Florida Institute of Technology HARRIS

samples.malware.org: Sample Sharing for the Next Decade

Dr. Richard Ford, Thomas Walsh, Dr. William Allen

Conclusion

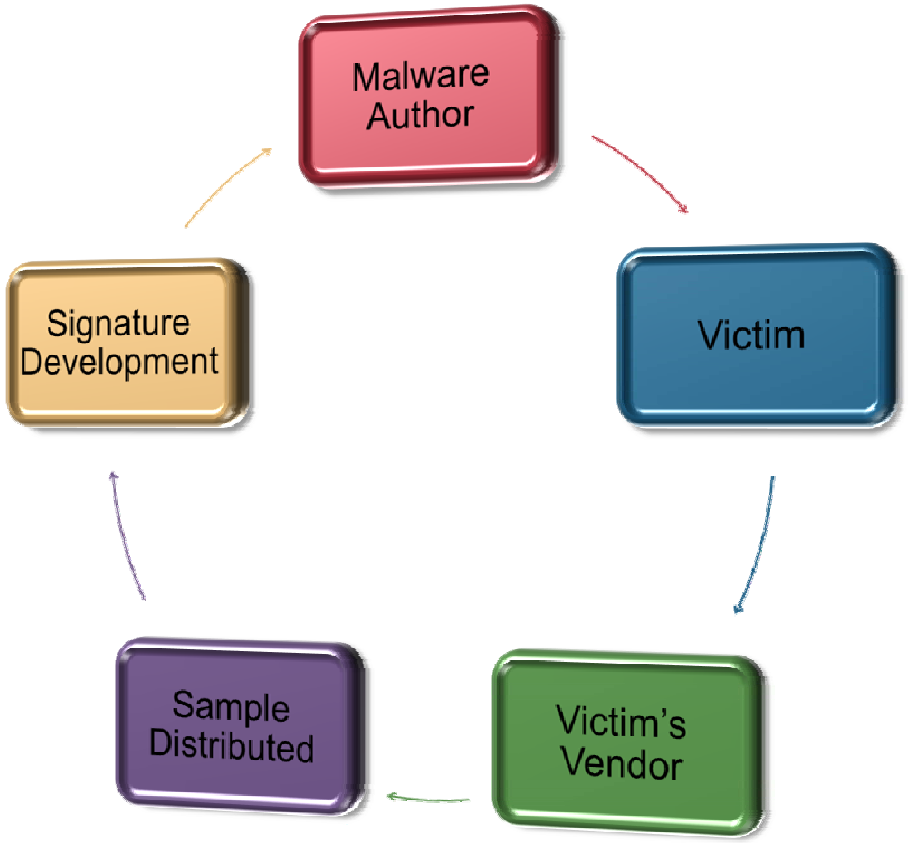
- Current multiscanner services aren't bad… but they also don't make life better
- "If you build it, they will come…" but quality is more important than quantity
- Economy of scale and prioritization of samples outweigh other considerations
- Investing in a system done right is money and time well spent

Getting there from Here

- Where we've been
- Where we are now
- Why that's a problem
- What do we really want?
- How to get it…

Virus Flow…

- Historically, the AV industry has revolved around sample access



Malware Collections Then (1995)

- CD20
  - And this WAS detected by a certain scanner manufacturer
- Size was everything, quality was lagging
- Fortunately…
  - Things slowed down
  - We all started to get bored!
  - Quality of collections began to improve
    - Focus on “in the wild”, VB100 etc.

