Supercomputing and Malware Analysis: Solving Threat Intelligence and Context

October 10th, 2012

Agenda

- Introduction
- Context of Threat Intelligence
- Scale of the Context Problem
- Solving the Context Problem
- ThreatGRID Technology Pieces
- Case Study: A Year In the life of a MD5

Introduction

- Wes Brown
  - Chief Architect of ThreatGRID, Inc.
  - Expert in malware analytics and automated analysis.
  - Engineering manager, scientist, engineer, chief washer
- ThreatGRID, Inc.
  - Provider of Actionable Threat Intelligence
  - Cloud based platform for Malware Analysis and Correlation
  - Built by malware/SOC analyst and incident responders for malware/SOC analyst and incident responders

Context of Threat Intelligence

- Given a potential sample, determine if it is a threat to the organization.
- Analyze the sample for behavioral and static traits.
- Compare the sample's behavioral and static traits against context.
- Using context, make a threat assessment.
- Utilize context and sample traits to create actionable intelligence.
- Apply actionable intelligence to protect organization.

Scale of the Context Problem

- The threat analyst needs access to the historical data for context to determine the threat that a sample poses to his organization.
- Performing analysis on 150,000 or more samples a day and storing the context to perform threat evaluations against.
  - ~5 million samples a month.
  - Billions of contextual traits a month.
  - Beyond the in-house capabilities of most organizations.

Requirements for a Solution

- Scale to analyzing hundreds of thousands of samples.
- Provide users with a timely analysis for near-realtime actionable intelligence.
- Coverage and Accuracy
  - Capture all transient activity possible.
  - Store all analysis artifacts generated per session.
  - Use multiple sources of data per session to correlate and counter evasion techniques.
  - Store traits in a fashion that is relatable and responsive.

What 100,000 Samples Mean

- Dynamic analysis at a rate of one sample a second.
  - One VM provisioned, started up, and terminated **every second.**
  - Postprocessing of multiple data sources and in-session correlation.
  - **600mb-1gb** of raw data produced **every sample**.
  - **1 petabyte** of raw data every 24 hours.
- Converting raw data into observations and traits.
  - Distill **1 petabyte** of data into analytics for database
  - Convert into up to 30,000 indexed rows *per* session.
  - **Half a billion rows** per day!

What 100,000 Samples Mean

1 petabyte of raw data = two full VNX racks **a day!**

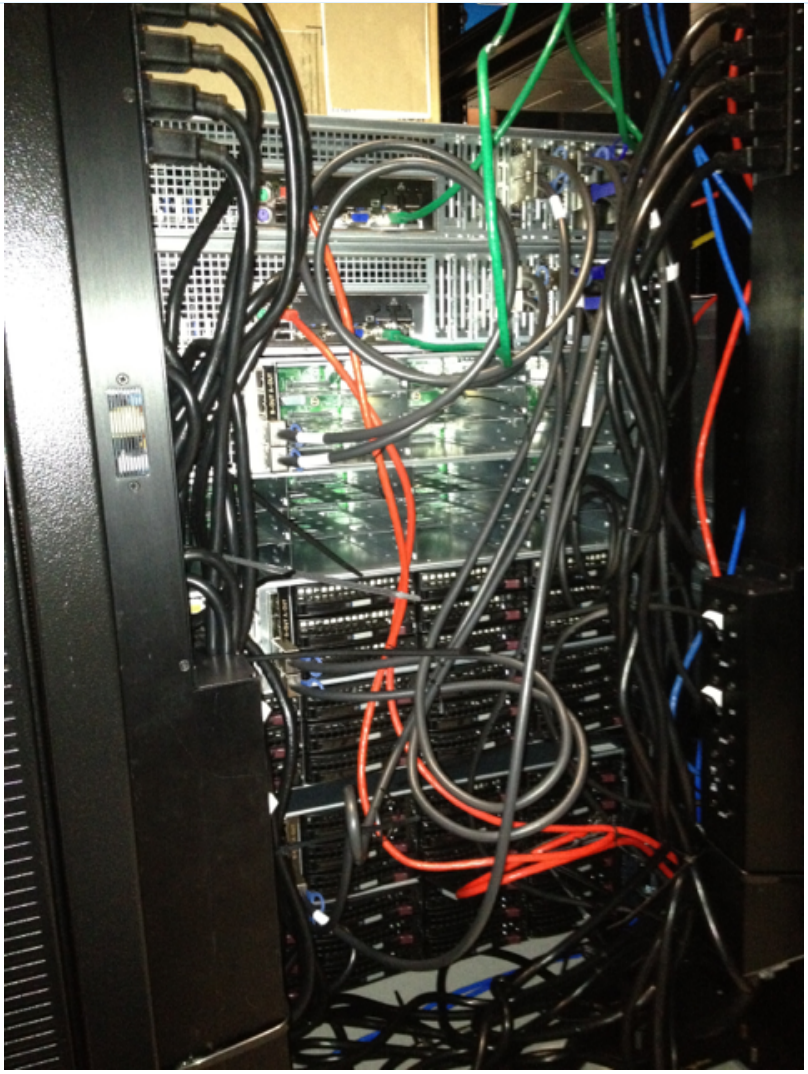
The Solution

Build a High Performance Computing cluster

- Our own in-house supercomputer!
  - Goal: Break into TOP500 list of fastest supercomputers.
- Scales up to ~1 million samples dynamically analyzed in 42u of rack space.
  - ~4,000 cores
  - 60 kw of power
- 40gbps Infiniband interconnect
  - Mesh topology - every node has a connection to every other node.
  - 80 terabit per second backplane throughput capabilities.
- 500 tb every 42u of rack space
  - 10GB+/sec of I/O capabilities per 42u, saturating Infiniband

