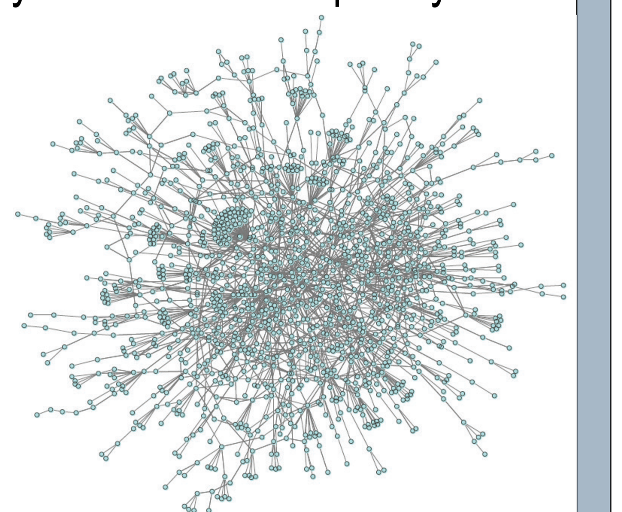
Anomaly Analysis and Visualization through Compressed Graphs

IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV), Seattle, Washington, October 14-15, 2012

Qi Liao (Central Michigan University), **Lei Shi** (Chinese Academy of Sciences) and **Xiaohua Sun** (Tongji University)

