Temporal Analytics for Predictive Cyber Threat Intelligence

[Tool Presentation Abstract]

Staffan Truvé
Recorded Future
truve@recordedfuture.com

ABSTRACT

Recorded Future has developed its *Temporal Analytics Engine* as a general purpose platform for harvesting and analyzing unstructured text from the open, deep, and dark web, and for transforming that content into a structured representation suitable for different analyses.

In this paper we present some of the key components of our system, and show how it has been adapted to the increasingly important domain of *cyber threat intelligence*.

We also describe how our data can be used for predictive analytics, e.g. to predict the likelihood of a product vulnerability being exploited or to assess the maliciousness of an IP address.

1. TEMPORAL ANALYTICS

Temporal Analytics is by Recorded Future's definition the entire process of going from unstructured text describing events in the world, via a structured representation of world events, to analyses that describe the current state of world affairs or predict future events. Our structured representation is based on *entities* (persons, places, organizations, technologies, etc.) and *events* that relate entities to an event type and a set of metadata such as the the source of an event mention, its publishing time, and a derived *event time*. Analysis of this data can range from finding individual event mentions to quantitative analysis of the number of mentions, publishing and event time patterns, and machine learning based predictive models. Our *Temporal Analytics Engine* today harvests and analyses millions of documents – ranging from short tweets to long articles and reports – every day. The coverage is very broad, from government and big media sites to blogs, forums, and social media. To provide relevant data for cyber threat analysis, we have broadened the scope of our harvesting to include cyber related media, blogs, feeds, forums, as well as paste sites (web sites used to share code, logs, data dumps etc. anonymously) and parts of the "dark web" – primarily Tor/Onion sites.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889294 .

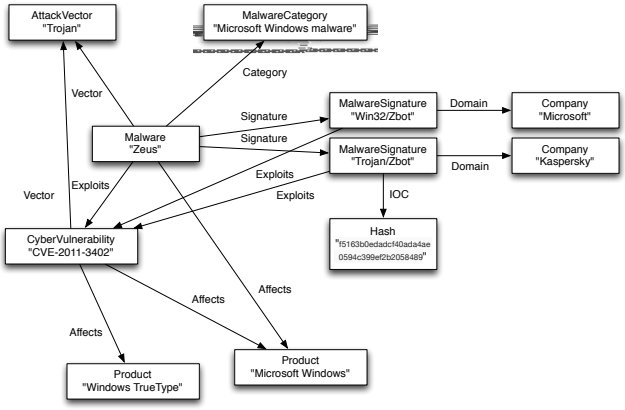

Figure 1: An example of the Cyber Ontology

2. CYBER ONTOLOGY

To adapt our system to cyber threat intelligence, we also added new entity and event types to represent this domain. In addition to standard named entities like persons, places, organizations etc. we added a set of new entity types of specific interest to the cyber domain. These include: malware, malware categories, attack vectors, vulnerabilities, and technical indicators such as hashes and file names. Figure 1 illustrates the entity ontology with some examples.

Recorded Future detects a large set of *events*, ranging from person and corporate related ones to geopolitical and environmental events. Each type has a set of (sometimes optional) named attributes.

To extend our world model to the cyber threat space, we added new event detectors for cyber attacks and vulnerability exploits. A cyber attack event relates an attacker to a target, and includes additional information about the attack method used and related hacktivist operation hashtags. All arguments are optional, but at least an attacker or a target must be identified.

3. DASHBOARDS AND TIMELINES

The data collected by Recorded Future can be analyzed in our web UI, which provides powerful ways of filtering and visualizing events, individually or in aggregated form. Figure 2 shows the cyber dashboard which gives an aggregated 60-day overview of cyber attack events, and Figure 3 is a line view illustrating the varying pattern of references to

Figure 2: Cyber Dashboard showing trending cyber attack events.

