OpenDNS

The Security Wolf of Wall Street: Fighting Crime with High-Frequency Classification and Natural Language Processing

Jeremiah O'Connor and Thibault Reuille

January 2016

OpenDNS is now part of Cisco.

CISCO

$ whois jeremiah

-Mad Scientist at OpenDNS/Cisco Labs
-M.S. in Computer Science from University of San Francisco
-Previously worked at Mandiant (IR/DNS Research), Evernote (AppSec/IR), Uber (Data Science)
-Career Goals: Solve interesting problems (Networking/Security, Bioinformatics, GPS Tracking, Video Games, etc.)
-Proud SFSPCA Pitbull Puppy owner

$ whois thibault

- Security Research Team at OpenDNS.
- Creator of OpenGraphiti.
- Focus: Data Visualization, 3D Graphics, Graph Theory and Real-time systems.

Presentation Agenda

- Introduction : Challenges & Hypothesis
- Real-Time Processing Fundamentals
- The Avalanche Project & The Research Pipeline
- Live Demo!
- Future Work

Introduction to Avalanche

OpenDNS

Challenges

I've got 99 problems but malware ain't one!

- We see a lot of traffic.
  - Needles in a haystack.
- Bad guys move fast.
  - The needles are playing hide-and-seek.
- Outdated information has less impact than hot news.
  - Slowpoke syndrome.
- Measuring the accuracy of our classifiers is not trivial.
  - How big is the base of the iceberg?

Hypothesis

To stream or not to stream.

- Most of our models can work in streaming.
  - Well, that's a strong statement.
- We can detect "anomalies" on the fly.
  - TSA is overrated anyway.
- We can have precise visibility over malicious activity.
  - Statistics + Dataviz = Win!
- We can talk about what nobody knows.
  - Wanna be famous?

REAL-TIME !

Real-Time, you said?

Different Levels of Constraints.

- "Soft"
  - Ex: Youtube / Netflix video streaming, Video Games, GPS …
- "Firm" :
  - Ex: Dishwasher, Audio DSP, Assembly line …
- "Hard" :
  - Ex: Airbag, UHFT Algorithmic Trading …
- "Critical" :
  - Ex: Missiles, Aircrafts, Nuclear Reactor …

- "Near Real-Time" : Network-induced indeterminism.

The Blackbox Abstraction

Real-Time vs High Performance.

T1 – T0 ~ 1 second
vs
T1 – T0 <= 2 seconds !!

Real-time != Fast

When Murphy meets the law of large numbers.

There's no such thing as "half water-proof".

Program

Runs fine 99% of the time
Probability of a crash : 1%

99% → 99% x 99% → 99% x 99% x 99% → … → 99% ^ N → … → ZERO !

At infinity, a program that SOMETIMES crashes is equivalent to a program that ALWAYS crashes!

OpenDNS

Key Design Points

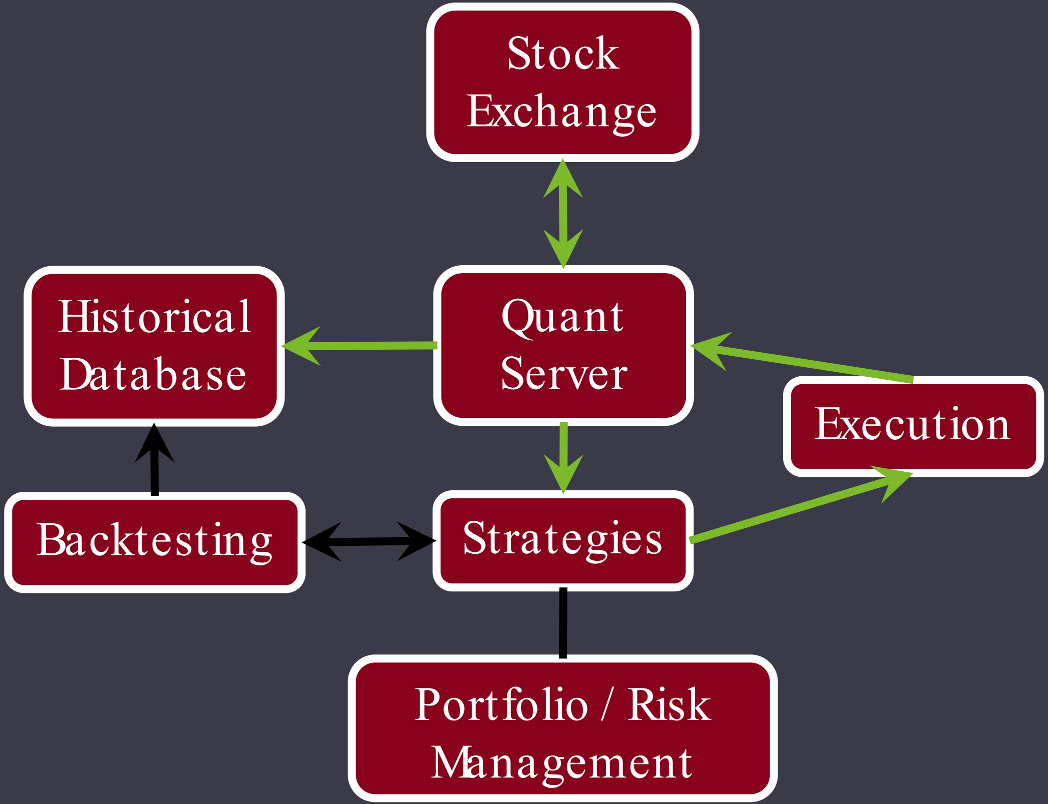
Things to consider when writing code.

- Fault Tolerancy
  - Rigorous code.
  - Flawless error handling.
  - Unit tests
  - Degraded Mode?
- Algorithm Complexity : What's your worst case?
  - Computing Time : Is it deterministic?
  - Parallelism & Concurrency : Context Switching, Deadlocks, Race Condition…
  - Memory Allocation : Static vs Dynamic
- Environment
  - Background jobs, RAM, CPUs, Parasites, Hardware Failures…


The Avalanche Project
OpenDNS

High Frequency Trading vs Traffic Classification

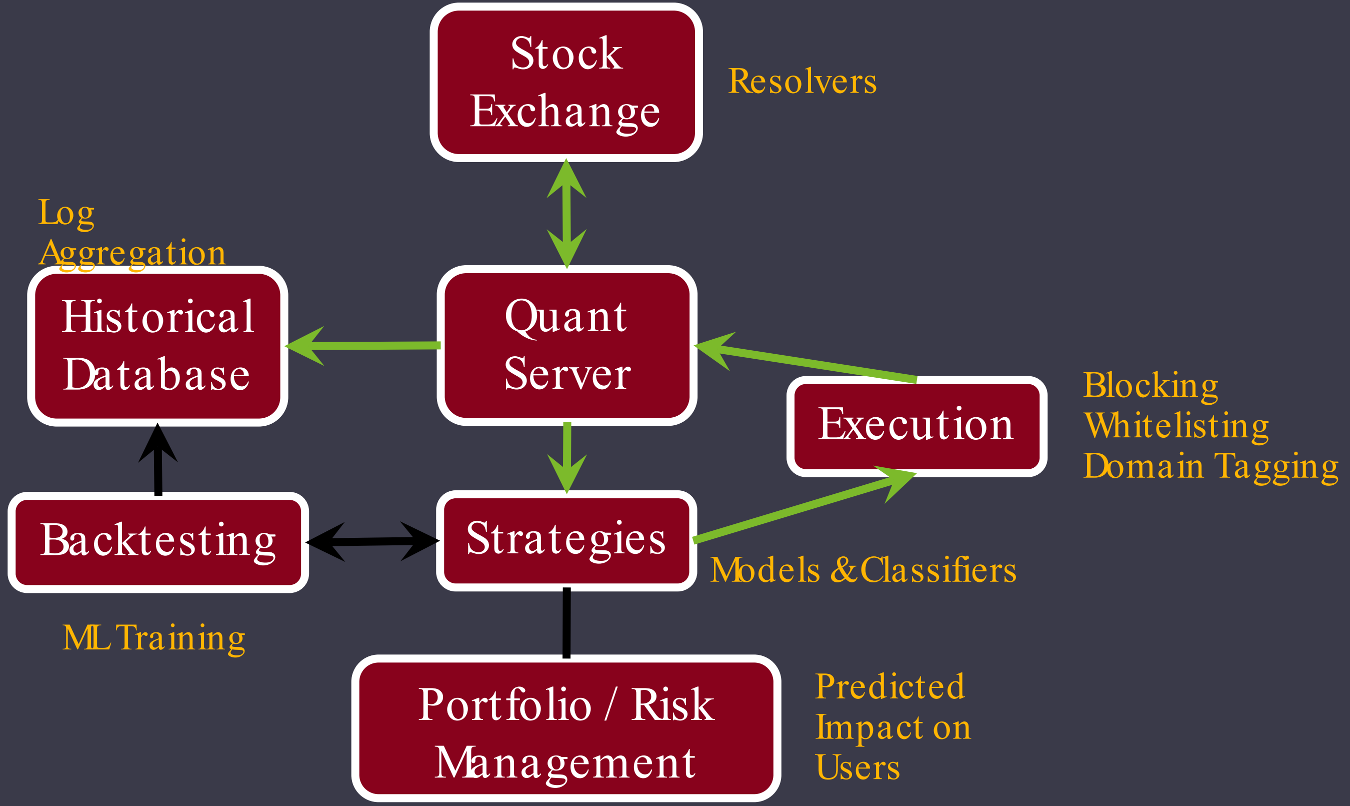
The Wolf of Wall Street



OpenDNS

High Frequency Trading vs Traffic Classification

The Wolf of Wall Street



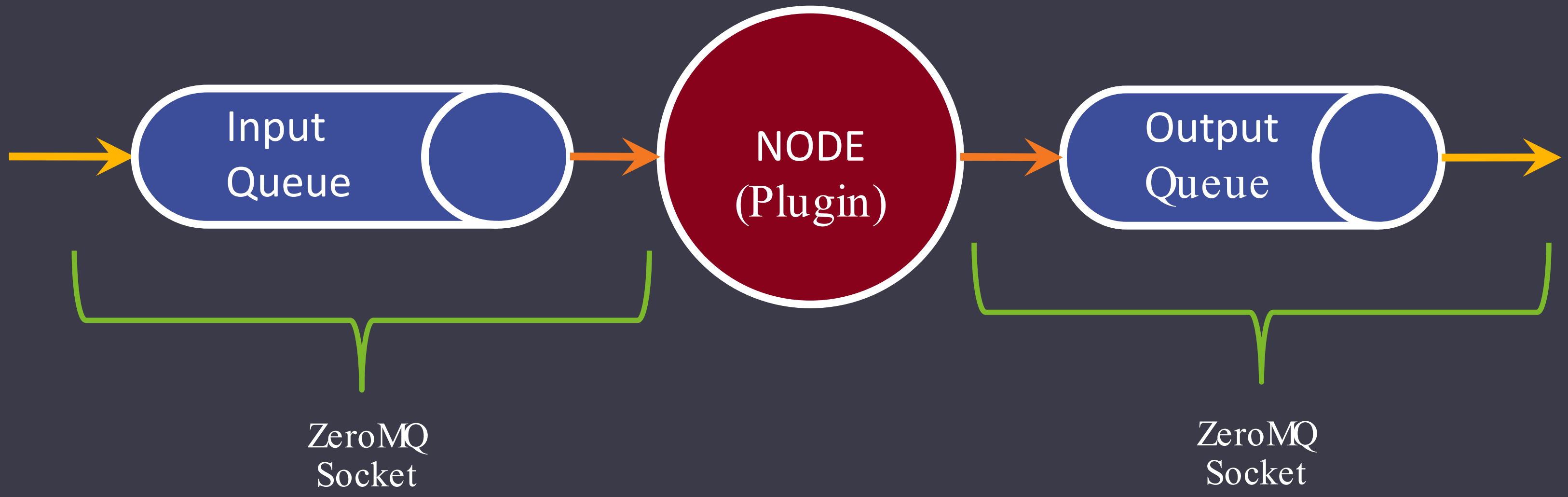
OpenDNS

What is Avalanche?

Overview and Technical Details.

- Open source project :
  - http://github.com/ThibaultReuille/avalanche
- "Real-time" data processing framework.
- Modular, parallel and distributed design.
- Written with Python and ZeroMQ.
- Platform for some OpenDNS models (Private) :
  - https://github.office.opendns.com/Research/avalanche-opendns
  - NLP-Rank
  - DNS Tunnelling
  - Talos DGA classifier and others (In progress)

Avalanche Design
Divide and Conquer

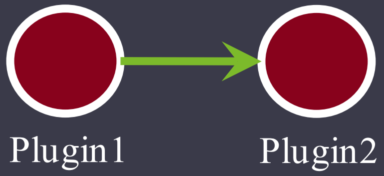


Avalanche Node

Plugin Template Code

```python
import json
import plugins.base

class Plugin1(plugins.base.Plugin):
    def __init__(self, info):
        # NOTE: The info argument contains the full node definition
        # written in the pipeline configuration file.
        pass

    def process_message(self, message):
        # NOTE : Here we can process the message, add field, remove, etc.
        # Retuning None drops the message from the pipeline.
        return message

class Plugin2(plugins.base.Plugin):
    def __init__(self, info):
        # NOTE: The info argument contains the full node definition
        # written in the pipeline configuration file.
        pass

    def run(self, node):
        # NOTE: Each node runs on its own thread/process,
        # Here we enter our infinite loop.
        while True:

            # NOTE: Read incoming data sent to our node
            data = node.input.recv()

            # NOTE: Parse it as a JSON message
            message = json.loads(data)

            # NOTE: This template plugin doesn't do anything except being a passthru filter.
            # This is where the processing would actually happen in a real processor.
            # You can send whatever data you like in the output stream. That can be a modified
            # version of the incoming messages or any other message of your creation.

            # NOTE: Send it back through the pipeline
            node.output.send_json(message)

if __name__ == "__main__":
    print("Please import this file!")
```

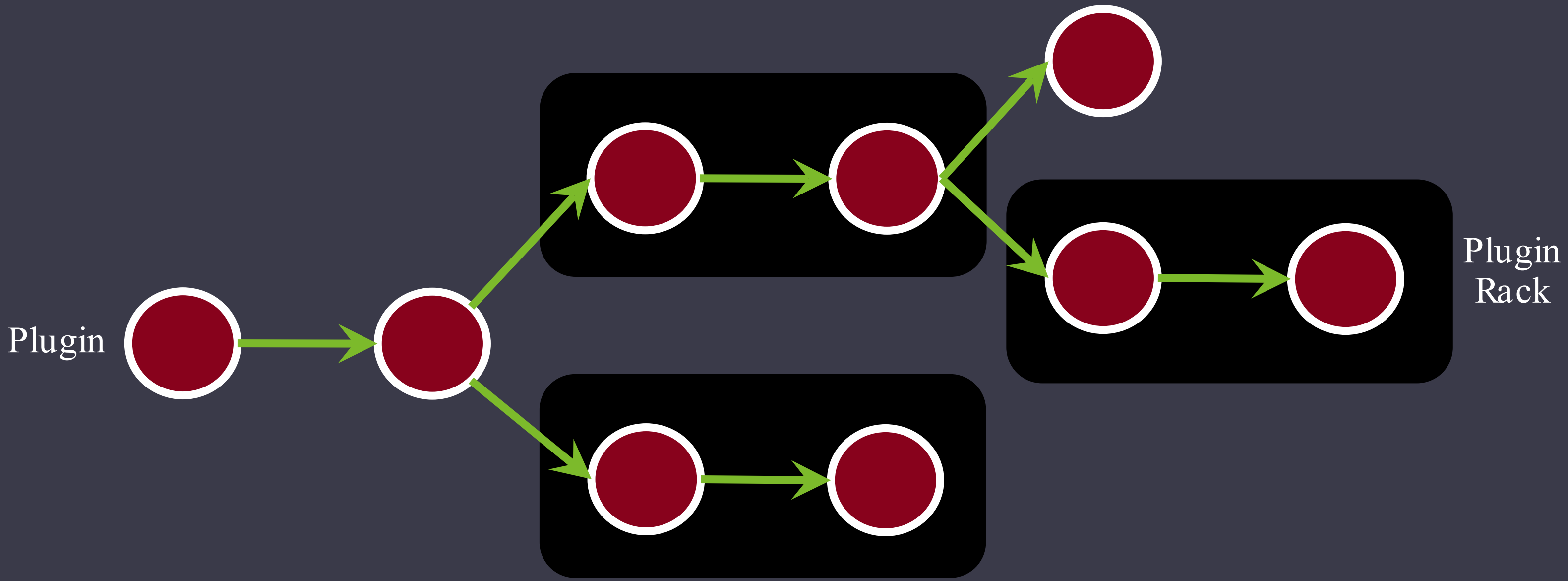
Avalanche Graph
Pipeline Definition




```
{
    "attributes" : {
        "plugins" : [
            { "name" : "plugin1", "filename" : "path/to/plugin1.py" },
            { "name" : "plugin2", "filename" : "path/to/plugin2.py" }
        ]
    },

    "nodes" : [
        {
            "id" : 0,
            "type" : "plugin1",
            "attributes" : {
                "my_data" : "my_value"
            }
        },

        {
            "id" : 1,
            "type" : "plugin2",
            "attributes" : {
                "other_data" : "other_value"
            }
        }
    ],

    "edges" : [
        { "id" : 0, "src" : 0, "dst" : 1 }
    ]
}
```

Avalanche Pipeline

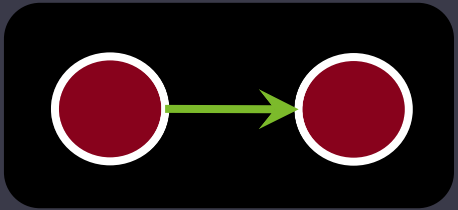
Divide and Conquer



OpenDNS

Avalanche Rack

Plugin Rack Definition



```
{
    "id" : 0,
    "type" : "rack",
    "plugins" :
    [
        {
            "type" : "plugin1",
            "attributes" : { "my_data" : "my_value" }
        },
        {
            "type" : "plugin2",
            "attributes" : { "other_data" : "other_value" }
        }
    ]
}
```

Run Avalanche

```
$ ./avalanche.py path/to/my_pipeline.json 10000
```

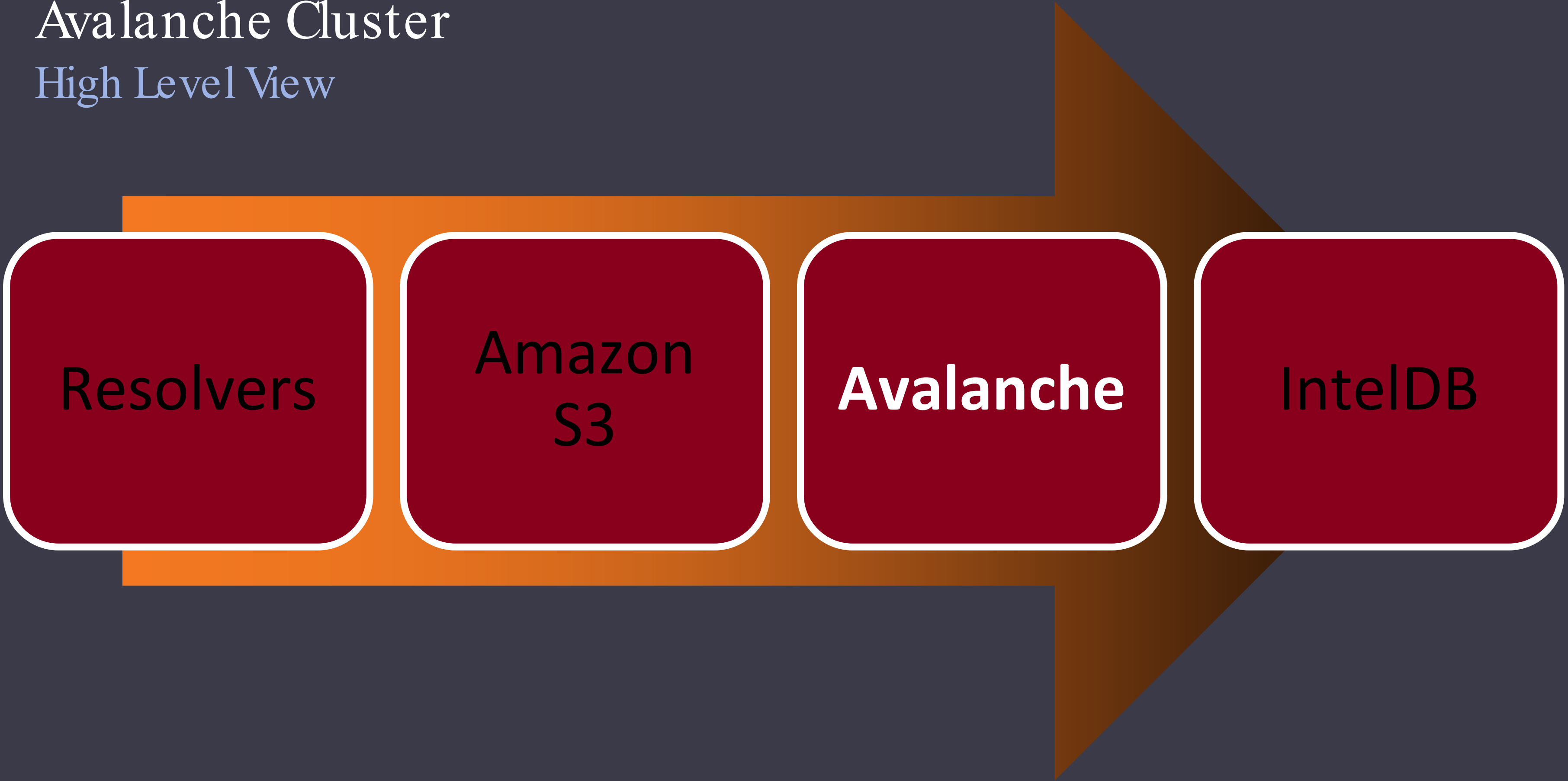
- Things you get for free :
  - Modularity.
  - Multi-Threading.
  - A library of plugins ready-to-use.
  - Reusability & collaboration.
  - An insanely fast messaging system.

The Research Pipeline

OpenDNS

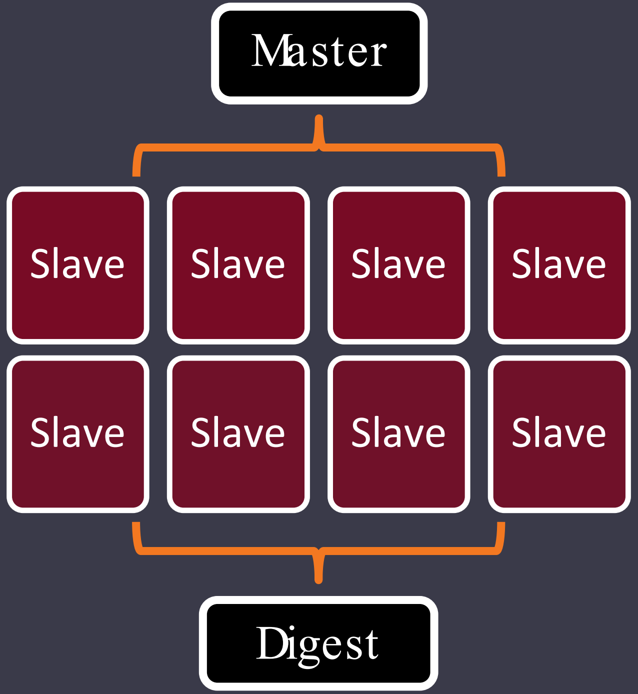
Avalanche Cluster

High Level View



Avalanche Cluster

- 8 Amazon instances
- Master distributes work
  - Round-robin
  - “Fire and forget”
- Slaves process the chunks
- 4 Avalanche pipelines
- Results are centralized



Cluster Management with Boto & Fabric

```
treuille — avalanche@ip-10-20-9-89: ~/avalanche-services — ssh — 168×24
avalanche@ip-10-20-9-89:~/avalanche-services$ ls
avalanche.pem  digest.sh  instances.py  instances.pyc  miner.conf  profile.conf  pusher.py  requirements.txt  results.json  stats.py
avalanche@ip-10-20-9-89:~/avalanche-services$ fab -f instances.py -i avalanche.pem -- uptime
[Instance:i-c029ac72, Instance:i-c129ac73, Instance:i-ca29ac78, Instance:i-cb29ac79, Instance:i-ce29ac7c, Instance:i-cf29ac7d, Instance:i-cd29ac7f, Instance:i-cc29ac7e]
[10.20.9.96] Executing task '<remainder>'
[10.20.9.96] run: uptime
[10.20.9.96] out:  17:25:59 up 21 days, 16:49,  1 user,  load average: 0.02, 1.63, 2.35
[10.20.9.96] out:

[10.20.9.97] Executing task '<remainder>'
[10.20.9.97] run: uptime
[10.20.9.97] out:  17:24:43 up 21 days, 16:48,  1 user,  load average: 6.19, 2.92, 2.34
[10.20.9.97] out:

[10.20.9.90] Executing task '<remainder>'
[10.20.9.90] run: uptime
[10.20.9.90] out:  17:25:29 up 21 days, 16:48,  1 user,  load average: 0.04, 1.61, 2.51
[10.20.9.90] out:

[10.20.9.91] Executing task '<remainder>'
[10.20.9.91] run: uptime
[10.20.9.91] out:  17:25:58 up 21 days, 16:49,  1 user,  load average: 0.04, 1.60, 1.82
[10.20.9.91] out:
```

https://github.office.opendns.com/Research/avalanche-services

OpenDNS

Traffic Speed vs Avalanche Pipeline

Numbers don't lie.

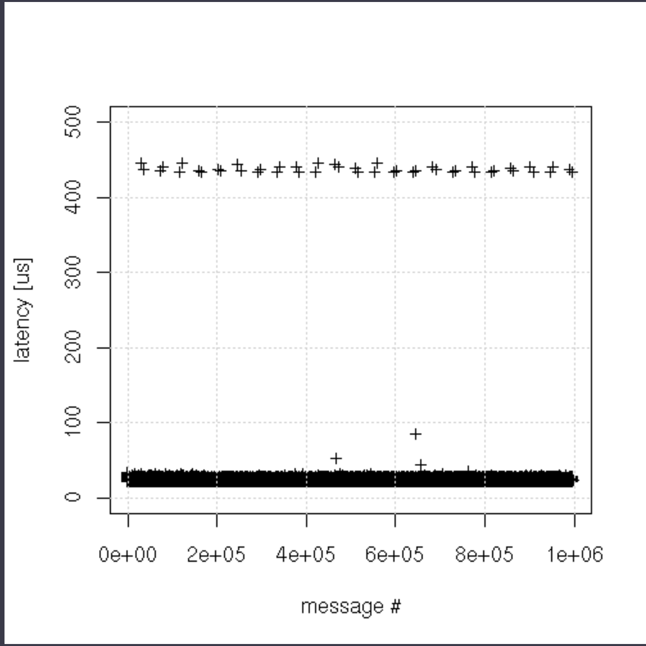
Queries / Chunk	Authlogs (AMS.m1)	Querylogs (AMS.m1)
Noon (UTC)	564 752	6 147 997
Midnight (UTC)	412 050	3 315 157
Queries / Second	**Authlogs (AMS.m1)**	**Querylogs (AMS.m1)**
Noon (UTC)	941.25	10246.66
Midnight (UTC)	686.75	5525.26

- Avalanche Benchmark :
  - ~30000 messages per second ⇔ 1 message every 33 microseconds.
  - 3 times faster than AMS.m1 query logs at peak time.

