Lawrence Livermore National Laboratory

Mining Proxy Logs: Finding Needles In Haystacks

2010-05-19

Matthew Myrick (myrick3@llnl.gov)

Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, CA 94551

This work performed under the auspices of the U.S. Department of Energy by
Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

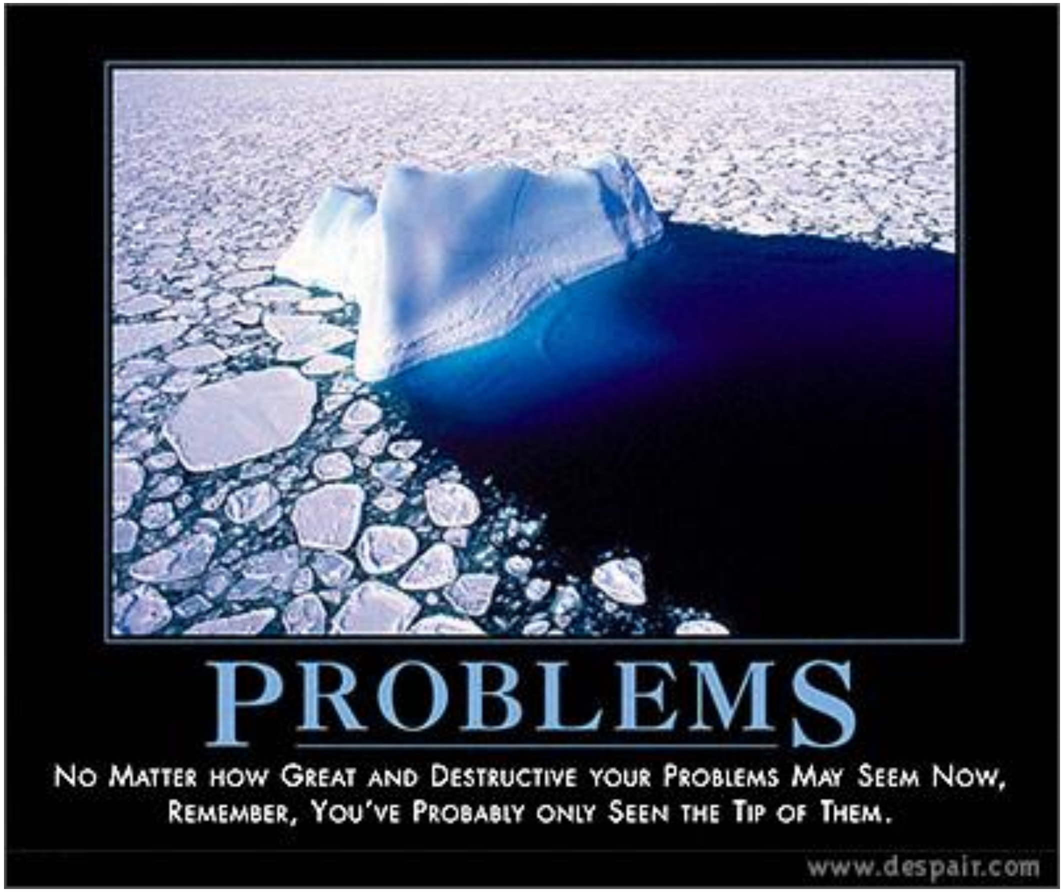
Disclaimer

- Our security infrastructure is a work in progress
  - This presentation is for educational purposes
  - This discussion pertains to our “Unclassified” environment ONLY
  - Hopefully we can make things better by learning from each other
    - If you see problems please say something

Have Fun!

