Preview of Award 1127425 - Annual Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1127425
Project Title:	SDCI Sec: SESv3 (Security Event System - Version 3)
PD/PI Name:	Douglas D Pearson, Principal Investigator
Submitting Official (if other than PD\PI):	Douglas D Pearson Principal Investigator
Submission Date:	07/25/2013
Recipient Organization:	Indiana University
Project/Grant Period:	08/01/2011 - 07/31/2014
Reporting Period:	08/01/2012 - 07/31/2013
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Douglas D Pearson

Accomplishments

*** What are the major goals of the project?**

The goals of the SESv3 project are threefold: (1) provide a near real-time path from the observation of security-related events, to the derivation of legitimate indicators, to the application of protections against threat; (2) provide a rich repository of threat and security event information with flexible analyst interface; and (3) to accomplish this in the context of a multi-institutional peer community of trusted security practitioners, and among multiple, similarly-focused trust communities.

- Provide capability (technical and policy) for inter-federation among trust communities
- Establish interoperation with private, commercial, and governmental threat mitigation communities and organizations
- Incorporate and correlate BGP data, accelerating the identification of rogue networks
- Correlate with passive DNS data, permitting the identification of malicious name servers associated with rogue and fastflux networks
- Provide for the automated, real-time hand-off of malware binaries to third-party anti-virus partners, and to third-party sandbox analysis, incorporating analysis results into the intelligence cloud
- Provide repository performance and scaling enhancements
- Provide additional operational interfaces for event and incident submission and query, integrating SES into common incident responder and analyst tools and workflows
- Provide linkage of unstructured data, e.g. e-mail discussion regarding Internet element identifiers
- Contribute to the evolution and programmatic use of standards-based data representations
- Provide support for IPv6-based correlations

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities: 1. CIF v1 was released to the community as a pilot for v2 (aka: SESv3)

A complete changelog of activity can be found here:
https://github.com/collectiveintel/cif-v1/blob/master/ChangeLog

The release of cif-v1 provides for migration from the v0 framework to the v2 framework (aka: SESv3), and provides a pilot of the interconnects of the v2 (SESv3) framework. This allows us to test our ideas against the live community, while providing direct value to them for their time, testing, and feedback.

The cif-v1 release is a composition of the following:

- libcif client libraries (used in building applications against the framework)
- libcif-dbi libraries (database interface)
- cif-smrt framework (intelligence parser)
- cif-router (provides the API style interface to the data-warehouse)
- cif-protocol (the core routing protocol for the framework)

The architecture baasd on these pieces allows for scaling of the framework in terms of:

- easier application integration abstracting the data-schemes, protocols and security mechanisms from the developer (eg: libcif)
- easier scaling of the framework (splitting the cif-router so it can be scaled in the same fashion traditional network architecture is)
- splitting out the core protocol, allowing others to build against the protocol in their own integration methodologies

With this release, we also abstracted the transports from the core interfaces allowing lower level developers to introduce new transports as they evolve (eg: HTTP/TLS, ZeroMQ. etc). This allows the framework to scale past traditional, more hard-coded barriers (eg: HTTP/TLS + REST, HTTP+SOAP, etc).

2. Malware specifications and prototypes completed

Independent of the v1 release, we've extended our lessons learned in protocol definition and implementation towards other known malware standards. While these standards didn't make it into the v1 release, they serve as a model for generation of a “cookie cutter” protocol implementation process, allowing the framework to rapidly evolve and implement newer protocols as the space evolves.

Examples of this have since been published to github for incorporation with v2 (SESv3).

- https://github.com/collectiveintel/mmdef-protoco
- https://github.com/collectiveintel/mmdef-pb-perl
- https://github.com/collectiveintel/mmdef-XmlToProtobuf

3. CIF v2 alpha completed against all major objectives (with the exception of passive dns and bgp integration, TBC: end-2013)

4. Documentation

Recognizing that documentation is critical to adoption and use, substantial effort has been invested to create thorough documentation. One hundred and seventy-five (175) wiki "pages" of CIF documentation cover topics including: overview to the CIF approach to information sharing; server installation (multiple Linux variants), configuration, initialization and operation; client installation and example use; integration with various other tools (commercial and open source); instruction and examples for ingesting various data sources; API description and use; cookbooks for deploying on Amazon EC2, sharing threat intelligence, and interfederated

sharing; advanced topics; and FAQ.