What 100,000 Samples Looks Like

What High-I/O Storage Looks Like



What This Means For You

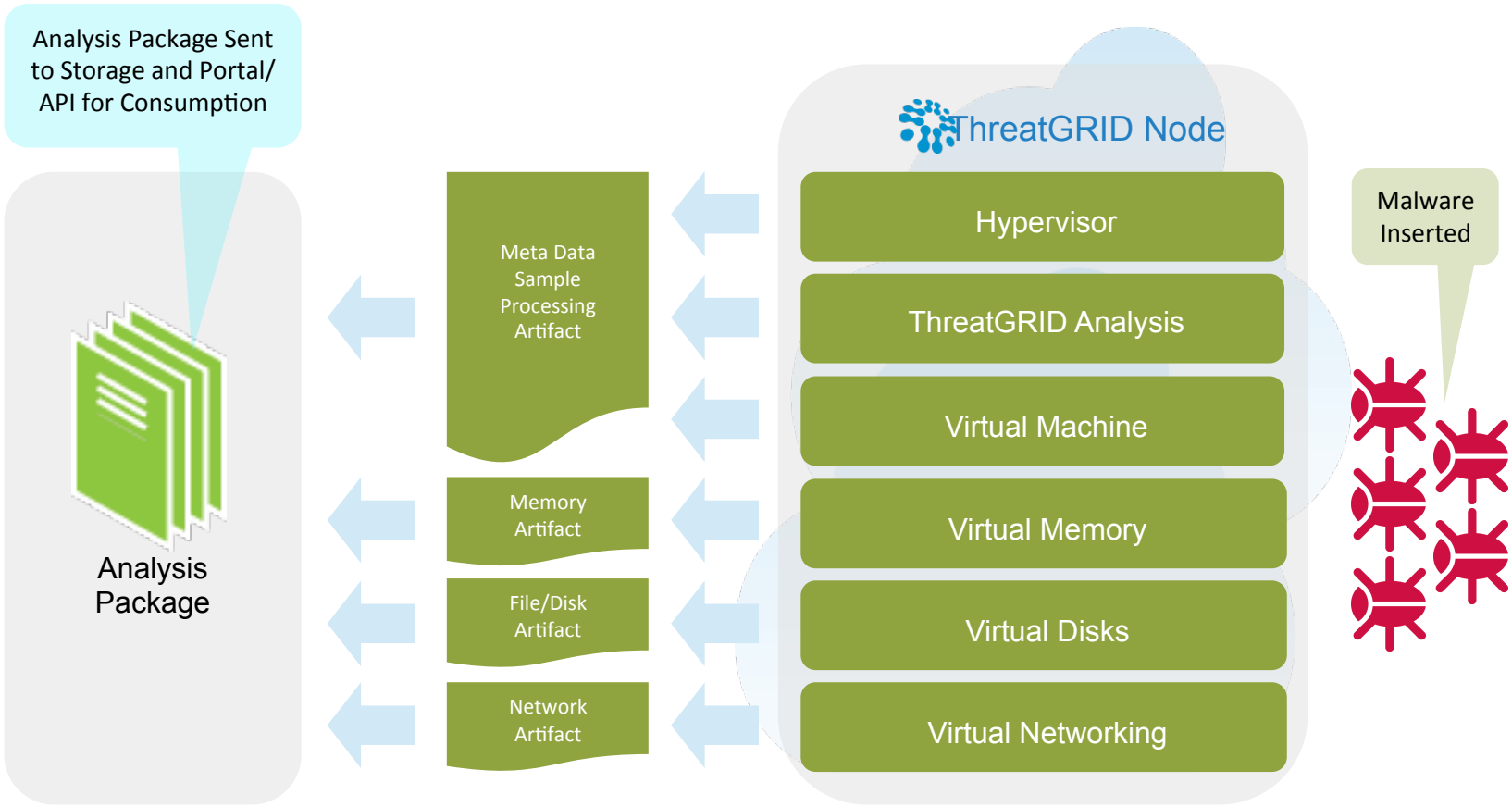
- Access to Supercomputing Resources
  - Dedicated to malware analysis and correlation.
  - You can submit samples of your own for correlation.
- Access to Data Correlation
  - Terabytes of data.
  - Trillions of rows of correlation between malware samples.
- Access to API
  - Integrate into your own infrastructure however you want.
  - RESTful API
    - Well documented
    - Multiple query parameters
- **Access to all analysis artifacts**

Technology Discussion

Technology Flow

- Threat Intelligence
  - Inbound Information
    - Blacklists, whitelists
    - Raw Malware Samples
      - Obtained from our own collectors, partners, customers and other feed sources
  - Processing
    - Digestion of inbound information
    - Processing of Malware Samples (Sandbox)
  - Correlation and Enrichment
    - Using information from multiple sources to enrich
  - Outbound Information
    - Individual Malware Sample Reports
    - Outbound feeds to subscribers

Artifact and Analysis Generation



ThreatGRID Kernel Monitor

- Kernel Monitor
  - Custom Windows NT kernel debugger
    - Programmable and scriptable
  - Undetectable via debugger detection
    - Do not use typical debugging techniques that are detectable.
  - Captures and logs system activity
    - Process activity
    - Registry, disk, network activity
  - High performance
    - VMs run at real time performance.
    - Hundreds of thousands of debugger exceptions during a session.

Block-level Disk Analysis

- Block-level Disk Analysis
  - All filesystem changes are written to a separate file
  - Parses NTFS filesystem for changes and extraction
  - Parses MBR and partition tables and detects changes
  - Detects changes that do not map to NTFS filesystem
    - Detects rootkits that hides things in raw areas of disk
  - Extracts to an archive all files changes for further analysis
  - Analyzes disks faster than CHKDSK!

Sample Processing

- Virtualization Environment
  - Does not use common virtualization platform
    - No debuggers
    - No special hooking DLLs
  - Does not tamper with or modify the OS
    - Standard Windows install with supporting applications
    - Support for multiple virtual machines and types
    - No Instrumentation in the virtual machine

Sample Processing

- Preserves all transient artifacts generated by malware
  - VM Snapshot
    - CPU and process state
  - Memory Dump
    - Memory artifacts at time of dump
  - Network Traffic
    - All network traffic (PCAP) generated by virtual machine
  - Filesystem
    - Filesystem changes at block disk level
    - Changed or added files can be extracted
  - Process Activity
    - Kernel system calls
    - Registry changes
    - Socket

Artifact Storage

- Permanently archives all sample artifacts
  - Stored in a custom in-house archive format
    - Efficient enough that **all** sample runs are stored and archived.
  - Efficient delta and compression algorithms
  - Can be retrieved and reprocessed
    - Database and information updated as processing technology improves
    - Better correlation of historic information with current trends

Filetype Support

- Supporting additional filetypes by creating handlers for each filetype
  - Support for
    - Windows PE Executable
    - Windows PE DLLs
    - Adobe PDF
    - Java JAR files
    - Flash
    - Microsoft Office
  - Supports archive formats
    - ZIP
    - Quarantine formats

Analysis JSON

- JSON
  - Direct Serialization of Data Structures
    - XML wraps lots of metadata
  - Moderately Human Readable
    - Skilled analyst can read the raw JSON
    - XML is not readable.
  - Machine readable
    - Import into your own dataset.
- Analysis JSON
  - Contains all analysis.
  - Specification available.