Overview

- Introduction
- Problems
- Landscape
- Solutions
- Future
- Conclusion


PROBLEMS
No Matter how Great and Destructive your Problems May Seem Now, Remember, You've Probably only Seen the Tip of Them.
www.despair.com

Introduction – About LLNL/My Team/Me

- LLNL – Livermore, CA (1 sq mile and 13 sq miles)
  - ~7000 Employees (including contractors)
  - ~20,000 computers
    - ~10,000 Access Internet
- My Team
  - Network Security Team (NST) - 8 people
    - Incident Management Team (IMT) - 4 people
  - IDS, IPS, Proxys, Firewalls, IR, Log Aggregation/Correlation, Pen Testing, Malware Analysis, Forensics, etc.
- Me
  - B.S./M.S. in Computer Science from CSU, Chico
  - Over 13 years w/ LLNL ~6 years full time
  - Currently hold a CISSP, BCCPA, GCIH, GPEN

Problems – What Are We Trying To Solve?

- How do we find “bad guys” on our networks???
  - There are a lot of users
  - There are a lot of computers
  - There is a lot of data
  - There is no consistency and centralized governance is lacking
- What do the “bad guys” look like???
  - I’ve never spoken to a “bad guy”
  - I’ve never met a “bad guy” in person
  - “Bad guys” means something different to different people
- Most of us now have a web proxy…now what???
  - It never works perfectly
  - Somebody is always blaming me for breaking their app ☹

Landscape - LLNL Proxy Deployment

Cyber Security Program

- Blue Coat Proxy SG's
  - Transparent forwarding deployment using WCCP
  - We proxy ALL egress traffic (4 ports)
    - Excluding mail, dns and things explicitly exempted
  - Protocols or enforced on their respective ports
  - Content filtering (BCWF)
  - A/V scanning (McAfee)
  - Internet authentication (ldaps)
  - By and large most data flows through our proxy!

Landscape - Log Format Details

- Blue Coat uses a format called ELFF (WC3 Extended)
  - Extend Log File Format (http://www.w3.org/TR/WD-logfile.html)
- date time time-taken c-ip cs-username cs-auth-group x-exception-id sc-filter-result cs-categories cs(Referer) sc-status s-action cs-method rs(Content-Type) cs-uri-scheme cs-host cs-uri-port cs-uri-path cs-uri-query cs-uri-extension cs(User-Agent) s-ip sc-bytes cs-bytes x-virus-id

Landscape - Log Format Details

- prefix (header) Describes a header data field. The valid prefixes are:
- c = Client
- s = Server
- r = Remote
- cs = Client to Server
- sc = Server to Client

http://www.bluecoat.com/http:/%252Fwww.bluecoat.com/support/self-service/6/ELFF_Format_Descriptions.html

Landscape - Log Format Details continued…

- LLNL format
- date time time-taken c-ip sc-status s-action sc-bytes cs-bytes cs-method cs-uri-scheme cs-host cs-ip cs-uri-port cs-uri-path cs-uri-query cs(Referer) cs-username cs-auth-group s-hierarchy s-supplier-name rs(Content-Type) cs(User-Agent) sc-filter-result cs-category x-virus-id s-ip s-sitename
- Customize your log format to best suite your needs

Landscape - Log Format Example

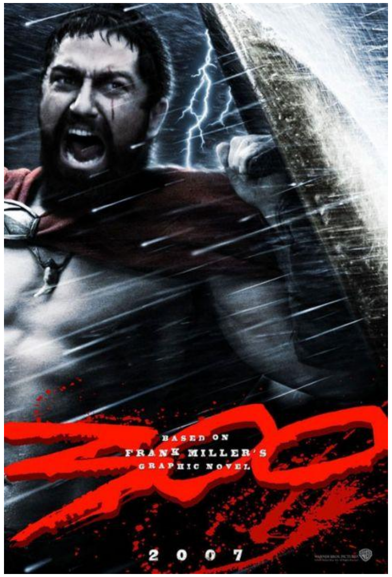
```
2010-04-20 07:00:40 225 1XX.115.109.XX 200 TCP_NC_MISS 332 533 GET http 116vistadrive.greatluxuryestate.com 65.18.172.67 80 /mlsmax/layout05/images/menu_div.gif - http://116vistadrive.greatluxuryestate.com/mlsmax/home.htm?mls=&vkey=&vid= linney1 - DIRECT 116vistadrive.greatluxuryestate.com image/gif "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3" OBSERVED "Real Estate" – 1XX.115.27.XX SG-HTTP-Service
```

Solutions – How can we solve our problems?

