MIRcon 2012

Painting the Data for Fun and Profit

William Ballenthin
Consultant
Mandiant

Agenda

- Introduction
- Exploring the Attacker Lifecycle
- Visually Reviewing Binary Files
- Making Sense of Malware Variants
- Q&A

Introduction

WILLI BALLENTHIN

- Mandiant Consultant
- Primarily Tasked with
  - Incident response
  - Forensics
  - Mobile application pen-testing
- @williballenthin
-

EXPLORING THE ATTACKER LIFECYCLE

Exploring the Attacker Lifecycle

- Problem Domain
  - During an IR, we collection many events, items
  - They're all related on a macro scale
  - And, if you're lucky, you're only dealing with one adversary...
- How can we digest the "big picture" of a compromise while still retaining access to the details?
- Timelines are an accepted approach, but are they scalable?

Motivating Example

- We're in the middle of an IR with ~5,000 hosts
- There are a few adversaries in the environment
- Fortunately, we have a number of tools available

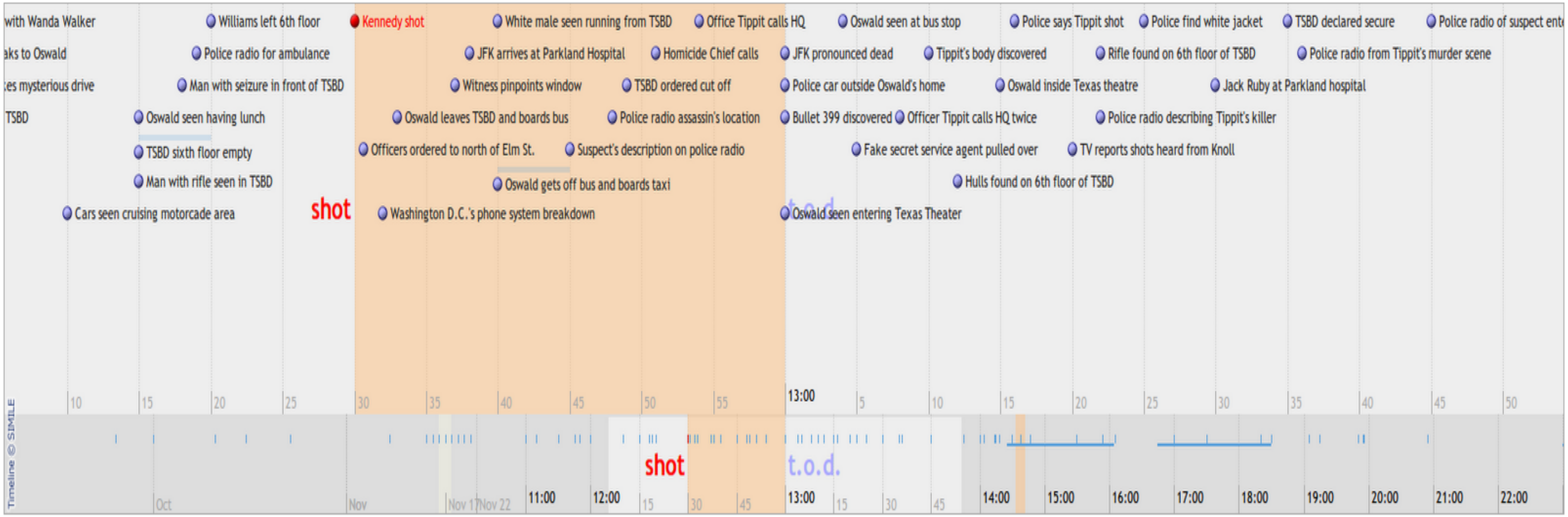
Potential Solutions

- Bodyfile/CSV/Excel
  - Handles a few hundred thousand entries
  - View is usually a simple grid
  - Data formatting?
- SIEM
  - Collects all the data, so its ready to go
  - Interface may be a bit... cumbersome

