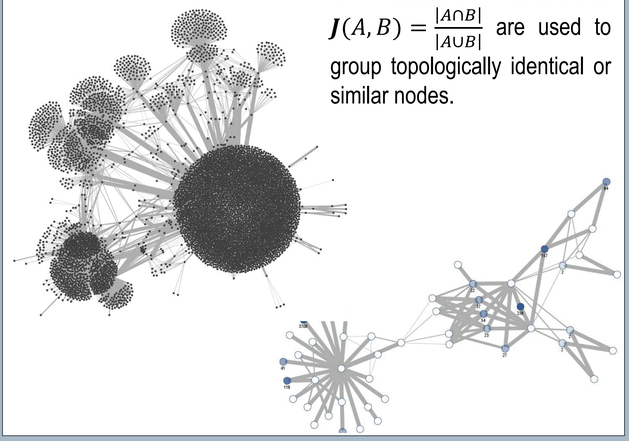
An interactive analysis and visualization: Anomaly icons aligned with timeline, each representing a different type of anomaly

Algorithms

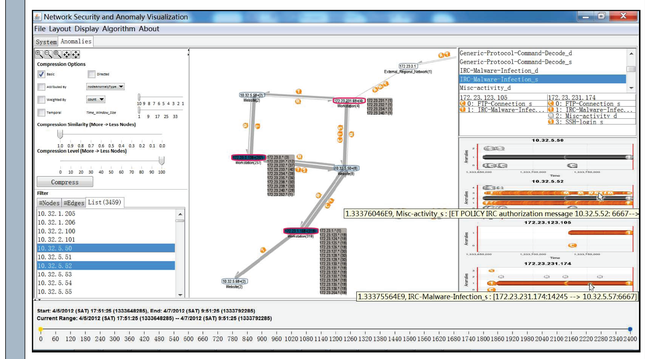
The basic idea of our approach is to aggregate nodes with similar connection patterns (e.g., same neighbor sets) in the graph together into groups and then constructs a new graph for visualization. Let W be the graph adjacency matrix where $w_{ij} > 0$ indicates a link from v_i to v_j, with w_{ij} denoting the link weight. In each row of W, $R_i = \{w_{i1},...,w_{in}\}$ denotes the row vector for node v_i, representing its connection pattern. On graph G, order its node list by the corresponding row vectors R_i (i= 1,...,n). For any collection of nodes with the same row vector (including the single outstanding node), aggregate them into a new mega-node $G_{vi} = \{v_{i1},...,v_{ik}\}$. All G_{vi} form the node set V^* for the compressed graph G^*. Also let $fv_i = v_{i1}$ denote the first sub-node in G_{vi}. The link set E^*in G^* are generated by simply replacing all $fv_i$with G_{vi}in the original link set, and removing all the links not incident to any fv_i. We also have extended our compression algorithm to support directed, weighted, and dynamic graphs by generalizing the definition of adjacency matrix and the corresponding row vectors.



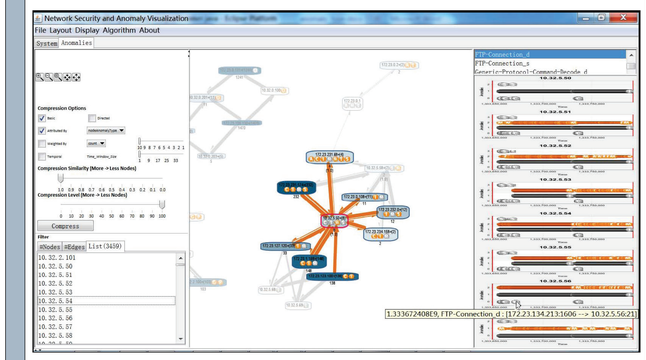
Node similarity scores based on Jaccard Coefficients $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ are used to group topologically identical or similar nodes.



Case Study -- VAST 2012 MC2



Botnet infection: Three group of workstations with heavy IRC traffic with the websites through port 6667. The 10.32.5.50-59 websites are botnet servers.



FTP and SSH connection attempts to websites 10.32.5.50-57. The related workstation machines are grouped by both neighbor set and node anomaly types.

Performance

Data	Nodes (before)	Edges (before)	Nodes (after)	Edges (after)	Rate* (Γ)	Time (compress)	Layout (before)	Layout (after)
Undirected sim=1	3460	48599	18	28	99.9%	0.437	3.151	0.078
Undirected sim=0.8	3460	48599	15	19	99.9%	0.515	3.151	0.062
Directed sim=1	3460	48599	102	1022	97.9%	0.328	3.151	0.125
Directed sim=0.8	3460	48599	57	403	99.2%	0.374	3.151	0.078

The compression rate is defined by Γ = 1-|E*| / |E|.

Video

A video demo showing interaction with the tool can be viewed from http://cps.cmich.edu/liao1q/video/VAST2012-MC2-Lei.wmv

References

[1] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. Van Wijk, J. Fekete, and D. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. Computer Graphics Forum, 30(6):1719–1749, September 2011.

[2] J. T. Bjørke, S. Nilsen, and M. Varga. Visualization of network structure by the application of hypernodes. International Journal of Approximate Reasoning, 51:275–293, 2010.