Case Study:
A year In the Life of a MD5

A Year In the Life of a MD5: Intro

- Malware is not static!
  - Behaviors can change day to day.
  - A session capture is a **snapshot** of behaviors that day.
  - Many intelligence vendors evaluate whether a given hash is ‘good’ or ‘bad’.
    - The **same hash** can be viewed as **bad** on one day, and trigger indicators of compromise.
    - The **same hash** can be **good** on another day and not trigger indicators of compromise.
    - A **known good** sample can change to a **unknown bad** sample, and if it is whitelisted, it will slip through the cracks.

A Year In the Life of a MD5: Sample

- IRC Test Sample
  - Internally called ‘irc-test.exe’
  - Discovered when searching PCAP output files from sandbox for IRC traffic to validate internal IRC protocol dissection code.
  - Uses IRC for command and control.
  - Originally **not detected** by antivirus.
- Basic Characteristics
  - Simple dropper
  - Uses IRC to obtain URLs to download and execute.

A Year In the Life of a MD5: Dropper

- Dropper
  - Drops different artifacts almost daily.
    - Zeus, Bugat, Virut, etc…
  - Each artifact behaves differently.
    - C&C, Persistence, Weakening, Obfuscation, etc…
  - Uses public IRC networks.
    - Long shelf life – HTTP Command and Control easy to take down.
- The Gift that Keeps Giving
  - Every run that drops a different artifact.
    - Generates new traffic to different networks.
    - Generates new behaviors to analyze.
    - New evasion techniques discovered.
  - Golden Goose

A Year in the Life of a MD5: AV

- Antivirus matches on Artifacts Dropped By Sample

W32.Virut.ca	5244
Trojan.Agent-291320	2825
W32.Virut-10	2774
W32.Virut.da	2105
W32.Virut.ci	527
W32.Virut.di	380
W32.Virut.sa	225
W32.Virut.ia	215
W32.Virut.Gen.D-148	152
Trojan.Agent-270551	143
W32.Trojan.Adload-8	78
W32.Virut.ii	67
Trojan.Downloader-130866	54

A Year in the Life of a MD5: SSDEEP

- Ssdeep was performed comparing every artifact to every artifact produced by this sample.
  - **~5million hits using ssdeep.**
  - **5 billion comparisons.**
    - **Done in 4 hours on one cluster node.**
- Used to correlate from known antivirus to discover related families that are not
- Too much data to display on this 8GB Core i7 MacBook!

.

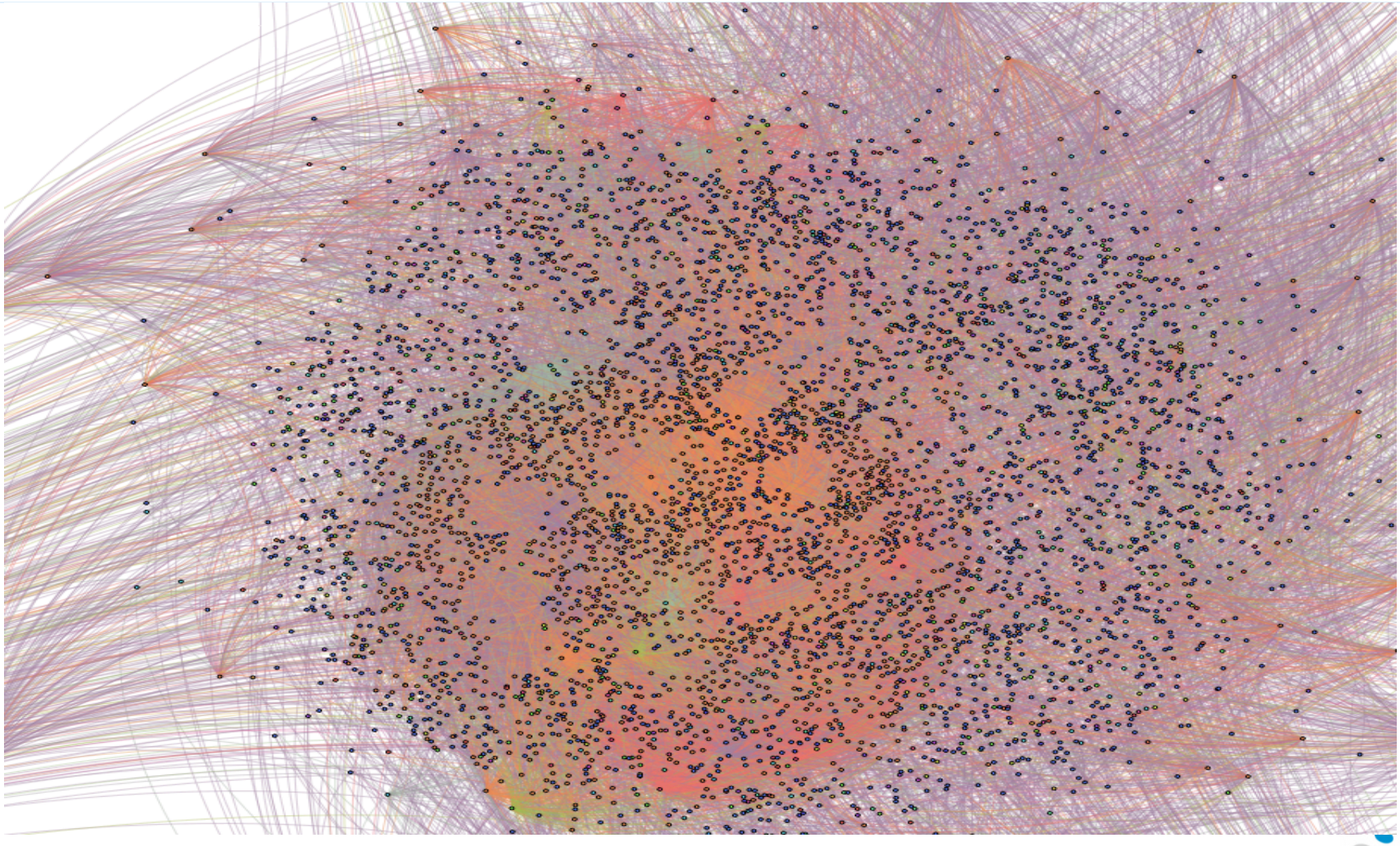
A Year In the Life of a MD5: Dropper

- Today
  - Contacts several different IP addresses.
  - Downloads a few artifacts.
- Past
  - Additional IP contacted not contacted today.
  - Different files dropped
    - Different SHA256
    - Different filenames
    - Different behaviors

