Data-Driven Threat Intelligence: Metrics on Indicator Dissemination and Sharing (#ddti)

Alex Pinto
Chief Data Scientist
Niddel / MLSec Project
@alexcpsec
@MLSecProject

Alexandre Sieira
CTO
Niddel
@AlexandreSieira
@NiddelCorp
MLSec Project / Niddel

- MLSec Project – research-focused branch of Niddel for open-source tools and community building
- Niddel builds Magnet, the Applied Threat Intelligence Platform focused on detecting breaches and malware activity
- Looking for trial prospects and research collaboration
- More info at:
  - niddel.com
  - mlsecproject.org
Agenda

- Cyber War... Threat Intel – What is it good for?
- Combine and TIQ-test
- Measuring indicators
- Threat Intelligence Sharing
- Future research direction (i.e. will work for data)

HT to @RCISCwenda
Presentation Metrics!!

50-ish Slides
3 Key Takeaways
2 Heartfelt and genuine defenses of Threat Intelligence Providers
1 Prediction on “The Future of Threat Intelligence Sharing”
What is TI good for (1) Attribution
Sony breach linked to Romanian external activist group

Executive Summary

On November 24, 2014, personally identifiable information about Sony Pictures Entertainment (SPE) employees and their dependents, e-mails between employees, information about executive salaries at the company, copies of unreleased Sony films, and other information, was obtained and released by a hacker group going under the moniker "Guardians of Peace" or "GOP".

Although the motives for the hack have yet to be revealed, the hack has been tied to the planned release of the film The Interview, which depicts an assassination attempt on North Korean leader Kim Jong-un, with the hackers threatening acts of terrorism if the film were to be released.

Recently, a team of 2 researchers from iDefense examined the evidence left behind by the attackers. This research has provided insight into the likely source of these attacks. Though not definitive, our analysis provides a much clearer picture and suggests an external activist group operating out of Romania is responsible for the data breach impacting Sony Pictures Entertainment. This disclosure casts further doubt on the FBI’s assertion that the attack was carried out by state-sponsored actors under the control of North Korea, a theory that has been all but discredited by a host of security professionals since the attack became public, including security product pre-sales engineer Nellie Nau.

Our product indicates a different, more sinister source behind the Sony attack.
— Nellie Nau, security product pre-sales engineer

The research team is quite certain, however, that the Guardians of Peace hacker group played no role in this attack. The clues left behind confirm that the group claiming responsibility were a fabrication to throw investigators off the trail and to mask the true source.

Links to Romania

The research team was able to reconstruct the attack from the ground up and discovered a number of IP addresses that are linked to other attacks that have been attributed to actors in Romania as well as the presence of Romanian text in the comment strings of the malware that was recovered during the forensic investigation. Some of these malware samples have also been used in Romanian attacks.

Additional signals intelligence acquired by the research team has also implicated an actor based in Romania. This intelligence is highly classified and cannot be released in a public document, but the research team has briefed investigators with the U.S. Federal Bureau of Investigation on their findings.
What is TI good for (2) – Cyber Maps!!

TY to @hrbrmstr for his work on
https://github.com/hrbrmstr/pewpew
What is TI good for anyway?

• (3) How about actual defense?
  • Strategic and tactical: planning
  • Technical indicators: DFIR and monitoring
Affirming the Consequent Fallacy

1. If A, then B.
2. B.
3. Therefore, A.

1. Evil malware talks to 8.8.8.8.
2. I see traffic to 8.8.8.8.
3. ZOMG, APT!!!
But this is a Data-Driven talk!
Combine and TIQ-Test

- **Combine** ([https://github.com/mlsecproject/combine](https://github.com/mlsecproject/combine))
  - Gathers TI data (ip/host) from Internet and local files
  - Normalizes the data and enriches it (AS / Geo / pDNS)
  - Can export to CSV, “tiq-test format” and CRITs
  - Coming Soon™: CybOX / STIX / SILK /ArcSight CEF

- **TIQ-Test** ([https://github.com/mlsecproject/tiq-test](https://github.com/mlsecproject/tiq-test))
  - Runs statistical summaries and tests on TI feeds
  - Generates charts based on the tests and summaries
  - Written in R (because you should learn a stat language)
```python
print(tiq.data.getAvailableDates("raw", "public_outbound"))
```