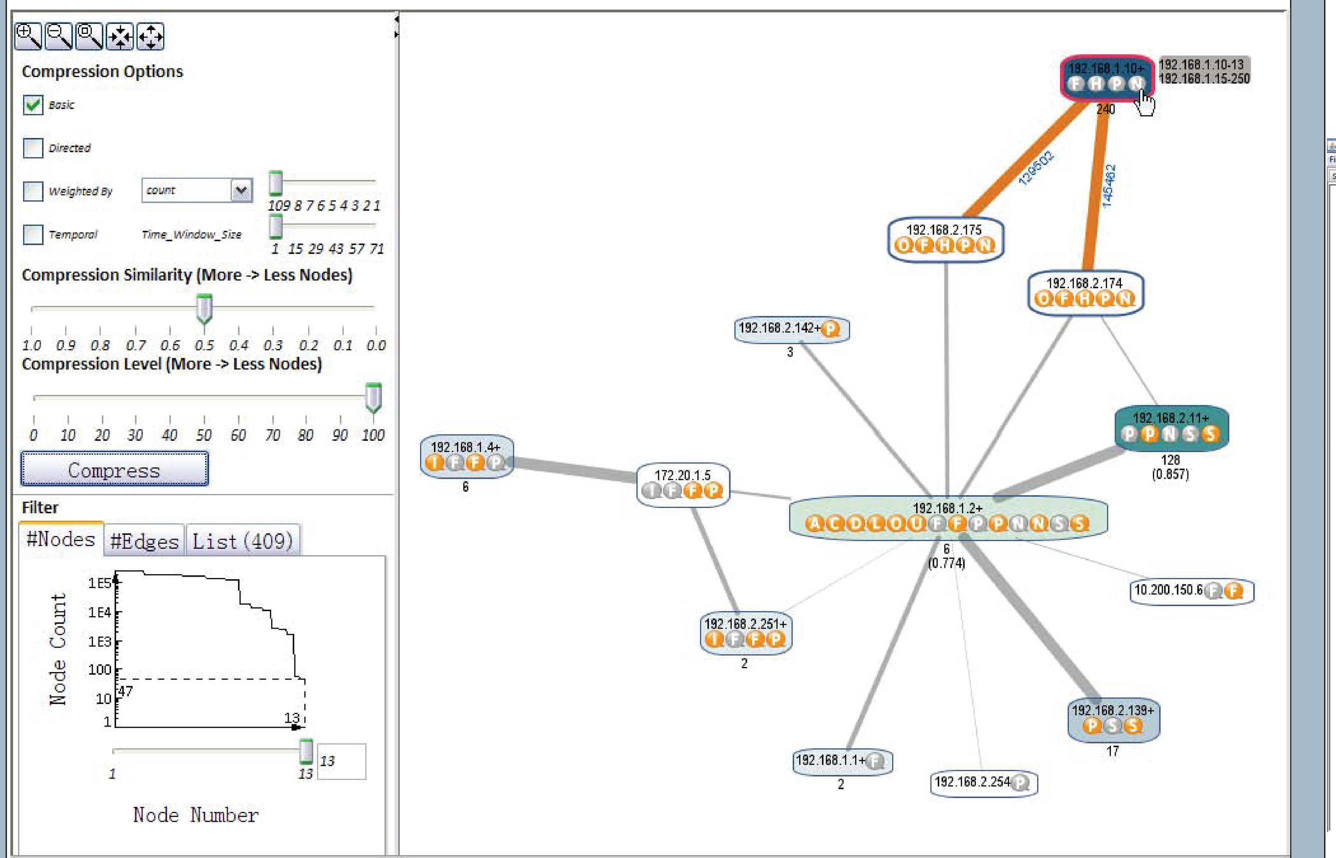
Problem

The size of today's networks has exploded. Notably, the Internet of Things (IoT) consists billions of devices ranging from smart phones, sensors, meters to appliances and vehicles. The changing behavior of people's social networking make one popular online SNS approaching one billion users. Understanding and managing such large networks is fundamentally harder because the dynamics and complexity does not increase linearly with the increase of the size. How to detect, and more importantly the root causes of, essential changes (i.e., anomalies) is urgently needed and would have substantial impact on many areas of research and applications, such as security and network management, etc.

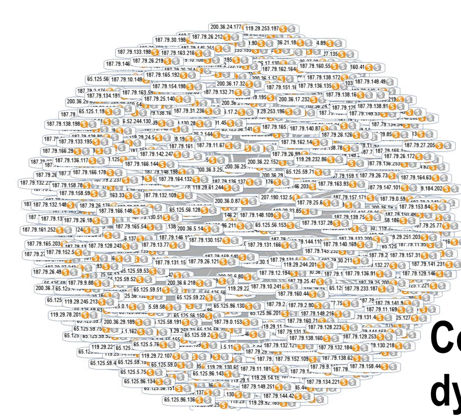


Visual analytics for "big graphs" [1][2]

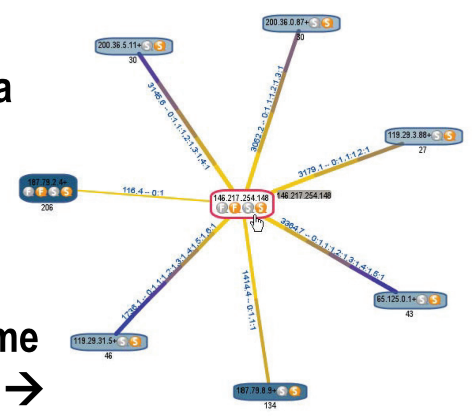
We propose a novel visual analytic technique called compressed graphs for effectively analysis of anomalies in large-scale dynamic graphs (big graphs). Compressed graphs can significantly reduce the size of original large graphs by a factor of 10 while reducing most clutter and crossing of nodes and edges. Unlike traditional clustering technique, compress graphs retain all topology structure of the original graphs.



The user interface for the compressed graph visualization: the left panel includes various controllers for the compression operation; the right panel is a compressed graph with appropriate color coding and anomaly icons.