- Most of us now have a web proxy…now what???
  - Centralize your logs
  - Modify your log format to suite your needs
- What do the “bad guys” look like???
  - Different types of bad guys, overlap, difficult to tell apart
    - Users
    - Criminals / Entrepreneurs
    - APT (Advanced Persistent Threat)
- How do we find “bad guys” on our networks???
  - Depends on which “bad guys” we’re looking for
    - Digest
    - Analyze
    - Scrutinize

Solutions – Overview

- Parse your logs with whatever makes you happy
  - My Proof of Concept codes are in Perl
    - Need a code reference I’ll share
  - You can use grep, awk, sed, cut, PHP, C, etc.
- Practical tips
  - Pay attention to http redirects
    - 301, 302, 3XX
  - Pay attention to referrer
    - Could contain search terms
    - Multi staged attacks are commonplace
  - Looking at logs after 5pm can be detrimental! -Monzy



Solutions – Overview Continued

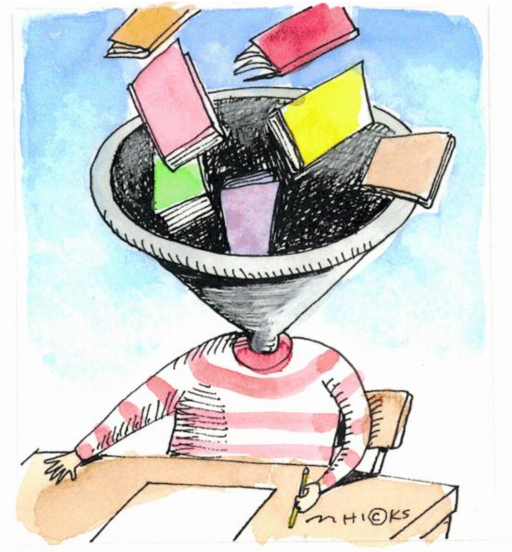
- Getting comfortable with the data
  - Machine learning algorithms are not mandatory
    - get www.010h45m.com/FreeAV2010.exe
- Our solutions will focus on the following
  - Simple statistics
    - summarization, mean, std. dev, etc.
  - User agents
  - Content Types
  - Compound Searches
  - Consult the oracle
    - a.k.a. google

Solution - Summarization

- 2010-04-20 07:00:40 225 1XX.115.109.XX 200 TCP_NC_MISS 332 533 GET http 116vistadrive.greatluxuryestate.com 65.18.172.67 80 /mlsmax/layout05/images/menu_div.gif - http://116vistadrive.greatluxuryestate.com/mlsmax/home.htm?mls=&vkey=&vid= linney1 - DIRECT 116vistadrive.greatluxuryestate.com image/gif "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3" OBSERVED "Real Estate" – 1XX.115.27.XX SG-HTTP-Service

Solution - Summarize logs

- Daily summary
  - Total HTTP users, Total FTP users, Top Sources, Top Destinations, Top Categories, Top Denied Sources, Top Spyware/Malware Sources, Top Spyware Effects, Top User Agents, Top IP getting images, Top IP performing POST’s
  - Top 15 Spyware/Malware Sources:
  - 1xx.115.226.xx : 48
  - 1xx.115.105.xxx : 47
  - 1xx.9.139.xx : 14
  - 1xx.9.139.xx : 12
  - 1xx.9.93.xx : 8
  - 1xx.115.105.xxx : 5
  - 1xx.115.105.xxx : 5
  - 1xx.9.135.xx : 2
  - 1xx.9.135.xx : 2
  - 1xx.115.62.xxx : 2
  - 1xx.9.135.xx : 1



- PoC bcsummary.pl
  - Daily summary of most of the above

Solution - Summarize all requests by TLD

- Top Level Domain (TLD)
  - I need to jump through hoops to travel physically
    - Virtually users are all over the map!
  - Summary of daily TLD's:
  - com : 15889009
  - net : 1883675
  - org : 679329
  - gov : 265059
  - edu : 125093
  - uk : 116674
  - us : 38544
  - de : 29788
  - it : 26079
  - tv : 24495
  - fr : 11703
  - ca : 11016
  - ru : 7621

  - PoC tldsummary.pl
  - summary by Top Level Domain
- Maybe you should block entire TLD's?