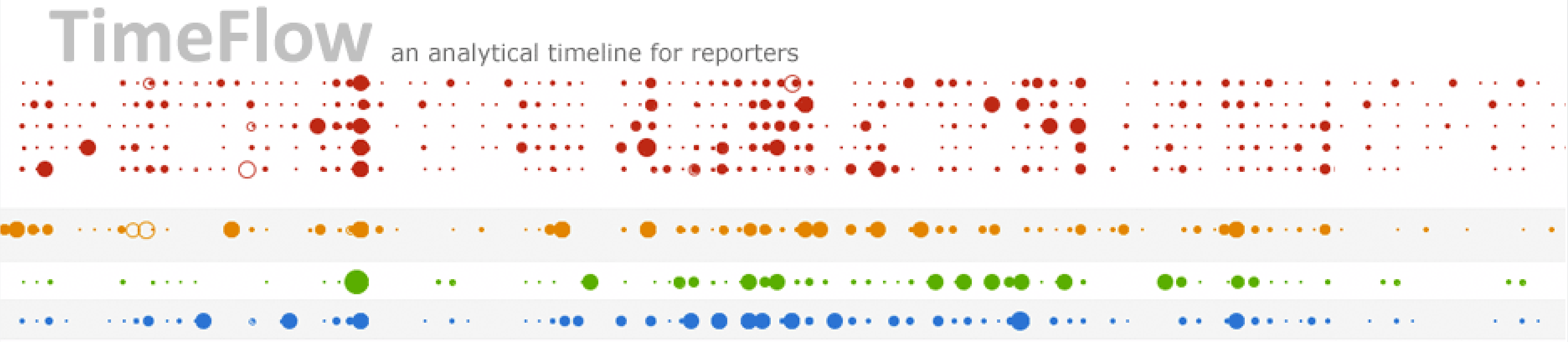
Potential Solutions

- Simile Widget
    - *Interactive* HTML + JavaScript widget
    - MIT libraries, http://www.simile-widgets.org/timeline/
    - Tons of fun to play with!
    - Does not scale to 10s of thousands of items
    - HTML page generation is required

Potential Solutions - Simile Widget


with Wanda Walker
Williams left 6th floor
Kennedy shot
White male seen running from TSBD
Office Tippit calls HQ
Oswald seen at bus stop
Police says Tippit shot
Police find white jacket
TSBD declared secure
Police radio of suspect ent
aks to Oswald
Police radio for ambulance
JFK arrives at Parkland Hospital
Homicide Chief calls
JFK pronounced dead
Tippit's body discovered
Rifle found on 6th floor of TSBD
Police radio from Tippit's murder scene
es mysterious drive
Man with seizure in front of TSBD
Witness pinpoints window
TSBD ordered cut off
Police car outside Oswald's home
Oswald inside Texas theatre
Jack Ruby at Parkland hospital
TSBD
Oswald seen having lunch
Oswald leaves TSBD and boards bus
Police radio assassin's location
Bullet 399 discovered
Officer Tippit calls HQ twice
Police radio describing Tippit's killer
TSBD sixth floor empty
Officers ordered to north of Elm St.
Suspect's description on police radio
Fake secret service agent pulled over
TV reports shots heard from Knoll
Man with rifle seen in TSBD
Oswald gets off bus and boards taxi
Hulls found on 6th floor of TSBD
Cars seen cruising motorcade area
shot
Washington D.C.'s phone system breakdown
t.o.d.
Oswald seen entering Texas Theater
10
15
20
25
30
35
40
45
50
55
13:00
5
10
15
20
25
30
35
40
45
50
Timeline © SIMILE
shot
t.o.d.
Oct
Nov
Nov 17
Nov 22
11:00
12:00
15
30
45
13:00
15
30
45
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00

