NortonLifeLock™

12 DEC 2019

APPLYING THE GUILT BY ASSOCIATION PRINCIPLE TO THREAT DETECTION WITH SPARSELY LABELED DATA

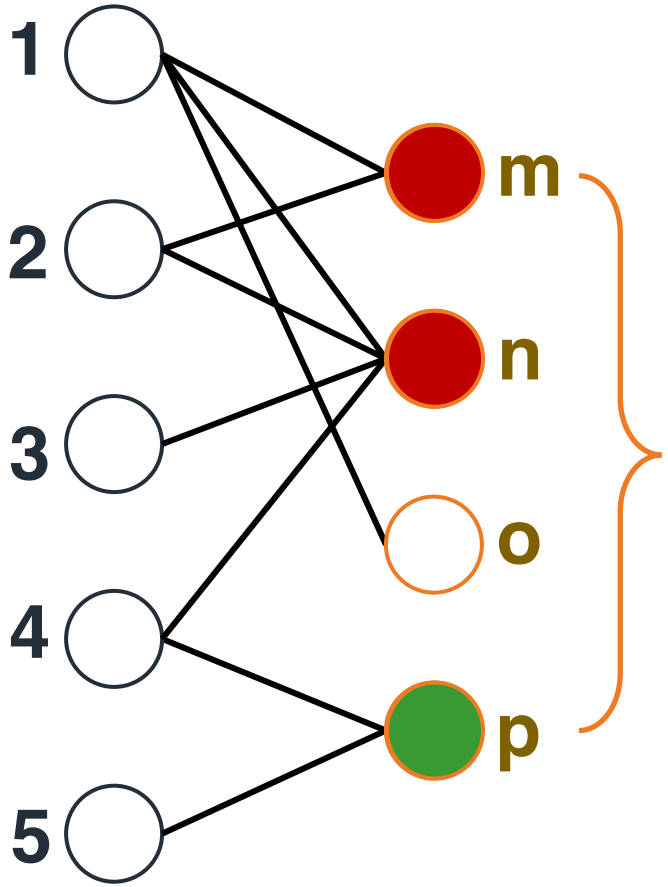
KEVIN ALEJANDRO ROUNDY
NORTONLIFELOCK RESEARCH GROUP (NRG)

“The Ass and His Purchaser” – by Aesop

Does Guilt By Association Apply in Computer Security?

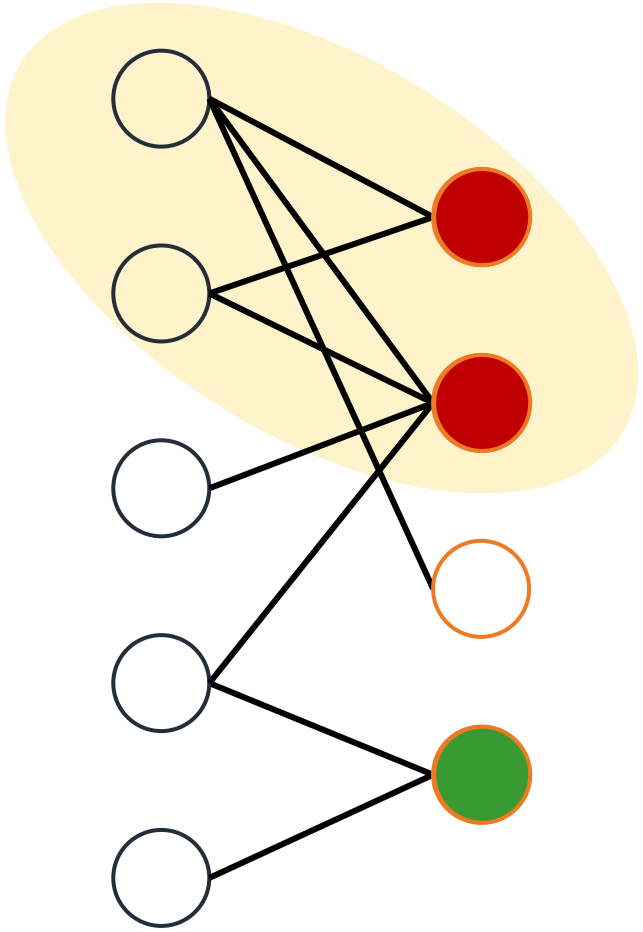
Relationship Data

Mobile Devices
PC's / Laptops
IP Addresses



Apps
Software .exe .dll files
Alerts from AV, IDS, FW, …

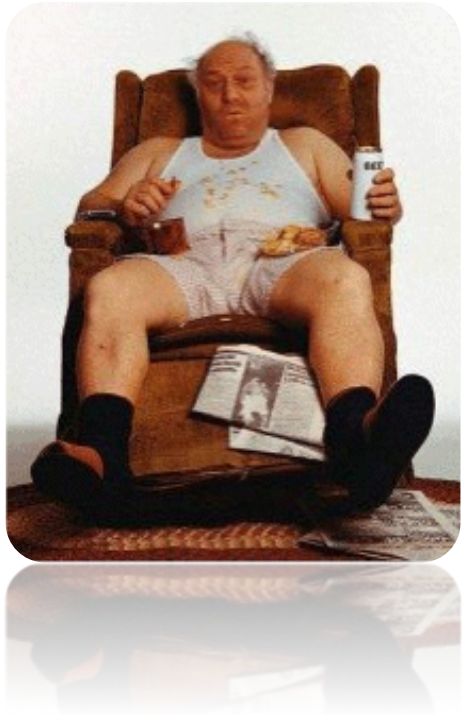
Challenges / Opportunities

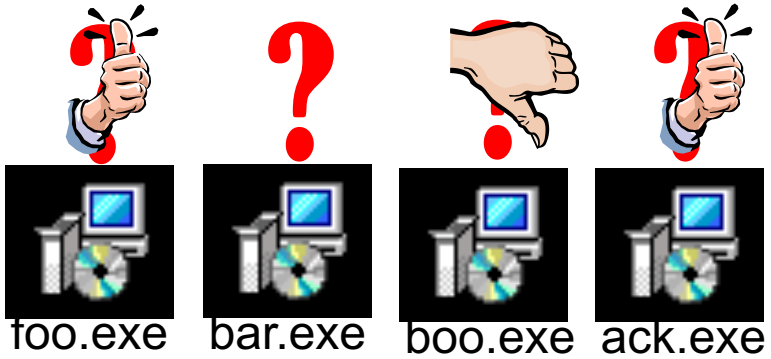


1. Billions of nodes, trillions of edges
2. Miniscule amounts of labeled data of only one class
3. Massive imbalance between classes
4. Nodes representing rare entities for which we have very few observations
5. Cliques in the bipartite graph

Machine Hygiene

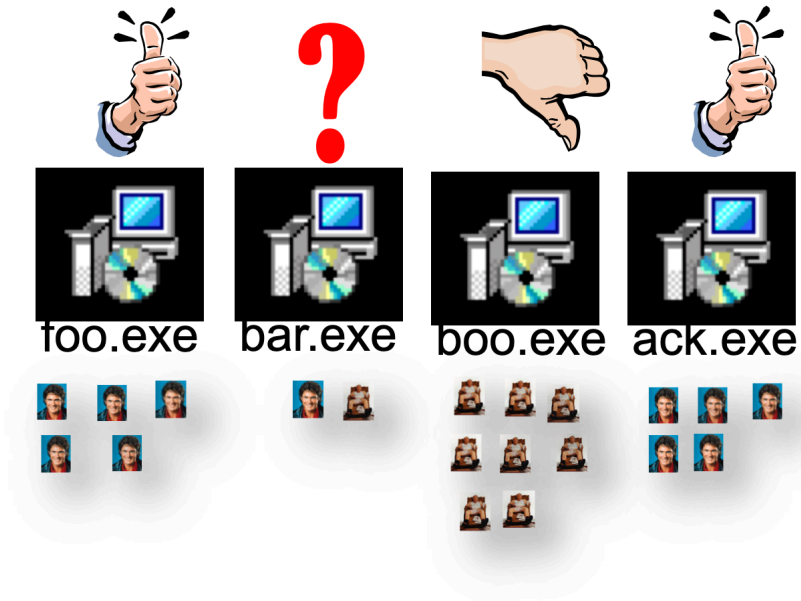
"Polonium: Tera-Scale Graph Mining and Inference for Malware Detection," by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.



