DEPARTMENT OF SECURITY AND CRIME SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

UCL

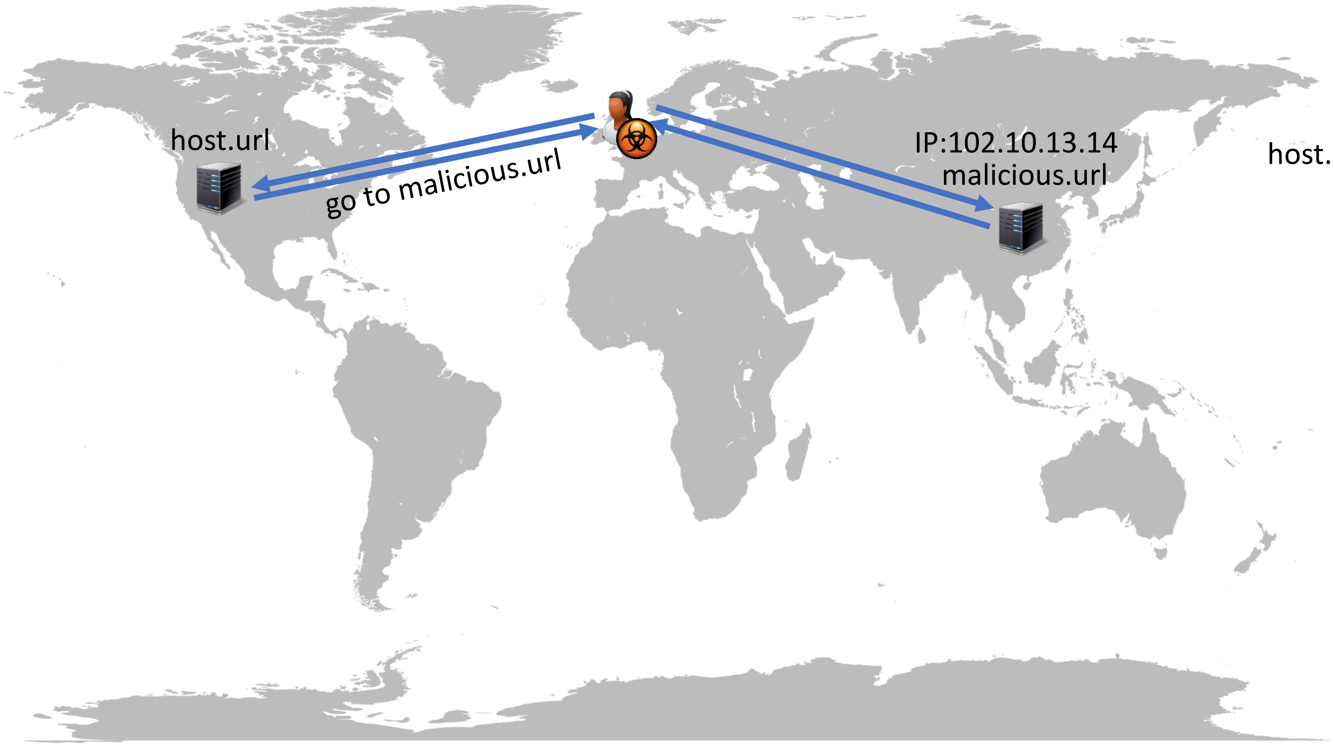
Waves of Malice: A Longitudinal Measurement of the Malicious File Delivery Ecosystem on the Web

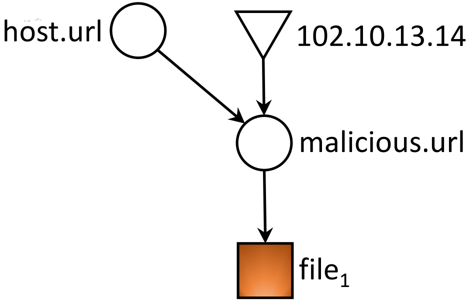
Colin C. Ife

Yun Shen, Steven J. Murdoch, Gianluca Stringhini

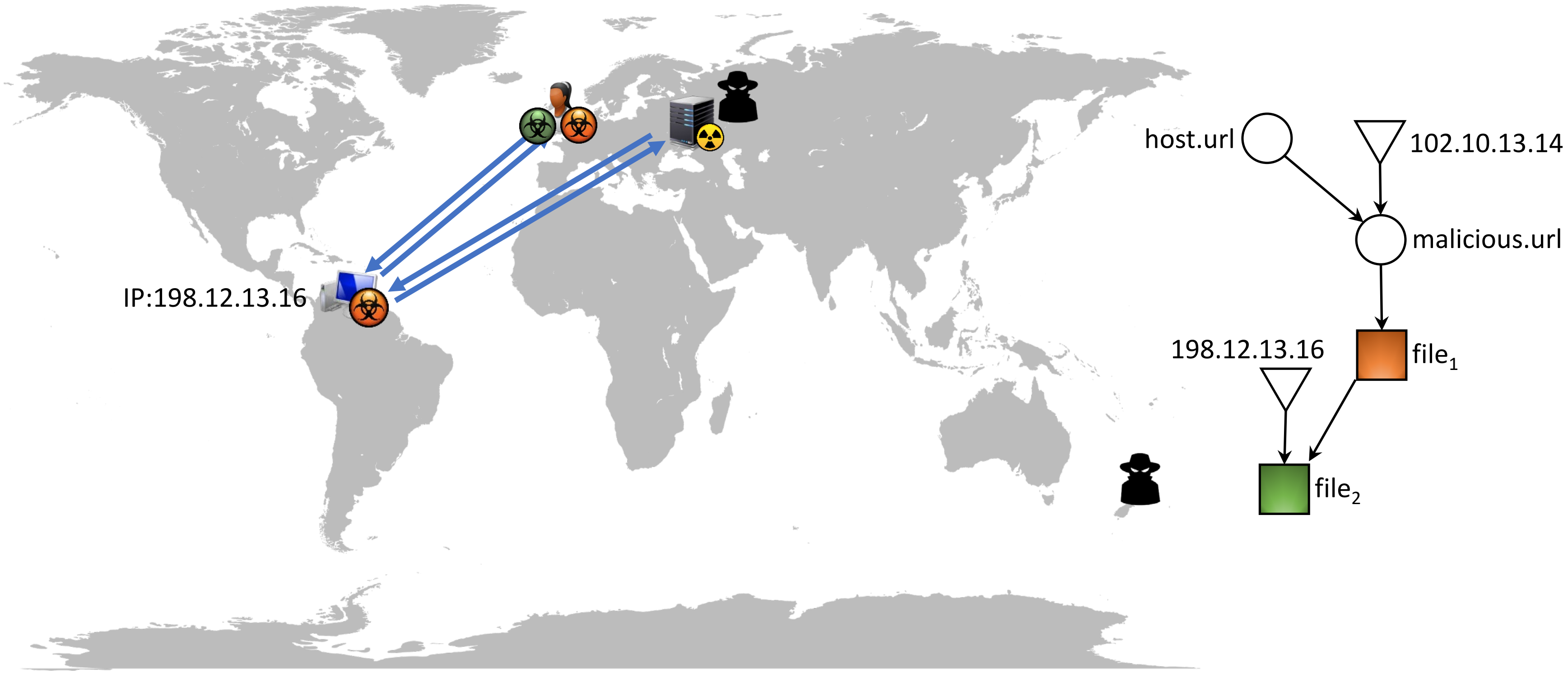
Symantec EPSRC THE ROYAL SOCIETY BOSTON UNIVERSITY

An example of a malicious file delivery event





An example of a malicious file delivery event



Research Aims

- ❑ **Analyze malware delivery networks (MDNs) from a global perspective and put other research into context**

Research Aims

- ❑ **Analyze malware delivery networks (MDNs) from a global perspective and put other research into context**
- ❑ **Answer important questions, such as:**

1. What does the malicious file delivery ecosystem look like?
2. How do the structures of networks delivering malware, potentially unwanted programs (PUP), or mixed payloads differ, if at all?
3. How do these infrastructures change over time?

