Internet Special Ops

Stalking Badness Through Data Mining

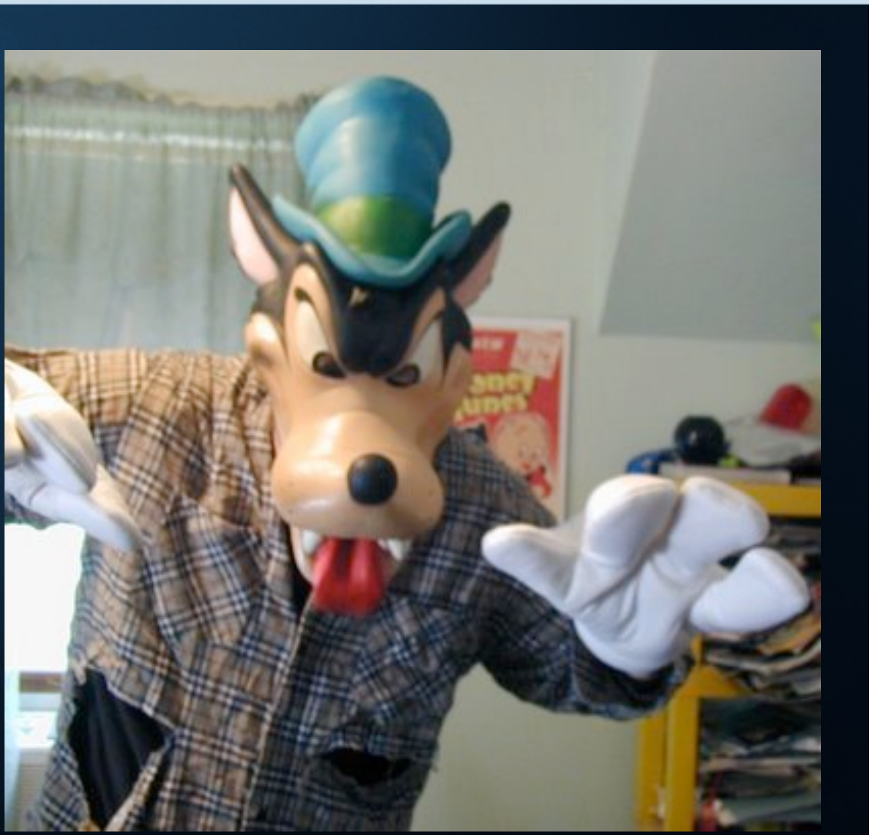
Paul Vixie

Andrew Fried

Dr. Chris Lee

Grandma has a problem

- An email or web banner offered her a free demo of the game Bejeweled 3D
- She clicked “yes” to download a program.
- New unrecognized malware?
- Anti-virus out of date or otherwise not effective?



Her PC is 0wned

- An error message is displayed. Oh well.
- Unknowing, she goes back to playing Bejeweled 2.
- PC is now under control of someone else.
- All she notices that its sluggish or slower than normal, but still usable.

What data can be collected

- Toolbar in her browser logged a query to the download site
  - Toolbar maintainers notice thousands of others have made similar visits today where none made before and log it.
- AV software logged the download and unsuccessful match against known malware
  - AV maintainers see several similar downloads across user base base on signature.
- Browser performed a DNS query to lookup website
  - ISP recursive server logs and shares Passive DNS information
  - Other ISPs see the same

What data can be collected

- Her PC started talking with C&C server on a high TCP port
  - ISP captured and shared netflow data for her sessions
  - DHCP logs track her PC's IP to her access device
- The next day, her PC starts sending out SPAM
  - IP address is different, but ISP tracks IP via DHCP logs to same access device
  - Recursive nameserver at ISP sees unusually high number of MX lookups from her IP.
  - Noted traffic flow on port 25 outbound has increased.
  - DNSBL sites start seeing manymore lookup requests based on her IP

What data can be collected

- More spam is sent
  - A spamtrap picks up a few of the messages sent by her PC
  - People using webmail started marking the messages as spam
  - URLs from the spam messages were submitted to SURBL
  - Similar emails are logged at mail service providers coming from lots of other IPs.
  - People started submitting messages to spamcop

What data can be collected

- Her PC starts probing nearby and remote networks for an attack vector
  - ISP netflow logs attempt to talk to bogus IPs
  - Darknet sensors pick up connection attempts
  - A military firewall gateway picks up connection attempts
  - A corporate firewall vendor sees logs from several customers' installations of probes from common sources.
  - Her PC successfully attacks an unpatched honeypot at a University research center.

What data can be collected

- Meanwhile, a day earlier, domains were registered at a registrar for a Pacific island.
  - All were registered at the same time
  - All have bogus registration information for an address between two casinos in Las Vegas
  - The domains were all purchased using the same credit card that had not yet been reported stolen – no chargebacks yet.
  - Malware links in spams use URLs in these domains.
  - Registrar logged CAPTCHA access during registration came from VPN service hosted in ex-Soviet republic.

What data can be collected

- The VPN service is hosted at an ISP in the same BGP AS number of some of the C&C servers.
  - Passive DNS collected from ISPs see other suspect domains (randomly created or containing known phishing keywords) on nearby IP addresses.
  - Web crawlers identify a similar header signature used on webservers hosted on several of the neighboring IPs.
  - Web crawlers found malware and phishing kits on some of the neighboring servers.

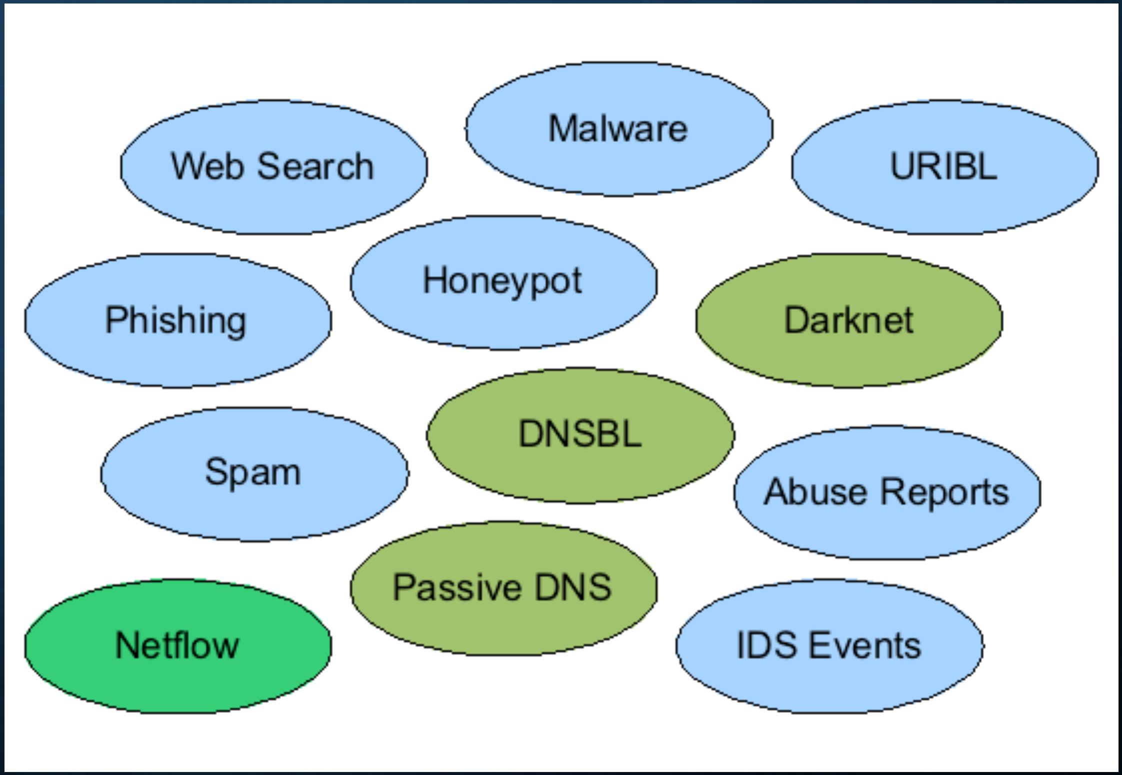
Do we collect it? Do we share it?