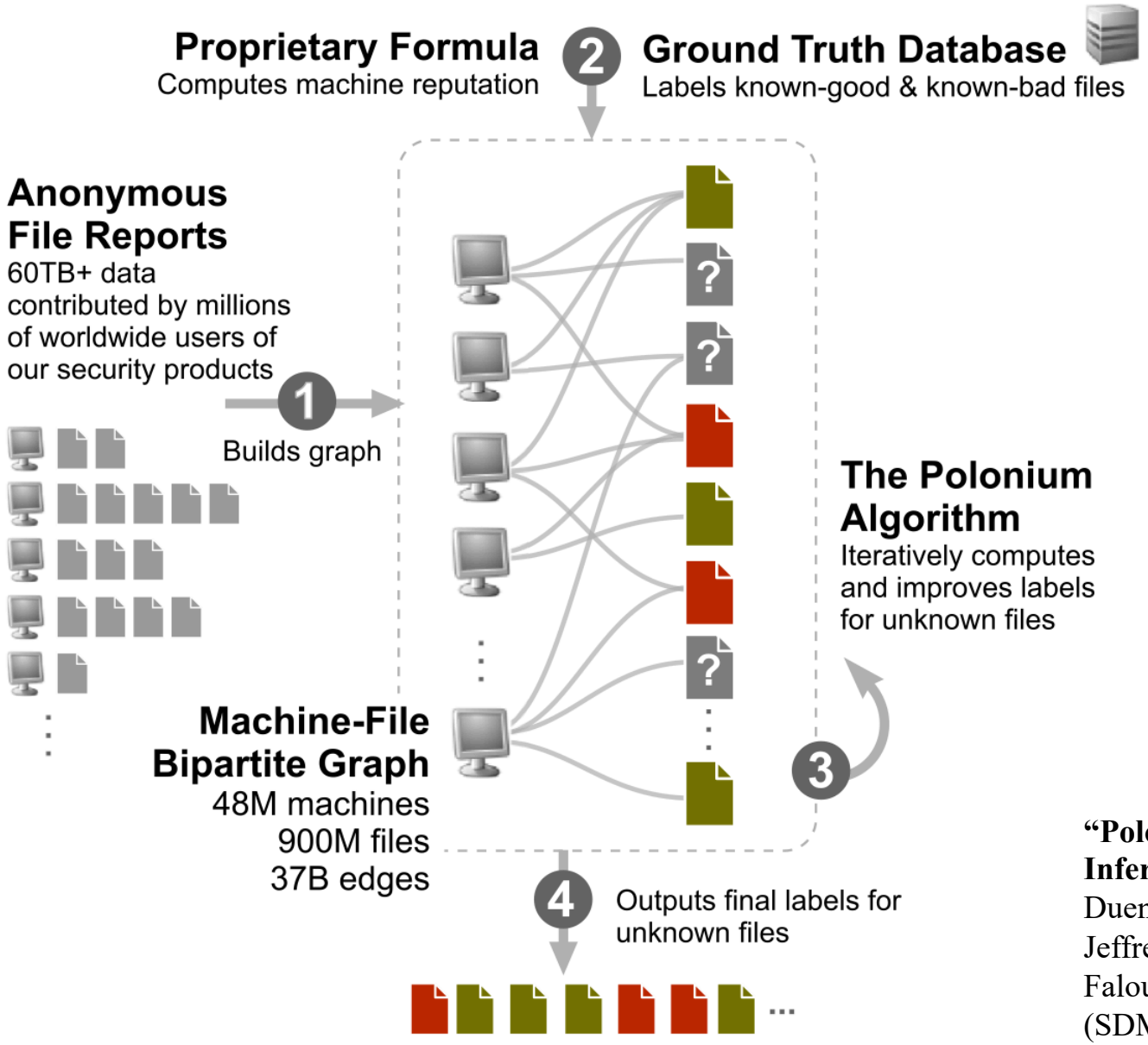


Machine Hygiene

"Polonium: Tera-Scale Graph Mining and Inference for Malware Detection," by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.



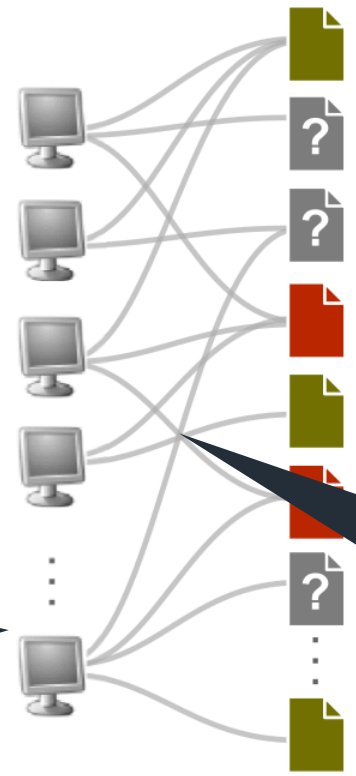
Polonium Technology Overview



"Polonium: Tera-Scale Graph Mining and Inference for Malware Detection," by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.

Belief Propagation Parameterized for Polonium

Belief propagation is a message passing algorithm for inference on graphs using Bayesian principles



File Prior
.99 for GT good, .01 for GT bad, function of prevalence for other files

Machine Prior
(aka node potential func)
is a function of the machine reputation score

Edge Potential
Machine reputation propagates along edges to files

era-Scale Graph Mining and Malware Detection," by g Chau, Carey Nachenberg, y Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.

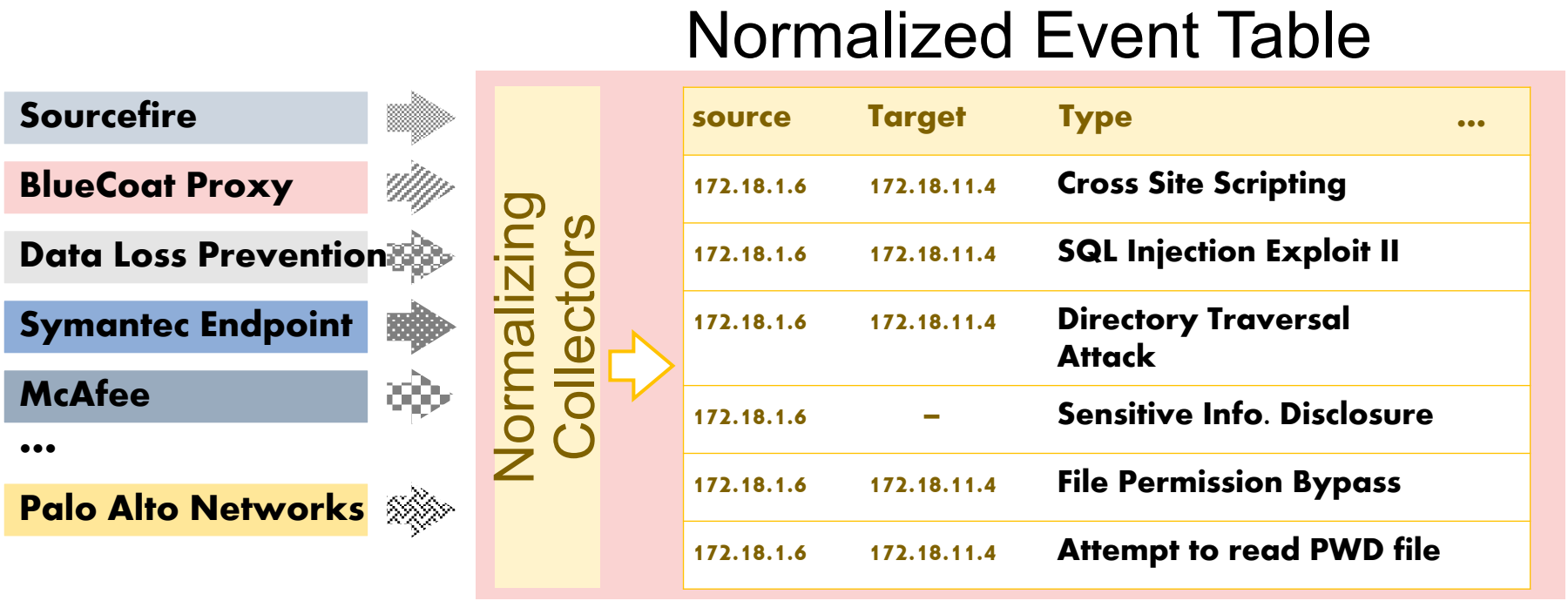
Smoke Detector

“Smoke Detector: Cross-Product Intrusion Detection With Weak Indicators,” Kevin A. Roundy, Acar Tamersoy, Michael Spertus, Michael Hart, Daniel Kats, Matteo Dell’Amico, Robert Scott. ACSAC 2017.

Managed Security Service Providers

MSSP running Security Operations Centers on behalf of 700 customers

100 products producing 2 trillion events per month, of 100,000 distinct event types



Challenges

Scale!

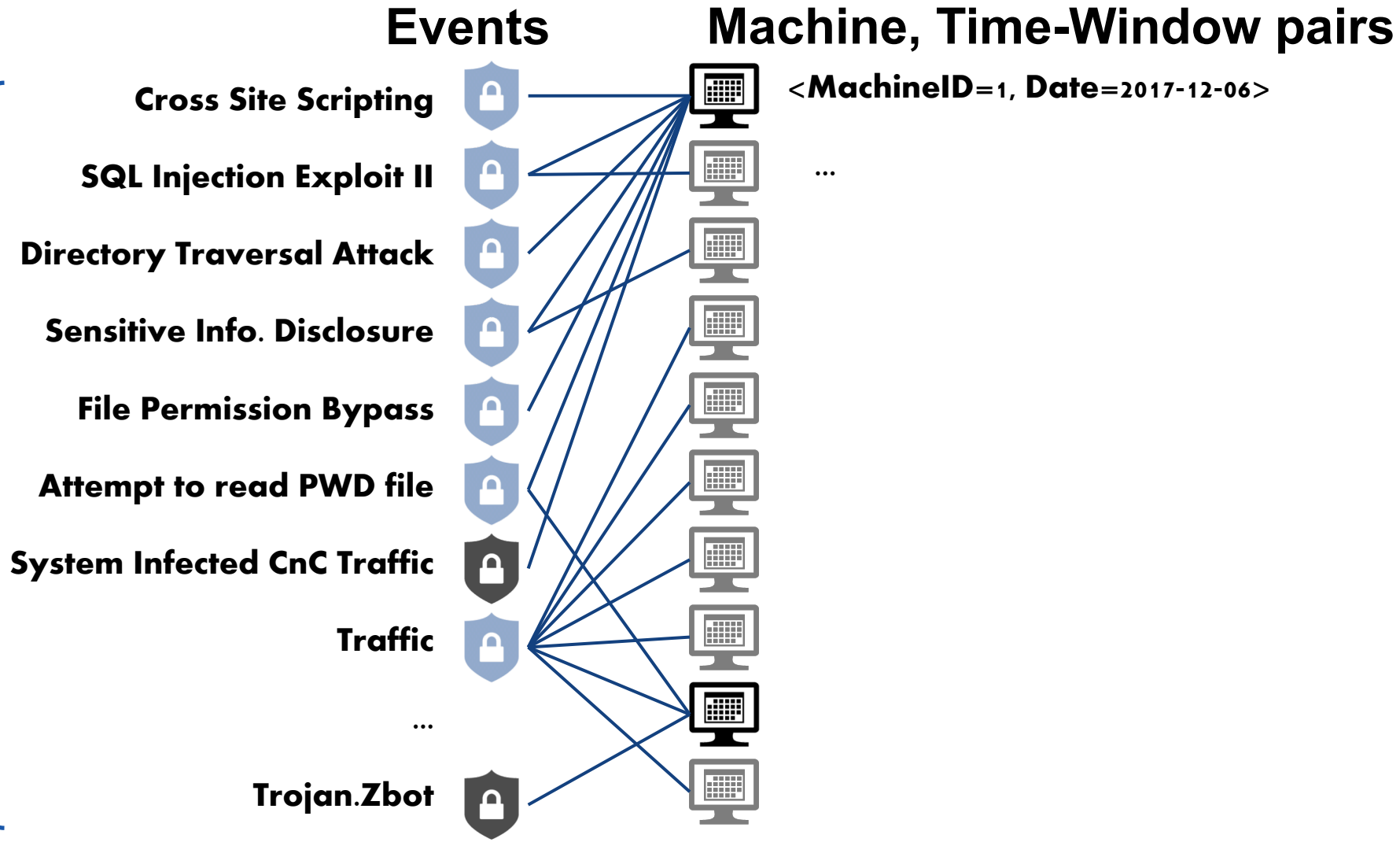
Security events are either **neutral or bad**, there are none that make you think a machine is not infected