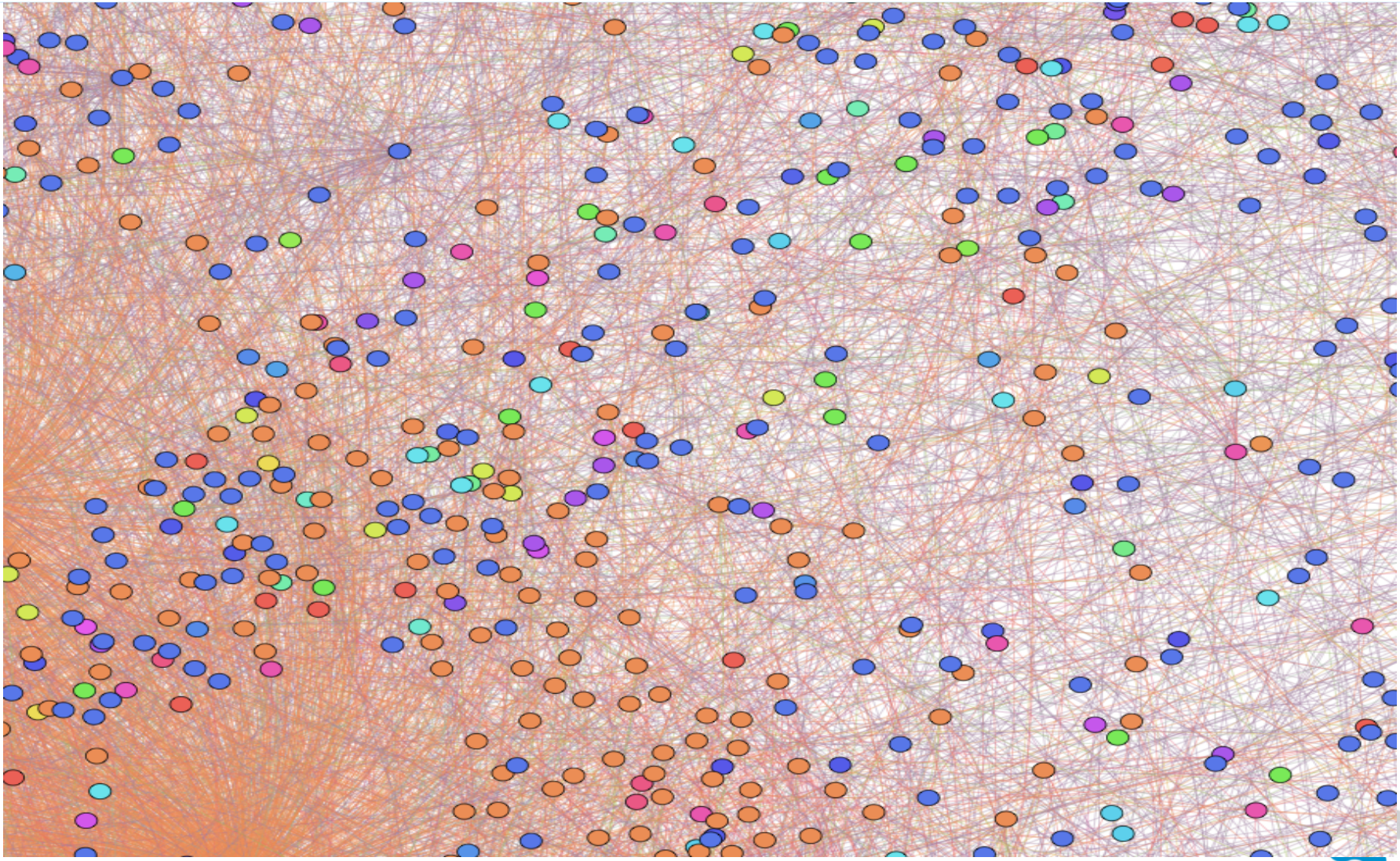
A Year in the Life of a MD5: Net

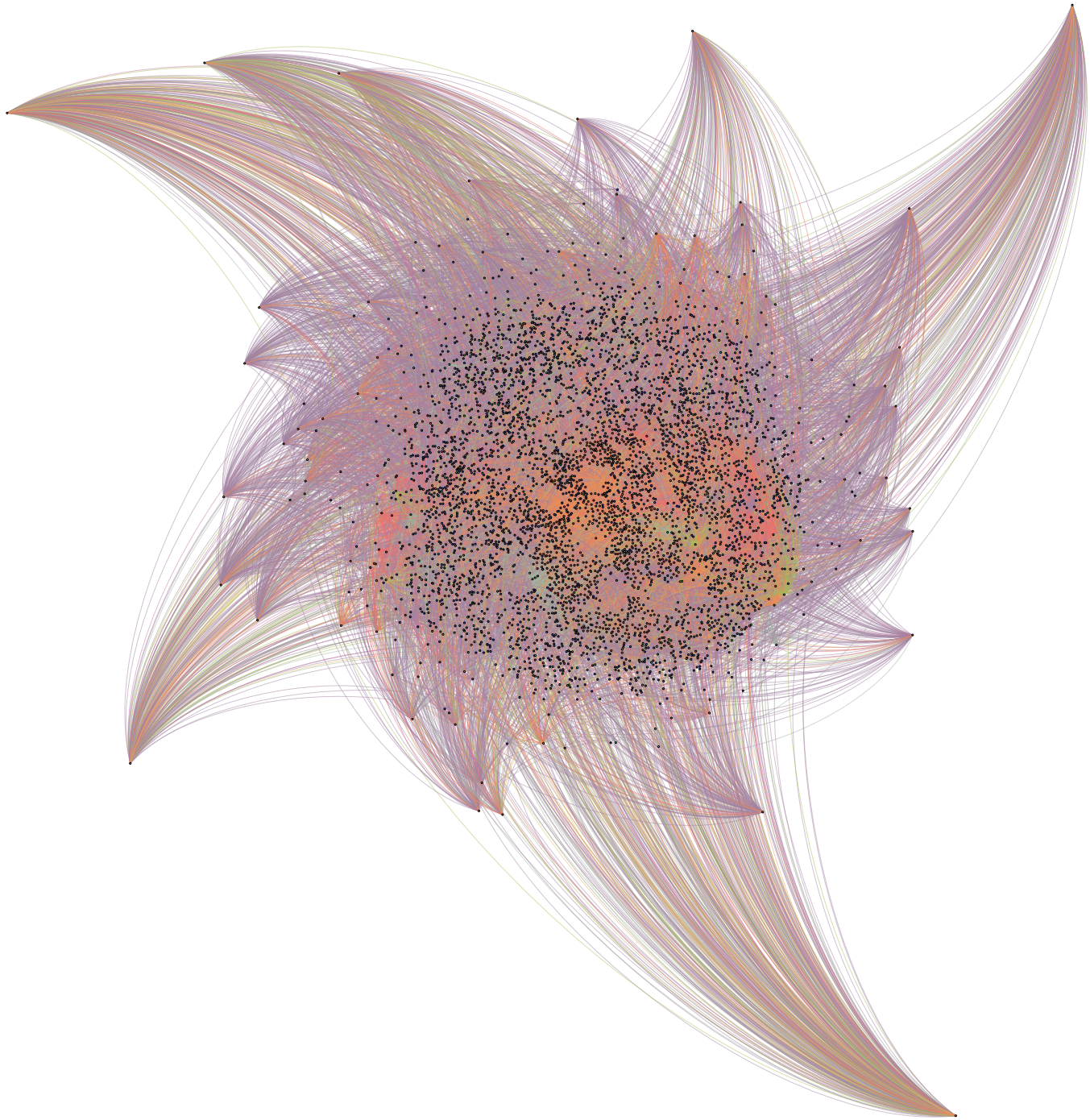
- 3653 Distinct IP Addresses
- More than 50 Countries
  - Hong Kong, Romania, Russia, Kazakhistan, Ireland, South Korea, United States
- Visualization of:
  - Distinct IP address – Node Circle