- Ideally: Security data is collected and either shared or made readily accessible in a trusted community in real time.
- Today: Security data is mostly discarded or at least not shared in a common framework.

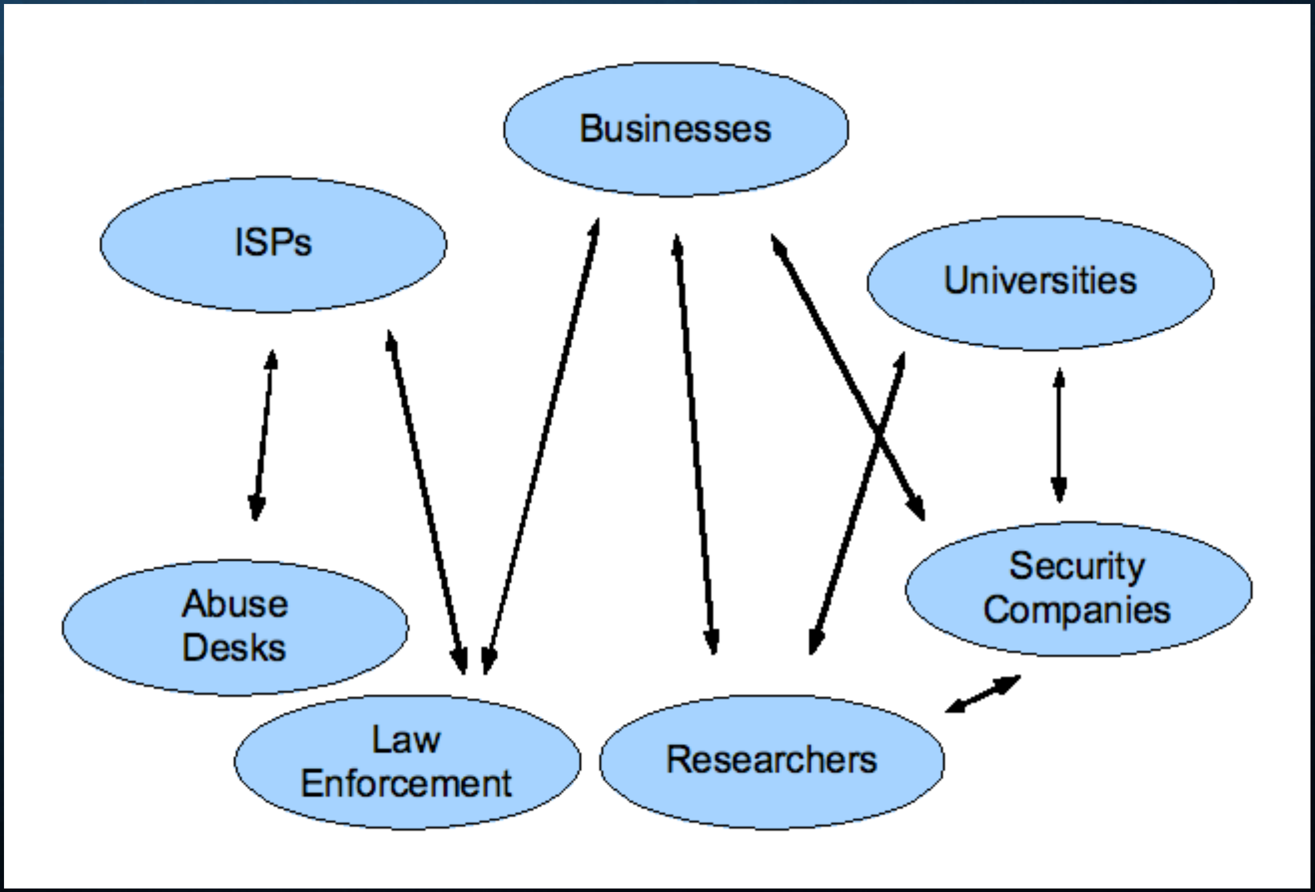
Challenges

- Miscreants operate behind the scenes on stolen or leased resources. They only need to organize within infrastructure for a short period of time to be effective.
- Unlike ISPs or user populations, they have nothing real to defend.
- Time window between allocation of resources and attack is shrinking.
- Asking peers on a security mailing list for information can take too long to be effective.

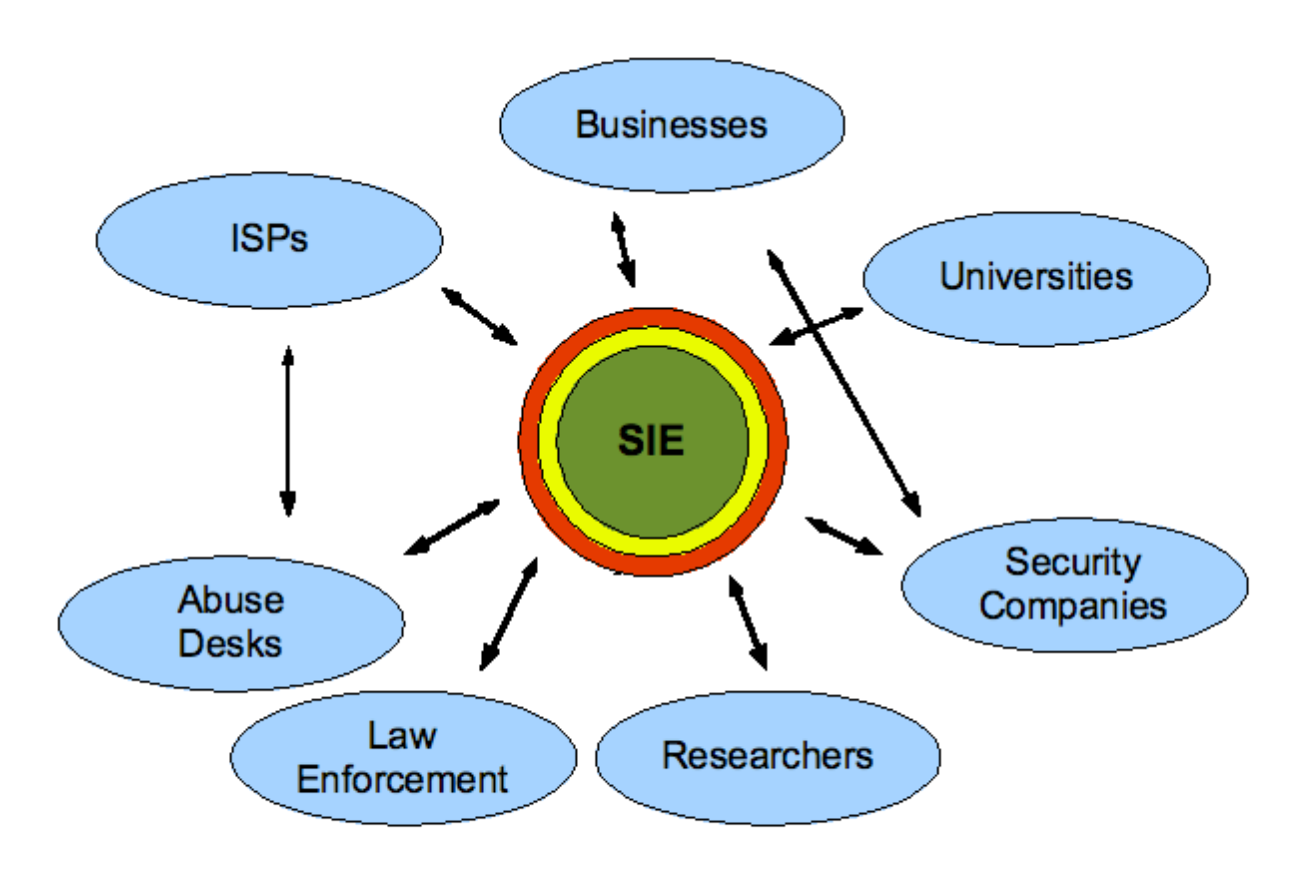
Disparate data types



Bi-lateral information flows



ISC SIE – enabling data mining



Efficient sharing within common legal/privacy framework

Data mining is the process of extracting hidden patterns from data

-- Wikipedia

Internet Special Ops

Stalking Badness Through Data Mining

Finding a “target” on the Internet requires the collection and analysis of unprecedented amounts of data from a variety of sources throughout the world