Enter: TimeFlow

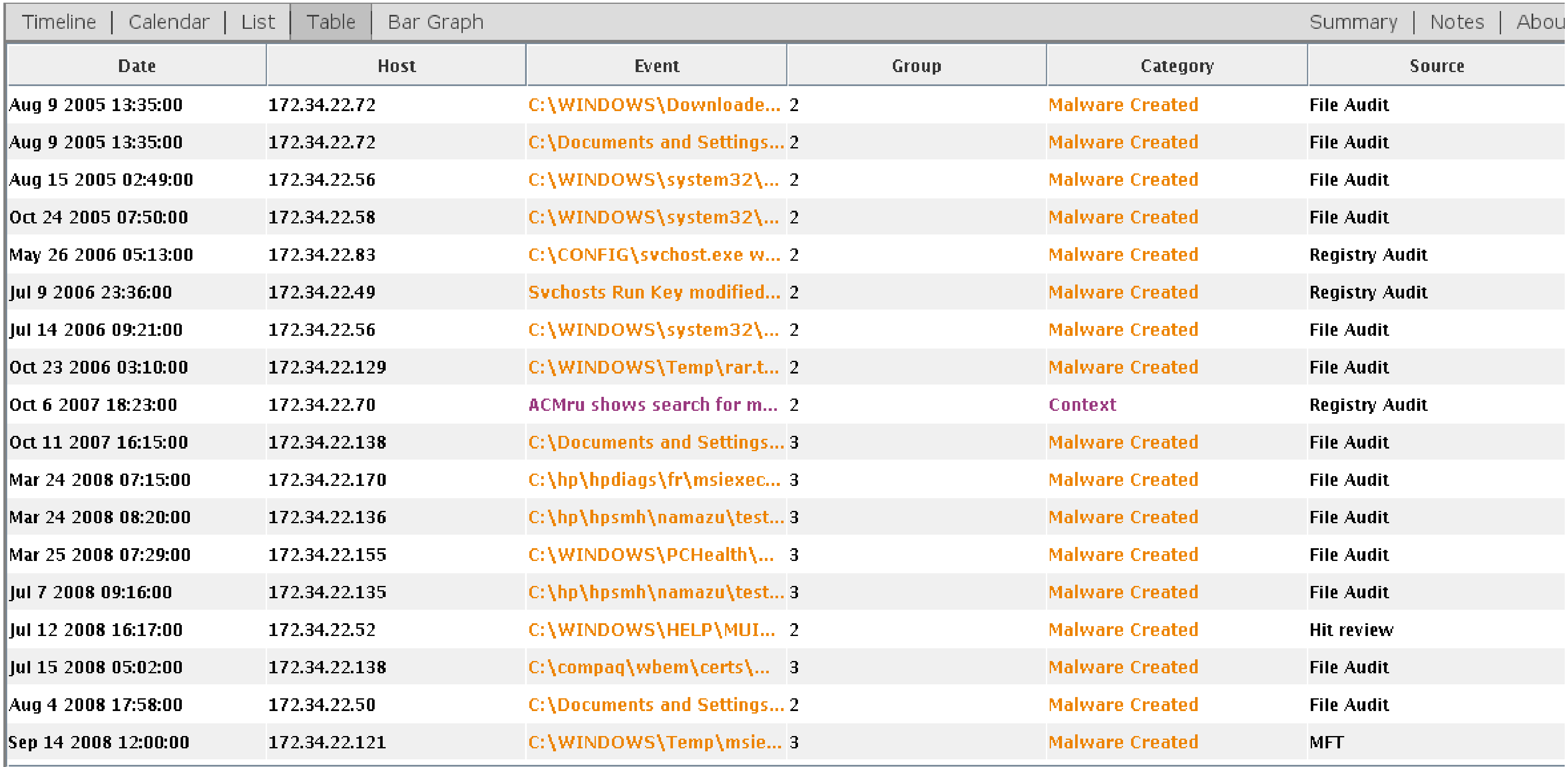


Enter: TimeFlow

- TimeFlow
  - http://flowingmedia.com/timeflow.html
  - Developed for journalists to reconstruct events
  - Extremely interactive
  - Slice-n-dice on fields
  - Supports long running events
  - A bunch of views
    - Timeline
    - Calendar
    - Bar chart
    - Table, List
  - Implemented in Java, provided as a single JAR