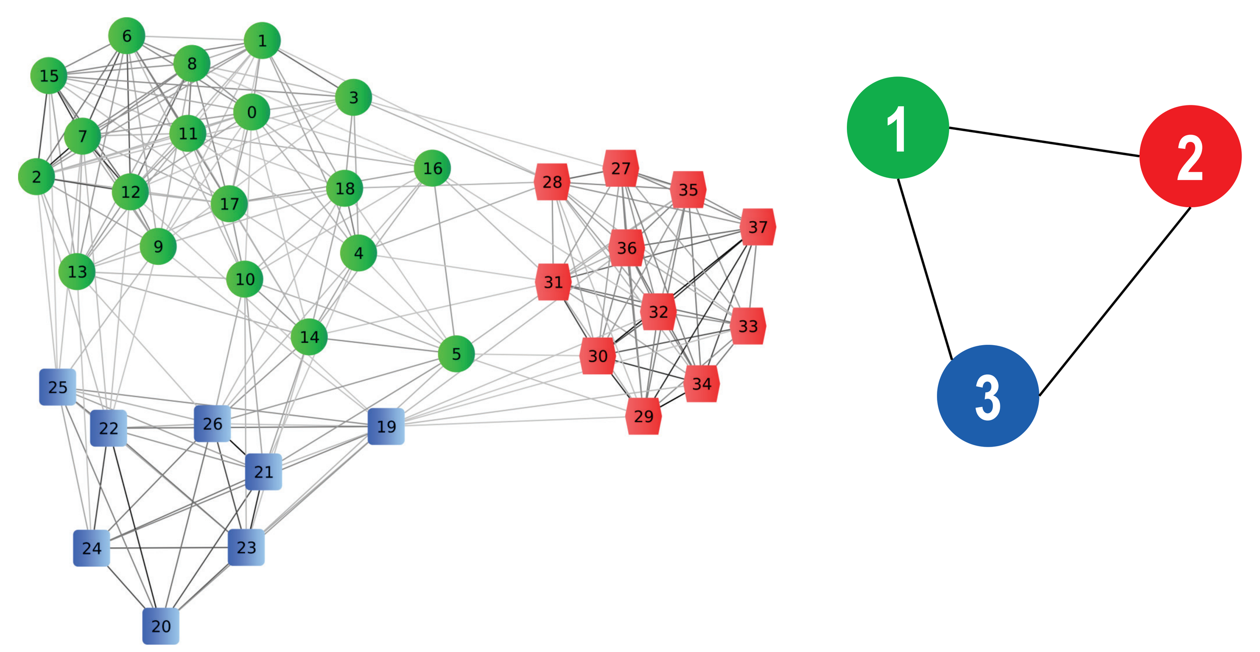


← Original honeypot data

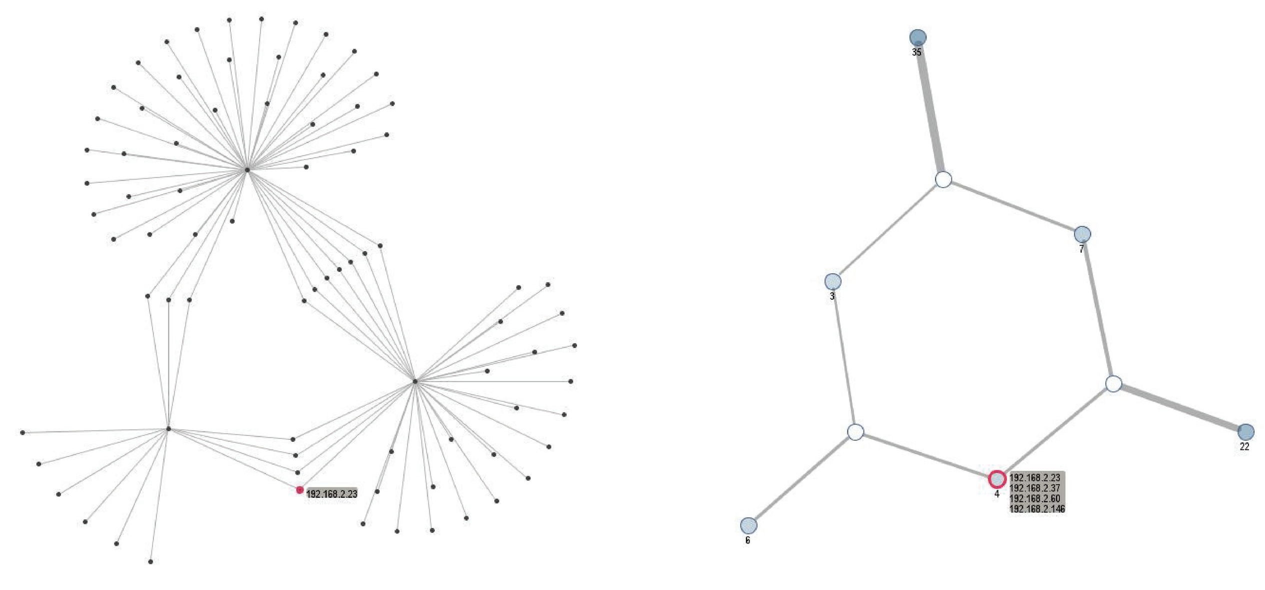


Compressed time dynamic graph →

Compressed Graphs