Related Work

- ❑ **Downloader (Dropper) Graphs** ***(Kwon et al., 2015; 2016; Rossow et al., 2013)***
- ❑ **Pay-per install (PPI) Networks** ***(Caballero et al., 2011)***
- ❑ **PUP Distribution** ***(Thomas, 2016; Kotzias et al., 2015; 2016)***
- ❑ **Other respects**

Our Study

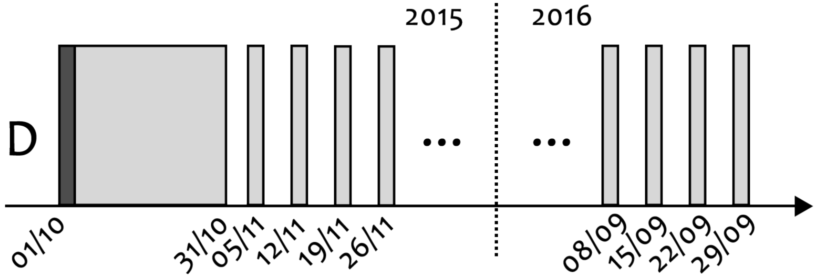
Analytical Approach

- ❑ **Snapshot study (24 hours)**
- ❑ **Longitudinal study (1 year)**

Dataset

❑ Symantec download telemetric data

❑ 129M download events (from 12M users)



❑ Focus on malicious files → Low reputation score

Dataset

A download event includes:

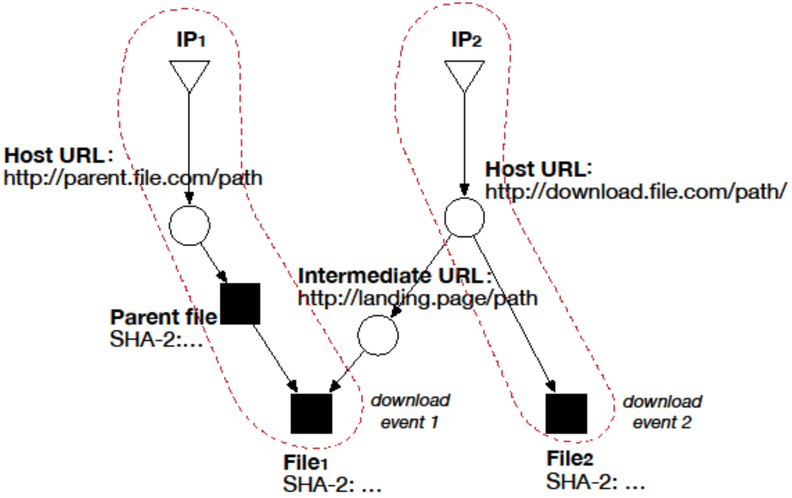
- Timestamp
- **SHA-2 of file (256 bits)**
- File name
- Size of file in bytes
- **Host URL**
- **Landing page URL** (after redirection from Host URL)
- **IP address** of server hosting file
- **Parent file SHA-2**
- **Landing page URL of parent file**

Data Representation

❑ Build a **directed graph** of download activity:

- Each unique file (SHA-2), host, or IP address are represented as **nodes**
- Downloads and network-level associations are represented as **directed edges**