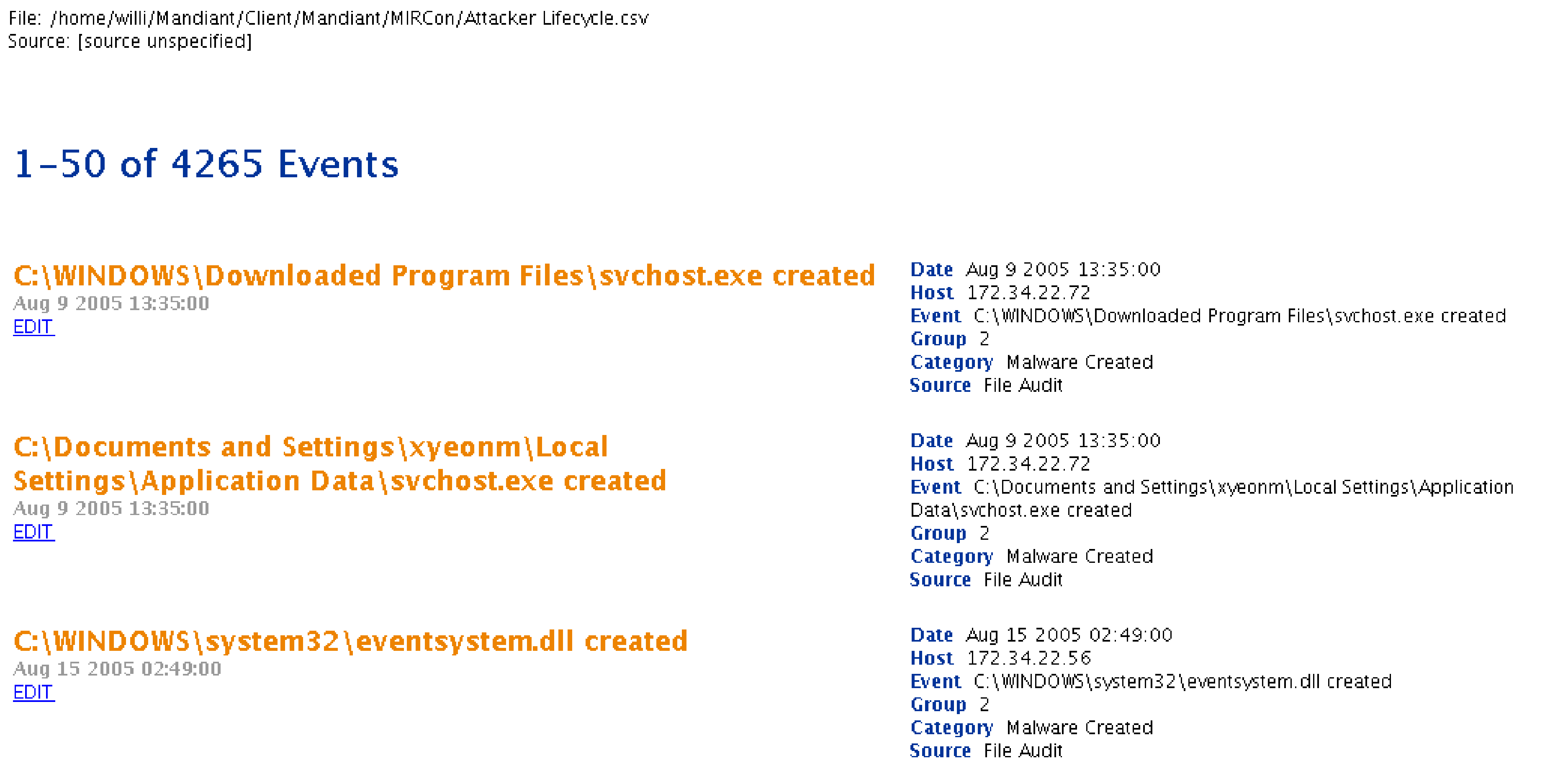
TimeFlow - As easy as a CSV

Example data: 4,265 events from ~2008 - 2010



Date	Host	Event	Group	Category	Source
Aug 9 2005 13:35:00	172.34.22.72	C:\WINDOWS\Downloade...	2	Malware Created	File Audit
Aug 9 2005 13:35:00	172.34.22.72	C:\Documents and Settings...	2	Malware Created	File Audit
Aug 15 2005 02:49:00	172.34.22.56	C:\WINDOWS\system32\...	2	Malware Created	File Audit
Oct 24 2005 07:50:00	172.34.22.58	C:\WINDOWS\system32\...	2	Malware Created	File Audit
May 26 2006 05:13:00	172.34.22.83	C:\CONFIG\svchost.exe w...	2	Malware Created	Registry Audit
Jul 9 2006 23:36:00	172.34.22.49	Svchosts Run Key modified...	2	Malware Created	Registry Audit
Jul 14 2006 09:21:00	172.34.22.56	C:\WINDOWS\system32\...	2	Malware Created	File Audit
Oct 23 2006 03:10:00	172.34.22.129	C:\WINDOWS\Temp\rar.t...	2	Malware Created	File Audit
Oct 6 2007 18:23:00	172.34.22.70	ACMru shows search for m...	2	Context	Registry Audit
Oct 11 2007 16:15:00	172.34.22.138	C:\Documents and Settings...	3	Malware Created	File Audit
Mar 24 2008 07:15:00	172.34.22.170	C:\hp\hpdiags\fr\msiexec...	3	Malware Created	File Audit
Mar 24 2008 08:20:00	172.34.22.136	C:\hp\hpsmh\namazu\test...	3	Malware Created	File Audit
Mar 25 2008 07:29:00	172.34.22.155	C:\WINDOWS\PCHealth\...	3	Malware Created	File Audit
Jul 7 2008 09:16:00	172.34.22.135	C:\hp\hpsmh\namazu\test...	3	Malware Created	File Audit
Jul 12 2008 16:17:00	172.34.22.52	C:\WINDOWS\HELP\MUI...	2	Malware Created	Hit review
Jul 15 2008 05:02:00	172.34.22.138	C:\compaq\wbem\certs\...	3	Malware Created	File Audit
Aug 4 2008 17:58:00	172.34.22.50	C:\Documents and Settings...	2	Malware Created	File Audit
Sep 14 2008 12:00:00	172.34.22.121	C:\WINDOWS\Temp\msie...	3	Malware Created	MFT