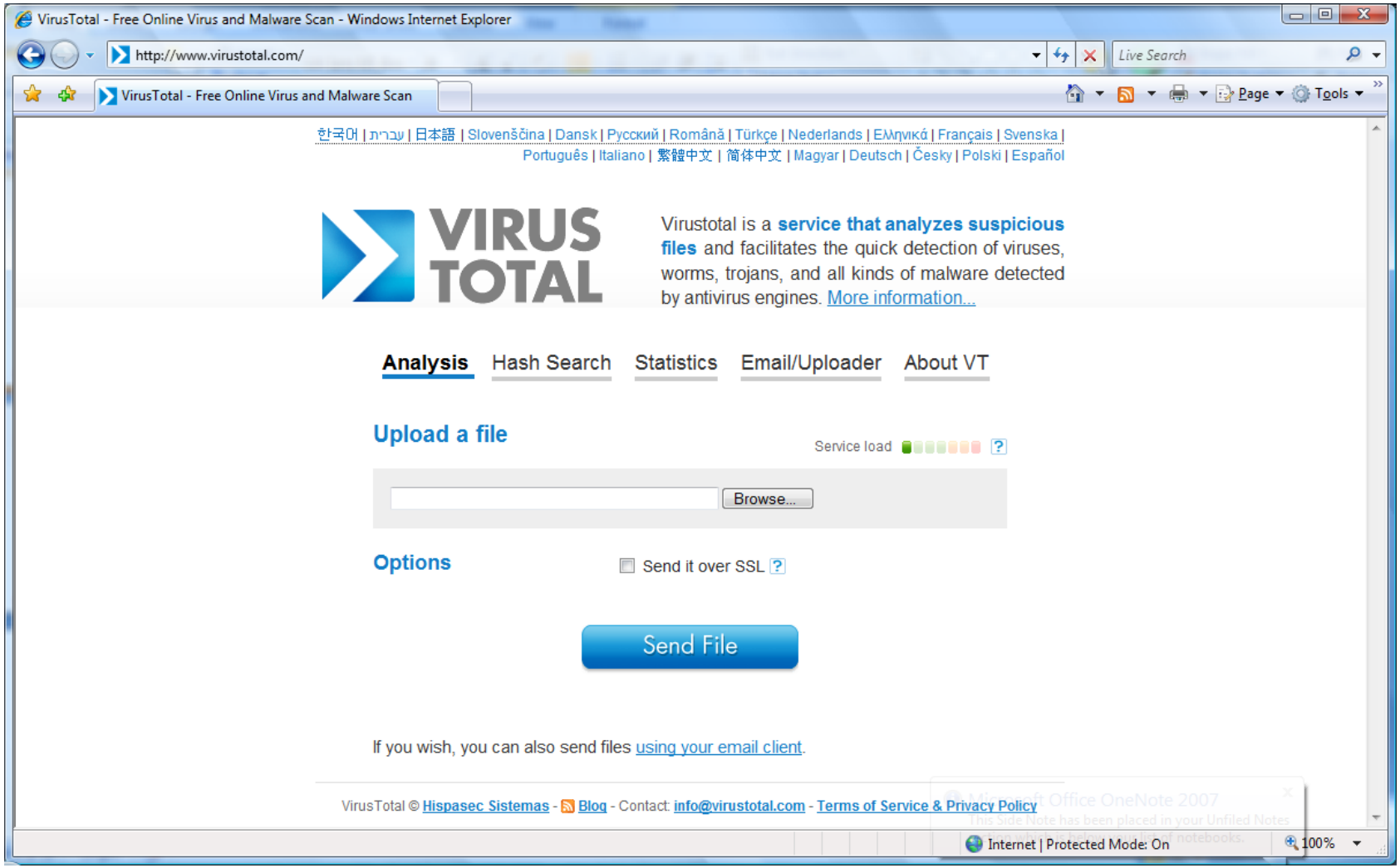
Malware Gets Interesting Again (2002-ish)

- Suddenly, malware became interesting
  - Trojans/Spyware/Adware change everything in terms of sample sharing
  - Fundamental shift in motivation
  - Fundamental change in technology
  - Packers become a huge problem
  - Having the *exact* sample a customer has is important…

