NIDDEL™

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

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

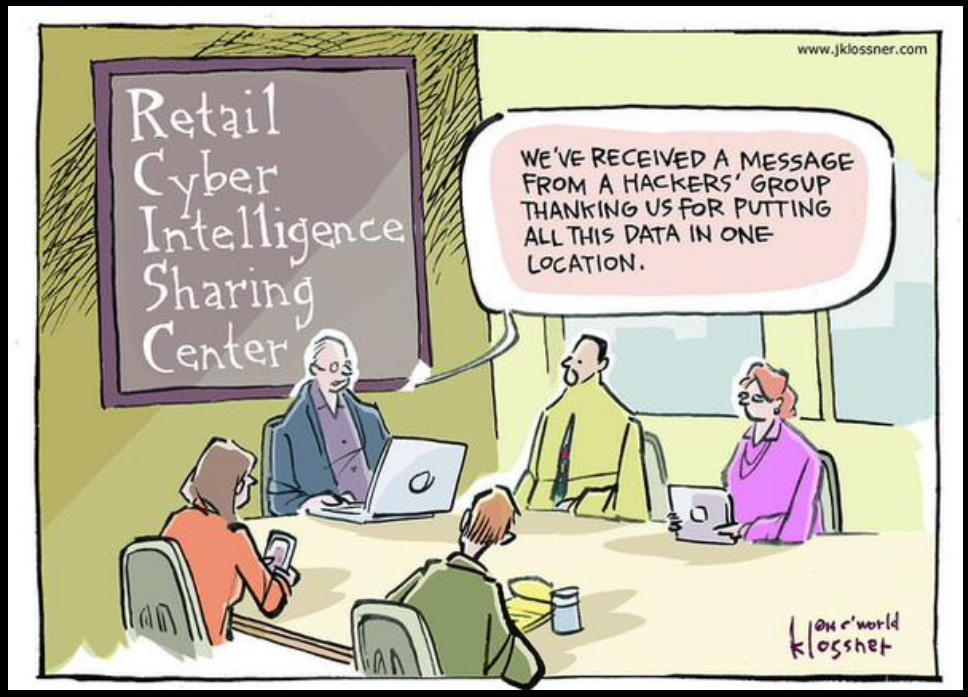
Alexandre Sieira
CTO
Niddel
@AlexandreSieira
@NiddelCorp

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 @RCISCwendy

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.attributed.to

TLP: White

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 diclosure 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 responsiblity 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.

TY to @bfist for his work on http://sony.attributed.to

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)
  - 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)
  - 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)

- https://github.com/mlsecproject/tiq-test-Summer2015

```
print(tiq.data.getAvailableDates("raw", "public_outbound"))
```

```
##   [1] "20140601" "20140602" "20140603" "20140604" "20140605" "20140606"
##   [7] "20140607" "20140608" "20140609" "20140610" "20140611" "20140612"
##  [13] "20140613" "20140614" "20140615" "20140616" "20140617" "20140618"
##  [19] "20140619" "20140620" "20140621" "20140622" "20140623" "20140624"
##  [25] "20140625" "20140626" "20140627" "20140628" "20140629" "20140630"
##  [31] "20140701" "20140702" "20140703" "20140704" "20140705" "20140706"
##  [37] "20140707" "20140708" "20140709" "20140710" "20140711" "20140712"
##  [43] "20140713" "20140714" "20140715" "20140716" "20140717" "20140718"
##  [49] "20140719" "20140720" "20140721" "20140722" "20140723" "20140724"
##  [55] "20140725" "20140726" "20140727" "20140728" "20140729" "20140730"
##  [61] "20140731" "20140801" "20140802" "20140803" "20140804" "20140805"
##  [67] "20140806" "20140807" "20140808" "20140809" "20140810" "20140811"
##  [73] "20140812" "20140813" "20140814" "20140815" "20140816" "20140817"
##  [79] "20140818" "20140819" "20140820" "20140821" "20140822" "20140823"
##  [85] "20140824" "20140825" "20140826" "20140827" "20140828" "20140829"
##  [91] "20140830" "20140831" "20140901" "20140902" "20140903" "20140904"
##  [97] "20140905" "20140906" "20140907" "20140908" "20140909" "20140910"
```

```
## [307] "20150403" "20150404" "20150405" "20150406" "20150407" "20150408"
## [313] "20150409" "20150410" "20150411" "20150412" "20150413" "20150414"
## [319] "20150415" "20150416" "20150417" "20150418" "20150419" "20150420"
## [325] "20150421" "20150422" "20150423" "20150424" "20150425" "20150426"
## [331] "20150427" "20150428" "20150429" "20150430" "20150501" "20150502"
## [337] "20150503" "20150504" "20150505" "20150506" "20150507" "20150508"
## [343] "20150509" "20150510" "20150511" "20150512" "20150513" "20150514"
## [349] "20150515" "20150516" "20150517" "20150518" "20150519" "20150520"
## [355] "20150521" "20150522" "20150523" "20150524" "20150525" "20150526"
## [361] "20150527" "20150528" "20150529" "20150530" "20150531"
```


2004

Using TIQ-TEST – Feeds Selected

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

```
outbound.ti = tiq.data.loadTI("raw", "public_outbound", "20150501")
unique(outbound.ti$source)
```

```
##  [1] "alienvault"          "bambenek"            "et_shadowserver_cnc"
##  [4] "feodo"               "kafeine"             "malcode"
##  [7] "malwared"            "malwaredomainlist"   "malwaredomains"
## [10] "malwaregroup"        "openphish"           "palevotracker"
## [13] "phishtank"           "sslbl"               "zeus"
```

We can do the same for the inbound data we have to see the sources we have available:

```
inbound.ti = tiq.data.loadTI("raw", "public_inbound", "20150501")
unique(inbound.ti$source)
```

```
##  [1] "alienvault"        "autoshun"          "blocklistde"
##  [4] "botscout"          "bruteforceblocker" "charleshaley"
##  [7] "ciarmy"            "dragonresearch"    "dshield"
## [10] "honeypot"          "openbl"            "packetmail"
## [13] "virbl"
```

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

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

```
##                         entity type direction     source       date
##      1:          103.18.247.72 IPv4  outbound alienvault 2015-05-01
##      2:         103.253.41.10 IPv4  outbound alienvault 2015-05-01
##      3:          103.6.196.92 IPv4  outbound alienvault 2015-05-01
##      4:          103.6.198.12 IPv4  outbound alienvault 2015-05-01
##      5:         103.9.103.141 IPv4  outbound alienvault 2015-05-01
##     ---
## 145195:          winscoft.com FQDN  outbound       zeus 2015-05-01
## 145196:  worldrecipeblogs.com FQDN  outbound       zeus 2015-05-01
## 145197:          www.nikey.cn FQDN  outbound       zeus 2015-05-01
## 145198: www.riverwalktrader.co.za FQDN  outbound       zeus 2015-05-01
## 145199:   zetes.vdsinside.com FQDN  outbound       zeus 2015-05-01
```

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)

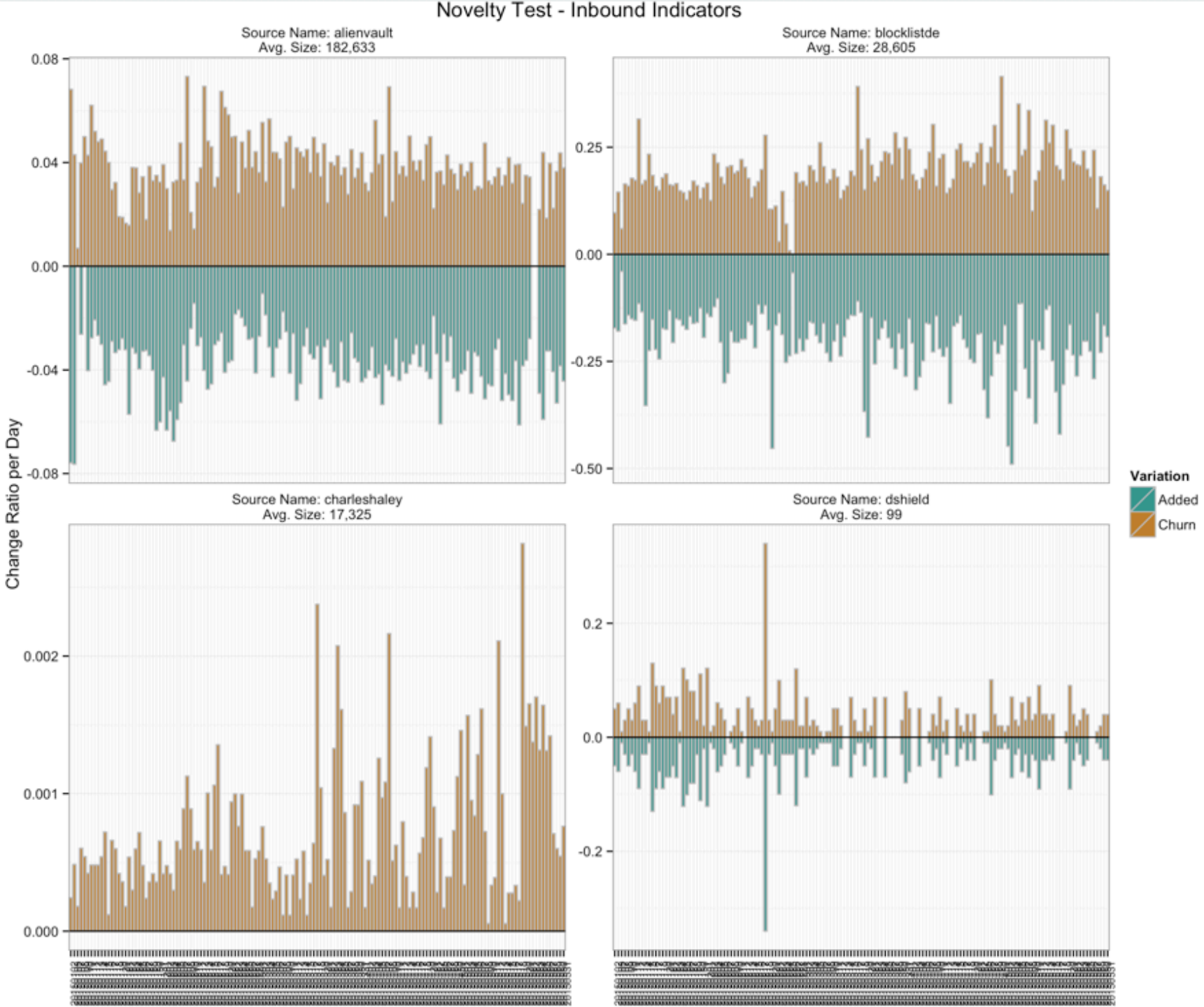
- For each IP record (including the ones from hostnames):
  - Add asnumber and asname (from MaxMind ASN DB)
  - Add country (from MaxMind GeoLite DB)
  - Add rhost (again from DNSDB) – most popular “PTR”

```
enrich.ti = tiq.data.loadTI("enriched", "public_outbound", "20150501")
enrich.ti = enrich.ti[, notes := NULL]
tail(enrich.ti)
```

```
##             entity type direction source       date asnumber
## 1:   94.76.211.87 IPv4  outbound   zeus 2015-05-01    29550
## 2: 95.211.243.120 IPv4  outbound   zeus 2015-05-01    60781
## 3: 95.211.243.123 IPv4  outbound   zeus 2015-05-01    60781
## 4: 95.211.243.125 IPv4  outbound   zeus 2015-05-01    60781
## 5: 98.131.185.136 IPv4  outbound   zeus 2015-05-01    32392
## 6: 98.131.185.136 IPv4  outbound   zeus 2015-05-01    32392
##                  asname country                          host
## 1:    Simply Transit Ltd      GB                            NA
## 2:         LeaseWeb B.V.      NL                            NA
## 3:         LeaseWeb B.V.      NL                            NA
## 4:         LeaseWeb B.V.      NL                            NA
## 5: Ecommerce Corporation      US                            NA
## 6: Ecommerce Corporation      US projects.globaltronics.net
##                              rhost
## 1: 94-76-211-87.static.as29550.net
## 2:                              NA
## 3:                              NA
## 4:                              NA
## 5:                              NA
## 6:                              NA
```

