Office of Naval Research (ONR)

Final
Research Performance Progress Report (RPPR)

Title: Software Forensic Archaeology for Cyber Attribution

Principal Investigator: Matthew Elder, Ph.D.
Institution: Johns Hopkins University / Applied Physics Laboratory (JHU/APL)

Contract Number: N000141812111

Distribution Statement

- DISTRIBUTION A. Approved for public release: distribution unlimited.

Accomplishments

What were the major goals and objectives of the project?

The major objective of this project was to enable attribution of cyber-attacks to state or non-state actors using malware artifacts, derived in part from the software development life cycle (SDLC), and analyzed using social network analysis (SNA) and relational algebra (RA) techniques. To achieve this primary objective, the project sought to achieve these supporting objectives:

- *Compare the malware SDLC with traditional/goodware SDLCs to develop predictive models using data sets that include input and output artifacts*
- *Determine the human-identifying artifacts in malware that survive the inherently reductive software construction process and contribute to attribution*
- *Discover social relationships between malware samples by using SNA, RA, and high-dimensional dynamic network analysis methodologies on malware artifacts*
- *Determine how software diversification and obfuscation techniques can hamper and help with malware attribution*

Successful attribution of cyber-attacks—especially attribution with respect to nation-state actors—would likely enable more options to achieve operational goals, such as better responses to these attacks for the United States Navy, broader Department of Defense, and civilian U.S. Government agencies. It is important to tease out the multiple actors involved in the production and deployment of malware in order to offer a range of targeting options to military, law enforcement, and other government officials. Cyber-attack attribution provides operational benefits for military cyber defense in that understanding the source of the attack enables better prioritization of security event handling from the flood of alerts and events currently seen by defenders. From a policy perspective, nation-state-level attribution would enable better decision making with respect to policy and planning, including the full range of engagement options with those nation-state adversaries, from deterrence to active response.

What was accomplished towards achieving these goals?

During the project we made progress and produced results across multiple lines of research efforts towards achieving the various project goals.

For our research on SDLC modeling and experimentation, we modeled various forms of SDLCs and implemented extraction of SDLC features from open source software repositories in GitHub. Example SDLCs that we modeled using UML (Unified Modeling Language) included the "hobby project" (a small, often single-developer software development effort), an "iterative" model (e.g., a "waterfall" model with cycles), and an "agile" software development model. Building on

this modeling work, we conducted experimentation with mining software repositories, extracting features for commits, releases, contributors, various file types, development time periods, and other key SDLC characteristics. SDLC modeling and research using open source software repositories, where ground truth authorship is known, enabled controlled experimentation on the artifacts that contribute to attribution of software developers.

We developed a capability to mine open source software repositories from GitHub, the most popular online software repository hosting site. This capability enabled us to identify 1,835 candidate Windows C or C++ repositories to build for experimentation and model validation. Windows is historically—and still currently—a prevalent platform for malware, and malware is often written in C or C++. Thus, these candidates enabled analysis of the SDLC on benign repositories with known ground truth that share important characteristics with the important form of malware that we investigated. Many of the activities, artifacts, and actors associated with traditional/benign software development are included in malware software development, with additions specific to malware, as depicted in Figure 1 below. We mined important SDLC artifacts from these GitHub repositories, including tags, releases, build information, and actor data. The tags and releases are important data points to understand the SDLC process over time. Build information enables us to trace artifacts from source to binary, which is ultimately the form in which malware – and software in general – arrives and resides on end host machines. The actor data is important to trace authorship attribution from source artifacts to features that can be extracted from the software binaries. Actor data also enabled us to characterize the important contributors to a software project, including code authors and committers.