  - Country of Origin – Color of Node Circle
- Working on adding visualizations like this as a standard feature.

A Year in the Life of a MD5: Net



A Year in the Life of a MD5: Net





A Year in the Life of a MD5: Correlation

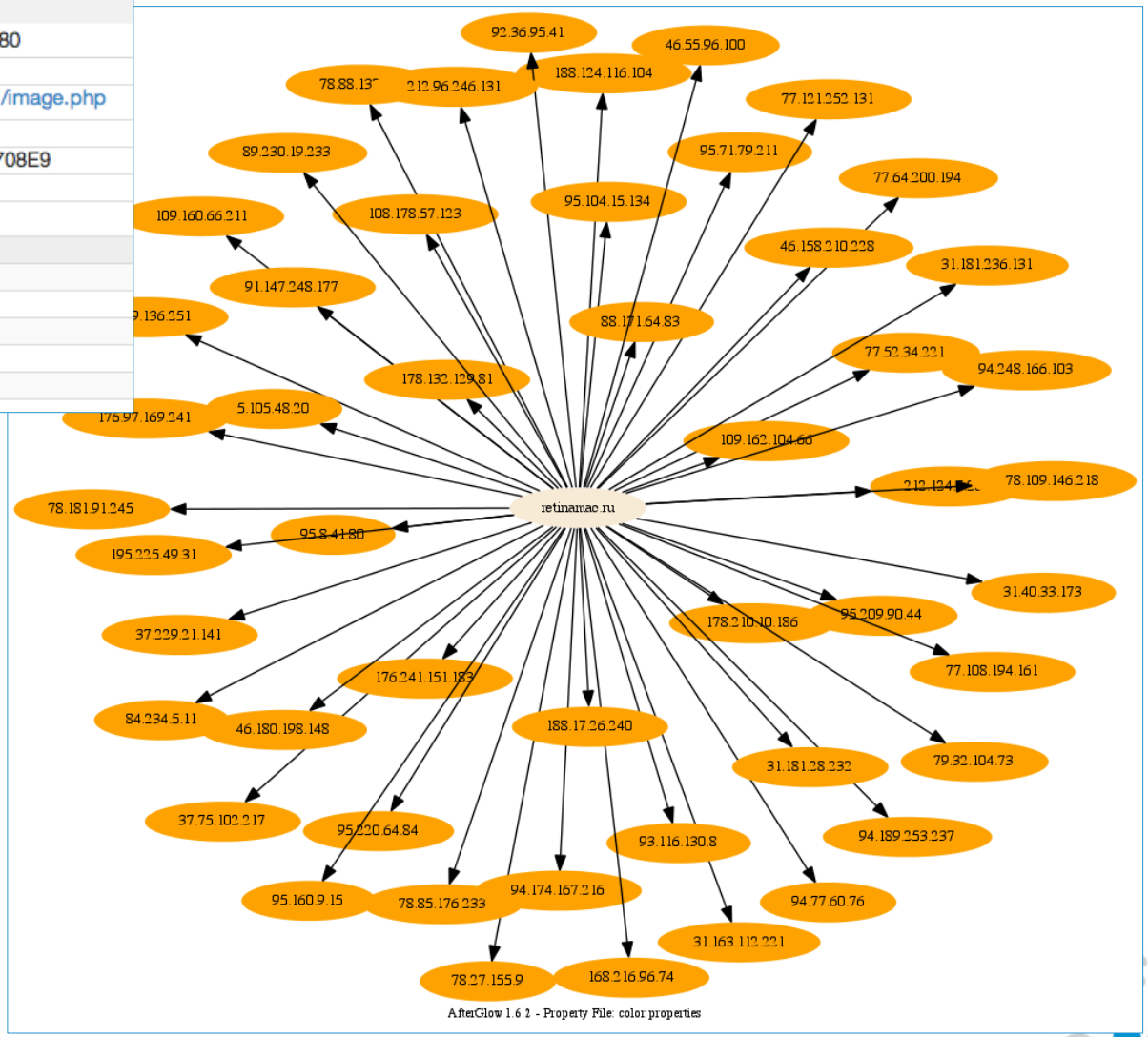
[-] POST http://retinamac.ru:80/and/image.php