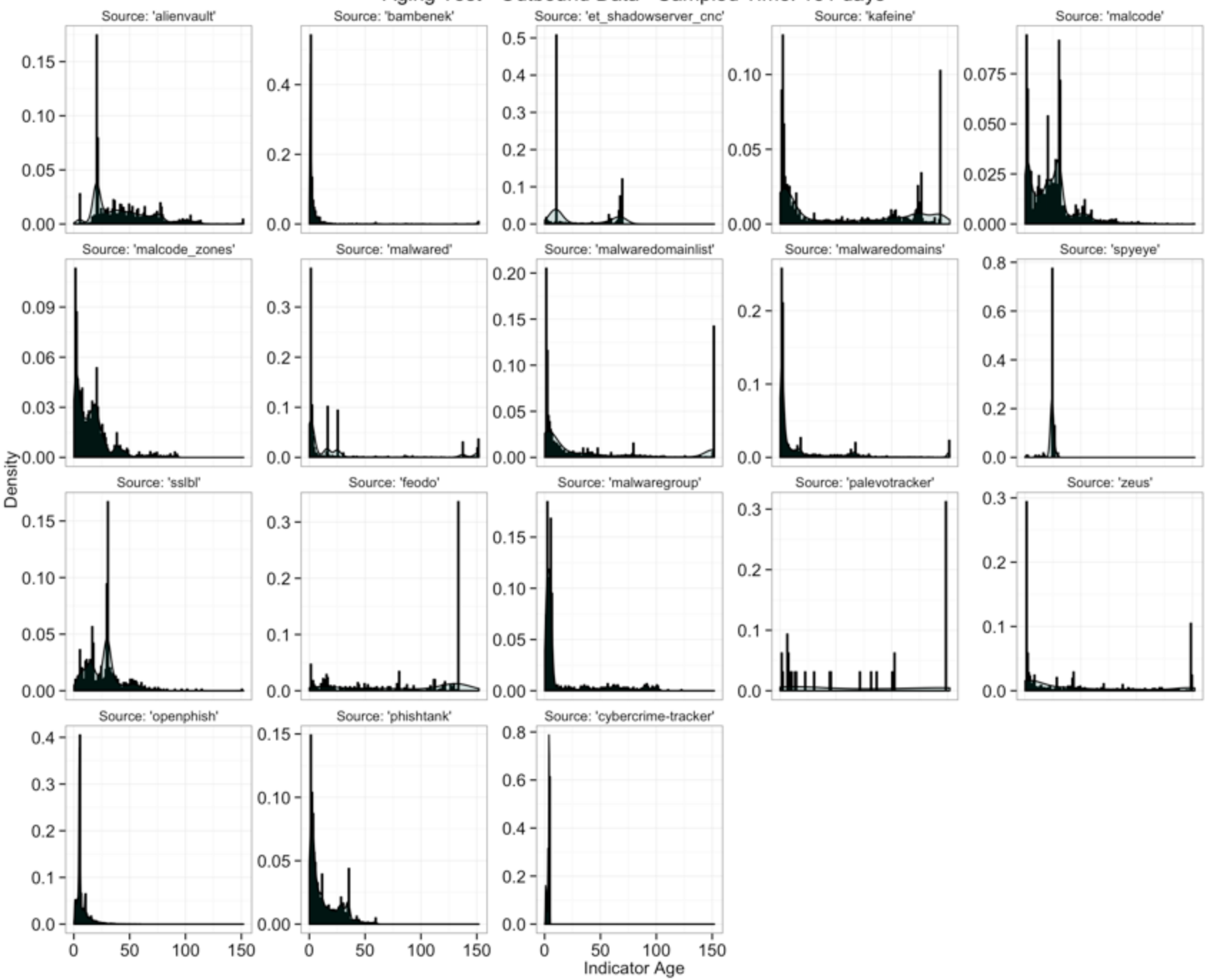
Novelty Test

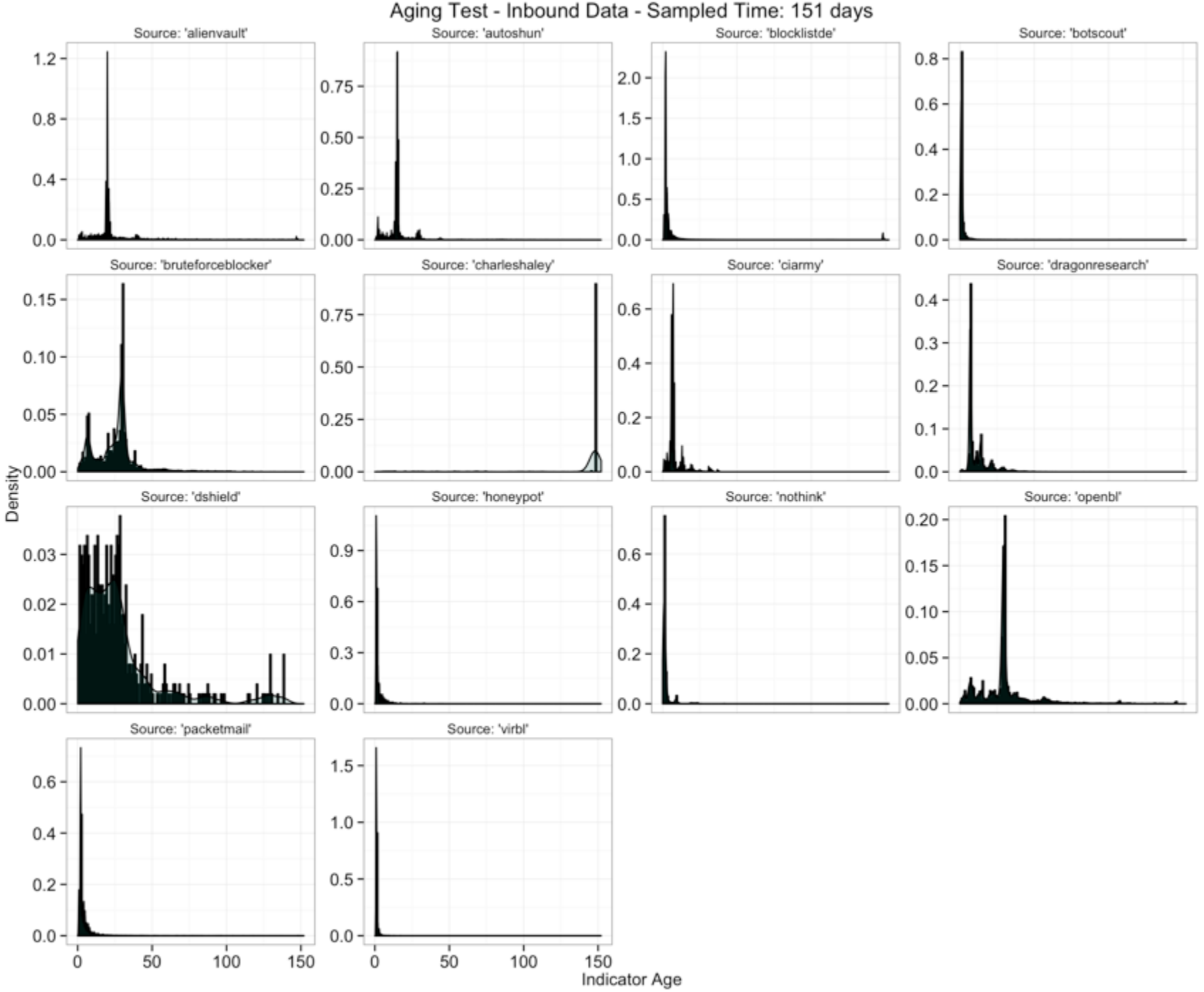
Measuring added and dropped indicators


Novelty Test - Inbound Indicators
Source Name: alienvault
Avg. Size: 182,633
Source Name: blocklistde
Avg. Size: 28,605
Source Name: charleshaley
Avg. Size: 17,325
Source Name: dshield
Avg. Size: 99
Change Ratio per Day
0.08
0.04
0.00
-0.04
-0.08
0.25
0.00
-0.25
-0.50
0.002
0.001
0.000
0.2
0.0
-0.2
Variation
Added
Churn

Aging Test

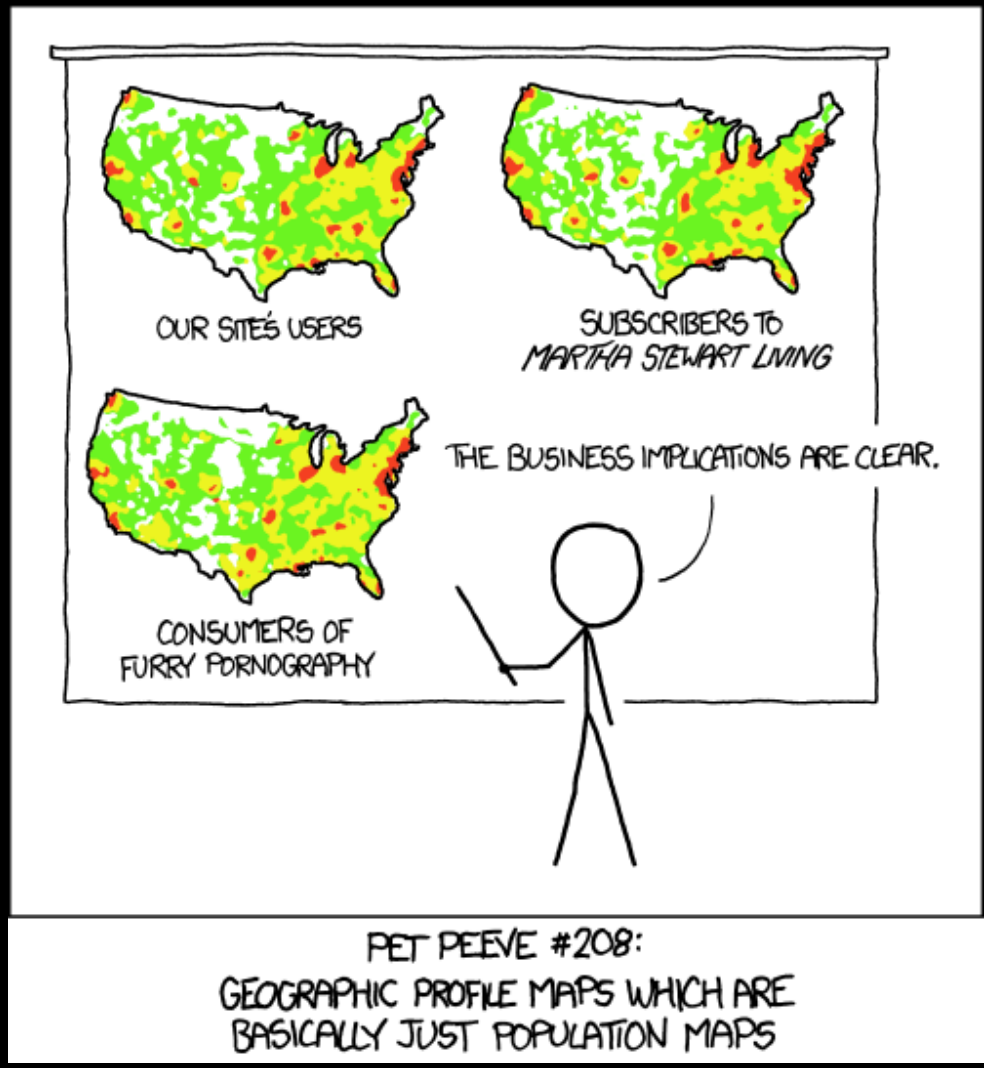
Is anyone cleaning this mess up eventually?