The starting page for CIF documentation can be found here https://code.google.com/p/collective-intelligence-framework/wiki/WhatisCIF

5. Community building (better process, etc)

During this second year of the project the CIF community (persons tracking, using, and/or contributing to the project) more than doubled from 80 to 180+ persons. Sustaining the large community has made significant demand on resources for Q/A, bugfixes, and documentation. This highlighted a need for a better community process, in terms of answering questions, writing better doc, more effective issue tracking and lowering the barrier of entry towards contributors.

With that, we've learned that breaking up and compartmentalizing the various pieces of the project, as well as leveraging more social constructs such as google groups, google code and github has allowed us to scale past what we'd be able to traditionally in-house.

This scale allows us to maintain over 60 code repositories as well as many cross development initiatives within some of the core repo's (eg: v1 vs v2 development). All while keeping them in sync, allowing easy “drive-by” contribution from our community members (and/or interns who require less access) and allowing us to seamlessly context-switch between various parts of the project on a regular basis.

The end result has been:

- easy cross-coordination between development branches
- more external contribution, everything from small patches to structural fixes
- easier search and retrieval of knowledge (eg: google groups searching of the community archives, making it easier to derive new FAQ’s).
- more efficient utilization of resources (we can do more with much less)

By engaging our user base early and being more transparent with them from alpha to release, we've forced ourselves into a more efficient, transparent development process.

With that, we are now able to rapidly develop and release newer code, protocols and documentation, shortening the development life-cycle from feedback through release.

Specific Objectives: NA

Significant Results: Doubled the size of the CIF community (from 80 to 180+).

Google protocol buffers work much faster than XML and even JSON.

Writing good, complete documentation is really hard, however is a critical driver for adoption.

Community building takes time, cycles, and focused resources.

Open-source projects of this scale require a sophisticated long-term business model, e.g. something between a foundation (501cX in the US), private equity fund, and a commercial business. Something that allows for multiple funding streams, keeps the technology open-source, and balances governance to the interest of the served community.

Scripting languages (perl, python, ruby, etc) are difficult to scale by their nature.

If you make it easy for your community to contribute back, they will, usually in ways that both expedite and slow down the work process. The end result is adoption growth and a more resilient platform. The difficult piece is getting things in place where the community is self-sustaining and not dependent on single-threaded resources.

The faster you get a release out, the faster the adoption of your project, the more feedback the project will get in return, the more successful the project will be.

Being as transparent as possible with a project (features and bugs) significantly lowers the barrier to entry for early adopters as well as contributions.

Key outcomes or Other achievements:

The release of cif-v1.

User growth.

3rd party presentations.

*** What opportunities for training and professional development has the project provided?**

A summer 2012 graduate student (Kevin Benton) under funding under TransPAC3 (NSF OCI award #0962968) worked to establish a pilot security event information sharing with the TransPAC3 APAN partner. The student implemented a testbed of multiple CIF instances, each receiving event information from its local 'community' (federation), and established inter-federation sharing among the instances. The installation, configuration, and implementation steps were carefully documented - as a cookbook for establishing CIF interfederation. The student worked with the APAN partner to guide their implementation of a CIF instance, and then established inter-federated sharing among the instances. Results of the work were presented to the 2013 Techs In Paradise conference.