<table>
<thead>
<tr>
<th></th>
<th>&quot;20140601&quot;</th>
<th>&quot;20140602&quot;</th>
<th>&quot;20140603&quot;</th>
<th>&quot;20140604&quot;</th>
<th>&quot;20140605&quot;</th>
<th>&quot;20140606&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>&quot;20140607&quot;</td>
<td>&quot;20140608&quot;</td>
<td>&quot;20140609&quot;</td>
<td>&quot;20140610&quot;</td>
<td>&quot;20140611&quot;</td>
<td>&quot;20140612&quot;</td>
</tr>
<tr>
<td>13</td>
<td>&quot;20140613&quot;</td>
<td>&quot;20140614&quot;</td>
<td>&quot;20140615&quot;</td>
<td>&quot;20140616&quot;</td>
<td>&quot;20140617&quot;</td>
<td>&quot;20140618&quot;</td>
</tr>
<tr>
<td>19</td>
<td>&quot;20140619&quot;</td>
<td>&quot;20140620&quot;</td>
<td>&quot;20140621&quot;</td>
<td>&quot;20140622&quot;</td>
<td>&quot;20140623&quot;</td>
<td>&quot;20140624&quot;</td>
</tr>
<tr>
<td>25</td>
<td>&quot;20140625&quot;</td>
<td>&quot;20140626&quot;</td>
<td>&quot;20140627&quot;</td>
<td>&quot;20140628&quot;</td>
<td>&quot;20140629&quot;</td>
<td>&quot;20140630&quot;</td>
</tr>
<tr>
<td>31</td>
<td>&quot;20140701&quot;</td>
<td>&quot;20140702&quot;</td>
<td>&quot;20140703&quot;</td>
<td>&quot;20140704&quot;</td>
<td>&quot;20140705&quot;</td>
<td>&quot;20140706&quot;</td>
</tr>
<tr>
<td>37</td>
<td>&quot;20140707&quot;</td>
<td>&quot;20140708&quot;</td>
<td>&quot;20140709&quot;</td>
<td>&quot;20140710&quot;</td>
<td>&quot;20140711&quot;</td>
<td>&quot;20140712&quot;</td>
</tr>
<tr>
<td>43</td>
<td>&quot;20140713&quot;</td>
<td>&quot;20140714&quot;</td>
<td>&quot;20140715&quot;</td>
<td>&quot;20140716&quot;</td>
<td>&quot;20140717&quot;</td>
<td>&quot;20140718&quot;</td>
</tr>
<tr>
<td>49</td>
<td>&quot;20140719&quot;</td>
<td>&quot;20140720&quot;</td>
<td>&quot;20140721&quot;</td>
<td>&quot;20140722&quot;</td>
<td>&quot;20140723&quot;</td>
<td>&quot;20140724&quot;</td>
</tr>
<tr>
<td>55</td>
<td>&quot;20140725&quot;</td>
<td>&quot;20140726&quot;</td>
<td>&quot;20140727&quot;</td>
<td>&quot;20140728&quot;</td>
<td>&quot;20140729&quot;</td>
<td>&quot;20140730&quot;</td>
</tr>
<tr>
<td>61</td>
<td>&quot;20140731&quot;</td>
<td>&quot;20140801&quot;</td>
<td>&quot;20140802&quot;</td>
<td>&quot;20140803&quot;</td>
<td>&quot;20140804&quot;</td>
<td>&quot;20140805&quot;</td>
</tr>
<tr>
<td>67</td>
<td>&quot;20140806&quot;</td>
<td>&quot;20140807&quot;</td>
<td>&quot;20140808&quot;</td>
<td>&quot;20140809&quot;</td>
<td>&quot;20140810&quot;</td>
<td>&quot;20140811&quot;</td>
</tr>
<tr>
<td>73</td>
<td>&quot;20140812&quot;</td>
<td>&quot;20140813&quot;</td>
<td>&quot;20140814&quot;</td>
<td>&quot;20140815&quot;</td>
<td>&quot;20140816&quot;</td>
<td>&quot;20140817&quot;</td>
</tr>
<tr>
<td>79</td>
<td>&quot;20140818&quot;</td>
<td>&quot;20140819&quot;</td>
<td>&quot;20140820&quot;</td>
