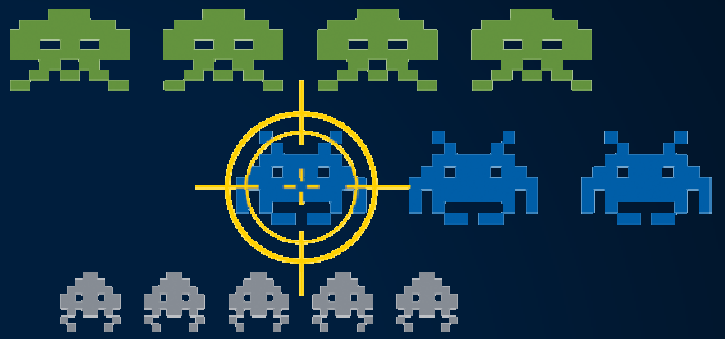


Norton from symantec

Reputation-based Security

Vijay Seshadri
Zulfikar Ramzan
Carey Nachenberg

Agenda – Reputation Based Security

- The Problem
- Reputation Concept
- Implementing Reputation
- Deploying Reputation
- Conclusion

Norton from symantec

The Problem

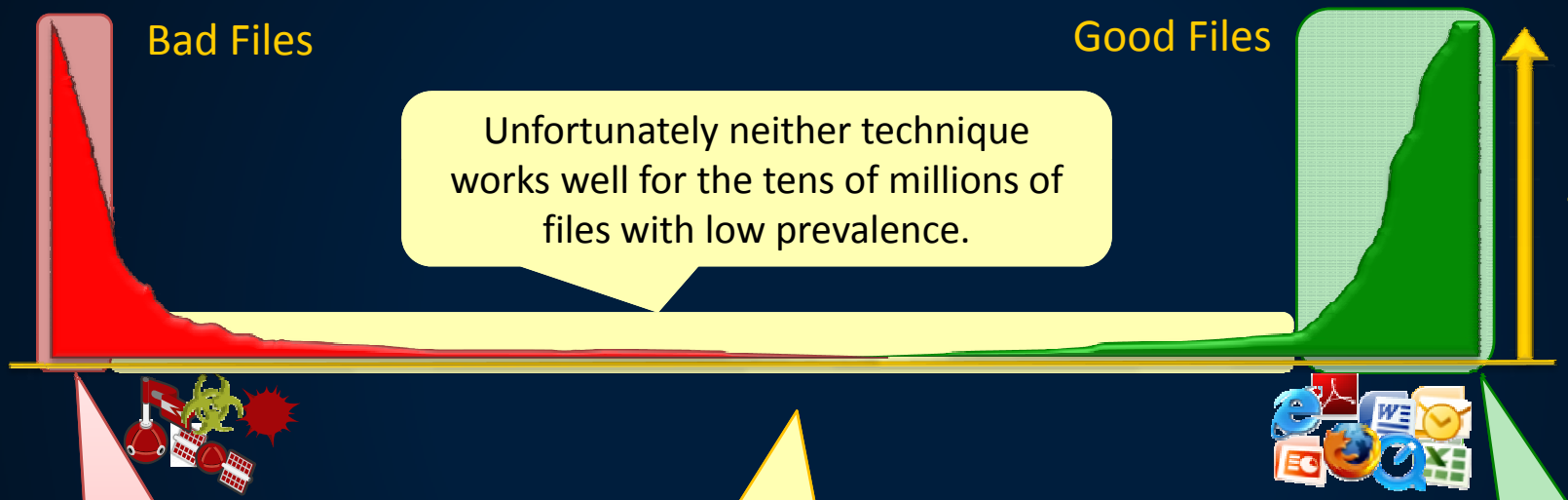
Mismatched odds

- Malware authors have switched
  - from mass distribution of a few threats
  - to micro distribution of millions of distinct threats
- They're targeting each user with a distinct new threat
  - Each has different instructions and a distinct fingerprint
- Stats:
  - 20k-40k new threats discovered per day
  - 120M distinct threats detected by Symantec over last 12 months
  - The average Vundo variant is distributed to 18 Symantec users!
  - The average Harakit variant is distributed to 1.6 Symantec users!
- With such micro distribution, what are the odds a security vendor will discover most of these threats?
- And if the vendor doesn't have a sample, how do they protect the customer?

Norton from symantec

No Existing Protection Addresses the “Long Tail”

Today, both good and bad software obey a long-tail distribution.