Subsequent to the TransPAC3 work, Kevin was retained under this SESv3 grant as a Graduate Assistant to perform work on the SESv3 deliverable: "provide for the automated, real-time hand-off of malware binaries to third-party anti-virus partners, and to third-party sandbox analysis, incorporating analysis results into the intelligence cloud". Kevin completed work to represent the Malware Metadata Exchange Format (MMDEF) in Protocol Buffers, developed a code library for manipulating the malware objects, adapted the open source Cuckoo malware sandbox for submission to CIF, and developed a VirusTotal/CIF interface. Kevin's final report on work performed is attached as file "Kevin_Benton_SESv3_GA_Report_2012.pdf".

*** How have the results been disseminated to communities of interest?**

1. CIF Community

- 180+ members
- https://groups.google.com/forum/?fromgroups#!forum/ci-framework

2. Communicating how others are employing CIF:

- Community Examples: https://code.google.com/p/collective-intelligence-framework/wiki/CommunityExamples

3. Reports

- ENISA: Proactive Detection of Network Security Incidents http://www.enisa.europa.eu/activities/cert/support/proactive-detection/proactive-detection-report
- CERT Polska http://www.cert.pl/PDF/MP-IST-111-18.pdf

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

We plan to ramp up development of our v2 (SESv3) code and start soliciting feedback from the community towards the end of 2013, much as we did with the v1 release. By the finish of our project, the v2 code should be production grade and have had some months of testing done as well as some forms of implementation in the cif-user community.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
Kevin_Benton_SESv3_GA_Report_2012.pdf	GA (Kevin Benton) final report, work accomplished	Douglas Pearson	07/23/2013

Products

Journals

Books

Book Chapters

Thesis/Dissertations

Conference Papers and Presentations

Other Publications

Wes Young, Principal Security Engineer, REN-ISAC (5/23/13). *Big-data: solved! … now what?*. AusCERT 2013: https://collective-intelligence-framework.googlecode.com/files/2013_auscert_wesyoung.pdf.

Status = OTHER; Acknowledgement of Federal Support = No

Wes Young, Principal Security Engineer, REN-ISAC (2/18/13). *intro to [effective] data-sharing, 101*. MAAWG seminar; https://code.google.com/p/collective-intelligence-framework/downloads/detail?name=2013_MAAWG_wesyoung.pdf.

Status = OTHER; Acknowledgement of Federal Support = No

Wes Young, Principal Security Engineer, REN-ISAC, Tom Millar, Chief, Communications, US-CERT (8/22/12). *Big Data, Collaborative Incident Response and Shared Situational Awareness in the Real World*. GFIRST; https://code.google.com/p/collective-intelligence-framework/downloads/detail?name=2012_SANS.pdf.

Status = OTHER; Acknowledgement of Federal Support = No

Wes Young, Principal Security Engineer, REN-ISAC (6/22/12). *Sharing data's hard, here's how we did it*. FIRST; https://code.google.com/p/collective-intelligence-framework/downloads/detail?name=2012_FIRST.pdf.

Status = OTHER; Acknowledgement of Federal Support = No

Technologies or Techniques

Nothing to report.

Patents

Nothing to report.

Inventions

Nothing to report.

Licenses
Nothing to report.

Websites

Title: collective-intelligence-framework

URL: http://code.google.com/p/collective-intelligence-framework

Description: All documentation concerning the project, and home for the user community.

Major work during this reporting period on:

developer workflow documentation
https://code.google.com/p/collective-intelligence-framework/wiki/DeveloperWorkflow_v1

community examples documentation
https://code.google.com/p/collective-intelligence-framework/wiki/CommunityExamples

and more specific OS install doc (CentOS, Ubuntu, Squeeze)
https://code.google.com/p/collective-intelligence-framework/wiki/ServerInstall_v1

Other Products

Product Type: Software or Netware

Description: Our first stable release of CIF, v1:

https://github.com/collectiveintel/cif-v1

https://code.google.com/p/collective-intelligence-framework/wiki/ServerInstall_v1

This is the first deployable iteration on the path to CIF v2 (aka: SESv3), currently soliciting community feedback for integration with CIFv2 (aka: SESv3).