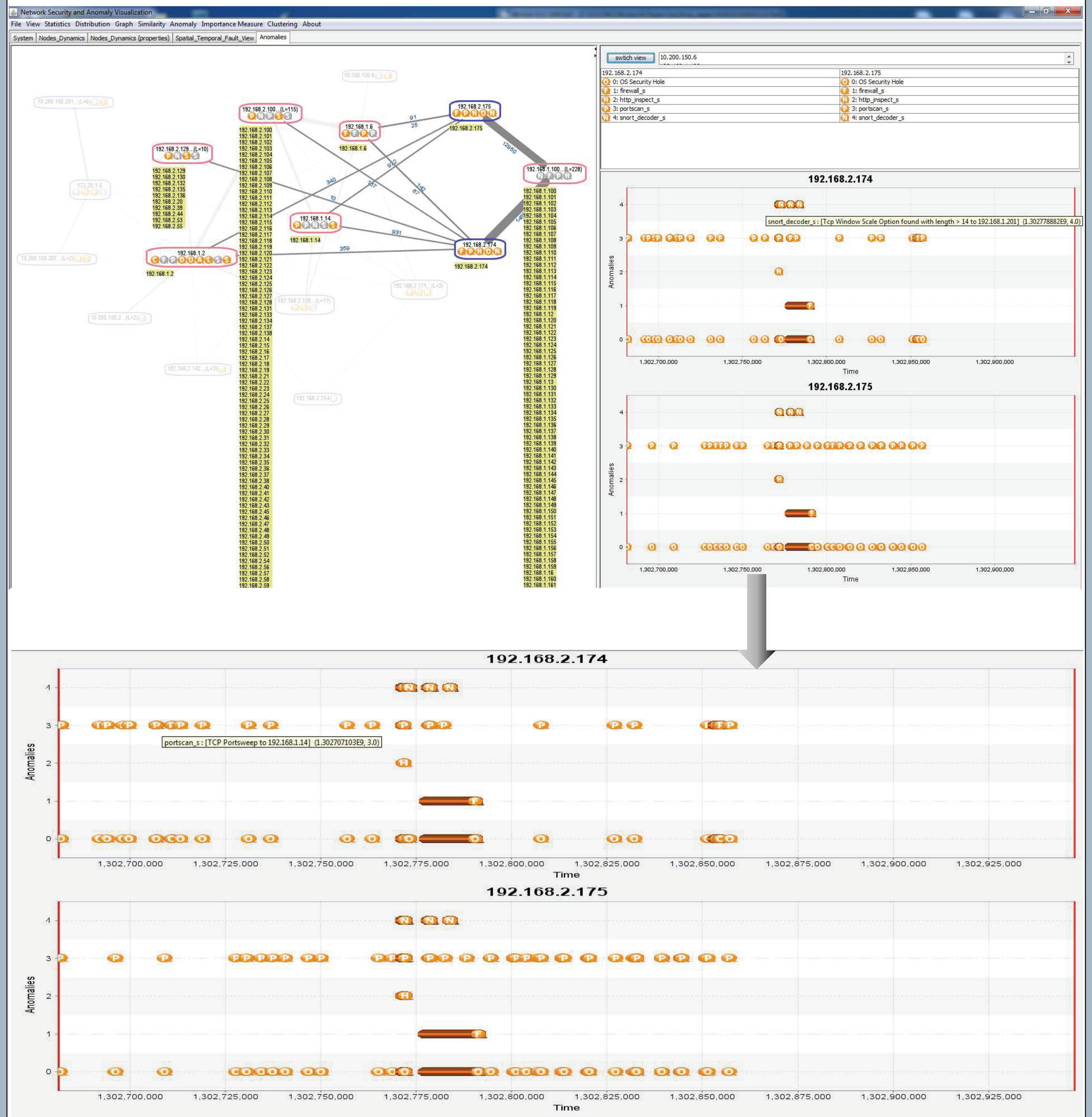


Graph clustering (or communities) may lose important topological information (e.g., edges within a community) during large graph analysis.