The **noisier** an event is, the more highly connected it is, and the **more influential** it is in the graph

Instance Count	Confidence	Signature Name
44,104,172,561	0.01	TCP Connection
21,396,843,738	0.00	Traffic
19,493,074,472	0.00	UPD Connection
15,161,094,870	0.00	Firewall
14,586,679,865	0.02	Teardown TCP Connection
6,905,911,034	0.00	Teardown UDP Connection
6,762,250,252	0.00	Flow Session Close
4,350,063,989	0.01	PIX-6-305012
3,790,388,695	0.01	Connection Discarded
2,626,258,118	0.02	TPC Cache Miss
2,381,224,704	0.00	HTTP Get
2,109,964,562	0.02	Packet Permitted
1,676,587,290	0.02	Connection Allowed

Approach: Learning from Semi-Supervised Data and Relationships

Weight events based on the probability with which they occur on infected machines



Part 2: Random Walk with Restart for Incident Ranking

A random particle starts traversing the graph from incident node *i* (restart node)

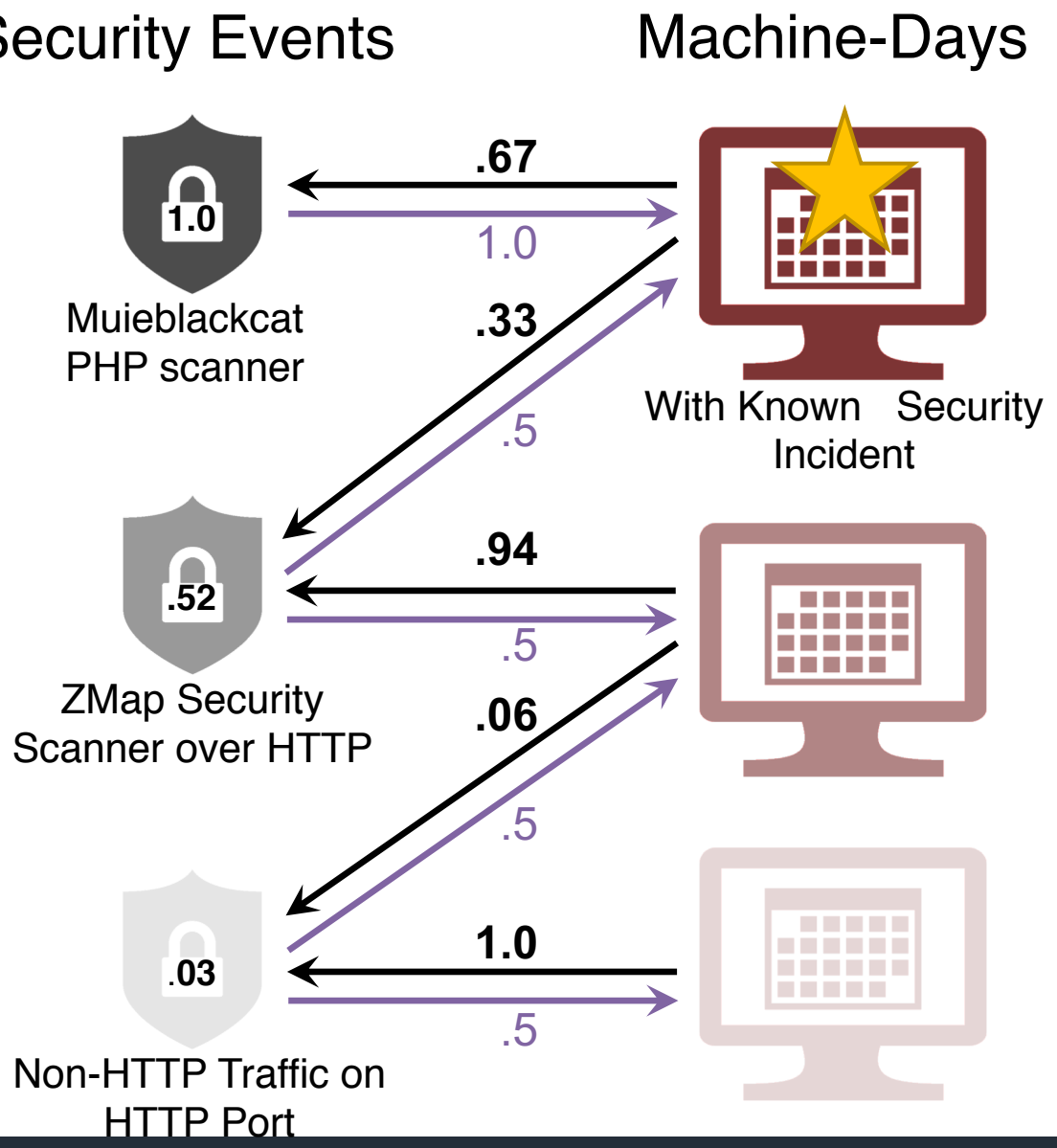
Transmits to an out-neighbor with a probability ∝ to the edge weights

At each step, transports back to node *i* with a small probability *c=.15*

Node *j*'s relevance to node *i* is the probability that the particle will be found at node *j*

RWR as a linear system:

$$\vec{r} = (1 - c)\tilde{W}\vec{r} + c\vec{e}$$



Creepware

“The Many Kinds of Creepware Used for Interpersonal Attacks,” by Kevin A. Roundy, Paula Barmaimon Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, Acar Tamersoy, to appear in IEEE Security & Privacy, May 2020.

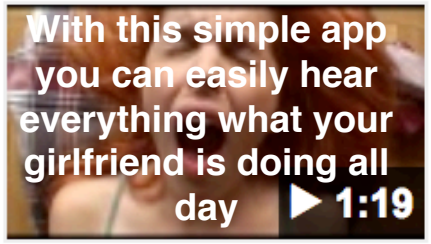
Intimate Partner Violence and Stalkerware

Quick Features List

- Track Phone Location
- Read Phone SMS Remotely
- Read Phone Contacts
- See Call History
- Track Internet Browsing History
- See All Photos Captured
- Record Apps Usage
- Auto Answer
- Back Up & Download CSV
- Free Update
- 100% Undetectable

How To Spy Your Girlfriend's Android Phone

https://www.youtube.com/watch?v=ZL6zWUSaE6Y



May 9, 2016 - Uploaded by Nagy Sándc

Link: https://play.google.com/store/apps

Have you ever ...

mSpy now selling phones pre-loaded with spyware

Cerberus v3.5 adds 'fake shutdown' feature to fool phone thieves

76

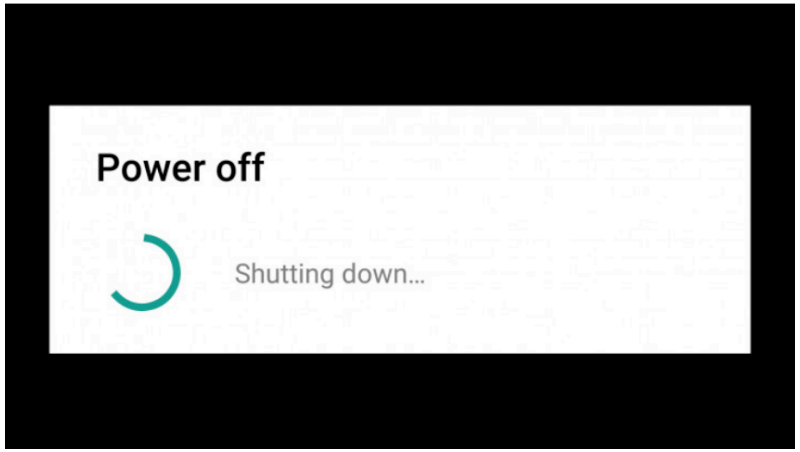
Ryan Whitwam

Aug 24, 2017

G+ 18 | f 237 | 42

Total Shares 297

APPLICATIONS | NEWS



Stalkerware Discovery Through Co-Occurrence

Sequences of installed apps

1. Track a Phone by Number
2. Find My Friends
3. Live Mobile Location Tracker
4. SMS from PC / Tablet & MMS Text Messaging Sync
5. HelloSpy