Data Mining

- Identification
- Collection
- Normalization
- Reduction
- Add Derivative Data
- Analysis
- Putting the pieces together

Example Data Sources

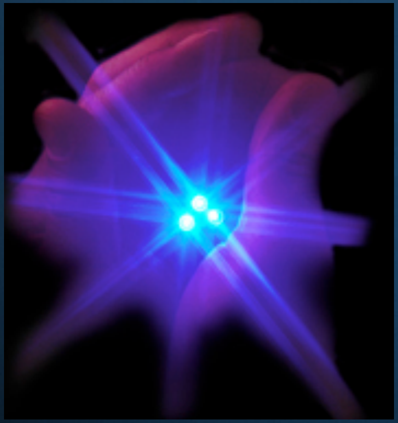
- Passive DNS – 12,000 per second
- Spamtrap Data – 3,500 per second
- Domain Registrations – 450,000 per day
- Tracking Nameservers – 2,600,000 per day
- BGP/ASN Data – 288,000 ASNs
- Malware Samples (unfortunately, a LOT!)
- Conficker Infected Hosts – over 5 million

The goal of data mining

The tools of the “trade”

- Bandwidth
- Storage
- Fast servers + RAM
- Databases
- Intuition & Ingenuity

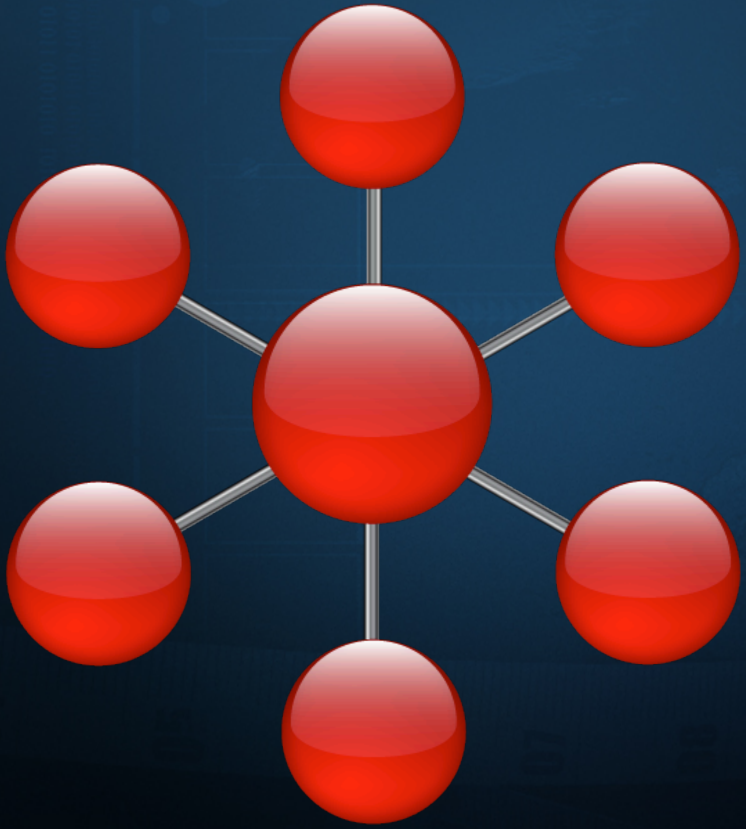
Data Normalization



- •Standard format
- •Common fields
- •“Relational Characteristics”
- •Compatible with database

Data Reduction

- Pruning Data
- Packing data (Integer vs IP)
- Summarization Tables



Derivative Data

Developing new datasets through relational characteristics of your original and possibly disparate processed data

Produces “3D” views of your data

Very effective method for trend analysis with relational databases

DNS is the central nervous system of the Internet.

Virtually all analysis of events on the Internet begin with DNS records, or more specifically, IP addresses. By themselves, an IP address identifies a single host. But what else can we learn from a lowly IP address?

Enumerating IP addresses

First, we can attempt to find the reverse arpa (PTR) records for a given IP address. That often tells us the domain name of the host.

Enumerating IP addresses

Next, we can identify who “owns” that IP address (registered netblock owner).

Enumerating IP addresses

In order to reach an address on the Internet, routers need to know how to route traffic to the subnet containing that address. BGP routing tables can provide us with that answer, providing both the ASN number and other netblocks served from the same ASN.

Enumerating IP addresses

GeoIP databases can assist us in determining the geographic location of the host. Data can include country, city and state and even latitude and longitude coordinates that can be used in distance calculations.

Enumerating IP addresses

IP addresses can also be associated to fully qualified domain names and authoritative nameservers through passive DNS (assuming PTR records are inaccurate or unavailable).

Enumerating IP addresses

Using a combination of both active and passive DNS, we can determine if an IP addresses appears in more than one published DNS resource record.

Enumerating IP addresses

Using SPAM trap data, we can determine if the IP address and enumerated domain name is appearing in SPAM and if the netblock appears in RBLs.

Tying the IP Pieces Together

DNS PTR records
Netblock owner via RIR records
ASN via BGP data
Location via GeoIP
FQDN via active and passive DNS
Authoritative nameserver(s) through enumeration
Appearance of domain in SPAM & RBLs

Internet Special Ops

Stalking Badness Through Data Mining

What kind of questions can we NOW ask of the data?

How many spam messages originate from a particular ASN?
What percentage of domains on a given nameserver are RBL’ed?
How many domains resolve back to a single IP address?
How many infected machines are located in { $country } ?
How many nameservers are hosted on a given IP address?
What domains is a given nameserver authoritative for?

How can we Use Passive DNS to Identify Fast Flux Botnets?

How can we Use Passive DNS to Identify Fast Flux Botnets?