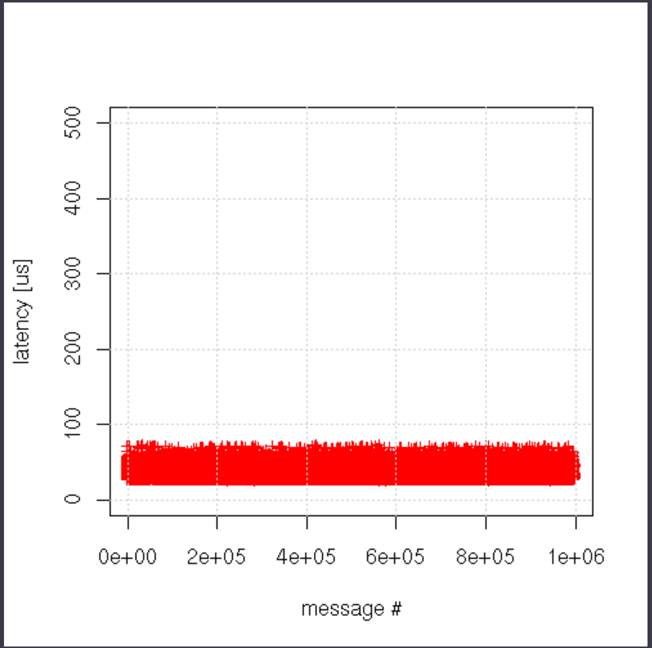
OpenDNS

ZeroMQ Performance Tests

Standard Linux Kernel

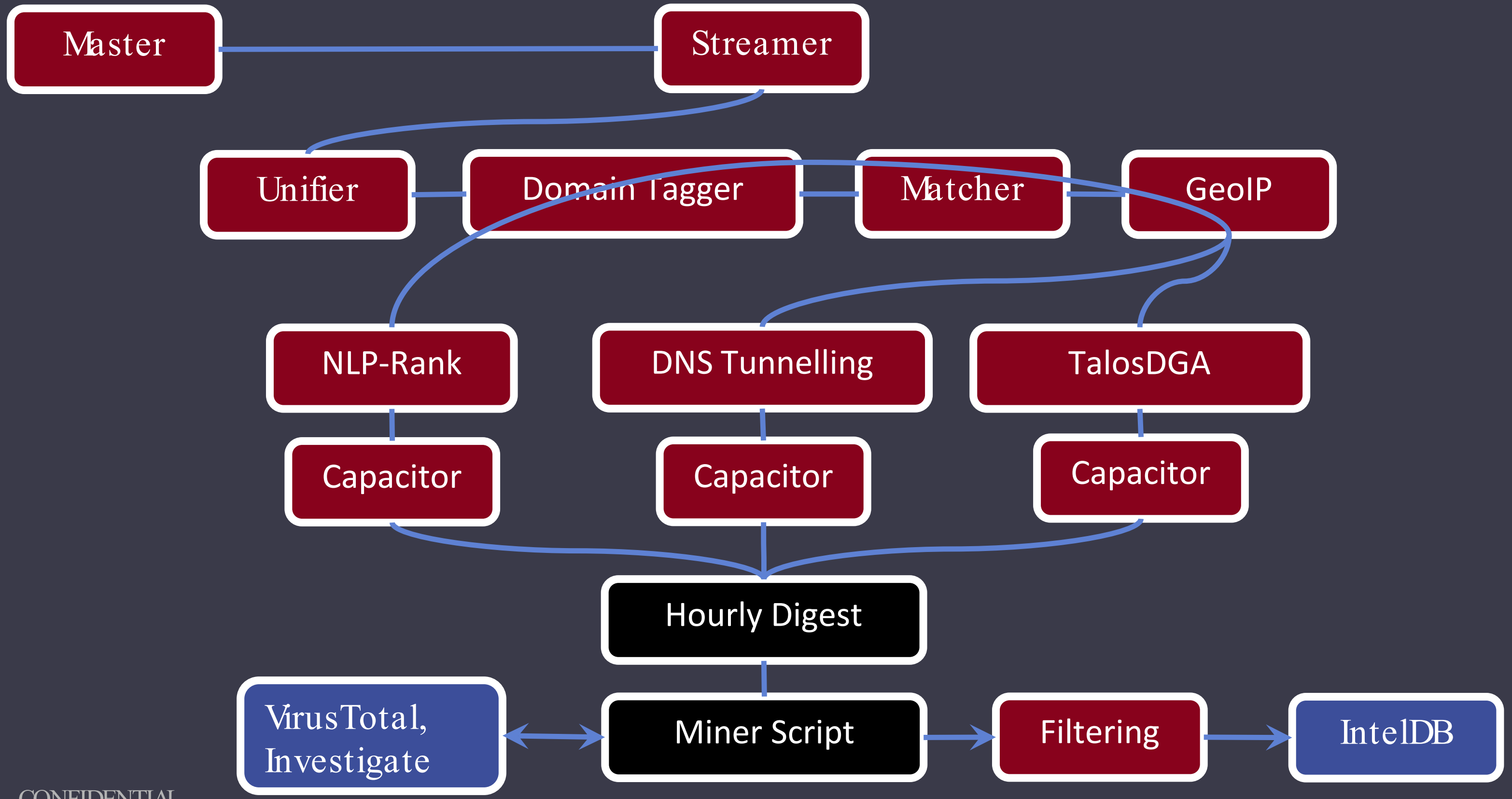


Real-Time Linux Kernel



Source: http://zeromq.org/results:rt-tests-v031

Slave Processing Pipeline



avalanche.r1.usw1.opendns.com/avalanche/

Index of /avalanche/

../		
dns-tunnelling/	06-Nov-2015 00:15	
nlp-rank/	06-Nov-2015 00:13	

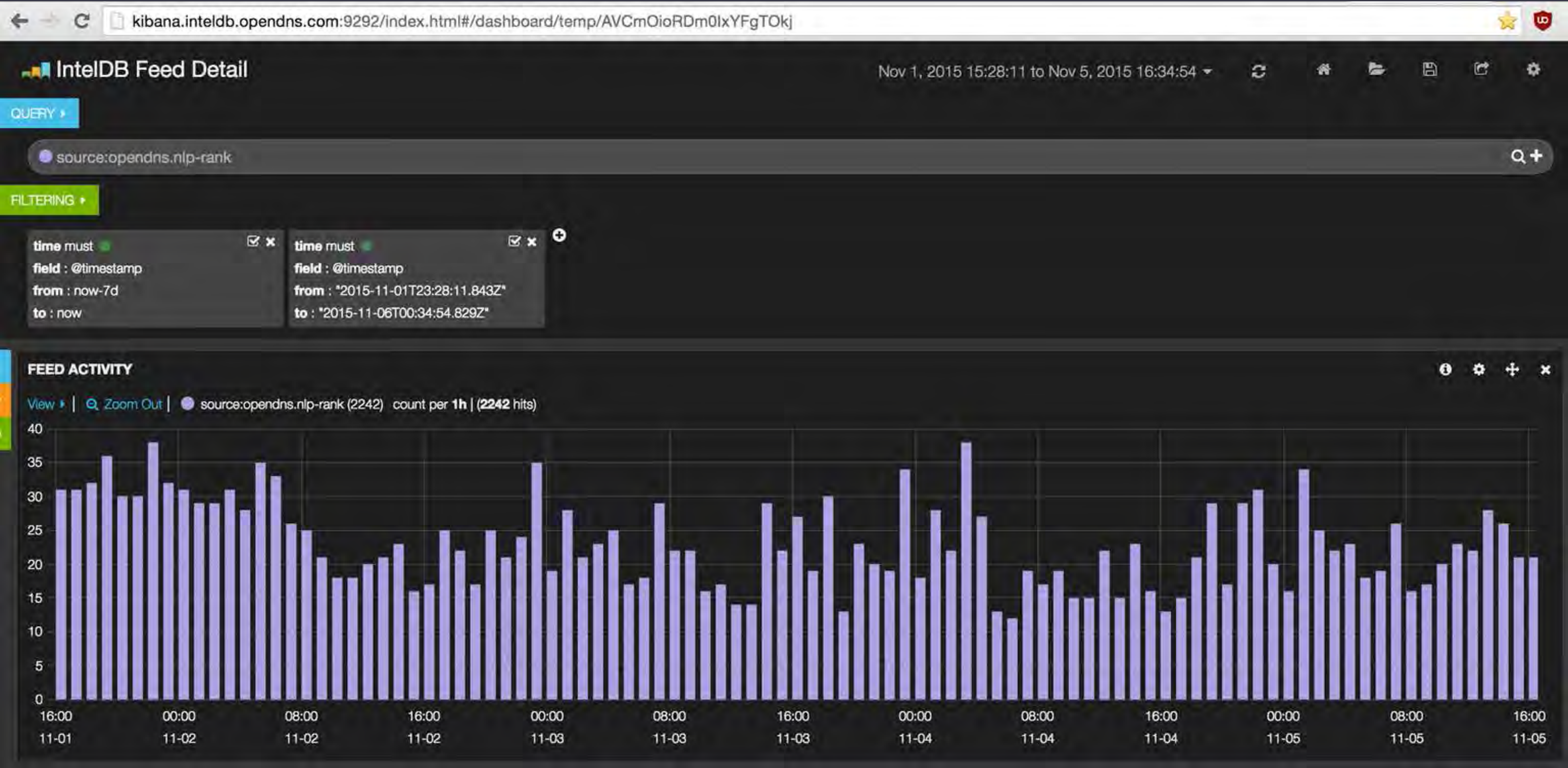
2015.11.05-19.00.01/	05-Nov-2015 19:13	-
2015.11.05-20.00.01/	05-Nov-2015 20:13	-
2015.11.05-21.00.01/	05-Nov-2015 21:12	-
2015.11.05-22.00.01/	05-Nov-2015 22:14	-
2015.11.05-23.00.01/	05-Nov-2015 23:13	-
2015.11.06-00.00.01/	06-Nov-2015 00:13	-
stats.txt	06-Nov-2015 00:14	718
total.txt	06-Nov-2015 00:14	5655720

```
--- Generic Statistics ---

214679 Elements: 188016 domains + 26663 missing data (Ignored).

. Blacklisted: 3867
. Greylisted: 182233
. Whitelisted: 1916

. VT positives >= 5 : 5222
. Unknown by VT : 176676
. Popularity >= 80.0 : 14

--- Detailed Statistics ---

. Blacklisted and VT >= 5 : 2185
. Blacklisted and unknown by VT : 1002
. Blacklisted and Popularity >= 80.0 : 0

. Greylisted and VT >= 5 : 2865
. Greylisted and unknown by VT : 174123
. Greylisted and Popularity >= 80.0 : 10

. Whitelisted and VT >= 5 : 172
. Whitelisted and unknown by VT : 1551
. Whitelisted and Popularity >= 80.0 : 4
```

Index of /avalanche/nlp-rank/2015.11.06-00.00.01/

../		
domains.txt	06-Nov-2015 00:13	9705
nlp-rank.10.20.9.90.csv	06-Nov-2015 00:12	153216
nlp-rank.10.20.9.91.csv	06-Nov-2015 00:11	141006
nlp-rank.10.20.9.92.csv	06-Nov-2015 00:10	108028
nlp-rank.10.20.9.93.csv	06-Nov-2015 00:09	87443
nlp-rank.10.20.9.94.csv	06-Nov-2015 00:13	158555
nlp-rank.10.20.9.95.csv	06-Nov-2015 00:11	140592
nlp-rank.10.20.9.96.csv	06-Nov-2015 00:10	114785
nlp-rank.10.20.9.97.csv	06-Nov-2015 00:08	77933
stats.txt	06-Nov-2015 00:13	613


```
#FQDN,depth,popularity,age,ips,prefixes,asns,countries,ttl_min,ttl_max,ttl_stddev,geo_sum,geo_mean,entropy,perplexity,
apple-winks.com,0,0.0,,1,1,1,1,600,600,0.0,0.0,0.0,3.2776134368191165,0.2739846357448707,0,6
ebay.login.com.5599.carsgoneby.aspmodel.info,0,0.0,,,,,,,,,,,3.0,0.6361674803007081,-1,6
ekosamazonia.com.br,0,7.169532493946863,,1,1,1,1,14400,14400,0.0,0.0,0.0,3.0220552088742,0.4266416677105029,-1,11
www.microsoftpartnerserverandcloud.com,0,50.50501253890862,,1,1,1,1,3600,3600,0.0,0.0,0.0,3.8029100796497266,0.5594928
serviceapple-support.bugs3.com,0,0.0,,1,1,1,1,14400,14400,0.0,0.0,0.0,2.321928094887362,0.524856068944591l,-1,9
secure2.store.apple.com-contacter-apple.jrrdy.com,0,11.363440150607609,,1,1,1,1,600,600,0.0,0.0,0.0,1.9219280948873623
ebooking.applewf.com,0,18.532972644554473,,1,1,1,1,3600,3600,0.0,0.0,0.0,2.5216406363433186,0.5095322471047489,1,10
yourjavascript.com,0,99.73011810869362,,5,3,2,3,30,300,133.30655317392907,9517.938306462407,3172.646102154136,3.521640
electricidadobera.com,0,11.363440150607609,,1,1,1,1,14400,14400,0.0,0.0,0.0,3.219528282299548,0.3663643606263674,1,11
login.ebay.com.account-limited.8619.redhoaglandhyundai_s5_129716198.aspmodel.info,0,,,,,,,,,,,,3.0,0.9851213341419353,
login.ebay.com.account-limited.3564.chris.aspmodel.info,0,0.0,,,,,,,,,,,3.0,0.6510072618562623,-1,6
drive.google.uploadeddocx.com,0,0.0,,1,1,1,1,600,600,0.0,0.0,0.0,3.0220552088742,0.6446774004795882,-1,8
paypalverification.co.vu,0,0.0,,1,1,1,1,60,60,0.0,0.0,0.0,1.0,0.5850301939830299,1,9
signin.ebay.com.ssl-protection.5724.jimmy.aspmodel.info,0,0.0,,,,,,,,,,,3.0,0.8053896409511141,-1,7
poypal.simply-winspace.fr,0,11.363440150607609,,1,1,1,1,900,900,0.0,0.0,0.0,3.506890595608518,0.7655825019506184,-1,13
verify-apple.ml,0,,,,,,,,,,,,3.2516291673878226,0.981196000857034,0,9
www.gooogle.com,0,68.25134144531397,,6609,314,249,81,300,300,0.0,1164166.5744639637,6577.21228510714,1.842370993177108
newpaypal.uni.me,0,0.0,,4,1,1,1,300,300,0.0,0.0,0.0,1.584962500721156,0.8364938372280273,1,8
bankofamerica.com.restore-pagenkt23nbrirz.bb01abc4net.com,0,0.0,,2,2,2,2,300,14400,7050.0,8106.479711160472,4053.23985
update-secure-signin-help-inc-confirm-apple-manage.srpschapper.org,0,7.169532493946863,,1,1,1,1,14400,14400,0.0,0.0,0.
questionnairepaypal03822.110mb.com,0,0.0,,1,1,1,1,21600,21600,0.0,0.0,0.0,1.9219280948873623,0.8078908438816185,1,12
```

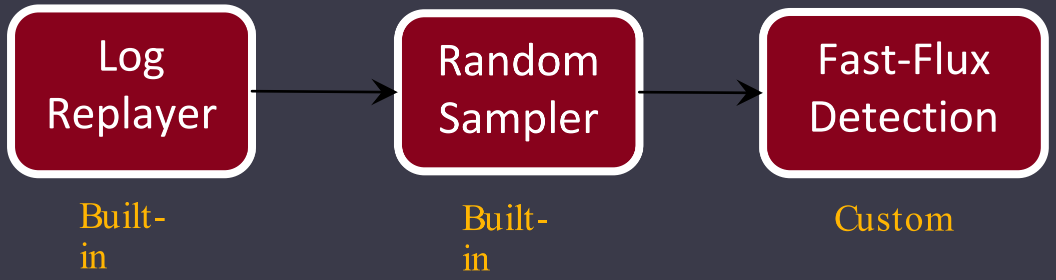

kibana.inteldb.opendns.com:9292/index.html#/dashboard/temp/AVCmOioRDm0lxYFgTOkj
IntelDB Feed Detail
Nov 1, 2015 15:28:11 to Nov 5, 2015 16:34:54
QUERY
source:opendns.nlp-rank
FILTERING
time must
field : @timestamp
from : now-7d
to : now
time must
field : @timestamp
from : "2015-11-01T23:28:11.843Z"
to : "2015-11-06T00:34:54.829Z"
FEED ACTIVITY
View | Zoom Out | source:opendns.nlp-rank (2242) count per 1h | (2242 hits)
40
35
30
25
20
15
10
5
0
16:00 11-01
00:00 11-02
08:00 11-02
16:00 11-02
00:00 11-03
08:00 11-03
16:00 11-03
00:00 11-04
08:00 11-04
16:00 11-04
00:00 11-05
08:00 11-05
16:00 11-05

Live Demo

Authlogs & Querylog Replaying

OpenDNS

Workshop : Simple Fast-Flux Detection Pipeline



OpenDNS

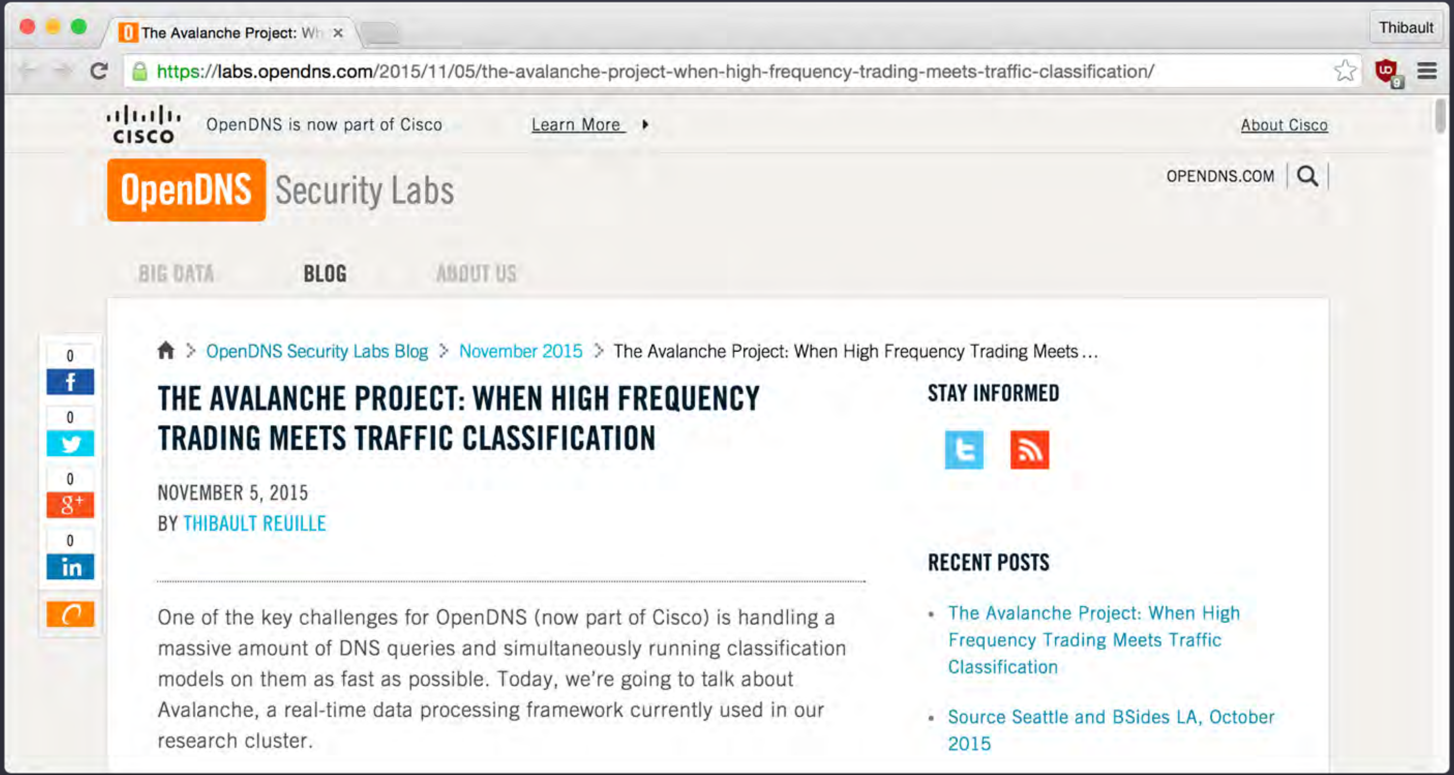
What's next?

Future Work

- More models!
- Cython or rewrite core in C/C++
  - Optimize model performance
- Use GPU grids :
  - OpenCL, GPU cluster
- Hackathon Idea :
  - Avalanche at the DNS resolver level
- More log visibility
  - Querylogs
  - Proxy logs

OpenDNS

Blog Post is Live.

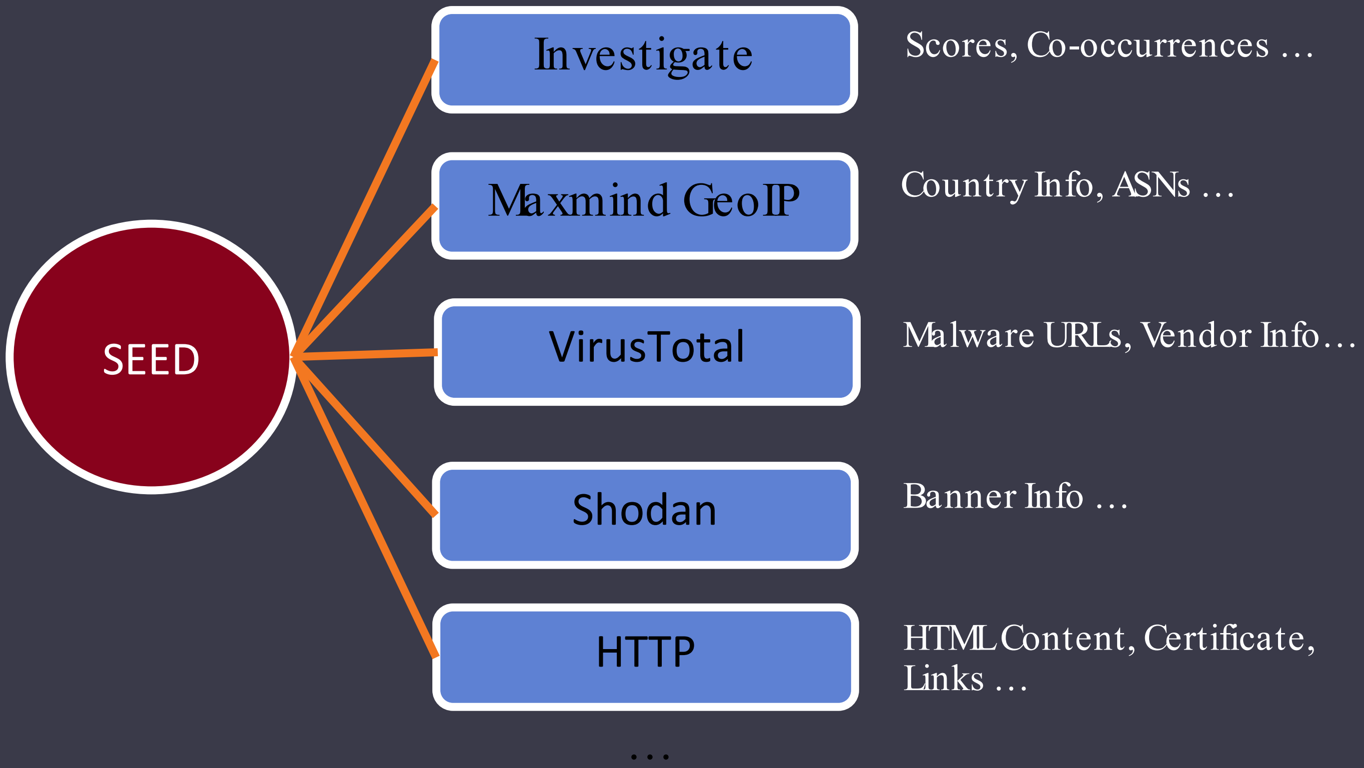


OpenDNS

Introduction to Miner/Graph-Oriented Data Mining

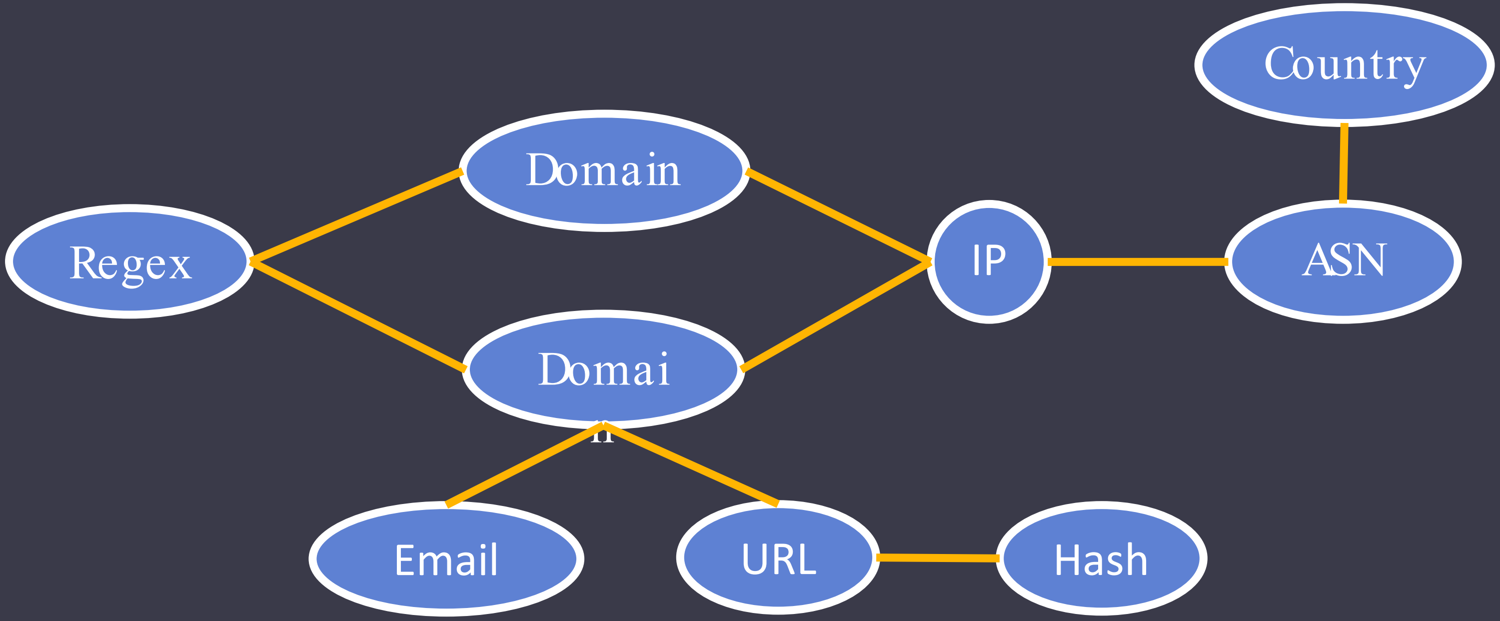
Interesting Data Sources …

- Domain
- URL
- IP
- ASN
- Hash
- Email
- Regex



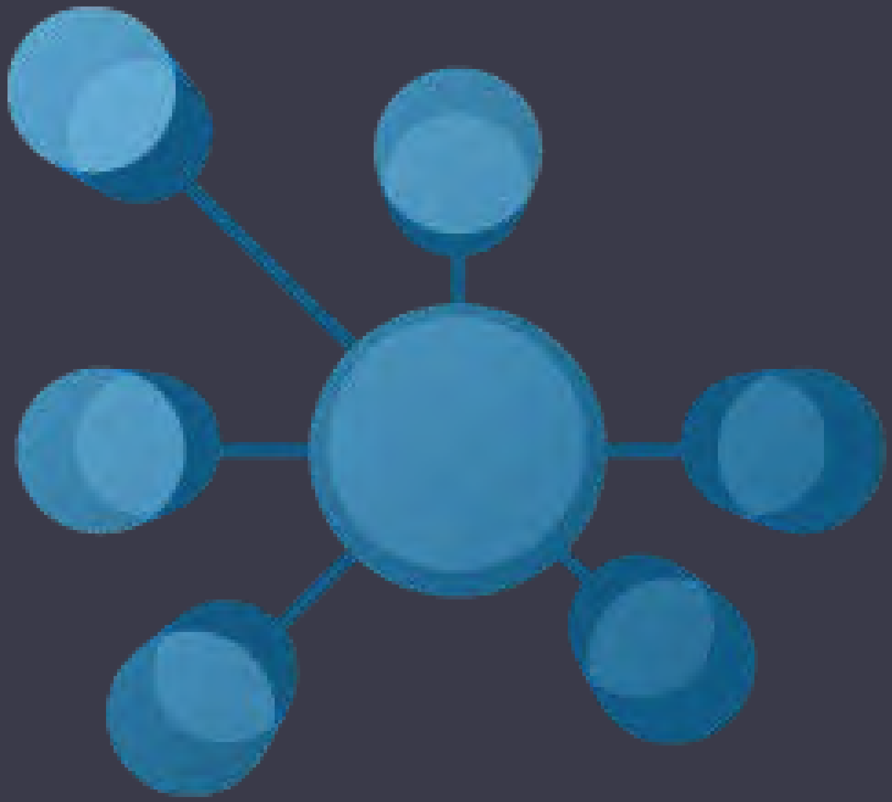
OpenDNS

Data Modeling Example

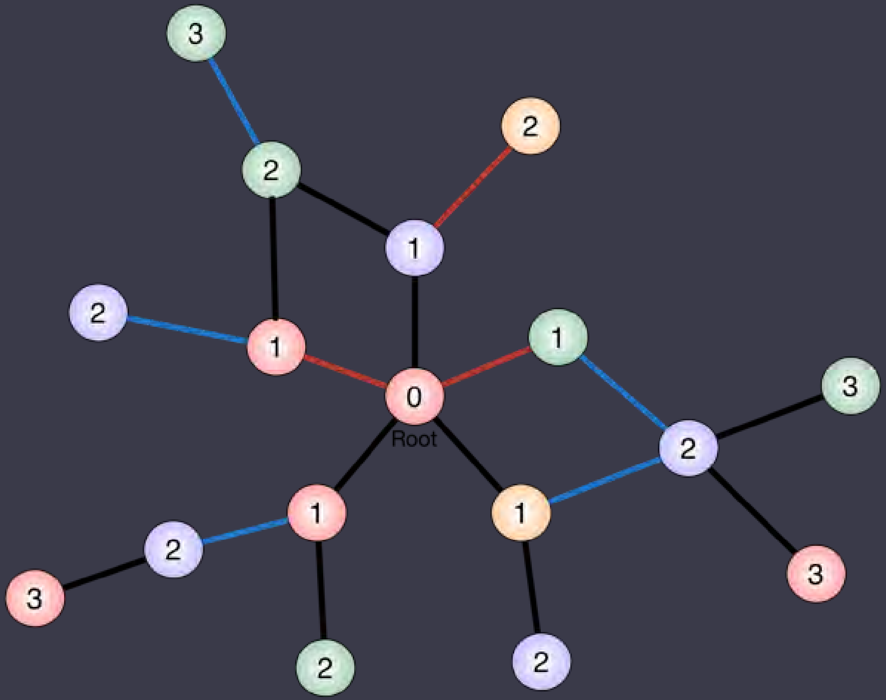


Knowledge

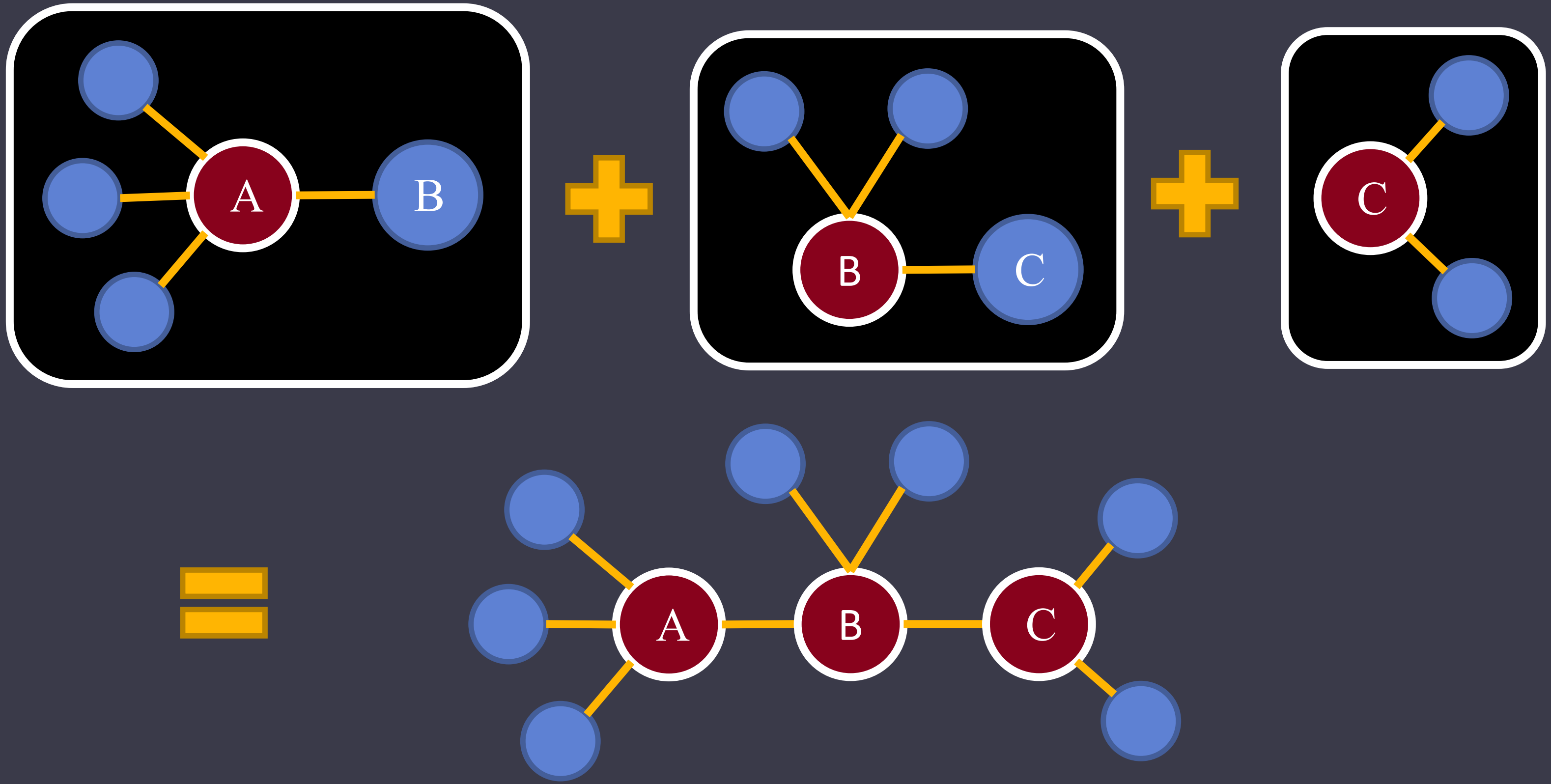
- Semantic Networks / Property Graph
- Node = Concept, Edge = Relationship
- Model of the Information
- Ontology : Model of the Model



Data Exploration : Breadth First Traversal

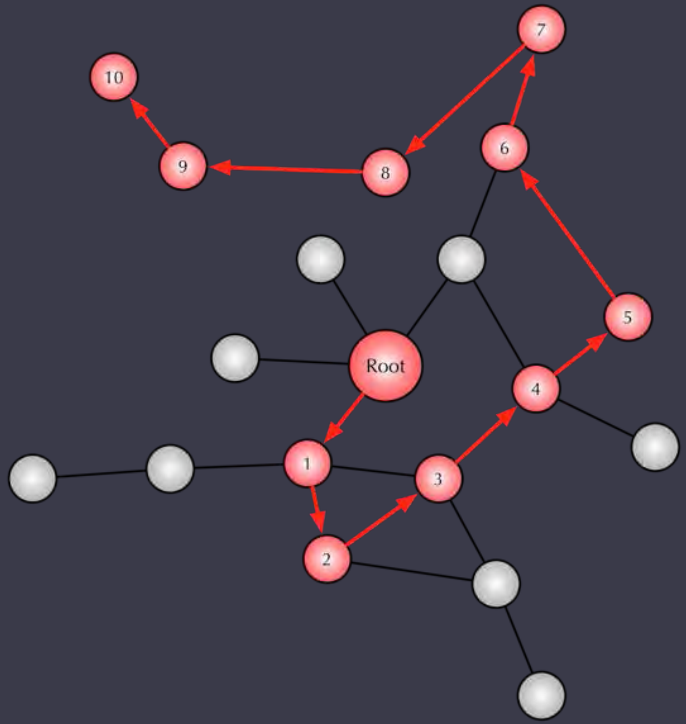


Multi-Threaded Breadth First Traversal



OpenDNS

Lambda Mining



- Functional Graph Exploration
- Rule Based / Thresholds / Topology based …
- Profiles for specific use cases
- Automated Smart Data Mining

NLPRank/Phishing Detection

OpenDNS Security Labs

Data Science ∩ Network Security

Big Security Data-

DNS Traffic:

~70B DNS requests per day

HTTP Traffic:

~10.1Mrequests per day

Daily Tasks:

-Detection Algorithms, Security Data Analysis, Distributed Systems, Big Data Engineering, Data Viz

Purpose:

Overview of our new model **NLPRank**:

-Fraud detection system using NLP techniques and traffic features to identify domain-squatting/brand spoofing in DNS/HTTP (a technique commonly used by phishing and APT CnCs).