1. GirlFriend Cell Tracker
2. System Services
3. Hidden Auto Call Recorder
4. Family Locator - GPS Tracking
5. SMS Forwarder

Quick Features List

- Track Phone Location
- Read Phone SMS Remotely
- Read Phone Contacts
- See Call History
- Track Internet Browsing History
- See All Photos Captured
- Record Apps Usage
- Auto Answer
- Back Up & Download CSV
- Free Update
- 100% Undetectable

Searching for Guilt By Association In a Pure Form

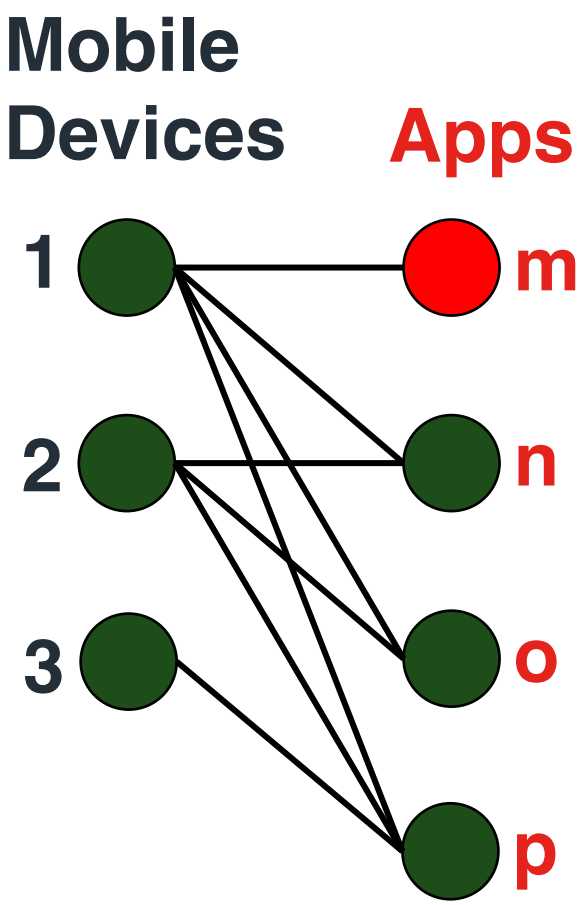
Prior information limited to 18 covert stalkerware apps

No features

Parameterless algorithm

Graph Mining Algorithm Walkthrough

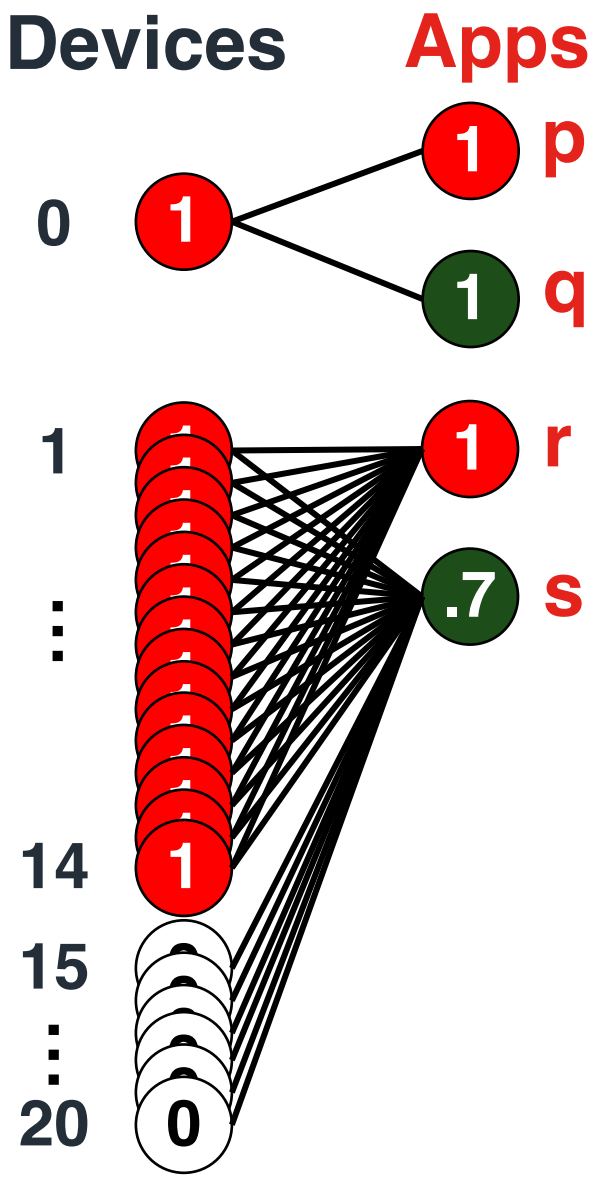
	Device scores			Normalized app scores			
Round	1	2	3	m	n	o	p
0	-	-	-	1	0	0	0
1	1	0	0	1	0.5	0.5	0.333
2	1	0.5	0.333	1	0.474	0.474	0.386
3	1	0.474	0.386	1	0.469	0.469	0.395
4	1	0.469	0.395	1	0.469	0.469	0.396
5	1	0.469	0.396	1	0.468	0.468	0.397
6	1	0.468	0.397	1	0.468	0.468	0.397
7	1	0.468	0.397	1	0.468	0.468	0.397



False Positive Woes

-App_q scores higher than App_s which is intuitively wrong

-The % of apps that are stalkerware related is unknown but very low, many rare apps like App_q co-occur with stalkerware by random chance, leading to high False Positive rates



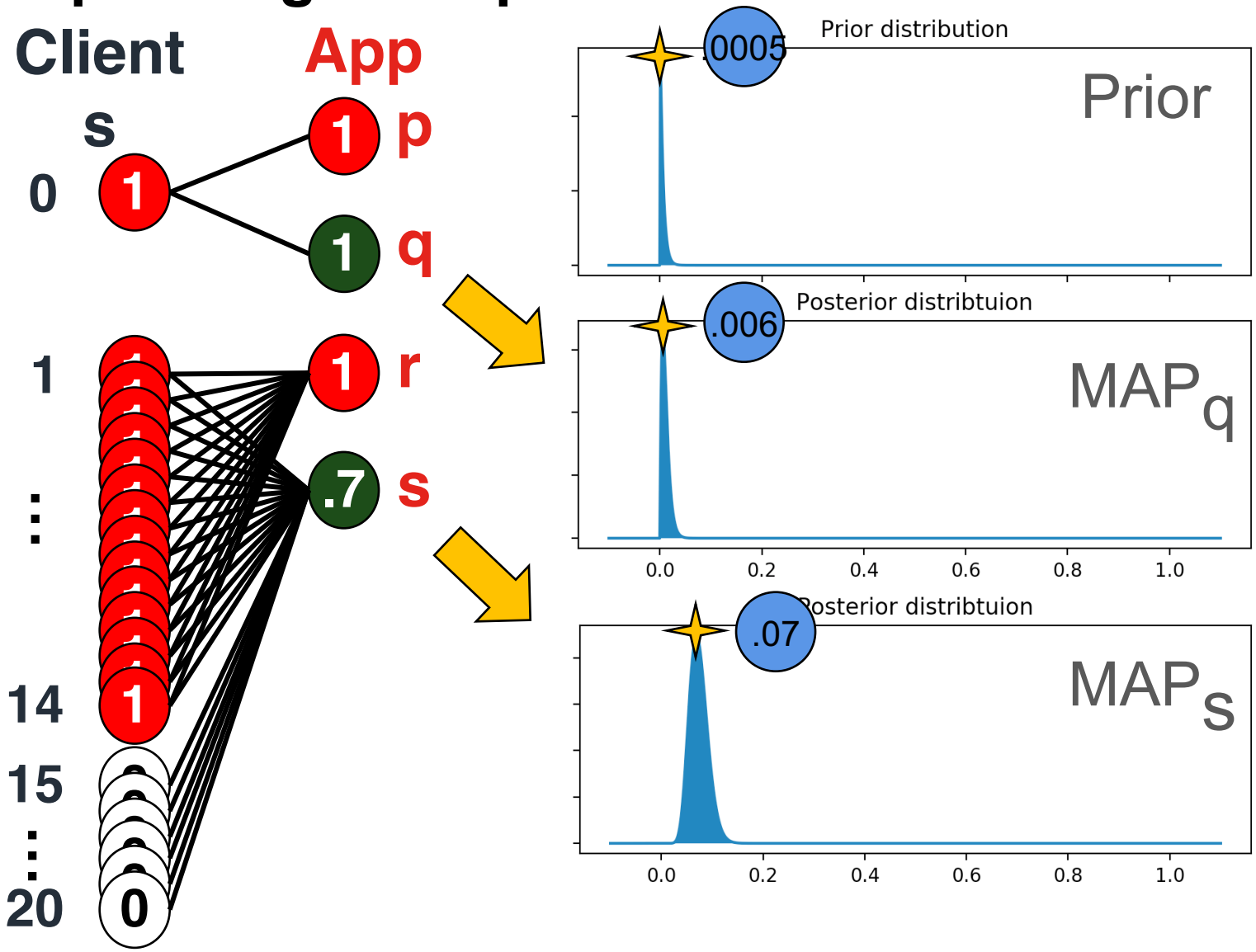
Controlling FP's By Incorporating a Skeptical Prior

If we model θ as a binomial distribution with a Beta prior, we get a closed form solution for the posterior, whose mode is the Maximum a Posteriori estimate.

