Lowering Cost of Data Exchange for Analysis and Defence

Paweł Pawliński
CERT Polska / NASK
pawel.pawlinski@cert.pl

Adam Kozakiewicz
NASK
adam.kozakiewicz@nask.pl

April 10, 2015

1 Motivations for exchange of security data

Using information from multiple sources for the purpose of mitigation of threats and proactive improvement of security posture is becoming a common practice in organizations of any size.

In general, we can distinguish two main complimentary approaches that leverage security-related information: defense and analysis. Straightforward application of information with the goal of defending a computer network is a natural way of securing organizations' infrastructure. For example, indicators of compromise allow to identify malicious activity like connections to botnet command and control (C&C) servers or presence of software implants on compromised machines. Other indicators are helpful to detect and disrupt attacks before any damage is done. Also there is a wealth of datasets collected from multiple internal systems – various types of logs, network traffic records, etc. – which can be correlated with other sources during incident response or, proactively, during so called *hunting operations*. [1]

Application of security information for defence in principle is available and, to some degree, even necessary for any organization. However, ones that have sufficient capabilities (e.g. security vendors, some national CERTs) can take advantage of the large number of available data sources and implement analytical processes that combine these datasets and extract additional valuable knowledge.

2 Cost of data source integration

For both of the aforementioned purposes, it would be desirable to get access to and integrate as many security data sources as possible (assuming that they have unique content), since each of them might eventually provide a valuable piece of information. Naturally, in practice there are limitations to the number of sources that an organization can ingest, most importantly the costs of integration and maintenance. Costs of adding a new source can be broken down into the following elements:

1. Administrative overhead associated with entering into a relationship with an external party.
2. Expenses associated with getting access, e.g. licence fee.
3. Cost of adapting existing infrastructure to accept a new transport mechanism, e.g. if the source uses XMPP for sending data, it must be supported by the local platform that is used for collection.
4. Similarly, the cost of adaptation for new data formats, e.g. creating parsers that will normalize incoming data into form suitable for local storage.
5. Evaluation of data quality, especially its accuracy and completeness. It is often a non-trivial task and in many cases organization perform only rudimentary checks of the incoming data.

Apart from the integration, there are ongoing maintenance costs, mostly associated with management and monitoring of existing sources. With sufficient automation, maintenance costs for multiple sources can be limited, and require human intervention only in case of a technical malfunction or, on a higher level, problems with the data itself. On the other hand, total integration costs are mostly proportional to the number of distinct data sources.

Since budget allocated for collection of information is usually limited, organizations tend to integrate a small number of external sources. Selection of these sources is often based on their subjectively perceived value, since comprehensive evaluation can be costly and there are few comparative studies on the subject (e.g. ENISA report from 2011 [2]).

More in-depth discussion of these issues can be found in the recent ENISA report on *actionable information*. [3]

3 Existing solutions

Traditionally, the most common mechanisms for data sharing was plain HTTP and email for transport and a multitude of data formats, often based on CSV serialization. However, it can be observed that the situation improves gradually as new solutions that can lower the cost of data exchange are gaining adoption in the security community. Below we present three current trends that in our opinion have most significant influence in this regard.

MITRE standards The Structured Threat Information eXpression (STIX) format [4] created by MITRE provides syntax and semantics (informal ontology) to encode a wide range of concepts related to security. The broad scope and increasing adoption in the community are two main factors that differentiate STIX from previous similar efforts like IODEF. The Trusted Automated eXchange of Indicator Information is a complementary standard that specifies transport layer for automated communication.

Proprietary APIs Security vendors and other information providers tend to base interfaces to their services on common web standards, in particular HTTP, REST, and JSON. Custom formats allow more flexibility for the providers and building on popular technologies make the implementation work easier on the receiving side. Nevertheless, there remains an issue of non-standardized and sometimes under-specified semantics, which causes additional effort of normalizing the incoming data.

Alternative formats There exist initiatives that attempt to fill the gap between standardized but complex formats, and custom proprietary solutions. An early example is the

WAPI interface (part of the WOMBAT project[1]) which provided a generic object-oriented data access API with built-in support for dataset introspection and discovery. n6 API is another solution, which was created by CERT Polska (part of NASK) for distribution of threat data feeds on a national level (the project is described in more detail in the next section). n6 design prioritizes simplicity of the query interface and response format, at the same time being generic enough to address a majority of information exchange scenarios.