<td>&quot;20140821&quot;</td>
<td>&quot;20140822&quot;</td>
<td>&quot;20140823&quot;</td>
</tr>
<tr>
<td>85</td>
<td>&quot;20140824&quot;</td>
<td>&quot;20140825&quot;</td>
<td>&quot;20140826&quot;</td>
<td>&quot;20140827&quot;</td>
<td>&quot;20140828&quot;</td>
<td>&quot;20140829&quot;</td>
</tr>
<tr>
<td>91</td>
<td>&quot;20140830&quot;</td>
<td>&quot;20140831&quot;</td>
<td>&quot;20140901&quot;</td>
<td>&quot;20140902&quot;</td>
<td>&quot;20140903&quot;</td>
<td>&quot;20140904&quot;</td>
</tr>
<tr>
<td>97</td>
<td>&quot;20140905&quot;</td>
<td>&quot;20140906&quot;</td>
<td>&quot;20140907&quot;</td>
<td>&quot;20140908&quot;</td>
<td>&quot;20140909&quot;</td>
<td>&quot;20140910&quot;</td>
</tr>
<tr>
<td></td>
<td>&quot;20150403&quot;</td>
<td>&quot;20150404&quot;</td>
<td>&quot;20150405&quot;</td>
<td>&quot;20150406&quot;</td>
<td>&quot;20150407&quot;</td>
<td>&quot;20150408&quot;</td>
</tr>
<tr>
<td>313</td>
<td>&quot;20150409&quot;</td>
<td>&quot;20150410&quot;</td>
<td>&quot;20150411&quot;</td>
<td>&quot;20150412&quot;</td>
<td>&quot;20150413&quot;</td>
<td>&quot;20150414&quot;</td>
</tr>
<tr>
<td>319</td>
<td>&quot;20150415&quot;</td>
<td>&quot;20150416&quot;</td>
<td>&quot;20150417&quot;</td>
<td>&quot;20150418&quot;</td>
<td>&quot;20150419&quot;</td>
<td>&quot;20150420&quot;</td>
</tr>
<tr>
<td>325</td>
<td>&quot;20150421&quot;</td>
<td>&quot;20150422&quot;</td>
<td>&quot;20150423&quot;</td>
<td>&quot;20150424&quot;</td>
<td>&quot;20150425&quot;</td>
<td>&quot;20150426&quot;</td>
</tr>
<tr>
<td>331</td>
<td>&quot;20150427&quot;</td>
<td>&quot;20150428&quot;</td>
<td>&quot;20150429&quot;</td>
<td>&quot;20150430&quot;</td>
<td>&quot;20150501&quot;</td>
<td>&quot;20150502&quot;</td>
</tr>
<tr>
<td>337</td>
<td>&quot;20150503&quot;</td>
<td>&quot;20150504&quot;</td>
<td>&quot;20150505&quot;</td>
<td>&quot;20150506&quot;</td>
<td>&quot;20150507&quot;</td>
<td>&quot;20150508&quot;</td>
</tr>
<tr>
<td>343</td>
<td>&quot;20150509&quot;</td>
<td>&quot;20150510&quot;</td>
<td>&quot;20150511&quot;</td>
<td>&quot;20150512&quot;</td>
<td>&quot;20150513&quot;</td>
<td>&quot;20150514&quot;</td>
</tr>
<tr>
<td>349</td>
<td>&quot;20150515&quot;</td>
<td>&quot;20150516&quot;</td>
<td>&quot;20150517&quot;</td>
<td>&quot;20150518&quot;</td>
<td>&quot;20150519&quot;</td>
<td>&quot;20150520&quot;</td>
</tr>
<tr>
<td>355</td>
<td>&quot;20150521&quot;</td>
<td>&quot;20150522&quot;</td>
<td>&quot;20150523&quot;</td>
<td>&quot;20150524&quot;</td>
<td>&quot;20150525&quot;</td>
<td>&quot;20150526&quot;</td>
</tr>
<tr>
<td>361</td>
<td>&quot;20150527&quot;</td>
<td>&quot;20150528&quot;</td>
<td>&quot;20150529&quot;</td>
<td>&quot;20150530&quot;</td>
<td>&quot;20150531&quot;</td>
<td></td>
</tr>
</tbody>
</table>
Using TIQ-TEST – Feeds Selected