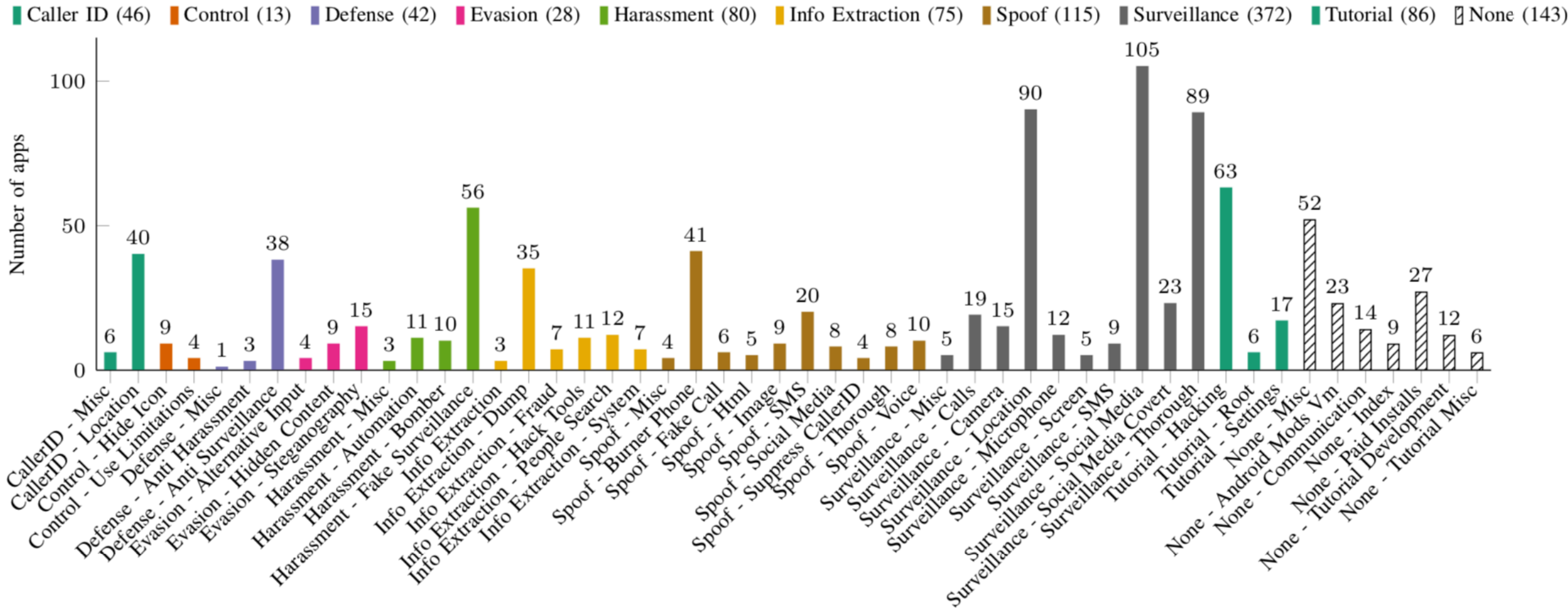
Mode of Beta (1.1,185)

$$= \frac{1.1 + \text{hits} - 1}{1.1 + 185 + \text{trials} - 2}$$

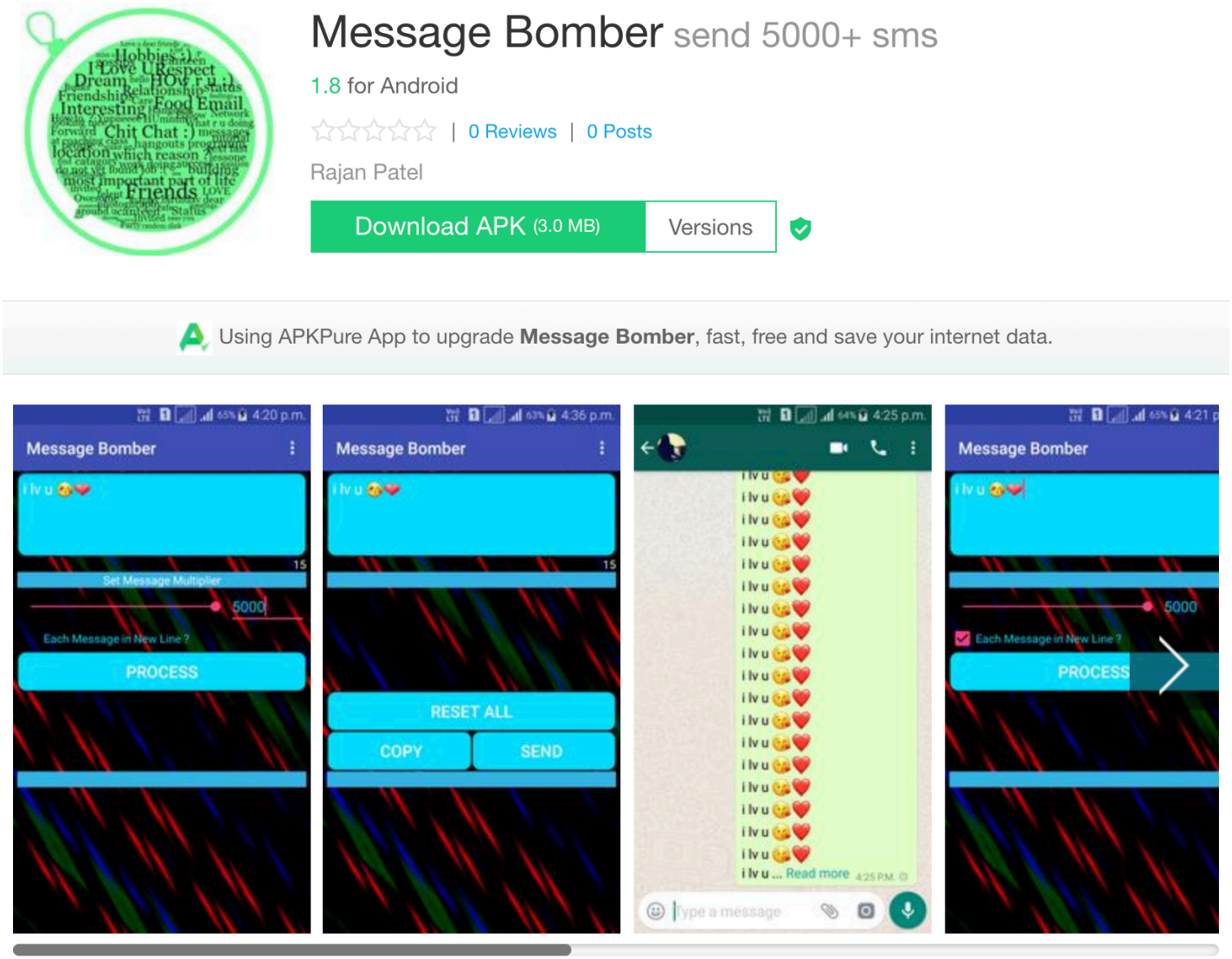
For a derivation see:
https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading14a.pdf



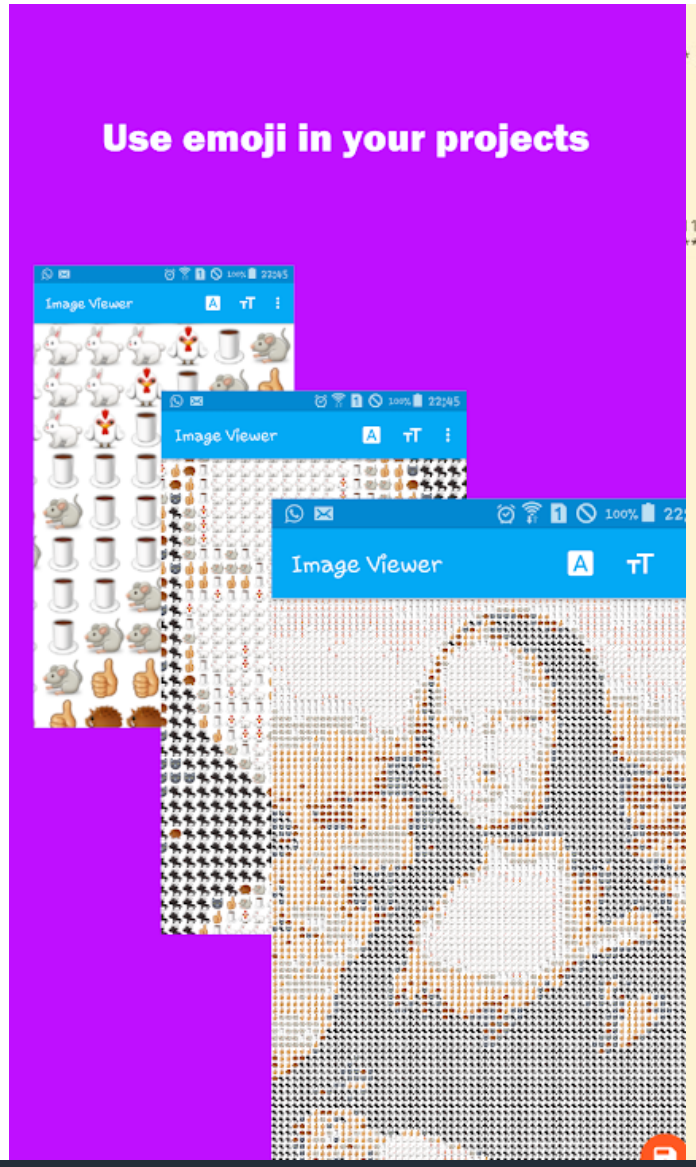
Sub-Category Counts in 2017 (similar distribution in 2018 and 2019)

