Carnegie Mellon University

Copyright 2015 Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM-0002555

Finding Related Malware Samples Using Run-Time Features

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Rhiannon Weaver

Establishing Trust in Software

Who are you?

What do you want to do?

(Are you lying about either?

Can I still trust you next week?)

How does this break down in the Cyber landscape?

- Unique identifiers of individuals
- Granularity of “identity”
- Rate of change

http://en.wikipedia.org/wiki/Jamie_Madrox

http://marvel.com/universe/Rogue

http://marvel.wikia.com/Raven_Darkholme_(Earth-11326)

Polling Question

Are you familiar with cryptographic hashes (MD5 and SHA256)?

Malware Data and Analysis

Who are you?

- Binary file
- UID = MD5 hash
- Trusted signed certificate

HP Revokes Digital Certificate Used to Sign Malware

by Liora R. Herman on November 20, 2014 | Leave a comment

Filed under Industry News and tagged digital, HP, Malware, Verisign.

LinkedIn 2 | Twitter | Google +1 | Facebook 1 | Reddit | Email

As reported by Krebs on Security, HP has performed the cyber security equivalent of a "my bad" by quietly advising customers of a digital certificate that had been used to sign malware in May 2010. The certificate, which was initially signed in error, was revoked by Verisign at HP's request on October 21, 2014.

HP detected the error after the malware, which assumed the name of a legitimate piece of software and bundled into a package, tried to connect to its command and control server. Originally, the compromised package was used by HP's internal teams for software testing. However, it eventually spread beyond the network by what the company believes was a mechanism designed to copy the malware.

Signatures

Static: can be defeated by polymorphism and packing

Dynamic: author must alter/obfuscate how the code interacts with the target system

Granularity of "individuals" is at the behavioral level

Increasingly we need to turn to "What do you want to do?" to augment "Who are you?"

http://www.seculert.com/blog/2014/11/hp-revokes-digital-certificate-used-to-sign-malware.html

Malware Data and Analysis

Tactics for Analysis

“Point-Query” starts with a file

- in-depth study of behavior via reverse engineering
- find more examples like this one

“Back-End” starts with a catalog

- clustering and fuzzy categorization
- find relationships among completely unknown and uncategorized files

My Research: exploring run-time features for “Back-End” malware analysis to direct the next in-depth reverse engineering analysis.

The CERT Artifact Catalog and Run-Time Environment

The CERT Artifact Catalog and Run-Time Environment

The CERT Artifact Catalog is a repository of suspicious and malicious files.

- Contributions of code have been automated and archived since 2000.
- 175m unique artifacts (binary files and related meta-data) as of June 2015

Anexa is an automated surface analysis tool that uses host-based instrumentation and forensics to observe malware activity and artifacts in the system.

- Registry, file, network, and service activity logging
- Configurable environments
- Anti-analysis mitigation
- Profiling to reduce noise

Dropped Files

Malware feature extracted during run-time analysis in Anexa:

- Start VM and take a snapshot of the file table (hash all files)
- Run Malware for 5 minutes
- Take another snapshot of the file table
- Any hashes added or modified are “dropped” by the malware

Some Questions of Interest:

1. What kind of relationship do dropped files have with the known malware families?

2. Can we use patterns in dropped files to help discover relationships among previously unstudied malware samples?

Data Set

Raw data: 1,993,810,620 records summarizing all Anexa runs through July 31 2014

- Malware Sample (MD5 hash)
- Dropped File (SHA256 hash)
- File Name
- File Path
- Date Run

50,911,788 unique malware samples by MD5 hash.

357,121,519 unique dropped files by SHA256 hash.