Server IP: 78.27.155.9 **Server Port:** 80

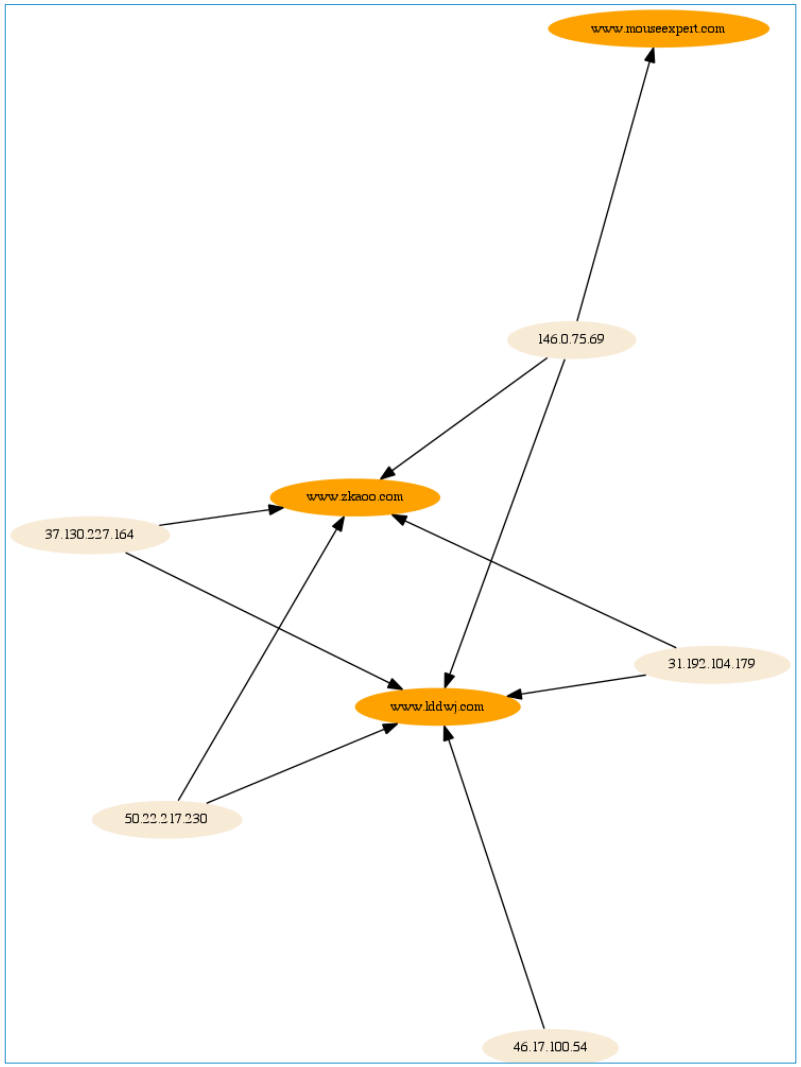
Method	POST
URL	http://retinamac.ru:80/and/image.php
Request	
Timestamp	1.349053256819708E9
Actual Encoding	ascii
Actual Content-type	text/plain

Header	Value
content-length	84
content-type	application/x-www-form-urlencoded
user-agent	Mozilla/4.0
host	retinamac.ru
connection	close

retinamac.ru



A Year in the Life of a MD5: Correlation



Domain: www.lddwj.com	
Name	www.lddwj.com
Sha256	732daa4b7b8ce54cb10ad8c5b32c3ac71f148e3a7f09d607dcf2a83b7881e1ce
MD5	511712c695cb250ba0fccbb55c15dc28

Related IPs View All

IP	Last Seen
37.130.227.164	10/8/12 21:05:27
146.0.75.69	9/5/12 20:44:16
46.17.100.54	8/3/12 17:47:21
31.192.104.179	7/9/12 17:29:50
1.1.1.1	4/19/12 01:58:50
50.22.217.230	4/12/12 19:18:24

www.lddwj.com

A Year In the Life of a MD5: Drilling Down

Domain: humanbodyfitness.com

Name	humanbodyfitness.com
Sha256	85b803700a2d354744a4ed36c73e7d86e39709da6db003a36beed001f7e8cd6f
MD5	c34aa9a32b810705b768c77818b0372a

Related IPs View All

IP	Last Seen
216.57.210.200	10/3/12 20:59:37

Hosted URLs View All

URL	Last Seen
http://humanbodyfitness.com:80/	Unknown
http://humanbodyfitness.com:80/unavailable.htm	Unknown
http://humanbodyfitness.com:80/exitjs.php	Unknown

Related Samples View All

Sample ID	Sha256	Relation	Time
23e59966ee81fc6a798a1a892684bf50	7b2b027289297b04...	http-requests	10/3/12 20:59:37
23e59966ee81fc6a798a1a892684bf50	7b2b027289297b04...	dns-lookup	10/3/12 20:59:37
9e92baaa48d9c8010f44f5571b5b2b05	7b2b027289297b04...	http-requests	10/1/12 22:59:45
9e92baaa48d9c8010f44f5571b5b2b05	7b2b027289297b04...	dns-lookup	10/1/12 22:59:45
132ae972c261e6eda69e69035858b909	7b2b027289297b04...	dns-lookup	8/28/12 18:59:05
132ae972c261e6eda69e69035858b909	7b2b027289297b04...	http-requests	8/28/12 18:59:05

A Year In the Life of a MD5: Correlation

Domains related to 216.57.210.200
Domain
funcarreferee.com
gluelaw.com
i.dotzup.com
diabeticdietplanmenu.com
www.moonslot.com
getnewcarquote.com
clapslot.com
seemslot.com
relieveemotionalpain.com
whomslot.com
humanbodyfitness.com
www.diabeticdietplanmenu.com
diabeticweightlossmenu.com
leaseprivatejet.com
dietplanscholesterol.com
privatejetsrent.com
www.clothescostume.com
bodyjewelrybuyer.com
marriagejudgment.com
solegame.com
myspahealth.com
lumpgame.com
manyslot.com
diabeticrecipesmenu.com
paymentpanel.com
relieveconstipationpain.com
biosolarfuel.com