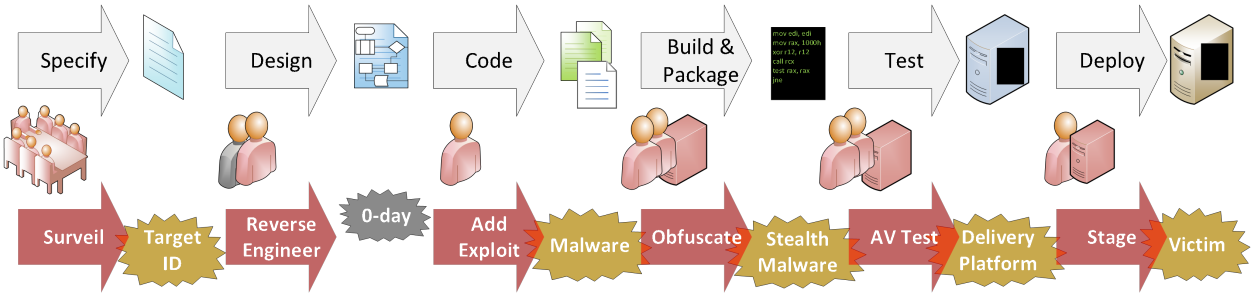

Figure 1. Software Development Life Cycle (SDLC) and Malware SDLC.

We continued research and development of a capability to mine and build open-source software repositories from GitHub for experimentation and model validation. We enhanced our GitHub mining capability with two components that enable further experimentation, Repo Extractor and Repo Builder, as depicted in Figure 2 below.

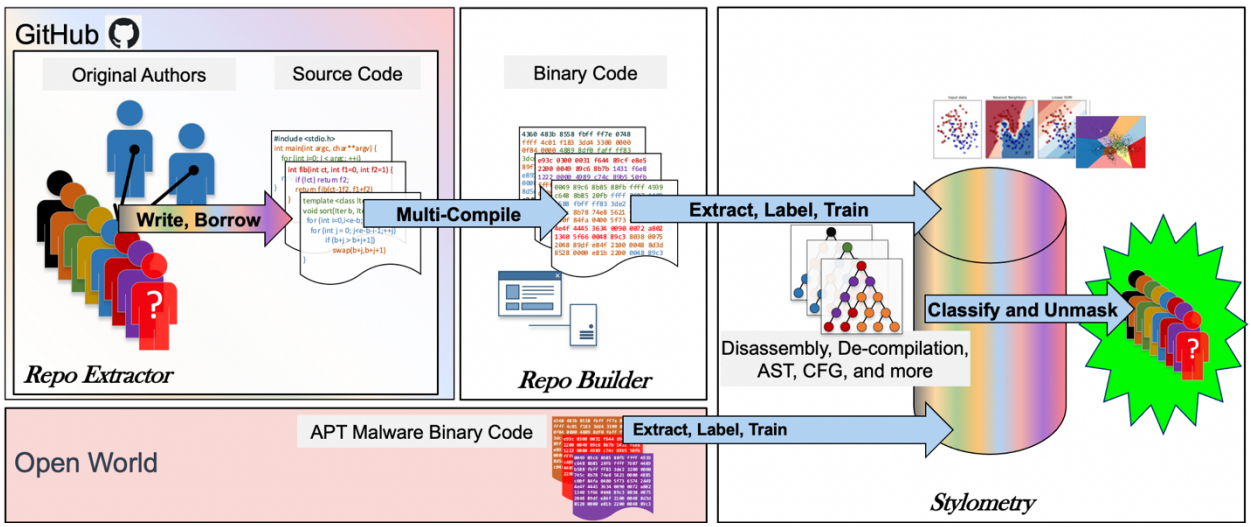

Figure 2. GitHub Experimentation using Repo Extractor, Repo Builder, and the Stylometry Pipeline.

