Data Breach Databases - Lot of Incidents, just few Data

Rocco Gagliardi
Defense Department, scip AG
roga@scip.ch
https://www.scip.ch

Marc Ruef (Editor)
Research Department, scip AG
maru@scip.ch
https://www.scip.ch

Abstract: We know for sure that we are constantly under attack, but by whom and how? To date, information on IT incidents is practically ridiculous compared to their number. The few public data present on the web are not structured, thus making it difficult to use. Today, CTI generic reports help to identify the most obvious critical issues, now we need details.

Keywords: APT, Assessment, CIA, CTI, Cybersecurity, Flash, Framework, GitHub, Governance, JSON

1. Preface

This paper was written in 2019 as part of a research project at scip AG, Switzerland. It was initially published online at *https://www.scip.ch/en/?labs.20190502* and is available in English and German. Providing our clients with innovative research for the information technology of the future is an essential part of our company culture.

2. Introduction

We are adding technology to every aspect of our lives; the net is becoming a medium like air and water; like them, it can carry threats. If a country is under strong attack, is it possible to have a complete disconnection from the Internet? Now, nobody knows what would happen; with 5G it will be increasingly probable to have a physical actor remotely controlled by an A.I. distributed at a transnational level: total or partial disconnection will become impractical. After having discarded the easy solution, the difficult one remains: we must *assess the risks*, *prevent accidents* and be *ready with the recovery procedure*.

But how can we decide the *priorities*? Cyber attackers are often looking backward more than forward, it is therefore sufficient to analyze what has happened so far and act accordingly. Sounds easy, but what has happened so far?

We all know so-called *Cyber Threat Intelligence* (CTI) Reports, we all read our yearly set of documents – full of statistics, charts, trends and comments – and we all do not have any idea about the underlying, often proprietary, data. But when we read the documents, we can note at least the following:

- In a world where we have tons of standards, everyone tend to use own identifiers (simple example: Vietnam, VietNam, Viet Nam)
- We have just a ridiculous amout of detailed data about incidents

3. The need of quality data

Until now *generic CTI reports* have helped to identify the most obvious critical issues, now it's time to go further, *we need details* about actors, victims, and vectors. Maintaining a list of incidents and statistics is useful, but to make data usable they must be normalized and include technical, industrial, and social aspects. Many lists and databases exists with data describing actors, victims, impacted assets, and other aspects of an attack, but normally they just describe what happened or categorize a very minimal number of information.

For specific tasks, in risk management process, it is useful to have *solid data and not just perceptions*. A good dataset should answer questions like what is the role of flash drives in incidents? How flash drives are involved in attacks to companies like mine?

4. Crafting our own CTI Reports

To craft our own report, we need raw high quality data; we can then observe them from our perspective and extract the facets we need. It is not enough to have *Financial* or *Malware* and another couple of columns to highlight – even with limited data – the patterns used in common attacks; we need to know more precisely who the actors are, who the victims, and what is being manipulated.

VERIS Database [1] is the most structured and complete database of incidents we have seen; incidents are analyzed, data – where known – are normalized and verified, so that the consistency is maintained. Each incident is described with four elements:

- **Actors**: Whose actions affected the asset? Includes type and country.
- **Actions**: What actions affected the asset? Includes type, vector, vulnerabilities.
- **Assets**: Which assets were affected? Database, Webserver, Fileserver, etc.
- **Attributes**: How the asset was affected? Which "C-I-A Triad" part is affected.

Furthermore, the victim is assigned to an industrial sector using the *NAICS standard* (but the conversion to SIC/ISIC and others is possible), this is fundamental to obtain data concerning specific companies in our sector.