A Year In the Life of a MD5: Drilling Down

IP: 83.133.119.197

ASN	13237 -- European Backbone of LambdaNet
Country	DE
Region	
City	

Related Domains View All

Domain	Last Seen
ilo.brenz.pl	10/7/12 13:23:11
irc.zief.pl	10/7/12 12:35:51
proxim.ircgalaxy.pl	10/7/12 06:24:00
f1.varpo.ru	10/7/12 04:04:01
n2.rolmi.ru	10/7/12 02:18:54
ru.brans.pl	10/7/12 02:03:09
dml.mlix.ru	10/7/12 02:01:42
sys.zief.pl	10/7/12 01:29:05
izc.idet.pl	10/7/12 01:27:04
mk.gimbs.ru	10/7/12 00:43:29

URLs View All

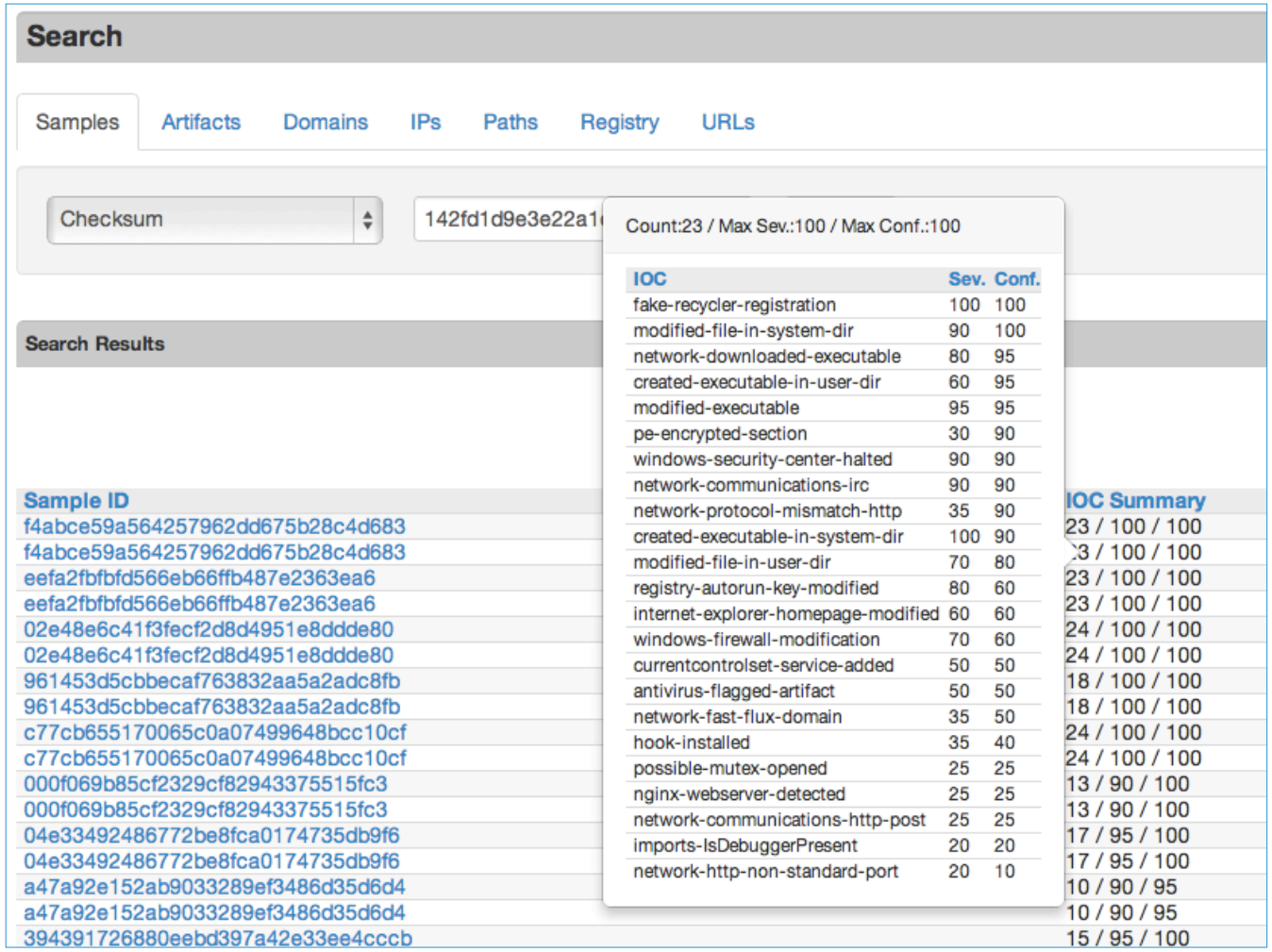
URL	Last Seen

Related Samples View All

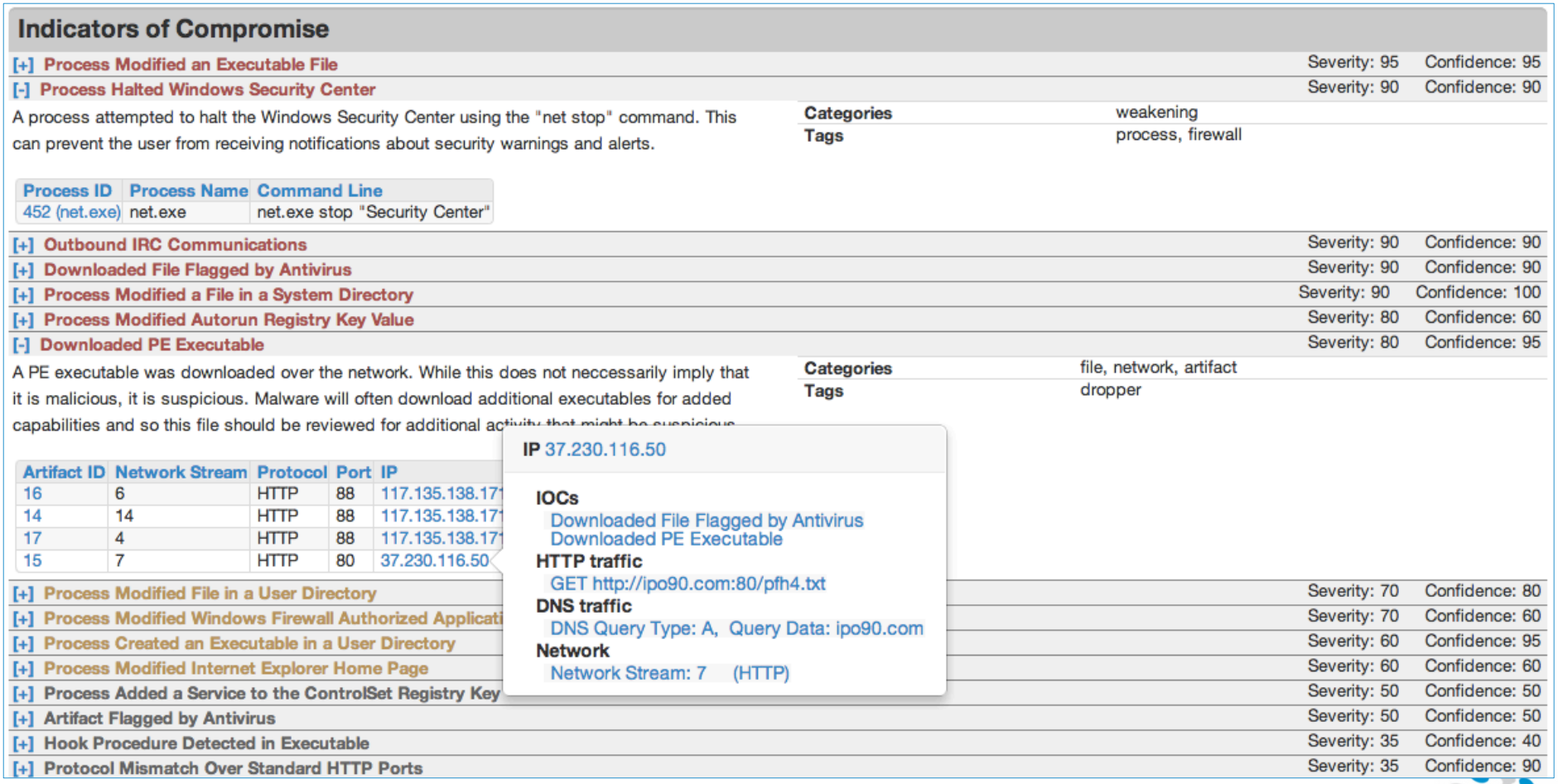
Sample ID	Sha256	Relation	Time
d669f3ca68dbf1ba41f66e312c64f619	e58885cde7143193...	network-stream-destination	10/7/12 13:23:11
d669f3ca68dbf1ba41f66e312c64f619	e58885cde7143193...	dns-lookup	10/7/12 13:23:11
4564928df523f67ea68a5ea4a71efed2	c23bec415390a0de...	dns-lookup	10/7/12 12:35:51
4564928df523f67ea68a5ea4a71efed2	c23bec415390a0de...	network-stream-destination	10/7/12 12:35:51
112832131286a32dad3dfc3362c33ea9	79f0faae9ae0f0a6...	dns-lookup	10/7/12 08:59:01
112832131286a32dad3dfc3362c33ea9	79f0faae9ae0f0a6...	network-stream-destination	10/7/12 08:59:01
d0a39fb464b690289937488476903fea	b4230ed6977cd48f...	network-stream-destination	10/7/12 06:24:00
d0a39fb464b690289937488476903fea	b4230ed6977cd48f...	dns-lookup	10/7/12 06:24:00
ad3ee89533cdf3b17e8442ea6f9cb9af	853249dcfb3a1725...	network-stream-destination	10/7/12 04:04:01
ad3ee89533cdf3b17e8442ea6f9cb9af	853249dcfb3a1725...	dns-lookup	10/7/12 04:04:01

Different Submitted Samples

A Year In the Life of a MD5: Indicators



A Year In the Life of a MD5: Indicators



Indicators of Compromise

[+] Process Modified an Executable File	Severity: 95	Confidence: 95
[-] Process Halted Windows Security Center	Severity: 90	Confidence: 90

A process attempted to halt the Windows Security Center using the "net stop" command. This can prevent the user from receiving notifications about security warnings and alerts.

Categories	weakening
Tags	process, firewall

Process ID	Process Name	Command Line
452 (net.exe)	net.exe	net.exe stop "Security Center"