#TeamPython

Natural Language Analyses with NLTK

NLP/Data Science:

-NLTK
-Scikit-Learn
-Gensim

Web Scraping:

-Beautiful Soup
-LXML

scikit learn

gensim
Gensim home
topic modelling for humans

System Origins

-OpenDNS Labs has detection models for commodity malware (ex. Botnet, Fast-Flux, DGA) need a model to detect targeted attacks

-Assigned to analyze DarkHotel data set

Question: How to detect "evil" in DNS records using lexical features of FQDN and <u>validate</u> results?

Human-Computer Interaction

Targeted Attacks: From a psychological perspective, if you were a high-profile exec for company what kind of links would you click on? What are your interests?

Commodity Phishing: Same psychology

Topics of interest:

-$$$, Bank Account/CCs, Financial

-News

-Security/Software updates

-Social Network

CHASE

PayPal

Bank of America

WELLS FARGO

Google

YAHOO!

AMERICAN EXPRESS

Heuristic #1 - ASN Filtering

OpenDNS

ASN Overview

-Autonomous System Number is basically like your neigborhood/zipcode on the internet
-Associated with Internet Service Provider
-Set of routers operating under specific or multiple routing protocol
-Domains exhibiting fraudulent behavior are observed to be hosted on ASN’s that are unassociated with the company they’re spoofing

OpenDNS

Examples

Expect a FQDN containing “adobe” to be associated with Adobe’s ASN (ex. ASNs 14365, 44786, etc.), or FQDN containing “java” and advertising an “update” be associated with Oracle ASN (ex. 41900, 1215, etc.)

So why then?

APT Example (Carbanak):

-adobe-update[.]net - ASN 44050, PIN-AS Petersberg Internet Network LLC in Russia

-update-java[.]net - ASN 44050, PIN-AS Petersberg Internet Network LLC in Russia

Commodity Phishing Examples:

Domain: securitycheck.paypal.com

ASN 20013, CYRUSONE -CyrusOne LLC, US

Domains: serviceupdate-paypal.com, updatesecurity-paypal.com,

ASN 32400 - Hostway Services, Inc.,US

The Usual Suspects..

1. CyrusOne LLC,US
2. Unified Layer,US
3. OVH OVH SAS,FR
4. GoDaddy.com, LLC,US
5. HostDime.com, Inc.,US
6. SoftLayer Technologies Inc.
7. HOSTINGER-AS Hostinger International Limited,LT
8. HETZNER-AS Hetzner Online AG,DE
9. Liquid Web, Inc.,US
10. CLOUDIE-AS-AP Cloudie Limited-AS number,HK

More Normalized...

1. OBTELECOM-NSK OOO Ob-Telecom, RU
2. GVO - Global Virtual Opportunities, US
3. CONFLUENCE-NETWORK-INC - Confluence Networks Inc, VG
4. CYRUSONE - CyrusOne LLC, US
5. VFMNL- AS Verotel International B.V., NL
6. NEOLABS- AS Neolabs Ltd., KZ
7. DEEPMEDIA- AS Deep Media / V.A.J. Bruijnes (sole proprietorship),NL
8. NEUSTAR- AS6 - NeuStar, Inc., US
9. VERISIGN- ILG1 - VeriSign Infrastructure & Operations, US
10. CIA- AS Bucan Holdings Pty Ltd, AU

ASN Filter + Whitelisting

1st step to take a big chunk out of the traffic, because text processing is computationally intensive

-Do a lot of ASN Analysis with other models (Dhia Mahjoub, PhD Graph Theory)

Authlogs come in -> Enricher node will look up ASN and include logs

- Create mapping of Brand Names to their legitimate ASNs
- Lookup domains/IPs as they come in

OpenDNS

Heuristic #2 - Defining Malicious Language Within FQDNs

Building Intuitions

-Eyeball Data

-Run basic text metrics on the data, gain intuitions about the data and extract important words/substrings in APT FQDN datasets

-APT domains exhibit similar lexical features to commodity phishing domains

-Important look at word co-occurrences (bigrams, trigrams, etc.)

OpenDNS

Building Intuitions

-From APT data sets extracted words from dictionary and applied stemming looking at word stats:

Top counts (stemmed): mail, news, soft, serv, updat, game, online, auto, port, host, free, login, link, secur, micro, support, yahoo

Bigram Collocations:

Words that often appear with each other

adobe-update

update-java[.]com

Idea:

brandname + ad-action word [.] tld

OpenDNS

Examples

Dark Hotel (Kaspersky):

- adobeupdates[.]com
- adobeplugs[.]net
- adoberegister[.]flashserv[.]net
- microsoft-xpupdate[.]com

Carbanak (Kasperksy):

- update-java[.]net
- adobe-update[.]net

APT 1 Domains (Mandiant):

- gmailboxes[.]com
- microsoft-update-info[.]com
- firefoxupdata[.]com

OpenDNS

NLP on FQDN

-Creating a “malicious language” derived from lexical features of FQDNs from APT/Phishing data sets
-Built corpus of domains similar to examples in previous slide
-Create custom dictionaries
Brandname Dictionary
Ex. google, gmail, paypal, yahoo, bankofamerica, wellsfargo
-Custom set of stemmed common malicious words
Ex. secur, updat, install, etc.
-Reason for stemming example: updat -> firefoxupdata[.]com (APT1)
-Apply Edit-Distance/Automata Theory on substrings to build spam language

OpenDNS

Heuristic #3- HTML Content Analysis

Recreating Researcher's Mind

When reviewing malicious domains what is typical methodology for review:

1) Visit site in Tor browser
2) Researcher processes information on site, looks for clues, gains summary
3) Makes decision whether site is legit/malicious

Specifically for Phishing Sites:

Human-Computer Interaction: What makes people fall for this?

Site will be near copy of legitimate site it's intending to spoof

How can we automate this process?

Can we apply document similarity algorithms?

Human-Computer Interaction

Examples from Apple Phishing page:

Title: Apple GSXLogin

Links:

https://iforgot.apple.com/cgi-bin/findYourAppleID.cgi?language=US-EN&app_id=157&s=548-548

https://id.apple.com/IDMSAccount/myAccount.html?appIdKey=45571f444c4f547116bfd052461b0b3ab1bc2b445a72138157ea8c5c82fed623&action=register&language=US-EN

Images:

Other Clues:

HTTrack - tool used to clone site

```
<!DOCTYPE HTML><html lang="">

<!-- Mirrored from tools.google.com/dlpage/drive/index.html by HTTrack Website Copier/3.x [XR&CO'2014], Tue, 23 Sep 2014 08:58:40 GMT -->

<!-- Added by HTTrack --><meta http-equiv="content-type" content="text/html;charset=utf-8" /><!-- /Added by HTTrack -->

<head><script type="text/javascript">

function utmx_section(){}function utmx(){}
```


OpenDNS

Preparing The Data

-Cleaning the Data

- -Stripping punctuation, symbols, unnecessary content
- -Normalizing the data
  - -Stemming (update, updating, updater →updat)
  - Feature Encoding

```
© Google    •
<a href="https://www.google.com/intl/en/policies/privacy/">
  Privacy Policy
</a>
```

Harder than it seems...

-Non-Trivial to extract relevant terms from HTML documents
-Dealing with malformed tags
-Lose data, dealing with HTML and JS
-Which tags to encode?
    -Title
    -Links
    -Images

Applied basic NLP Algos..but need more samples for training!!

OpenDNS

More Headaches

Legit USAA Site:

<title>USAA Military Home, Life & Auto Insurance | Banking & Investing</title>

Many USAA Phishing Sites:

<title>USAA Military Home, Life & Auto Insurance | B
Investing</title>

USAA Phishing Page:

<title>USAAMilitary Home, Life & Auto Insurance</title>

Success Identifying All Different Types of Attacks

Success in Training:

Detecting:

Careto

APT Domains Darkhotel/Carbanak/APT1 etc.

AJAX Hacking Group/Flying Kitten infostealer C&C

Operation Pawn Storm

Operation Saffron Rose

and more...

Success on Live Data:

Exploit Kit

Fast-Flux

And new stuff..

Interesting Results

Carbanak (banking trojan) came out in February:

2015-01-23 14:52:58 -- a96e74b8-b052-4f42-a517-d7273d4f13e7

NLPRank High-Risk Results (FQDNs)

cdneu.windows8downloadscdn.com
update-java.net

OpenDNS

Interesting Results

symantecupdates.com

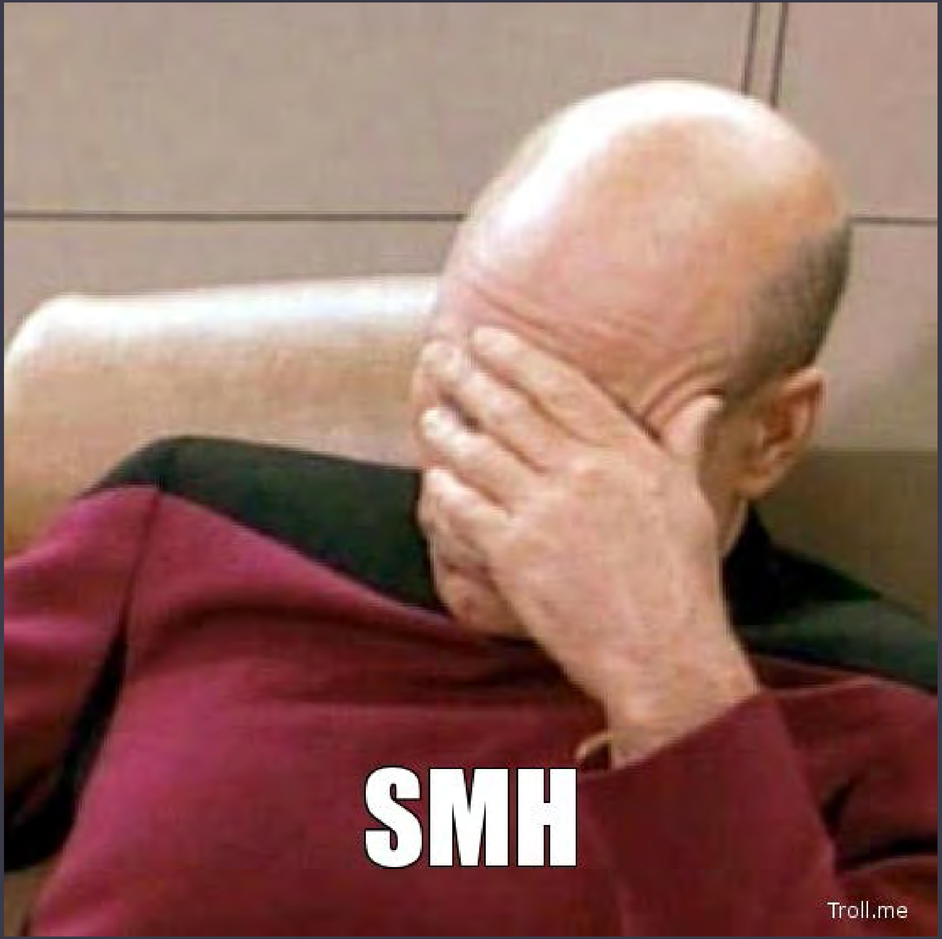
Whois information

Registration date	2013-09-03 00:00:00 +0000
Registrar name	GODADDY.COM, LLC
Registrant	li ning < li2384826402@yahoo.com >
Registrant contact address	guangdongsheng guangzhoushi Alabama UNITED STATES

Li Ning From guangdongsheng guangzhoushi Alabama???
Let's investigate all domains associated with that email address...

21,533 Domains???

crowcasinovip.biz mybestbrand.biz mybestbrands.biz huarenceluewangzhi.com icbczay.com boyinbocai5.com
haoyunc3.com bocaiwangzhenqianpingtai.com zuqiubocaiwangzhan7.com weinisirenyulecheng94.com
xinquanxunwang244.com dfjdh.com yaojiyulecheng9898.com wanbaoluyulecheng94.com xinpujingyule15.com
toabaao.com jinbaiyiyulecheng26.com toubakd.com tiantianleyulecheng61.com wangziyulecheng33.com
yezonghuiyulecheng82.com bocwry.com huangguantouzhuwangzhanwangzhi86.com huangguanwangquaomen29.com
haiwangxingylc1664.com yinghuangylc727.com bocaiasd.com changjianggjylc.com jinmaylcoiu.com
yazhougjylc.com huangguanxin2wang32.com benchixsyl.com zhucecaipiaosongcaijin.com ceoylcdf.com
zhucesongcaijindewangzhan62.com aomenduchangyouxiyounaxie30.com mengtekaluoylcb.com
guojihuangguanyule40.com huangguantiyupingtai93.com huangguanxianjinwangxinyu37.com
aomenduchangpaixing27.com 500wanylcyu.com dajihuiylc686.com ruifengguojiyy.com makeboluoylcb.com
jincaigjylc.com xindongfangylc869.com aomenduchangzainali50.com wangshangyulekaihusongcaijin.com
huangguanxjwkh.com jinbangylc77.com baijialeqo.com yataigjylc.com baishenggjylcwe.com bocaigongsiqe.com
wufagjylc.com moerbenylckk.com bogouylc1663.com huangguandailiwangzhi23.com bojueylcpo.com
bocaiwangzhanqe.com taoataao.com bbhunas.com sjzd36.com sjpt63.com bjlkh33.com
baijialebishengtouzhujiqiao20.com xijialiansaijifenbang57.com baijialeyule86.com xijiapaiming46.com
aomenbaijialechangying76.com baijialeyulepingtai34.com wangshangbaijialekaihusongcaijin76.com
ouzhouwudaliansaipaiming53.com wudaliansaitedian39.com baijialekaihusong50caijin17.com baijialeguize52.com
zhibobazuqiuzhibo2.com zuqiubifenqiutan88.com dejiasaichengbiao88.com zuqiuba85.com mahuitqzzjw83.com
sjzd01.com weixingjianting29.com cwanpp.com xingboyulezaixian86.com mwqpah.com
jiankongpingtairuanjian43.com zhenqianyulechengguanwang63.com njdyyytj.com fanheer.com 999coin.com
shenganna74.com jackwolfskinsalejp.com zaozhuangcq.com bjl7788.com ruhejiankongshouji2.com
aomenduchangyingqianliao75.com shoujidingweichaxunruanjian12.com shoujijiantingshebei46.com aomen916.com
shoujikajiantingqi77.com zhenqianyouxipaixing2.com rysevw.com wanzhenqianwangzhan36.com vrcgw.com
feilvbinshengannayulecheng20.com duchangyingqianmijue81.com zzvqo.com



OpenDNS

Sakula/ThreatConnect Report

Domain Name: TOPSEC2014.COM	Domain Name: TOPSEC2014.COM
Registry Domain ID: 1857525015_DOMAIN_COM-VRSN	Registry Domain ID: 1857525015_DOMAIN_COM-VRSN
Registrar WHOIS Server: whois.godaddy.com	Registrar WHOIS Server: whois.godaddy.com
Registrar URL: http://www.godaddy.com	Registrar URL: http://www.godaddy.com
Update Date:	Update Date: 2014-05-06 04:52:21
Creation Date: 2014-05-06 04:48:49	Creation Date: 2014-05-06 04:48:49
Registrar Registration Expiration Date: 2015-05-06 04:48:49	Registrar Registration Expiration Date: 2015-05-06 04:48:49
Registrar: GoDaddy.com, LLC	Registrar: GoDaddy.com, LLC
Registrar IANA ID: 146	Registrar IANA ID: 146
Registrar Abuse Contact Email: abuse@godaddy.com	Registrar Abuse Contact Email: abuse@godaddy.com
Registrar Abuse Contact Phone: +1.480-624-2505	Registrar Abuse Contact Phone: +1.480-624-2505
Domain Status: ok	Domain Status: clientTransferProhibited
	Domain Status: clientUpdateProhibited
	Domain Status: clientRenewProhibited
	Domain Status: clientDeleteProhibited
Registry Registrant ID:	Registry Registrant ID:
Registrant Name: li ning	Registrant Name: Top Sec
Registrant Organization:	Registrant Organization: TopSec
Registrant Street: guangdongsheng	Registrant Street: china
Registrant City: guangzhoushi	Registrant City: china
Registrant State/Province: Alabama	Registrant State/Province: china
Registrant Postal Code: 54152	Registrant Postal Code: 100000
Registrant Country: United States	Registrant Country: China
Registrant Phone: +1.4805428751	Registrant Phone: +1.82776666
Registrant Phone Ext:	Registrant Phone Ext:
Registrant Fax:	Registrant Fax:
Registrant Fax Ext:	Registrant Fax Ext:
Registrant Email: li2384826402@yahoo.com	Registrant Email: TopSec_2014@163.com

More BlueCross/Premera

Found these:

adobeupdated[.]com
gmail-msg[.]com
intel-update[.]com
vmwaresupportcenter[.]info

Didn't catch these but definitely capable:

prennera[.]com
we11point[.]com.

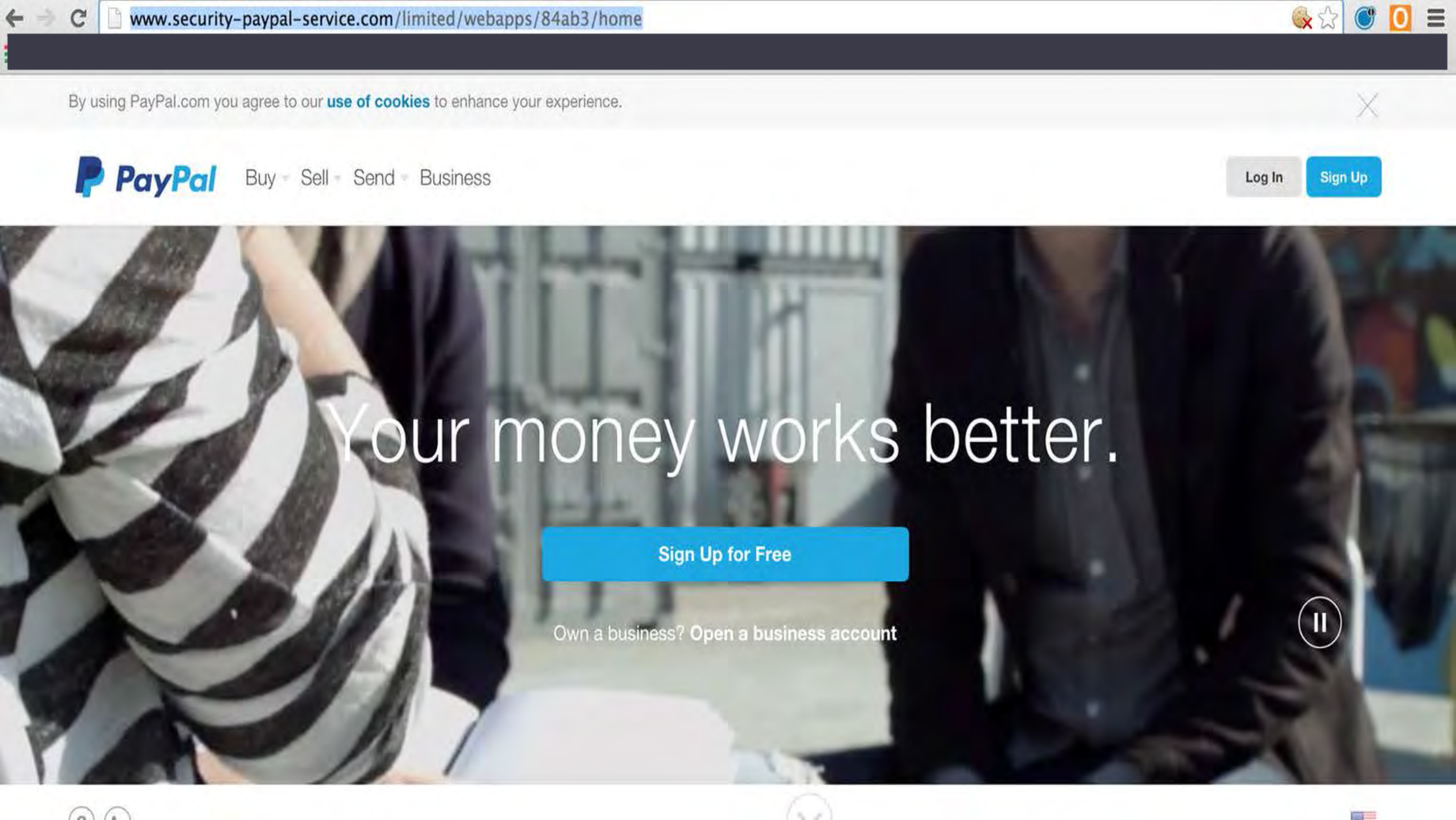
OpenDNS

Interesting Results

Way to filter into parked/suspended pages??

1. Parked Pages
   a. Interesting patterns among terms of parked pages, examples:
      i. www[.]iniciar-sesion-gmail[.]com
         1. Important Terms (stemmed) : fjccheck1catchexcept, click, trydocumentcooki, proceed
      ii. ww2.content.archiveofourown.orgamazon.com
         1. Important Terms (stemmed) : fjccheck1catchexcept, click, trydocumentcooki, proceed
      iii. android.clients.google.com.www.smartbrosettings.net,
         1. Important Terms (stemmed) : fjccheck1catchexcept, click, trydocumentcooki, proceed
2. Suspended Pages
   a. "Suspend" relayed as most important terms, example:
      i. FQDN: xbmcwindows[.]com
         1. Important Terms: **'suspend'**,'arial', normal, solid'

OpenDNS


www.security-paypal-service.com/limited/webapps/84ab3/home
By using PayPal.com you agree to our use of cookies to enhance your experience.
PayPal
Buy
Sell
Send
Business
Log In
Sign Up
Your money works better.
Sign Up for Free
Own a business? Open a business account


facebooklogin-facebook.com
Your computer will restart to
complete these updates.
Restart
facebook Login
Email or Phone
Password
Log In
Keep me logged in
Forgot your password?
Login on Facebook
Google
Twitter
Yahoo
Hotmail

Combining Detection Models

OpenDNS Security Labs Blog > September 2015 > Phishing, Spiking, and Bad Hosting

PHISHING, SPIKING, AND BAD HOSTING

SEPTEMBER 14, 2015
BY DHIA MAHJOUB, JEREMIAH O'CONNOR, THIBAULT REUILLE AND THOMAS MATHEW

At OpenDNS Labs we have developed a number of predictive models to hunt down evil on the Internet. We have discussed in previous blogs and conferences our algorithms NLPRank [1][2][3], Spike detector [4][5][6], and malicious IP space/rogue host detectors [7][8](section 14)[9][10][11][12][13][14][15].

In this blog we will discuss how we integrate all of these detection models to improve detection coverage of current threats and walk through a few interesting examples.

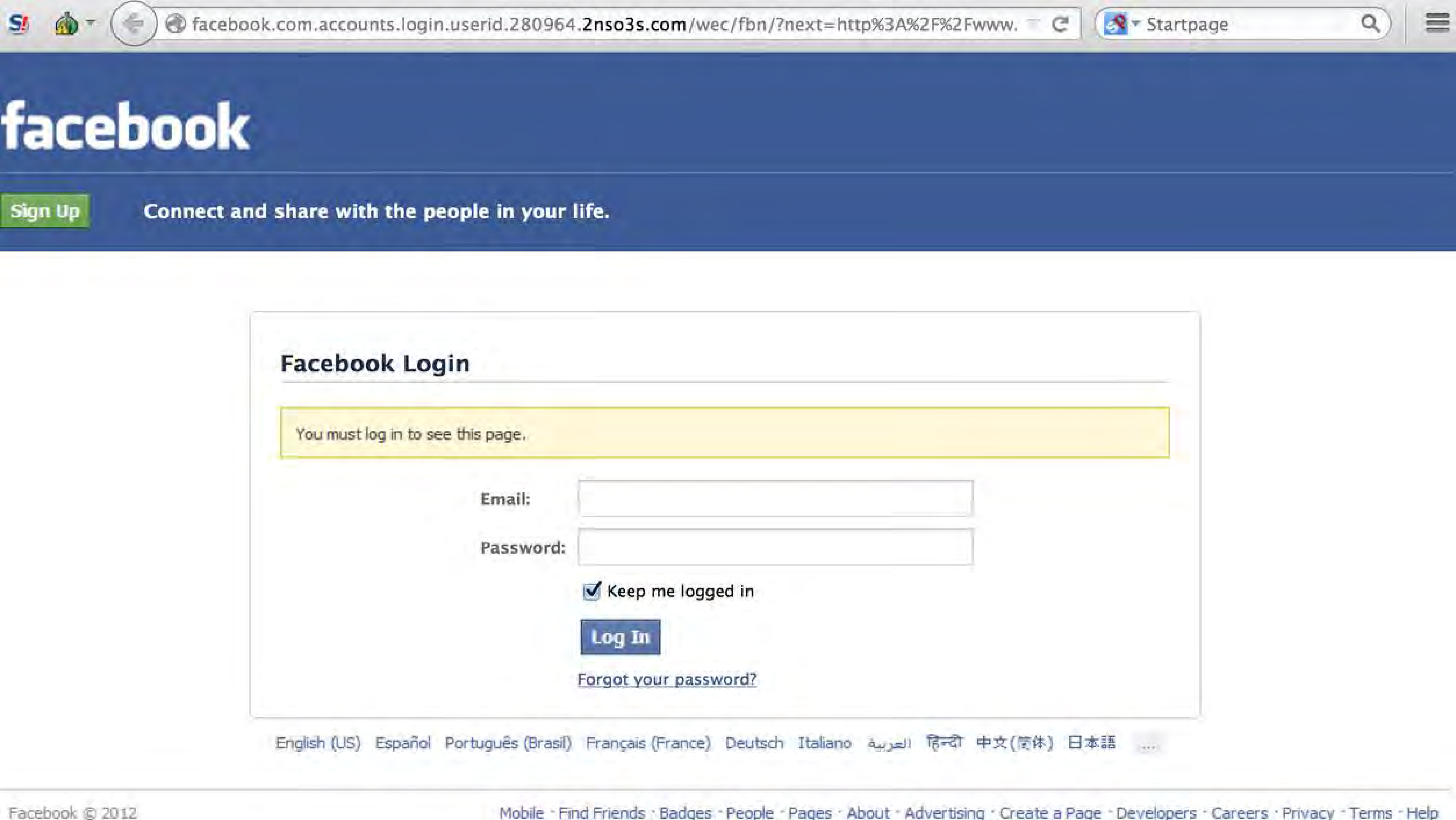
PHISHING AND SPIKES

One of the recent samples we have found was a Facebook phishing campaign that was surfaced by our real-time alert system. Our model NLPRank detected the campaign of Facebook phishing sites spoofing Facebook under the second-level domain (2LD) 2nso3s[.]com.