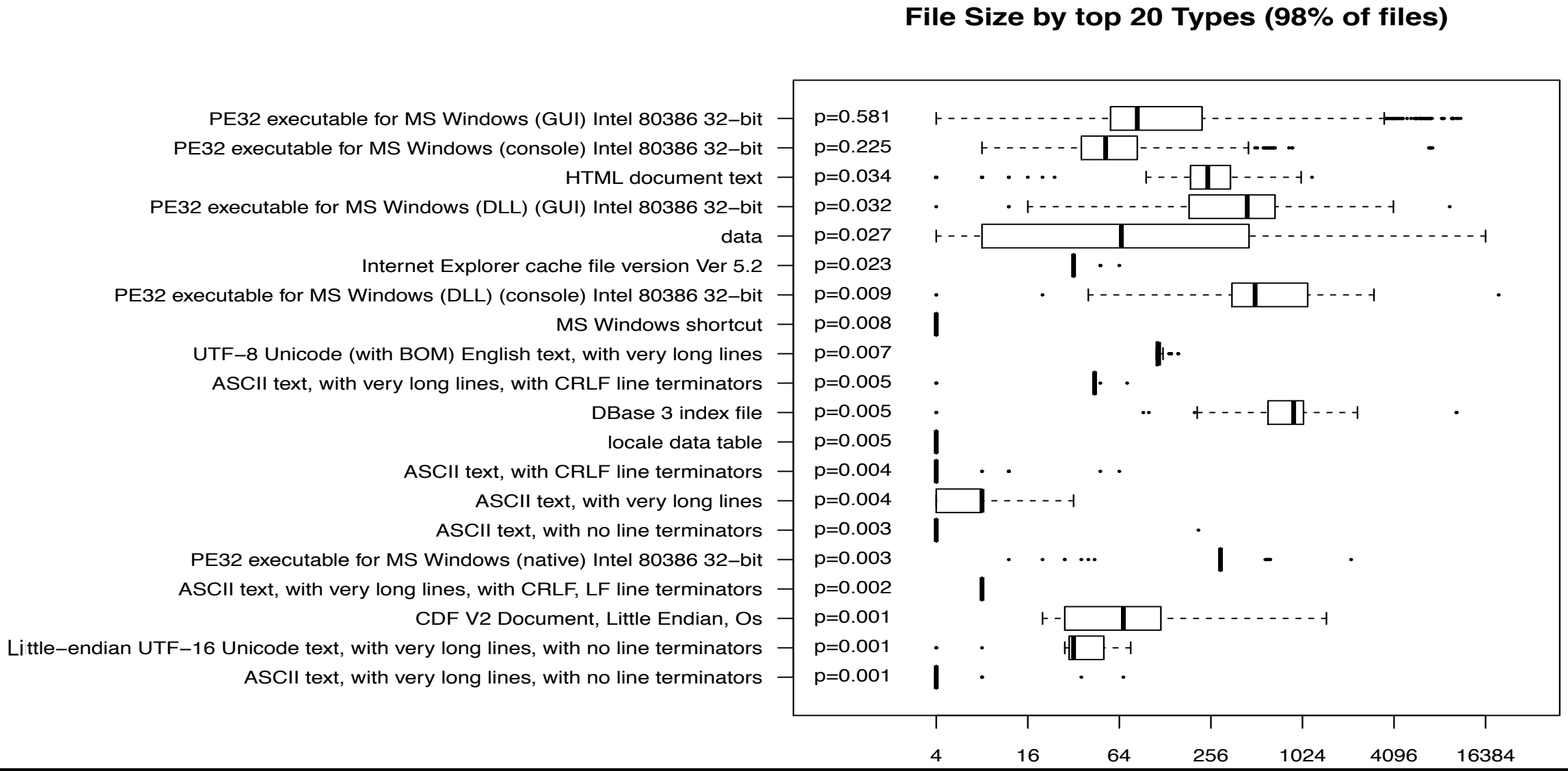
Data Set

Multiple Sample-to-File links exist because of paths and filenames

- Record separate list of filenames per dropped file
- Summarize unique Sample-Dropped File pairs (min, max Date run, #Paths)
- Drops us down to just 1.8billion records!

15,463,445 unique files dropped by multiple malware samples (4.33% of all files)

Dropped Files: Types and Sizes



Polling Question

Would you like to focus on techniques for:

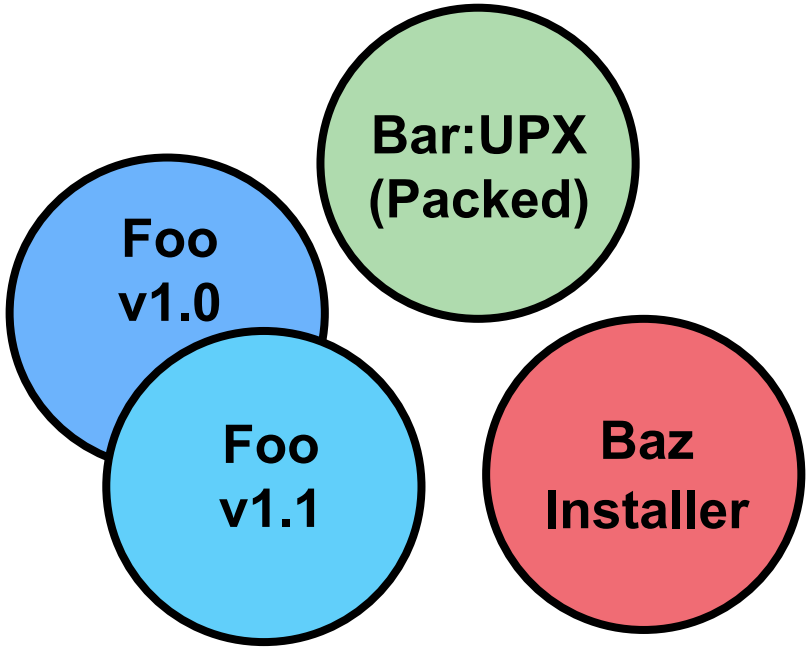
- Finding more examples of known malware families
- Finding related but completely uncategorized/unknown files

Leveraging Knowns: Finding “More Like X”

Dropped files and the Knowns List

Knowns List:

High fidelity, consistent, repeatable, and verifiable Yara signatures of malware families, batch run on new samples as ingested.



Approximately 10% of the samples in the catalog are categorized as known.

3,089,179 of the files run through Anexa (6%) are categorized as known.

- 62 million unique dropped files
- 2.2 million dropped by >1 sample (3.5%)

Which Files are Specific to a Family?

For each family and each dropped file, calculate:

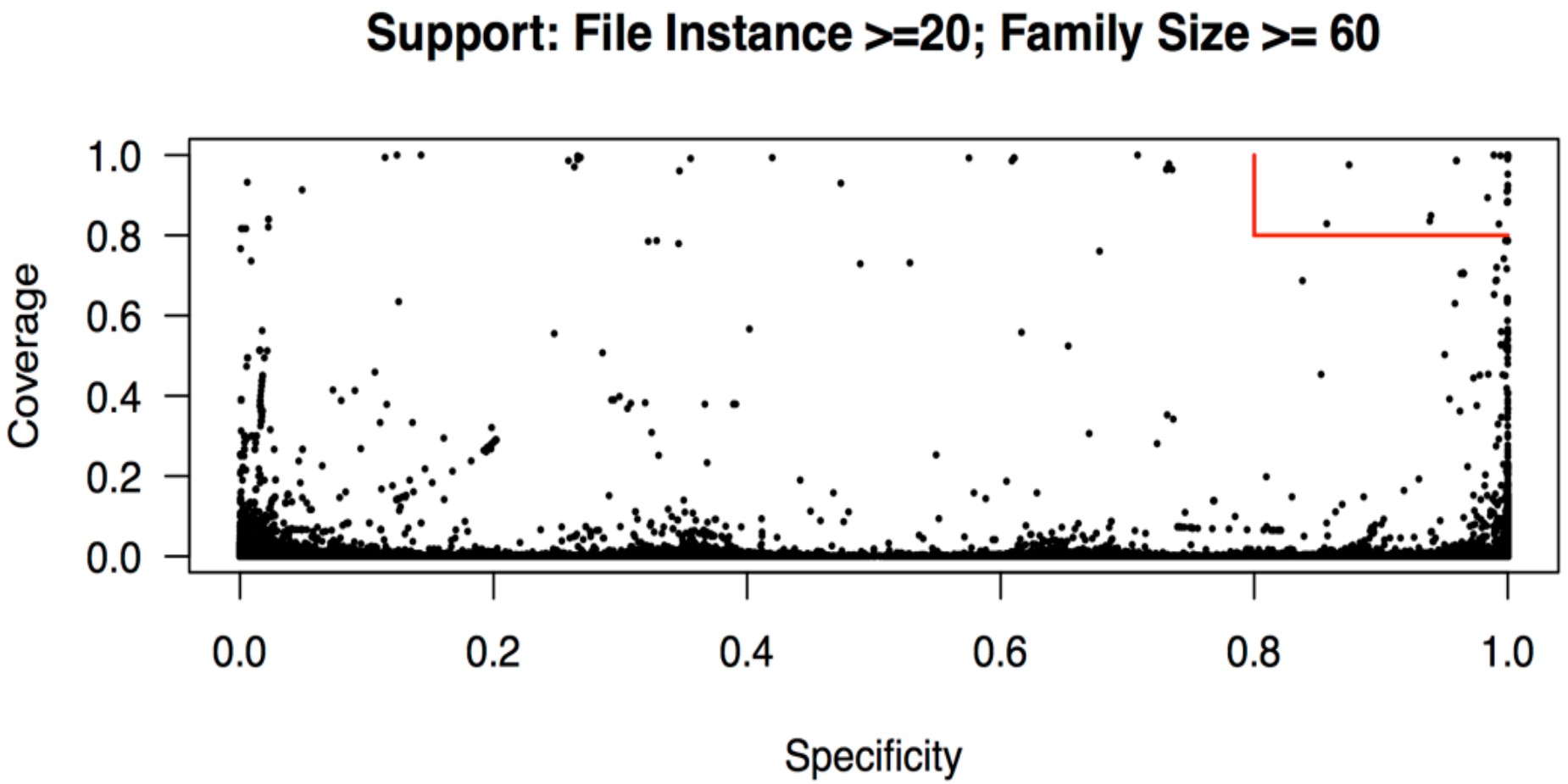
Specificity: the percent of all known drops of the file that appeared in the family