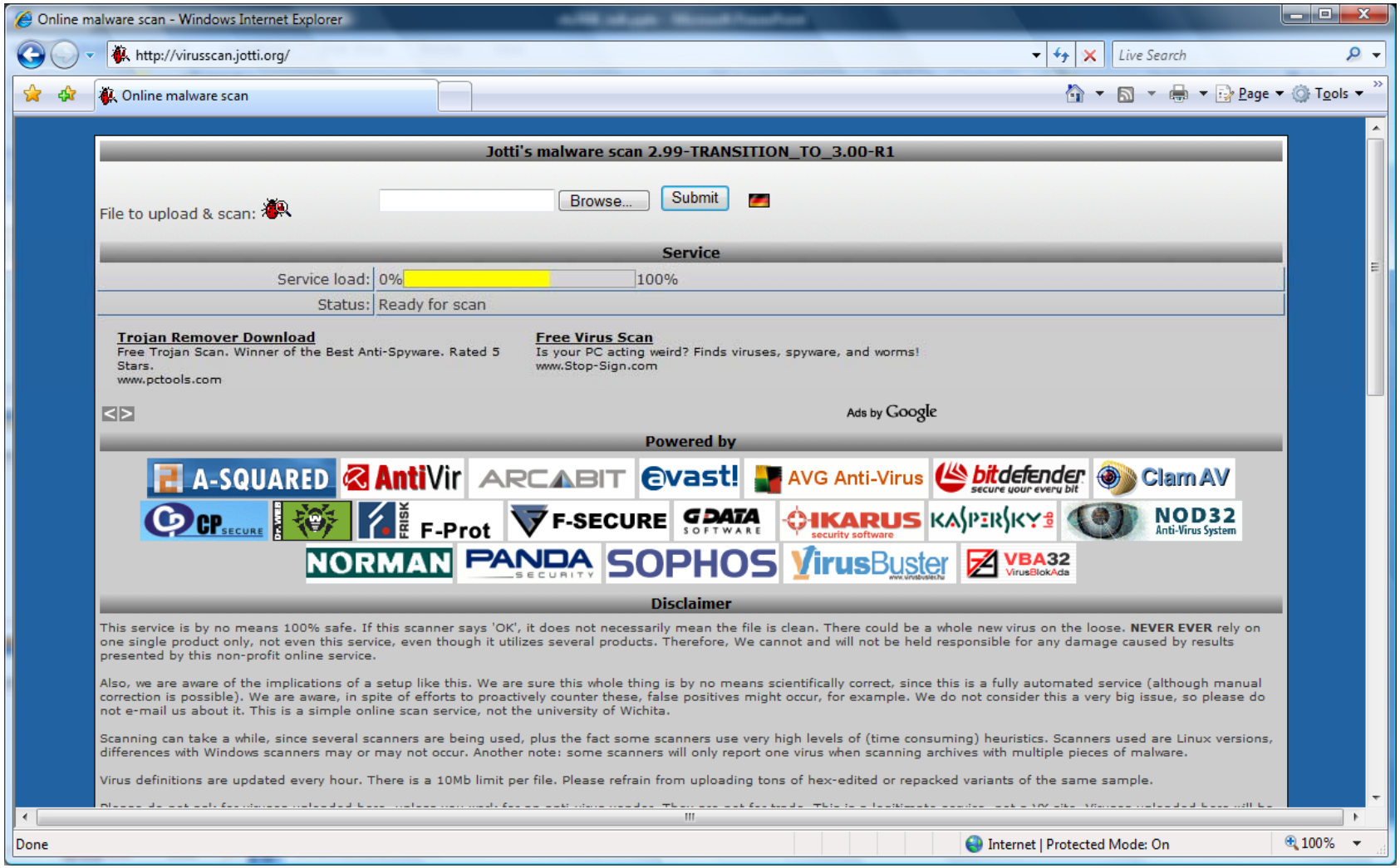
Along Came MultiScanners

- Suddenly, users were getting new malware again
- “Is this file actually malware?”
- Best known:
  - VirusTotal
  - Jotti

VirusTotal


VirusTotal - Free Online Virus and Malware Scan - Windows Internet Explorer
http://www.virustotal.com/
Live Search
VirusTotal - Free Online Virus and Malware Scan
Page
Tools
한국어 | עברית | 日本語 | Slovenščina | Dansk | Русский | Română | Türkçe | Nederlands | Ελληνικά | Français | Svenska | Português | Italiano | 繁體中文 | 简体中文 | Magyar | Deutsch | Česky | Polski | Español
VIRUS TOTAL
Virustotal is a service that analyzes suspicious files and facilitates the quick detection of viruses, worms, trojans, and all kinds of malware detected by antivirus engines. More information...
Analysis
Hash Search
Statistics
Email/Uploader
About VT
Upload a file
Service load
Browse...
Options
Send it over SSL
Send File
If you wish, you can also send files using your email client.
VirusTotal © Hispasec Sistemas - Blog - Contact: info@virustotal.com - Terms of Service & Privacy Policy
Internet | Protected Mode: On
100%