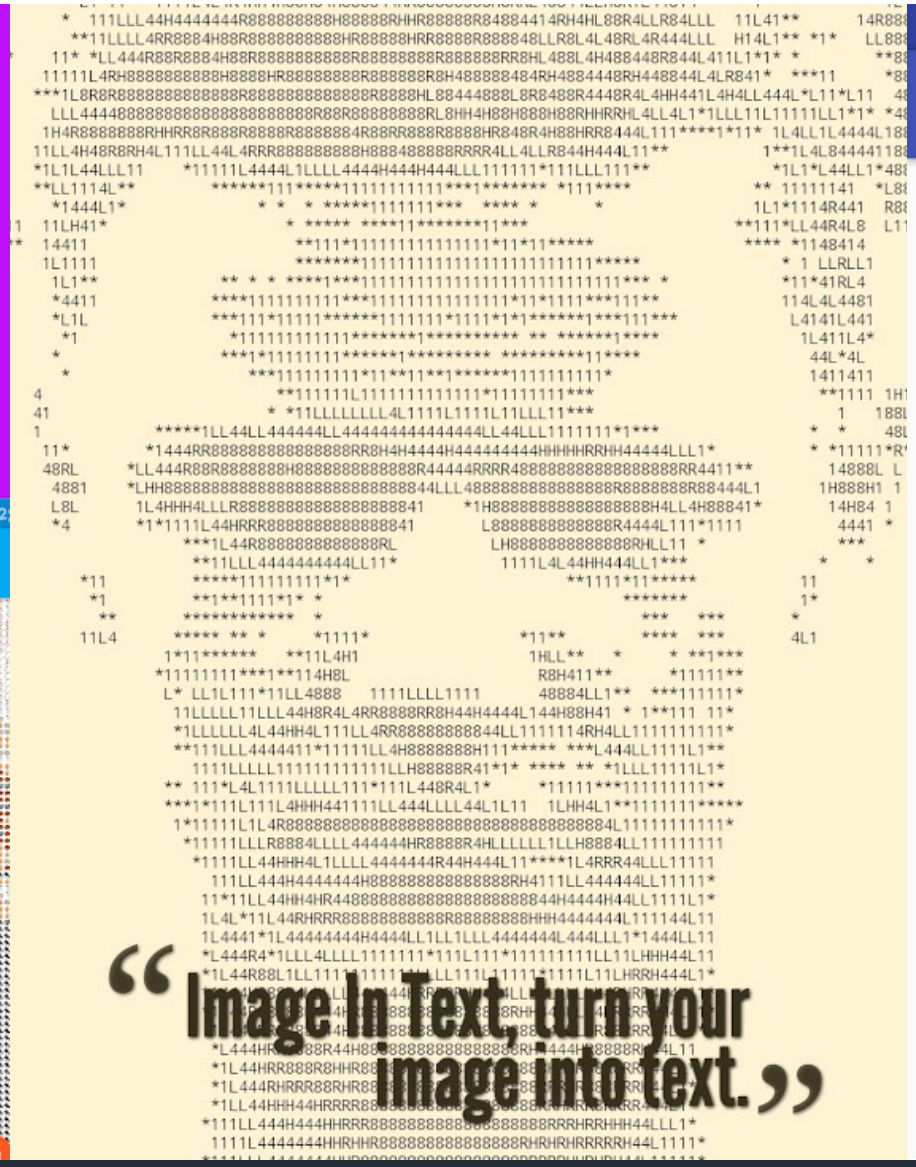


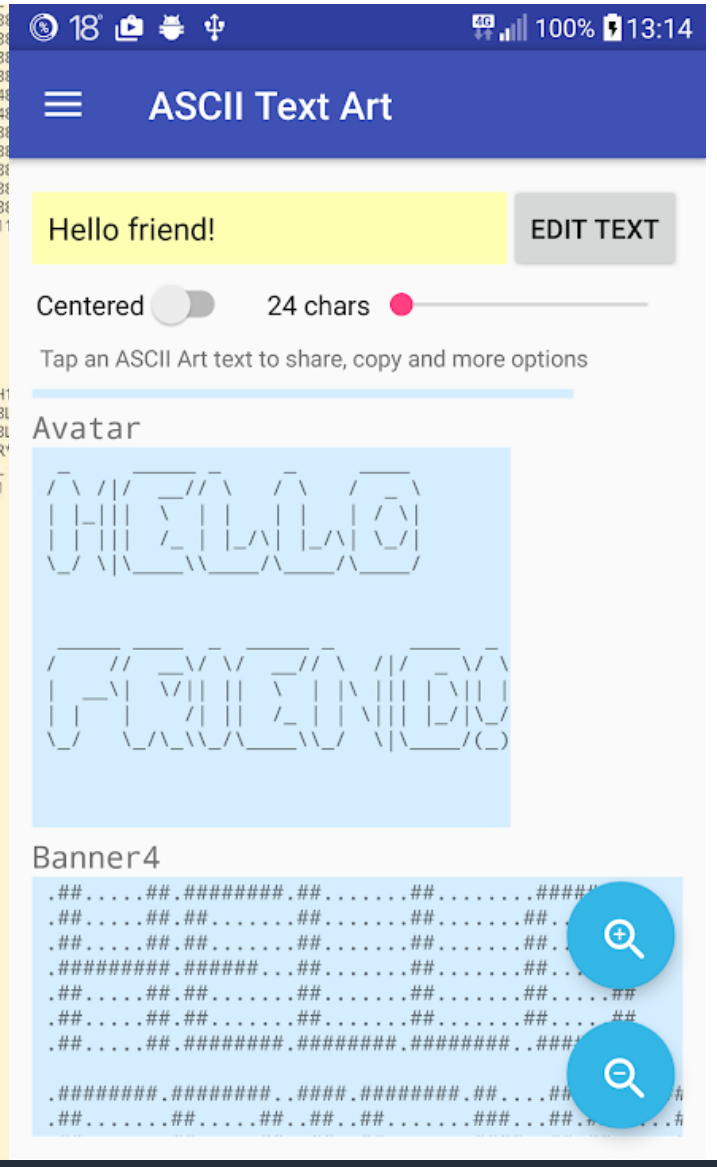
Bombing


Message Bomber send 5000+ sms
1.8 for Android
0 Reviews | 0 Posts
Rajan Patel
Download APK (3.0 MB)
Versions
Using APKPure App to upgrade Message Bomber, fast, free and save your internet data.
Message Bomber
i lv u
Set Message Multiplier
5000
Each Message in New Line ?
PROCESS
RESET ALL
COPY
SEND
Type a message