Coverage: the percent of all samples in the family that dropped the file.

Support:

- File Instance: total number of times the file was dropped by a known sample
- Family Size: total number of samples in the family.

Which Files are Specific to a Family?



What is Nearly 1 to 1?

1. cea7b3b4a7faa87fb6f191b9f0ba02a2
2. f0d450a1b8ff06f7393f6f1498f1d4b6
3. 9b357bfdb4281db1109c25006511c9df
4. 2f561b02a49376e3679acd5975e3790a

ID	AKA	Family	Spec.	Cov.	File Instance	Family Size	#Unknown
1	malware.ico	Cyclops:UPX	0.994	0.998	67587	67298	10652
2	t.asp t[1].asp	Scourge +ScourgeInstaller	1.000	0.994	56070	56377	608
3	start[1].htm	Firestar:UPX	0.857	0.829	35995	37223	2666
4	blan46aa.rra blank.gic.gif dot.gif loading1.gif (10 more)	Fenris:UPX	0.999	0.909	30950	34009	279078

Looking Closer

File	Family	First Seen	Last Seen	Samples
cea7b3b4a7faa87fb6f191b9f0ba02a2	Cyclops:UPX	2010/12/03	2014/03/25	67201
cea7b3b4a7faa87fb6f191b9f0ba02a2	Cyclops	2010/12/03	2014/06/25	386
cea7b3b4a7faa87fb6f191b9f0ba02a2	UNKNOWN	2010/12/02	2014/08/06	10652
f0d450a1b8ff06f7393f6f1498f1d4b6	Scourge+ScourgeIns.	2012/08/22	2013/03/01	56070
f0d450a1b8ff06f7393f6f1498f1d4b6	UNKNOWN	2012/09/13	2012/11/01	608
9b357bfdb4281db1109c25006511c9df	Firestar:UPX	2011/04/06	2014/03/20	30855
9b357bfdb4281db1109c25006511c9df	Firestar:ASPro.	2012/04/21	2013/04/11	5140
9b357bfdb4281db1109c25006511c9df	UNKNOWN	2012/08/25	2014/06/15	2666
2f561b02a49376e3679acd5975e3790a	Fenris:UPX	2011/07/31	2014/02/27	30929
2f561b02a49376e3679acd5975e3790a	MagnetoInfected	2011/07/26	2014/02/18	8
2f561b02a49376e3679acd5975e3790a	Redwing	2012/09/24	2014/04/18	6
2f561b02a49376e3679acd5975e3790a	Rhodey	2011/09/28	2013/05/01	4
2f561b02a49376e3679acd5975e3790a	Angel	2011/01/25	2013/05/31	3
2f561b02a49376e3679acd5975e3790a	UNKNOWN	2010/12/05	2014/07/22	279078

Looking Closer

```
6475d5ecc14fea09774be55723d2d52 aka "at11.job      at12.job at8.job  at9.job
autorun.inf      perflib_perfdata_e8.dat   regedt32.sys     ~df13e8.tmp      ~df1cee.tmp
~df1f05.tmp      ~df1f17.tmp      ~df207e.tmp      ~df22eb.tmp      ~df268f.tmp      ~df2bc4.tmp
~df2d30.tmp      ~df2e8b.tmp      ~df3055.tmp      ~df31f2.tmp      ~df3401.tmp      ~df3c0d.tmp
~df3fd4.tmp      ~df411c.tmp      ~df4635.tmp      ~df58e8.tmp      "
```