[+] Outbound IRC Communications	Severity: 90	Confidence: 90
[+] Downloaded File Flagged by Antivirus	Severity: 90	Confidence: 90
[+] Process Modified a File in a System Directory	Severity: 90	Confidence: 100
[+] Process Modified Autorun Registry Key Value	Severity: 80	Confidence: 60
[-] Downloaded PE Executable	Severity: 80	Confidence: 95

A PE executable was downloaded over the network. While this does not neccessarily imply that it is malicious, it is suspicious. Malware will often download additional executables for added capabilities and so this file should be reviewed for additional activity that might be suspicious.

Categories	file, network, artifact
Tags	dropper

Artifact ID	Network Stream	Protocol	Port	IP
16	6	HTTP	88	117.135.138.171
14	14	HTTP	88	117.135.138.171
17	4	HTTP	88	117.135.138.171
15	7	HTTP	80	37.230.116.50

IP 37.230.116.50

IOCs
- Downloaded File Flagged by Antivirus
- Downloaded PE Executable

HTTP traffic
- GET http://ipo90.com:80/pfh4.txt

DNS traffic
- DNS Query Type: A, Query Data: ipo90.com

Network
- Network Stream: 7 (HTTP)

[+] Process Modified File in a User Directory	Severity: 70	Confidence: 80
[+] Process Modified Windows Firewall Authorized Applicati	Severity: 70	Confidence: 60
[+] Process Created an Executable in a User Directory	Severity: 60	Confidence: 95
[+] Process Modified Internet Explorer Home Page	Severity: 60	Confidence: 60
[+] Process Added a Service to the ControlSet Registry Key	Severity: 50	Confidence: 50
[+] Artifact Flagged by Antivirus	Severity: 50	Confidence: 50
[+] Hook Procedure Detected in Executable	Severity: 35	Confidence: 40
[+] Protocol Mismatch Over Standard HTTP Ports	Severity: 35	Confidence: 90

A Year In the Life of a MD5: Lessons

- IRC Protocol Disassembly
  - Command and Control
- Handling extreme static and disk cases
  - Thousands of PE files dropped.
  - Increased PE analysis performance.
  - Increased disk analysis performance.
- Handling extreme network cases
  - Thousands of network streams.
  - Improved performance.
- Handling evasion
  - PE disassembler bomb attack

Finis

Questions?

- Wes Brown
- Chief Architect, ThreatGRID, Inc.
- wes@threatgrid.com