Other:

Participants

Research Experience for Undergraduates (REU) funding

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Wes Young	Other Professional	12
Gabriel Iovino	Other Professional	6
Chris O'Donnell	Other Professional	3
Douglas D Pearson	PD/PI	1
Kevin Benton	Graduate Student (research assistant)	2

What other organizations have been involved as partners?

Name	Location
CMU Software Engineering Institute	Pittsburg, PA
Nlckel City Software	Buffalo NY

Have other collaborators or contacts been involved? Y

Impacts

What is the impact on the development of the principal discipline(s) of the project?

Traditionally, most data-sharing in the operational space happens either via human conversations (email, blogs, im, etc) or via the transfer of simple key-pair values (csv, json, etc). With the release of CIF v0 and v1, we've prototyped a way to represent complex data using the IETF RFC 5070 standard ("IODEF") in a faster and more scalable fashion. Traditionally this has been something left to the slow and sluggish forms of XML, which is why most of the operational community has averted it for the past decade. With these recent releases we've developed a model where user interfaces can still interact with the system in their native types (plain language, or simple key-pairs), but the system can rapidly translate these into complex data-types for correlation, etc.

This model provides example in adapting these paradigms to other standards and technology, it is not limited to protocol buffers or IODEF. The libraries were developed in a way that is easily replicable to other types of technologies and standards. We've developed a "cookie cutter" pipeline process for developing and releasing standardized protocols as the security space rapidly evolves.

What is the impact on other disciplines?

To-date, most "big-data" technologies focus on how to rapidly compute massive amounts of simple key-pair values, our goal is to break that paradigm and allow for the computation and correlation of largely, multi-dimensional complex data leveraging something that scales better than traditional XML and computationally intense XML analytics.

Our framework is built to be non-specific to security intelligence, and more towards a generic data-sharing platform, which over time can be easily adapted to any other discipline as a storage and routing platform for normalizing, warehousing and correlating complex intelligence.

What is the impact on the development of human resources?

Lower barrier to entry when researching threat intel for students, researchers, and new professionals

Lower barrier to entry when researching programming paradigms to scale data-warehouses and messaging frameworks for students, researchers, and new professionals

Provides community for new professionals, students and researchers to learn from, ask questions, test theories.

What is the impact on physical resources that form infrastructure?

This framework lowers the cost of protecting infrastructure from malicious threats, and allows for the interfederation of threat intelligence between entities that own and operate infrastructure.

What is the impact on institutional resources that form infrastructure?

This framework lowers the cost of protecting infrastructure from malicious threats, and allows for the interfederation of threat intelligence between entities that own and operate infrastructure.

What is the impact on information resources that form infrastructure?

This framework lowers the cost of protecting infrastructure from malicious threats, and allows for the interfederation of threat intelligence between entities that own and operate infrastructure.

What is the impact on technology transfer?

Enables commercial or open-source projects to build these concepts into their products

Enables commercial or open-source projects to openly integrate with these technologies

Allows technologists to re-architect the platform for their own internal use

Gives other open source projects a model for building out their technology

Provides components that other projects can leverage in building their own technology (libraries, etc).

What is the impact on society beyond science and technology?

Gives the security community something tangible to work through the international data-sharing problem

Lowers the cost of security, enables focus on other innovations

Raises the cost of doing business for the bad guys

Lowers the cost of remediation and cleanup

Provides a model for sharing data between federations

Provides a model in which other open-source projects can leverage to generate their own communities of users

---

Changes

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.

Special Requirements

Responses to any special reporting requirements specified in the award terms and conditions, as well as any award specific reporting requirements.

Identification of open source license used:

Each of the components is specifically licensed in relation to it's technology (the generally accepted community standard for that type of code). The licenses vary from: BSD, LGPL and GPL. The default licenses are CCv3 (content, presentations) and LGPL (code).