Original graph **Compressed Graph**

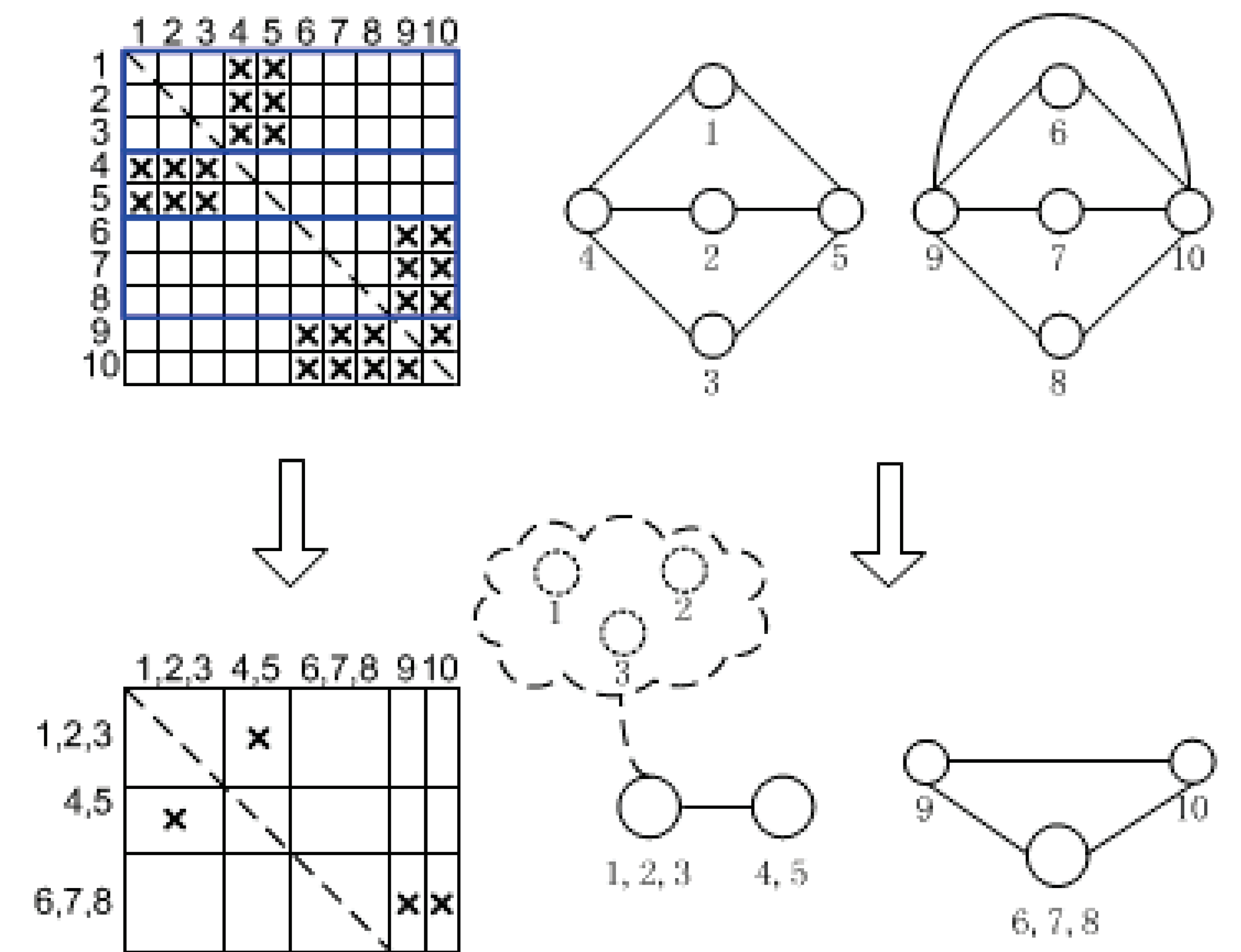
Illustration of concept of topology-preserving compressed graph on a basic subgraph.