Multiple IP addresses / low TTLs
Generally hosted on compromised boxes
Geographically dispersed
Newly registered domain names

From our LIVE feed of 12,000 records per second:

Pull out host names with 3 or more “A” records
Determine ASN for each IP
Determine ratio of ASN to IP
Add “points” for TTL of 300 or less
Score of .6 or higher good indicator

Internet Special Ops

Stalking Badness Through Data Mining

From a feed of newly registered domain names:

- Perform bulk IP lookups
- Flag domains appearing in SPAM traps
- Flag domains with 3 or more IP addresses
- Flag domains containing “paypal”, “bank”, etc.
- Flag domains with “bad” nameservers
- Flag domains resolving to known BOT IPs
- Flag domains from known “bad” ASNs

Even Fancier data mining techniques:

Identify nameservers with a high ratio of newly registered domains

Identify IP addresses with multiple nameservers that have a “significant” percentage of RBL hits

Identify nameservers that are authoritative for numerous domains that exhibit similar domain name characteristics (ratio of consonants, length, etc)

Internet Special Ops

Stalking Badness Through Data Mining

Using BGP / ASN / IP and Domain Data

Identify hosts resolving to newly advertised ASNs
Identify hosts resolving to BOGON addresses
Identify netblocks that “move” over a period of time

Sample scan results:

```
aaa-pharmacystore.com|6|6|1.00|N
best-buy-pharmacyonline.com|6|6|1.00|N
bmw50.com|10|10|1.00|N
ciglm.com|13|9|0.69|N
mdclr.com|17|13|0.76|N
mdclr.com|17|14|0.82|N
mltjd.com|12|9|0.75|N
mltjd.com|12|9|0.75|N
mzkta.com|14|14|1.00|N
nrzce.com|16|12|0.75|N
rsurt.com|17|11|0.65|Y
rsurt.com|17|11|0.65|Y
```

Internet Special Ops

Stalking Badness Through Data Mining

Sample scan results:

```
aaa-pharmacystore.com|6|6|1.00|N
best-buy-pharmacyonline.com|6|6|1.00|N
bmw50.com|10|10|1.00|N
ciglm.com|13|9|0.69|N
mdclr.com|17|13|0.76|N
mdclr.com|17|14|0.82|N
mltjd.com|12|9|0.75|N
mltjd.com|12|9|0.75|N
mzkta.com|14|14|1.00|N
nrzce.com|16|12|0.75|N
rsurt.com|17|11|0.65|Y
rsurt.com|17|11|0.65|Y   <-   LET'S LOOK AT THIS ONE
```

Internet Special Ops

Stalking Badness Through Data Mining

rsurt.com|17|11|0.65|Y <- LET'S LOOK AT THIS ONE

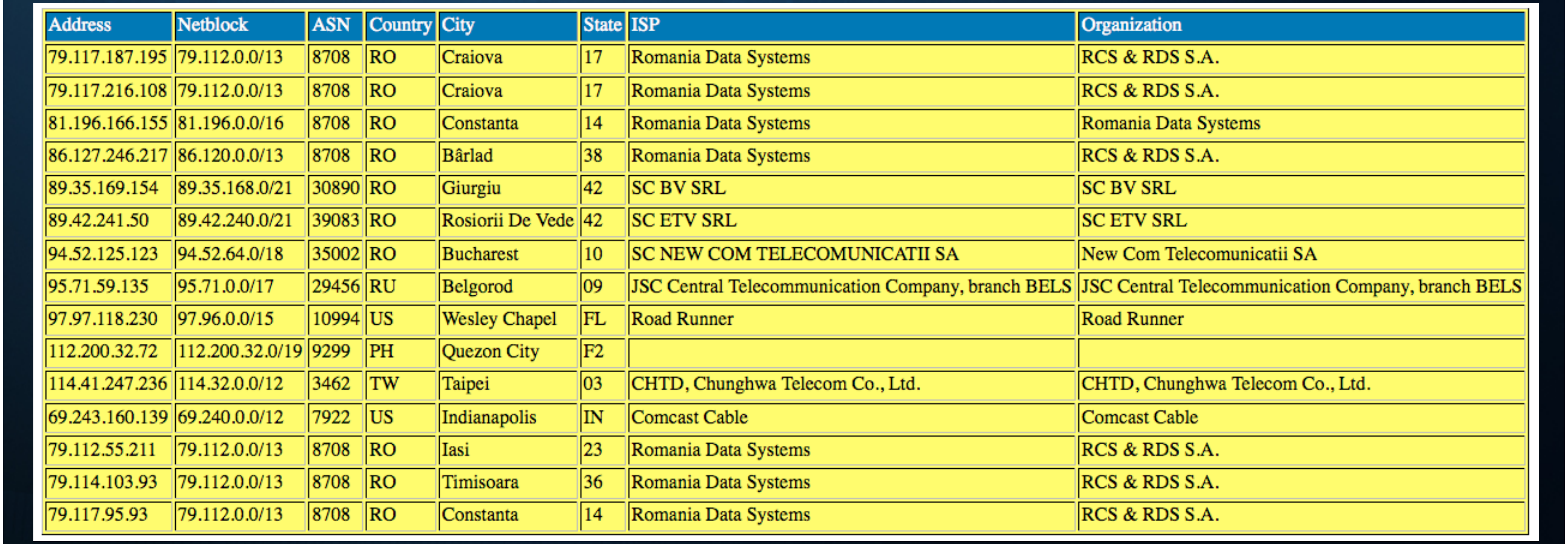
IP addresses:

79.117.187.195
81.196.166.155
89.35.169.154
94.52.125.123
97.97.118.230
114.41.247.236
79.112.55.211
79.115.69.195
79.117.95.93

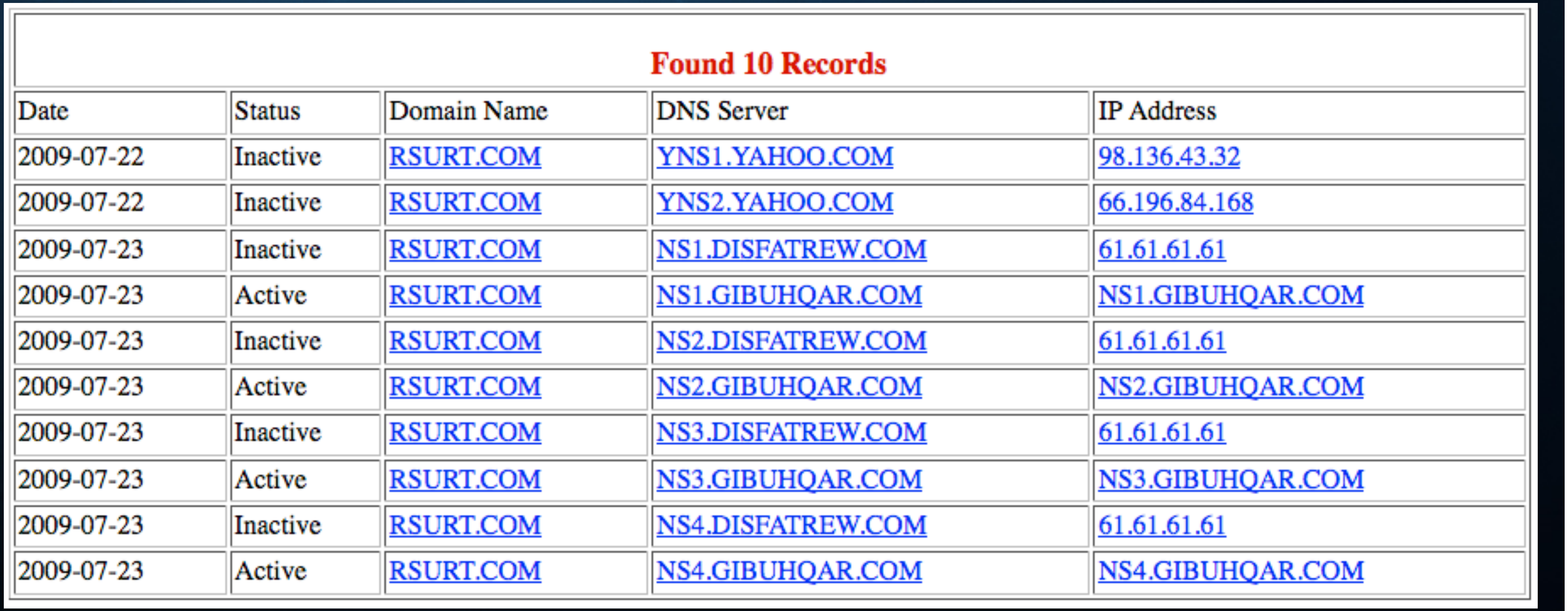
79.117.216.108
86.127.246.217
89.42.241.50
95.71.59.135
112.200.32.72
69.243.160.139
79.114.103.93
79.115.113.35