Repo Extractor associates source code lines in a GitHub project with the author responsible for authorship at a particular point in time (e.g., a commit associated with a release tag). Repo Builder identifies GitHub repositories with build environment artifacts, such as makefiles or Microsoft compiler configuration files, to assess the "buildability" of the software project. Of the 1,835 Windows C/C++ repositories previously identified for experimentation, 465 had a Microsoft build file indicating possible buildability along with release tags to identify specific points in time within the repository history to be built. The outputs of Repo Extractor and Repo Builder are being fed to the stylometry pipeline – a framework for classifying authorship attribution using programmer style, described in previous reports and publications and previously applied to "open world" APT malware binary code – thus enabling end-to-end attribution experimentation on real-world software. This experimentation framework using GitHub repositories was presented as a working paper at the Applied Computer Security Association Conference (ACSAC) in December 2020. We have also implemented multi-author source code attribution labeling as part of the Repo Builder, which associates the disambiguated actor identity with each line of source code and associated compiler-generated binary code. The stylometry pipeline uses these labels to isolate contributions to each disambiguated actor. This capability has been documented in the provisional application for patent, listed in the Technology Transfer section under the provisional patent filing "Binary Attribution Labeling for Multi-Author Software (BALMS)," Docket No. 6069-2111. In the process of developing and experimenting with this end-to-end framework, we learned much about the intricacies of building real-world software from open-source repositories, as well as the technical challenges in associating build artifacts from program binaries back to source code lines in the original repository and further back to GitHub actor. For future work we seek to continue to conduct experimentation mapping binary stylometry features back to source code using both GitHub and Google Code Jam programming data.

We also extracted and analyzed actor data from the GitHub repositories of interest. The actor data are important to trace authorship attribution from source artifacts to features that can be extracted from the software binaries. Actor data also enabled us to characterize the important contributors to a software project, including code authors and committers. We analyzed the actor information associated with the repositories of interest and conducted social network analysis on the network of repositories and actors. We identified the impact of different actor identity disambiguation and aggregation strategies on the author and committer social networks, given the different/multiple names and e-mail addresses used by contributors to GitHub. This investigation was presented at the Sunbelt social network analysis virtual conference in July 2020.

In the process of mining GitHub for candidate software to build for experimentation, we discovered a surprising number of malware binaries in the candidate repositories. While GitHub is known to host repositories with malware for research and educational purposes, the vast majority of our candidate repositories had nothing to do with malware according to their topic tags. We developed an extension capability to investigate these suspicious/malicious binaries, submitting these files to VirusTotal for analysis. VirusTotal is a malware information service that aggregates the detection results of over 70 anti-virus (AV) products. VirusTotal is a popular resource for malware analysis researchers to assess whether a suspicious file is known malware, given the composite answer provided by the detection capabilities of the various AV engines. We processed binaries using VirusTotal with the process depicted in Figure 3 below. Some files had been submitted to VirusTotal previously at some point in the past – we recorded the prior scan results and requested reanalysis for the latest scan results. Some files were unknown by VirusTotal and we uploaded those files for analysis. In the 1,835 candidate GitHub repositories we processed, we extracted 24,395 unique binary files from the repository heads and histories. Of those 24,395 binary files, 4,335 could be considered suspicious, with at least one malicious detection in VirusTotal. Of the 1,835 repositories studied, 593 contained binaries and 314 had at least one binary with at least one recent malicious detection. We reported these suspicious/malicious binaries to GitHub, out of an abundance of due care in exercising responsible disclosure. We published and presented this research at the International Conference on Information Systems Security and Privacy (ICISSP) in February 2021. We continued exploration of malware data on GitHub and the implications from a software supply chain perspective, and we presented this follow-on work at the Malware Technical Exchange Meeting in July 2021.

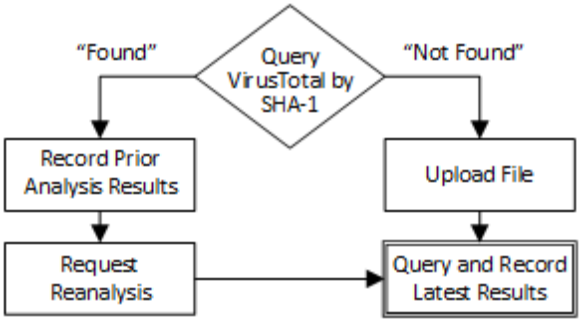

Figure 3. VirusTotal Query Flowchart for GitHub Binaries.