Working demonstration description and results:

https://code.google.com/p/collective-intelligence-framework/wiki/ServerInstall_v1
https://code.google.com/p/collective-intelligence-framework/wiki/CommunityExamples
https://code.google.com/p/collective-intelligence-framework/wiki/Preso#3rd_Party_Presentations

Identification of end users engaged, trialing, and/or using the system:

End user activity for the entirety of the project can be found here:
https://groups.google.com/forum/?fromgroups#!forum/ci-framework

More specifically, some mid-level adopters have cited their view of the project with respect to their operational needs:
https://groups.google.com/forum/?fromgroups#!topic/ci-framework/wGpnN5MxTwI

Among those:

Brian Smith-Sweeney
Assistant Director of Global Security Operations, New York University

I help run the IT security program at NYU; we're the largest private University in the US and have a presence in ~20 countries, with a significant footprint in Abu Dhabi, China, Australia, and a number of places in Europe.

We are currently using CIF to share data with a reasonably well-known private trust community, in two ways: 1) we integrate data into our monitoring infrastructure (dns log monitoring, netflow monitoring, etc.) to find bad stuff automatically, and 2) our analysts use CIF clients via browser plugins and/or ArcSight integration manually during investigations.

There is a non-zero chance we will someday use it to share internal threat data among peers at our global sites, but that's purely in idea stage.

Jose E Hernandez
Prolexic Technologies (DDoS Mitigation)

We utilize CIF to cross reference threat feeds with data collected from attacks we witness and mitigate. Although we have an in house threat DDoS only feed that we publish out to our customer we would like to integrate it with CIF as a long term goal. Also I host a public instance of CIF for whom ever wants to give it a try at https://feed.josehelps.com and for personal use.

Matt Carothers
Cox Communications (ISP)

We will use the Collective Intelligence Framework to aggregate and normalize a variety of threat intelligence feeds. We receive feeds of infected subscribers from several sources, and CIF will allow us to put them all in one place where we can query them easily. We also plan to use CIF to publish our own data back out to the community. For instance, we see millions of malware-infected hosts on other providers' networks communicating with hosts on our own, and we plan

to provide those to the responsible ISPs in the form of a CIF feed.

David Kovar
Ernst and Young

My name is David Kovar and I am a manager with Ernst and Young. My focus is mostly on incident response and forensics, but I'm trying to get a reasonable grounding on the threat intel side. I'm also a major proponent of open source solutions and tools, and am concerned about the growth in the threat intel industry at a time when we need to share more rather than lock intel behind very large financial contracts. But I understand capitalism, the costs of collecting data, and costs making it consumable.

Interesting times.
When I first looked it months ago, CIF struck me as a superb platform for turning a lot of data collections into data that was much easier to work with and it has lived up to that expectation and more. I'm enjoying watching CIF, and the CIF community, grow all while enjoying learning how to apply CIF to the many different clients we work with.

The NSF grant that supports CIF is probably some of the most effective and impactful money the US government has spent on cyber security.

Date of initial use of nmi build and test:

Initial perl based modules tested via standard perl test system (cpantesters.org):
http://www.cpantesters.org/distro/N/Net-DNS-Match.html#Net-DNS-Match-0.04
http://www.cpantesters.org/distro/N/Net-Abuse-Utils-Spamhaus.html#Net-Abuse-Utils-Spamhaus-0.04
http://www.cpantesters.org/distro/L/LWPx-ParanoidAgent.html#LWPx-ParanoidAgent-1.09
http://www.cpantesters.org/distro/I/Iodef-Pb-Simple.html#Iodef-Pb-Simple-0.18

After releases stabilize, certain modules will be rewritten in C for performance and stability reasons. At that the, the NMI Build and Test framework will be leveraged, among other open alternatives.

Milestones reached against stated project goals

+ CIF v1 released to the community as a pilot for v2 (aka: SESv3): https://github.com/collectiveintel/cif-v1/blob/master/ChangeLog
+ CIF v2 alpha completed against all major objectives (with the exception of passive dns and bgp integration, TBC: end-2013)
+ to release CIF v2 beta-1 end of 2013
+ malware specifications and prototype's completed
+ milestones can be tracked to maturity here: https://github.com/collectiveintel/cif-v2/issues/milestones