TimeFlow - Review, Edit Data



File: /home/willi/Mandiant/Client/Mandiant/MIRCon/Attacker Lifecycle.csv
Source: [source unspecified]

1-50 of 4265 Events

C:\WINDOWS\Downloaded Program Files\svchost.exe created
Aug 9 2005 13:35:00
EDIT

Date Aug 9 2005 13:35:00
Host 172.34.22.72
Event C:\WINDOWS\Downloaded Program Files\svchost.exe created
Group 2
Category Malware Created
Source File Audit

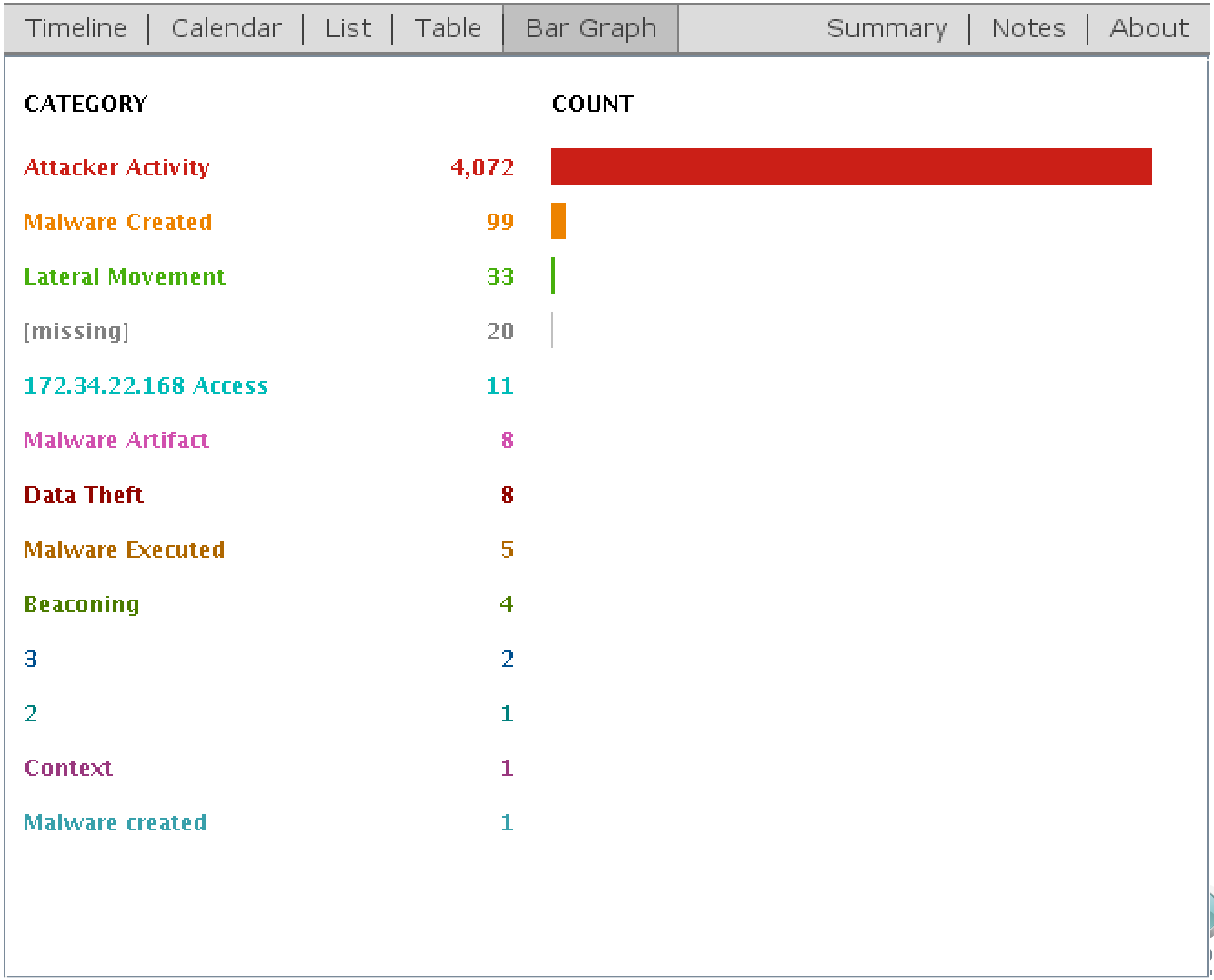
C:\Documents and Settings\xyeonm\Local Settings\Application Data\svchost.exe created
Aug 9 2005 13:35:00
EDIT