Bombing Amplification Apps



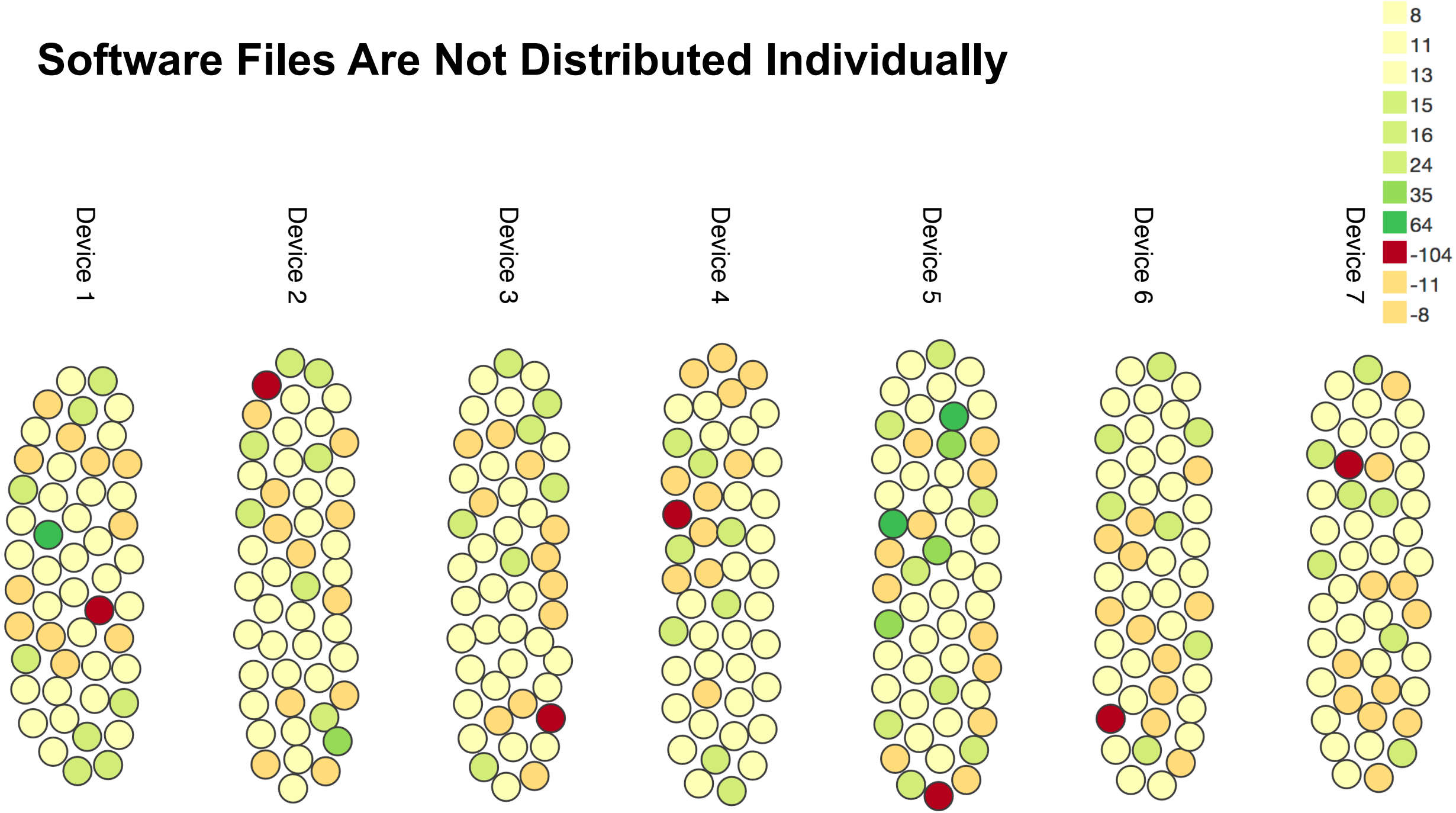




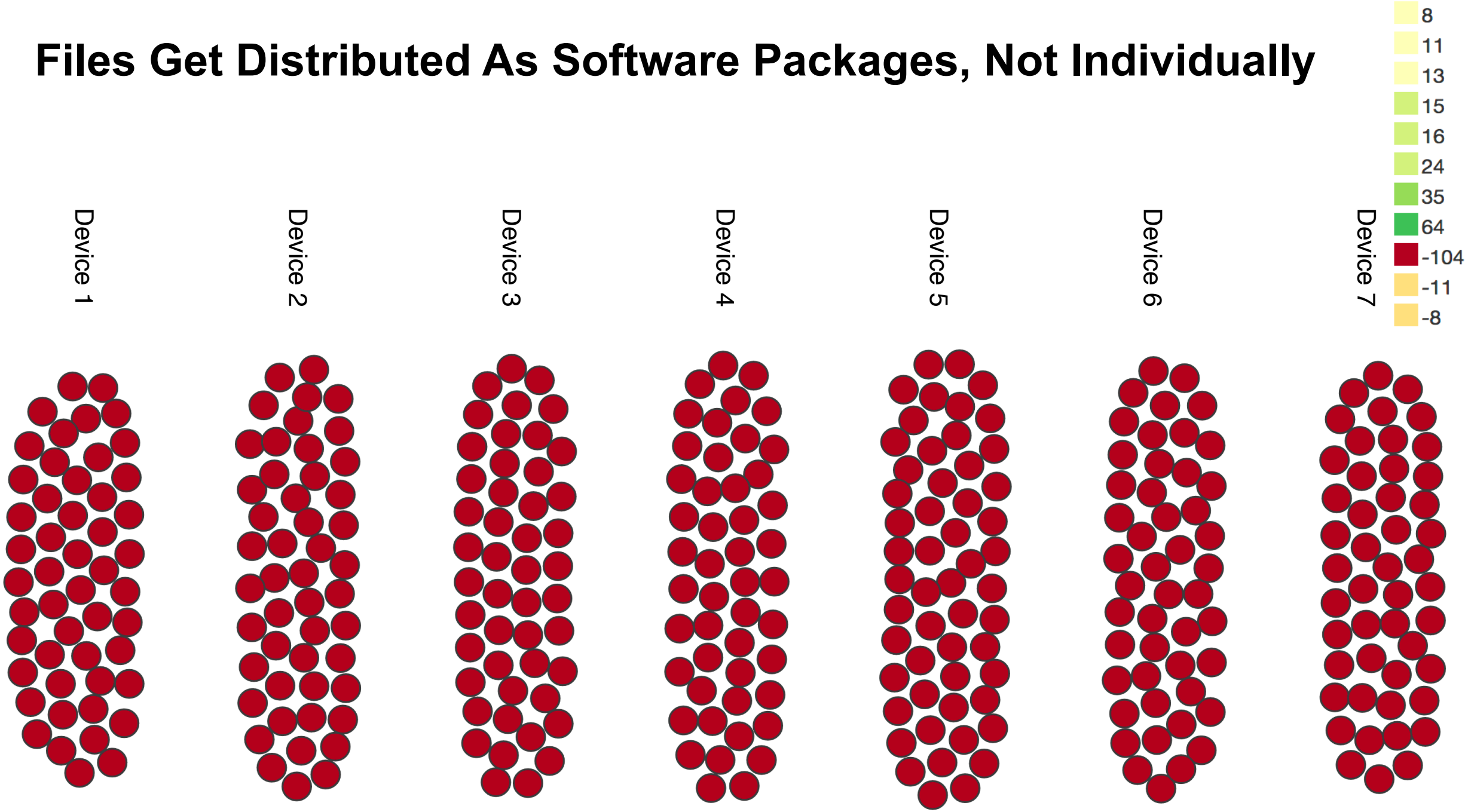
Cliques = Software Packages?

“Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs,” by Acar Tamersoy, Kevin A. Roundy, Duen Horng Chau. Knowledge and Data Mining Conference (KDD), Industry Track, 2011.

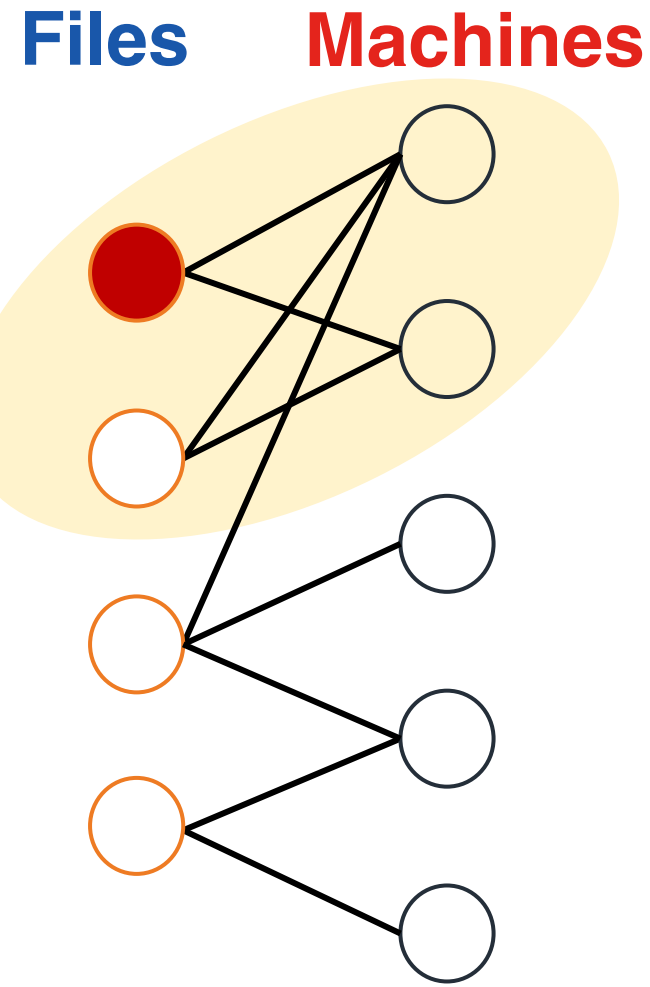
Software Files Are Not Distributed Individually



Files Get Distributed As Software Packages, Not Individually



Aesop: Software Packages Form Cliques in the Install Graph



File	Set of machines containing the file
f_1	$M_{f_1} = \{m_2, m_4, m_5, m_8\}$
f_2	$M_{f_2} = \{m_3, m_5, m_7\}$
f_3	$M_{f_3} = \{m_1, m_3, m_5, m_6, m_7\}$

Comparing File Install Lists at Scale

11 Billion files and 400 million machines

- Popular files occur up to 200,000 times
- Huge numbers of rare files (**10.5 billion files** occur only once)