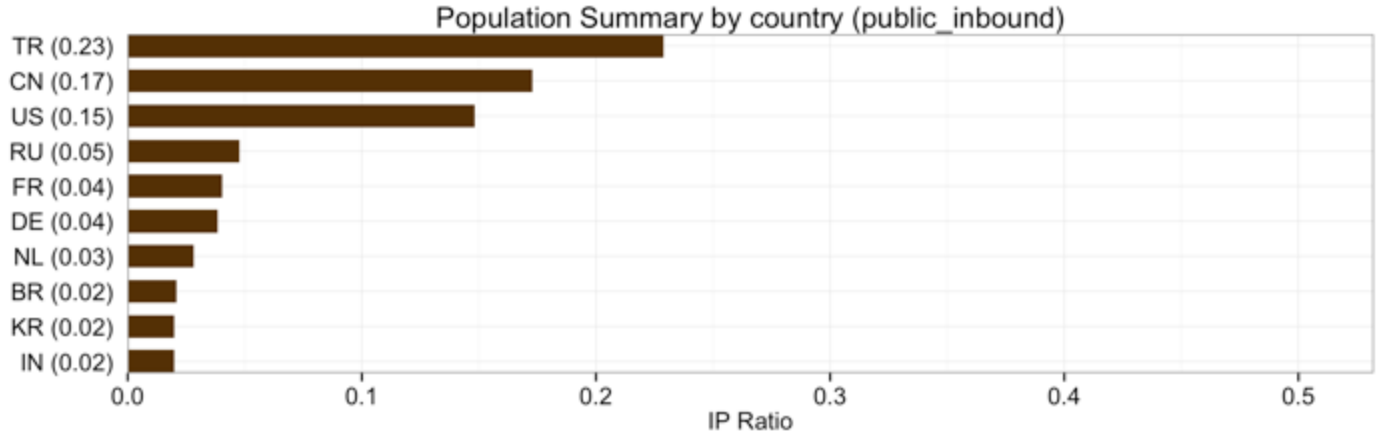

Aging Test - Outbound Data - Sampled Time: 151 days
Source: 'alienvault'
Source: 'bambenek'
Source: 'et_shadowserver_cnc'
Source: 'kafeine'
Source: 'malcode'
Source: 'malcode_zones'
Source: 'malwared'
Source: 'malwaredomainlist'
Source: 'malwaredomains'
Source: 'spyeye'
Source: 'sslbl'
Source: 'feodo'
Source: 'malwaregroup'
Source: 'palevotracker'
Source: 'zeus'
Source: 'openphish'
Source: 'phishtank'
Source: 'cybercrime-tracker'
Density
Indicator Age

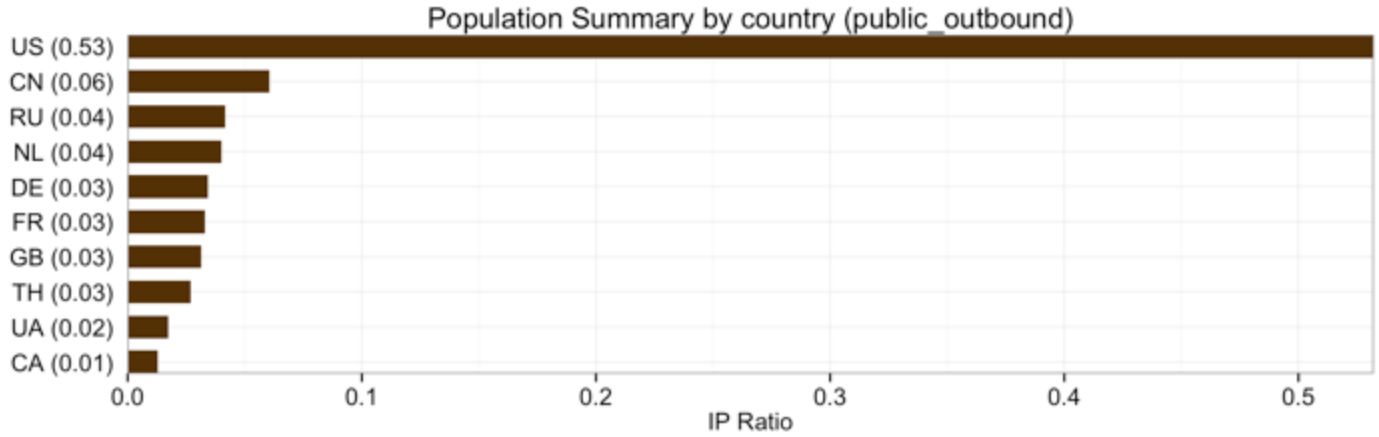

Aging Test - Inbound Data - Sampled Time: 151 days
Source: 'alienvault'
Source: 'autoshun'
Source: 'blocklistde'
Source: 'botscout'
Source: 'bruteforceblocker'
Source: 'charleshaley'
Source: 'ciarmy'
Source: 'dragonresearch'
Source: 'dshield'
Source: 'honeypot'
Source: 'nothink'
Source: 'openbl'
Source: 'packetmail'
Source: 'virbl'
Density
Indicator Age

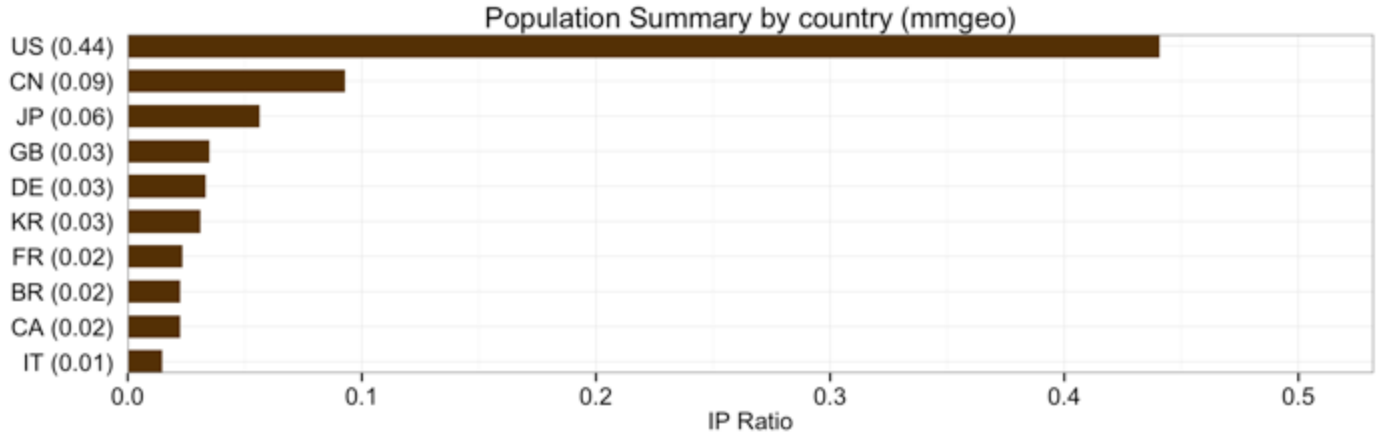
Population Test

- 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!




Population Summary by country (public_inbound)
TR (0.23)
CN (0.17)
US (0.15)
RU (0.05)
FR (0.04)
DE (0.04)
NL (0.03)
BR (0.02)
KR (0.02)
IN (0.02)
0.0
0.1
0.2
0.3
0.4
0.5
IP Ratio


Population Summary by country (public_outbound)
US (0.53)
CN (0.06)
RU (0.04)
NL (0.04)
DE (0.03)
FR (0.03)
GB (0.03)
TH (0.03)
UA (0.02)
CA (0.01)
0.0
0.1
0.2
0.3
0.4
0.5
IP Ratio


Population Summary by country (mmgeo)
US (0.44)
CN (0.09)
JP (0.06)
GB (0.03)
DE (0.03)
KR (0.03)
FR (0.02)
BR (0.02)
CA (0.02)
IT (0.01)
0.0
0.1
0.2
0.3
0.4
0.5
IP Ratio