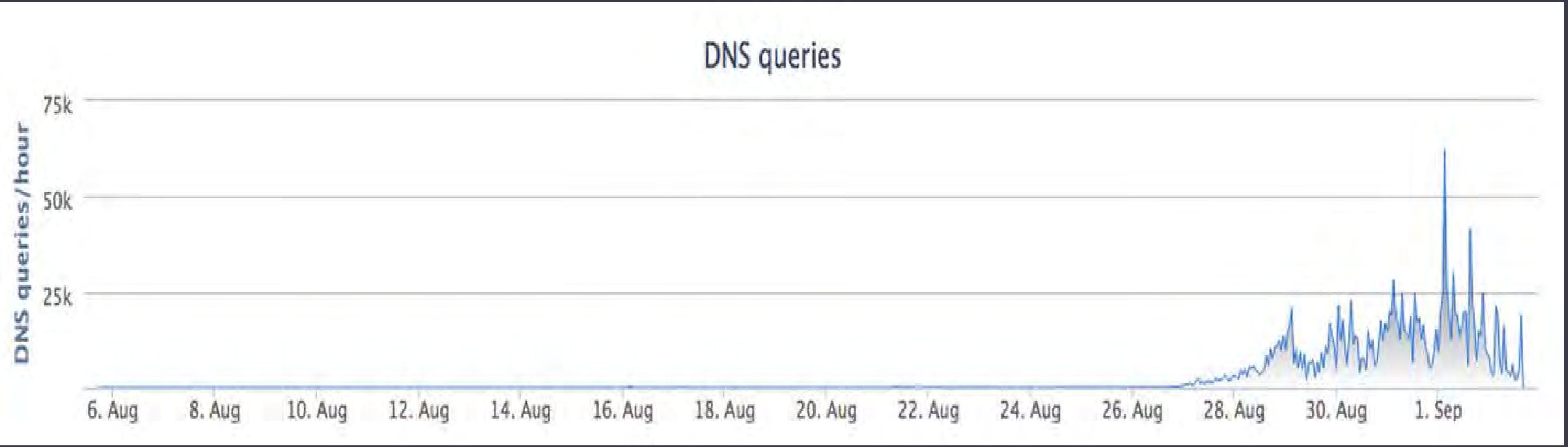
For this particular domain, when visiting the 2LD, 2nso3s[.]com from your browser, you would be directed to a URL that looks like:

http://facebook[.]com.accounts[.]login[.]userid[.]280964[.]2nso3s[.]com/we
next=http%3A%2F%2Fwww.facebook.com%2videos%2F%3A%4A%4ID%1/

As we can see in the path of the URL the next page routes you directly to


facebook.com.accounts.login.userid.280964.2nso3s.com/wec/fbn/?next=http%3A%2F%2Fwww.
Startpage
facebook
Sign Up
Connect and share with the people in your life.
Facebook Login
You must log in to see this page.
Email:
Password:
Keep me logged in
Log In
Forgot your password?
English (US) Español Português (Brasil) Français (France) Deutsch Italiano العربية हिन्दी 中文(简体) 日本語 ...
Facebook © 2012
Mobile · Find Friends · Badges · People · Pages · About · Advertising · Create a Page · Developers · Careers · Privacy · Terms · Help

Traffic for 2nso3s.com

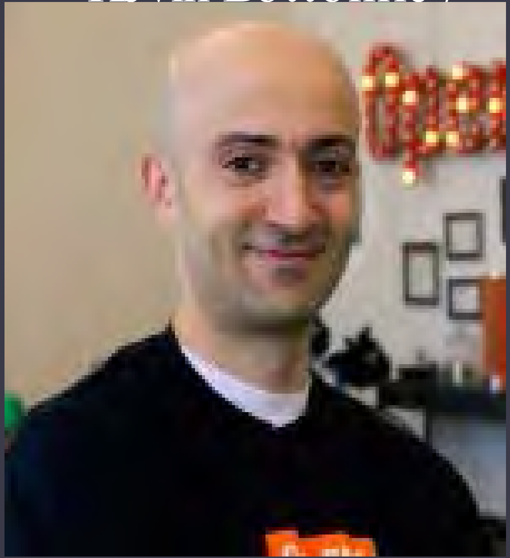


PhishTank® Out of the Net, into the Tank.

Vinny Lariza

Kevin Bottomley

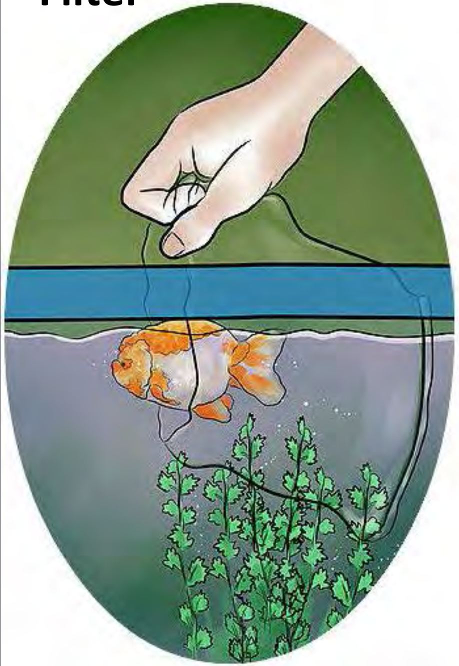
Dhia Mahjoub

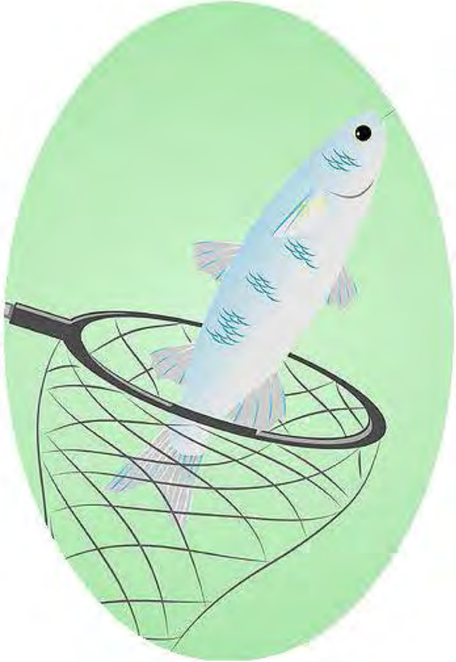


OpenDNS

How Phishtank Works

Submit --------------------> Vote! -------------> Categorize ----------------> Filter





Identifying Problem

-PhishTank has Cult Following in Security Community

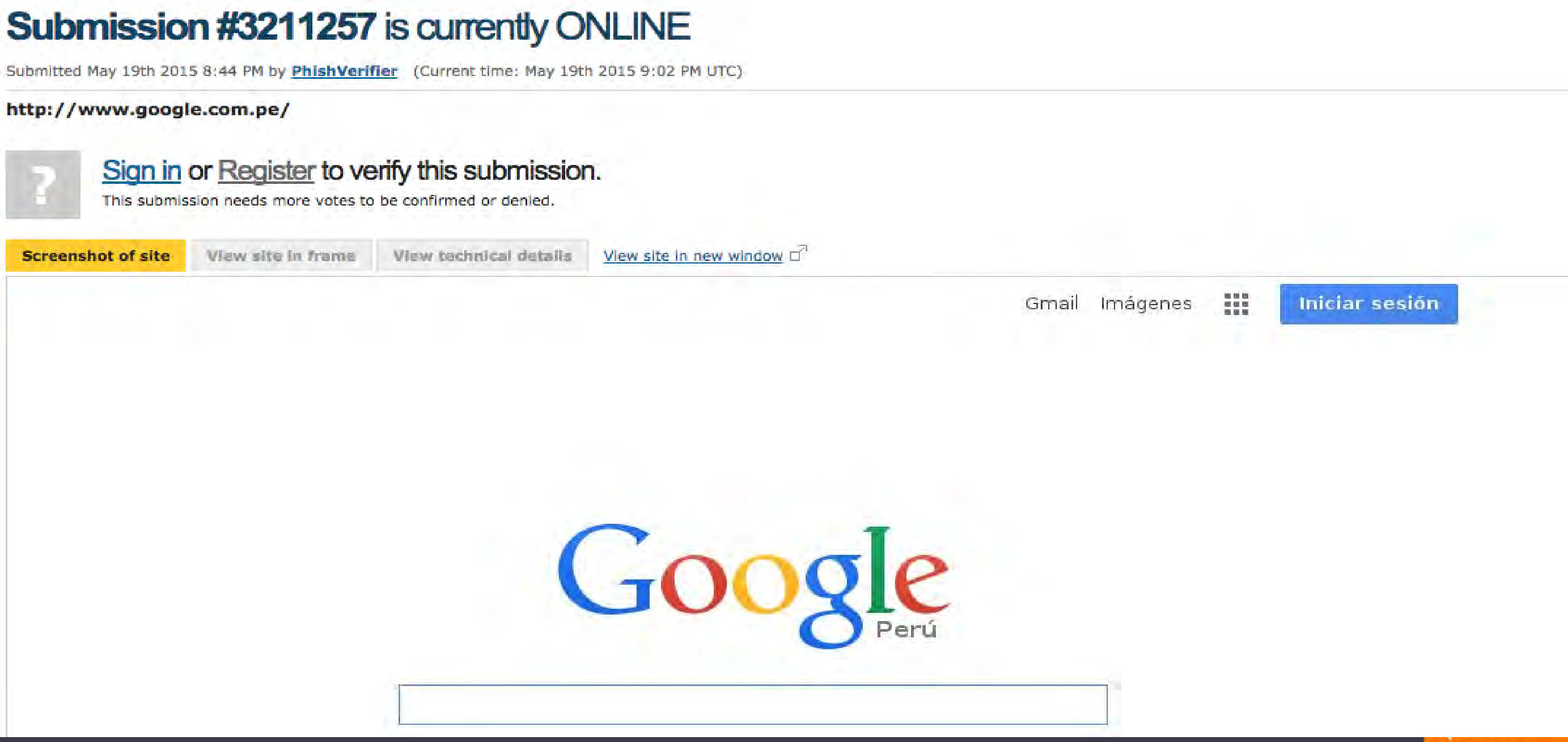
- People always asking about it conferences, security parties, LinkedIn etc.

-Identifying spoofed brands of phishing URL's in real-time / as they are submitted is necessary for reducing the amount of false positives in the PhishTank feed

-Reducing the amount of time from submission to approval

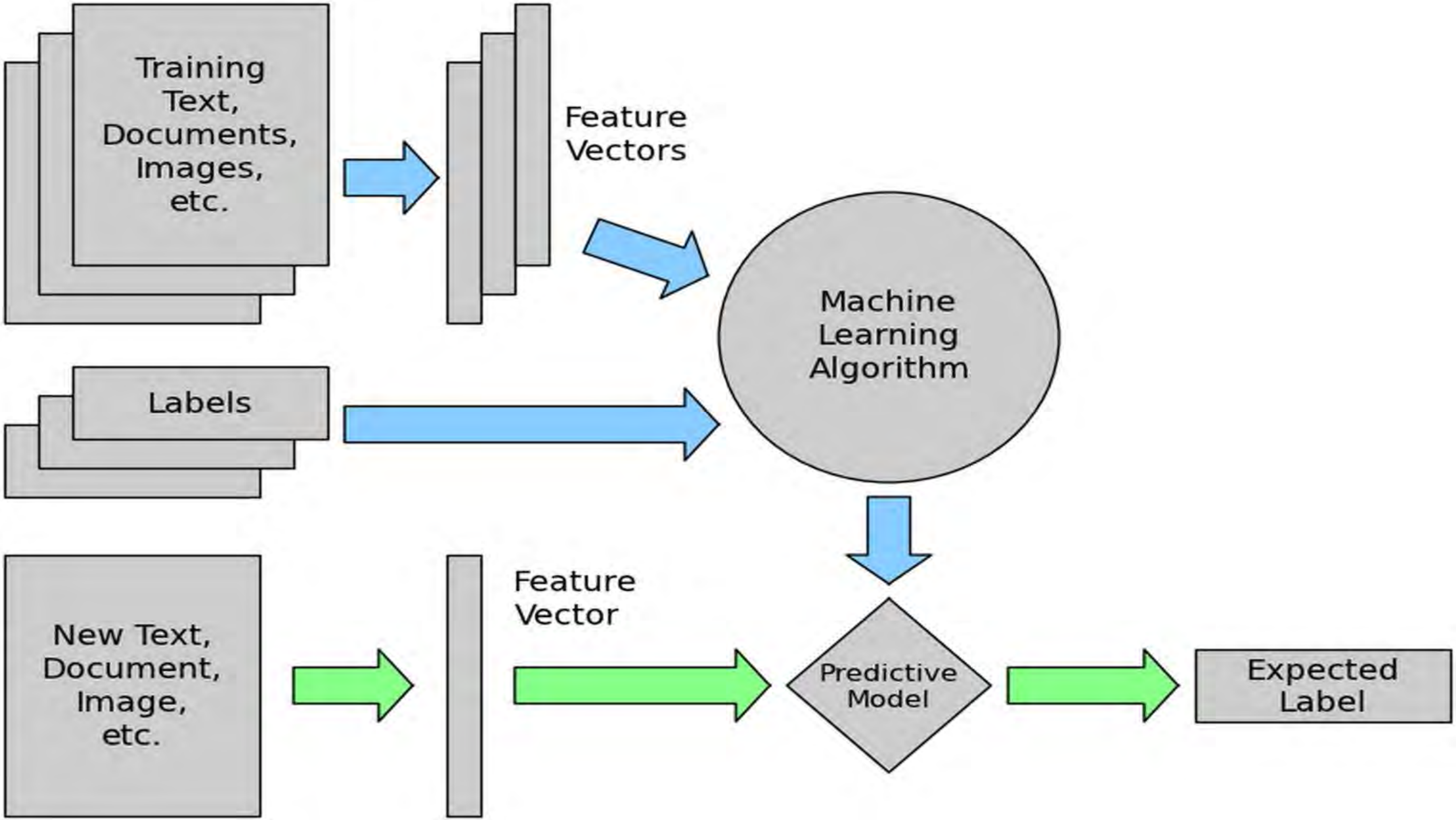
-IMO: Phishtank= giant training set for sec data scientists

Examples of False Positives



Hypothesis:

-Using IR/NLP techniques to gain a summary of the web page is a problem that has already been solved algorithmically ex. search
-Similar to way Netflix recommends movies based on user history, can we recommend what brand name the phish is by content of the page?
-Lets apply these same techniques to identify commodity phishing pages
Hypothesis: We can identify Phishing pages by using IR/Topic Modeling techniques, and auto-label Phishtank submissions as they come in


Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Topic Modeling

-Methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Discover the hidden themes of collection.
2. Annotate the documents according to themes.
3. Use annotations to organize, summarize, search, make predictions.

-Great for building recommender systems

-Used as features for a classifier

Building Corpus

-Built Corpus of HTML Content of Phishing pages, ex. WellsFargo, Paypal, Amazon, Apple, Bank of America, from Phishtank
Only Focused on Big Name Brands
- Data Collection, although at times tedious, become very intimate with the data
-See all kinds of variations of Phishes
90s Paypal vs. 2000s Paypal vs. 2015 Paypal
Christian Mingle Phishing?

TF-IDF

Input: Word Count Vector From Terms in HTML Document (Query), Word Count Matrix over a collection (Corpus)

TF-IDF - Show how important word is to a collection

Balance between: Frequency of Term and Rarity over all documents

Term-Frequency: # of times term t, appears in the document d

-Term Relevance does not increase proportional with term-frequency

Inverse-Document Frequency: the # of documents that contain term t

TFIDF - tf-weight * idf-weight

TFIDF - Increases with number of occurrences within a document, and rarity of term over all documents

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

LSA/LSI

Latent Semantic Analysis: analyzing documents to find underlying concepts/meaning from them (clustering algorithm)

Uses singular value decomposition (reduce dimensionality) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text.

Hard because of variations in English language, synonyms, ambiguities

some words have different meanings when used in context

-Uses Bag of Words Model (Ordering doesn't matter)

-Using n-grams can help identify associations using co-occurrences

Helps with normalization of data

Bigrams: San Francisco -> san_francisco, Sign In -> sign_in

LSA/LSI

Input: X, count matrix (or TFIDF), where m (rows) is number of terms, and n is number of documents

When we do decomposition, have to pick a value k, which represents the number of topics/concepts

Process: Decompose X into 3 matrices, U, S, V(T)

U = m x k matrix, where m = terms, k = concepts

S = k x k diagonal matrix. Elements are amount of variation … t.

V(transpose) = k x n matrix, where k = concepts, n = docume…

$$X \approx USV^T$$

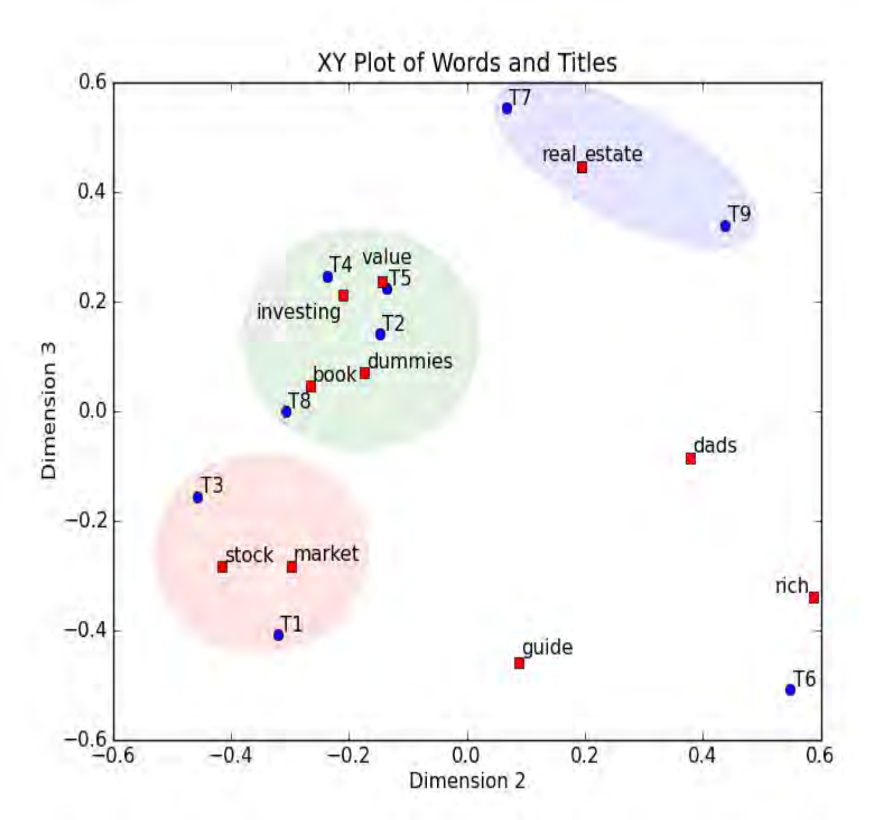
OpenDNS

LSA/LSI Example

[redacted]

1. [redacted]
2. [redacted]
3. [redacted]

1. [redacted]



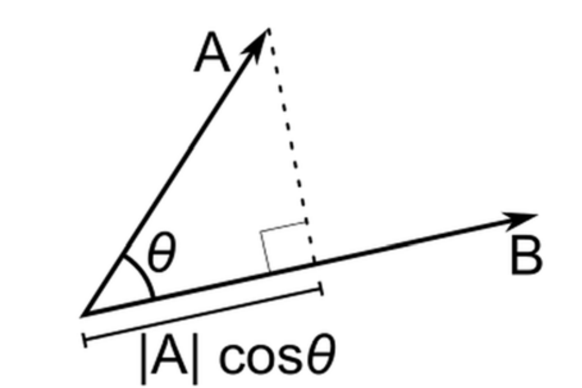
Cosine Distance

Word counts of the documents (HTML Content) form vectors
Cosine is normalized dot product of the vectors
Compute Cosine Distance from the components of the 2 vectors

i. Cosine Similarity to Phishing Pages in the Corpus
    1. Transform terms of HTML document into vectors and Corpus (Phishing) documents to vectors
    2. Find angle (Cosine Similarity) between input HTML document term vector and Corpus documents
    3. Return ranking of the sites with the most similar HTML Documents in Corpus

OpenDNS

Cosine Distance b/t Vectors



Cosine distance between two vectors:

In[1]:= CosineDistance[{a, b, c}, {x, y, z}]

Out[1]= $1 - \frac{a\,x + b\,y + c\,z}{\sqrt{\mathrm{Abs}[a]^2 + \mathrm{Abs}[b]^2 + \mathrm{Abs}[c]^2}\ \sqrt{\mathrm{Abs}[x]^2 + \mathrm{Abs}[y]^2 + \mathrm{Abs}[z]^2}}$

Auto-Labeling Brand Results:

Sample Output (Document Handle, Document (Cosine) Similarity Score, Brand/FQDN of URL):
Input URL/Query: WellsFargo/fitac.com.tr.html
(61, 0.99899197) WellsFargo/wellsfargo.com.html
(62, 0.99890876) WellsFargo/usam.edu.sv.html
(60, 0.9984659) WellsFargo/school76.irkutsk.ru.html
(59, 0.98146677) WellsFargo/theweddingcollection.gg.html
(63, 0.97453147) WellsFargo/exin.ba.html
Input URL/Query: Chase/www.nutrem.mx.html
(76, 0.98566723) Chase/bororooil.com.html
(75, 0.92363083) Chase/chaseonline.chase.com.html
(27, 0.92042124) BankOfAmerica/createcrafts.ph.html
(25, 0.92009199) BankOfAmerica/actautismoman.com
(74, 0.91776139) Chase/www.zac.or.tz.html

OpenDNS

Auto-Labeling Brand Results:

Sample of Brand Names from Incoming Phishtank Stream
467 Total Samples - 78 in Corpus, 389 Test
353 hitting as Top recommendation, 18 out of remaining 36 in Top 5
Still along the same Topic/Theme, ex. (Bank/Finance, Mail, Social)
371 / 389 (With additional weighting tests, work in progress)
Some Brands have higher accuracy than others (Wells Fargo, BofA)

Auto-Labeling Brand Results:

ACCURACY: 0.989112354453
PRECISION 0.907455012853
RECALL 0.907455012853
SENSITIVIY 0.907455012853
SPECIFICTY 0.994215938303
TPR 0.907455012853
FPR 0.00578406169666
X, Y(Best 0,1) (0.005784061696658127, 0.9074550128534704)
BALANCED F1 MEASURE 0.907455012853

OpenDNS

Beyond Phishtank

-DNS data is not the ideal match for this data...HTTP traffic much better fit

Why? When doing lookups, landing on index page, most often phishing page is not on index page

-Within DNS, necessary to build crawler

Question: But there's so much traffic, are we going to do GET request for every URL???

OpenDNS

OpenDNS Intelligent Proxy

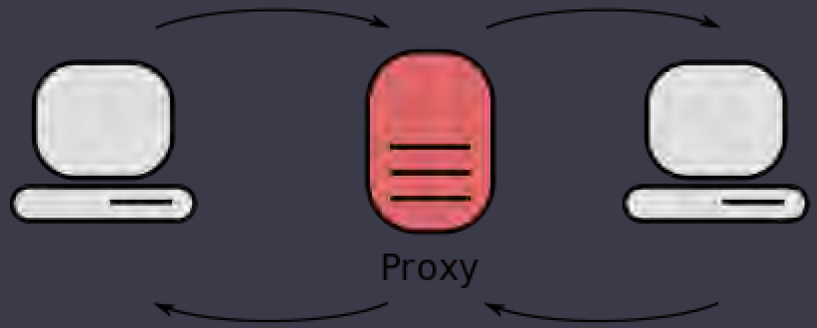
What is the Intelligent Proxy?

-Awesome Team!!

-Man in the Middle

-Greylisting

-Next step in OpenDNS Security



Dedicated vs. Compromised Examples

Dedicated:

update-java[.]net, adobe-update[.]net, http://wellsinfo.net/login

Compromised:

Domain: wwelllssssfffarrgo.webzdarma.cz.html

http://dandraghicescu.ro/dbox/dpbx/dpbx/

http://school76.irkutsk.ru/language/Wellsfargo/online.htm

http://createcrafts.ph/bankofamerica.com.update.login.in.info/de17792ab89754c6b0a58d767a6985fc/

http://www.kingdomhome.com.au/wp-admin/wellsfargo.zip/wellsfargo-online.server/details.html

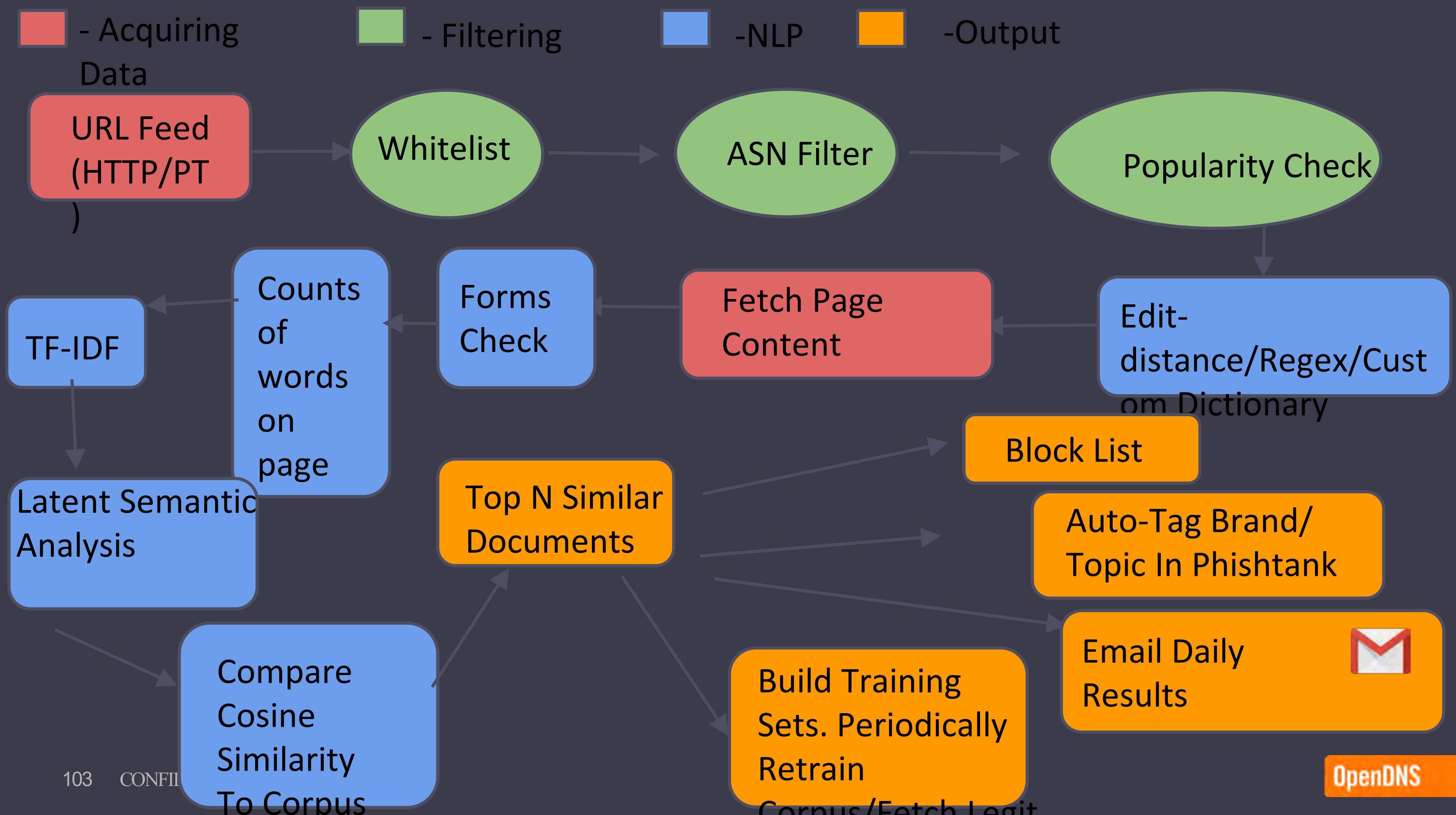
http://wellsfargoonline.pfwv.com.br/wellsfargo/

http://www.cityroo.com/sarasoa/wellsfargo/wellsfargo-online.php

http://wellsfargo.com.billing.account.updatemyaccount.**wellsfrago.**com.onlineaccounts.upgrade.online.billing.account.update.nlineaccounts.upgrade.online.billing.account.update.kowafdfsfs.net

http://comosecuraladiabetes.com/wp-admin/js/well.htm

OpenDNS


- Acquiring Data
- Filtering
-NLP
-Output
URL Feed (HTTP/PT)
Whitelist
ASN Filter
Popularity Check
Edit-distance/Regex/Custom Dictionary
Fetch Page Content
Forms Check
Counts of words on page
TF-IDF
Latent Semantic Analysis
Compare Cosine Similarity To Corpus
Top N Similar Documents
Block List
Auto-Tag Brand/ Topic In Phishtank
Email Daily Results
Build Training Sets. Periodically Retrain Corpus/Fetch Legit

OpenDNS

Conclusion

- Agile Research: Building, Testing, Tuning, Iterating
- Different Algorithms, LSA as Feature
- Topic Modeling on More Content (LDA, seasons)
- More Features (SimHashing, HTML content encoding)
- Data Collection/Building Corpus
- Filtering FPs
- Spark Streaming!
- United States ODNS=-1009US0; 62/167,178

OpenDNS is
now part of Cisco.

QUESTIONS?

@jmoconnor415
jeremiah@opendns.com
jeoconno@cisco.com

@ThibaultReuille
thibault@opendns.com
treuille@cisco.com

OpenDNS

The Security Wolf of Wall Street: Fighting Crime with High-Frequency Classification and Natural Language Processing

Jeremiah O'Connor and Thibault Reuille

January 2016

OpenDNS is now part of Cisco.

CISCO

$ whois jeremiah

-Mad Scientist at OpenDNS/Cisco Labs
-M.S. in Computer Science from University of San Francisco
-Previously worked at Mandiant (IR/DNS Research), Evernote (AppSec/IR), Uber (Data Science)
-Career Goals: Solve interesting problems (Networking/Security, Bioinformatics, GPS Tracking, Video Games, etc.)
-Proud SFSPCA Pitbull Puppy owner

OpenDNS

$ whois thibault

- Security Research Team at OpenDNS.
- Creator of OpenGraphiti.
- Focus: Data Visualization, 3D Graphics, Graph Theory and Real-time systems.