Internet Special Ops

Stalking Badness Through Data Mining



Address	Netblock	ASN	Country	City	State	ISP	Organization
79.117.187.195	79.112.0.0/13	8708	RO	Craiova	17	Romania Data Systems	RCS & RDS S.A.
79.117.216.108	79.112.0.0/13	8708	RO	Craiova	17	Romania Data Systems	RCS & RDS S.A.
81.196.166.155	81.196.0.0/16	8708	RO	Constanta	14	Romania Data Systems	Romania Data Systems
86.127.246.217	86.120.0.0/13	8708	RO	Bârlad	38	Romania Data Systems	RCS & RDS S.A.
89.35.169.154	89.35.168.0/21	30890	RO	Giurgiu	42	SC BV SRL	SC BV SRL
89.42.241.50	89.42.240.0/21	39083	RO	Rosiorii De Vede	42	SC ETV SRL	SC ETV SRL
94.52.125.123	94.52.64.0/18	35002	RO	Bucharest	10	SC NEW COM TELECOMUNICATII SA	New Com Telecomunicatii SA
95.71.59.135	95.71.0.0/17	29456	RU	Belgorod	09	JSC Central Telecommunication Company, branch BELS	JSC Central Telecommunication Company, branch BELS
97.97.118.230	97.96.0.0/15	10994	US	Wesley Chapel	FL	Road Runner	Road Runner
112.200.32.72	112.200.32.0/19	9299	PH	Quezon City	F2		
114.41.247.236	114.32.0.0/12	3462	TW	Taipei	03	CHTD, Chunghwa Telecom Co., Ltd.	CHTD, Chunghwa Telecom Co., Ltd.
69.243.160.139	69.240.0.0/12	7922	US	Indianapolis	IN	Comcast Cable	Comcast Cable
79.112.55.211	79.112.0.0/13	8708	RO	Iasi	23	Romania Data Systems	RCS & RDS S.A.
79.114.103.93	79.112.0.0/13	8708	RO	Timisoara	36	Romania Data Systems	RCS & RDS S.A.
79.117.95.93	79.112.0.0/13	8708	RO	Constanta	14	Romania Data Systems	RCS & RDS S.A.

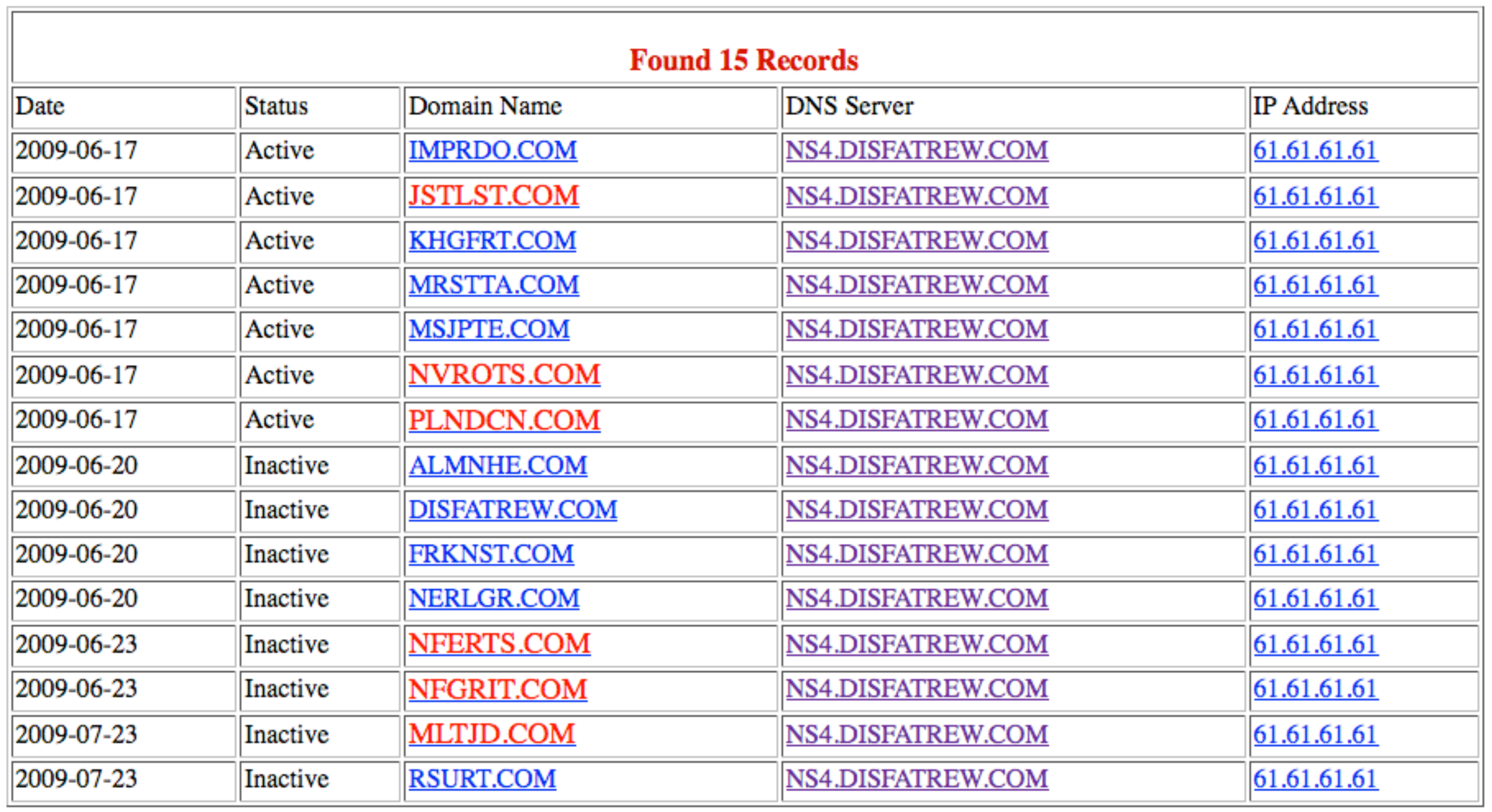


Found 10 Records

Date	Status	Domain Name	DNS Server	IP Address
2009-07-22	Inactive	RSURT.COM	YNS1.YAHOO.COM	98.136.43.32
2009-07-22	Inactive	RSURT.COM	YNS2.YAHOO.COM	66.196.84.168
2009-07-23	Inactive	RSURT.COM	NS1.DISFATREW.COM	61.61.61.61
2009-07-23	Active	RSURT.COM	NS1.GIBUHQAR.COM	NS1.GIBUHQAR.COM
2009-07-23	Inactive	RSURT.COM	NS2.DISFATREW.COM	61.61.61.61
2009-07-23	Active	RSURT.COM	NS2.GIBUHQAR.COM	NS2.GIBUHQAR.COM
2009-07-23	Inactive	RSURT.COM	NS3.DISFATREW.COM	61.61.61.61
2009-07-23	Active	RSURT.COM	NS3.GIBUHQAR.COM	NS3.GIBUHQAR.COM
2009-07-23	Inactive	RSURT.COM	NS4.DISFATREW.COM	61.61.61.61
2009-07-23	Active	RSURT.COM	NS4.GIBUHQAR.COM	NS4.GIBUHQAR.COM

Internet Special Ops

Stalking Badness Through Data Mining



Found 15 Records

Date	Status	Domain Name	DNS Server	IP Address
2009-06-17	Active	IMPRDO.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-17	Active	JSTLST.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-17	Active	KHGFRT.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-17	Active	MRSTTA.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-17	Active	MSJPTE.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-17	Active	NVROTS.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-17	Active	PLNDCN.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-20	Inactive	ALMNHE.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-20	Inactive	DISFATREW.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-20	Inactive	FRKNST.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-20	Inactive	NERLGR.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-23	Inactive	NFERTS.COM	NS4.DISFATREW.COM	61.61.61.61
2009-06-23	Inactive	NFGRIT.COM	NS4.DISFATREW.COM	61.61.61.61
2009-07-23	Inactive	MLTJD.COM	NS4.DISFATREW.COM	61.61.61.61
2009-07-23	Inactive	RSURT.COM	NS4.DISFATREW.COM	61.61.61.61

rsurt.com|17|11|0.65|Y <- LET'S LOOK AT THIS ONE

Any other domains using the same IP addresses in todays list?