For our research on malware artifacts, we implemented a binary stylometry approach to identify code features for attribution, and we conducted experimentation with various malware families attributed to nation-state and/or advanced persistent threat (APT) groups, as well as commodity malware from non-state threat actors. We analyzed the various malware samples using the tools of the binary stylometry pipeline and performed attribution using the associated machine learning classification approach. The results were mixed. For samples that could be processed by the tools of the pipeline, the classification results attributing to APT malware threat actors were good (over 90%). These positive results mirror the results of applying binary stylometry to benign software in controlled experiments, i.e., projects with a single known author, with multiple implementations of the same project by different authors. However, a significant number of malware samples could not be processed by at least one tool in the binary stylometry toolchain due to obfuscation techniques employed by the malware, e.g., executable packing. These results were presented at the 2020 Malware Technical Exchange Meeting.

For our research on social network analysis of malware, we applied SNA techniques to malware family data for cluster detection and comparison of similarity analysis tool results. This initial research demonstrated that pairwise comparison results from similarity analysis tools could be turned into an adjacency matrix and explored using social network analysis tools. It was presented at the North American Social Network (NASN) conference. Follow-on work explored the application of socio-cultural cognitive mapping (SCM) and weighted consensus graphs (WCG) to syntactic and semantic features of malware to investigate network attributes. We conducted experimentation on one advanced persistent threat (APT) malware family, Sakula, and compared SCM-WCG results to clusters produced by a widely used similarity analysis tool, ssdeep – cluster membership was similar but network structure was different, as shown in Figures 4 and 5 below. This research was presented at the Sunbelt SNA conference and follow-on research was published at the Social Computing, Behavioral-Cultural Modeling & Prediction and Behavioral Representation in Modeling and Simulation (SBP-BRiMS) conference. That paper won the best late-breaking paper award at SBP-BRiMS, and an extended version of the paper was published in the Computational and Mathematical Organization Theory (CMOT) journal.

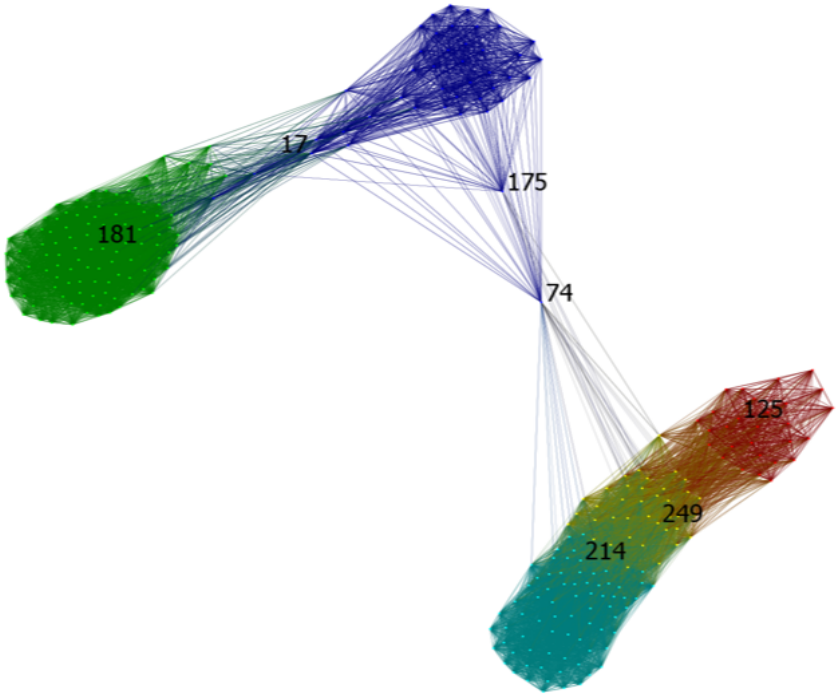

Figure 4. Socio-cultural Cognitive Mapping (SCM) results on Sakula malware family.