Solution – User Agents

-
  ```
  2010-04-20 07:00:40 225 1XX.115.109.XX 200 TCP_NC_MISS
  332 533 GET http 116vistadrive.greatluxuryestate.com
  65.18.172.67 80 /mlsmax/layout05/images/menu_div.gif -
  http://116vistadrive.greatluxuryestate.com/mlsmax/home.htm?mls=
  &vkey=&vid= linney1 - DIRECT
  116vistadrive.greatluxuryestate.com image/gif "Mozilla/5.0
  (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3)
  Gecko/20100401 Firefox/3.6.3" OBSERVED "Real Estate" –
  1XX.115.27.XX SG-HTTP-Service
  ```

- Most things identify themselves
  - Or at least try to mimic common agents
    - Even malware or unwanted software

Solution - User Agents

- Recent interesting user agents
  - Possible trojan
    - "QI.exe"
    - "TwcToolbarIe7"
    - "weather2004"
  - Possible piracy/MPAA issues
    - "AnyDVD" (DVD cloning software)
  - Possible PII issues
    - TurboTax2009.r07.005 VFNetwork/438.14 Darwin/9.8.0 (i386) (MacBook5%2C1)
  - Possible Data Loss Prevention
    - dotmacsyncclient259 CFNetwork/438.14 Darwin/9.8.0 (i386) (MacBook3%2C1)
  - Possible malware or IT issue
    - "Immunet Updater"
    - "MSDW" //Thank you Monzy,Danny Quist, Kevin Hall
      - Microsoft Dr. Watson (sqm.microsoft.com, sqm.msn.com, watson.microsoft.com)

Solution - User Agents Continued

- Possible Waste Fraud and Abuse (WFA)
  - “BattlefieldBadCompany2Updater”
  - “Solitaire III Build 33”
  - “AppleTV/1.1”
- Possible attack victims
  - “honeyd/1.5b”
  - “Winamp/5.551” //Integer overflow exploit
    - Integer overflow (www.milw0rm.com/exploits/8783)
  - WordPress/2.7; http://localhost
- Possibly anything
  - “ie8ish“
  - “blah”

- Handy lookup tool (user-agents.org)
- PoC uasummary.pl
  - Summarizes user agents and list in descending order by number of occurrences

Solution – Content Types

- 2010-04-20 07:00:40 225 1XX.115.109.XX 200 TCP_NC_MISS 332 533 GET http 116vistadrive.greatluxuryestate.com 65.18.172.67 80 /mlsmax/layout05/images/menu_div.gif - http://116vistadrive.greatluxuryestate.com/mlsmax/home.htm?mls=&vkey=&vid= linney1 - DIRECT 116vistadrive.greatluxuryestate.com image/gif "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3" OBSERVED "Real Estate" – 1XX.115.27.XX SG-HTTP-Service

Solution - Content-Type (a.k.a. MIME)

- HTTP Protocol (originally designed for SMTP)
  - Used to identify the type of information that a file contains.
    - Specified by web server using “Content-type:” header
- Common examples

text/html	.html	Web Page
image/png	.png	PNG-format image
image/jpeg	.jpeg	JPEG-format image
audio/mpeg	.mp3	MPEG Audio File
application	.exe	Executable content

Content-Type continued…

- Focus on executable content
  - "application/octet-stream“
    - Beware of .ico (i.e. favicon.ico)
  - "application/x-msdownload”
  - application/x-msdos-program

Or

- Ends in
  - .exe
  - .msi
  - .pif
  - .scr
  - etc.