Date Aug 9 2005 13:35:00
Host 172.34.22.72
Event C:\Documents and Settings\xyeonm\Local Settings\Application Data\svchost.exe created
Group 2
Category Malware Created
Source File Audit

C:\WINDOWS\system32\eventsystem.dll created
Aug 15 2005 02:49:00
EDIT

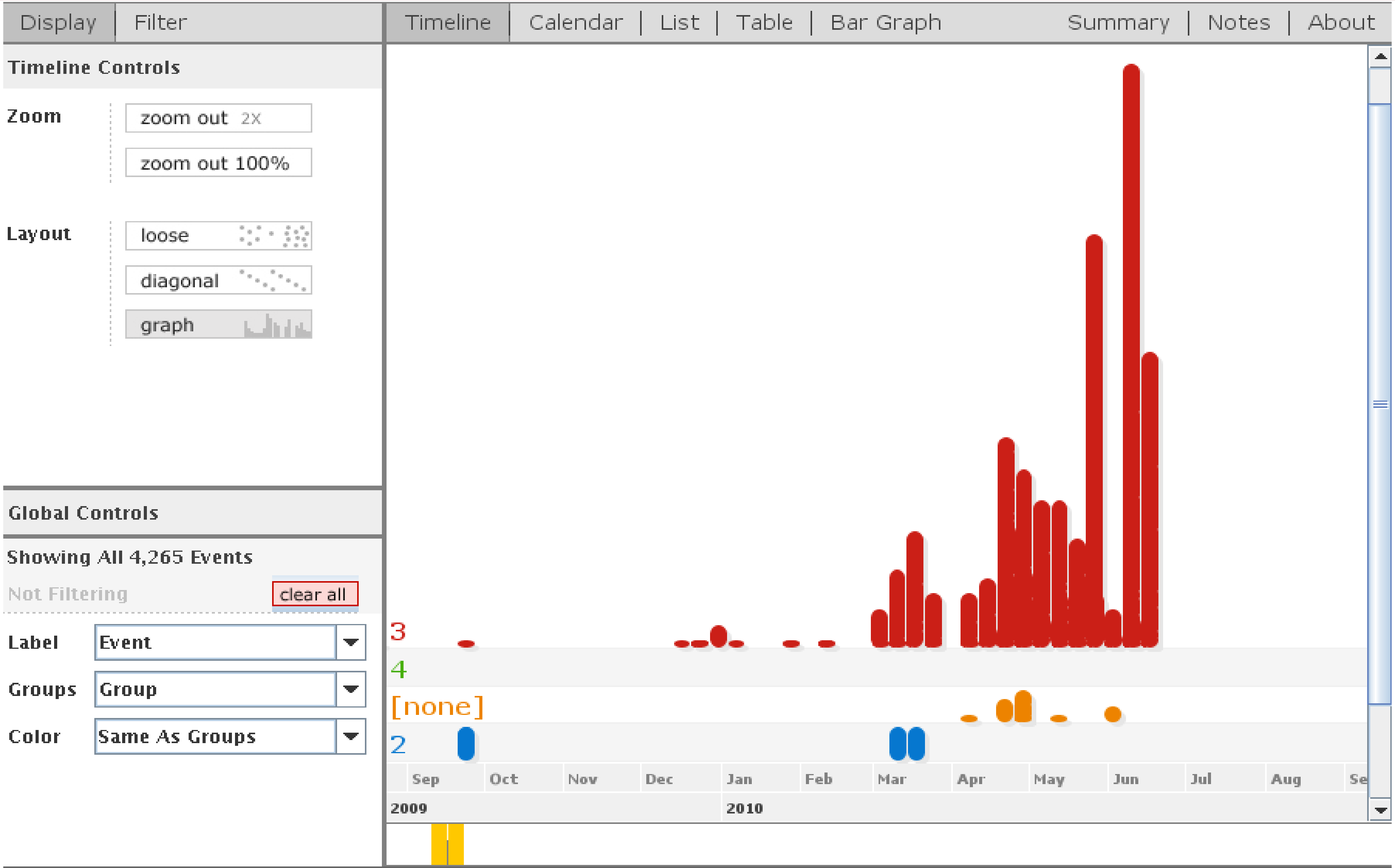
Date Aug 15 2005 02:49:00
Host 172.34.22.56
Event C:\WINDOWS\system32\eventsystem.dll created
Group 2
Category Malware Created
Source File Audit