❑ More **integrated** and **holistic** than past works

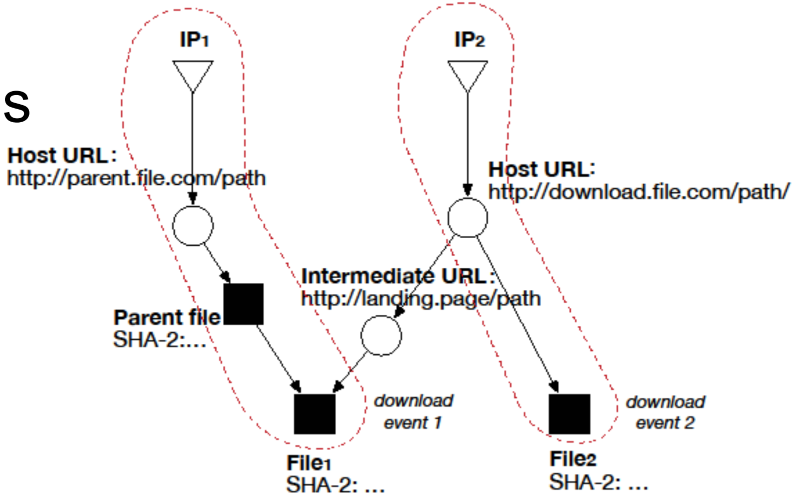


24-hour Snapshot Study

Snapshot Methodology

❑ Separating Components

- Identifying interacting operations
- Attributing infrastructure to actors

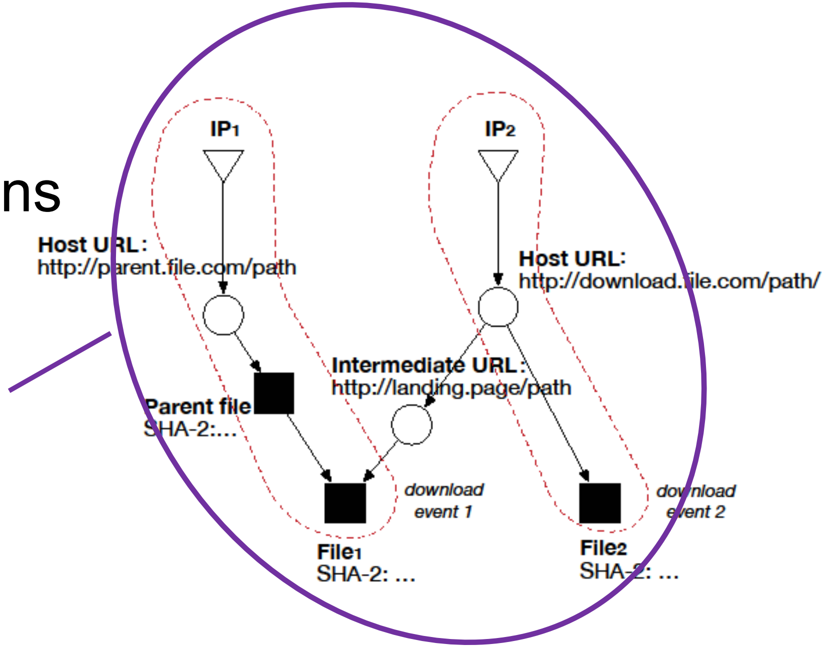


Snapshot Methodology

❑ Separating Components

- Identifying interacting operations
- Attributing infrastructure to actors

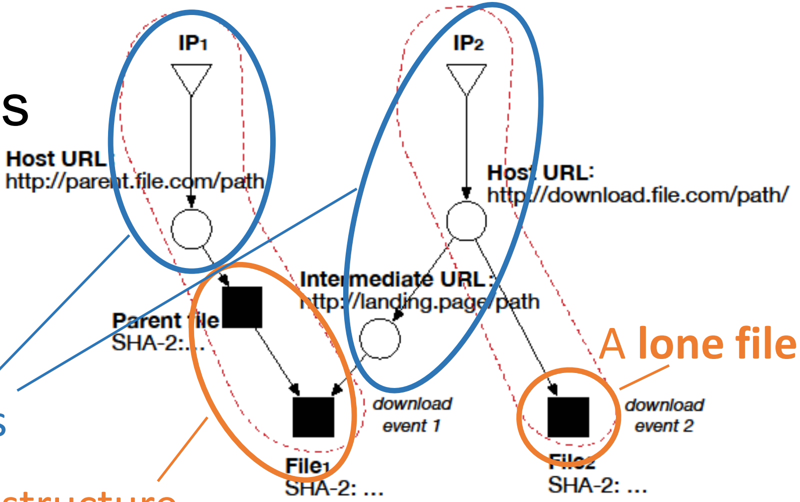
A component (weakly connected) – a single delivery operation, or two?



Snapshot Methodology

❑ Separating Components

- Identifying interacting operations
- Attributing infrastructure to actors



A lone file

Network (server-side) infrastructures

File (client-side) infrastructure

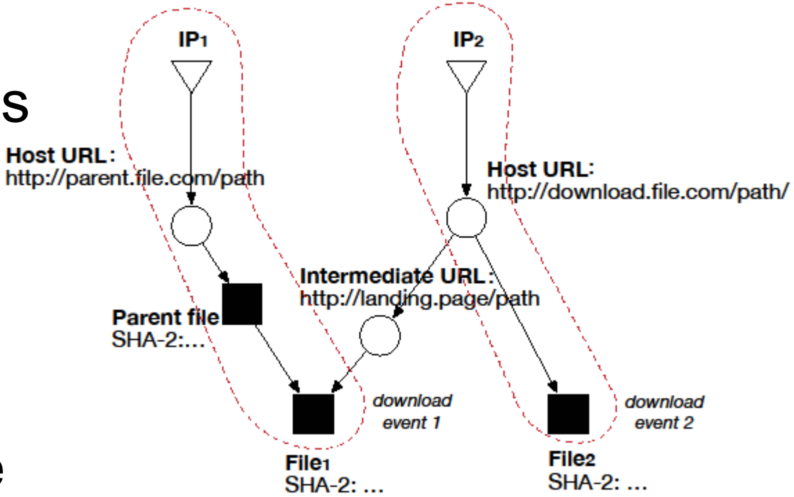
Snapshot Methodology

❑ Separating Components

- Identifying interacting operations
- Attributing infrastructure to actors

❑ File Classification

- Identifying malware, PUP, or unknown files/clusters, using the *VirusTotal* database and *AVClass* labeler *(Sebastian et al., 2016)*

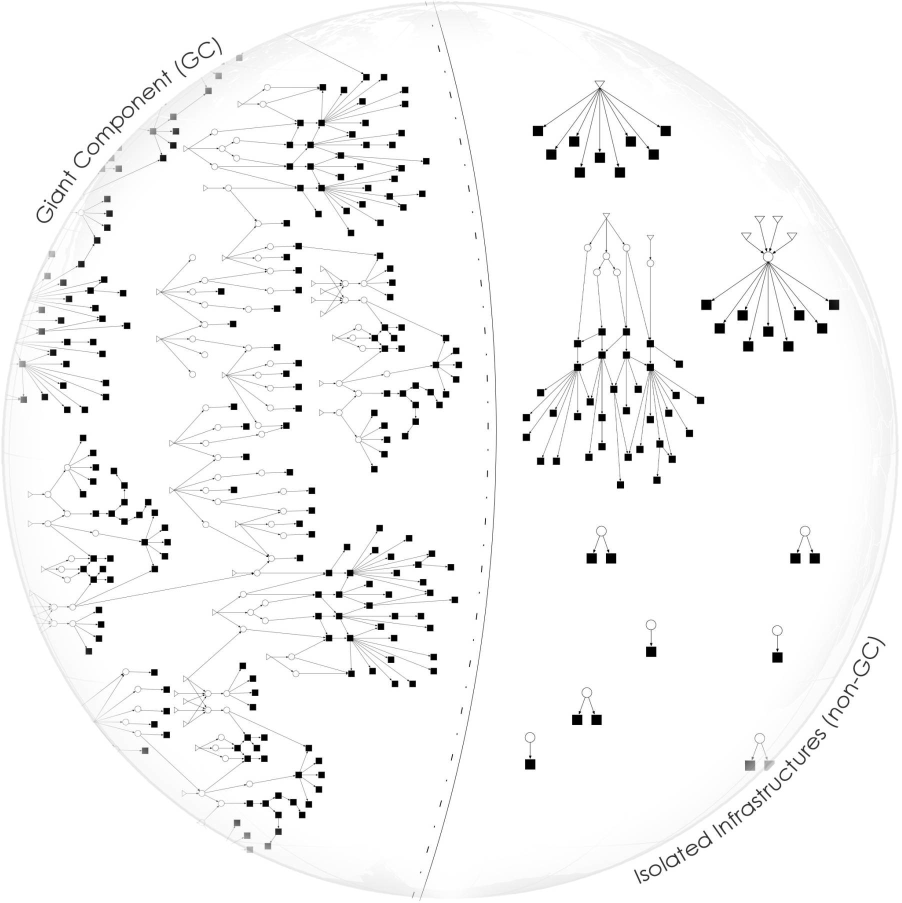


What We Found

Some Initial Statistics

- ❑ **Graph of 1.6M nodes, 1.9M edges:**
  - ▪ **965K unique files, 603K URLs (131K FQDNs), and 92K IPs**
  - ▪ **1.6M download events**