Jotti


Online malware scan - Windows Internet Explorer
http://virusscan.jotti.org/
Live Search
Online malware scan
Page
Tools
Jotti's malware scan 2.99-TRANSITION_TO_3.00-R1
File to upload & scan:
Browse...
Submit
Service
Service load: 0% 100%
Status: Ready for scan
Trojan Remover Download
Free Trojan Scan. Winner of the Best Anti-Spyware. Rated 5 Stars.
www.pctools.com
Free Virus Scan
Is your PC acting weird? Finds viruses, spyware, and worms!
www.Stop-Sign.com
Ads by Google
Powered by
A-SQUARED
AntiVir
ARCABIT
avast!
AVG Anti-Virus
bitdefender
ClamAV
CP SECURE
F-Prot
F-SECURE
G DATA SOFTWARE
IKARUS security software
KASPERSKY lab
NOD32 Anti-Virus System
NORMAN
PANDA SECURITY
SOPHOS
VirusBuster
VBA32
Disclaimer
This service is by no means 100% safe. If this scanner says 'OK', it does not necessarily mean the file is clean. There could be a whole new virus on the loose. NEVER EVER rely on one single product only, not even this service, even though it utilizes several products. Therefore, We cannot and will not be held responsible for any damage caused by results presented by this non-profit online service.
Also, we are aware of the implications of a setup like this. We are sure this whole thing is by no means scientifically correct, since this is a fully automated service (although manual correction is possible). We are aware, in spite of efforts to proactively counter these, false positives might occur, for example. We do not consider this a very big issue, so please do not e-mail us about it. This is a simple online scan service, not the university of Wichita.
Scanning can take a while, since several scanners are being used, plus the fact some scanners use very high levels of (time consuming) heuristics. Scanners used are Linux versions, differences with Windows scanners may or may not occur. Another note: some scanners will only report one virus when scanning archives with multiple pieces of malware.
Virus definitions are updated every hour. There is a 10Mb limit per file. Please refrain from uploading tons of hex-edited or repacked variants of the same sample.
Done
Internet | Protected Mode: On
100%

Problems with Existing Multiscanners

- Opportunity For Abuse
  - Beta testing malware
  - Deliberate false alarms
  - Clean sample glut
- Misleading Results!
- Not really very helpful for the industry

Malware Beta Testing

- Any serious malware developer would want to use a multiscanner
- Submit sample, see who misses, move on to the next
- Unrealistic to think this would stop happening if the multiscanners went away

Deliberate False Alarms

- Triggering an alarm in a competitor's product is going to force lengthy manual examination
  - Essentially, set up a Denial of Service attack on the industry!

Clean Sample Glut

- Wasting time and energy…
  - Take a copy of Notepad
  - Pack it with Themida or your packer of choice
  - Upload it to a multiscanner
  - Result: a sample that probably will end up getting tipped into the analysis queue for vendors *especially* when someone decides to detect it
    - Remember CD20?

What's Bad is Good

- Perhaps the worst problem:
  - *Detection of malware in a file is almost always perceived as a Good Thing™ even if the file turns out to be clean*
  - i.e. A scanner with its heuristics turned on to maximum and lots of false positives will tend to look better on a multiscanner than its more useful competitors
  - We all know this… but many users don't

What do we want?

- Helps the user
  - Tells them what they need to know, not what they ask
- Helps the vendor
  - Gives them what they need to do their job

What users want…

- Information
  - Who detects what…
  - But more isn't more: sample != always bad
- But there's more
  - How who gets what: transparency
    - What versions were used?
    - What options were used?
    - What level of heuristics?
    - *How this changes as a function of time?*
    - *How this varies by geography?*

Vendors…

- Don’t want to have to detect everything
  - CD20
- Don’t think all samples are created equal
  - User A: Kicking the tires, recompiling, experimenting
  - User B: Huge bank (are there any left?)