Family	First Seen	Last Seen	Samples
Becatron:UPX	2010/12/02	2014/06/18	100918
Becatron:VB	2011/01/20	2014/06/17	36200
Becatron+Blizzard+Mesmero:UPX	2011/01/28	2013/12/24	37
Becatron+Fenris:UPX	2012/06/02	2014/04/11	27
Becatron+Mesmero:UPX	2011/02/15	2014/01/09	16
Becatron+Blizzard:UPX	2011/02/04	2011/06/21	16
Becatron+MagnetoInfected:UPX	2013/07/04	2013/11/11	10
Becatron+Blizzard+Mesmero:VB	2011/07/19	2012/10/19	7
Avengers+Becatron:UPX	2012/05/31	2014/04/10	5
Becatron+Blizzard:VB	2011/09/03	2011/11/09	3
Becatron+Psylocke_v2:UPX	2012/09/04	2012/09/09	3
Becatron+Mesmero:VB	2012/08/25	2013/05/30	2
Becatron+Redwing:UPX	2012/09/14	2012/09/25	2
Becatron+Gambit+Gambit_Installer:UPX	2013/09/05	2014/04/03	2
Becatron+Fenris:VB	2012/09/26	2013/12/20	2
Becatron+Gambit:UPX	2014/02/13	2014/02/13	1
Becatron+Gambit+Gambit_Installer:VB	2014/04/11	2014/04/11	1
UNKNOWN	2011/08/03	2014/07/30	3992

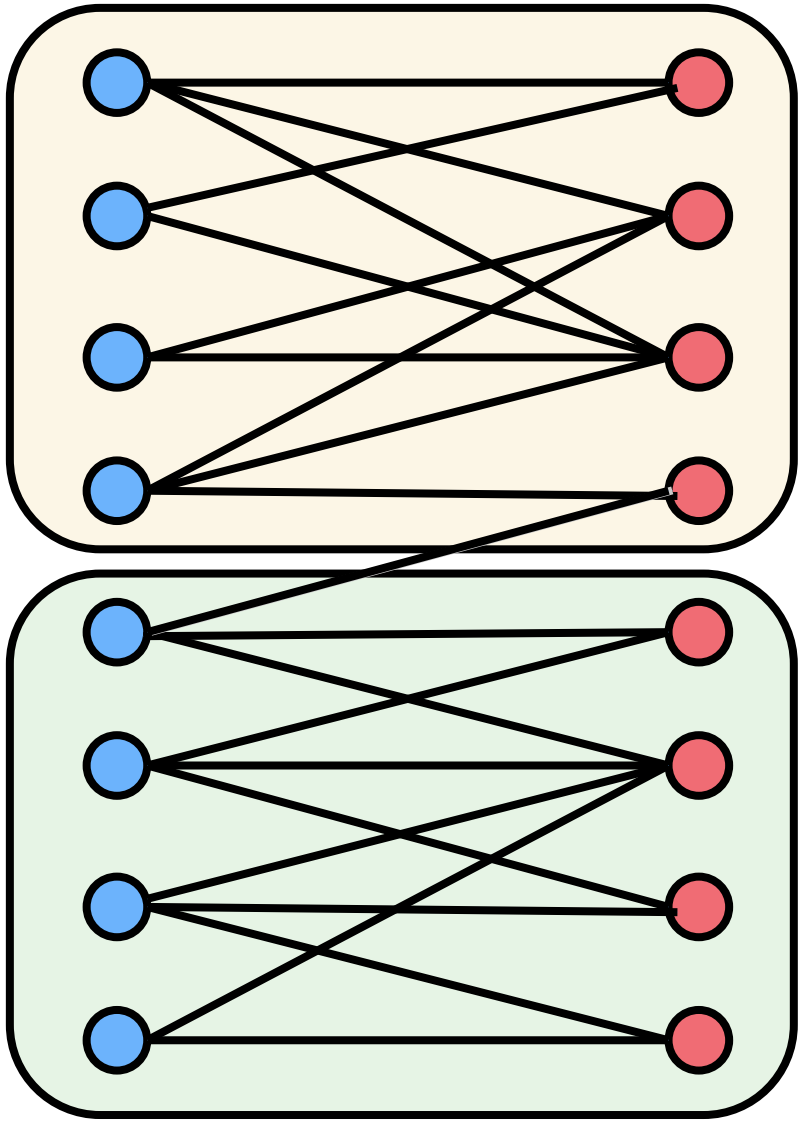
New Directions: Uncovering Potentially New Families

Community Detection

Links between Malware and Dropped Files form a graph.