1. ciglm.com
2. nrzce.com
3. rsurt.com
4. mltjd.com
5. mdclr.com
6. mzkta.com
7. dsrth.com
8. mltjd.com
9. mdclr.com
10. rsurt.com

Internet Special Ops

Stalking Badness Through Data Mining

- Badness leaves a trail
- Data mining techniques find that trail
- Effective mitigation requires timely and effective detection

Conficker: Phase 1

First Blood

OMG! My Network's on Fire

- Early January Conficker.B started shutting down networks with password attempts
- The security community takes notice and starts sinkholing domains
- A lot of time was spent obtaining domains, researching the other domains, and keeping up with traffic

Phase 2

Band of Brothers

Let’s unionize

- Researchers talk to each other and ask to share data.
- Cost of domains accumulating
- We ask Support Intelligence to “WhiteTaste” for us
- Cabal is dubbed

Conficker: Phase 3

The Great Escape

ICANN, so we can

- ICANN leads the coordination of registries
- Cost now bourne by TLDs, not researchers
- Data is centralized, PR is coordinated
- Massive reporting to affected networks
- Conficker Working Group is born

Conficker: Phase 4

Apocalypse Now

The Final Countdown

- Conficker.C is released, uses 116 TLDs
- Lots of evasion techniques
- News: the Internet will self-destruct.. Goodbye.
- ICANN/CWG coordinate with the affected TLDs
- The world is saved, right?

Conficker: Phase 5

A Very Long Engagement

Tug of War

- Drama is over, but now we need to fight back
- Look at the data and find out where to place our efforts
- Organize our troops and attack
- Let’s look at the numbers…

Conficker: Data Collection

The Island of Dr. Moreau

It’s log, log, it’s big, it’s heavy, it’s wood.

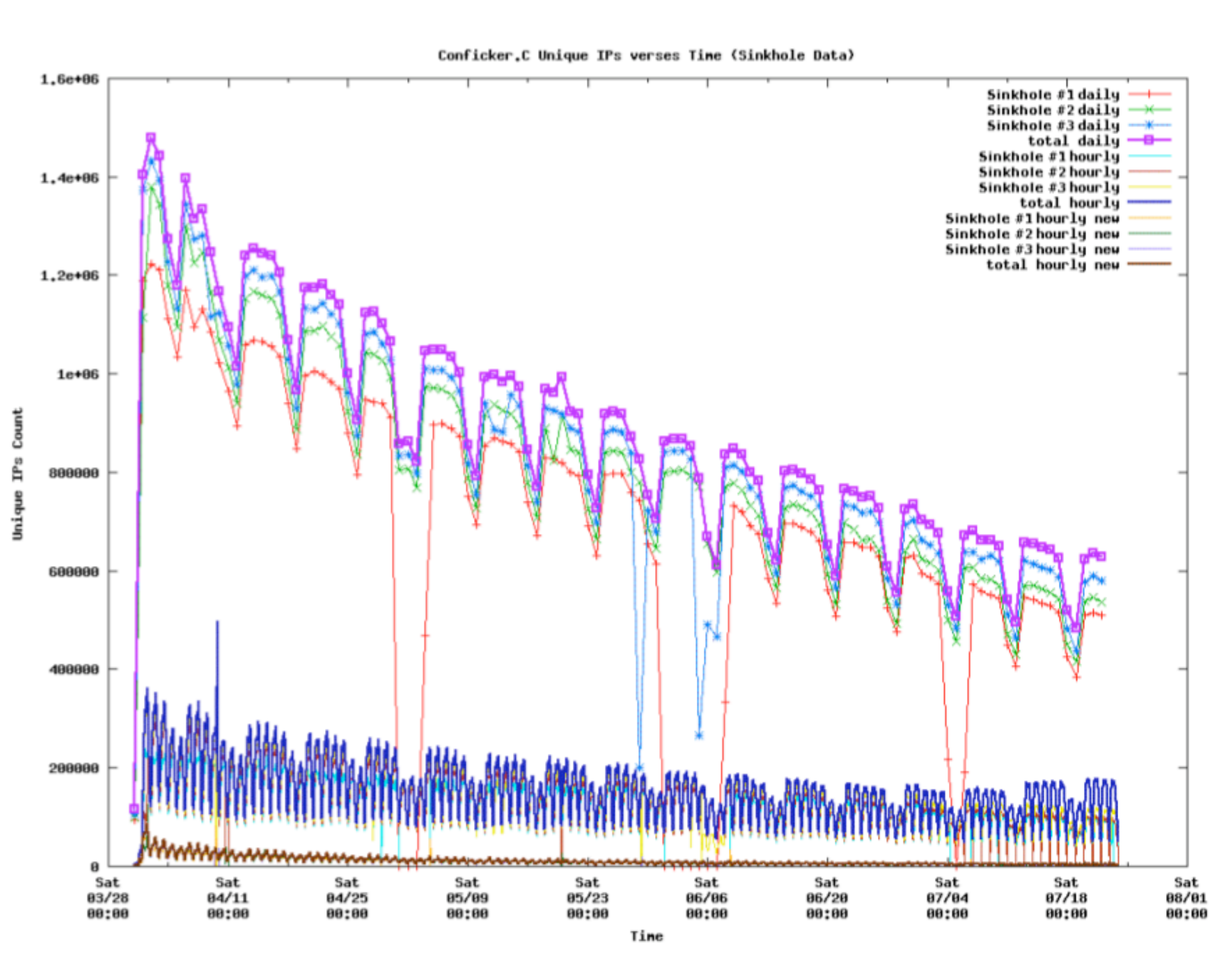
- Every contributor of sinkhole logs had a different format and collected in different ways.
- Standard “operations” format was agreed upon to help with parsing and reporting.
- Analysis techniques were ad hoc.

Conficker:

The Numbers

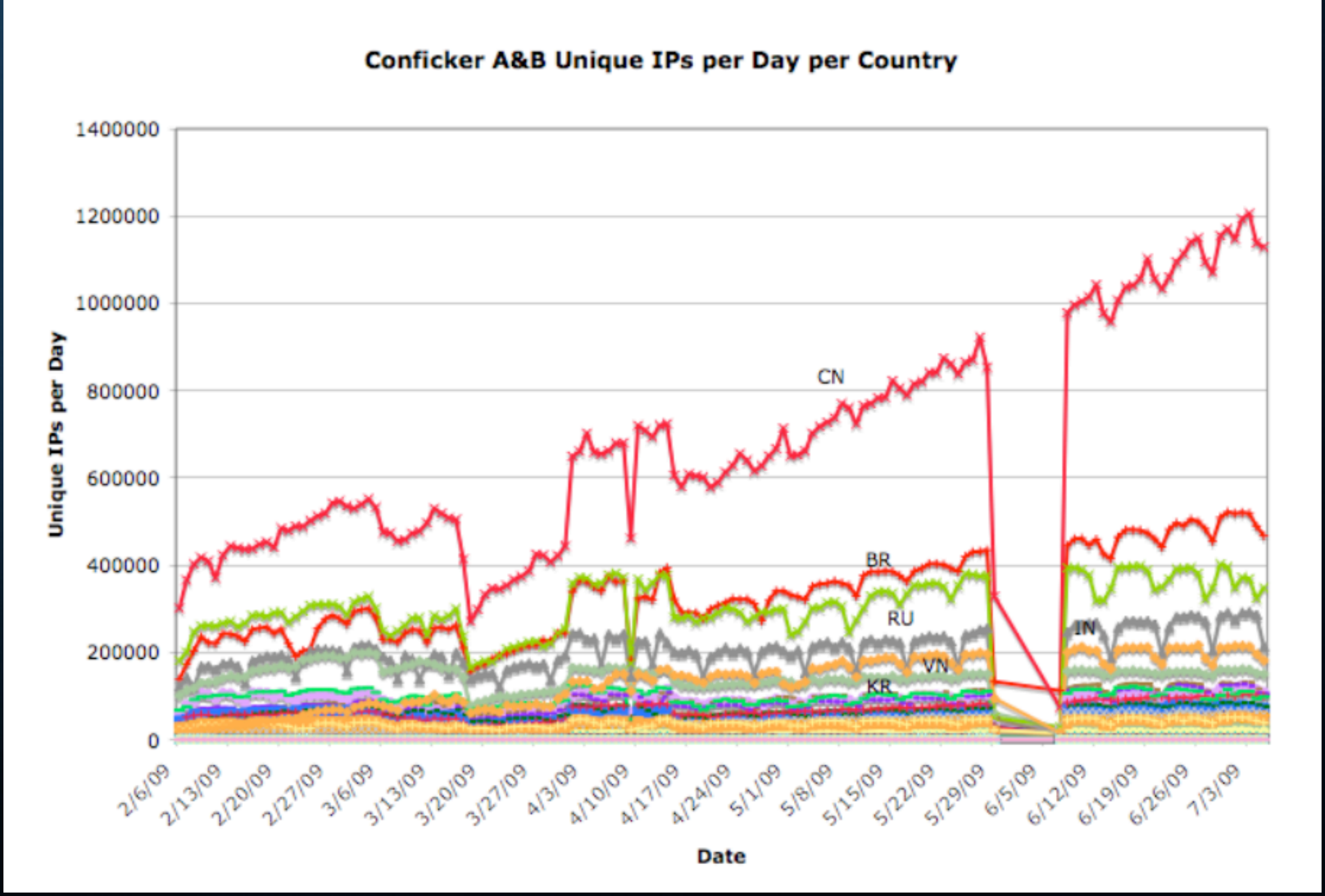
Internet Special Ops

Stalking Badness Through Data Mining



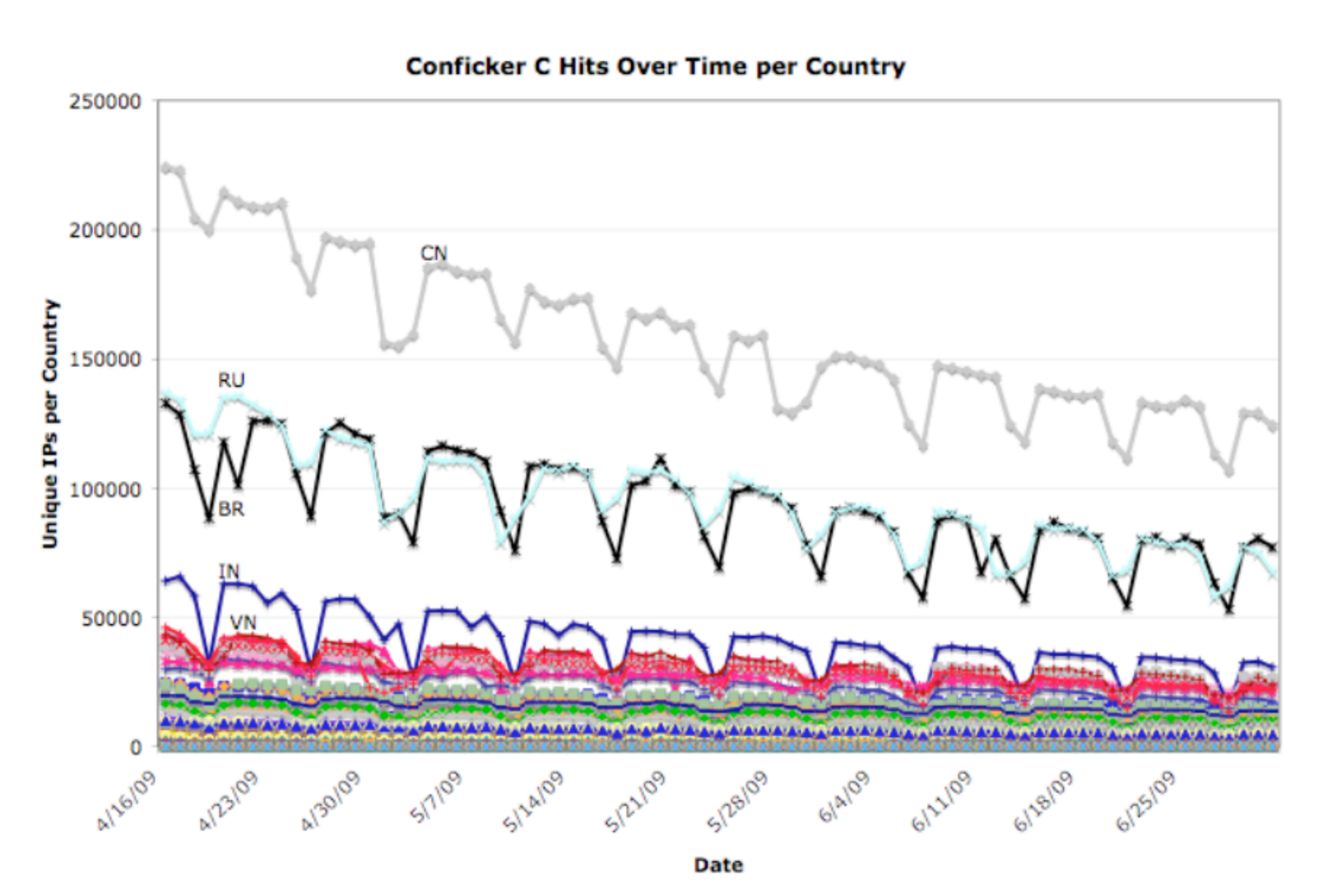
Internet Special Ops

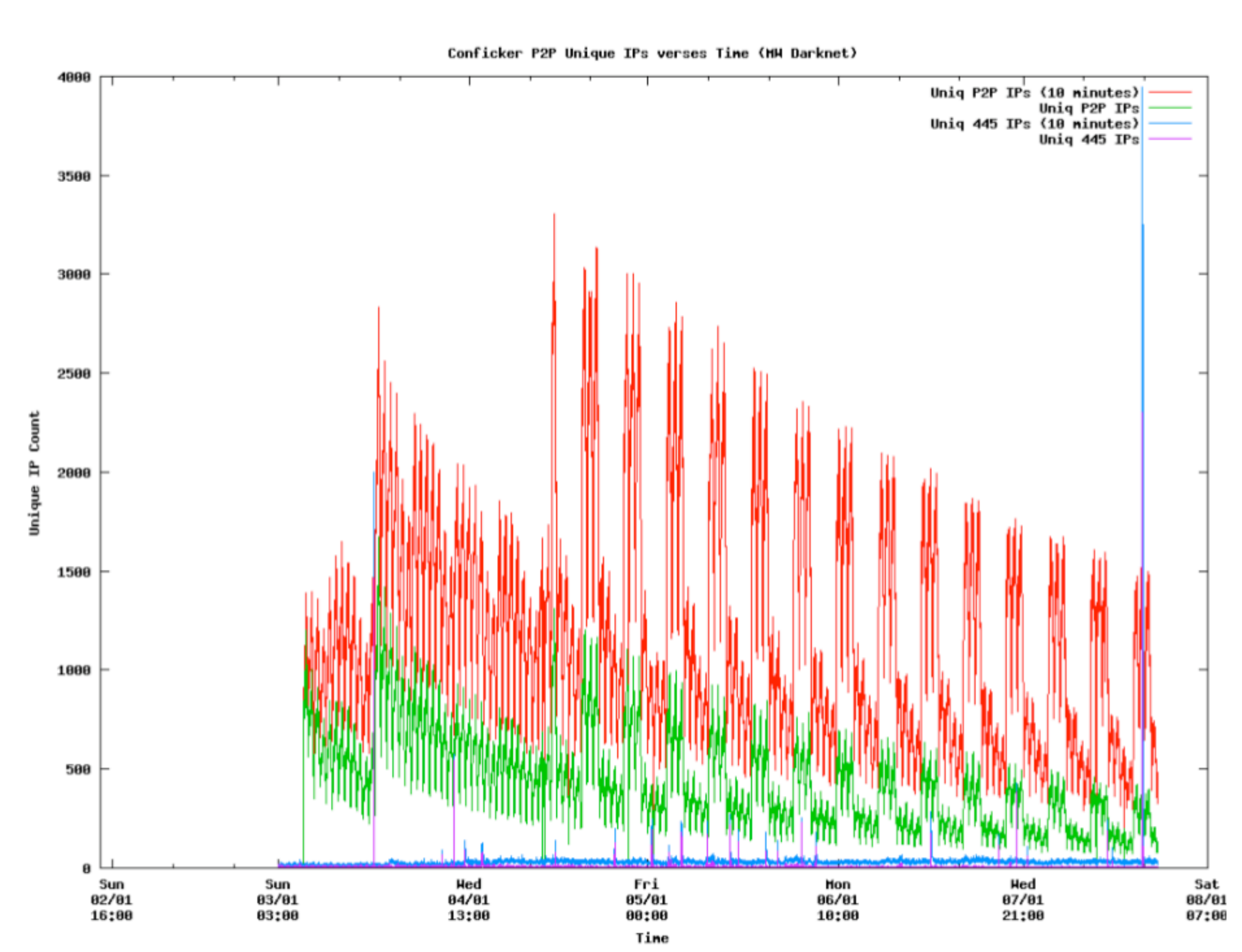
Stalking Badness Through Data Mining



Internet Special Ops

Stalking Badness Through Data Mining




Conficker P2P Unique IPs verses Time (MW Darknet)
Uniq P2P IPs (10 minutes) 🟥
Uniq P2P IPs 🟩
Uniq 445 IPs (10 minutes) 🟦
Uniq 445 IPs 🟪
Unique IP Count
4000
3500
3000
2500
2000
1500
1000
500
0
Sun 02/01 16:00
Sun 03/01 03:00
Wed 04/01 13:00
Fri 05/01 00:00
Mon 06/01 10:00
Wed 07/01 21:00
Sat 08/01 07:00
Time

It’s Log (Lyrics)

What rolls down stairs
alone or in pairs,
and over your neighbor's dog?
What's great for a snack,
And fits on your back?
It's log, log, log

It’s Log (Lyrics)

It's log, it's log,
It's big, it's heavy, it's wood.