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
```

```
##     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
## 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
```

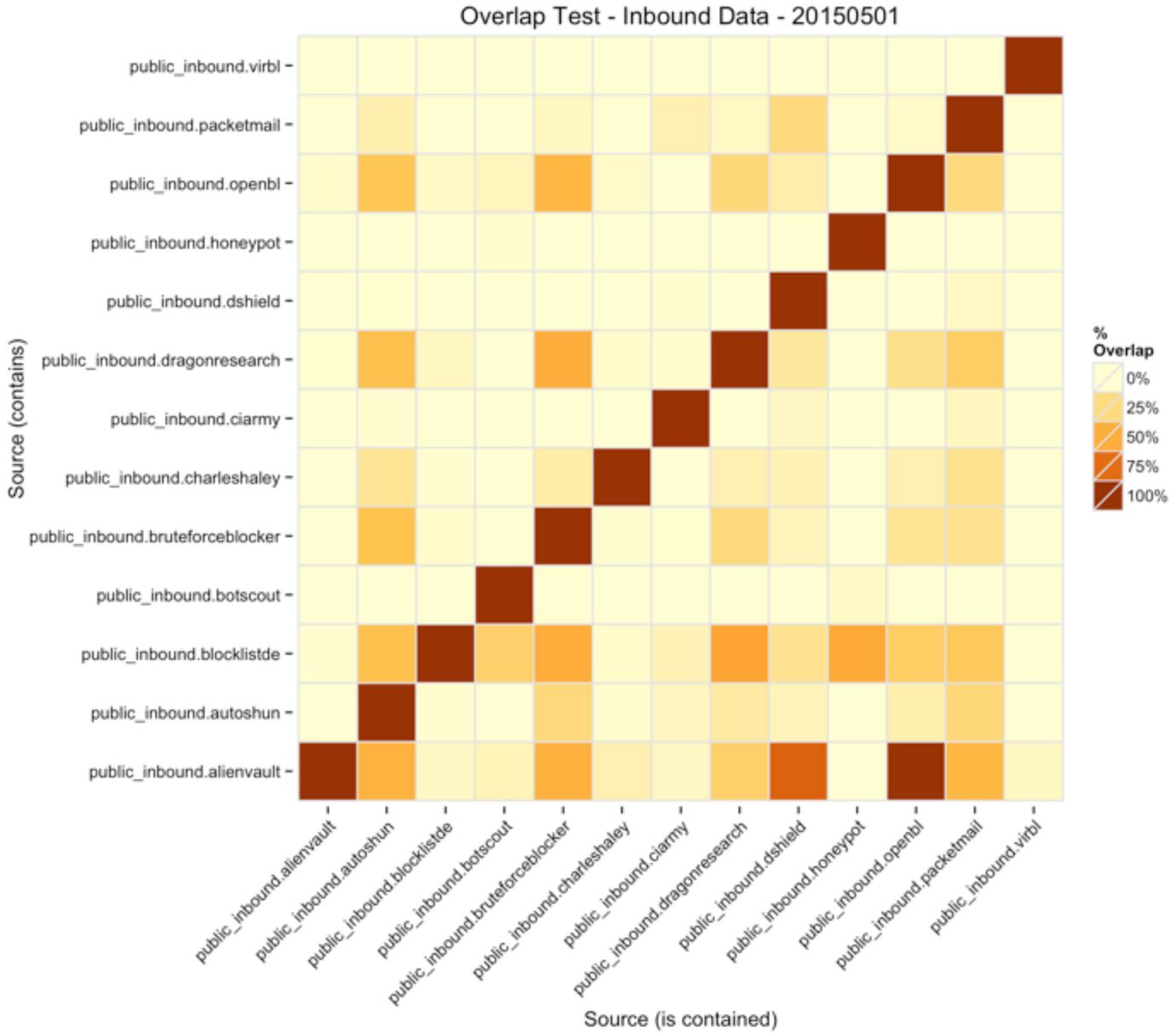
```
##     country conf.int.start conf.int.end     p.value
## 1:       DE   -0.001333158  0.003429626 0.3980818
```

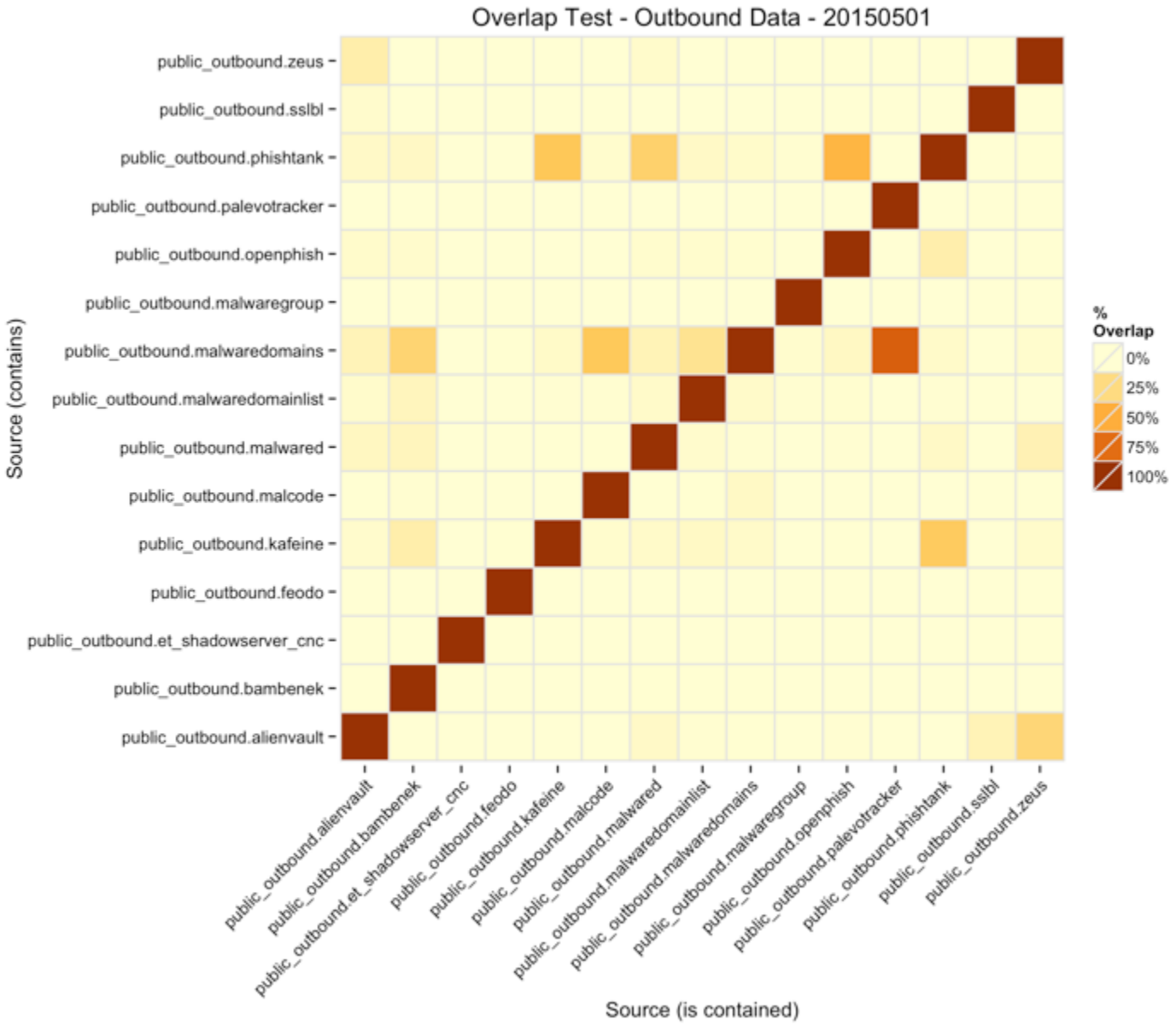
```
tests[p.value > 0.05/10]
```

```
##     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

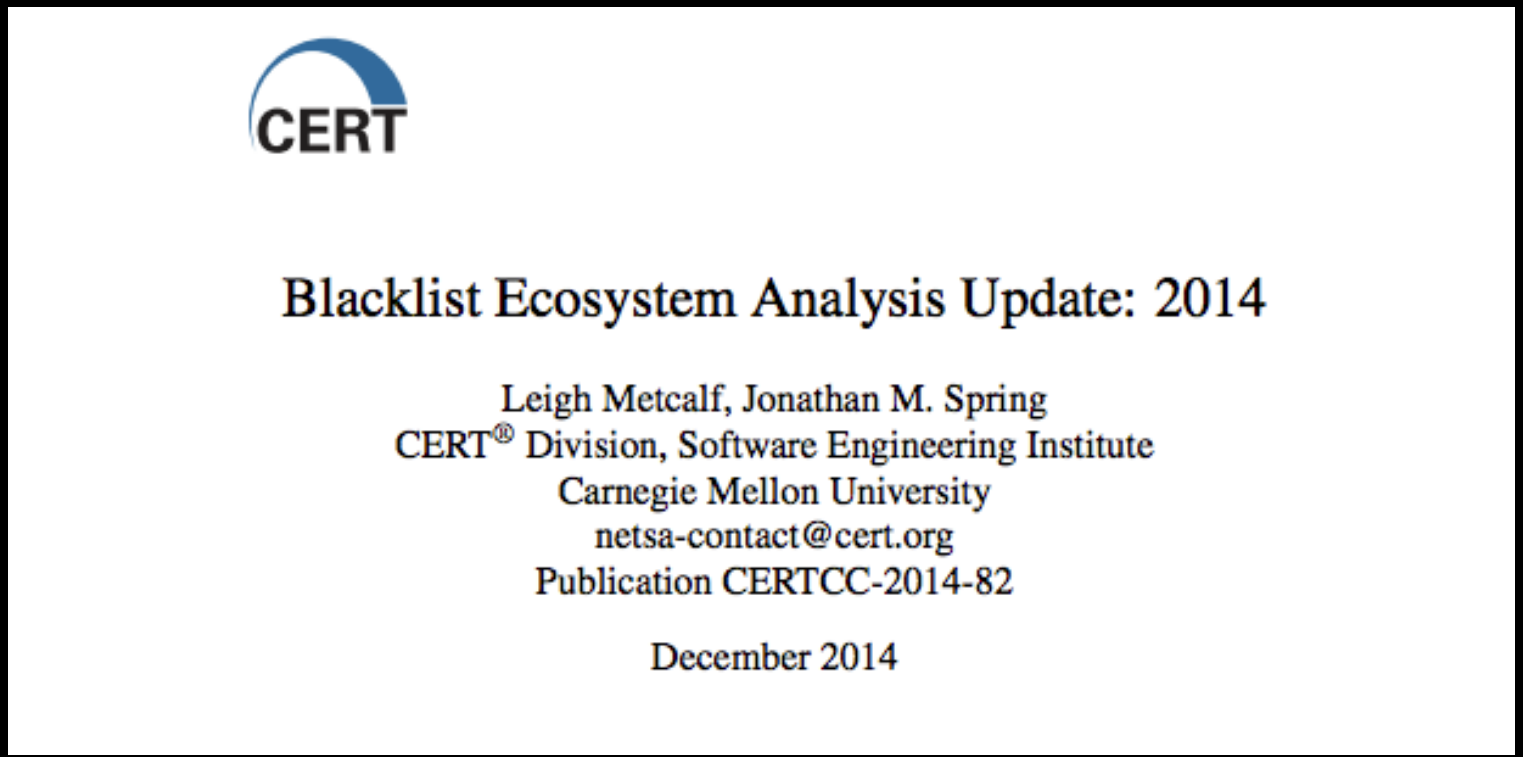

Overlap Test - Inbound Data - 20150501
Source (contains)
public_inbound.virbl
public_inbound.packetmail
public_inbound.openbl
public_inbound.honeypot
public_inbound.dshield
public_inbound.dragonresearch
public_inbound.ciarmy
public_inbound.charleshaley
public_inbound.bruteforceblocker
public_inbound.botscout
public_inbound.blocklistde
public_inbound.autoshun
public_inbound.alienvault
Source (is contained)
% Overlap
0%
25%
50%
75%
100%