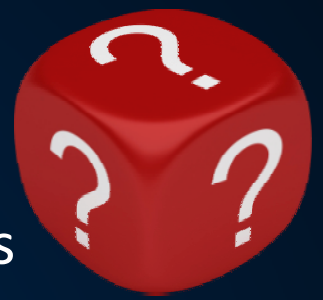
Blacklisting works well here.

For this long tail a new technique is needed.

Whitelisting works well here.

Norton from symantec

Is the “Cloud” the Answer?



- Clouds only contain what the AV industry sees
  - Most threats today exist in very small numbers
  - **Remember that long tail!**
  - What are the odds that a security vendor is going to discover a threat that targets just one or two PCs?
  - Low! And if they know nothing about such a threat, neither will their cloud!
- So “The Cloud” by itself doesn’t add much
  - It’s just the old signature-based model, but very slightly sped up

- *Most cloud based systems still fail to address the long tail*

Norton from symantec

‘Ah-Ha’ Moment (Part 1)

- Three years ago we started thinking…
- There’s got to be a better way!
- Could there possibly be a third option?

- What about using a *reputation-based approach*?

- Companies like ebay, Amazon and Zagat have had great success with this approach!
- Symantec has > 130M active users
- Couldn’t we somehow leverage this massive installed user base to compute file reputations?

Norton from symantec

Reputation-based Security

- **Idea: Use a "reputation-based" approach to "rate" applications**

- **Classic reputation approach (e.g., Amazon.com)**
  - Customers rate items they purchase
  - Over time, products build a reputation
  - "I only buy items with at least 4 stars."

- **One potential reputation approach:**
  - Symantec users rate apps they download
  - Over time, applications build a reputation
  - "Symantec Endpoint Protection only allows users to run programs with at least a 4 stars."

- **Challenge for security vendors:**
  - Unlike Amazon.com we can't ask our users "How do you rate this file?"
  - Most users don't know if software they download is good or bad!
  - So, how do we derive reputation scores for applications without prompting the user?

- **Through our research efforts, we believe we've found a way to…**
  - anonymously harvest useful reputation information
  - about websites and applications
  - from our tens of millions of opt-in users
  - without requiring them to explicitly tell us anything or make any judgments

How Can Reputation Help with Security?

- Traditional AV delivers ratings on only the subset of files that trigger signatures or heuristics
- In contrast, reputation gives us useful data about *every* file known to the reputation system
- And the data is not just black-and-white but nuanced

This file is trending bad

This file is trending good

- This data can be used to drive policy-based lockdown, supplement existing detection algorithms and inform the user

Norton from symantec

Reputation is Complementary to Fingerprint AV

- Traditional fingerprints are totally fooled by polymorphism
  - Attackers can simply change malware logic until existing fingerprints are missed
- Other the other hand, polymorphism sticks out like a sore thumb in a reputation system
  - Most good apps have at least a medium number of users
  - Most good apps have some longevity
  - Polymorphic bad apps have exactly the opposite demographics
- Benefit:
  - If the attackers polymorph their threats, they yield a lower reputation
  - If the attackers don't polymorph their threats:
    - traditional blacklisting works great
    - We can compute a more accurate reputation based on the larger number of users
- *Reputation is derived by a statistical inference engine analyzing data from 35 million+ users, NOT by deriving it just from signatures*.

Norton from symantec

The Impact of Reputation

Traditional Products

100% confidence

Do not scan

Allow

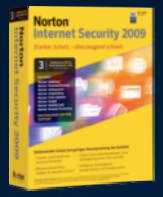
Potentially miss many malicious files

(Behavior blocking and IPS may catch these)