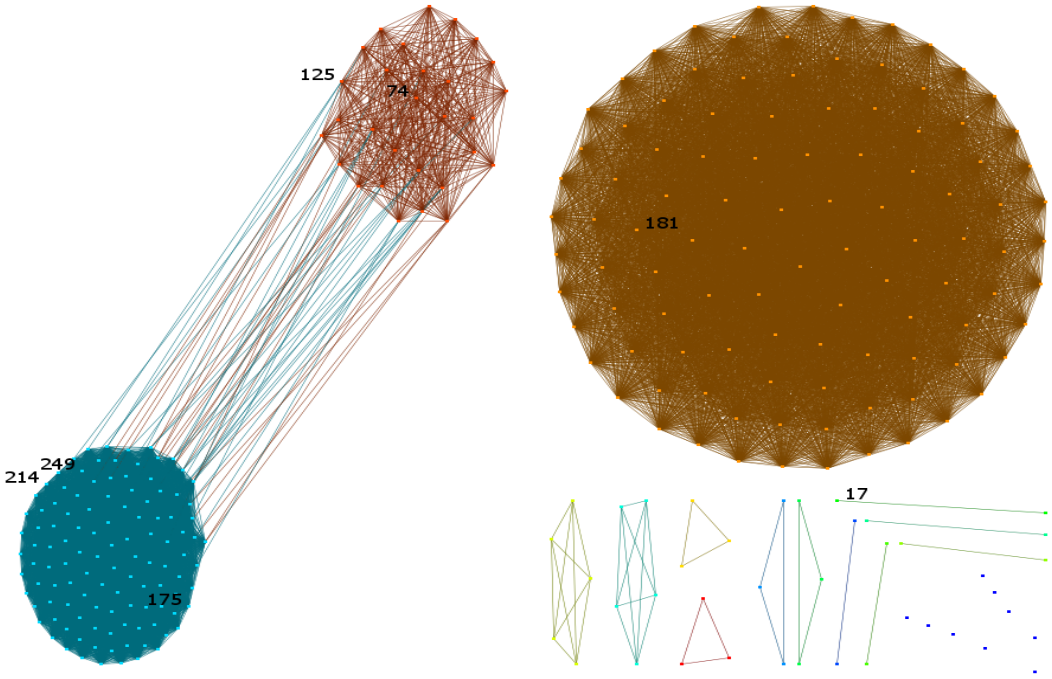

Figure 5. ssdeep results on Sakula malware family.

We continued experimentation and exploration of social network analysis techniques on malware and stylometry data for attribution. In collaboration with CMU, we analyzed the binary stylometry features from the malware attributed to four Advanced Persistent Threat (APT) groups, which we had explored in previous publications. The network analysis multiclustering approaches were able to differentiate the malware from certain APT groups in an unsupervised manner, but malware from multiple groups were combined into the same cluster in some analyses. We hypothesize that in some cases, the network analysis techniques are clustering based on elements of functionality that have survived in the stylometry features. This work was accepted and presented at the 3rd North American Social Networks (NASN) conference in January 2021. We are continuing this exploration using the original Google Code Jam programming data, in which the functionality of each programming submission is known, and thus ground truth data can be used to determine whether clusters correspond to programmer, program type, or something else. We also extended the previous clustering analysis of malware feature data to utilize Cross-View Influence Clustering (CVIC). CVIC iterates through two main steps: 1) update the cluster assignments for each view by diffusion in a social influence model with the cluster association matrix and view graphs, and 2) combine the updated cluster assignments from each view into a new cluster association matrix. Once the iterations are complete, the result is a new cluster association matrix that relates each entity to each of the clusters. This matrix naturally forms a bipartite graph of all of the entities by the cluster labels, which can be clustered by any bipartite graph clustering technique. The clustering goodness test presented interesting results, showing that the hybrid integration method, CVIC, performs the better in terms of finding clusters that agree with the original malware family labels.

Lastly, for our research on software diversity, we explored the effectiveness of various similarity analysis tools (ssdeep, sdhash, tlsh, LZJD, impfuzzy) on detecting diversity within malware families. We also conducted research measuring the impact of development variables on how different binaries appear with respect to binary analysis, an important consideration in malware attribution. This work was published at the 2018 IEEE MILCOM conference.