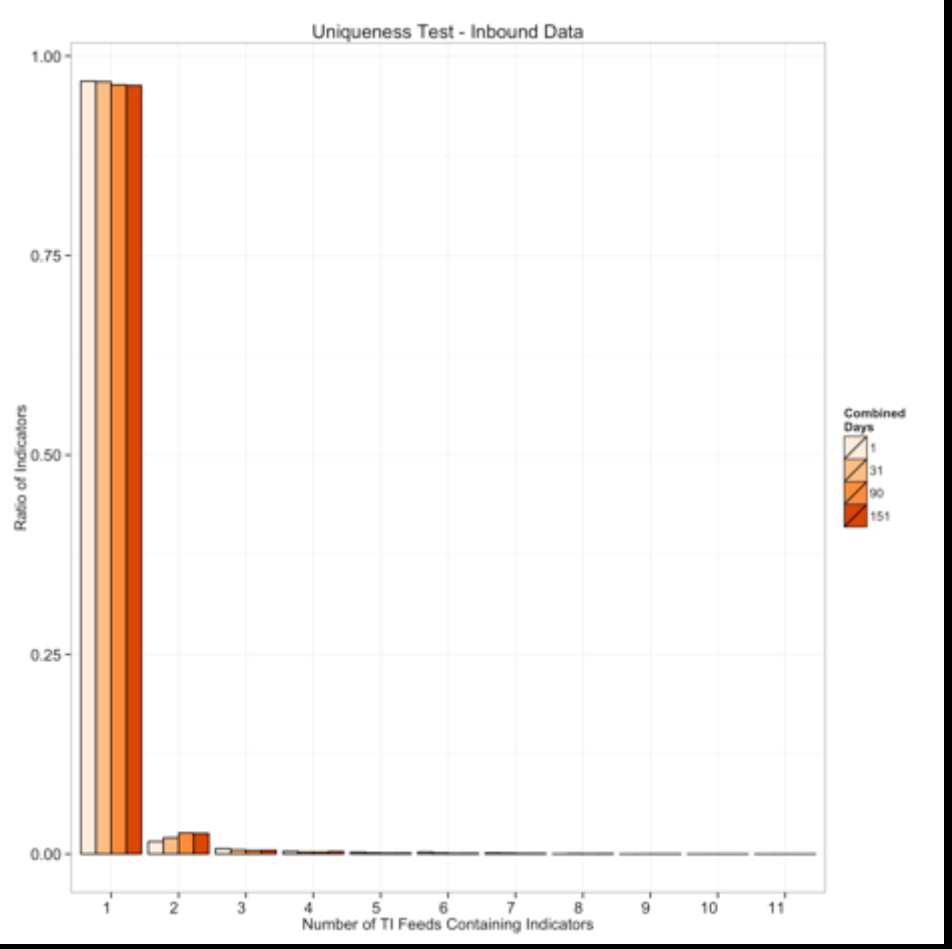

Overlap Test - Outbound Data - 20150501
Source (contains)
public_outbound.zeus
public_outbound.sslbl
public_outbound.phishtank
public_outbound.palevotracker
public_outbound.openphish
public_outbound.malwaregroup
public_outbound.malwaredomains
public_outbound.malwaredomainlist
public_outbound.malwared
public_outbound.malcode
public_outbound.kafeine
public_outbound.feodo
public_outbound.et_shadowserver_cnc
public_outbound.bambenek
public_outbound.alienvault
Source (is contained)
% Overlap
0%
25%
50%
75%
100%

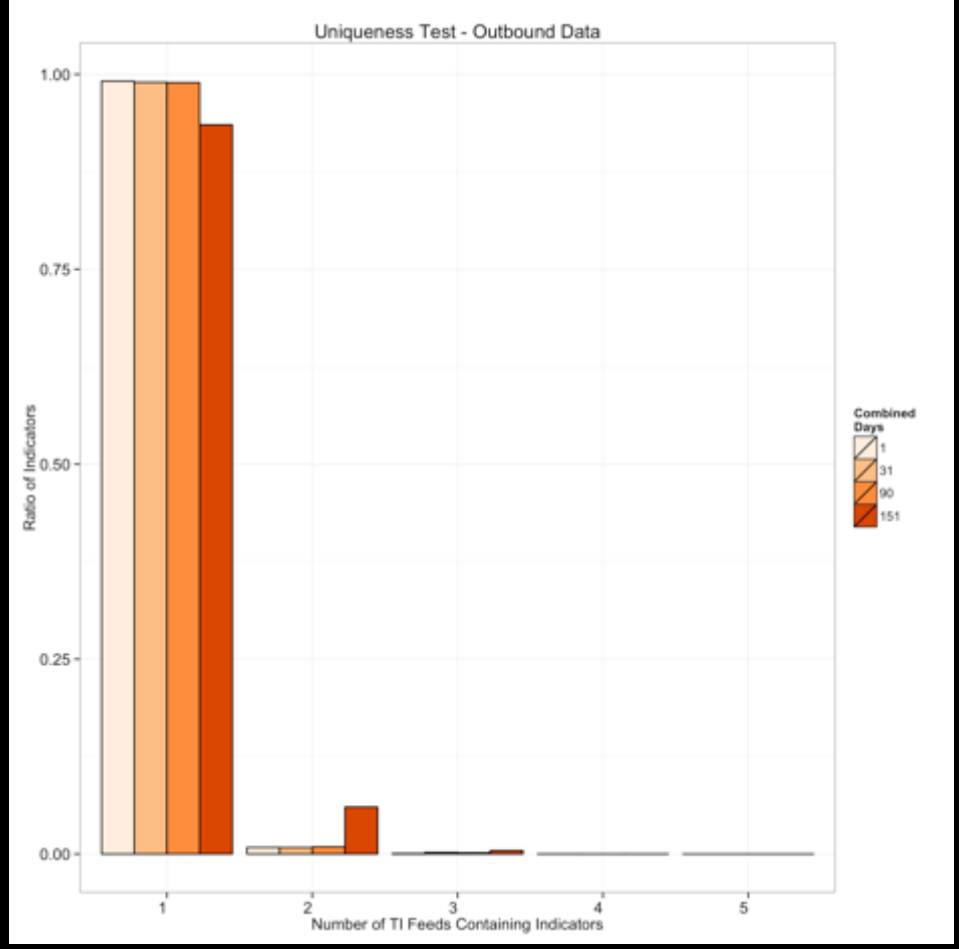
Uniqueness Test

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”








```
##     count     ratio days
## 1:      1 0.9684775    1
## 2:      1 0.9678683   31
## 3:      1 0.9639037   90
## 4:      1 0.9631740  151
```

```
##     count     ratio days
## 1:      1 0.9912258    1
## 2:      1 0.9898420   31
## 3:      1 0.9893606   90
## 4:      1 0.9352627  151
```

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

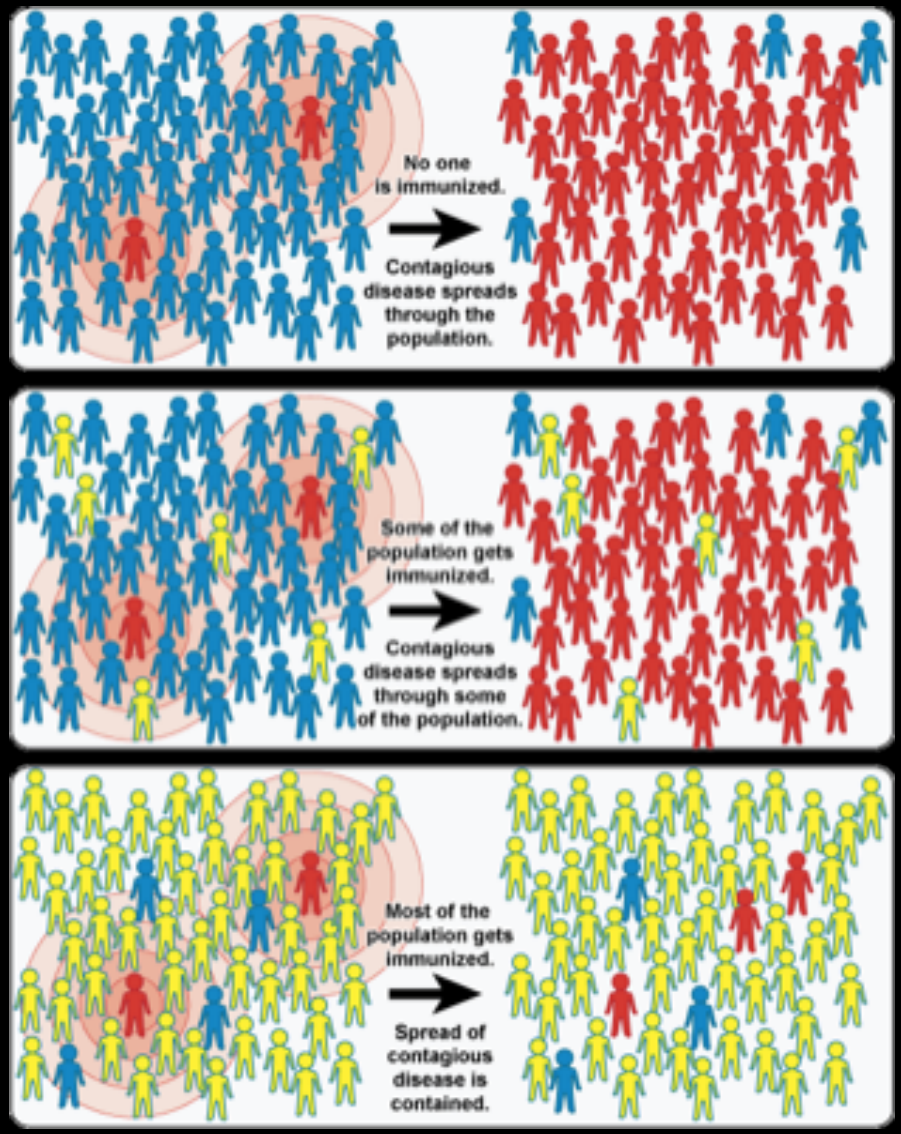

gifbin.com

Key Takeaway #2

"These are the problems Threat Intelligence Sharing is here to solve!”

Right?

Herd Immunity, is it?



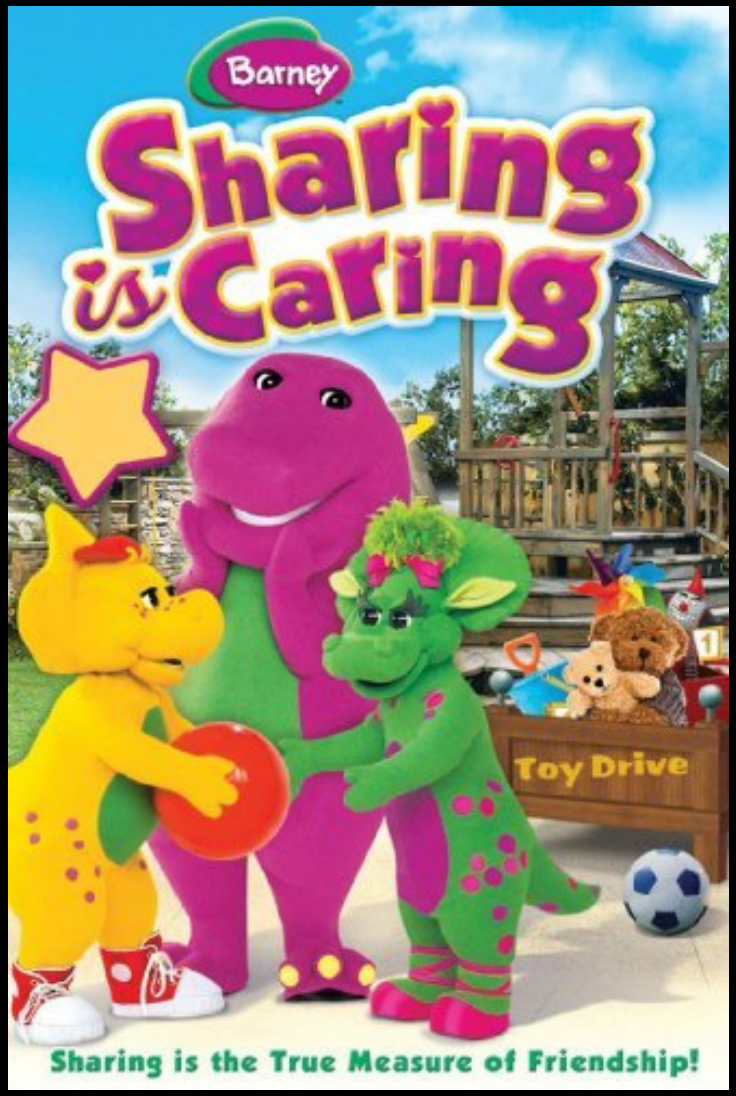
Source:
www.vaccines.gov

Herd Immunity…

… would imply that other people in your sharing community being immune to malware A meant your likelihood of infection from it wa 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-deeee-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…