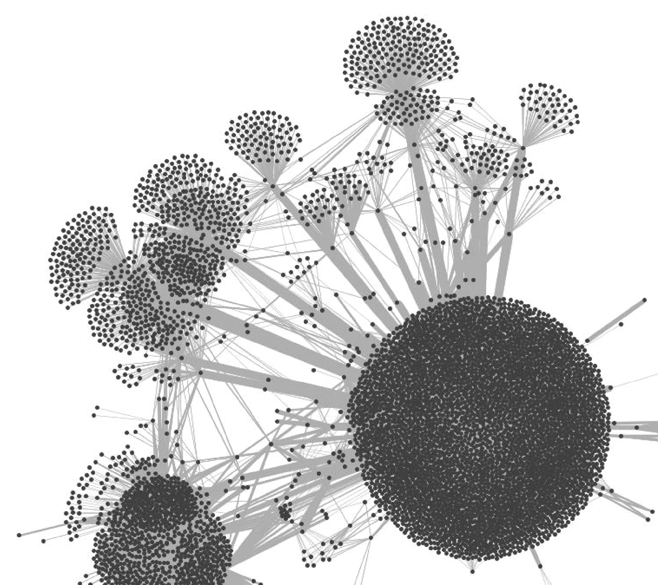


An interactive analysis and visualization: Anomaly icons aligned with timeline, each representing a different type of anomaly

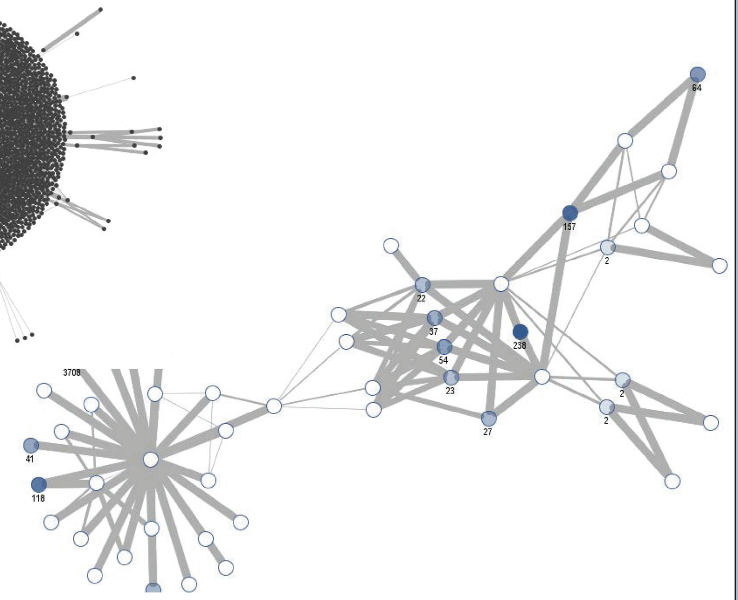
Algorithms

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, $R_i = \{w_{i1},\ldots,w_{in}\}$ denotes the row vector for node v_i, representing its connection pattern. For graph G, order its node list by the corresponding row vectors R_i ($i = 1,\ldots,n$). For any collection of nodes with the same row vector (including the single outstanding node), aggregate them into a new mega-node $G_{vi} = \{v_{i1},\ldots,v_{ik}\}$. 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.