What opportunities for training and professional development did the project provide?

This project has supported a number of training and professional development activities:

- *Three United States Naval Academy undergraduate students worked on malware similarity analysis experimentation as part of their internships at Johns Hopkins University Applied Physics Laboratory (JHU/APL) during the summer of 2018.*
- *Two high students worked on malware similarity analysis experimentation for their ASPIRE internships (a JHU/APL STEM program) during the 2018-2019 academic school year.*
- *A team of undergraduate and graduate student researchers from the University of Maryland Baltimore County (UMBC) worked on malware similarity analysis for their INSuRE project (an NSF cyber security research and education program) during the fall of 2018.*

- *A team of undergraduate and graduate student researchers from the University of Arizona worked on mining software repositories for their INSuRE project during the spring of 2019.*
- *A team of undergraduate and graduate student researchers from the University of Texas Dallas (UTD) worked on malware attribution for their INSuRE project during the spring of 2019.*
- *A team of undergraduate and graduate student researchers from the University of Houston worked on malware attribution for their INSuRE project during the spring of 2019.*
- *A team of undergraduate and graduate student researchers from Johns Hopkins University worked on malware analysis and classification for their Information Security Institute capstone project during the fall of 2019.*
- *A team of undergraduate and graduate student researchers from Dakota State University worked on malware attribution for their INSuRE project (an NSF cyber security research and education program) during the fall of 2019.*
- *A team of undergraduate and graduate student researchers from the University of Alabama Birmingham worked on malware analysis for their INSuRE during the spring of 2020.*
- *A team of undergraduate and graduate student researchers from the University of Alabama Birmingham worked on mining software repositories for vulnerable software for their INSuRE project during the spring of 2020.*
- *A team of undergraduate and graduate student researchers from the University of Texas Dallas worked on malware analysis for their NSF INSuRE project during the Fall of 2020.*
- *A team of undergraduate and graduate student researchers from the University of Maryland Baltimore County worked on mining software repositories for vulnerable software for their NSF INSuRE project during the Fall of 2020.*
- *A high school student conducted an ASPIRE internship (a JHU/APL STEM internship program) on malware analysis during the 2020-2021 school year.*

How were the results disseminated to communities of interest?

Over the course of the project, results were disseminated via publication and presentation within the social network analysis, computer security, malware analysis, and military/defense communities. The following are our publications and presentations to those communities:

- *Ian McCulloh, Janis Butkevics, and Matthew Elder. "Russian Toaster: Social networks of helpful software developers", Sunbelt Conference, June 2018.*
- *Emily Brown, William La Cholter, Brian Ahr, and Matthew Elder. "Measuring Impact of Construction Variables on Diversified Software Binaries", IEEE MILCOM, October 2018.*
- *Matthew Elder, Marc Johnson, Ian McCulloh, and Tony Johnson. "Network Analysis of Malware Similarity Data", North American Social Networks (NASN), November 2018.*
- *Tony Johnson, Matthew Elder, and Kathleen Carley. "A Network Approach to Malware Variant Similarity Analysis", Sunbelt Conference, June 2019.*