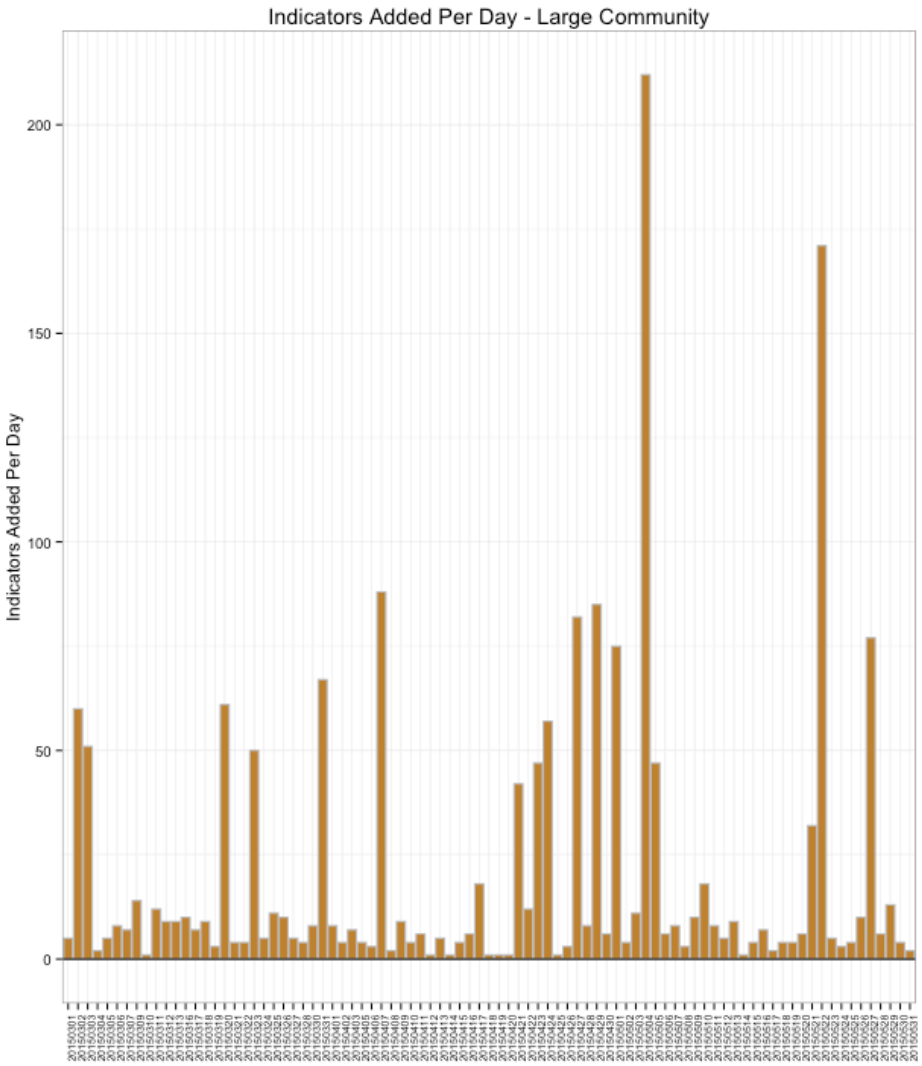
… and also the sharing communities that chose to remain anonymous. You know who you are, and we ❤ you too.

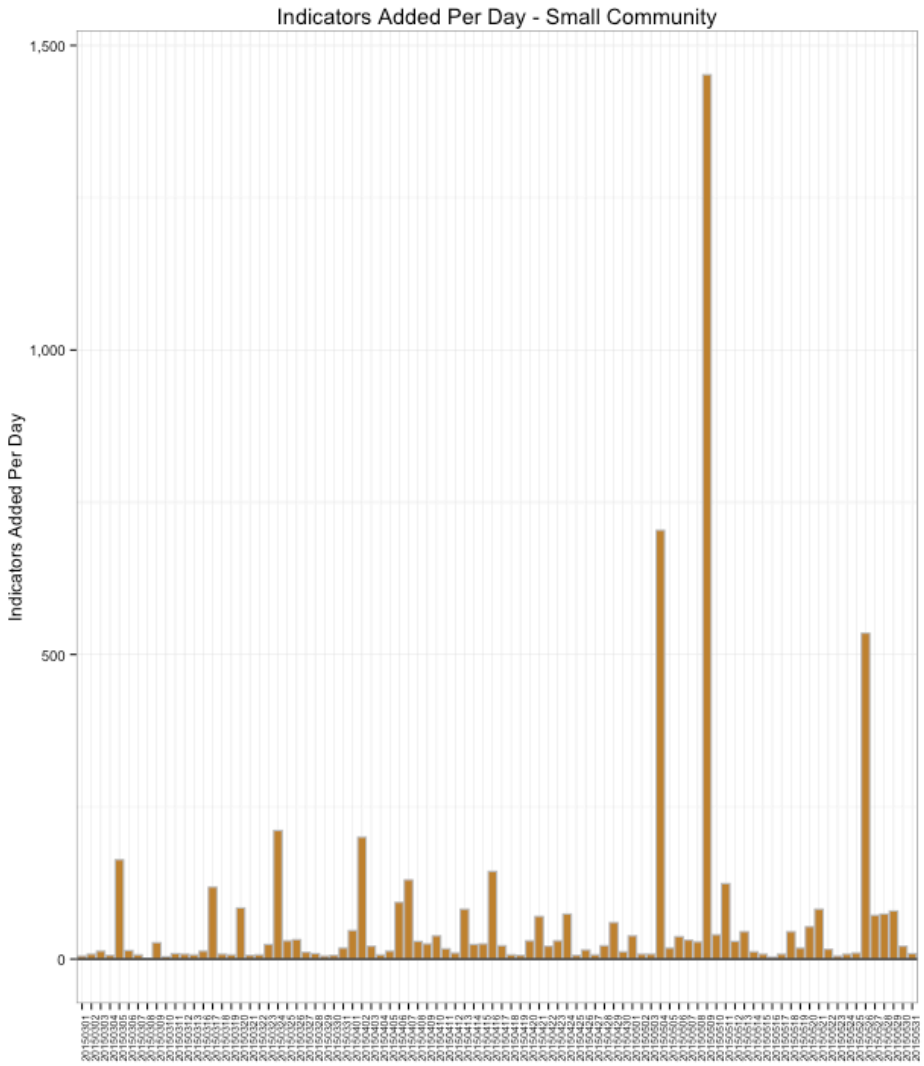
Threat Intelligence Sharing – Data

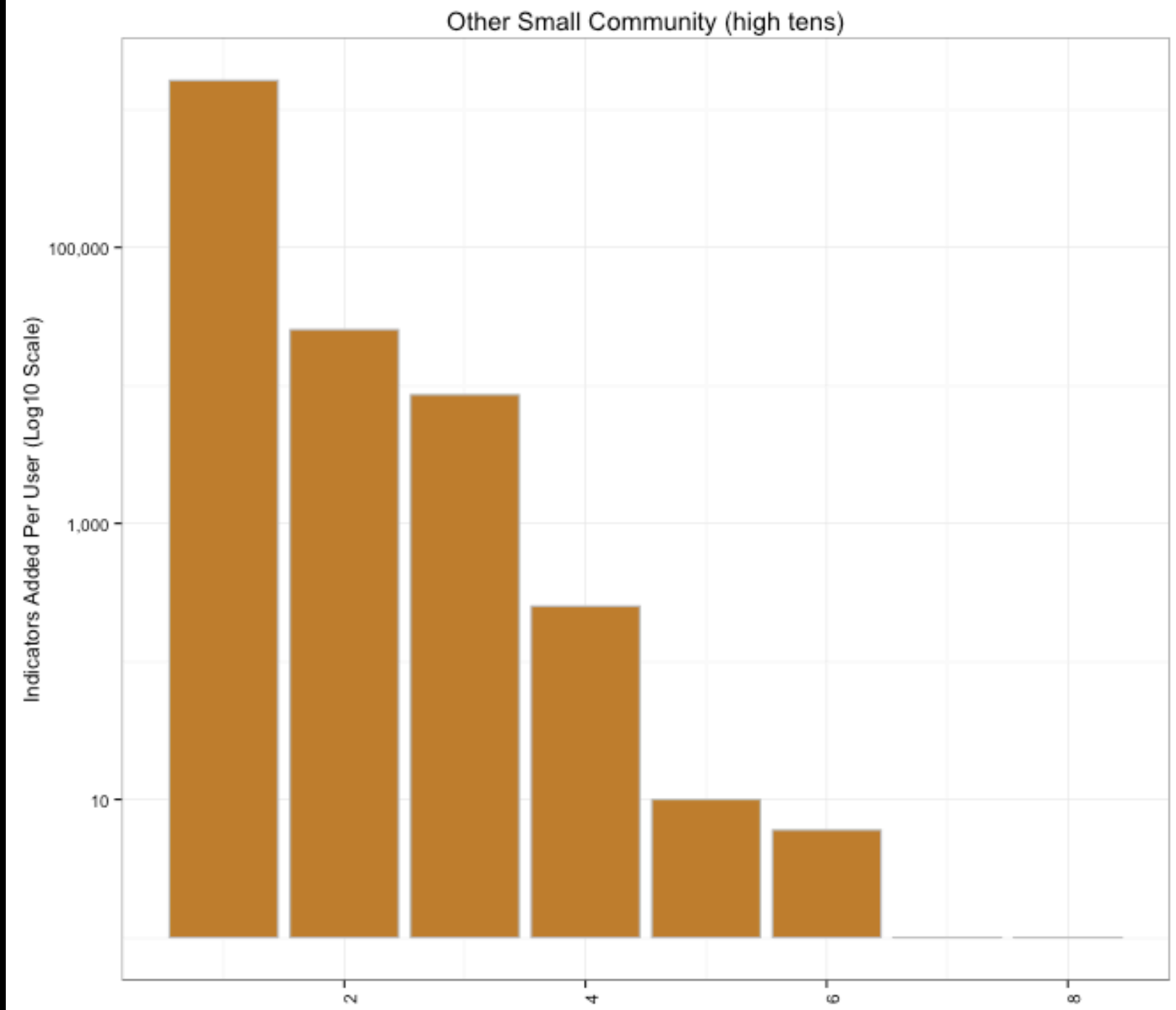
From a period of 2015-03-01 to 2015-05-31:

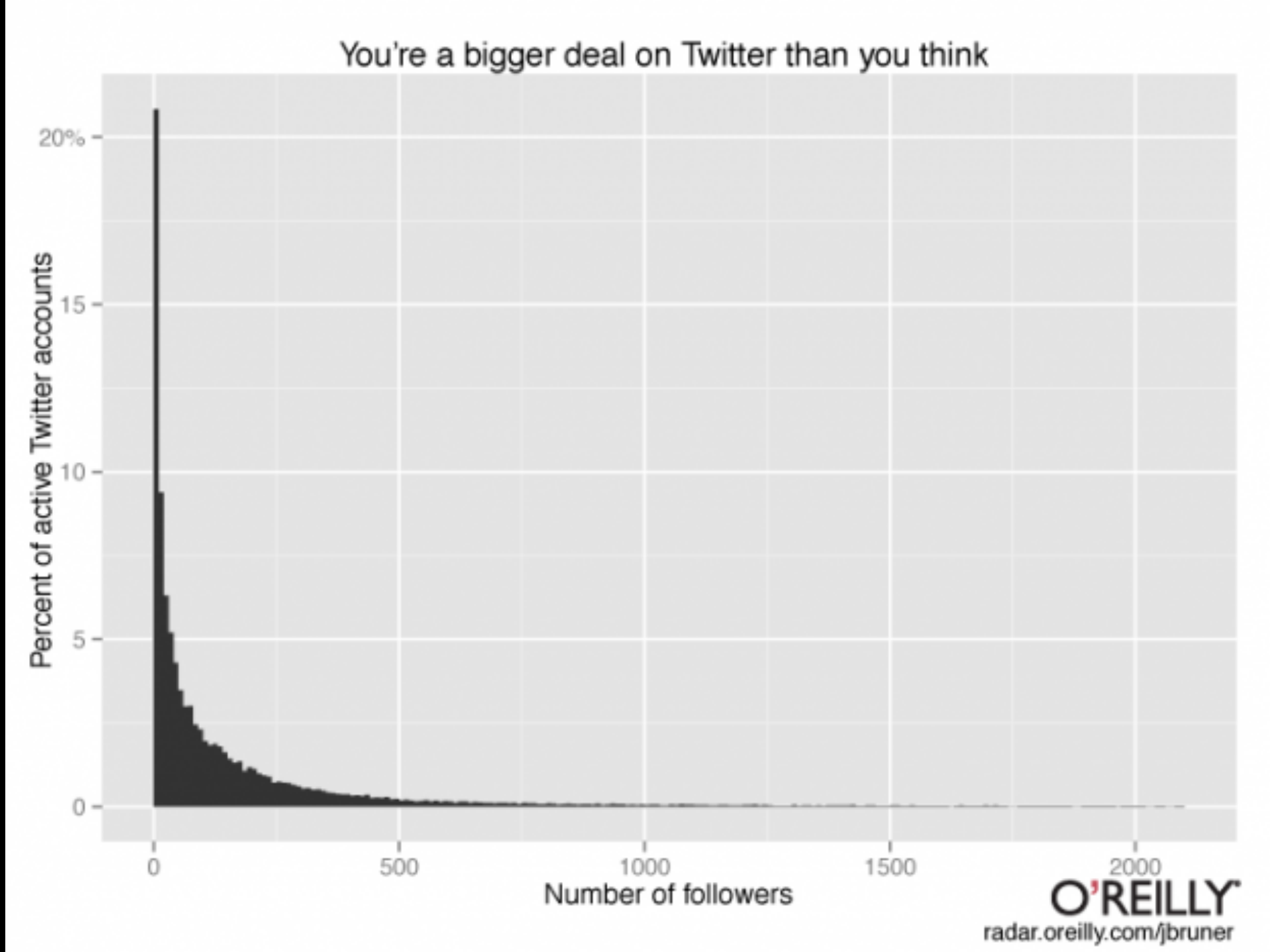
- Number of Indicators Shared
  - Per day
  - Per member

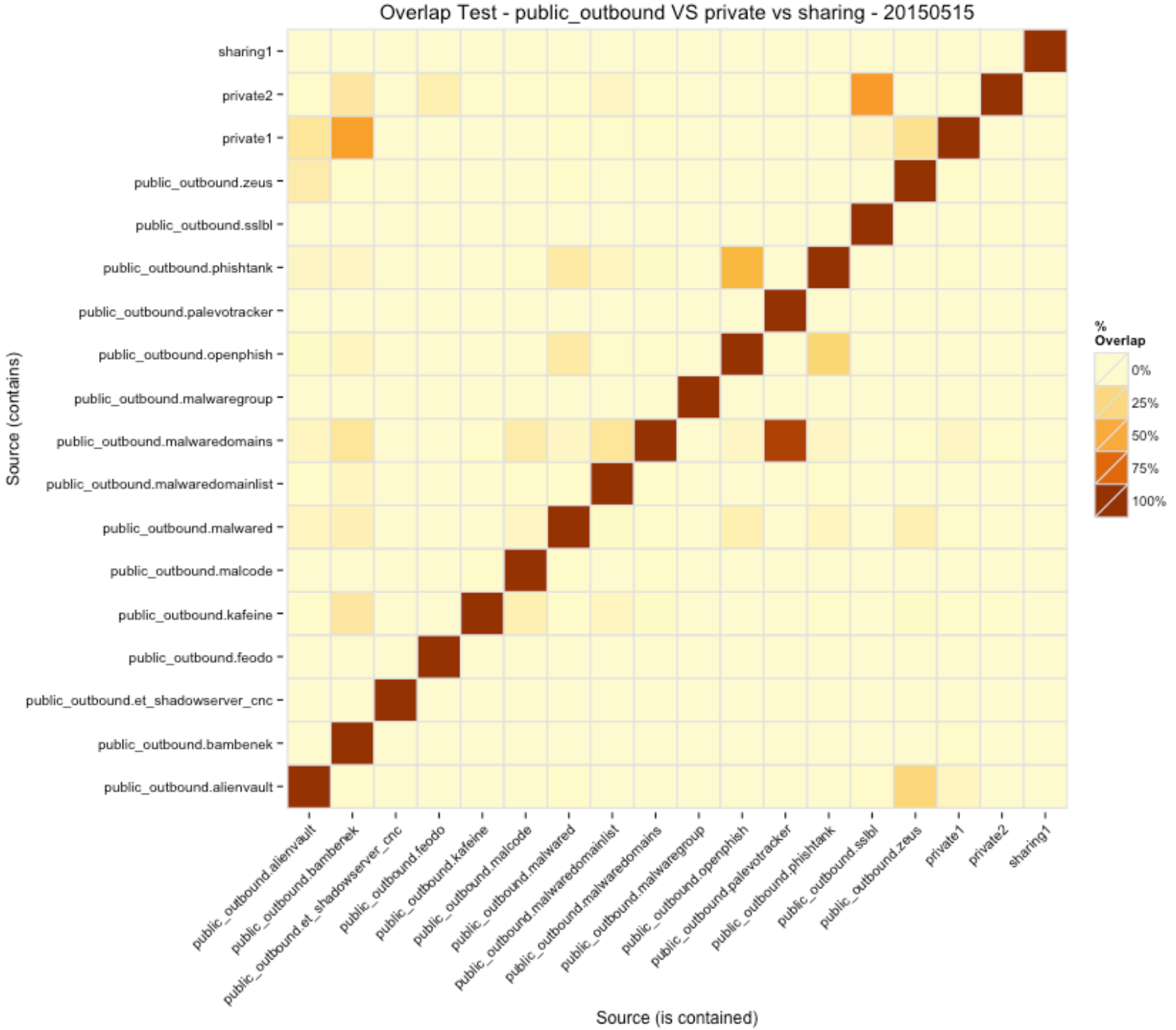
Not sharing this data – privacy concerns for the members and communities

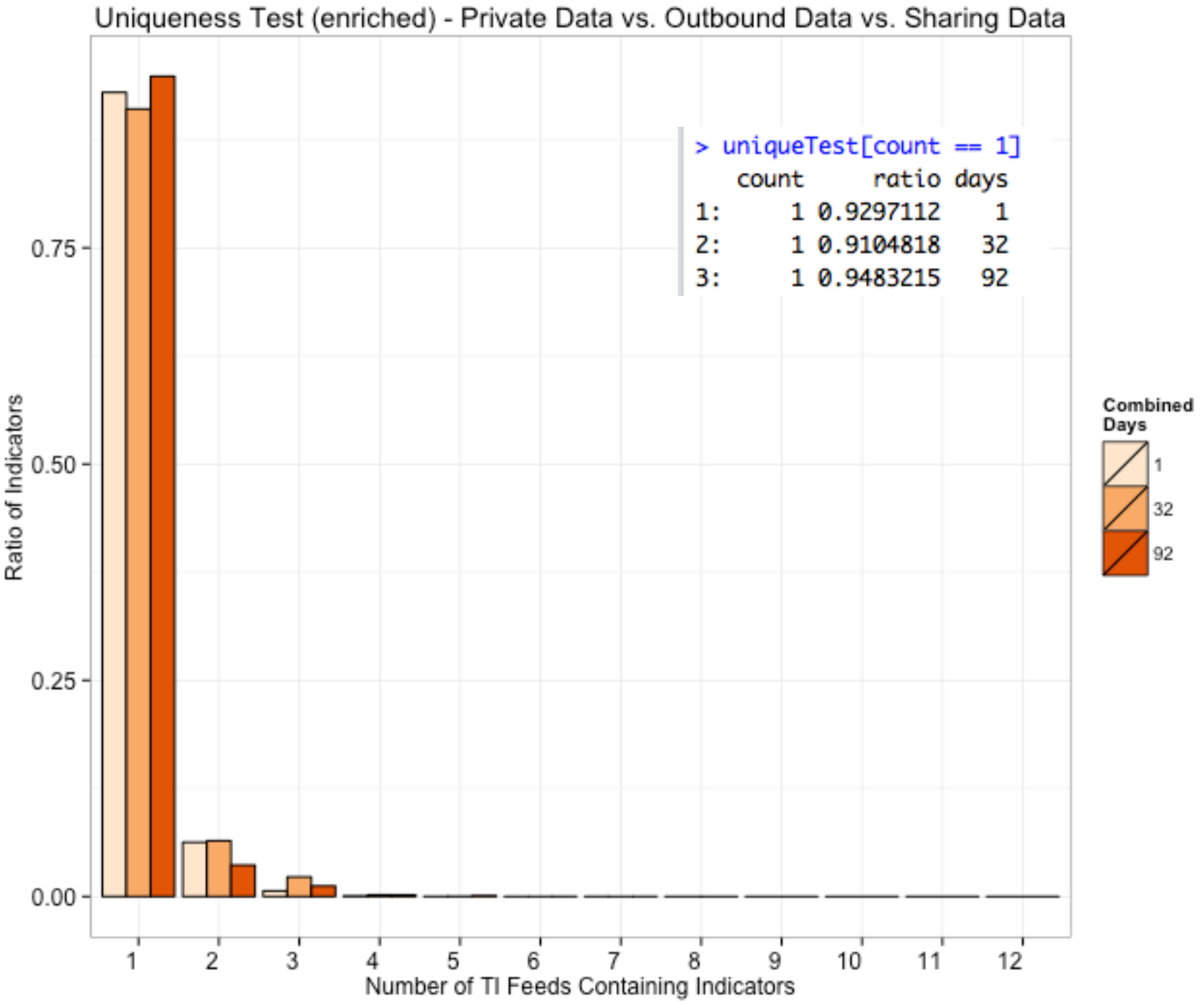

Indicators Added Per Day - Large Community
Indicators Added Per Day
200
150
100
50
0


Indicators Added Per Day - Small Community
Indicators Added Per Day
1,500
1,000
500
0


Other Small Community (high tens)
Indicators Added Per User (Log10 Scale)
100,000
1,000
10
2
4
6
8