Preliminary Results

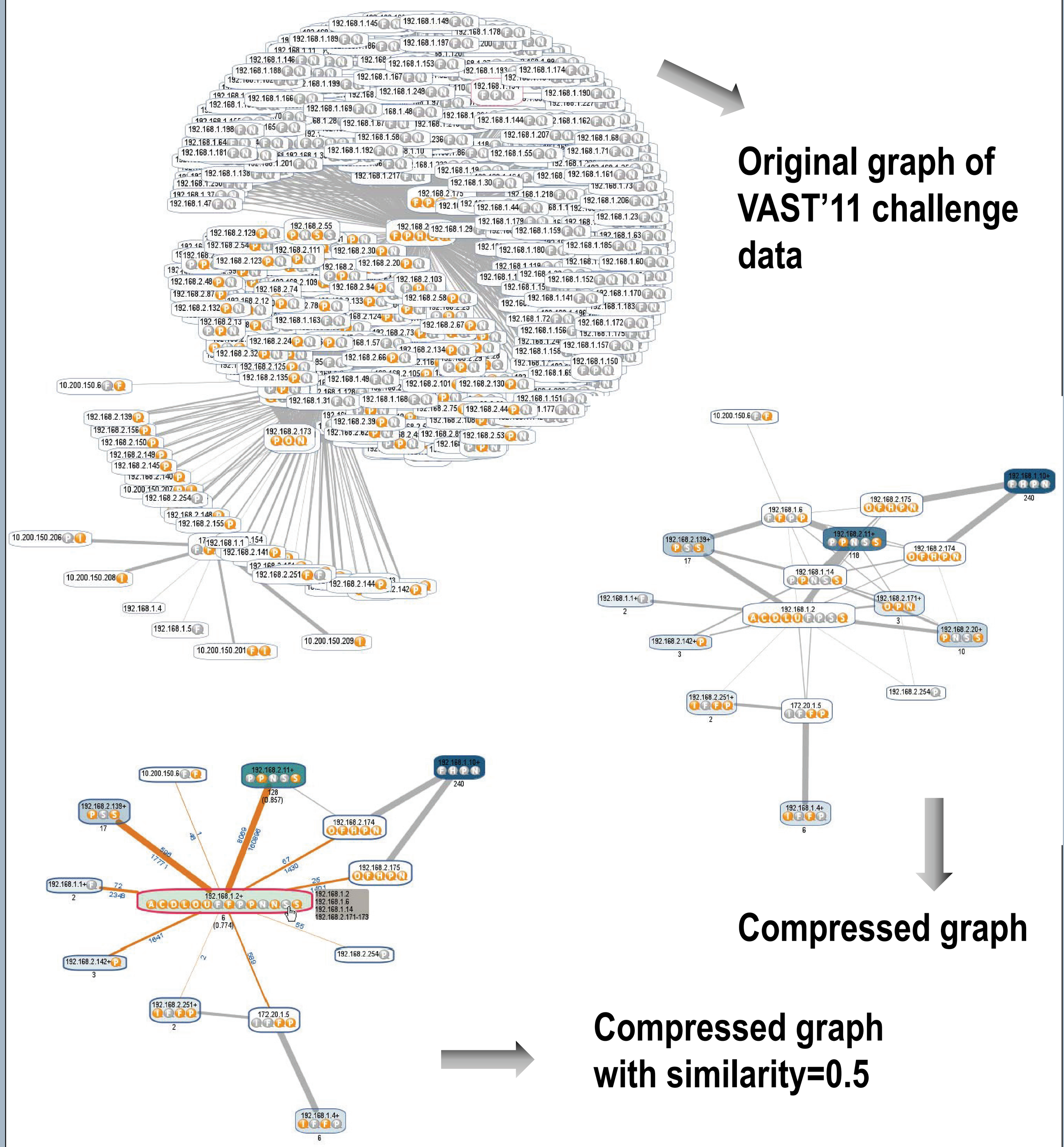
Performance evaluation on VAST11 Challenge dataset

Data	Nodes (before)	Edges (before)	Nodes (after)	Edges (after)	Rate* (Γ)	Time (compress)	Layout (before)	Layout (after)
Undirected sim=1	409	1613	17	50	96.9%	0.007	0.24	0.084
Undirected sim=0.8	409	1613	16	39	97.6%	0.012	0.242	0.088
Undirected sim=0.5	409	1613	13	23	98.6%	0.006	0.25	0.079
Directed sim=1	409	1613	26	82	94.9%	0.005	0.245	0.084

Performance evaluation on Honeypot dataset

Data	Nodes (before)	Edges (before)	Nodes (after)	Edges (after)	Rate* (Γ)	Time (compress)	Layout (before)	Layout (after)
Undirected	15380	16353	2	2	99.9%	0.123	10.179	0.079
Undirected weighted #bin=10	15380	16353	5	8	99.9%	0.151	10.179	0.692
Undirected	44668	45582	2	2	99.9%	0.401	35.09	0.08
Undirected	1051595	1158150	2	2	99.9%	4.56	(500) 36.404	0.026
Undirected dynamic #win=1	43602	47752	9	16	99.9%	1.27	33.504	0.1
Undirected dynamic weighted #win=1 #bin=10	43602	47752	105	208	99.6%	1.102	33.504	0.946

The compression rate is defined by $\Gamma = 1 - |E^*| / |E|$.



Original graph of VAST'11 challenge data

Compressed graph

Compressed graph with similarity=0.5

A video demo showing interaction with the tool can be viewed from http://cps.cmich.edu/liao1q/video/LDAVCompressedGraphs.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] P. C. Wong, P. Mackey, K. A. Cook, R. M. Rohrer, H. Foote, and M. Whiting. A multi-level middle-out cross-zooming approach for large graph analytics. In Proceedings IEEE Symposium on Visual Analytics Science and Technology (VAST), pages 147–154, Salt Lake City, UT, October 12-13 2009.