- *Timothy Davison, Matthew Elder, and William La Cholter. "Malware Attribution Using Binary Stylometry", Malware Technical Exchange Meeting (MTEM), July 2019.*
- *Matthew Elder and William La Cholter. "Malware and Software Diversity", Malware Technical Exchange Meeting (MTEM), July 2019.*
- *Iain Cruickshank, Tony Johnson, Timothy Davison, Matthew Elder, and Kathleen Carley. "Detecting Malware Communities Using Socio-Cognitive Mapping", SBP-BRIMS, July 2019.*
- *Matthew Elder and William La Cholter. "Malware Authors Are Just Writing Software: What Can the Software Development Life Cycle and Social Network Analysis Teach Us About Malware Attribution?", USENIX Hot Topics in Security (HotSec), August 2019.*
- *Matthew Elder, William La Cholter, Tony Johnson, and Kathleen Carley. "Software Forensic Archaeology for Cyber Attribution", USENIX Security Poster Session, August 2019.*
- *Iain Cruickshank, Tony Johnson, Timothy Davison, Matthew Elder, and Kathleen Carley. "Detecting Malware Communities Using Socio-Cognitive Mapping", Computational and Mathematical Organization Theory (CMOT), January 2020.*
- *Iain Cruickshank and Kathleen M. Carley, 2020, "Analysis of Malware Communities Using Multi-Modal Features," IEEE Access 8(1):77435-77448, April 2020.*
- *William La Cholter, Matthew Elder, Tony Johnson, and Kathleen Carley. "Social Network Analysis of Windows C/C++ Contributors on GitHub", Sunbelt Conference, July 2020.*
- *William La Cholter, Matthew Elder, Timothy Davison, and Antonius Stalick. "Software Forensic Archaeology for Cyber Attribution", Applied Computer Security Association Conference (ACSAC) work-in-progress paper, December 2020.*
- *Peter Hamilton, William La Cholter, Matthew Elder, and Timothy Davison. "Authorship Attribution of Software Using Code Stylometry", Decoded (the JHU/APL reverse engineering magazine), January 2021.*
- *Matthew Elder, Kathleen Carley, Tim Davison, Iain Cruickshank, and Tony Johnson. "Network Analysis of Malware Stylometry Features for Attribution", North American Social Networks (NASN), January 2021.*
- *William La Cholter, Matthew Elder, and Antonius Stalick. "Windows Malware Binaries in C/C++ GitHub Repositories: Prevalence and Lessons Learned", International Conference on Information Systems Security and Privacy (ICISSP), February 2021.*
- *William La Cholter, Matthew Elder, Timothy Davison, and Antonius Stalick. "Malware on GitHub: Supply Chain Threats and Research Opportunities", Malware Technical Exchange Meeting, July 2021.*

Dr. Elder also disseminated results via an invited keynote entitled "Malware Attribution: Research and Perspectives" at the Malware Technical Exchange Meeting (MTEM) in July 2019, a lightning talk at the Malware Technical Exchange Meeting in July 2018, and a poster at the Johns Hopkins University Workshop for Assured Autonomy in October 2018.

JHU/APL also hosted a Software Diversity Technical Exchange Meeting in October 2018 with approximately 60 participants from government, academia, and industry, in which Dr. Elder

presented a talk and led a discussion on software diversity experimentation, and Mr. La Cholter presented a talk and led a discussion on software diversity and cyber operations.

Dr. Carley disseminated results via the following invited keynote or plenary talks:

- *"Network Analysis and AI for Humanitarian Assistance and Disaster Response," AI and HADR workshop, AI Visionary Talk, Carnegie Mellon University, September 2018.*
- *"Social Cyber-Security," Keynote MOBICOM, Delhi, India, October 2018.*
- *"Social Cyber-Security," Keynote IEEE Intelligence and Security Informatics, Florida International University, November 2018.*
- *"Social Cyber-Security Dynamics," Keynote Computational Social Science – Quo Vadis? An Interdisciplinary Symposium Honoring Kathleen M. Carley, University of Zurich, April 2019.*
- *"Information Maneuver Assessment," Keynote StratComAPAC2019, Singapore, April 2019.*
- *"Social Cyber-Security Dynamics," Keynote SocialSens, Montreal, Canada, April 2019.*
- *"Social Cybersecurity," Plenary, DHS Conference on Disinformation, July 2019.*
- *"Artificial Intelligence and Social Cybersecurity," Keynote - AAAI Symposium on AI in Government and Public Sector, Washington DC, November 2019.*
- *"Advances and Opportunities in Social-cybersecurity," Plenary, DHS Foreign Interference, Washington DC, December 2019.*
- *"Analysis of Social Networks for Intelligence," Army War College Workshop, Carnegie Mellon University, Pittsburgh, PA, February 2020.*
- *"The Power of High Dimensional Networks," North American Social Networks (NASN), January 2021.*