Communities are clusters of highly related nodes.

Community detection is less strict than calculating Connected Components

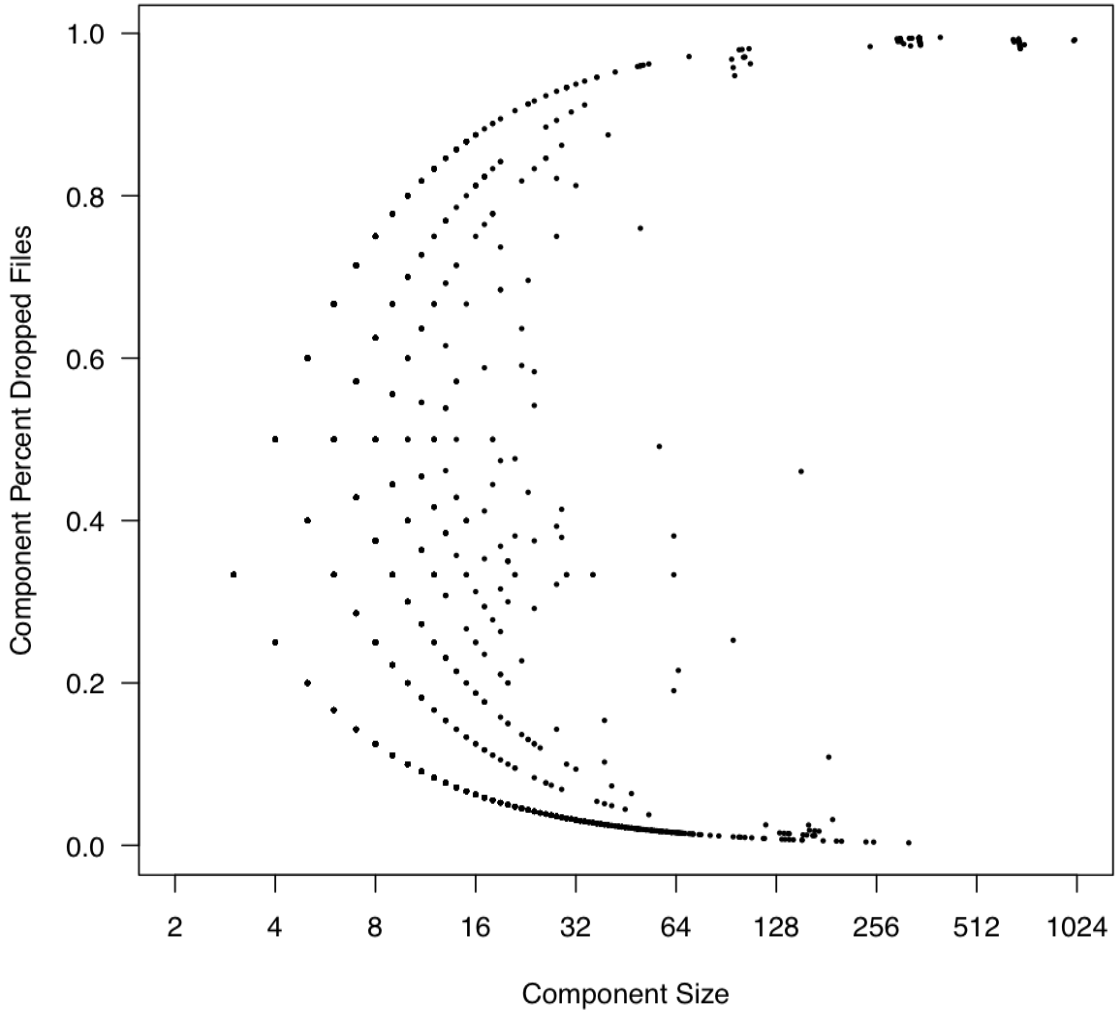


Connected Components

63.6m nodes, 350m edges, 35 minutes Graphchi on MacBook Pro

	Count	Size
1	1	63311055
2	1	1009
3	1	1002
4	1	712
5	2	692
6	1	691
7	1	690
8	1	689
9	2	687
10	2	686
11	3	685
12	3	684
13	1	663
14	1	662
15	1	659
16	1	398
17	2	347
18	4	346
19	7	345
20	16	344