Content-Type continued…

- PoC executables.pl
  - Look for everything that ends in .exe
    - Exempt items from trusted domains
      - Less than 100 domains for my enterprise
  - Interesting examples
  - www.hotfile.com/free-games-download/Treasure_Puzzle.exe
  - www.inovikov.net/srs/rgaSetup_Release_3.205.007.exe
  - download.uniblue.com/aff/rb/fdm/registrybooster.exe
  - koti.mbnet.fi/vaultec/files/miscellaneous/MinGWStudioSetup-2.05r2.exe
- The possibilities are endless

Solution – Compound Searches

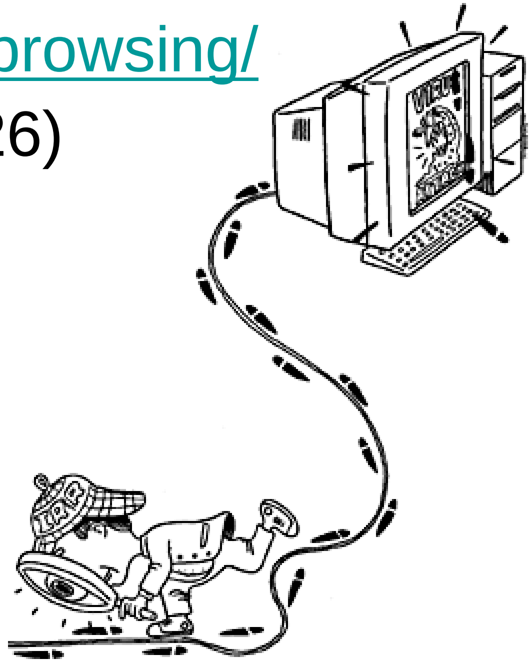
- Mix and match all of the tools previous tools
  - This where scripting languages come in handy!
- Pay close attention to some TLDs, specifically .ca!
  - If the destination ends in .ca
    - If the mime type is "application/octet-stream"
      - Print the log line
- Check out executables coming from category of “none”
  - If content type is "application/octet-stream” or "application/x-msdownload”
    - If category is “none”
      - If the file doesn’t end in .ico
        - » Print the log line

Solution – Compound Searches

- Examine requests to IP’s categorized as “none”
- If the destination host is the same as the destination IP
  - If the category is “none”
    - If this isn’t FTP
      - Print the log line
- PoC quickie.pl (does this + more)
  - Simple canned compound queries of interest
  - Useful for looking for things quickly
    - Great for APT indicators

Solution - Google Safe Browsing API

- “API that enables client applications to check URLs against Google's constantly updated blacklists of suspected phishing and malware pages. Isolates machine from Internet”
  - http://code.google.com/apis/safebrowsing/
    - ~300,227 domains (2010-04-26)
  - Built into Firefox
  - FREE
- Reactive
  - Checks nightly
  - PoC safebrows-canonical.pl
    - Monzy Merza & Adam Sealey



Future

- Continue automation/scripting
  - Proxy Log IDS (P.L.I.D.S.)
    - Share indicators/ideas that work!
- Other Antivirus Scanners
  - Experimenting with Avira/Kaspersky
    - Add to current ICAP group
- Other Content filters
  - Procuring McAfee Smartfilter
- Do More with HTTPS
  - Deny (Category=NONE && Untrusted issuer)
  - Intercept everything
  - Archive HTTPS files

Conclusion

- Further Reading
  - Dr Anton Chuvakin (chuvakin.blogspot.com)
  - Joe Griffin (http://www.sans.org/reading_room/whitepapers/malicious/mining_for_malware_theres_gold_in_them_thar_proxy_logs_32959)
- Any Questions???

- My Contact Information
  - Email: myrick3@llnl.gov (Entrust/PGP)
  - Office: 925.422.0361
- Thank you for your time ☺