Block

Exact-match signatures and Heuristics

100% confidence



100% confidence

Norton Trusted

Exact-match signatures and Heuristics

100% confidence

Our Next-gen Products

100% confidence

Norton Trusted

Norton Reputation Safe

Norton Reputation Warning

Norton Reputation Malicious

Exact-match signatures and Heuristics

100% confidence

Do not scan

Scan and allow

Warn against

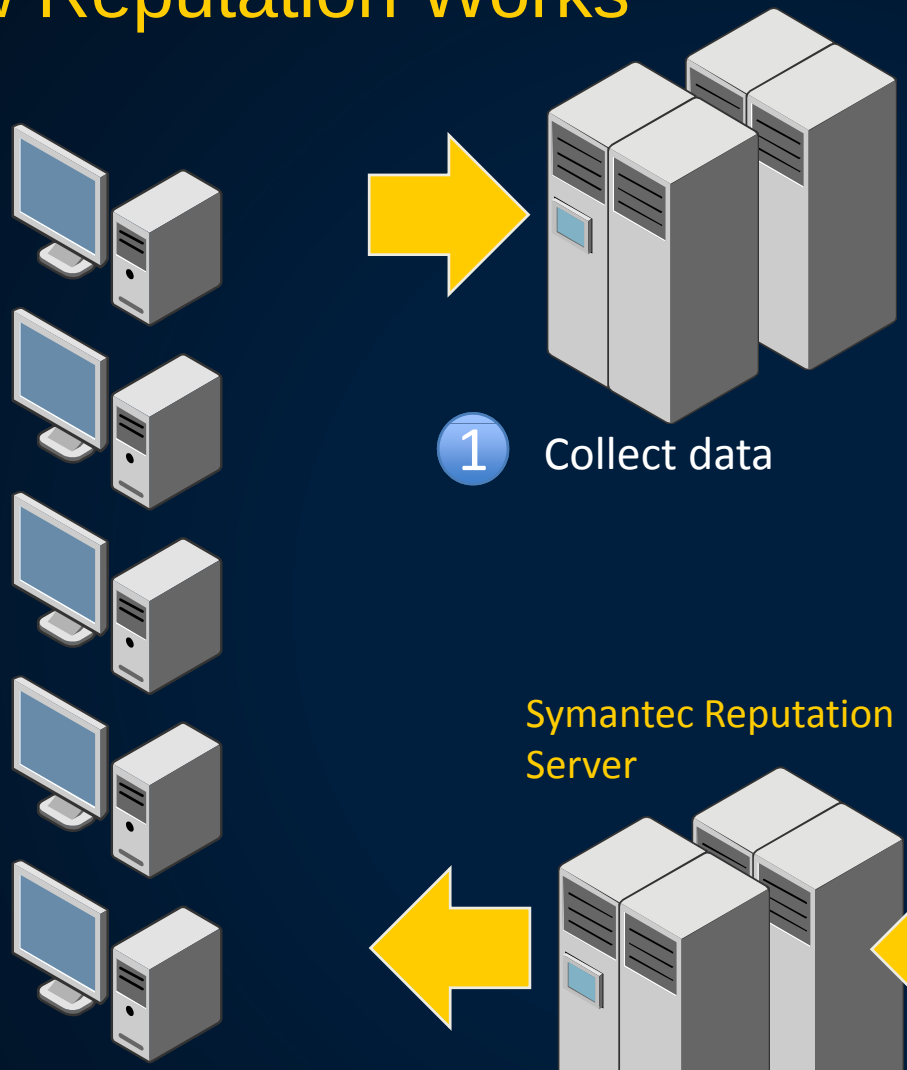
(few users, recently released, not from a major publisher URL)

Block

Norton from symantec

Implementing Reputation

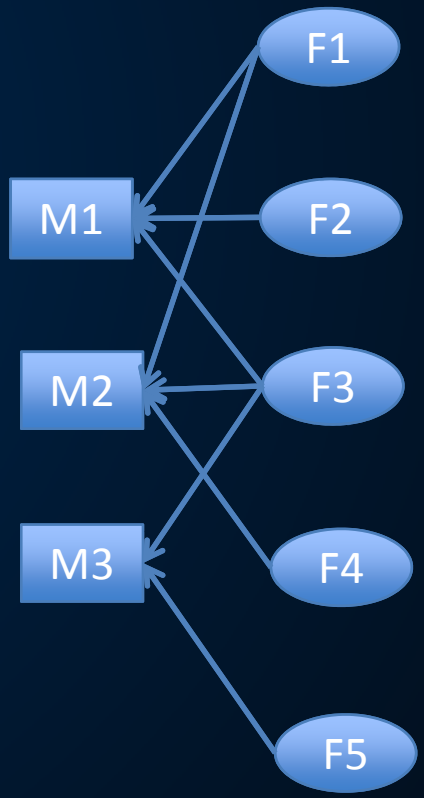
How Reputation Works



Norton from symantec

Calculating Reputation

- Our reputation algorithm is in many ways similar to Google's PageRank algorithm
- Here's a flavor for the approach
  - Inputs: prevalence, age, provenance, demographics of software adopters
    - All files identified by SHA2 hashes for security
    - Reputation computed not based on the contents of each file but rather who's using each file!
  - Scale: 35+M millions submitting users
  - Data structure: huge, sparse bipartite graph of machines and files
  - Outputs: Disposition (good/bad), confidence level, time-to-live