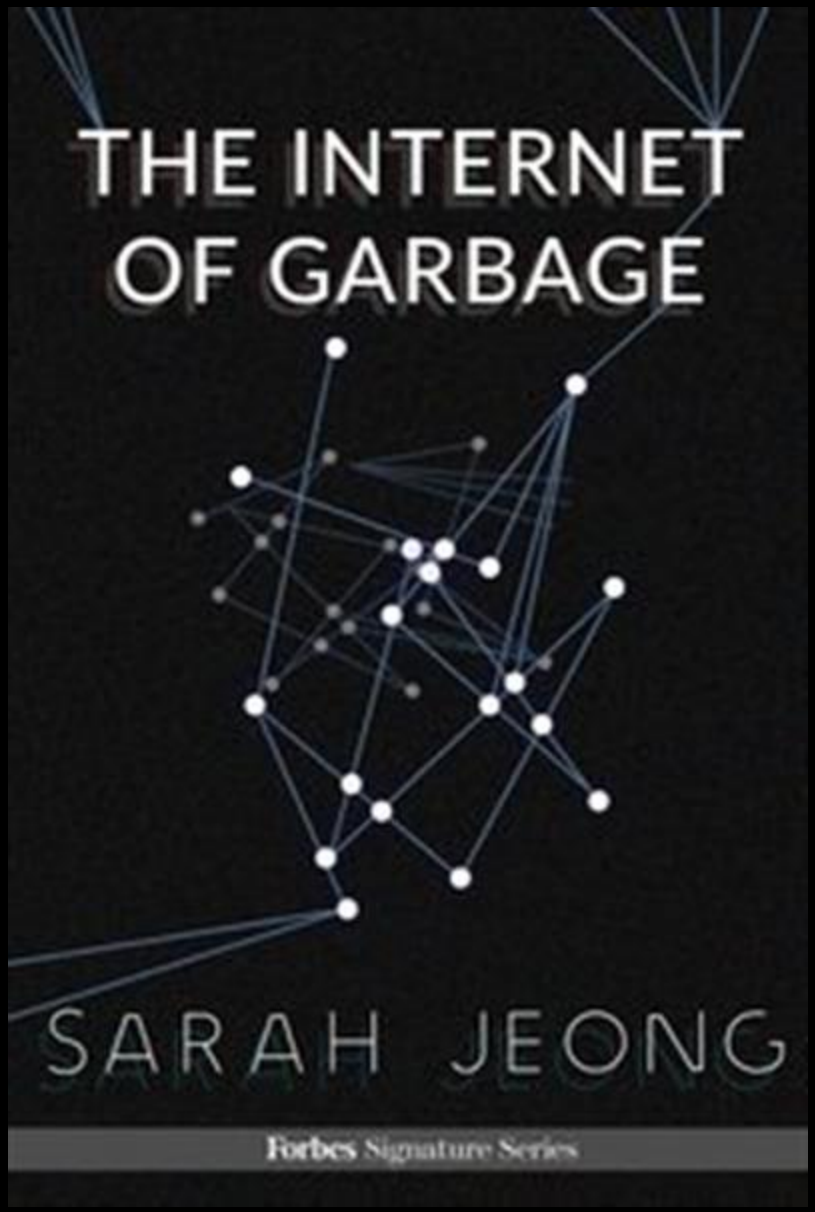

You're a bigger deal on Twitter than you think
20%
15
10
5
0
Percent of active Twitter accounts
0
500
1000
1500
2000
Number of followers
O'REILLY
radar.oreilly.com/jbruner


Overlap Test - public_outbound VS private vs sharing - 20150515
Source (contains)
sharing1
private2
private1
public_outbound.zeus
public_outbound.sslbl
public_outbound.phishtank
public_outbound.palevotracker
public_outbound.openphish
public_outbound.malwaregroup
public_outbound.malwaredomains
public_outbound.malwaredomainlist
public_outbound.malwared
public_outbound.malcode
public_outbound.kafeine
public_outbound.feodo
public_outbound.et_shadowserver_cnc
public_outbound.bambenek
public_outbound.alienvault
Source (is contained)
% Overlap
0%
25%
50%
75%
100%


Uniqueness Test (enriched) - Private Data vs. Outbound Data vs. Sharing Data
> uniqueTest[count == 1]
count ratio days
1: 1 0.9297112 1
2: 1 0.9104818 32
3: 1 0.9483215 92
Ratio of Indicators
0.00
0.25
0.50
0.75
1
2
3
4
5
6
7
8
9
10
11
12
Number of TI Feeds Containing Indicators
Combined Days
1
32
92


same

MATURITY?



“Reddit of Threat Intelligence”?

451 Research

Spotlight

Threat intelligence: only for the 1%?

Analyst: Scott Crawford 1 Jul, 2015

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)

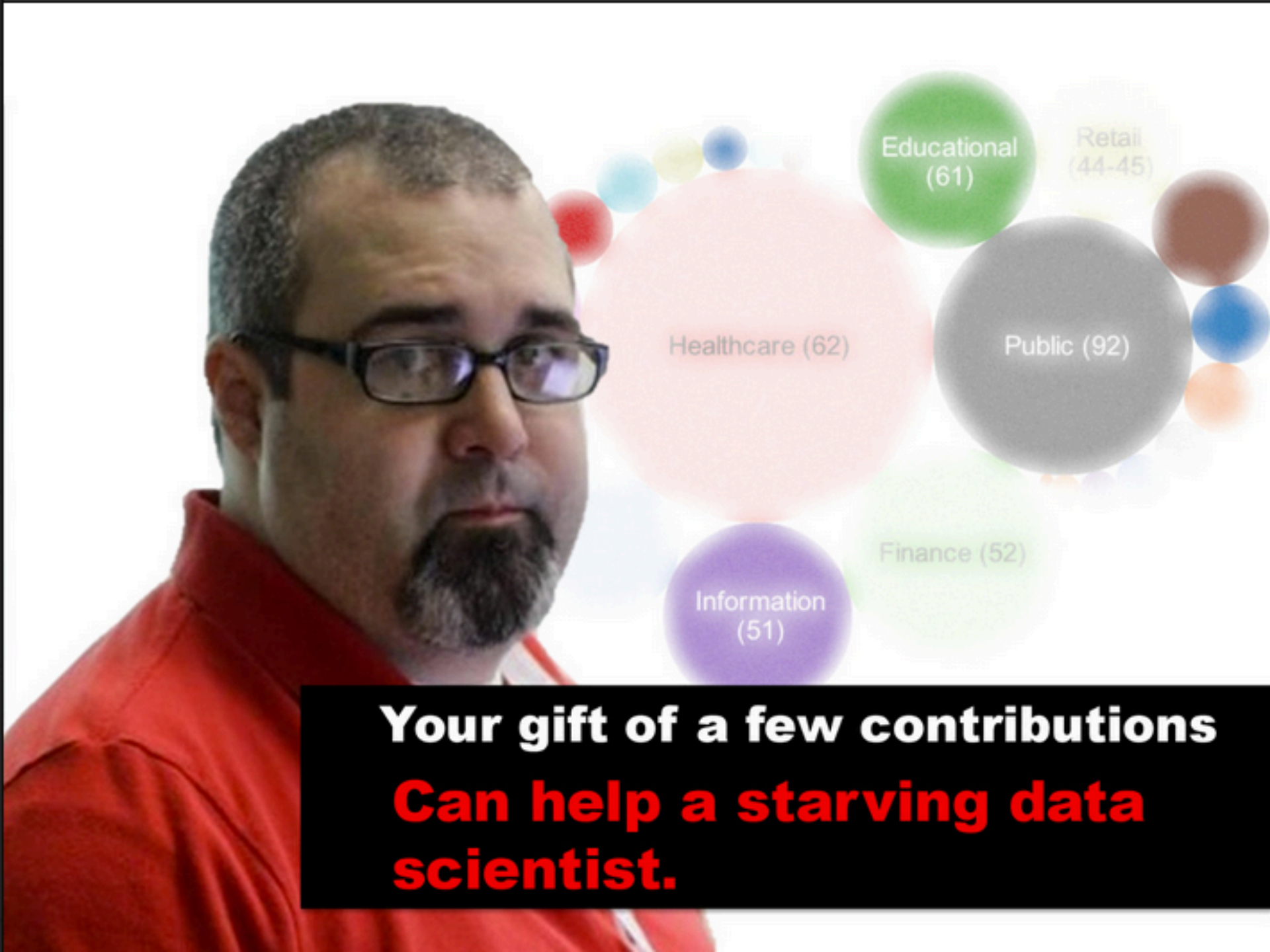
TELEMETRY > CONTENT

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!

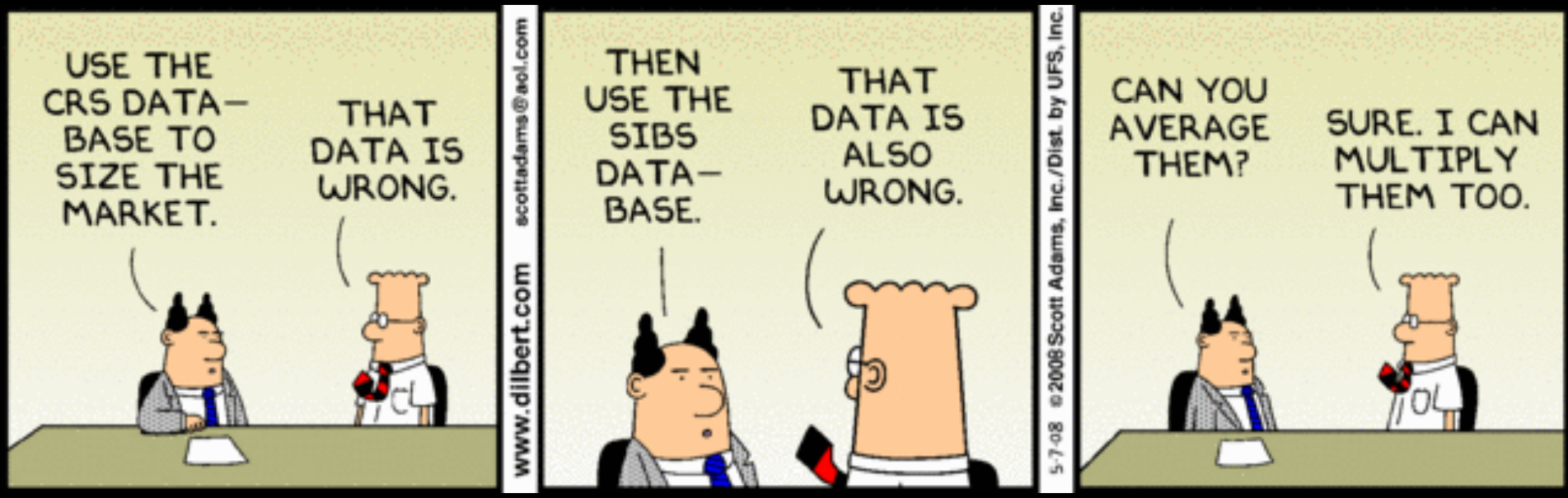

Educational (61)
Retail (44-45)
Healthcare (62)
Public (92)
Finance (52)
Information (51)
Your gift of a few contributions
Can help a starving data scientist.

Thanks!

- Q&A?
- Feedback!

Alex Pinto
@alexcpsec
@MLSecProject
@NiddelCorp

Alexandre Sieira
@AlexandreSieira
@MLSecProject
@NiddelCorp



”The measure of intelligence is the ability to change."

- Albert Einstein