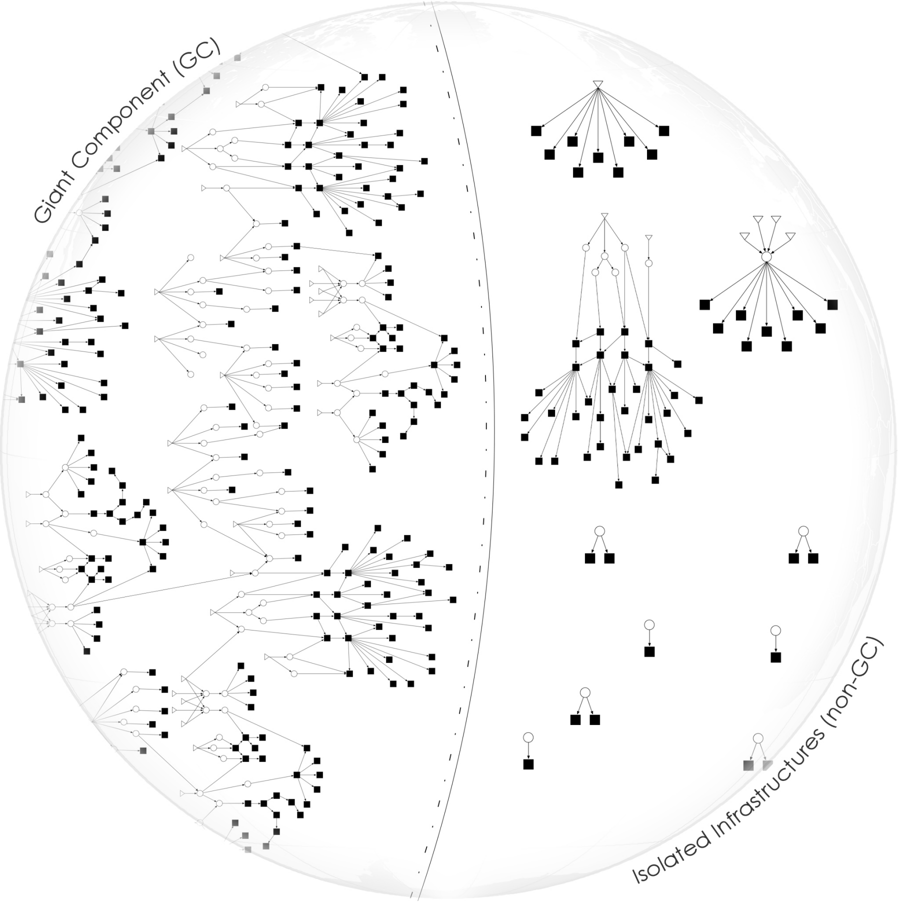
The Giant Component (GC)



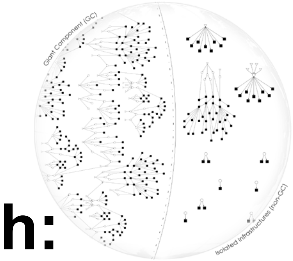
❑ **GC accounts for 80% of suspicious downloads**

- 1.3M nodes, 1.6M edges
- Next largest component is 2K nodes
- 58K total components

❑ **A massive operation (unlikely), or a well-connected marketplace?**



The Giant Component: Verification



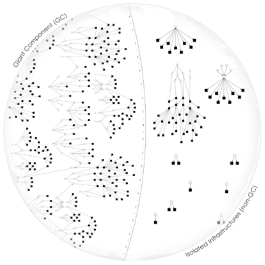
❑ We assess the validity of this finding through:

- Graph percolation/robustness experiments *(Callaway et al., 2000)*
- Rebuilding the graph without IPs and repeating graph percolation
- Blacklist popular effective second-level domains (e2LDs) → rule out shared use of popular IPs and e2LDs (e.g. Amazon EC2 instances)

❑ We find that the GC persists:

- 31% of its connectivity due to IPs, and **20% of GC** (180K file nodes) survives total removal of all server-side nodes.
- Persists over the course of the entire year’s data.

The Giant Component: Backbone



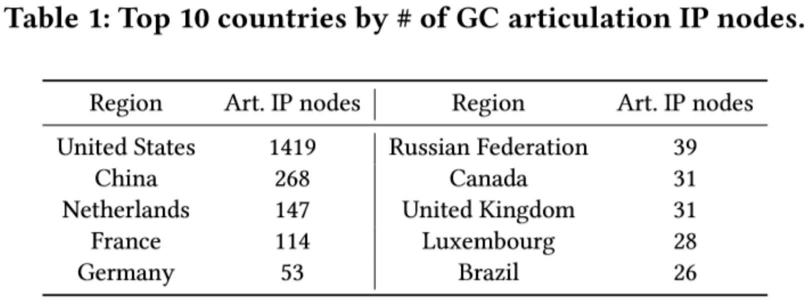


Table 1: Top 10 countries by # of GC articulation IP nodes.

Region	Art. IP nodes	Region	Art. IP nodes
United States	1419	Russian Federation	39
China	268	Canada	31
Netherlands	147	United Kingdom	31
France	114	Luxembourg	28
Germany	53	Brazil	26

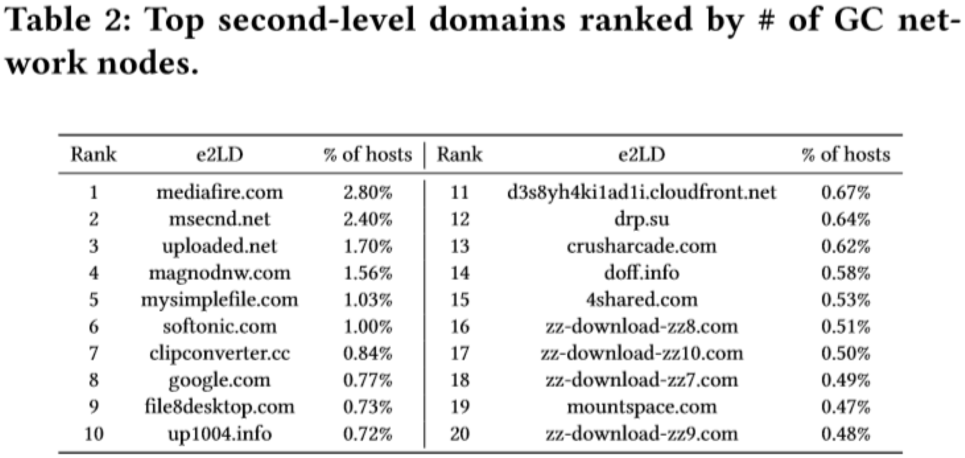
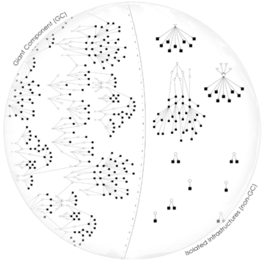


Table 2: Top second-level domains ranked by # of GC network nodes.

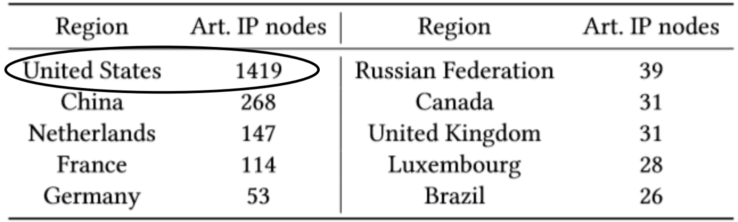
Rank	e2LD	% of hosts	Rank	e2LD	% of hosts
1	mediafire.com	2.80%	11	d3s8yh4ki1ad1i.cloudfront.net	0.67%
2	msecnd.net	2.40%	12	drp.su	0.64%
3	uploaded.net	1.70%	13	crusharcade.com	0.62%
4	magnodnw.com	1.56%	14	doff.info	0.58%
5	mysimplefile.com	1.03%	15	4shared.com	0.53%
6	softonic.com	1.00%	16	zz-download-zz8.com	0.51%
7	clipconverter.cc	0.84%	17	zz-download-zz10.com	0.50%
8	google.com	0.77%	18	zz-download-zz7.com	0.49%
9	file8desktop.com	0.73%	19	mountspace.com	0.47%
10	up1004.info	0.72%	20	zz-download-zz9.com	0.48%

The Giant Component: Backbone



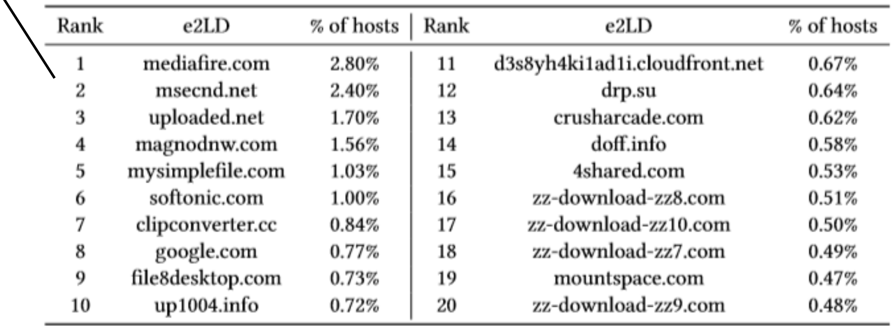
Some well-known services: MediaFire, Windows Azure CDN (msecnd.net), Softonic, Google,…

Table 1: Top 10 countries by # of GC articulation IP nodes.