OpenDNS

Presentation Agenda

- Introduction : Challenges & Hypothesis
- Real-Time Processing Fundamentals
- The Avalanche Project & The Research Pipeline
- Live Demo!
- Future Work

Introduction to Avalanche

OpenDNS

Challenges

I’ve got 99 problems but malware ain’t one!

- We see a lot of traffic.
  - Needles in a haystack.
- Bad guys move fast.
  - The needles are playing hide-and-seek.
- Outdated information has less impact than hot news.
  - Slowpoke syndrome.
- Measuring the accuracy of our classifiers is not trivial.
  - How big is the base of the iceberg?

Hypothesis

To stream or not to stream.

- Most of our models can work in streaming.
  - Well, that's a strong statement.
- We can detect "anomalies" on the fly.
  - TSA is overrated anyway.
- We can have precise visibility over malicious activity.
  - Statistics + Dataviz = Win!
- We can talk about what nobody knows.
  - Wanna be famous?

REAL-TIME !

Real-Time, you said?

Different Levels of Constraints.

- "Soft"
  - Ex: Youtube / Netflix video streaming, Video Games, GPS …
- "Firm" :
  - Ex: Dishwasher, Audio DSP, Assembly line …
- "Hard" :
  - Ex: Airbag, UHFT Algorithmic Trading …
- "Critical" :
  - Ex: Missiles, Aircrafts, Nuclear Reactor …

- "Near Real-Time" : Network-induced indeterminism.

The Blackbox Abstraction

Real-Time vs High Performance.

T1 – T0 ~ 1 second
vs
T1 – T0 <= 2 seconds !!

Real-time != Fast

When Murphy meets the law of large numbers.

There's no such thing as "half water-proof".

Program

Runs fine 99% of the time
Probability of a crash : 1%

At infinity, a program that SOMETIMES crashes is equivalent to a program that ALWAYS crashes!

OpenDNS

Key Design Points

Things to consider when writing code.

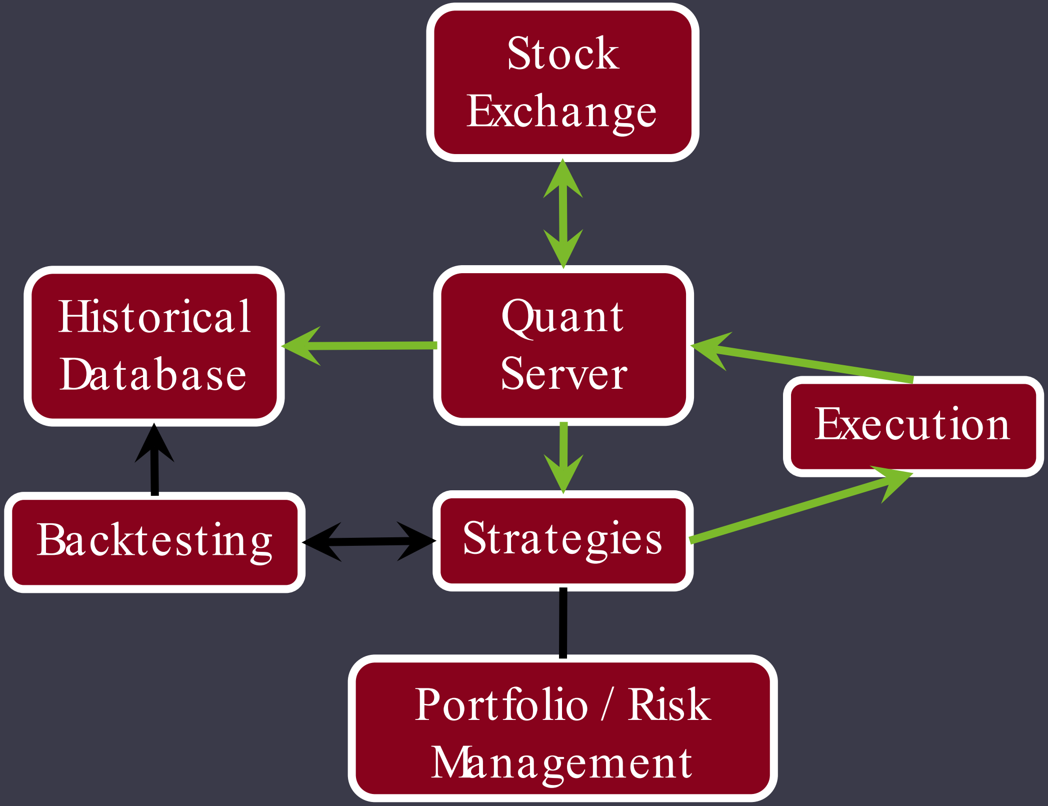
- Fault Tolerancy
  - Rigorous code.
  - Flawless error handling.
  - Unit tests
  - Degraded Mode?
- Algorithm Complexity : What's your worst case?
  - Computing Time : Is it deterministic?
  - Parallelism & Concurrency : Context Switching, Deadlocks, Race Condition…
  - Memory Allocation : Static vs Dynamic
- Environment
  - Background jobs, RAM, CPUs, Parasites, Hardware Failures…

The Avalanche Project

OpenDNS

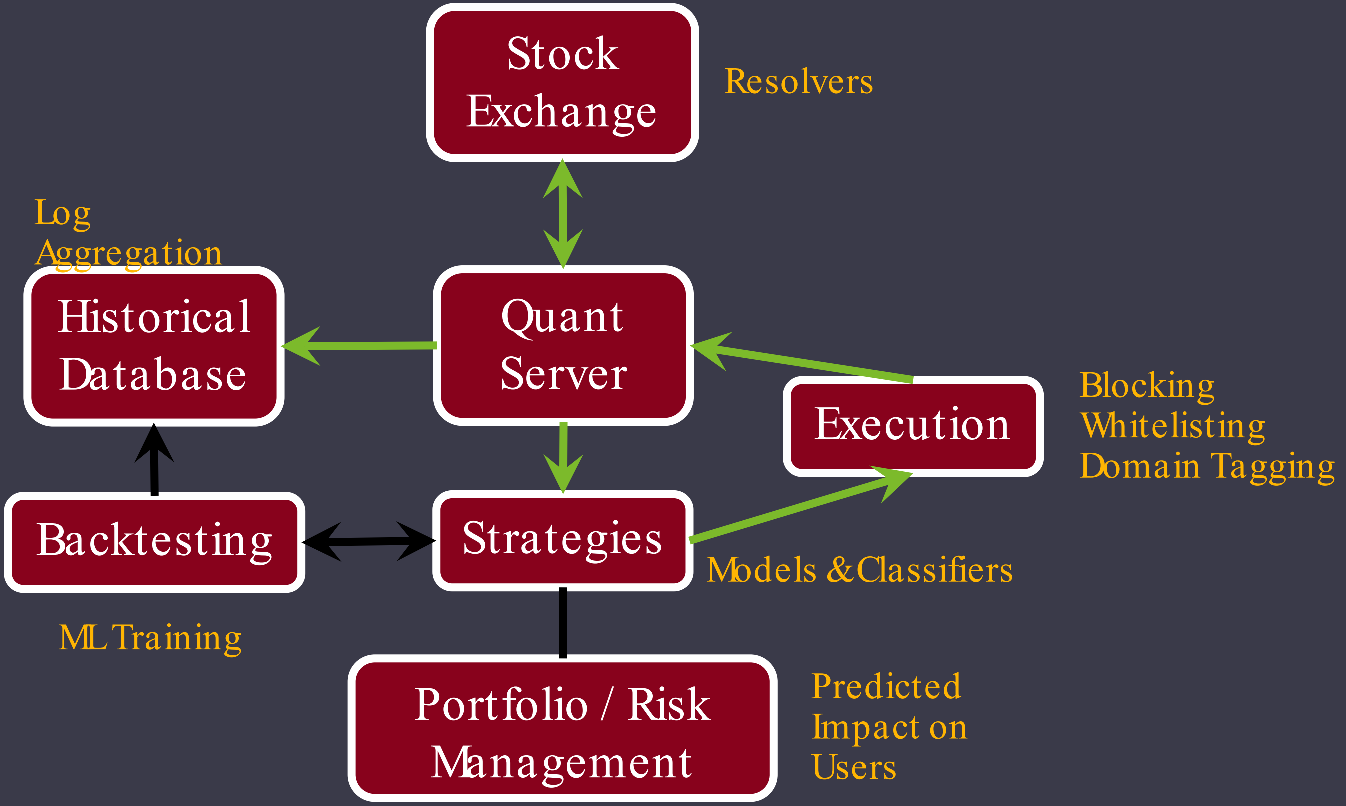
High Frequency Trading vs Traffic Classification

The Wolf of Wall Street



High Frequency Trading vs Traffic Classification

The Wolf of Wall Street



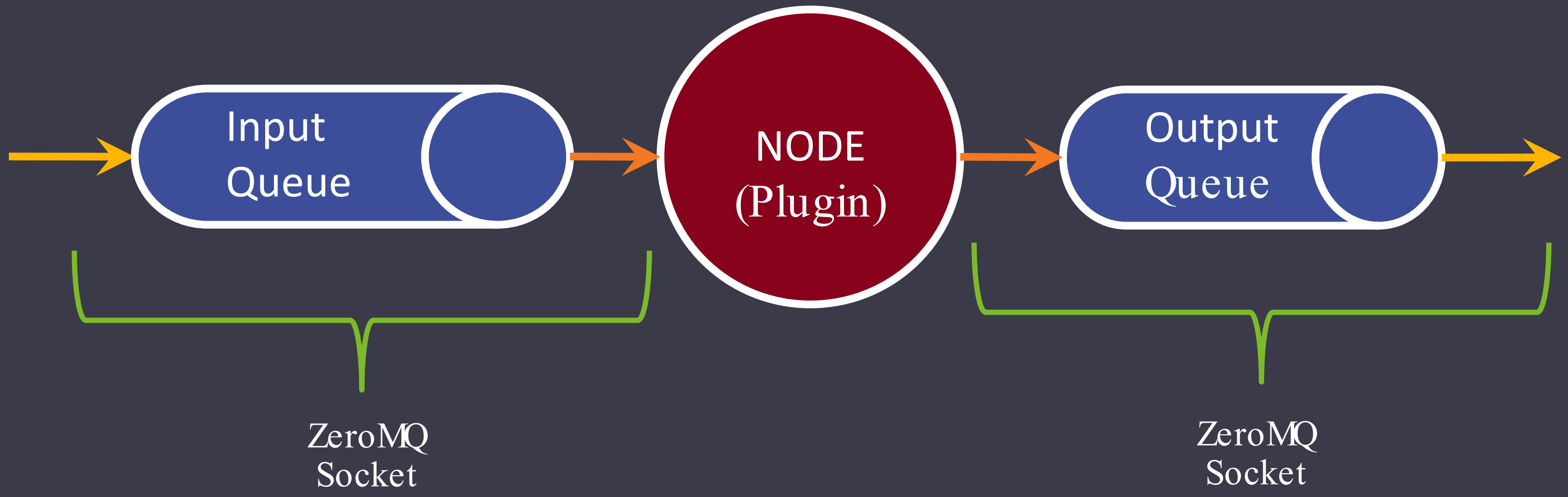
What is Avalanche?

Overview and Technical Details.

- Open source project :
  - http://github.com/ThibaultReuille/avalanche
- “Real-time” data processing framework.
- Modular, parallel and distributed design.
- Written with Python and ZeroMQ.
- Platform for some OpenDNS models (Private) :
  - https://github.office.opendns.com/Research/avalanche-opendns
  - NLP-Rank
  - DNS Tunnelling
  - Talos DGA classifier and others (In progress)

Avalanche Design

Divide and Conquer



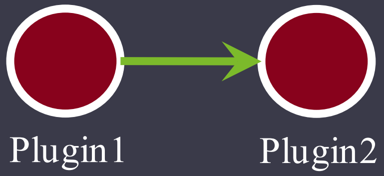
OpenDNS

Avalanche Node
Plugin Template Code

```python
import json
import plugins.base

class Plugin1(plugins.base.Plugin):
    def __init__(self, info):
        # NOTE: The info argument contains the full node definition
        # written in the pipeline configuration file.
        pass

    def process_message(self, message):
        # NOTE : Here we can process the message, add field, remove, etc.
        # Retuning None drops the message from the pipeline.
        return message

class Plugin2(plugins.base.Plugin):
    def __init__(self, info):
        # NOTE: The info argument contains the full node definition
        # written in the pipeline configuration file.
        pass

    def run(self, node):
        # NOTE: Each node runs on its own thread/process,
        # Here we enter our infinite loop.
        while True:

            # NOTE: Read incoming data sent to our node
            data = node.input.recv()

            # NOTE: Parse it as a JSON message
            message = json.loads(data)

            # NOTE: This template plugin doesn't do anything except being a passthru filter.
            # This is where the processing would actually happen in a real processor.
            # You can send whatever data you like in the output stream. That can be a modified
            # version of the incoming messages or any other message of your creation.

            # NOTE: Send it back through the pipeline
            node.output.send_json(message)

if __name__ == "__main__":
    print("Please import this file!")
```

Avalanche Graph

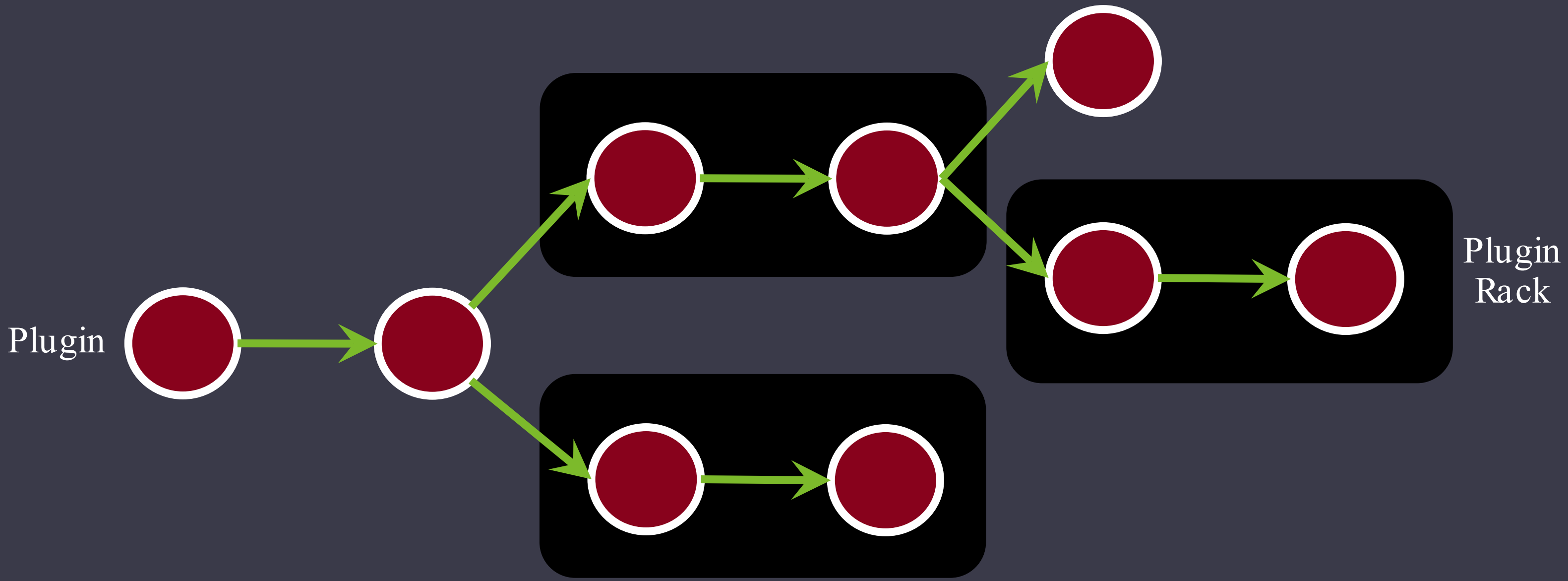
Pipeline Definition




```
{
    "attributes" : {
        "plugins" : [
            { "name" : "plugin1", "filename" : "path/to/plugin1.py" },
            { "name" : "plugin2", "filename" : "path/to/plugin2.py" }
        ]
    },

    "nodes" : [
        {
            "id" : 0,
            "type" : "plugin1",
            "attributes" : {
                "my_data" : "my_value"
            }
        },

        {
            "id" : 1,
            "type" : "plugin2",
            "attributes" : {
                "other_data" : "other_value"
            }
        }
    ],

    "edges" : [
        { "id" : 0, "src" : 0, "dst" : 1 }
    ]
}
```

Avalanche Pipeline

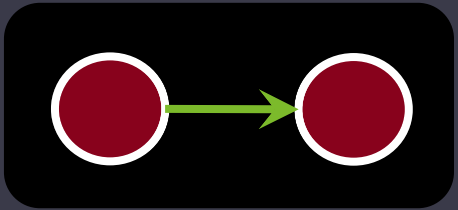
Divide and Conquer



OpenDNS

Avalanche Rack

Plugin Rack Definition



```
{
    "id" : 0,
    "type" : "rack",
    "plugins" :
    [
        {
            "type" : "plugin1",
            "attributes" : { "my_data" : "my_value" }
        },
        {
            "type" : "plugin2",
            "attributes" : { "other_data" : "other_value" }
        }
    ]
}
```


OpenDNS

Run Avalanche

```
$ ./avalanche.py path/to/my_pipeline.json 10000
```

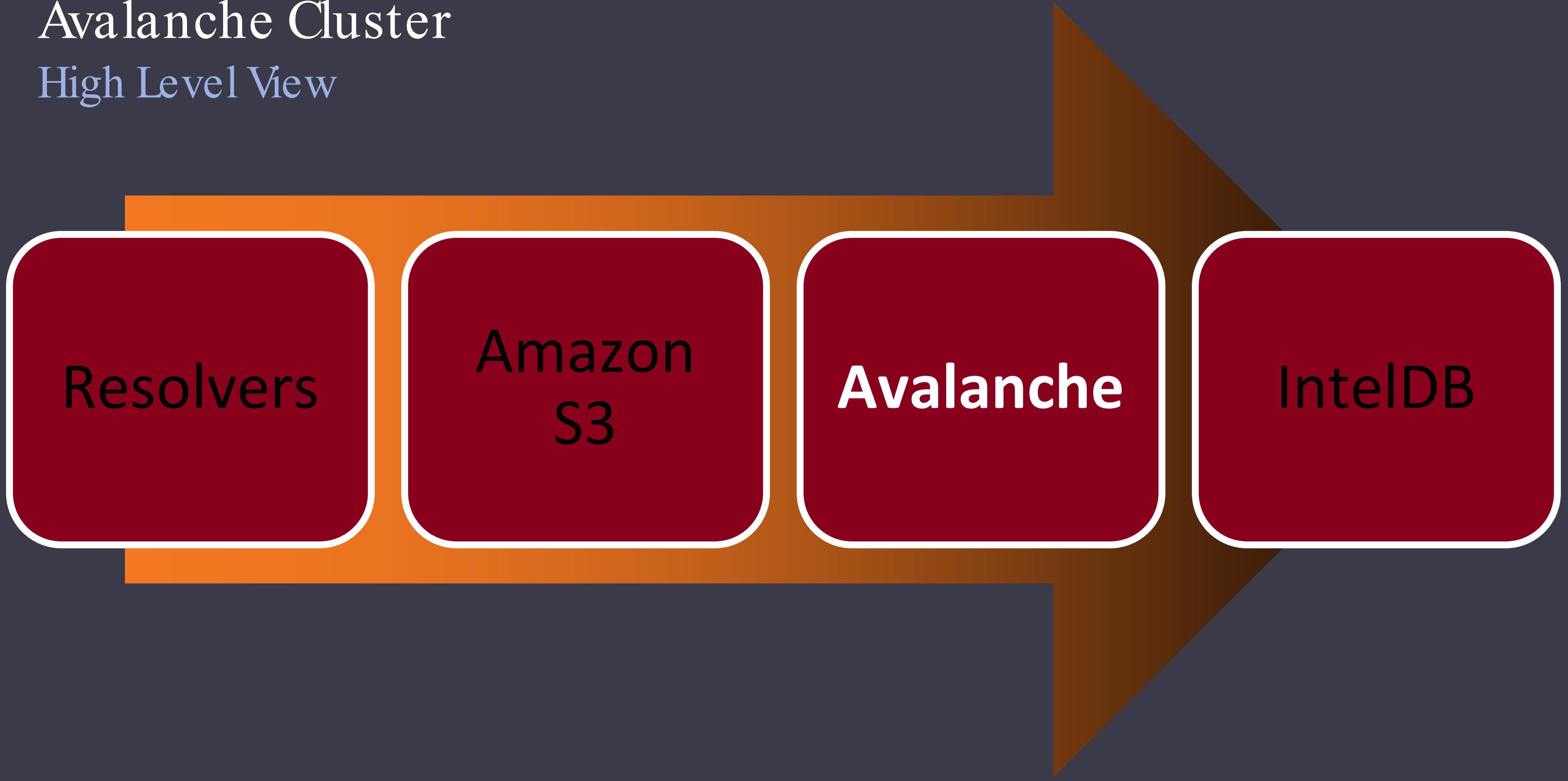
- Things you get for free :
  - Modularity.
  - Multi-Threading.
  - A library of plugins ready-to-use.
  - Reusability & collaboration.
  - An insanely fast messaging system.

OpenDNS

The Research Pipeline

OpenDNS

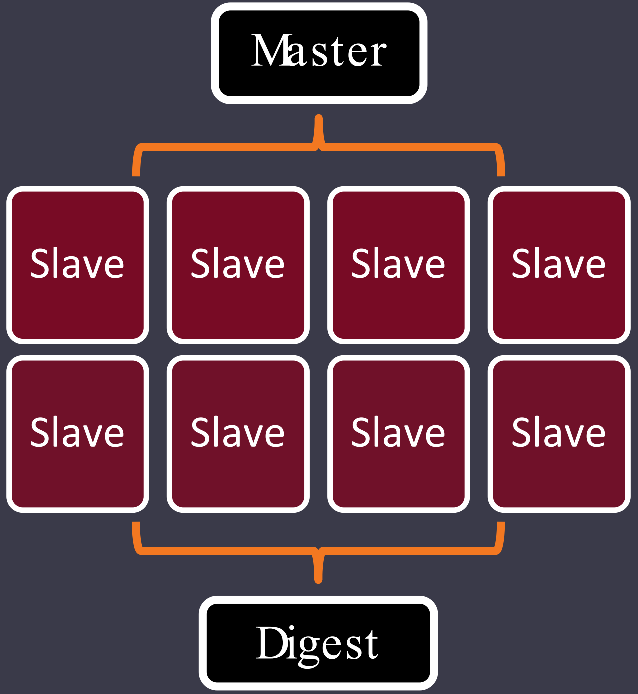
Avalanche Cluster
High Level View



OpenDNS

Avalanche Cluster

- 8 Amazon instances
- Master distributes work
  - Round-robin
  - "Fire and forget"
- Slaves process the chunks
- 4 Avalanche pipelines
- Results are centralized



Cluster Management with Boto & Fabric

```
treuille — avalanche@ip-10-20-9-89: ~/avalanche-services — ssh — 168×24
avalanche@ip-10-20-9-89:~/avalanche-services$ ls
avalanche.pem  digest.sh  instances.py  instances.pyc  miner.conf  profile.conf  pusher.py  requirements.txt  results.json  stats.py
avalanche@ip-10-20-9-89:~/avalanche-services$ fab -f instances.py -i avalanche.pem -- uptime
[Instance:i-c029ac72, Instance:i-c129ac73, Instance:i-ca29ac78, Instance:i-cb29ac79, Instance:i-ce29ac7c, Instance:i-cf29ac7d, Instance:i-cd29ac7f, Instance:i-cc29ac7e]
[10.20.9.96] Executing task '<remainder>'
[10.20.9.96] run: uptime
[10.20.9.96] out:  17:25:59 up 21 days, 16:49,  1 user,  load average: 0.02, 1.63, 2.35
[10.20.9.96] out:

[10.20.9.97] Executing task '<remainder>'
[10.20.9.97] run: uptime
[10.20.9.97] out:  17:24:43 up 21 days, 16:48,  1 user,  load average: 6.19, 2.92, 2.34
[10.20.9.97] out:

[10.20.9.90] Executing task '<remainder>'
[10.20.9.90] run: uptime
[10.20.9.90] out:  17:25:29 up 21 days, 16:48,  1 user,  load average: 0.04, 1.61, 2.51
[10.20.9.90] out:

[10.20.9.91] Executing task '<remainder>'
[10.20.9.91] run: uptime
[10.20.9.91] out:  17:25:58 up 21 days, 16:49,  1 user,  load average: 0.04, 1.60, 1.82
[10.20.9.91] out:
```

https://github.office.opendns.com/Research/avalanche-services

Traffic Speed vs Avalanche Pipeline

Numbers don't lie.

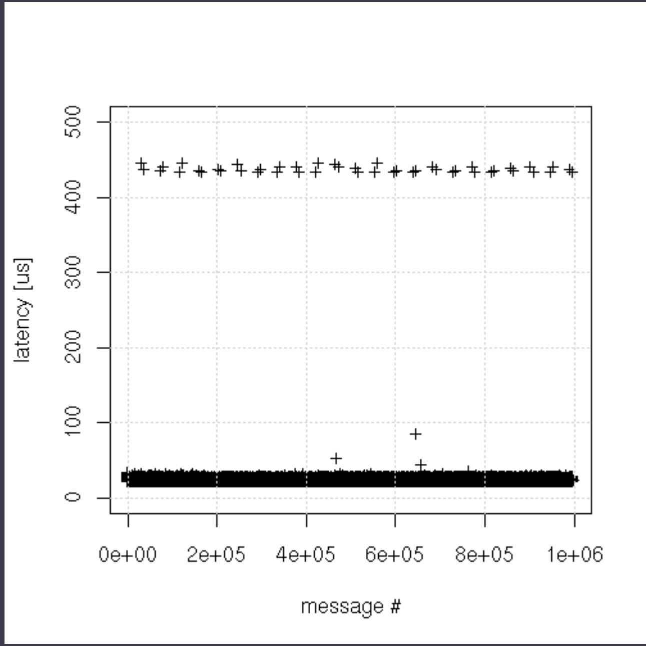
Queries / Chunk	Authlogs (AMS.m1)	Querylogs (AMS.m1)
Noon (UTC)	564 752	6 147 997
Midnight (UTC)	412 050	3 315 157
Queries / Second	**Authlogs (AMS.m1)**	**Querylogs (AMS.m1)**
Noon (UTC)	941.25	<u>10246.66</u>
Midnight (UTC)	686.75	<u>5525.26</u>

- Avalanche Benchmark :
  - ~30000 messages per second ⇔ 1 message every 33 microseconds.
  - 3 times faster than AMS.m1 query logs at peak time.