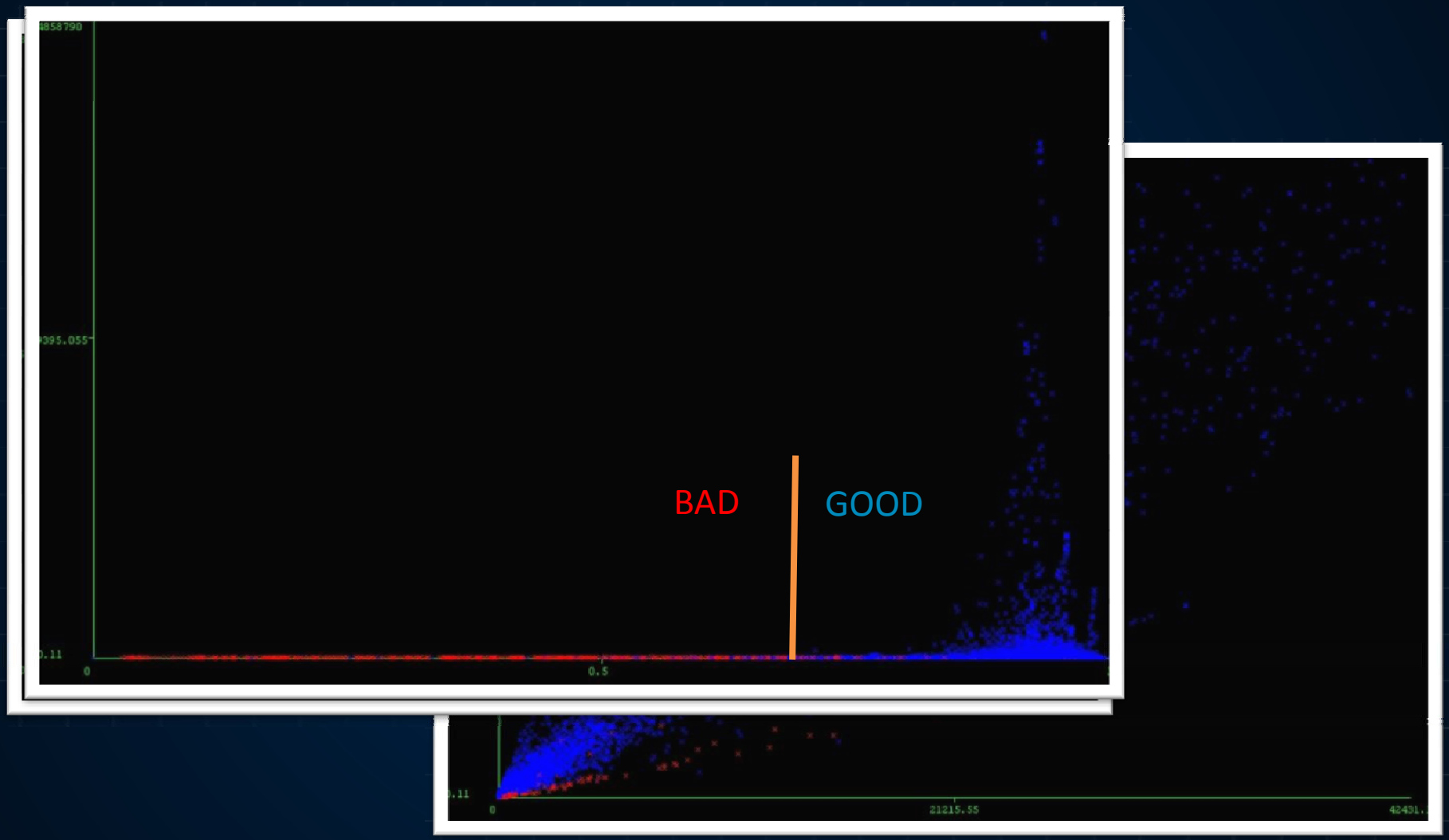


Data Import & Computation

- Challenges:
  - Potentially billions of files across hundreds of millions of users need to be tracked
  - Raw data that we receive from customers on the order of hundreds of gigabytes/day – compressed!
  - Need to perform complex mathematical computations on raw data for the statistical learning engine.
  - Need a Highly scalable and available computation system!
- Solutions:
  - Follow iterative design of the compute and storage clusters to improve mathematical calculations.
    - Optimized computation framework that can work produce/update incremental features for files.
  - We adapted our reputation algorithms to a distributed compute model (map-reduce like system).
    - Break a large bi-partite graph (files and machines) into smaller file hash-based partitions
    - Multiple partitions are mapped to physical nodes
    - We scale the system by adding more nodes