Region	Art. IP nodes	Region	Art. IP nodes
United States	1419	Russian Federation	39
China	268	Canada	31
Netherlands	147	United Kingdom	31
France	114	Luxembourg	28
Germany	53	Brazil	26

Table 2: Top second-level domains ranked by # of GC network nodes.

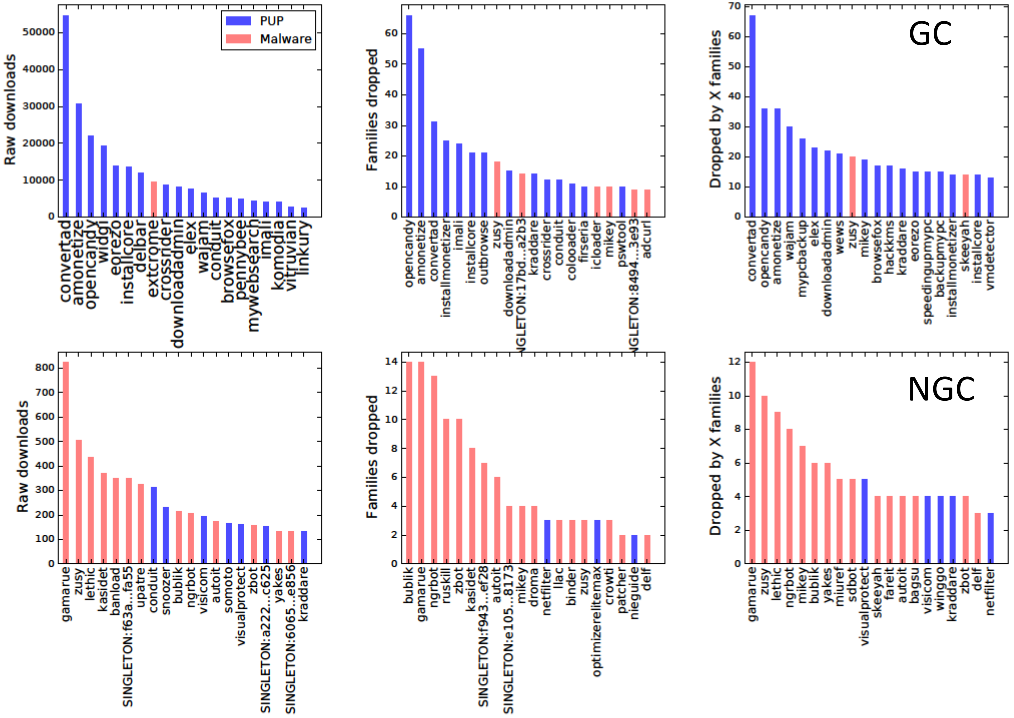


Rank	e2LD	% of hosts	Rank	e2LD	% of hosts
1	mediafire.com	2.80%	11	d3s8yh4ki1ad1i.cloudfront.net	0.67%
2	msecnd.net	2.40%	12	drp.su	0.64%
3	uploaded.net	1.70%	13	crusharcade.com	0.62%
4	magnodnw.com	1.56%	14	doff.info	0.58%
5	mysimplefile.com	1.03%	15	4shared.com	0.53%
6	softonic.com	1.00%	16	zz-download-zz8.com	0.51%
7	clipconverter.cc	0.84%	17	zz-download-zz10.com	0.50%
8	google.com	0.77%	18	zz-download-zz7.com	0.49%
9	file8desktop.com	0.73%	19	mountspace.com	0.47%
10	up1004.info	0.72%	20	zz-download-zz9.com	0.48%

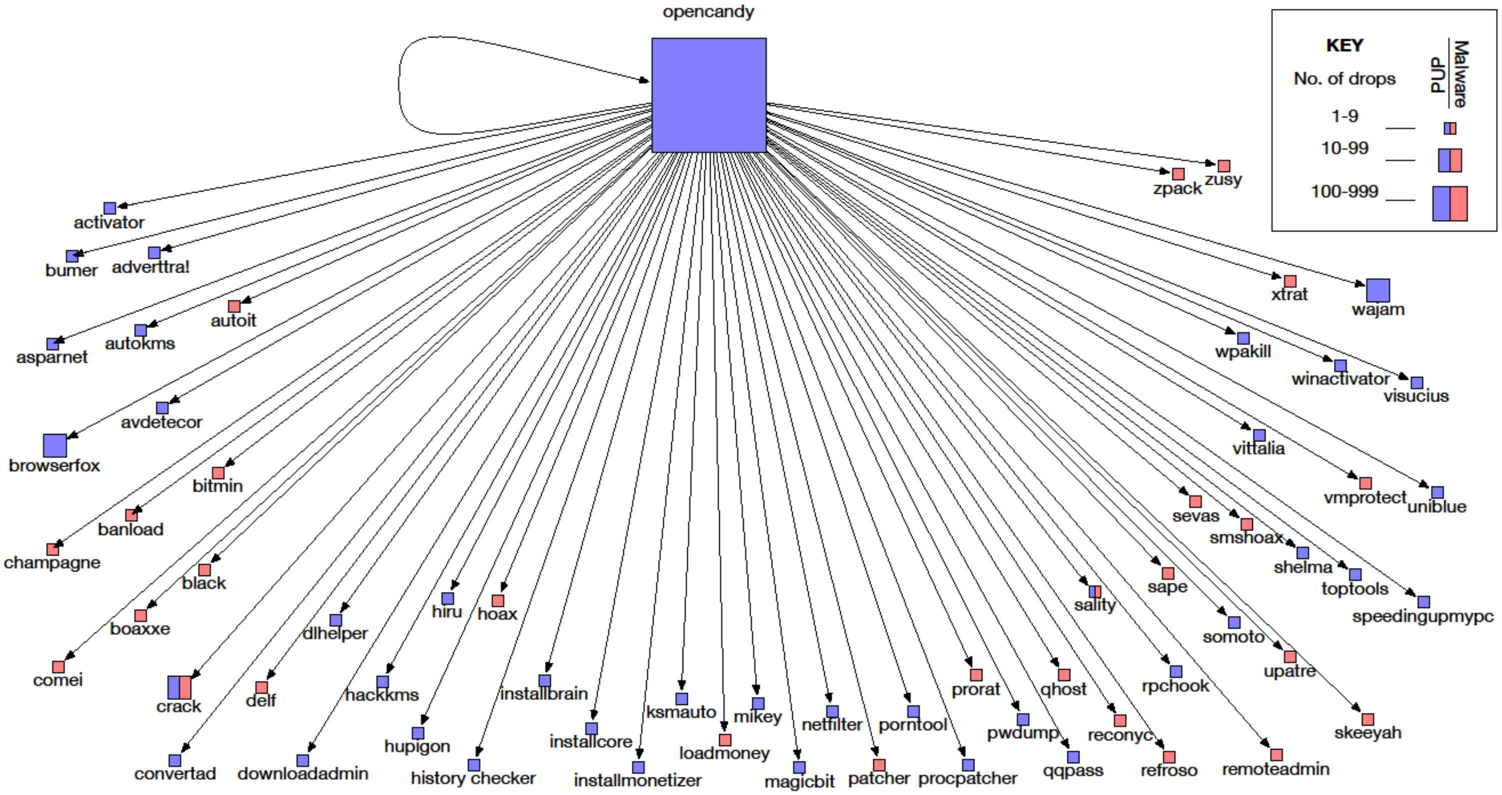
File Distributions of GC and NGC

- GC predominantly a **PUP Ecosystem**, while NGC predominantly a **Malware Ecosystem.**