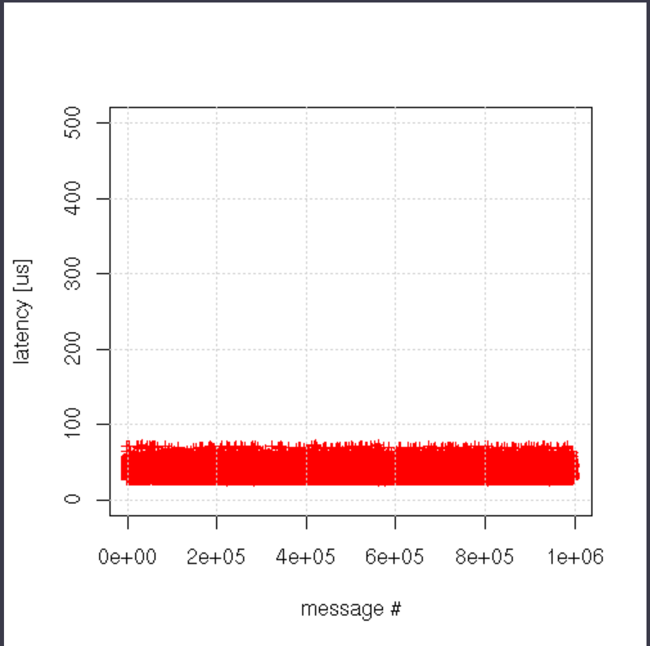
OpenDNS

ZeroMQ Performance Tests

Standard Linux Kernel

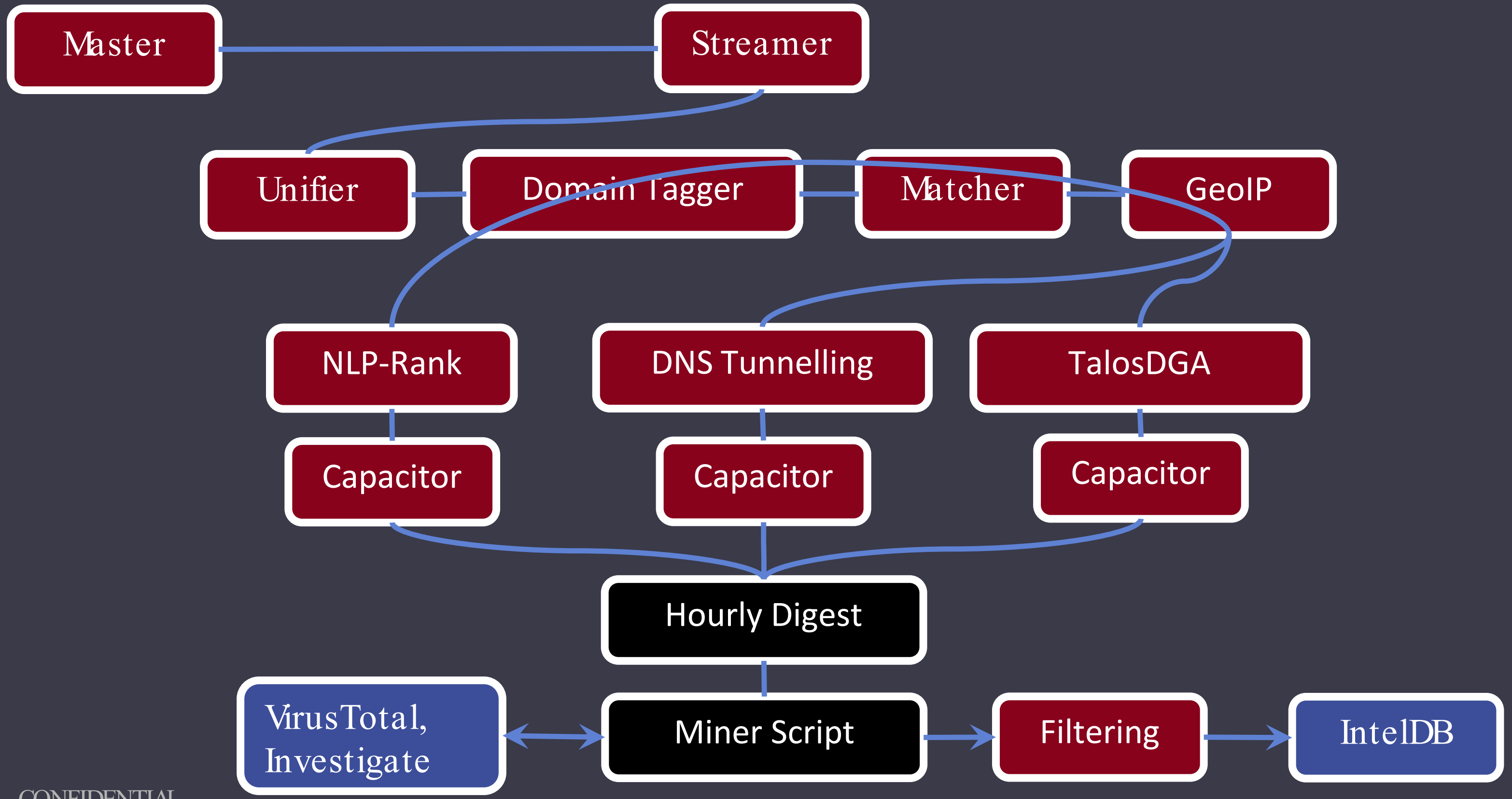


Real-Time Linux Kernel



Source: http://zeromq.org/results:rt-tests-v031

Slave Processing Pipeline



OpenDNS

```
avalanche.r1.usw1.opendns.com/avalanche/
```

Index of /avalanche/

```
../
dns-tunnelling/                         06-Nov-2015 00:15
nlp-rank/                               06-Nov-2015 00:13
```

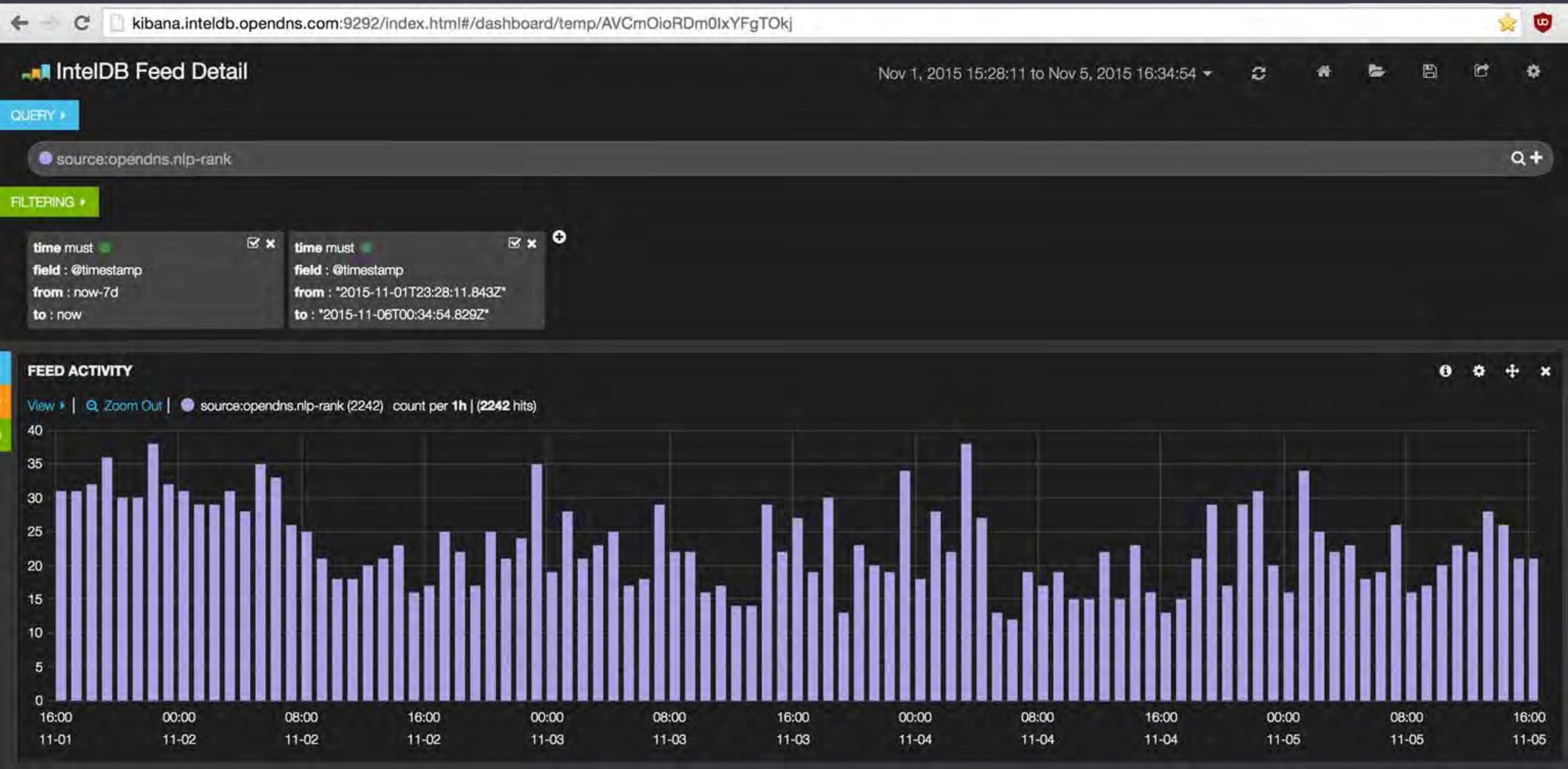
```
2015.11.05-19.00.01/                    05-Nov-2015 19:13        -
2015.11.05-20.00.01/                    05-Nov-2015 20:13        -
2015.11.05-21.00.01/                    05-Nov-2015 21:12        -
2015.11.05-22.00.01/                    05-Nov-2015 22:14        -
2015.11.05-23.00.01/                    05-Nov-2015 23:13        -
2015.11.06-00.00.01/                    06-Nov-2015 00:13        -
stats.txt                               06-Nov-2015 00:14      718
total.txt                               06-Nov-2015 00:14  5655720
```

Index of /avalanche/nlp-rank/2015.11.06-00.00.01/

```
../
domains.txt                     06-Nov-2015 00:13      9705
nlp-rank.10.20.9.90.csv         06-Nov-2015 00:12    153216
nlp-rank.10.20.9.91.csv         06-Nov-2015 00:11    141006
nlp-rank.10.20.9.92.csv         06-Nov-2015 00:10    108028
nlp-rank.10.20.9.93.csv         06-Nov-2015 00:09     87443
nlp-rank.10.20.9.94.csv         06-Nov-2015 00:13    158555
nlp-rank.10.20.9.95.csv         06-Nov-2015 00:11    140592
nlp-rank.10.20.9.96.csv         06-Nov-2015 00:10    114785
nlp-rank.10.20.9.97.csv         06-Nov-2015 00:08     77933
stats.txt                       06-Nov-2015 00:13       613
```

```
--- Generic Statistics ---

214679 Elements: 188016 domains + 26663 missing data (Ignored).

. Blacklisted: 3867
. Greylisted: 182233
. Whitelisted: 1916

. VT positives >= 5 : 5222
. Unknown by VT : 176676
. Popularity >= 80.0 : 14

--- Detailed Statistics ---

. Blacklisted and VT >= 5 : 2185
. Blacklisted and unknown by VT : 1002
. Blacklisted and Popularity >= 80.0 : 0

. Greylisted and VT >= 5 : 2865
. Greylisted and unknown by VT : 174123
. Greylisted and Popularity >= 80.0 : 10

. Whitelisted and VT >= 5 : 172
. Whitelisted and unknown by VT : 1551
. Whitelisted and Popularity >= 80.0 : 4
```

```
#FQDN,depth,popularity,age,ips,prefixes,asns,countries,ttl_min,ttl_max,ttl_stddev,geo_sum,geo_mean,entropy,perplexity,
apple-winks.com,0,0.0,,1,1,1,1,600,600,0.0,0.0,0.0,3.2776134368191165,0.2739846357448707,0,6
ebay.login.com.5599.carsgoneby.aspmodel.info,0,0.0,,,,,,,,,,,3.0,0.6361674803007081,-1,6
ekosamazonia.com.br,0,7.169532493946863,,1,1,1,1,14400,14400,0.0,0.0,0.0,3.0220552088742,0.4266416677105029,-1,11
www.microsoftpartnerserverandcloud.com,0,50.50501253890862,,1,1,1,1,3600,3600,0.0,0.0,0.0,3.8029100796497266,0.5594928
serviceapple-support.bugs3.com,0,0.0,,1,1,1,1,14400,14400,0.0,0.0,0.0,2.321928094887362,0.524856068944591l,-1,9
secure2.store.apple.com-contacter-apple.jrrdy.com,0,11.363440150607609,,1,1,1,1,600,600,0.0,0.0,0.0,1.9219280948873623
ebooking.applewf.com,0,18.532972644554473,,1,1,1,1,3600,3600,0.0,0.0,0.0,2.5216406363433186,0.5095322471047489,1,10
yourjavascript.com,0,99.73011810869362,,5,3,2,3,30,300,133.30655317392907,9517.938306462407,3172.646102154136,3.521640
electricidadobera.com,0,11.363440150607609,,1,1,1,1,14400,14400,0.0,0.0,0.0,3.219528282299548,0.3663643606263674,1,11
login.ebay.com.account-limited.8619.redhoaglandhyundai_s5_129716198.aspmodel.info,0,,,,,,,,,,,,3.0,0.9851213341419353,
login.ebay.com.account-limited.3564.chris.aspmodel.info,0,0.0,,,,,,,,,,,3.0,0.6510072618562623,-1,6
drive.google.uploadeddocx.com,0,0.0,,1,1,1,1,600,600,0.0,0.0,0.0,3.0220552088742,0.6446774004795882,-1,8
paypalverification.co.vu,0,0.0,,1,1,1,1,60,60,0.0,0.0,0.0,1.0,0.5850301939830299,1,9
signin.ebay.com.ssl-protection.5724.jimmy.aspmodel.info,0,0.0,,,,,,,,,,,3.0,0.8053896409511141,-1,7
poypal.simply-winspace.fr,0,11.363440150607609,,1,1,1,1,900,900,0.0,0.0,0.0,3.506890595608518,0.7655825019506184,-1,13
verify-apple.ml,0,,,,,,,,,,,,3.2516291673878226,0.981196000857034,0,9
www.gooogle.com,0,68.25134144531397,,6609,314,249,81,300,300,0.0,1164166.5744639637,6577.21228510714,1.842370993177108
newpaypal.uni.me,0,0.0,,4,1,1,1,300,300,0.0,0.0,0.0,1.584962500721156,0.8364938372280273,1,8
bankofamerica.com.restore-pagenkt23nbrirz.bb01abc4net.com,0,0.0,,2,2,2,2,300,14400,7050.0,8106.479711160472,4053.23985
update-secure-signin-help-inc-confirm-apple-manage.srpschapper.org,0,7.169532493946863,,1,1,1,1,14400,14400,0.0,0.0,0.0,
questionnairepaypal03822.110mb.com,0,0.0,,1,1,1,1,21600,21600,0.0,0.0,0.0,1.9219280948873623,0.8078908438816185,1,12
```



kibana.inteldb.opendns.com:9292/index.html#/dashboard/temp/AVCmOioRDm0IxYFgTOkj
IntelDB Feed Detail
Nov 1, 2015 15:28:11 to Nov 5, 2015 16:34:54
QUERY
source:opendns.nlp-rank
FILTERING
time must
field : @timestamp
from : now-7d
to : now
time must
field : @timestamp
from : "2015-11-01T23:28:11.843Z"
to : "2015-11-06T00:34:54.829Z"
FEED ACTIVITY
View | Zoom Out | source:opendns.nlp-rank (2242) count per 1h | (2242 hits)
40
35
30
25
20
15
10
5
0
16:00 11-01
00:00 11-02
08:00 11-02
16:00 11-02
00:00 11-03
08:00 11-03
16:00 11-03
00:00 11-04
08:00 11-04
16:00 11-04
00:00 11-05
08:00 11-05
16:00 11-05

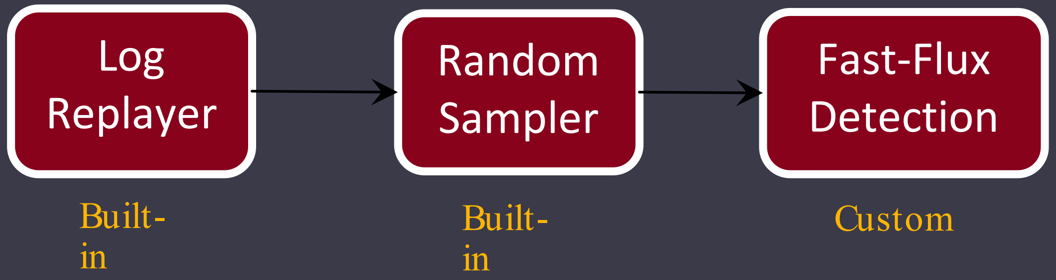
OpenDNS

Live Demo

OpenDNS

Authlogs & Querylog Replaying

Workshop : Simple Fast-Flux Detection Pipeline

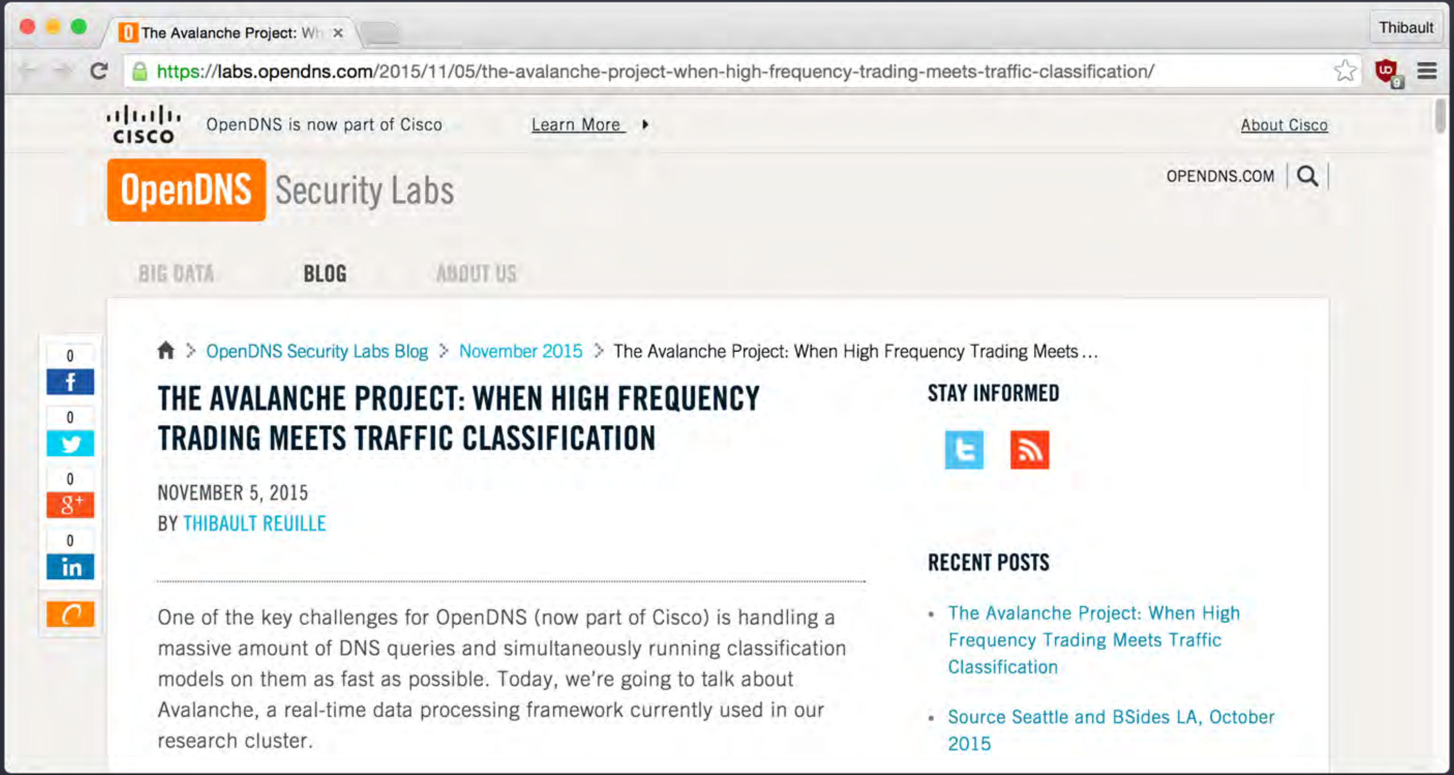


What's next?

Future Work

- More models!
- Cython or rewrite core in C/C++
  - Optimize model performance
- Use GPU grids :
  - OpenCL, GPU cluster
- Hackathon Idea :
  - Avalanche at the DNS resolver level
- More log visibility
  - Querylogs
  - Proxy logs

Blog Post is Live.



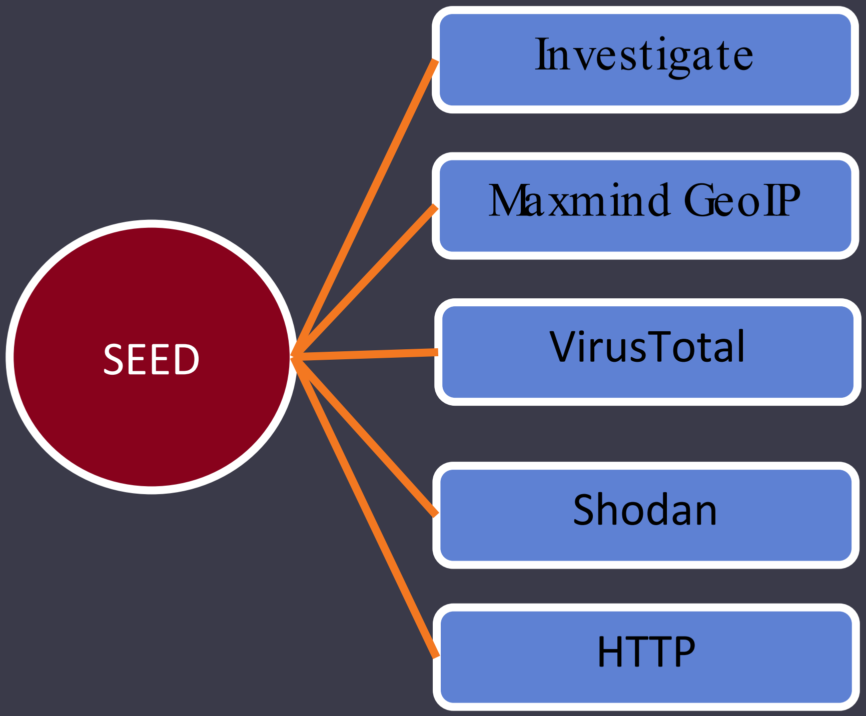
OpenDNS

Introduction to Miner/Graph-Oriented Data Mining

OpenDNS

Interesting Data Sources …

- Domain
- URL
- IP
- ASN
- Hash
- Email
- Regex



Scores, Co-occurrences …

Country Info, ASNs …

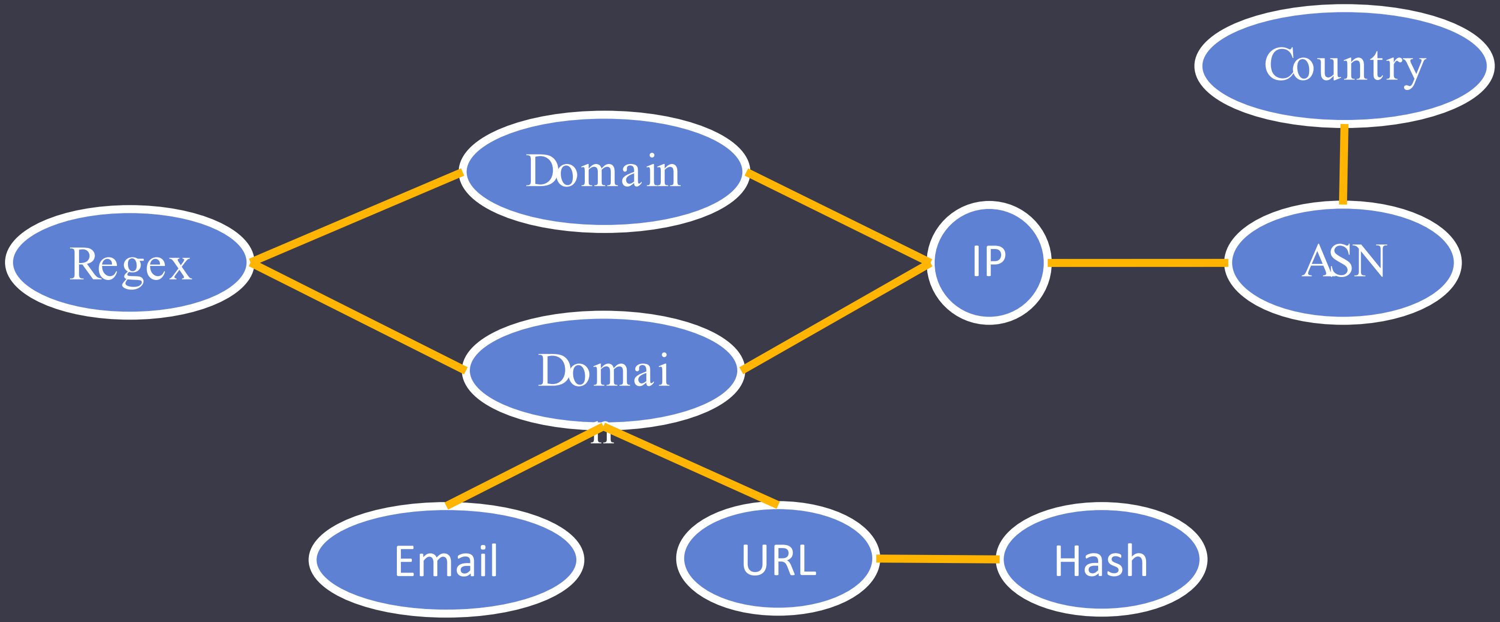
Malware URLs, Vendor Info…

Banner Info …

HTML Content, Certificate, Links …

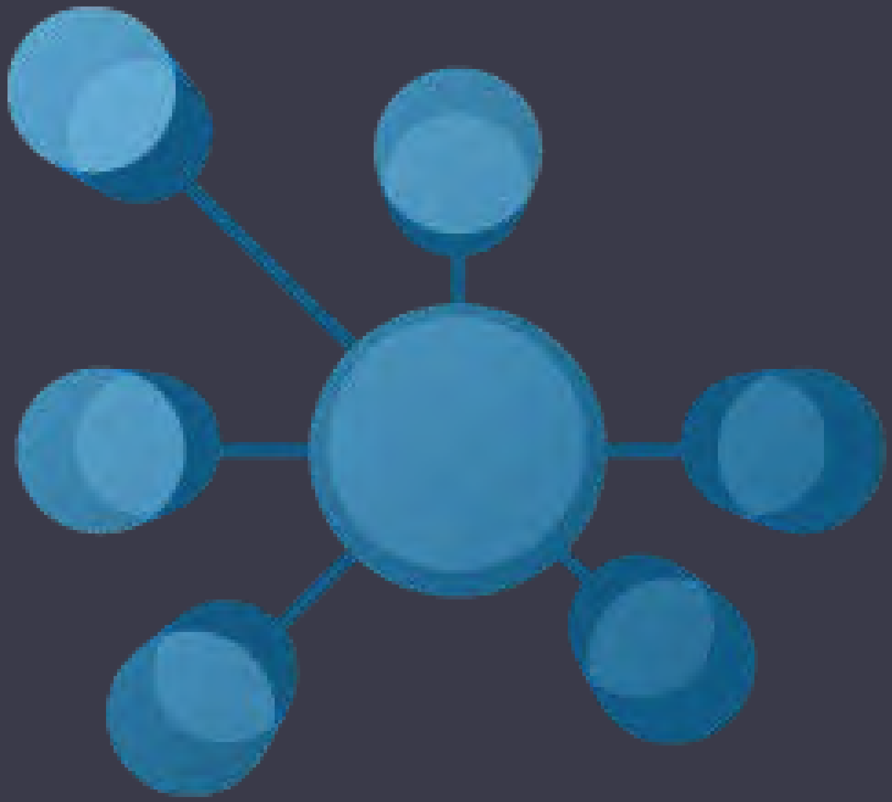
…

Data Modeling Example

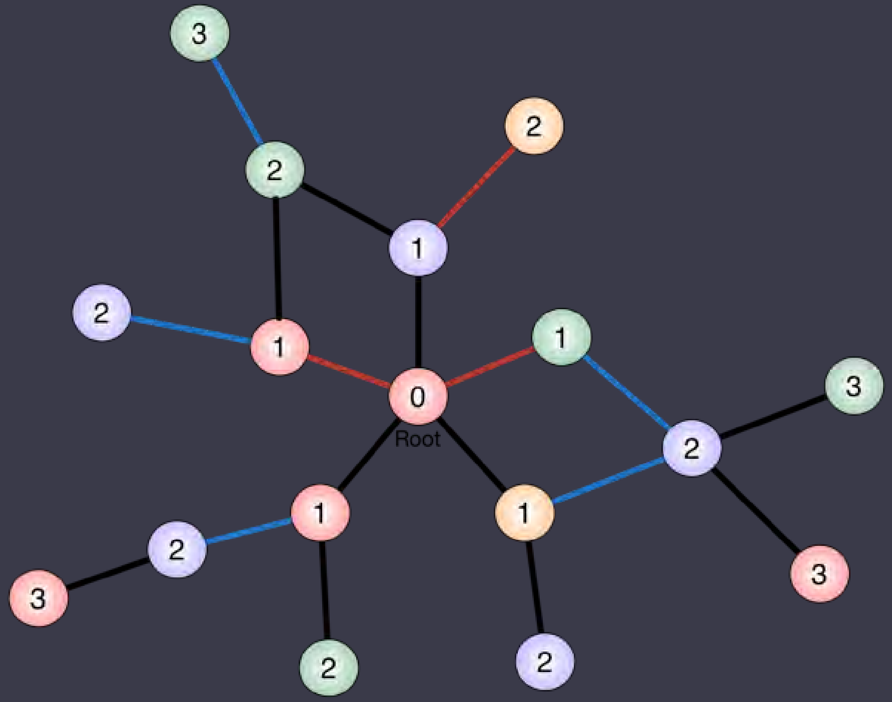


Knowledge

- Semantic Networks / Property Graph
- Node = Concept, Edge = Relationship
- Model of the Information
- Ontology : Model of the Model

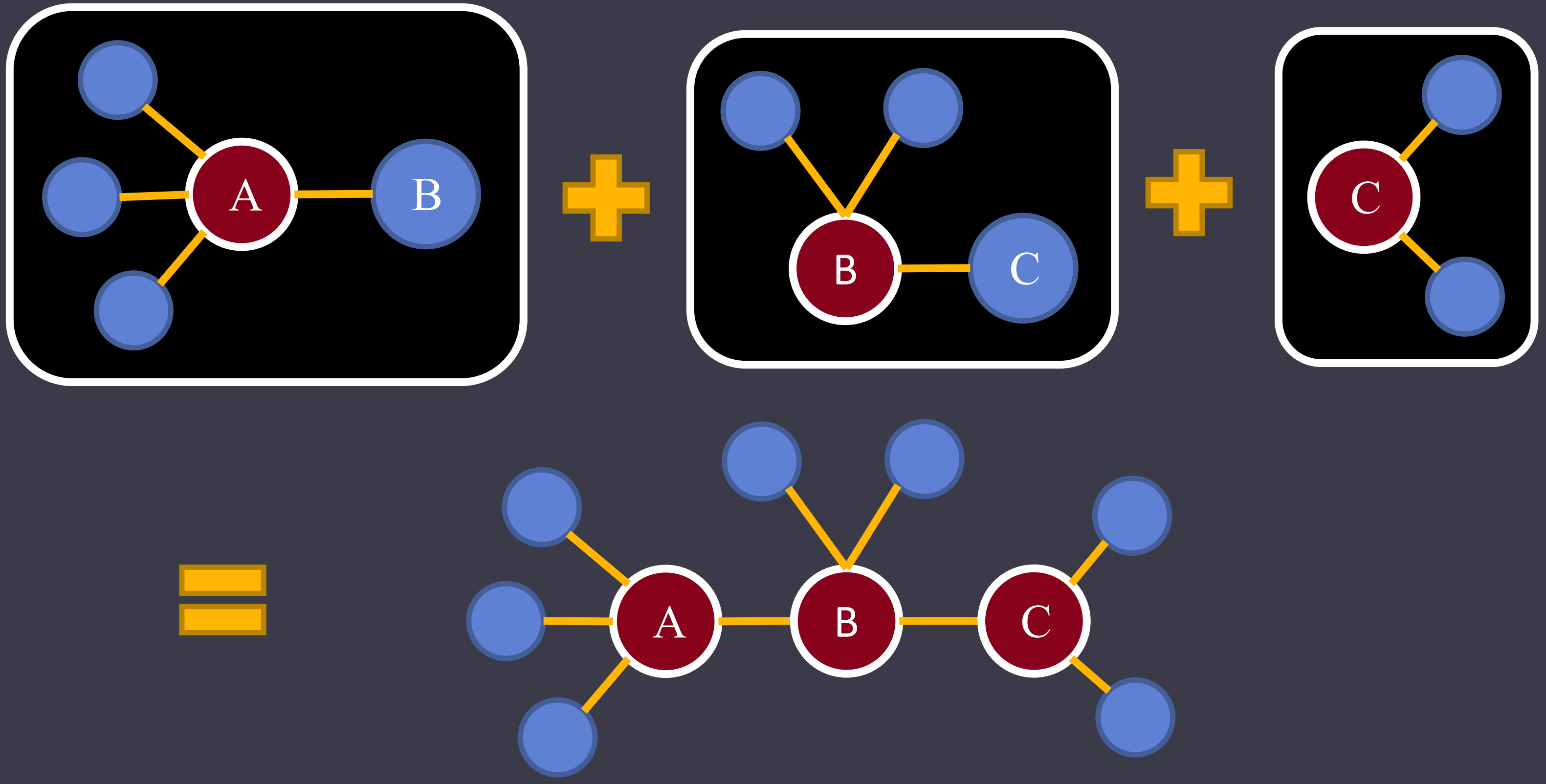


Data Exploration : Breadth First Traversal

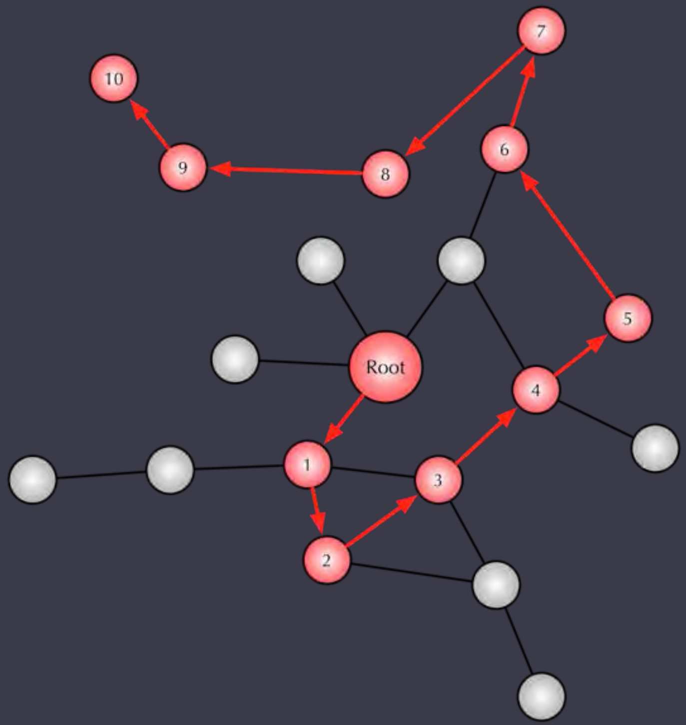


OpenDNS

Multi-Threaded Breadth First Traversal



Lambda Mining



- Functional Graph Exploration
- Rule Based / Thresholds / Topology based …
- Profiles for specific use cases
- Automated Smart Data Mining

NLPRank/Phishing Detection

OpenDNS Security Labs

Data Science ∩ Network Security

Big Security Data-
DNS Traffic:
~70B DNS requests per day
HTTP Traffic:
~10.1Mrequests per day

Daily Tasks:

-Detection Algorithms, Security Data Analysis, Distributed Systems, Big Data Engineering, Data Viz

OpenDNS

Purpose:

Overview of our new model **NLPRank**:

-Fraud detection system using NLP techniques and traffic features to identify domain-squatting/brand spoofing in DNS/HTTP (a technique commonly used by phishing and APT CnCs).