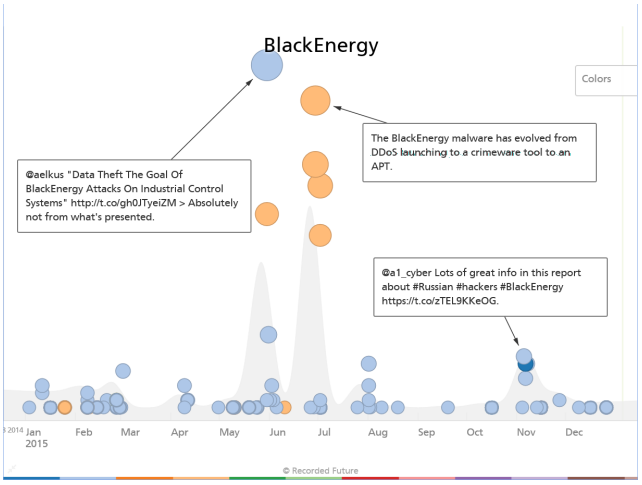

Figure 3: 2015 timeline for BlackEnergy malware.

the BlackEnergy malware during all of 2015. The user can interactively filter and zoom to drill down into details.

4. PREDICTIVE ANALYTICS

Our data can be used directly for some predictive analytics, since certain future events such as hacktivist attacks are mentioned ahead of time (see Figure 4 as an example), thus giving the defenders advance warning of upcoming threats. For other use cases, predictive models need to be constructed based on historic data. We provide two examples of such predictive models in the cyber threat domain.

Predicting Exploits

One interesting use of our data is to build a model for predicting which published product vulnerabilities are most likely to be exploited. In [1] we describe how a Support Vector Machine classifier can be trained to predict the likelihood of a vulnerability being exploited, with an accuracy of 0.83 (precision = 0.82, recall = 0.84), when trained on a data set consisting of 7,528 samples from 2010-01-01 - 2014-12-31, with an equal amount of exploited and unexploited vulnerabilities. One noteworthy result is also that the text used to describe vulnerabilities is more informative than the parameters assigned in the National Vulnerability Database (nvd.nist.gov).

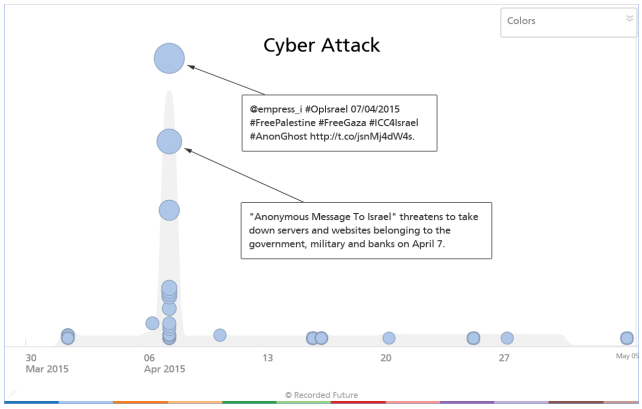

Figure 4: Future cyber attacks known on March 31st 2015 – a planned Anonymous attack on Israel on April 7th is the major known future event.

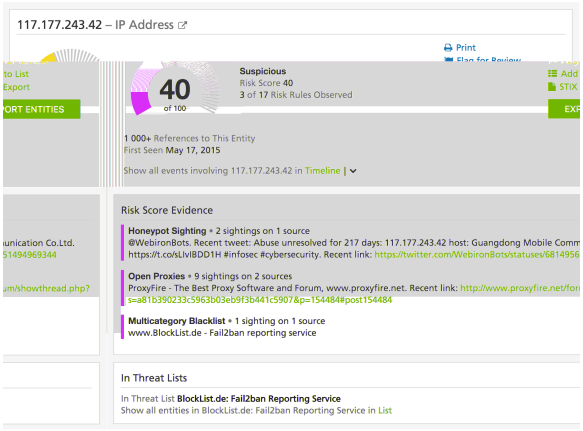

Figure 5: Risk scoring and entity information for IP address

Maliciousness of IP addresses

A second example of analytics based on our data is to assign a *risk score* to an IP address. Risk scores are valuable for threat analysts who want to get a quick aggregate score, to decide whether to explore an IP in further detail, and can also be used to automate blocking of IP addresses by network security equipment. Figure 5 show an information card for an IP address, with its risk score and associated information.

5. CONCLUSIONS

We have presented the basics of the *Temporal Analytics Engine*, and described how it has been adapted to the cyber threat intelligence domain. We have also illustrated its use in predictive analytics by showing two examples related to the prediction of cyber exploits and the assessment of maliciousness of IP addresses.

6. REFERENCES

[1] M. Edkrantz and A. Said. Predicting cyber vulnerability exploits with machine learning. In *Scandinavian Conference on Artificial Intelligence*, pages 48–57, 2015.