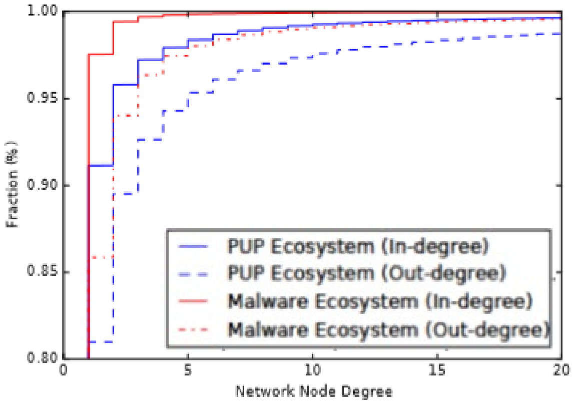
- PUP-to-malware ratios:
  - **5:1 (SHA-2s) and 17:2 (raw downloads) in the wild**
  - **8:1 (SHA-2s) and 11:1 (raw downloads) in GC**
  - **1:1.78 (SHA-2s) and 1:2.15 (raw downloads) in NGC**

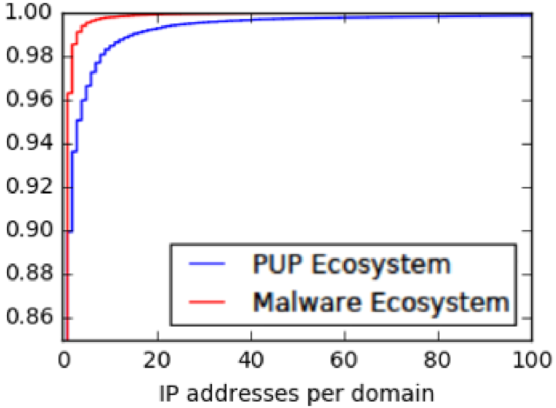


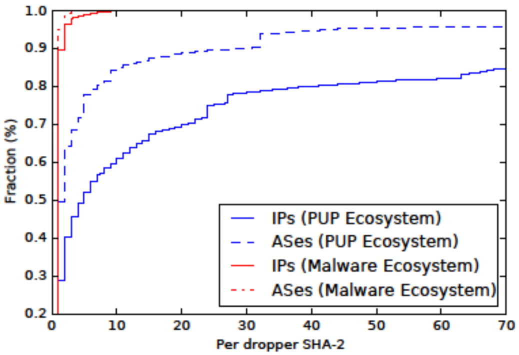
Case Study: Opencandy Operation



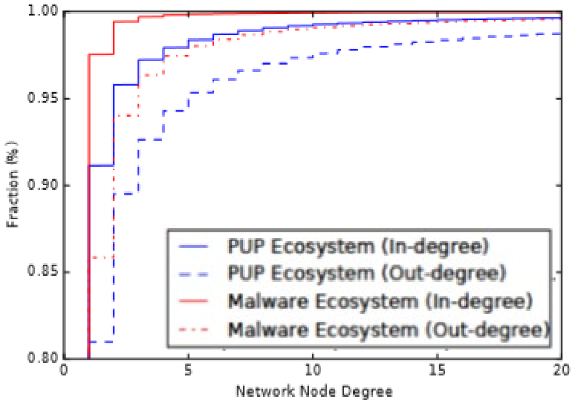
Comparing Ecosystem Structures (1)

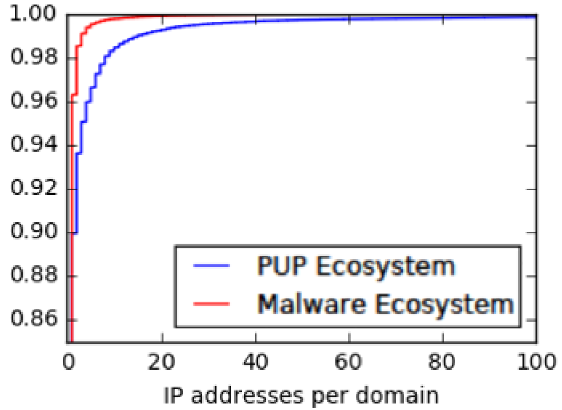


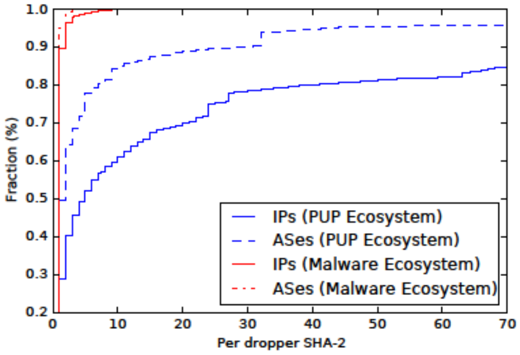




Comparing Ecosystem Structures (1)