Results

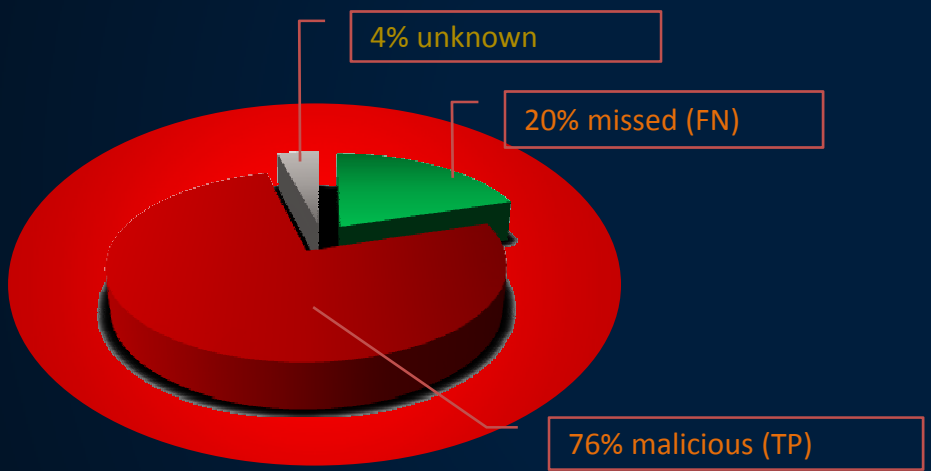
Initial Results



Initial Detection Results

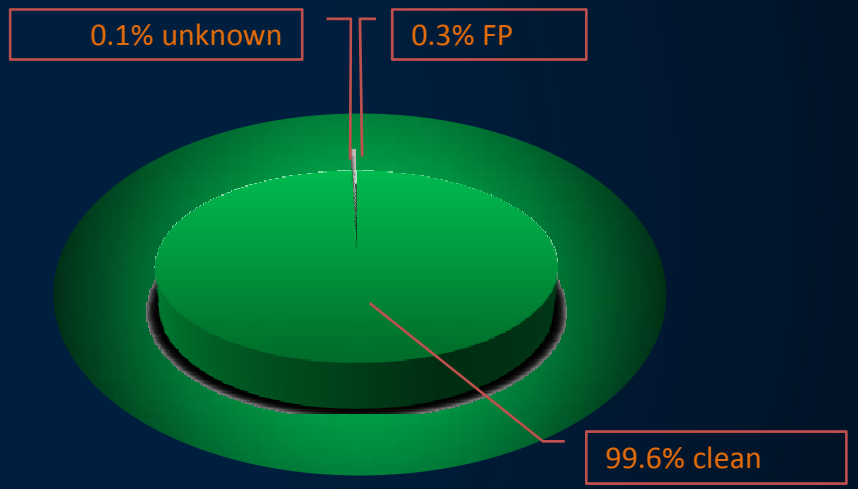
All files are classified as either Good, Bad or Unknown.