- Dataset was separated into “inbound” and “outbound”

TY to @kafeine and John Bambenek for access to their feeds
Using TIQ-TEST – Data Prep

- Extract the “raw” information from indicator feeds
- Both IP addresses and hostnames were extracted

```python
outbound.ti = tiq.data.loadTI("raw", "public_outbound", "20150501")
outbound.ti[, list(entity, type, direction, source, date)]
```

<table>
<thead>
<tr>
<th>#</th>
<th>entity</th>
<th>type</th>
<th>direction</th>
<th>source</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>103.18.247.72</td>
<td>IPv4</td>
<td>outbound</td>
<td>alienvault</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>2</td>
<td>103.253.41.10</td>
<td>IPv4</td>
<td>outbound</td>
<td>alienvault</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>3</td>
<td>103.6.196.92</td>
<td>IPv4</td>
<td>outbound</td>
<td>alienvault</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>4</td>
<td>103.6.198.12</td>
<td>IPv4</td>
<td>outbound</td>
<td>alienvault</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>5</td>
<td>103.9.103.141</td>
<td>IPv4</td>
<td>outbound</td>
<td>alienvault</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>145195</td>
<td>winscoft.com</td>
<td>FQDN</td>
<td>outbound</td>
<td>zeus</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>145196</td>
<td>worldrecipeblogs.com</td>
<td>FQDN</td>
<td>outbound</td>
<td>zeus</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>145197</td>
<td><a href="http://www.nikey.cn">www.nikey.cn</a></td>
<td>FQDN</td>
<td>outbound</td>
<td>zeus</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>145198</td>
<td><a href="http://www.riverwalktrader.co.za">www.riverwalktrader.co.za</a></td>
<td>FQDN</td>
<td>outbound</td>
<td>zeus</td>
<td>2015-05-01</td>
</tr>
<tr>
<td>145199</td>
<td>zetes.vdsinside.com</td>
<td>FQDN</td>
<td>outbound</td>
<td>zeus</td>
<td>2015-05-01</td>
</tr>
</tbody>
</table>
Using TIQ-TEST – Data Prep

- Convert the hostname data to IP addresses:
  - Active IP addresses for the respective date (“A” query)
  - Passive DNS from Farsight Security (DNSDB)

- For each IP record (including the ones from hostnames):
  - Add `asnumber` and `asname` (from MaxMind ASN DB)
  - Add `country` (from MaxMind GeoLite DB)
  - Add `rhost` (again from DNSDB) – most popular “PTR”
enrich.ti = tiq.data.loadTI("enriched", "public_outbound", "20150501")
enrich.ti = enrich.ti[, notes := NULL]
tail(enrich.ti)
Novelty Test
Measuring added and dropped indicators
Aging Test

Is anyone cleaning this mess up eventually?
Population Test

• Let us use the ASN and GeoIP databases that we used to enrich our data as a reference of the “true” population.

• But, but, human beings are unpredictable! We will never be able to forecast this! 😁

PET PEEVE #208: GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS.
Is your sampling poll as random as you think?
Can we get a better look?

• Statistical inference-based comparison models (hypothesis testing)
  • Exact binomial tests (when we have the “true” pop)
  • Chi-squared proportion tests (similar to independence tests)
outbound.pop = tiq.test.extractPopulationFromTI("public_outbound", "country", 
   date = "20150501", 
   select.sources=NULL, 
   split.ti=FALSE)

complete.pop = tiq.data.loadPopulation("mmgeo", "country")
tests = tiq.test.populationInference(complete.pop$mmgeo, 
   outbound.pop$public_outbound, "country", 
   exact = TRUE, top=10)

## country     conf.int.start conf.int.end       p.value  
## 1:   US  0.084870546   0.09783018   2.384509e-169
## 2:   RU  0.026186375   0.03139187   6.353991e-208

## 1:   US  0.084870546   0.09783018   2.384509e-169
## 2:   RU  0.026186375   0.03139187   6.353991e-208
## 3:   NL  0.023978511   0.02910542   5.195447e-173

## country     conf.int.start conf.int.end       p.value  
## 1:   CN  -0.035268623  -0.029053639   3.245893e-71
## 2:   CA  -0.010799505  -0.007832391   2.723407e-25