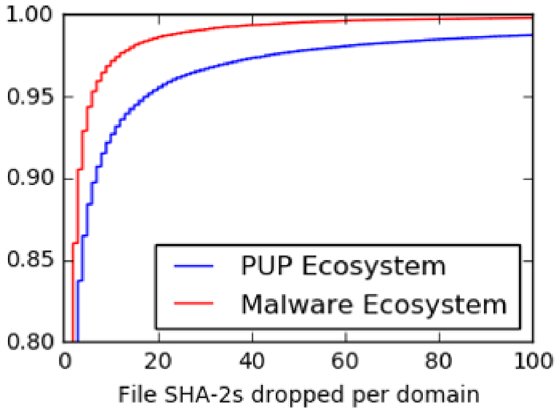


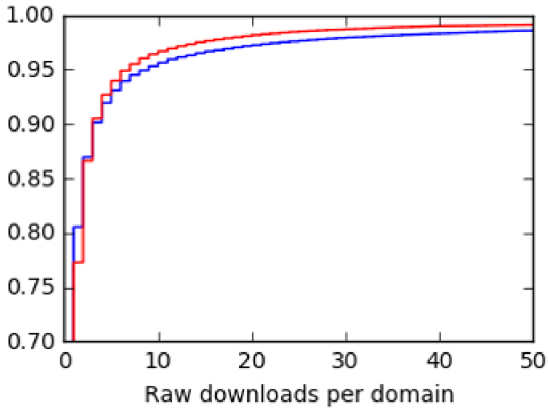


❑ **PUP Ecosystem: higher IP/AS usage and more URL redirections**

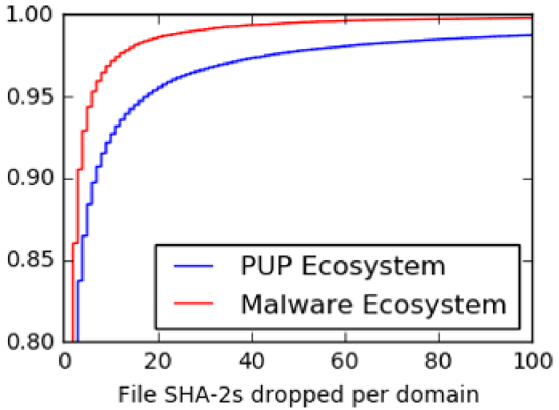
➔ **Higher CDN usage? Fast flux?**

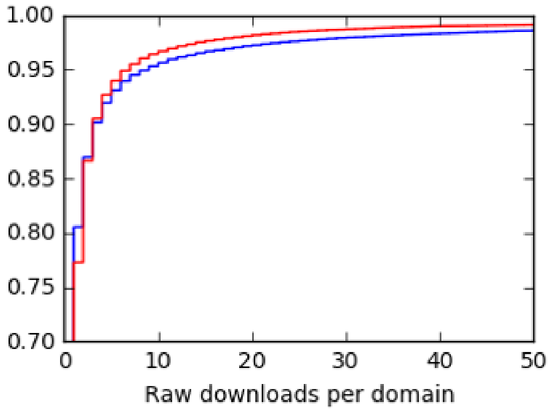
Comparing Ecosystem Structures (2)





Comparing Ecosystem Structures (2)





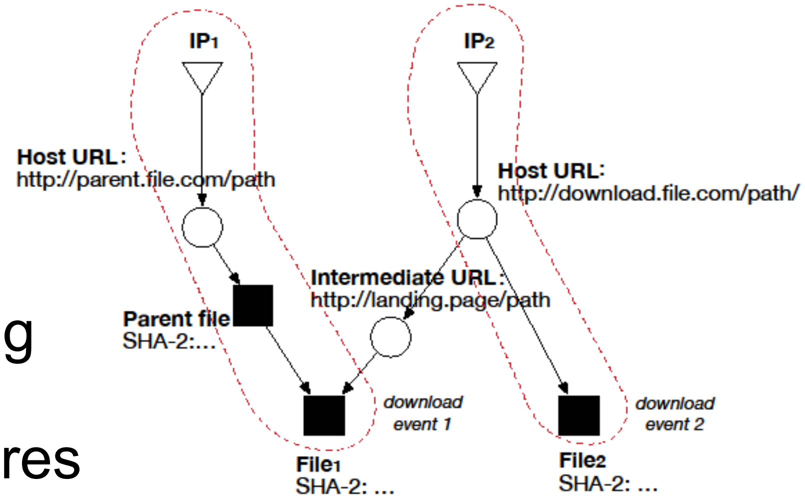
❑ **Malware Ecosystem: fewer SHA-2s dropped per domain but similar # of raw downloads**

➔ **Lower CDN usage? Evasive techniques?**

Longitudinal Study

Longitudinal Methodology

- ❑ **Snapshot Processing**
  - ▪ Repeat snapshot generation process
- ❑ **Component Tracking**
  - ▪ Generate signatures for tracking server-side (network-only) and client-side (file-only) infrastructures
  - ▪ Track these infrastructures in time



Infrastructure Churn

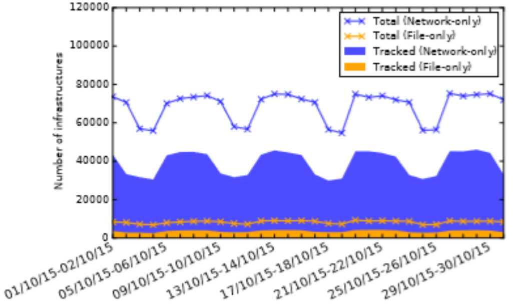


Figure 8: Daily churn of delivery infrastructures over a month.

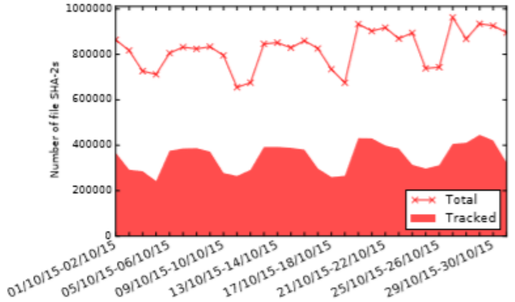


Figure 9: Daily churn of lone file SHA-2s over a month.

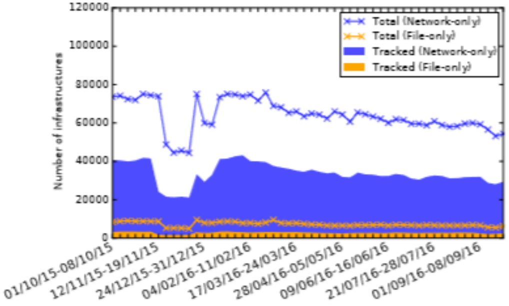


Figure 10: Weekly churn of delivery infrastructures over a year.

Infrastructure Churn

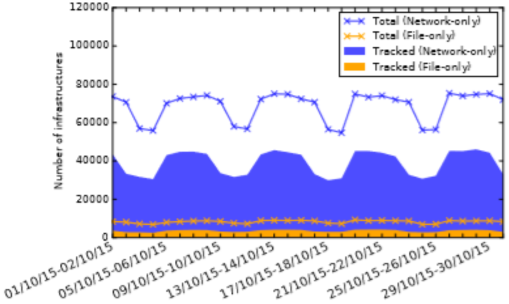


Figure 8: Daily churn of delivery infrastructures over a month.

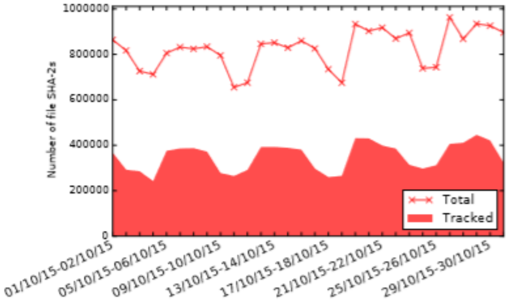


Figure 9: Daily churn of lone file SHA-2s over a month.

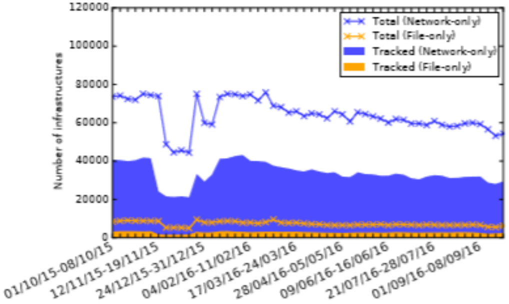


Figure 10: Weekly churn of delivery infrastructures over a year.

Cyclic pattern: Infection activity could follow the availability of victims -> Routine Activity Theory *(Cohen and Felson, 1979)*

Big dip in activity between mid-Nov and mid-Dec ’15

Infrastructure Lifespans

- 75% network infrastructures active for at least 6 weeks.
- 26% network and 10% file infrastructures active for a year.