Results on Malware



208,291 queries on malicious files

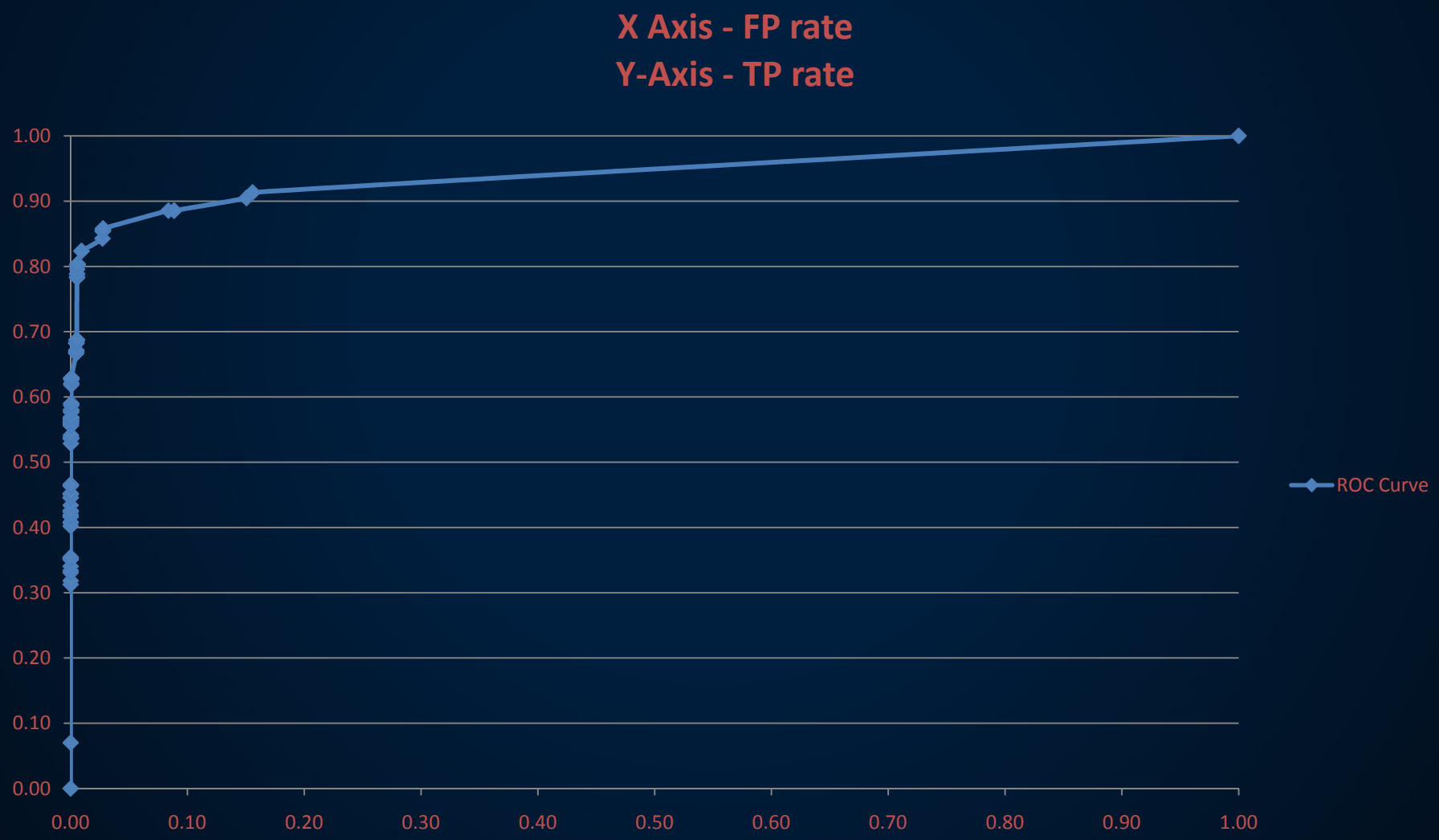
Results on Goodware



52,563,538 queries on clean files

These results are above and beyond any protection provided by classical fingerprinting, heuristics and behavior blocking.

ROC Curve


X Axis - FP rate
Y-Axis - TP rate
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
ROC Curve

Reputation Challenges

- All cloud-based systems are subject to DDoS attacks
  - Reputation-based systems are hosted in the cloud
  - Subject to the same attacks as fingerprint-based clouds
  - DDoS attacks on the cloud – affects non-reputation systems too!
- Gaming is the largest concern for reputation
  - Attacker submits under multiple IDs (Sybil Attacks)
  - Attackers suppress info submitted to us
  - Attackers submit forged info to us
  - Attackers use a botnet to scale their attacks
- Solutions
  - Use cyptography to identify users and prevent invalid submissions
  - Perform back-end analytics to detect strange behavior & inconsistencies
  - Use encryption to ensure secrecy and authentication
  - Statistical machine learning framework tolerates some noise in the data

Deploying Reputation

Reputation
Use Cases

Five distinct use cases:

- **Scan Avoidance**
- **Aggressive Heuristics**
- **False Positive Mitigation**
- **Download Protection**
- **Enterprise Policy-based Lockdown**

Checks Reputation when:

SEP identifies high-reputation software.

Our client can now permanently exclude good programs from further scanning.

Drastically reduces AV overhead (60-90% of actively-used files skipped in NAV '09)

A local heuristic suspects malware.

Client blocks when the file also has a bad reputation. Heuristics and Reputation are orthogonal so we combine them to catch more malware without higher FPs.

A generic signature or behavior blocking detects a new threat.

Reputation data is used to reduce the risk of a highly prevalent False Positive detection.

Block low reputation files automatically and provide context to users for other files

The Client blocks when the file has a bad reputation.

The user is provided with actionable data to help them decide if a download is legit.

The administrator can create custom blocking policies based on their unique risk profile.