## country     conf.int.start conf.int.end       p.value  
## 1:   CN  -0.035268623  -0.029053639   3.245893e-71
## 2:   CA  -0.010799505  -0.007832391   2.723407e-25

## country     conf.int.start conf.int.end       p.value  
## 1:   DE  -0.001333158   0.003429626   0.3980818

tests[p.value > 0.05/10]

## country     conf.int.start conf.int.end       p.value  
## 1:   DE  -0.001333158   0.003429626   0.3980818
Overlap Test

More data can be better, but make sure it is not the same data
Overlap Test - Outbound Data - 20150501

- public_outbound.zeus -
- public_outbound.sslbl -
- public_outbound.phishtank -
- public_outbound.palevotrack -
- public_outbound.opentracker -
- public_outbound.malwarephиш -
- public_outbound.malwaregroup -
- public_outbound.malwaredomains -
- public_outbound.malwaredomainlist -
- public_outbound.malware -
- public_outbound.malcode -
- public_outbound.kafeine -
- public_outbound.feodo -
- public_outbound.et_shadowserver_cnc -
- public_outbound.bambenek -
- public_outbound.alienvault -

Source (contains)

Source (is contained)

% Overlap
- 0%
- 25%
- 50%
- 75%
- 100%
Uniqueness Test
Uniqueness Test

- “Domain-based indicators are unique to one list between 96.16% and 97.37%”
- “IP-based indicators are unique to one list between 82.46% and 95.24% of the time”
I hate quoting myself, but...

It is hard to draw a positive conclusion from these metrics, and it seems to suggest that if threat intelligence indicators were really able to help an enterprise defense strategy, one would need to have access to **all of the feeds from all of the providers** to be able to get the “best” possible coverage. This would be a Herculean task for any organization, and given the results of our analysis, the result would still be **incomplete intelligence**. There is a need for companies to be able to apply their threat intelligence to their environment in smarter ways so that even if we cannot see inside the whole lake, we can forecast which parts of it are more likely to have a lot of fish we still haven’t caught.
Key Takeaway #1

MORE ! = BETTER

Threat Intelligence Indicator Feeds

Threat Intelligence Program
try some delicious and healthy snacks at our concession stand

Intermission
Key Takeaway #2
"These are the problems Threat Intelligence Sharing is here to solve!"

Right?
Herd Immunity...

... would imply that other people in your sharing community being immune to malware A meant your likelihood of infection from it was negligible regardless of controls you applied.
Threat Intelligence Sharing

- How many indicators are being shared?
- How many members do actually share and how many just leech?
- Can we measure that? What a super-deeeee-duper idea!
Threat Intelligence Sharing

We would like to thank the kind contribution of data from the fine folks at Facebook Threat Exchange and Threat Connect...

... and also the sharing communities that chose to remain anonymous. You know who you are, and we ❤️ you too.
Threat Intelligence Sharing – Data

From a period of 2015-03-01 to 2015-05-31:
- Number of Indicators Shared
  - Per day
  - Per member

Not sharing this data – privacy concerns for the members and communities
You're a bigger deal on Twitter than you think.
same
MATURITY?
“Reddit of Threat Intelligence”? 
Threat intelligence has become a booming area of information security, and with good reason. Attackers have the luxury of exploiting whichever weaknesses in a target best serve their intent. Defenders, on the other hand, must make the most of limited resources to defend all the most vulnerable aspects of critical information assets. Understanding the nature of current threats and adversary intent is essential to knowing how and where to place the most effective bets on defense.
'How can sharing make me better understand what are attacks that “are targeted” and what are “commodity”?'
Key Takeaway #3
(Also Prediction #1)

TELEMETRY > CONTENT
More Takeaways (I lied)

• Analyze your data. Extract more value from it!
• If you ABSOLUTELY HAVE TO buy Threat Intelligence or data, evaluate it first.

• Try the sample data, replicate the experiments:
  • https://github.com/mlsecproject/tiq-test-Summer2015
  • http://rpubs.com/alexcpsec/tiq-test-Summer2015

• Share data with us. I’ll make sure it gets proper exercise!
Your gift of a few contributions can help a starving data scientist.
"The measure of intelligence is the ability to change."

- Albert Einstein