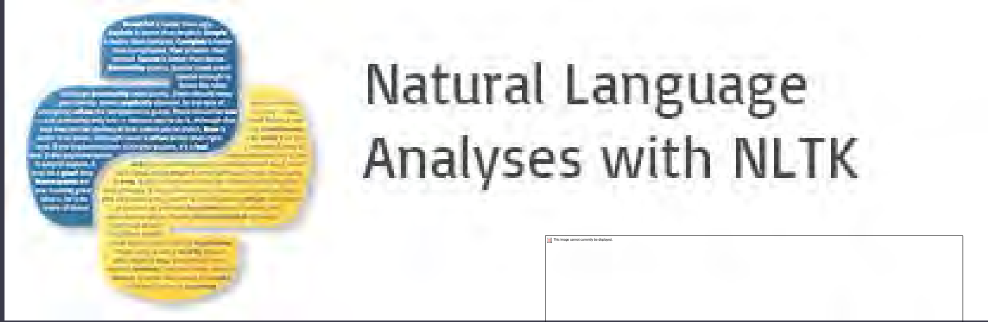
#TeamPython

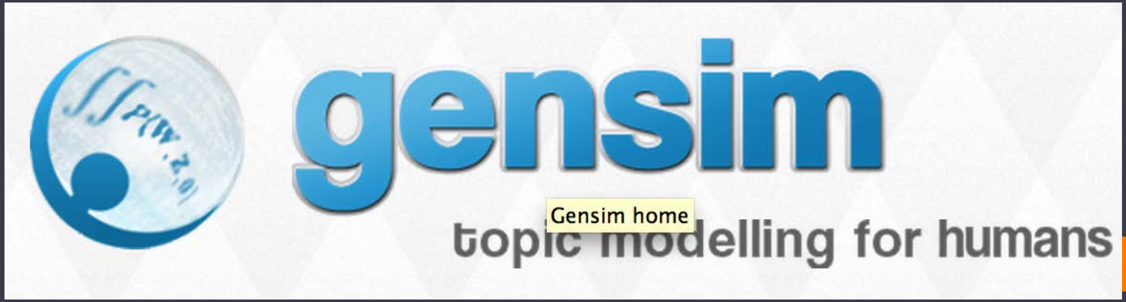
NLP/Data Science:

-NLTK
-Scikit-Learn
-Gensim

Web Scraping:

-Beautiful Soup
-LXML





System Origins

-OpenDNS Labs has detection models for commodity malware (ex. Botnet, Fast-Flux, DGA) need a model to detect targeted attacks

-Assigned to analyze DarkHotel data set

Question: How to detect "evil" in DNS records using lexical features of FQDN and validate results?

Human-Computer Interaction

Targeted Attacks: From a psychological perspective, if you were a high-profile exec for company what kind of links would you click on? What are your interests?

Commodity Phishing: Same psychology

Topics of interest:

-$$$, Bank Account/CCs, Financial

-News

-Security/Software updates

-Social Network

CHASE

PayPal

Bank of America

WELLS FARGO

Google

YAHOO!

AMERICAN EXPRESS

Heuristic #1- ASN Filtering

ASN Overview

-Autonomous System Number is basically like your neigborhood/zipcode on the internet
-Associated with Internet Service Provider
-Set of routers operating under specific or multiple routing protocol
-Domains exhibiting fraudulent behavior are observed to be hosted on ASN's that are unassociated with the company they're spoofing

OpenDNS

Examples

Expect a FQDN containing “adobe” to be associated with Adobe’s ASN (ex. ASNs 14365, 44786, etc.), or FQDN containing “java” and advertising an “update” be associated with Oracle ASN (ex. 41900, 1215, etc.)

So why then?

APT Example (Carbanak):

-adobe-update[.]net - ASN 44050, PIN-AS Petersberg Internet Network LLC in Russia

-update-java[.]net - ASN 44050, PIN-AS Petersberg Internet Network LLC in Russia

Commodity Phishing Examples:

Domain: securitycheck.paypal.com

ASN 20013, CYRUSONE -CyrusOne LLC, US

Domains: serviceupdate-paypal.com, updatesecurity-paypal.com,

ASN 32400 - Hostway Services, Inc.,US

OpenDNS

The Usual Suspects..

1. CyrusOne LLC,US
2. Unified Layer,US
3. OVH OVH SAS,FR
4. GoDaddy.com, LLC,US
5. HostDime.com, Inc.,US
6. SoftLayer Technologies Inc.
7. HOSTINGER-AS Hostinger International Limited,LT
8. HETZNER-AS Hetzner Online AG,DE
9. Liquid Web, Inc.,US
10. CLOUDIE-AS-AP Cloudie Limited-AS number,HK

OpenDNS

More Normalized…

1. OBTELECOM-NSK OOO Ob-Telecom, RU
2. GVO - Global Virtual Opportunities, US
3. CONFLUENCE-NETWORK-INC - Confluence Networks Inc, VG
4. CYRUSONE - CyrusOne LLC, US
5. VFMNL- AS Verotel International B.V., NL
6. NEOLABS- AS Neolabs Ltd., KZ
7. DEEPMEDIA- AS Deep Media / V.A.J. Bruijnes (sole proprietorship),NL
8. NEUSTAR- AS6 - NeuStar, Inc., US
9. VERISIGN- ILG1 - VeriSign Infrastructure & Operations, US
10. CIA- AS Bucan Holdings Pty Ltd, AU

OpenDNS

ASN Filter + Whitelisting

1st step to take a big chunk out of the traffic, because text processing is computationally intensive

-Do a lot of ASN Analysis with other models (Dhia Mahjoub, PhD Graph Theory)

Authlogs come in -> Enricher node will look up ASN and include logs

- Create mapping of Brand Names to their legitimate ASNs
- Lookup domains/IPs as they come in

OpenDNS

Heuristic #2 - Defining Malicious Language Within FQDNs

Building Intuitions

-Eyeball Data

-Run basic text metrics on the data, gain intuitions about the data and extract important words/substrings in APT FQDN datasets

-APT domains exhibit similar lexical features to commodity phishing domains

-Important look at word co-occurrences (bigrams, trigrams, etc.)

Building Intuitions

-From APT data sets extracted words from dictionary and applied stemming looking at word stats:

Top counts (stemmed): mail, news, soft, serv, updat, game, online, auto, port, host, free, login, link, secur, micro, support, yahoo

Bigram Collocations:

Words that often appear with each other

adobe-update

update-java[.]com

Idea:

brandname + ad-action word [.] tld

Examples

Dark Hotel (Kaspersky):

- adobeupdates[.]com
- adobeplugs[.]net
- adoberegister[.]flashserv[.]net
- microsoft-xpupdate[.]com

Carbanak (Kasperksy):

- update-java[.]net
- adobe-update[.]net

APT 1 Domains (Mandiant):

- gmailboxes[.]com
- microsoft-update-info[.]com
- firefoxupdata[.]com

NLP on FQDN

-Creating a “malicious language” derived from lexical features of FQDNs from APT/Phishing data sets
-Built corpus of domains similar to examples in previous slide
-Create custom dictionaries

Brandname Dictionary

Ex. google, gmail, paypal, yahoo, bankofamerica, wellsfargo

-Custom set of stemmed common malicious words

Ex. secur, updat, install, etc.

-Reason for stemming example: updat -> firefoxupdata[.]com (APT1)
-Apply Edit-Distance/Automata Theory on substrings to build spam language

Heuristic #3- HTML Content Analysis

Recreating Researcher's Mind

When reviewing malicious domains what is typical methodology for review:

1) Visit site in Tor browser
2) Researcher processes information on site, looks for clues, gains summary
3) Makes decision whether site is legit/malicious

Specifically for Phishing Sites:

Human-Computer Interaction: What makes people fall for this?

Site will be near copy of legitimate site it's intending to spoof

How can we automate this process?

Can we apply document similarity algorithms?

Human-Computer Interaction

Examples from Apple Phishing page:

Title: Apple GSXLogin

Links:

https://iforgot.apple.com/cgi-bin/findYourAppleID.cgi?language=US-EN&app_id=157&s=548-548

https://id.apple.com/IDMSAccount/myAccount.html?appIdKey=45571f444c4f547116bfd052461b0b3ab1bc2b445a72138157ea8c5c82fed623&action=register&language=US-EN

Images:

Other Clues:

HTTrack - tool used to clone site

```
<!DOCTYPE HTML><html lang="">

<!-- Mirrored from tools.google.com/dlpage/drive/index.html by HTTrack Website Copier/3.x [XR&CO'2014], Tue, 23 Sep 2014 08:58:40 GMT -->

<!-- Added by HTTrack --><meta http-equiv="content-type" content="text/html;charset=utf-8" /><!-- /Added by HTTrack -->

<head><script type="text/javascript">

function utmx_section(){}function utmx(){}
```

Preparing The Data

-Cleaning the Data

-Stripping punctuation, symbols, unnecessary content

-Normalizing the data

-Stemming (update, updating, updater →updat)

Feature Encoding

```
© Google    ·
<a href="https://www.google.com/intl/en/policies/privacy/">
  Privacy Policy
</a>
```

Harder than it seems...

-Non-Trivial to extract relevant terms from HTML documents
-Dealing with malformed tags
-Lose data, dealing with HTML and JS
-Which tags to encode?
  -Title
  -Links
  -Images

Applied basic NLP Algos..but need more samples for training!!

OpenDNS

More Headaches

Legit USAA Site:

<title>USAA Military Home, Life & Auto Insurance | Banking & Investing</title>

Many USAA Phishing Sites:

<title>USAA Military Home, Life & Auto Insurance | B
Investing</title>

USAA Phishing Page:

<title>USAA Military Home, Life & Auto Insurance</title>

Success Identifying All Different Types of Attacks

Success in Training:

Detecting:
Careto
APT Domains Darkhotel/Carbanak/APT1 etc.
AJAX Hacking Group/Flying Kitten infostealer C&C
Operation Pawn Storm
Operation Saffron Rose
and more...

Success on Live Data:

Exploit Kit
Fast-Flux
And new stuff..

Interesting Results

Carbanak (banking trojan) came out in February:

2015-01-23 14:52:58 -- a96e74b8-b052-4f42-a517-d7273d4f13e7

NLPRank High-Risk Results (FQDNs)

cdneu.windows8downloadscdn.com
update-java.net

OpenDNS

Interesting Results

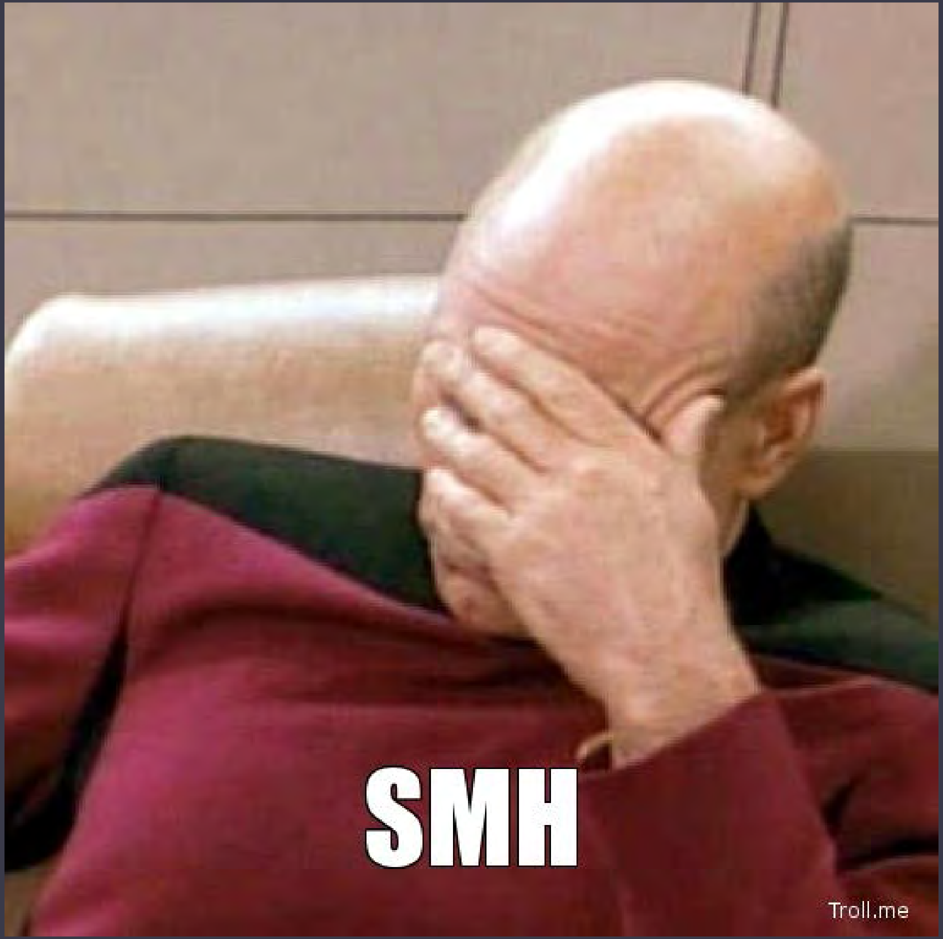
symantecupdates.com

Whois information

Registration date	2013-09-03 00:00:00 +0000
Registrar name	GODADDY.COM, LLC
Registrant	li ning < li2384826402@yahoo.com >
Registrant contact address	guangdongsheng guangzhoushi Alabama UNITED STATES

21,533 Domains???

crowcasinovip.biz mybestbrand.biz mybestbrands.biz huarenceluewangzhi.com icbczay.com boyinbocai5.com haoyunc3.com bocaiwangzhenqianpingtai.com zuqiubocaiwangzhan7.com weinisirenyulecheng94.com xinquanxunwang244.com dfjdh.com yaojiyulecheng9898.com wanbaoluyulecheng94.com xinpujingyule15.com toabaao.com jinbaiyiyulecheng26.com toubakd.com tiantianleyulecheng61.com wangziyulecheng33.com yezonghuiyulecheng82.com bocwry.com huangguantouzhuwangzhanwangzhi86.com huangguanwangquaomen29.com haiwangxingylc1664.com yinghuangylc727.com bocaiasd.com changjianggjylc.com jinmaylcoiu.com yazhougjylc.com huangguanxin2wang32.com benchixsyl.com zhucecaipiaosongcaijin.com ceoylcdf.com zhucesongcaijindewangzhan62.com aomenduchangyouxiyounaxie30.com mengtekaluoylcb.com guojihuangguanyule40.com huangguantiyupingtai93.com huangguanxianjinwangxinyu37.com aomenduchangpaixing27.com 500wanylcyu.com dajihuiylc686.com ruifengguojiyy.com makeboluoylcb.com jincaigjylc.com xindongfangylc869.com aomenduchangzainali50.com wangshangyulekaihusongcaijin.com huangguanxjwkh.com jinbangylc77.com baijialeqo.com yataigjylc.com baishenggjylcwe.com bocaigongsiqe.com wufagjylc.com moerbenylckk.com bogouylc1663.com huangguandailiwangzhi23.com bojueylcpo.com bocaiwangzhanqe.com taoataao.com bbhunas.com sjzd36.com sjpt63.com bjlkh33.com baijialebishengtouzhujiqiao20.com xijialiansaijifenbang57.com baijialeyule86.com xijiapaiming46.com aomenbaijialechangying76.com baijialeyulepingtai34.com wangshangbaijialekaihusongcaijin76.com ouzhouwudaliansaipaiming53.com wudaliansaitedian39.com baijialekaihusong50caijin17.com baijialeguize52.com zhibobazuqiuzhibo2.com zuqiubifenqiutan88.com dejiasaichengbiao88.com zuqiuba85.com mahuitqzzjw83.com sjzd01.com weixingjianting29.com cwanpp.com xingboyulezaixian86.com mwqpah.com jiankongpingtairuanjian43.com zhenqianyulechengguanwang63.com njdyyytj.com fanheer.com 999coin.com shenganna74.com jackwolfskinsalejp.com zaozhuangcq.com bjl7788.com ruhejiankongshouji2.com aomenduchangyingqianliao75.com shoujidingweichaxunruanjian12.com shoujijiantingshebei46.com aomen916.com shoujikajiantingqi77.com zhenqianyouxipaixing2.com rysevw.com wanzhenqianwangzhan36.com vrcgw.com feilvbinshengannayulecheng20.com duchangyingqianmijue81.com zzvqo.com


SMH
Troll.me

Sakula/ThreatConnect Report

1	Domain Name: TOPSEC2014.COM	1	Domain Name: TOPSEC2014.COM
2	Registry Domain ID: 1857525015_DOMAIN_COM-VRSN	2	Registry Domain ID: 1857525015_DOMAIN_COM-VRSN
3	Registrar WHOIS Server: whois.godaddy.com	3	Registrar WHOIS Server: whois.godaddy.com
4	Registrar URL: http://www.godaddy.com	4	Registrar URL: http://www.godaddy.com
5	Update Date:	5	Update Date: 2014-05-06 04:52:21
6	Creation Date: 2014-05-06 04:48:49	6	Creation Date: 2014-05-06 04:48:49
7	Registrar Registration Expiration Date: 2015-05-06 04:48:49	7	Registrar Registration Expiration Date: 2015-05-06 04:48:49
8	Registrar: GoDaddy.com, LLC	8	Registrar: GoDaddy.com, LLC
9	Registrar IANA ID: 146	9	Registrar IANA ID: 146
10	Registrar Abuse Contact Email: abuse@godaddy.com	10	Registrar Abuse Contact Email: abuse@godaddy.com
11	Registrar Abuse Contact Phone: +1.480-624-2505	11	Registrar Abuse Contact Phone: +1.480-624-2505
12	Domain Status: ok	12	Domain Status: clientTransferProhibited
		13	Domain Status: clientUpdateProhibited
		14	Domain Status: clientRenewProhibited
		15	Domain Status: clientDeleteProhibited
13	Registry Registrant ID:	16	Registry Registrant ID:
14	Registrant Name: li ning	17	Registrant Name: Top Sec
15	Registrant Organization:	18	Registrant Organization: TopSec
16	Registrant Street: guangdongsheng	19	Registrant Street: china
17	Registrant City: guangzhoushi	20	Registrant City: china
18	Registrant State/Province: Alabama	21	Registrant State/Province: china
19	Registrant Postal Code: 54152	22	Registrant Postal Code: 100000
20	Registrant Country: United States	23	Registrant Country: China
21	Registrant Phone: +1.4805428751	24	Registrant Phone: +1.82776666
22	Registrant Phone Ext:	25	Registrant Phone Ext:
23	Registrant Fax:	26	Registrant Fax:
24	Registrant Fax Ext:	27	Registrant Fax Ext:
25	Registrant Email: li2384826402@yahoo.com	28	Registrant Email: TopSec_2014@163.com

More BlueCross/Premera

Found these:

adobeupdated[.]com
gmail-msg[.]com
intel-update[.]com
vmwaresupportcenter[.]info

Didn't catch these but definitely capable:

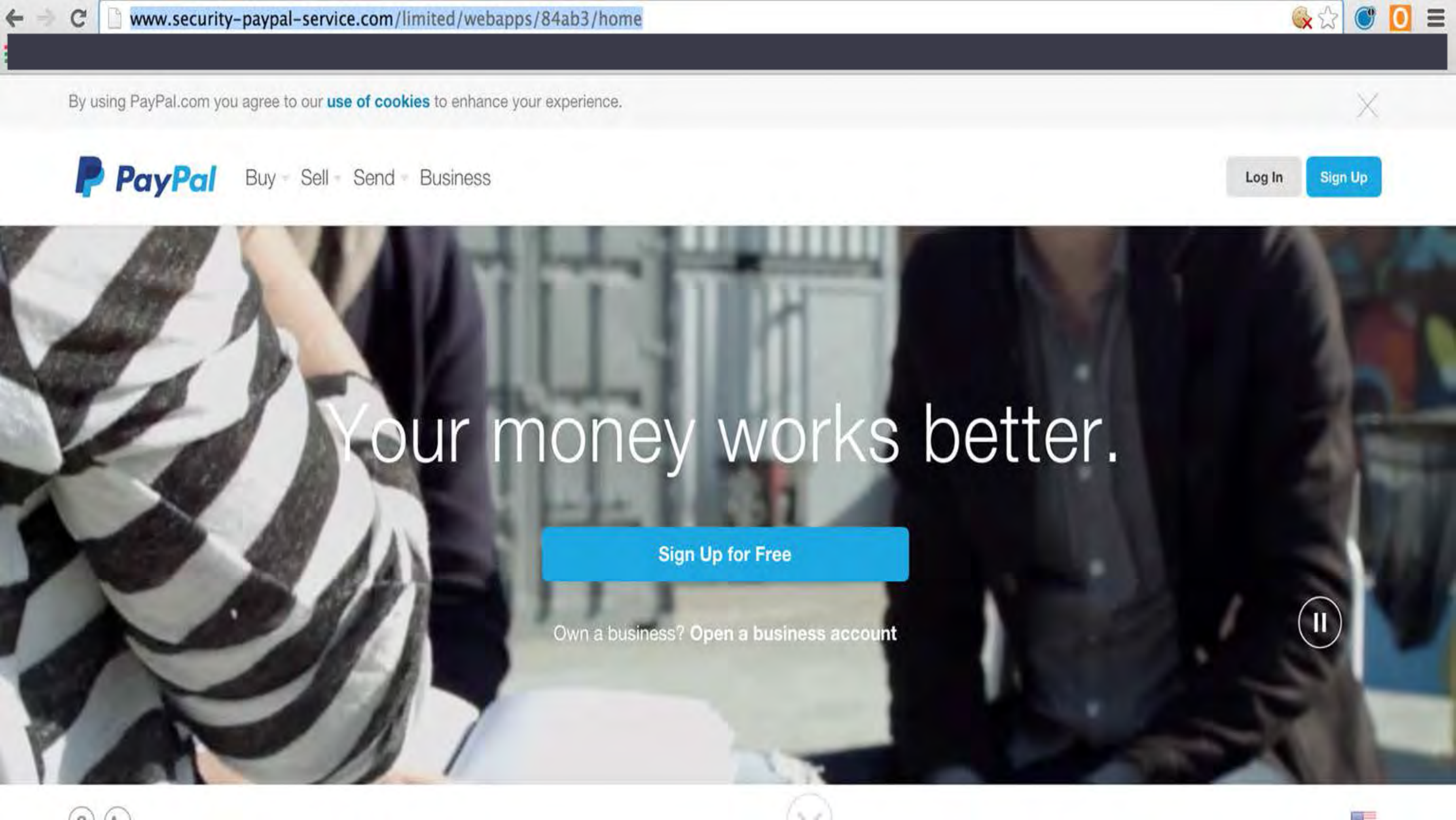
prennera[.]com
we11point[.]com.

Interesting Results

Way to filter into parked/suspended pages??

1. Parked Pages
   a. Interesting patterns among terms of parked pages, examples:
      i. www[.]iniciar-sesion-gmail[.]com
         1. Important Terms (stemmed) : fjccheck1catchexcept, click, trydocumentcooki, proceed
      ii. ww2.content.archiveofourown.orgamazon.com
         1. Important Terms (stemmed) : fjccheck1catchexcept, click, trydocumentcooki, proceed
      iii. android.clients.google.com.www.smartbrosettings.net,
         1. Important Terms (stemmed) : fjccheck1catchexcept, click, trydocumentcooki, proceed
2. Suspended Pages
   a. “Suspend” relayed as most important terms, example:
      i. FQDN: xbmcwindows[.]com
         1. Important Terms: **'suspend'**,'arial', normal, solid'

OpenDNS


www.security-paypal-service.com/limited/webapps/84ab3/home
By using PayPal.com you agree to our use of cookies to enhance your experience.
PayPal
Buy
Sell
Send
Business
Log In
Sign Up
Your money works better.
Sign Up for Free
Own a business? Open a business account


facebooklogin-facebook.com
Your computer will restart to complete these updates.
Restart
facebook Login
Email or Phone
Password
Log In
Keep me logged in
Forgot your password?
Login on Facebook
Google
Twitter
Yahoo
Hotmail

Combining Detection Models

OpenDNS Security Labs Blog > September 2015 > Phishing, Spiking, and Bad Hosting

PHISHING, SPIKING, AND BAD HOSTING

SEPTEMBER 14, 2015
BY DHIA MAHJOUB, JEREMIAH O'CONNOR, THIBAULT REUILLE AND THOMAS MATHEW

At OpenDNS Labs we have developed a number of predictive models to hunt down evil on the Internet. We have discussed in previous blogs and conferences our algorithms NLPRank [1][2][3], Spike detector [4][5][6], and malicious IP space/rogue host detectors [7][8](section 14)[9][10][11][12][13][14][15].

In this blog we will discuss how we integrate all of these detection models to improve detection coverage of current threats and walk through a few interesting examples.

PHISHING AND SPIKES

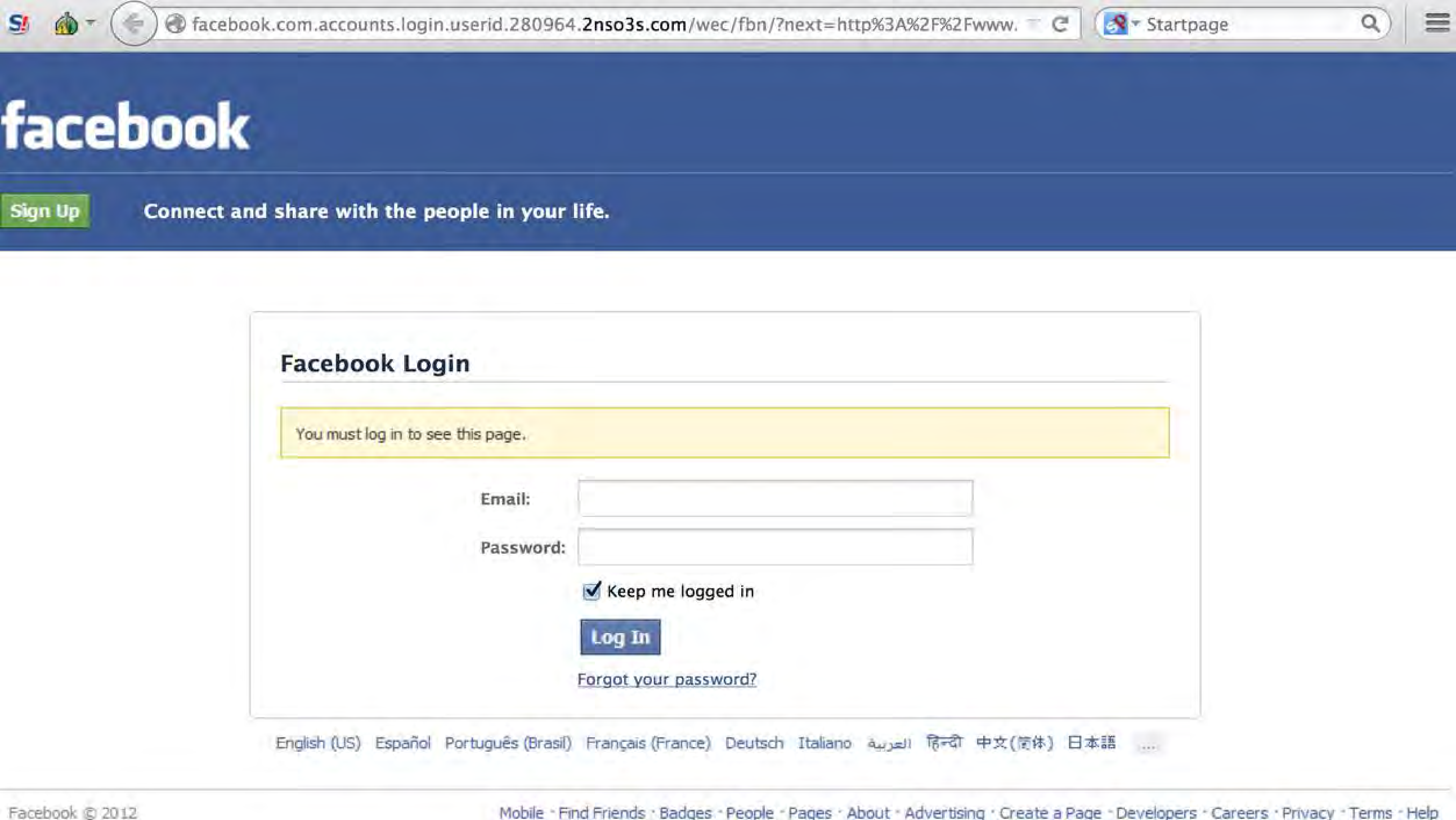
One of the recent samples we have found was a Facebook phishing campaign that was surfaced by our real-time alert system. Our model NLPRank detected the campaign of Facebook phishing sites spoofing Facebook under the second-level domain (2LD) 2nso3s[.]com.

