Iain J. Cruickshank
icruicks@Andrew.cmu.edu

Dr. Anthony Johnson
anthony.johnson@jhuapl.edu

Timothy Davison
timothy.davison@jhuapl.edu

Dr. Matthew Elder
matthew.elder@jhuapl.edu

Dr. Kathleen M. Carley
kathleen.carley@cs.cmu.edu

Carnegie Mellon University
www.casos.cs.cmu.edu

Detecting Malware Communities Using Socio-cultural Cognitive Mapping

Malware is a Pervasive Cyber Threat

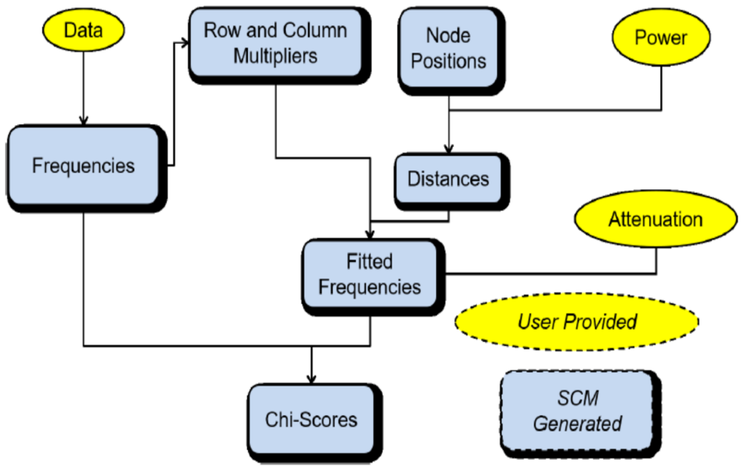
- Cyber-attacks are a critical problem for society, and malware is used by attackers in a large class of cyber-attacks
- The vast majority of malware samples are actually variants of existing malware. As such, many malware samples could be characterized as being a malware community or family with distinct relations between samples.

Sakula Malware Data

- The Sakula family of malware represents variations of remote access Trojan (RAT) tools that seek to target victims in the aerospace, government, healthcare, and technology sectors.
- We used a deep neural network-based extraction method uses four static feature domains to create a 1,204 length feature vector for each sample

Social Network Analysis (SNA) for Characterizing Malware

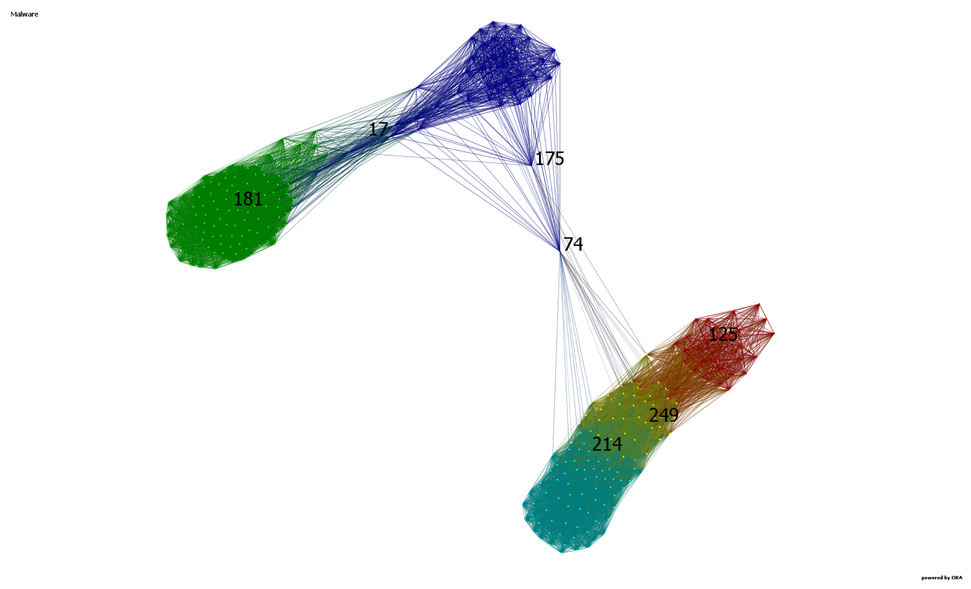
- We used a three step procedure to characterize malware communities: 1) Extract features from the malware, 2) Use Socio-cultural Cognitive Mapping to place each of the malware samples into a latent space by their features, 3) Use a Weighted Consensus Graph to extract the latent graph of the points in the latent space, and then analyze the latent network.
- Socio-cultural Cognitive Mapping (SCM): Uses shared counts between entities to map the entities into a lower-dimensional, metric space such that similarities are preserved



- Weighted Consensus Graph (WCG): for each entity add in a weighted edge (weighting by Weighted Jaccard Index of the neighborhoods of the endpoint of the edge) between each of the k-nearest neighbors (kNN) of that entity

SCM + WCG latent network and communities

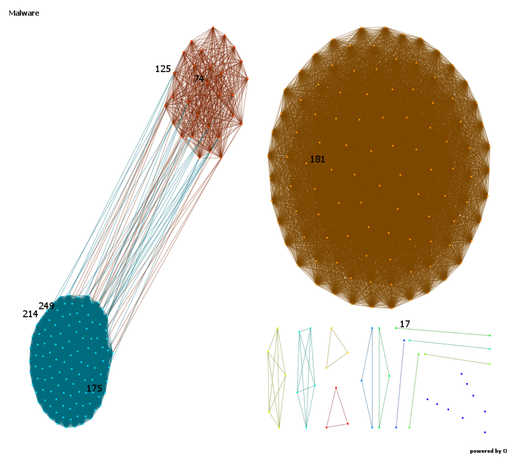
- The latent network of the malware samples displayed distinct, heterogenous sub-communities



- Those nodes that are weak members of any group, 74 and 175, were also the highest in centrality betweenness. It is likely these were anomalous samples with respect to the other malware samples
- The most central node of the green was 181, of the light blue was 214, of the dark blue was 17, of the yellow was 249, and of the red was 125. These were likely proto-typical nodes for each of their respective subgroups.

Comparison to *ssdeep* derived network

- In order to gain better insight into the learned latent network, we then compared the network produced by SCM + WCG to the data found by a commonly used similarity analysis tool for malware analysis, *ssdeep*



- The subgroups had a reasonably high degree of overlap, with an Adjusted Mutual Information value of 0.7103.
- The two latent networks also had different topologies, with the SCM + WCG having greater connectivity and heterogeneity in degree distribution.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship (DGE 1745016), Department of Defense Minerva Initiative (N00014-15-1-2797), and Office of Naval Research Multidisciplinary University Research Initiative (N00014-17-1-2675 and N00014-18-2-106). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Department of Defense, or the Office of Naval Research.

institute for SOFTWARE RESEARCH