78756 components

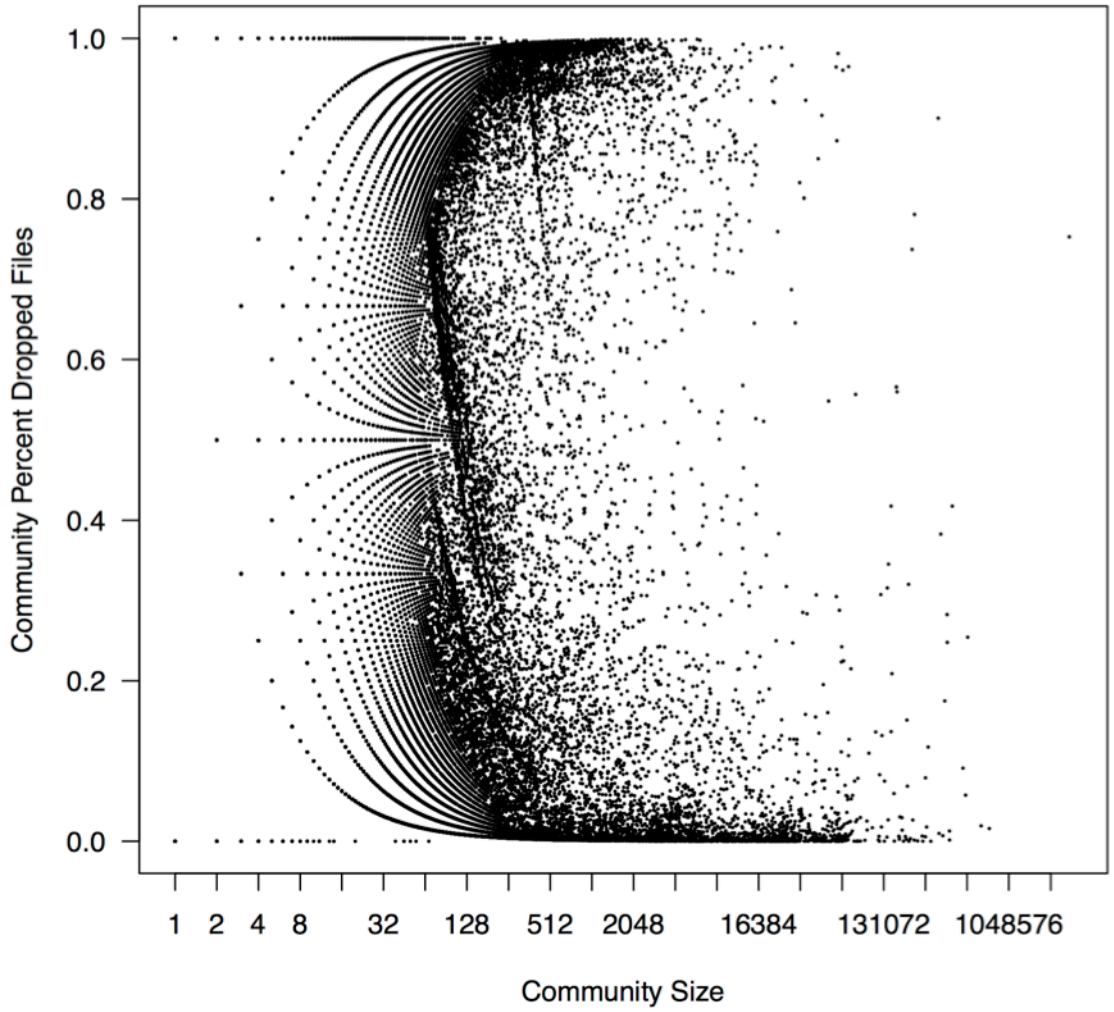


Community Detection

63.6m nodes, 350m edges, 45 minutes Graphchi on MacBook Pro, 2 minutes in Hadoop/Spark

	Count	Size
1	1	2853713
2	1	758666
3	1	662108
4	1	529684
5	1	511513
6	1	489196
7	1	409880
8	1	391510
9	1	391009
10	1	377154
11	1	376032
12	1	361557
13	1	350148
14	1	338525
15	1	328812
16	1	324644
17	1	286737
18	1	274181
19	1	267238
20	1	262437

632531 communities



Do the communities make sense?