It's log, it's log, it's better than bad,
it's good.

It’s Log (Lyrics)

Everyone wants a log
You're gonna love it, log
Come on and get your log
Everyone needs a log
log log log

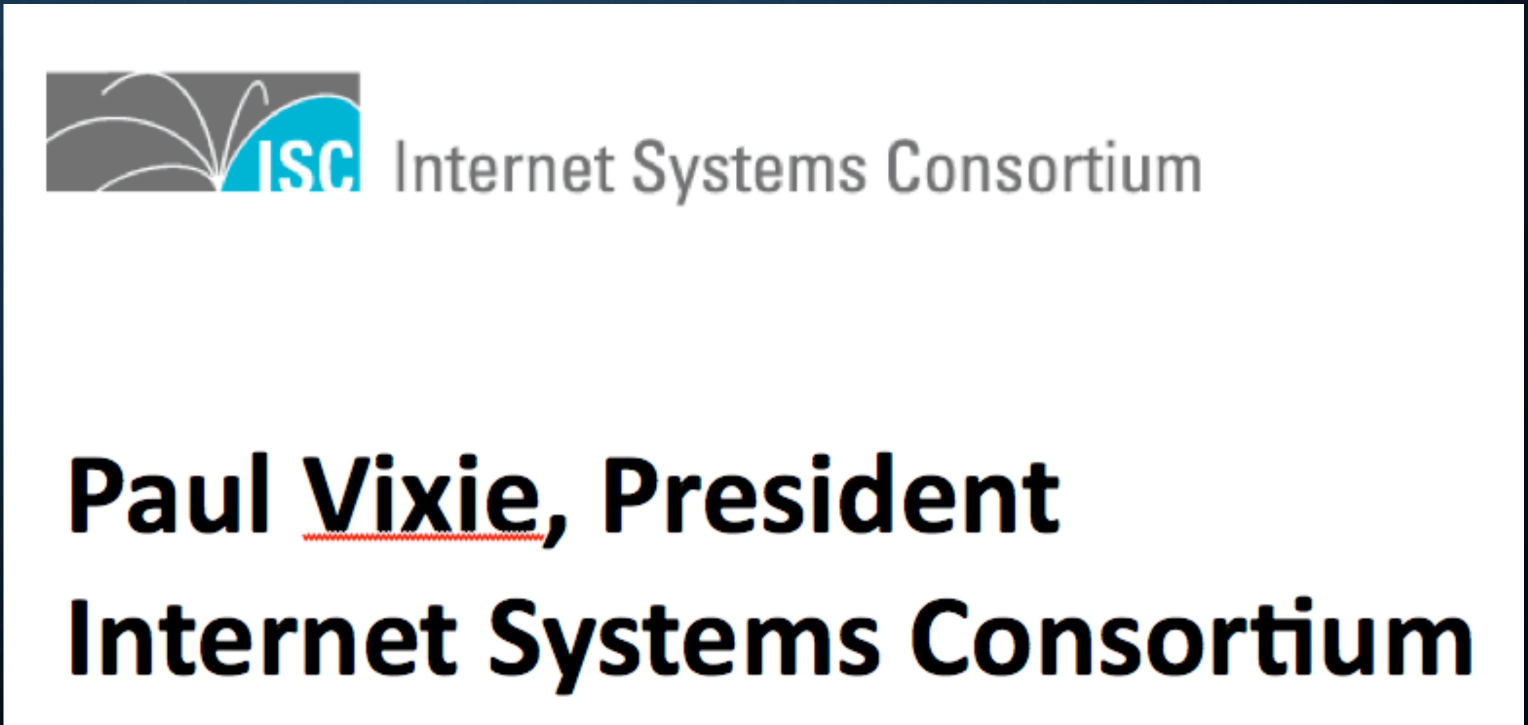
It’s Log (Lyrics)

whistle

LOG FROM BLAMMO

Internet Special Ops

Stalking Badness Through Data Mining



Internet Systems Consortium

Paul Vixie, President

Internet Systems Consortium

ISC Internet Systems Consortium

shadowserver

Dr. Chris Lee
Internet Systems Consortium
Shadowserver

Internet Special Ops

Stalking Badness Through Data Mining

ISC Internet Systems Consortium

Cutter Consortium

●●● ACCESS TO THE EXPERTS

SURBL

Deteque

Andrew Fried

Internet Systems Consortium, Cutter Consortium, Deteque, SURBL

Internet Special Ops

Stalking Badness Through Data Mining

Paul Vixie

Andrew Fried

Dr. Chris Lee