Dr. Carley also disseminated results via the following invited talks:

- *"Social Cybersecurity," Fed Cybersecurity R&D IWG, Washington D.C.-based Teleconference, June 2019.*
- *"Social Cyber-Security Dynamics," SMA speaker, Washington D.C.-based Teleconference, June 2019.*
- *"Social Cybersecurity: Assessing Information Maneuvers in Cyber-Space," MITRE Teleconference, October 2019.*

Furthermore, Dr. Carley delivered a number of other related invited talks, keynotes, and plenary presentations applying the same multi-view clustering methods used on malware data to social media data, describing these results related to social cybersecurity, disinformation, and social influence campaigns.

What do you plan to do during the next reporting period to accomplish the goals and objectives?

N/A – – the project was completed on November 30, 2021.

What honors or awards were received under this project in this reporting period?

2019, SBP-BRiMS Best late-breaking paper: Iain Cruickshank, Tony Johnson, Timothy Davison, Matthew Elder, and Kathleen Carley. "Detecting Malware Communities Using Socio-Cognitive Mapping".
2019, Kathleen M. Carley, Honorary Doctorate, "Doktor ehrenhalber" Faculty of Business, Economics and Informatics at University of Zurich, Zurich Switzerland
2019, Kathleen M. Carley, Computational Social Science – Quo Vadis? An Interdisciplinary Symposium Honoring Kathleen M. Carley, University of Zurich, Zurich Switzerland
2020, Kathleen M. Carley Ranked 286 in the world and 189 in United States of top 1000 Scientists in Computer Science in Guide2Research

Technology Transfer

Provisional patent application: Docket No. 6069-2111; "Binary Attribution Labeling for Multi-Author Software (BALMS)," William La Cholter, Timothy Davison, Matthew Elder, Peter Hamilton, and Antonius Stalick; October 23, 2020; provisional application.

Efforts are ongoing to release to the open source community the code used for mining GitHub in the paper "Windows Malware Binaries in C/C++ GitHub Repositories: Prevalence and Lessons Learned" presented at the International Conference on Information Systems Security and Privacy (ICISSP) conference in February 2021.

Participants

The participants that worked one person month or more on the project (not including students) were the following:

- *Matthew Elder, Ph.D.*
- *Ian McCulloh, Ph.D.*
- *William La Cholter*
- *Kathleen Carley, Ph.D.*
- *Tony Johnson, Ph.D.*
- *Tim Davison*
- *Michael Kelbaugh*
- *Jenna Nuth*
- *Anton Stalick*
- *Nithya Prakash*
- *Ian Crossett*
- *Samantha Cooley*
- *Anna Yaroslaski*

Students

Undergraduate: Ben Valdes, Chris Park, Juliet Yu, Kevin Bilzer, Mitchell Dzurick, Oksana Carlson, Stephen Ballenger, Alan Padilla, Henry Tran, Tu Van Nguyen, Kevin Chen, Alec Davila, Karol Pierre, Sydney Medina, Juhi Patel, Naomi Albert, Elias Enamorado, Benjamin Padgette, Anshika Patel

Graduate: Iain Cruickshank (graduated with Ph.D., August 2020), Stephen Dipple, Christine Sowa, R.J. Joyce, Seamus Burke, Tim Zhang, Michael Tran, Yu Qiu, Cheng Xu, Zichao Yang, Max Gorbachevsky, Dylan Johnson, Trevor Lewis, Travis Barron, Jeffery Cantrell, Darren Gregory, Jeffrey Hogan, Katie Estes, Taylor McCrary, Parth Patel, Nada Sager, Kavya Kurup, Dhruv Mehta, Pritom Das Radheshyam

Products

Please see the previous section above describing how results were disseminated for the listing of products.

Acknowledgment:

This work was funded by the Minerva Research Initiative and is sponsored by the Department of the Navy, Office of Naval Research under ONR award number N00014-18-1-2111. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.