The Client filters all downloads according to the administrator's Policy

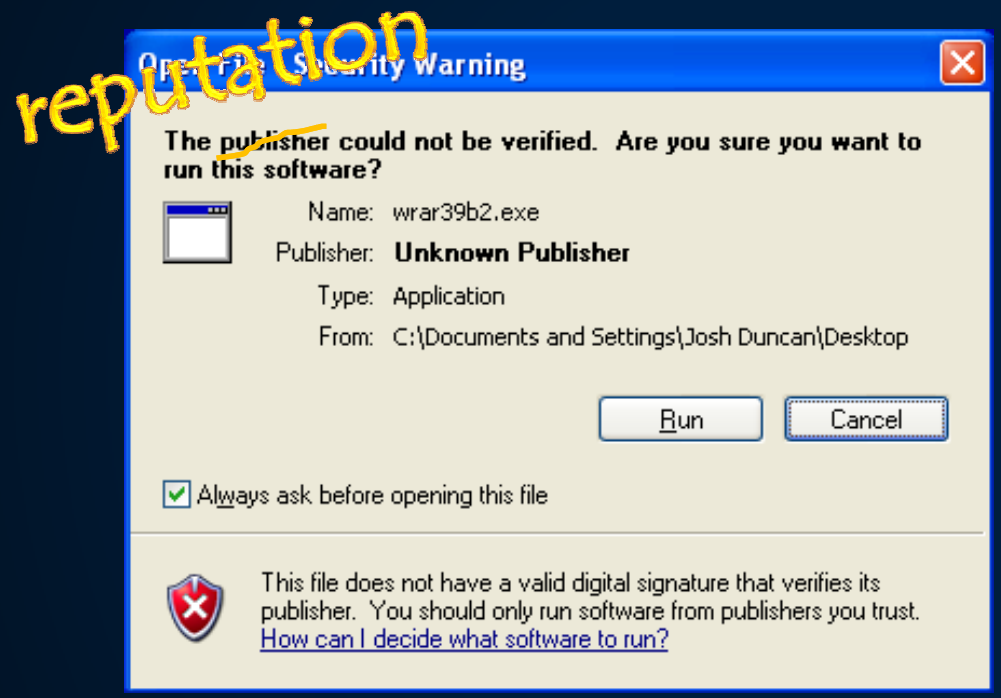
The differentiator is not the cloud itself but the information that the cloud delivers

Reactive Cloud Scanning

Our reputation cloud is also used to deliver our traditional signatures. This helps to address the periods in between traditional in-field signature updates.

How useful is a warning?

We've all seen dialog boxes like this…



that lead to…

…user confusion and frustration

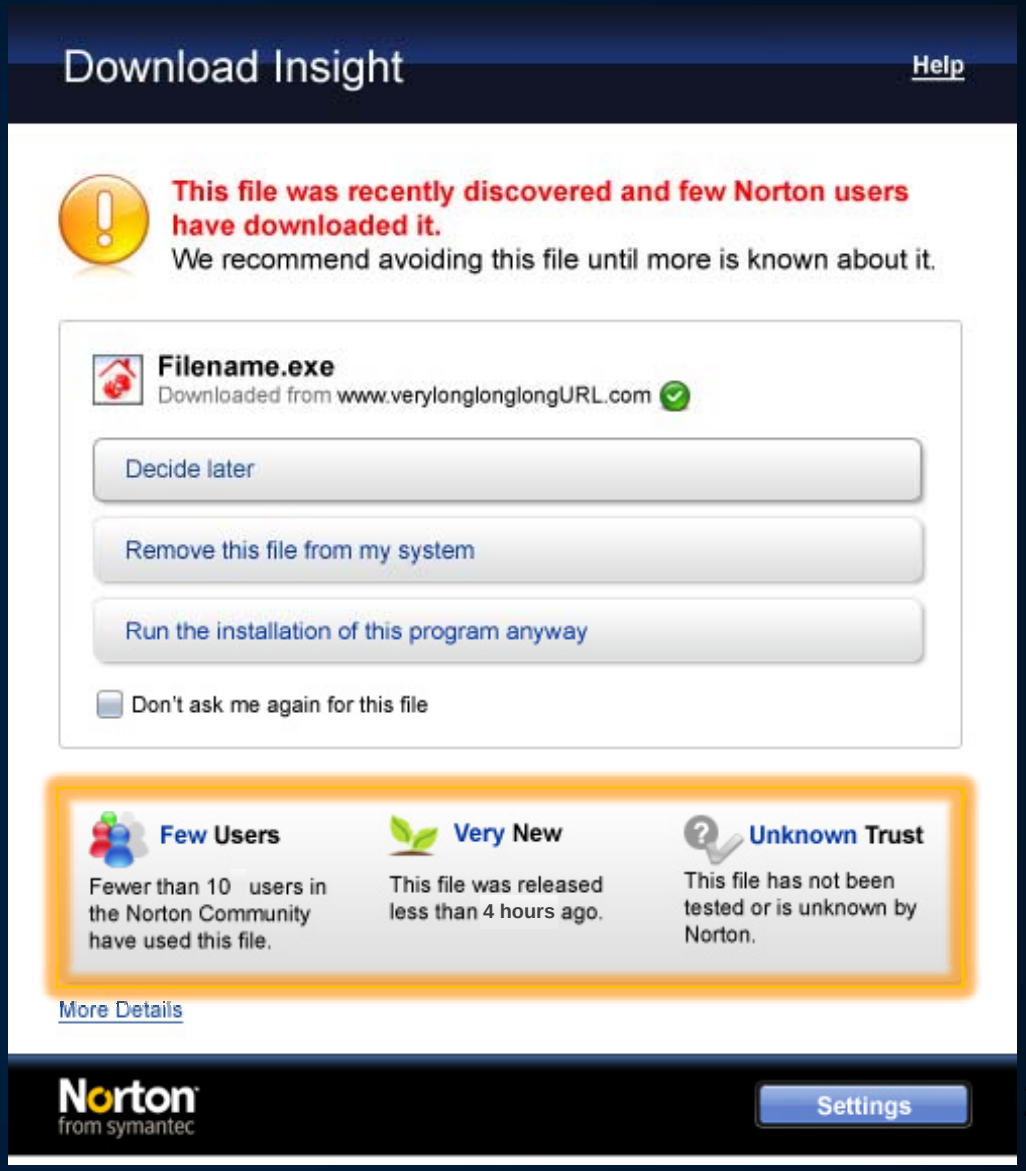
So dialog like this doesn't give a user real information to make an informed decision