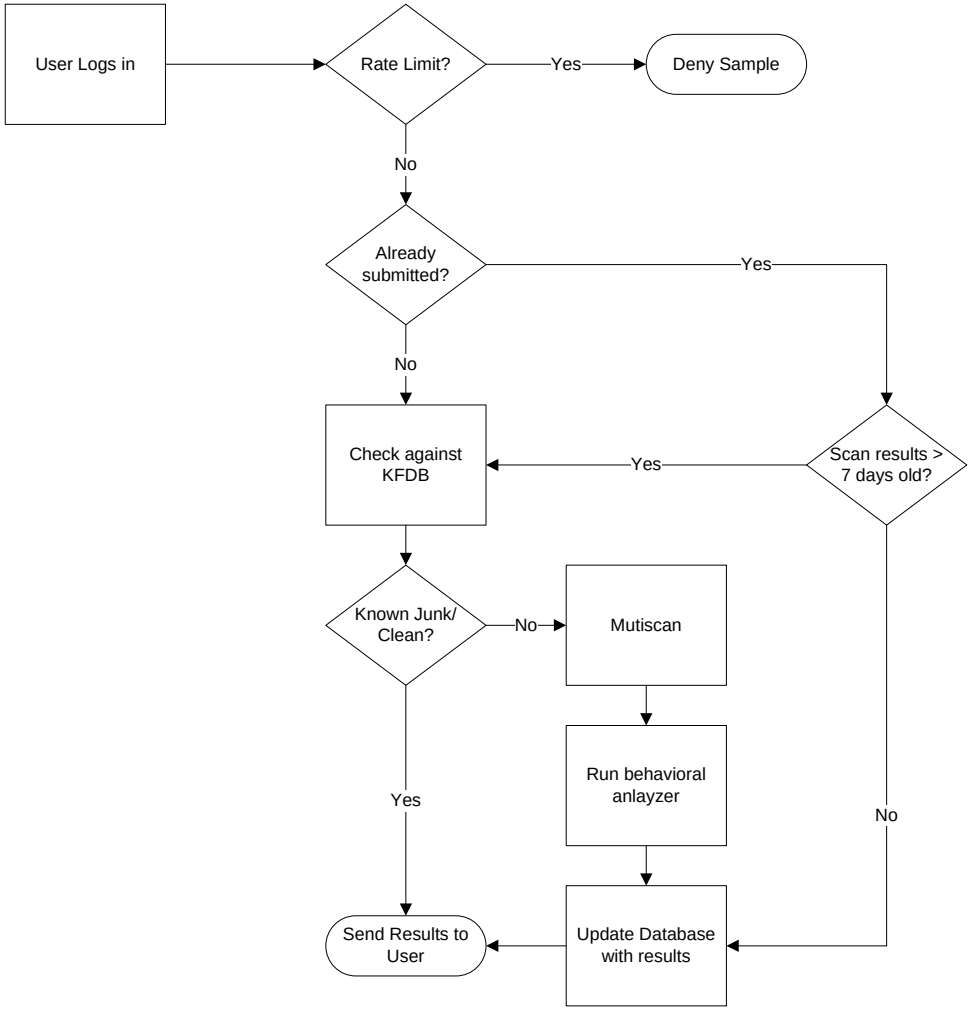
Top-Level Difference

- No anonymous sample submission
  - Users must have a valid email address to submit a sample
  - Users are grouped into tiers based on the anonymity they seek
- Why?
  - Provides the ability to rate limit a user
  - Helps us look at trending from particular users
  - *Dramatically improves sample provenance*