- Malware operations last a median of 5 weeks vs. 3 weeks for PUP operations.
- Sample size issues with ‘Mixed’

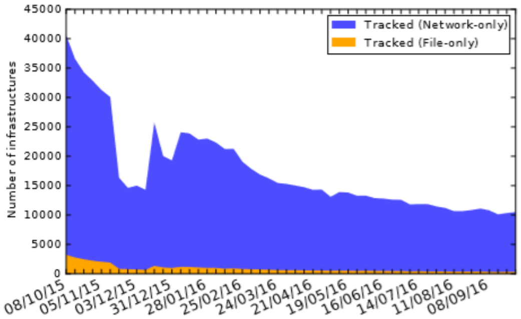


Figure 11: Lifespan of delivery infrastructures tracked from 1st October 2015, over a year.

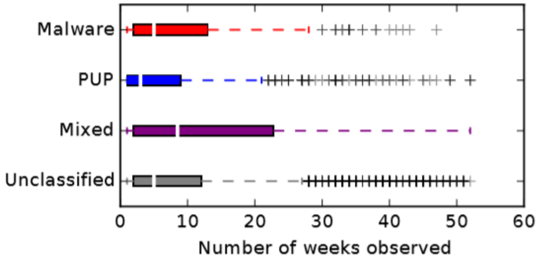


Figure 12: Box plots showing the lifespan of file delivery infrastructures.

Case Study: Dyre Takedown Operation

Source: Symantec, 2016

Case Study: Dyre Takedown Operation

- **Dyre** was a financial fraud trojan controlled by a cybercriminal group and installed by the **Upatre** dropper.
- After the takedown operation by Russian LEA in Nov ‘15, Symantec report virtual cessation of Dyre and Upatre activity.
- In our analysis, we found a significant drop in Upatre activity, but also in the activity of other popular PPI droppers and malware families at the same time:
- Amonetize, Installcore, Eorezo, Convertad PUP PPIs as well as Neshta malware.
- Shared infrastructure? Business relationships?

Discussion

Implication of Findings

- Both legitimate and malicious services involved in unwanted software delivery → **inform benign services to tighten security practices; takedown illegitimate ones**

Implication of Findings

- Both legitimate and malicious services involved in unwanted software delivery → **inform benign services to tighten security practices; takedown illegitimate ones**
- IPs from the US are core to the PUP Ecosystem → **most effective target for Internet service provider (ISP) takedowns?**

Implication of Findings

- Both legitimate and malicious services involved in unwanted software delivery → **inform benign services to tighten security practices; takedown illegitimate ones**
- IPs from the US are core to the PUP Ecosystem → **most effective target for Internet service provider (ISP) takedowns?**
- 26% of network infrastructures survive over a year → **these IP addresses and servers are stable, so focus on these (blacklists, takedowns, improve hygiene)**

Limitations

- Data collection biases (geographic, behavioural, etc.)

Limitations

- Data collection biases (geographic, behavioural, etc.)
- Ground-truth: only 10% of our snapshot dataset was covered by VirusTotal

Limitations

- Data collection biases (geographic, behavioural, etc.)
- Ground-truth: only 10% of our snapshot dataset was covered by VirusTotal
- Analysis: cannot see other *inter-URL connections* or *infection vectors*; malware can rapidly change their SHA-2s (re-packing)

Future Works

- ❑ **Repeatability studies**
  - ▪ other company (or open-source) data; more recent data; mobile downloads
- ❑ **Detecting botnets by graph evolution**
- ❑ **Evaluating current mitigations and identifying better ones through data-driven analysis**

Conclusion

❑ **Comprehensive data-driven analysis of MDNs on the Web, with a methodology to identify its key elements**

Conclusion

- ❑ Comprehensive data-driven analysis of MDNs on the Web, with a methodology to identify its key elements
- ❑ Two disjoint ecosystems, with the (stable) PUP Ecosystem conducting lion's share of suspicious downloads

Conclusion

❑ Comprehensive data-driven analysis of MDNs on the Web, with a methodology to identify its key elements

❑ Two disjoint ecosystems, with the (stable) PUP Ecosystem conducting lion’s share of suspicious downloads

❑ Estimated ratios of PUP-to-malware in the wild and differentiated in the two ecosystems’ characteristics

Conclusion

❑ Comprehensive data-driven analysis of MDNs on the Web, with a methodology to identify its key elements

❑ Two disjoint ecosystems, with the (stable) PUP Ecosystem conducting lion’s share of suspicious downloads

❑ Estimated ratios of PUP-to-malware in the wild and differentiated in the two ecosystems’ characteristics

❑ Found that most network hosts are volatile, but 26% are stable for over a year

Conclusion

❑ Comprehensive data-driven analysis of MDNs on the Web, with a methodology to identify its key elements

❑ Two disjoint ecosystems, with the (stable) PUP Ecosystem conducting lion’s share of suspicious downloads

❑ Estimated ratios of PUP-to-malware in the wild and differentiated in the two ecosystems’ characteristics

❑ Found that most network hosts are volatile, but 26% are stable for over a year

@ColinIfe

colinife.com

colin.ife@ucl.ac.uk

Thank you for listening!

References

❑ B. J. Kwon, J. Mondal, J. Jang, L. Bilge, and T. Dumitras. The dropper effect: Insights into malware distribution with downloader graph analytics. In ACM Conference on Computer and Communications Security (CCS), 2015.

❑ B. J. Kwon, V. Srinivas, A. Deshpande, and T. Dumitraş. Catching worms, trojan horses and pups: Unsupervised detection of silent delivery campaigns. arXiv preprint arXiv:1611.02787, 2016.

❑ C. Rossow, C. Dietrich, and H. Bos. Large-scale analysis of malware downloaders. In Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA), 2013.

❑ J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring pay-per-install: The commoditization of malware distribution. In USENIX Security Symposium, 2011.

❑ K. Thomas, J. Crespo, J.-M. Picod, C. Phillips, C. Sharp, M.-A. Decoste, A. Tofigh, M.-A. Courteau, L. Ballard, R. Shield, N. Jagpal, M. Abu Rajab, P. Mavrommatis, N. Provos, E. Bursztein, and D. McCoy. Investigating Commercial Pay-Per-Install and the Distribution of Unwanted Software. In USENIX Security Symposium, 2016.

❑ P. Kotzias, S. Matic, R. Rivera, and J. Caballero. Certified PUP: Abuse in authenticode code signing. In ACM Conference on Computer and Communications Security (CCS), 2015.

❑ P. Kotzias, L. Bilge, and J. Caballero. Measuring PUP Prevalence and PUP Distribution through Pay-Per-Install Services. In USENIX Security Symposium, 2016.

❑ M. Sebastián, R. Rivera, P. Kotzias, and J. Caballero. Avclass: A tool for massive malware labeling. In International Symposium on Research in Attacks, Intrusions, and Defenses (RAID), 2016.

❑ L. E. Cohen and M. Felson. Social change and crime rate trends: A routine activity approach. American sociological review, 1979.

❑ Symantec. Dyre: Operations of bank fraud group grind to halt following takedown. https://www.symantec.com/connect/blogs/ dyre-operations-bank-fraud-group-grind-halt-following-takedown, 2016. [Online; accessed 11-August-2017].