Norton from symantec

Symantec Reputation

Actionable data for all files

- When Symantec can't predict a definitive 'good' or 'bad' result, Reputation still provides the user with actionable data
  - How many other Norton users have this file?
  - When was the first time a Norton user saw the file?



Norton from symantec

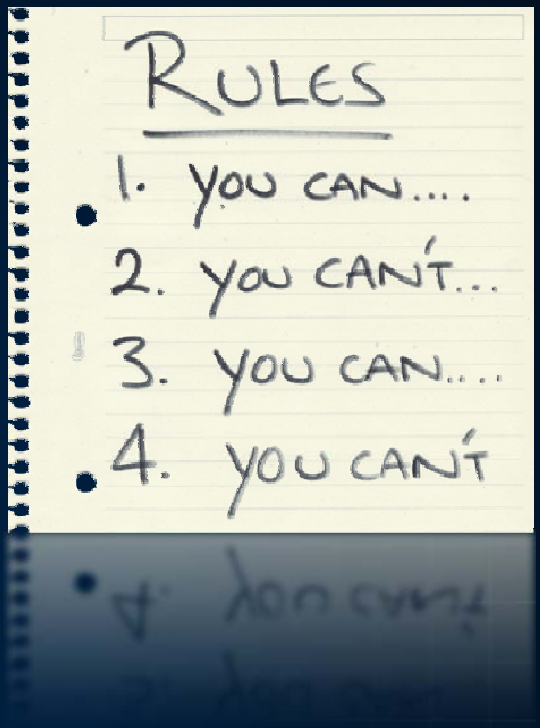
Enterprise Policy-based Lockdown

- You will have the ability to specify your own blocking policy for *all downloaded files*
  - "My finance department can only install high-reputation software used by at least 10,000 other users and available for at least 2 weeks"
  - "Our help-desk employees can install good-reputation software with at least 100 other users."
  - "Our CEO can install anything she likes."
- Given that most malware strains are distributed to less than 20 users, such policies would virtually eliminate traditional malware in the enterprise!

Conclusion

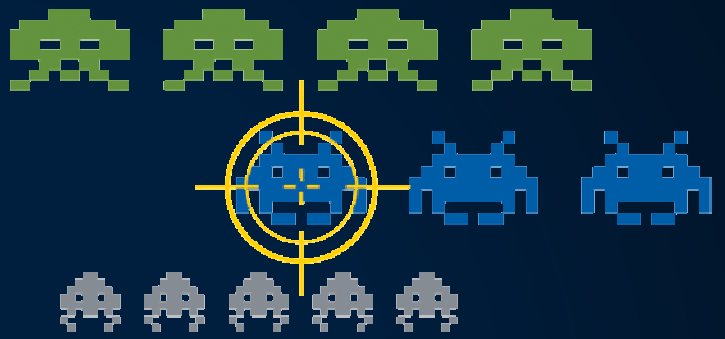
Reputation Changes the Game

- Reputation changes the rules of the malware /anti-malware game
- We no longer need to rely solely on traditional signatures
- We now use metadata from tens of millions of users to identify malware
- Attackers can no longer evade us by tweaking their malware



Norton from symantec

Questions



Norton
from symantec

THANK YOU