4 Case study: n6 platform

Development of the n6 platform[2] started in 2011 to unify and systematize data sharing processes in CERT Polska, which has the role of the national CERT for Poland. Ability of national CERTs to tackle threats to its constituency are usually limited, hence such CERTs are naturally focused on forwarding information to entities that are affected and ones that can take concrete mitigation steps. As a part of its mission CERT Polska aims to be an information broker for Polish entities both in the private and the public sector, providing them with high-quality information, free of charge.

n6 is used to collect, manage, and distribute multiple types of security feeds, including infection data, C&C servers, attack sources, malware targets. Recipients use the REST API, which exposes a unified, normalized output format for over 50 datasets integrated in the platform. The native output format is based on a simple JSON structure, which can be easily integrated with constituents' internal systems. To provide compatibility with existing systems and further lower the cost of integration, IODEF and normalized CSV are also offered as alternative data formats.

At the time of writing, almost 300 organizations are subscribed to the platform and receive pre-filtered, relevant data feeds.

5 Case study: NECOMA project

NECOMA[3] is a joint European-Japanese project aiming to improve the security data analysis capability and the applicability of analysis results to defence. The project takes advantage of a multitude of diverse data sources available within the consortium, spanning from infrastructure to end-point layers. An important aspect of NECOMA is application of cross-layer analysis techniques to extract new information from combination of different datasets.

In this approach, developers of analysis modules must overcome the cost of integrating multiple data sources, which means that a common data access method is especially important. The NECOMA consortium chose to use the n6 REST API as the basis of its inter-organizational automated data sharing mechanism. Main advantages of using the n6 API was its simple implementation (both on the consumer and producer side) and flexibility, which meant that the API could be easily adapted for new types of information. Moreover, thanks to basing the solution on plain HTTP with JSON serialization, the API has low overhead.

This new role resulted in a number of extensions to the n6 platform. The most important was development of a server-side SDK[4] (Software Development Kit). It further simplifies implementation of an n6-compatible API on top of arbitrary datasets. The software has been released on an open-source licence and is accompanied by a complete step-by-step integration guide. SDK was successfuly used to integrate a complex data warehouse developed by the Japanese part of consortium - MATATABI - with the n6 platform maintained by CERT Polska.

The REST API itself was extended to accomodate some of new types of data exchanged within the project. Additionally, a new stream API was introduced to enable subscription-based real time feeds.

The easy access to various data proved crucial for the work on threat analysis. Many new analysis methods were proposed and tried, connecting different datasets. For example an algorithm for malicious campaign detection based on frequent pattern trees (FP-trees) processes seamlessly malicious URLs or domain names from multiple sources. A new method for C&C communication detection in sandbox traffic dumps makes use of the n6 interface to import learning data from various sources. A graph-based cross-layer analysis (still work in progress) mixes data from a large number of sources using common tokens as search keys. All these approaches scale in regard to the number of sources only in presence of a simple, common interface for data access.

The n6 interface proved flexible enough to enable it to be used in other parts of the system as well. Apart from delivering data from datasets to analysis modules, the API can be used to deliver results of the analyses to defence mechanisms for mitigation purposes and to provide feedback from these mechanisms to the NECOMA platform.

The experience from NECOMA confirmed the that lowering interoperability costs on the producer and consumer side facilitates development of new analysis methods.

6 Conclusions

In this paper we discussed barriers limiting data sharing in the security community in the context of the cost of integrating data sources and selected approaches that may allow to lower this cost, facilitating collaboration. We also presented two real-world use cases that demonstrate how our solution for data sharing – the n6 platform and associated SDK – is used to exchange large datasets both for operational and research purposes.

References

[1] Jones, G. M., Stogoski, J., "ALTernatives to Signatures (ALTS)", CERT Coordination Center, Software Engineering Institute, 2014. http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=296146

[2] Gorzelak, K., Grudziecki, T., Jacewicz, P., Jaroszewski, P., Juszczyk, Ł. and Kijewski, P., "Proactive Detection of Network Security Incidents", ENISA, 2011.

[3] Pawliński, P., Jaroszewski, P., Kijewski, P., Siewierski, Ł., Jacewicz, P. and Zielony, P., "Actionable Information for Security Incident Response", ENISA, 2015.

[4] Barnum, S., "Standardizing cyber threat intelligence information with the Structured Threat Information eXpression (STIX^{TM})", version 1.1, MITRE Corporation, 2014.

[1] Worldwide Observatory of Malicious Behaviors and Attack Threats, EC FP7 grant no. FP7-ICT-216026-WOMBAT, 2008-2011

[2] http://www.cert.pl/projekty#n6

[3] Nippon-European Cyberdefense-Oriented Multilayer threat Analysis, EC FP7 grant no. 608533

[4] https://github.com/CERT-Polska/n6sdk