As example, in *Manufacturing* we have *Petroleum and Coal Products Manufacturing* and *Chemical Manufacturing* and although similar, there can be many differences, especially in IT standards. Taking a quick look at the data, we can spot the pattern used in attacks, the differences, and adjust our risk matrix:

Pivoting data from the *Sector* perspective:

Sector	**Asset Variety**	**Action Variety**	**Actor**
3240 – Petroleum and Coal Products Manufacturing	S – Database	Misconfiguration	Internal
3250 – Chemical Manufacturing	U – Laptop	Theft	External
3250 – Chemical Manufacturing	M – Payment card	Possession abuse	Internal

Pivoting data from the *Asset* perspective:

Asset Variety	**Actor**	**Action Variety**	**C-I-A Impact**
S – Database	External	Knowledge abuse	CIA
S – Database	Internal	Misconfiguration	-IA
S – Database	Internal	Abuse of functionality	-IA

Although they may seem like hot water, these are facts! Different from perceptions and with a high value as starting point or as integration in a risk management framework.

The *VERIS Database* is available as *JSON* file, very easy to parse with R, Phyton/Panda/Jupiter, or other tools. For our risk analysis, we prefer to *flatten* the JSON in CSV, extract the parameters we are interested in and pivot the data with worksheets.

5. Other resources

As mentioned before, *VERIS Database* is not the only data source in the net. Here some database of incidents:

- *Hackmageddon* [2]: Hackmageddon is a list of incidents maintained by Paolo Passeri. Lot of statistics.
- *Data Breach Database* [3]: Database specialized on data breaches, just a list of events with fancy design. The key, here, is the number of records exfiltrated.
- *RISI Online Incident Database* [4]: Incidents database with event description and minimal assessment. No longer updated.
- *DataLossDB* [5]: Blog about incidents. No structure at all. No longer maintained.
- *Wikipedia list of data breaches* [6]: List of incidents with minimal structure. *DataBreaches.net* [7]: List of incidents aggregated from various sources. Not structured.

Here some other CTI resources:

- *Confronting an 'Axis of cyber'* [8]: Very interesting and actual report focusing on "the behavior of the usual suspects".
- *Global Cyber Strategies Index* [9]: Links all cyber strategies documents created by countries for different categories: military, critical infrastructures, crime, national strategy, privacy, content, commerce.
- *Center for Strategic & International Studies* [10]: A bipartisan, nonprofit policy research organization dedicated to providing strategic insights and policy solutions to help decisionmakers chart a course toward a better world.
- *Securelist* [11]: Kaspersky lab resources.
- *APT Groups and Operations* [12]: A spreadsheet normalizing the public available data about APT Groups and actions.
- *Awesome-Threat-Intelligence* [13]: A curated list of CTI resources.

6. Summary

When doing intelligence work, models are sought that take into account multiple variables; intuition is a good thing, but concrete facts are needed to support decisions on how much and where allocate resources. It is not an easy job mostly because, even with continuous incidents happening, there are not enough public data, nor analyzed and organized in an effective manner. *Several initiatives have been undertaken* [14] in recent years to at least be able to share data, but we have still a long way ahead.

7. External Links

[1] http://veriscommunity.net/vcdb.html
[2] https://www.hackmageddon.com/
[3] https://breachlevelindex.com/data-breach-database
[4] https://www.risidata.com/Database/event_date/desc
[5] https://blog.datalossdb.org/2016/02/
[6] https://en.wikipedia.org/wiki/List_of_data_breaches
[7] https://www.databreaches.net/
[8] https://www.ispionline.it/en/pubblicazione/confronting-axis-cyber-21458
[9] https://csis-prod.s3.amazonaws.com/s3fs-public/Cyber%20Regulation%20Index%20V2%20%28002%29.pdf
[10] https://www.csis.org/programs/cybersecurity-and-governance/technology-policy-program/other-projects-cybersecurity
[11] https://securelist.com/
[12] http://apt.threattracking.com
[13] https://github.com/hslatman/awesome-threat-intelligence

[14] https://www.us-cert.gov/Information-Sharing-Specifications-Cybersecurity