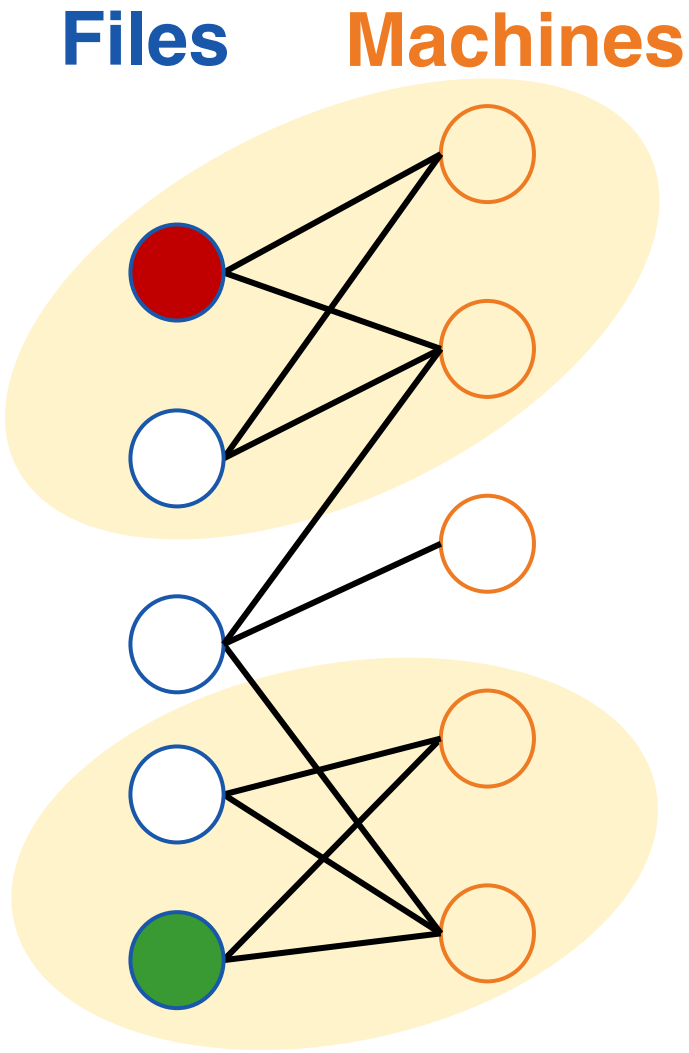
MinHashing

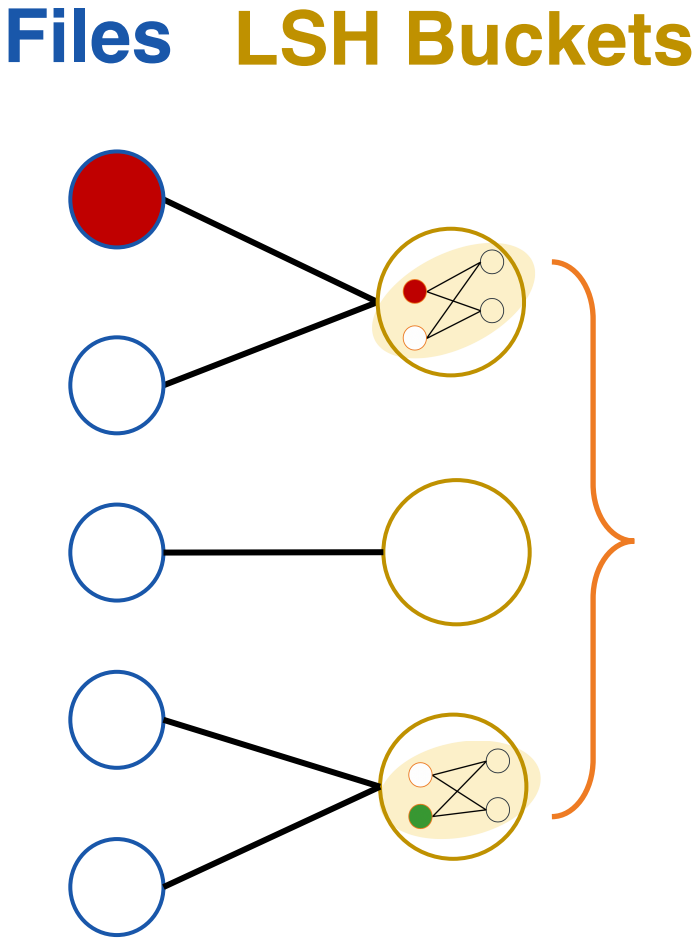
- Map each machine to a random numeric ID with a hash function
- A file's **minhash is** the smallest machine ID in its install list

Locality Sensitive Hashing

- We compute minhash **r** times for each file, and a string composed of the **r** minhashes is the file's hash bucket
- Create **b** bands of **r** minhashes to get more coverage

Graph Transformation Within LSH Buckets





Buckets exhibit homophily, we propagate labels within buckets

Aesop Detections on its First Day…

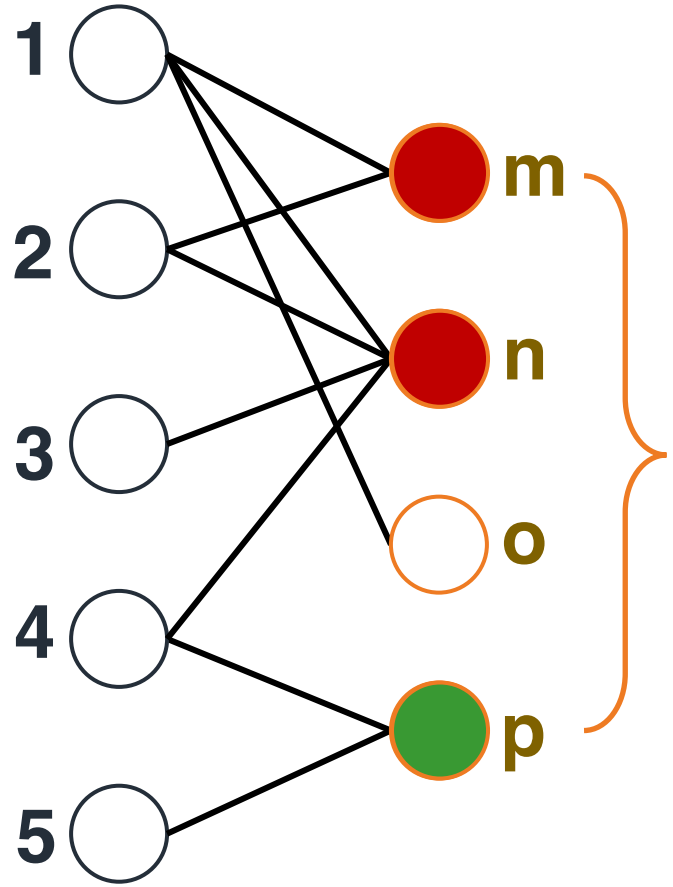
700 million benign files

1 million malware files

Summary

Guilt By Association Unlocks Power in Relationship Data

Mobile Devices
PC's / Laptops
IP Addresses



Apps
Software .exe .dll files
Alerts from AV, IDS, FW, …

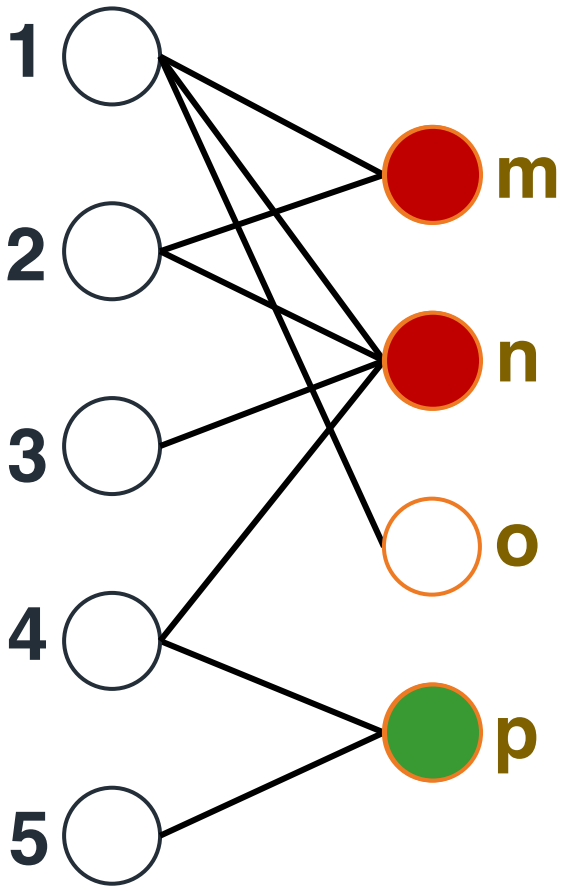
Many Guilt-Based Security Systems @ NortonLifeLock / Symantec

Polonium

"Polonium: Tera-Scale Graph Mining and Inference for Malware Detection," by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.

Aesop

"Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs," by Acar Tamersoy, Kevin A. Roundy, Duen Horng Chau. Knowledge and Data Mining Conference (KDD), Industry Track, 2011.



Creepware

"The Many Kinds of Creepware Used for Interpersonal Attacks," by Kevin A. Roundy, Paula Barmaimon Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, Acar Tamersoy, to appear in IEEE Security & Privacy, May 2020.

Smoke Detector

"Smoke Detector: Cross-Product Intrusion Detection With Weak Indicators," Kevin A. Roundy, Acar Tamersoy, Michael Spertus, Michael Hart, Daniel Kats, Matteo Dell'Amico, Robert Scott. ACSAC 2017.

Adage

"Generating Graph Snapshots from Streaming Edge Data," International World Wide Web Conference (WWW) Poster Paper, 2016.

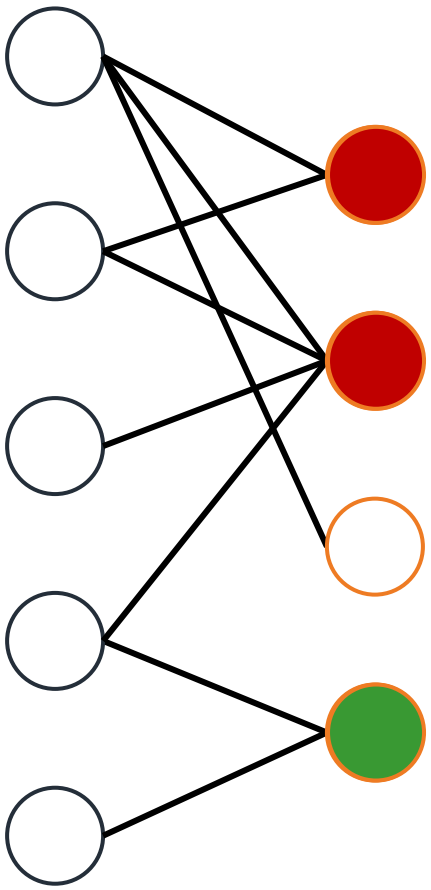
THANK YOU!

Polonium

"Polonium: Tera-Scale Graph Mining and Inference for Malware Detection," by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.

Aesop

"Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs," by Acar Tamersoy, Kevin A. Roundy, Duen Horng Chau. Knowledge and Data Mining Conference (KDD), Industry Track, 2011.



Creepware

"The Many Kinds of Creepware Used for Interpersonal Attacks," by Kevin A. Roundy, Paula Barmaimon Mendelberg, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, Acar Tamersoy, to appear in IEEE Security & Privacy, May 2020.

Smoke Detector

"Smoke Detector: Cross-Product Intrusion Detection With Weak Indicators," Kevin A. Roundy, Acar Tamersoy, Michael Spertus, Michael Hart, Daniel Kats, Matteo Dell'Amico, Robert Scott. ACSAC 2017.

Adage

"Generating Graph Snapshots from Streaming Edge Data," International World Wide Web Conference (WWW) Poster Paper, 2016.