Families:

```
35806 UNKNOWN
```

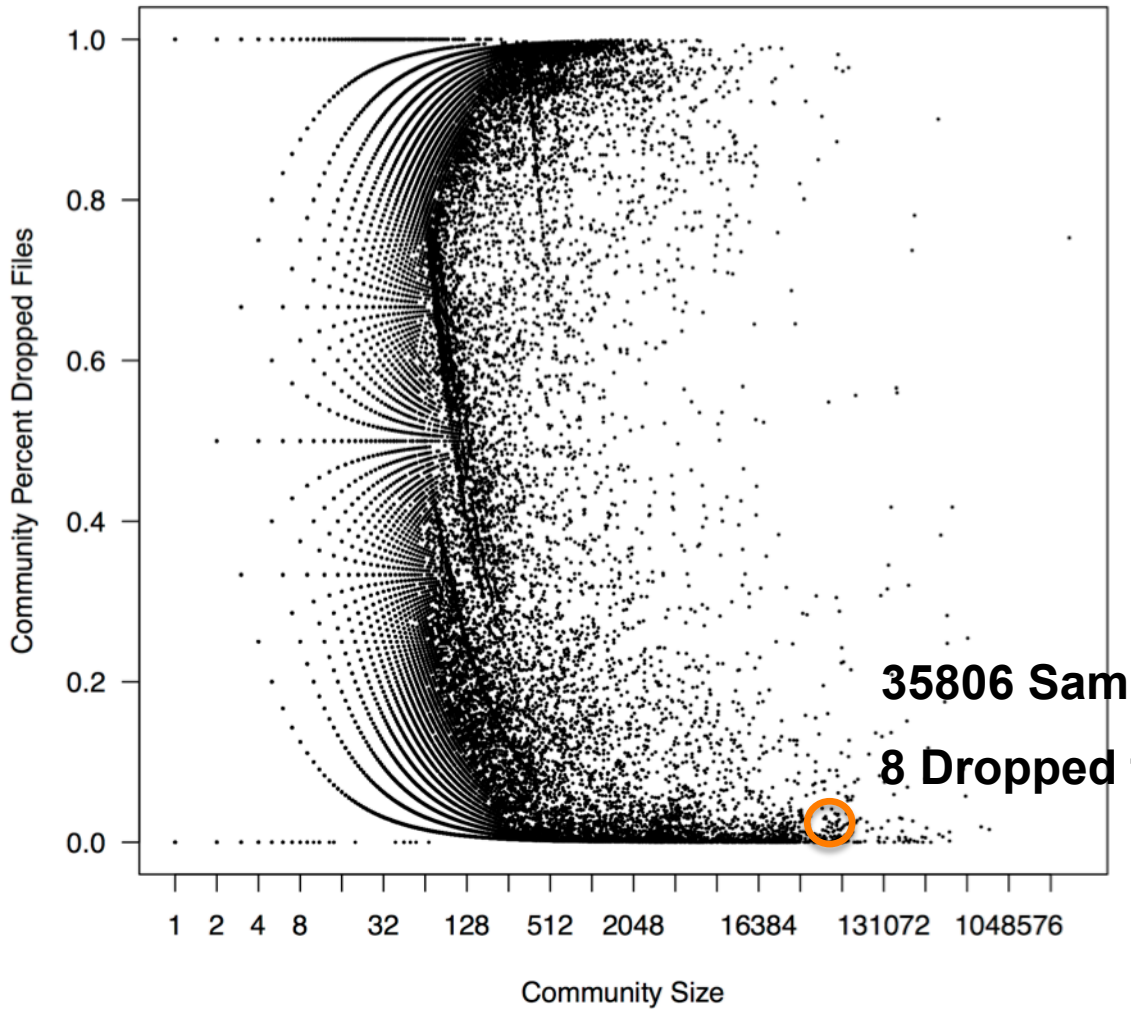
Filenames:

```
3 _setup.dll
1 custom.dll
1 readme.txt
1 setup.ico
1 wbemcore.log
```

Samples:

```
35887 custom.dll
35887 readme.txt
31524 _setup.dll
31524 setup.ico
```

Only these 35K files (run between Mar 30 and Jun 30 2014) use these specific readme.txt, custom.dll



Take-Home Points

Trusting Software

Software uniqueness and identity are intertwined more closely with behavior than in the physical world.

Analysis efforts benefit when both dynamic and static features are taken into account.

Fuzzy “Back-End” efforts can help to direct effort and resources for more costly, “Point-Query” high fidelity analysis.