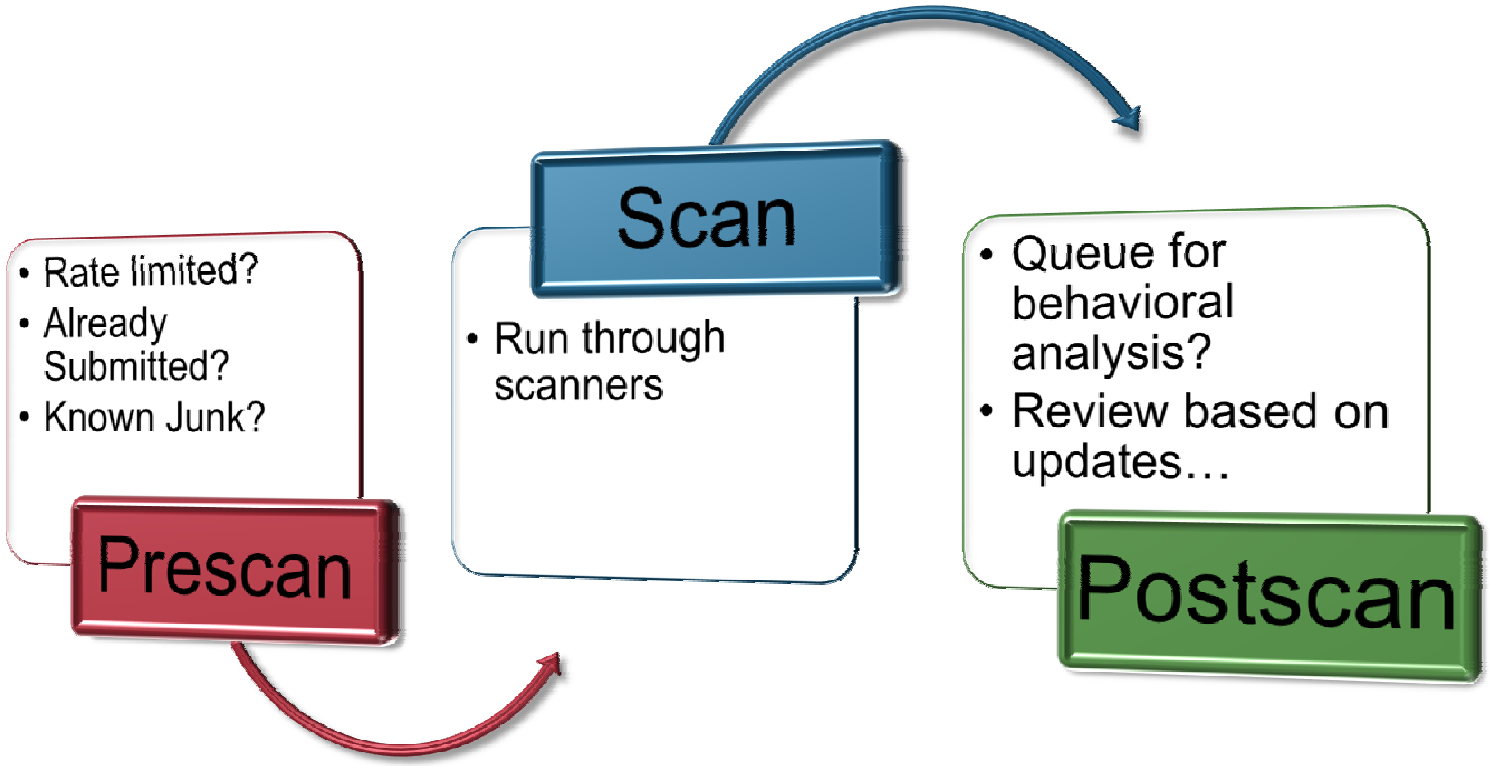
Tiers

- No anonymous sample submission
  - Users complete a CAPTCHA, remember?
  - Doesn't have to be perfect
- User Tiers
  - Tier I: "We know who you are, and you don't mind us sharing that"
    - Advantage to user: priority.
  - Tier II: "We know who you are, but you want us to keep that private"
    - Privacy, less prioritization, may show sector (e.g. business based on email domain)
  - Tier III: "You're a random email address"
    - Typically, least important samples
  - Isn't this pretty manual? Yes, but hopefully that's not an insurmountable problem

How it Works


User Logs in
Rate Limit?
Yes
Deny Sample
No
Already submitted?
Yes
No
Check against KFDB
Yes
Scan results > 7 days old?
Known Junk/ Clean?
No
Mutiscan
Run behavioral anlayzer
Yes
No
Send Results to User
Update Database with results

Before, During, After



When Idle

- Rescan the collection when the system load is predicted to be pretty low
- Update users' scan results and mail out these updates if requested
  - Better service to user – more utility
  - Interesting data in terms of update information in its own right

Challenges?

- Technically, very few
- The devil is in execution
- Driving traffic
- Attracting vendors
- What to add, how to setup?

Marketing

- If you build is *some* will come
- However, we can do a lot to get the word out
  - Corporate customers
  - Conferences
  - Blogs
  - Selfishness is a good motivat
  - Leverage sites like theinternetprotectors.com

- And the “junk” samples aren’t as interesting anyway

Does it Matter?

- With the number of samples screaming upward does it matter?
- Yes!
  - Users want it, and let's face it, the anti-malware industry needs a *lot* of PR help
  - We have 17,000,000 samples (or however many it is today), but how many of these are really different?
    - Different MD5 isn't a fundamentally different sample
  - Aside: the way we're counting is probably really misleading!

A Better Multiscanner Opens Doors

- Interesting opportunity for longitudinal analysis
  - Using VMs lets us archive the state of scanners over a long period of time “as installed”
  - Using pseudonymity lets us “classify” Tier III users automatically
    - “These 17 users are really the same guy”
    - Lots of opportunity for data mining

Samples Advisory Board

- Create more interaction between vendors, power users, and us
- Vendor list: implementation issues
- Advisory board: issues that we need to work through that are not always technical…
- Feedback is *critical*
  - Nobody has all the right answers

Enhancements

- First step is to get to equivalence
- Next, there are lots of more sophisticated things to do
  - “On the metal” analysis/Sandboxing
  - Active triage based on load/source/heuristics etc.
    - Filter as analysis techniques get more expensive
  - XML sample descriptions based on industry input
  - Back-end sftp mechanism for bulk upload/download for vendors
  - Move toward more sophisticated tests and analysis under “real world” conditions *if load allows*

Future?

- Add one new scanner every week to the samples system
  - When we get to 10 scanners, go live
- Set up an advisory board as soon as there is sufficient interest
- Raise minimal money to cover students and hardware
- And we're off to the races

Questions?