For this particular domain, when visiting the 2LD, 2nso3s[.]com from your browser, you would be directed to a URL that looks like:

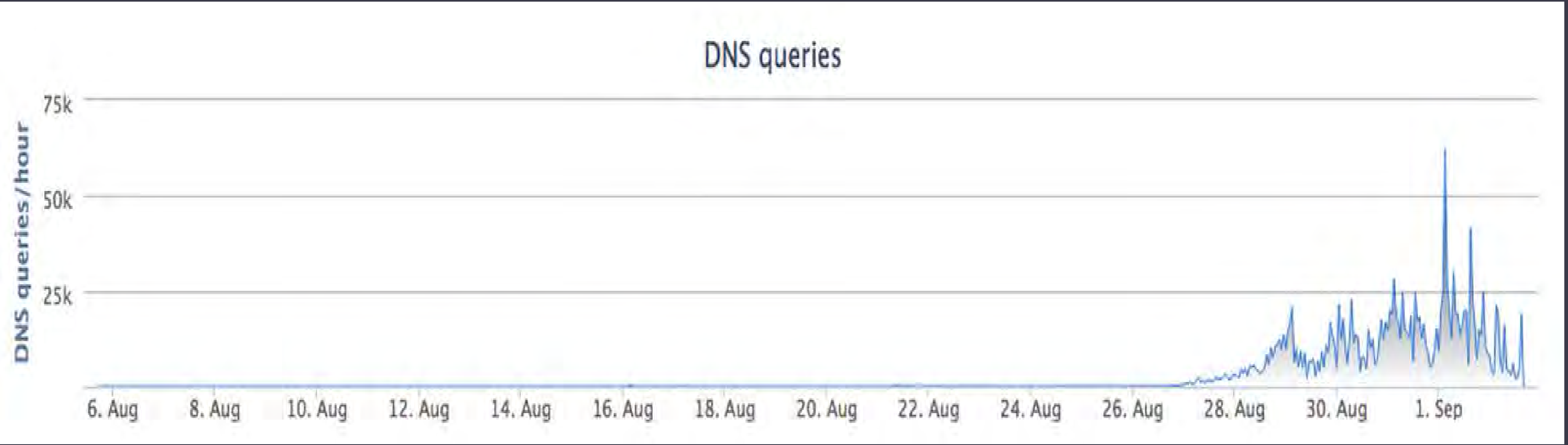
http://facebook[.]com.accounts[.]login[.]userid[.]280964[.]2nso3s[.]com/we
next=http%3A%2F%2Fwww.facebook.com%2videos%2F%3A%4A%4ID%1/

As we can see in the path of the URL the next page routes you directly to

OpenDNS


facebook.com.accounts.login.userid.280964.2nso3s.com/wec/fbn/?next=http%3A%2F%2Fwww.
Startpage
facebook
Sign Up
Connect and share with the people in your life.
Facebook Login
You must log in to see this page.
Email:
Password:
Keep me logged in
Log In
Forgot your password?
English (US) Español Português (Brasil) Français (France) Deutsch Italiano العربية हिन्दी 中文(简体) 日本語 ...
Facebook © 2012
Mobile · Find Friends · Badges · People · Pages · About · Advertising · Create a Page · Developers · Careers · Privacy · Terms · Help

Traffic for 2nso3s.com



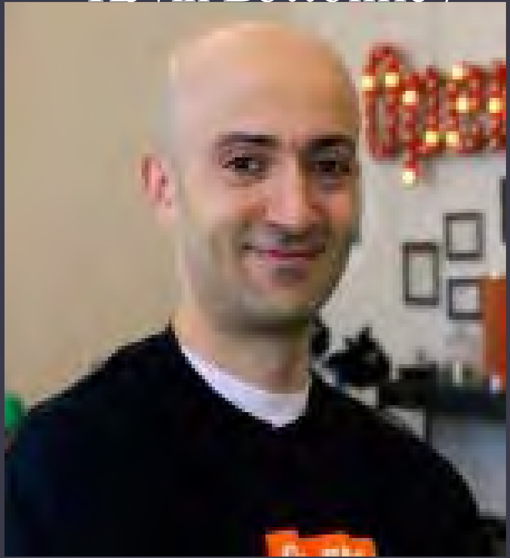
OpenDNS

PhishTank® Out of the Net, into the Tank.

Vinny Lariza

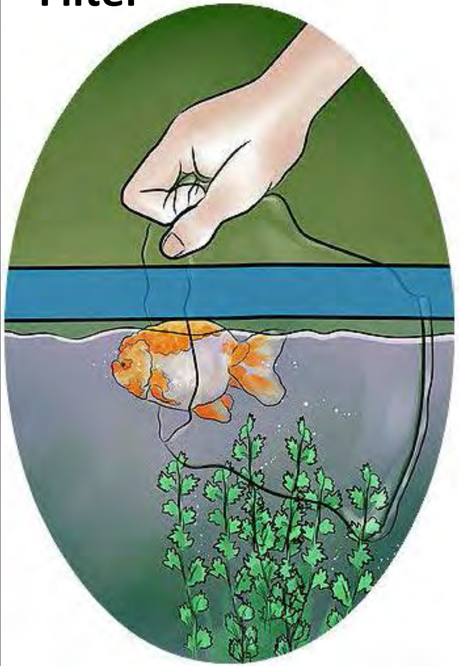
Kevin Bottomley

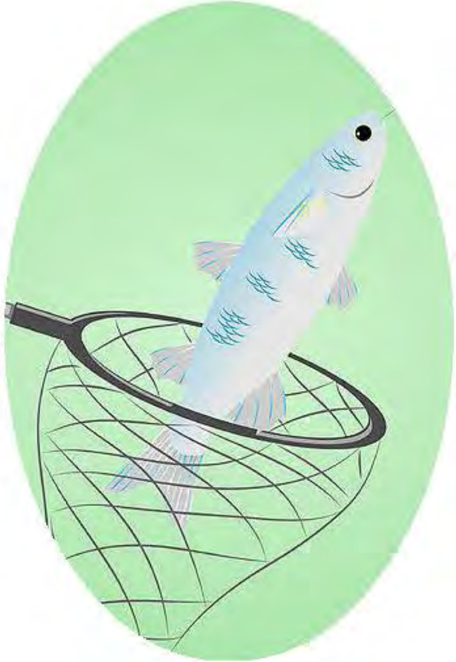
Dhia Mahjoub



How Phishtank Works

Submit ---------------------> Vote! -------------> Categorize -----------------> Filter





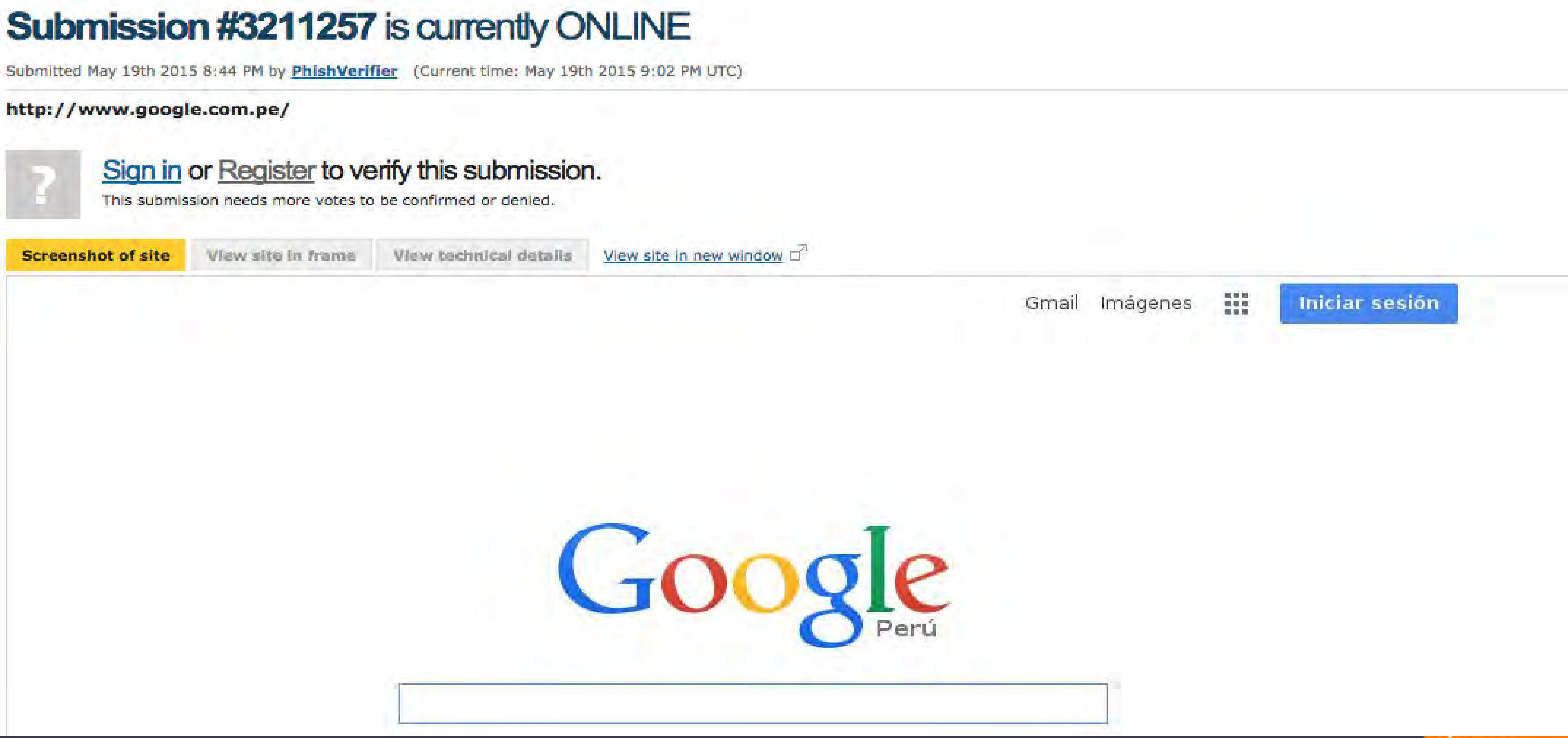
Identifying Problem

-PhishTank has Cult Following in Security Community
- People always asking about it conferences, security parties, LinkedIn etc.

-Identifying spoofed brands of phishing URL's in real-time / as they are submitted is necessary for reducing the amount of false positives in the PhishTank feed
-Reducing the amount of time from submission to approval
-IMO: Phishtank= giant training set for sec data scientists

OpenDNS

Examples of False Positives



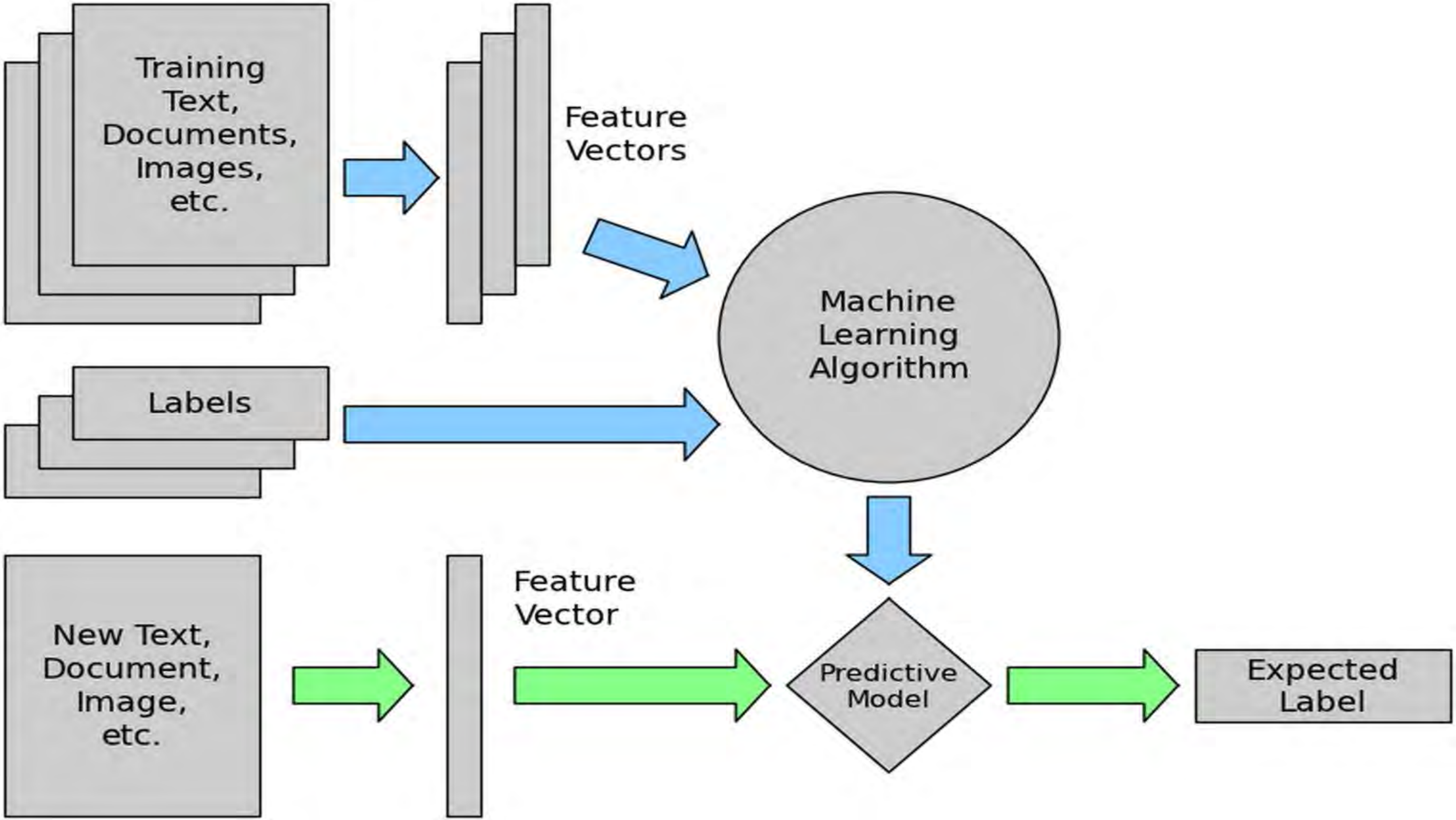
Hypothesis:

-Using IR/NLP techniques to gain a summary of the web page is a problem that has already been solved algorithmically ex. search

-Similar to way Netflix recommends movies based on user history, can we recommend what brand name the phish is by content of the page?

-Lets apply these same techniques to identify commodity phishing pages

Hypothesis: We can identify Phishing pages by using IR/Topic Modeling techniques, and auto-label Phishtank submissions as they come in


Training Text, Documents, Images, etc.
Feature Vectors
Labels
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Topic Modeling

- Methods for automatically organizing, understanding, searching, and summarizing large electronic archives.
  1. Discover the hidden themes of collection.
  2. Annotate the documents according to themes.
  3. Use annotations to organize, summarize, search, make predictions.
- Great for building recommender systems
- Used as features for a classifier

Building Corpus

-Built Corpus of HTML Content of Phishing pages, ex. WellsFargo, Paypal, Amazon, Apple, Bank of America, from Phishtank
Only Focused on Big Name Brands
- Data Collection, although at times tedious, become very intimate with the data
-See all kinds of variations of Phishes
90s Paypal vs. 2000s Paypal vs. 2015 Paypal
Christian Mingle Phishing?

TF-IDF

Input: Word Count Vector From Terms in HTML Document (Query), Word Count Matrix over a collection (Corpus)

TF-IDF - Show how important word is to a collection

Balance between: Frequency of Term and Rarity over all documents

Term-Frequency: # of times term t, appears in the document d

-Term Relevance does not increase proportional with term-frequency

Inverse-Document Frequency: the # of documents that contain term t

TFIDF - tf-weight * idf-weight

TFIDF - Increases with number of occurrences within a document, and rarity of term over all documents

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

OpenDNS

LSA/LSI

Latent Semantic Analysis: analyzing documents to find underlying concepts/meaning from them (clustering algorithm)

Uses singular value decomposition (reduce dimensionality) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text.

Hard because of variations in English language, synonyms, ambiguities

some words have different meanings when used in context

-Uses Bag of Words Model (Ordering doesn't matter)

-Using n-grams can help identify associations using co-occurrences

Helps with normalization of data

Bigrams: San Francisco -> san_francisco, Sign In -> sign_in

OpenDNS

LSA/LSI

Input: X, count matrix (or TFIDF), where m (rows) is number of terms, and n is number of documents

When we do decomposition, have to pick a value k, which represents the number of topics/concepts

Process: Decompose X into 3 matrices, U, S, V(T)

U= m x k matrix, where m = terms, k = concepts

S= k x k diagonal matrix. Elements are amount of variatior t.

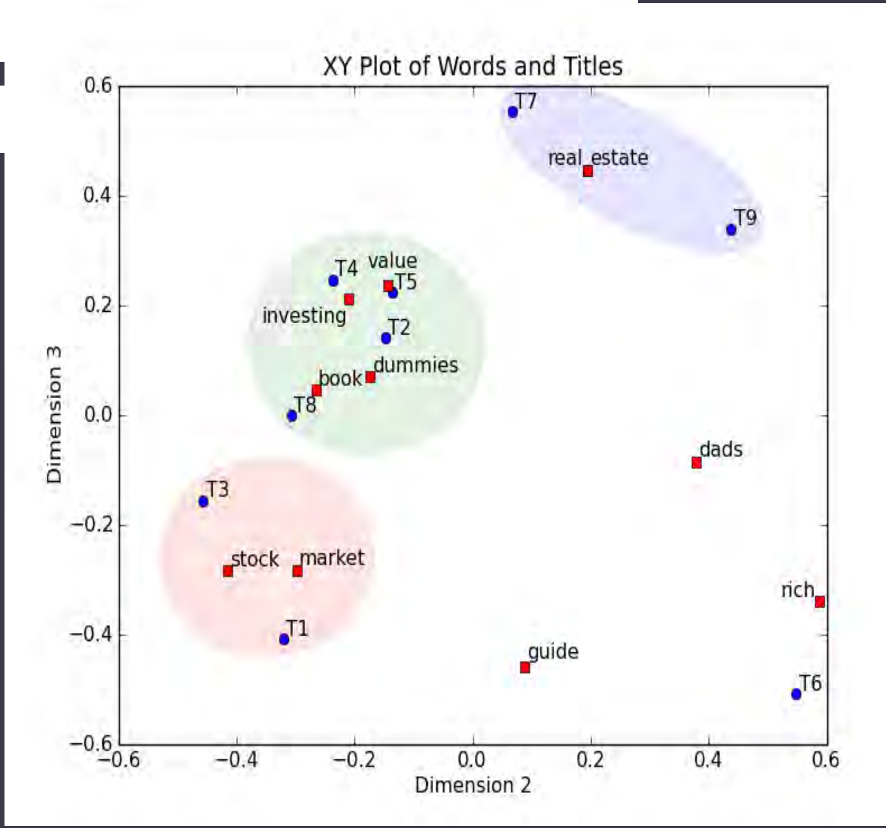
V(transpose)= k x n matrix, where k= concepts, n=docume

$$X \approx USV^T$$

LSA/LSI Example

1.
2.
3.

1.

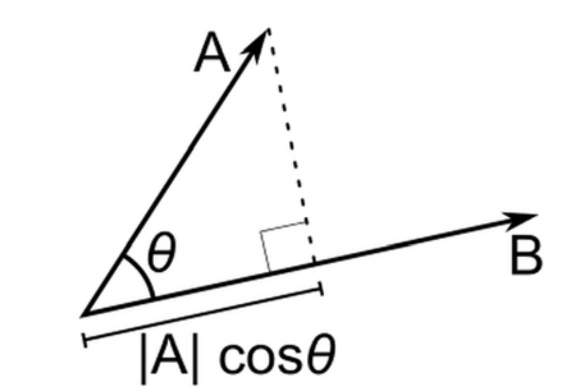


Cosine Distance

Word counts of the documents (HTML Content) form vectors
Cosine is normalized dot product of the vectors
Compute Cosine Distance from the components of the 2 vectors

i. Cosine Similarity to Phishing Pages in the Corpus
   1. Transform terms of HTML document into vectors and Corpus (Phishing) documents to vectors
   2. Find angle (Cosine Similarity) between input HTML document term vector and Corpus documents
   3. Return ranking of the sites with the most similar HTML Documents in Corpus

Cosine Distance b/t Vectors



Cosine distance between two vectors:

In[1]:= CosineDistance[{a, b, c}, {x, y, z}]

Out[1]= $$1 - \frac{a\,x + b\,y + c\,z}{\sqrt{\mathrm{Abs}[a]^2 + \mathrm{Abs}[b]^2 + \mathrm{Abs}[c]^2}\ \sqrt{\mathrm{Abs}[x]^2 + \mathrm{Abs}[y]^2 + \mathrm{Abs}[z]^2}}$$

OpenDNS

Auto-Labeling Brand Results:

Sample Output (Document Handle, Document (Cosine) Similarity Score, Brand/FQDN of URL):
Input URL/Query: WellsFargo/fitac.com.tr.html
(61, 0.99899197) WellsFargo/wellsfargo.com.html
(62, 0.99890876) WellsFargo/usam.edu.sv.html
(60, 0.9984659) WellsFargo/school76.irkutsk.ru.html
(59, 0.98146677) WellsFargo/theweddingcollection.gg.html
(63, 0.97453147) WellsFargo/exin.ba.html
Input URL/Query: Chase/www.nutrem.mx.html
(76, 0.98566723) Chase/bororooil.com.html
(75, 0.92363083) Chase/chaseonline.chase.com.html
(27, 0.92042124) BankOfAmerica/createcrafts.ph.html
(25, 0.92009199) BankOfAmerica/actautismoman.com
(74, 0.91776139) Chase/www.zac.or.tz.html

Auto-Labeling Brand Results:

Sample of Brand Names from Incoming Phishtank Stream
467 Total Samples - 78 in Corpus, 389 Test
353 hitting as Top recommendation, 18 out of remaining 36 in Top 5
Still along the same Topic/Theme, ex. (Bank/Finance, Mail, Social)
371 / 389 (With additional weighting tests, work in progress)
Some Brands have higher accuracy than others (Wells Fargo, BofA)

Auto-Labeling Brand Results:

ACCURACY: 0.989112354453
PRECISION 0.907455012853
RECALL 0.907455012853
SENSITIVIY 0.907455012853
SPECIFICTY 0.994215938303
TPR 0.907455012853
FPR 0.00578406169666
X, Y(Best 0,1) (0.005784061696658127, 0.9074550128534704)
BALANCED F1 MEASURE 0.907455012853

OpenDNS

Beyond Phishtank

-DNS data is not the ideal match for this data...HTTP traffic much better fit

Why? When doing lookups, landing on index page, most often phishing page is not on index page

-Within DNS, necessary to build crawler

Question: But there's so much traffic, are we going to do GET request for every URL???

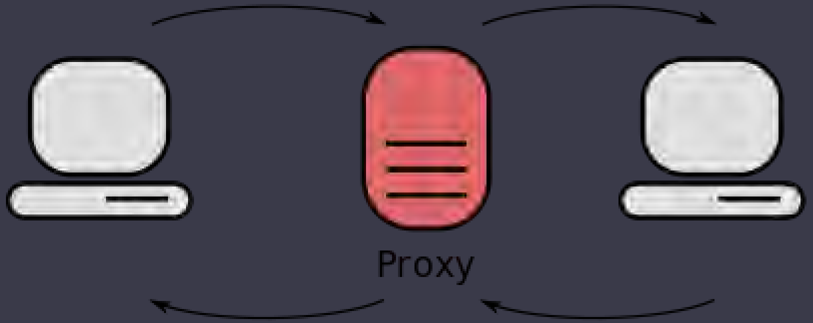
OpenDNS

OpenDNS Intelligent Proxy

What is the Intelligent Proxy?

-Awesome Team!!

- -Man in the Middle
- -Greylisting
- -Next step in OpenDNS Security



OpenDNS

Dedicated vs. Compromised Examples

Dedicated:

update-java[.]net, adobe-update[.]net, http://wellsinfo.net/login

Compromised:

Domain: wwelllssssfffarrgo.webzdarma.cz.html

http://dandraghicescu.ro/dbox/dpbx/dpbx/

http://school76.irkutsk.ru/language/Wellsfargo/online.htm

http://createcrafts.ph/bankofamerica.com.update.login.in.info/de17792ab89754c6b0a58d767a6985fc/

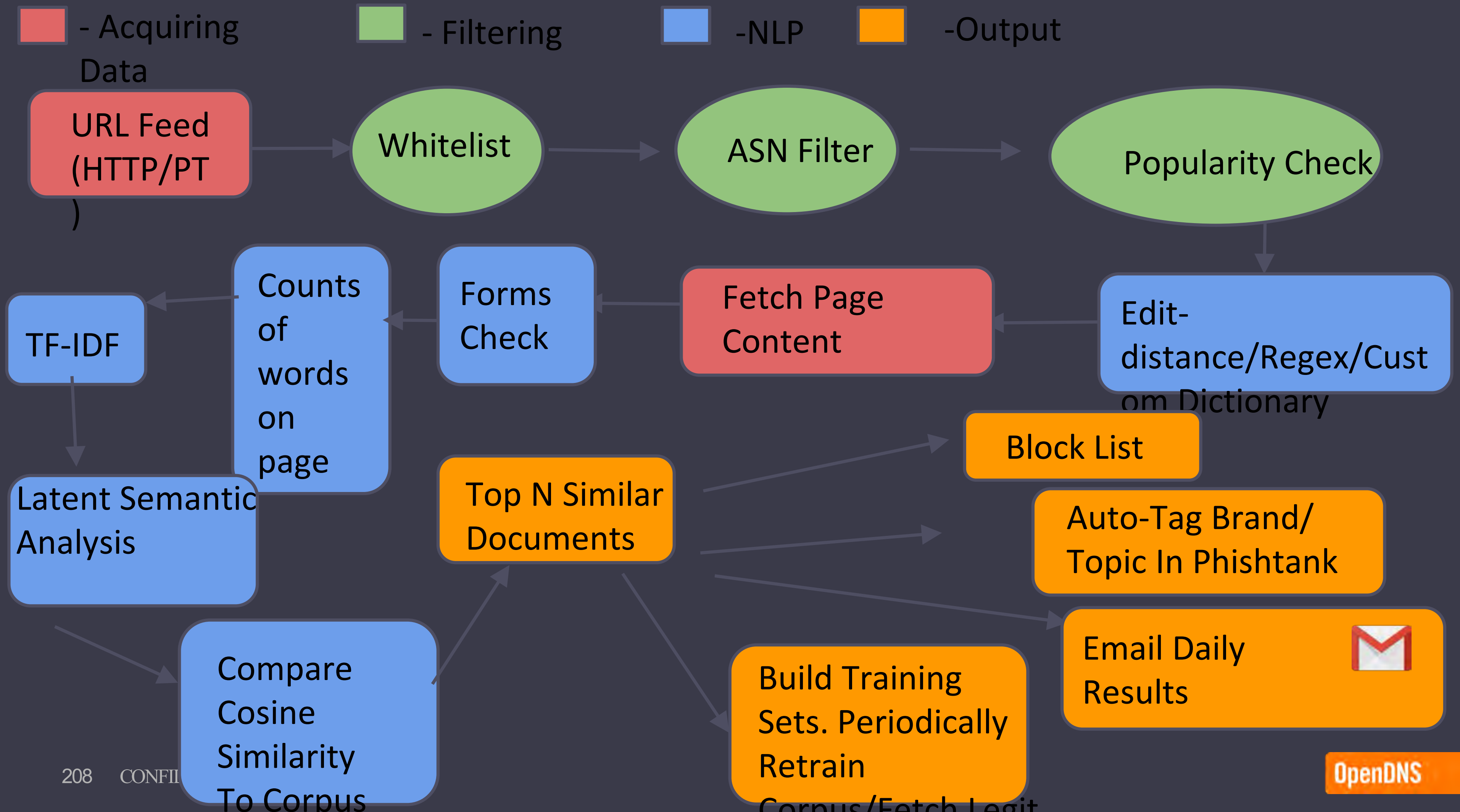
http://www.kingdomhome.com.au/wp-admin/wellsfargo.zip/wellsfargo-online.server/details.html

http://wellsfargoonline.pfwv.com.br/wellsfargo/

http://www.cityroo.com/sarasoa/wellsfargo/wellsfargo-online.php

http://wellsfargo.com.billing.account.updatemyaccount.**wellsfrago.**com.onlineaccounts.upgrade.online.billing.account.update.nlineaccounts.upgrade.online.billing.account.update.kowafdfsfs.net

http://comosecuraladiabetes.com/wp-admin/js/well.htm


- Acquiring Data
- Filtering
-NLP
-Output
URL Feed (HTTP/PT)
Whitelist
ASN Filter
Popularity Check
Edit-distance/Regex/Custom Dictionary
Fetch Page Content
Forms Check
Counts of words on page
TF-IDF
Latent Semantic Analysis
Compare Cosine Similarity To Corpus
Top N Similar Documents
Block List
Auto-Tag Brand/ Topic In Phishtank
Email Daily Results
Build Training Sets. Periodically Retrain Corpus/Fetch Legit
OpenDNS

Conclusion

- Agile Research: Building, Testing, Tuning, Iterating
- Different Algorithms, LSA as Feature
- Topic Modeling on More Content (LDA, seasons)
- More Features (SimHashing, HTML content encoding)
- Data Collection/Building Corpus
- Filtering FPs
- Spark Streaming!
- United States ODNS=-1009US0; 62/167,178

OpenDNS

OpenDNS

OpenDNS is
now part of Cisco.

cisco

QUESTIONS?

@jmoconnor415
jeremiah@opendns.com
jeoconno@cisco.com

@ThibaultReuille
thibault@opendns.com
treuille@cisco.com