TimeFlow - Summarize and Stack



CATEGORY	COUNT
Attacker Activity	4,072
Malware Created	99
Lateral Movement	33
[missing]	20
172.34.22.168 Access	11
Malware Artifact	8
Data Theft	8
Malware Executed	5
Beaconing	4
3	2
2	1
Context	1
Malware created	1

TimeFlow - Summarize and Timeline



Showing All 4,265 Events

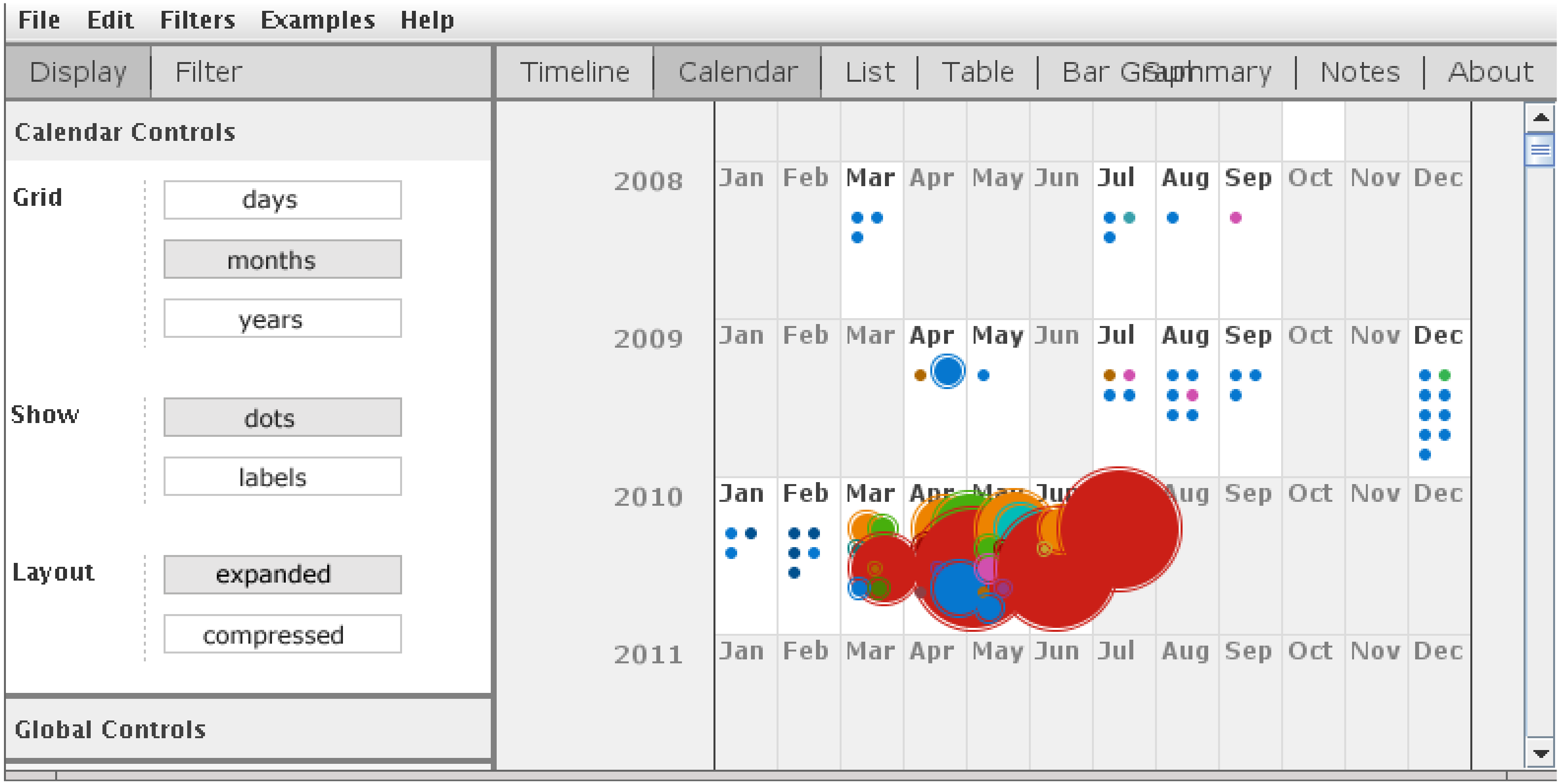
Not Filtering

Label: Event

Groups: Group

Color: Same As Groups

TimeFlow - Events over Time

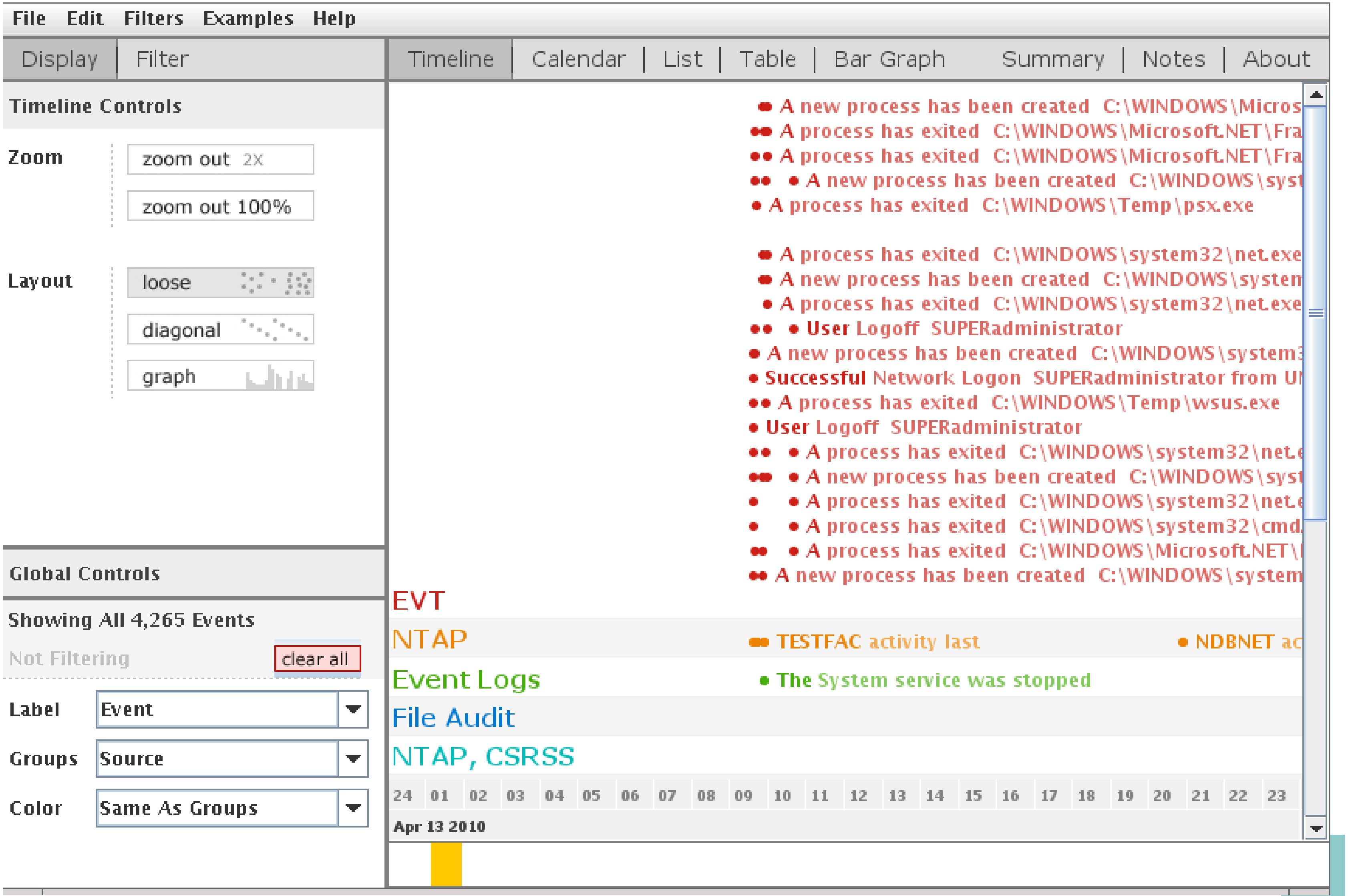


Grid: months

Show: dots

Layout: expanded

TimeFlow - Interact with the Timeline



Showing All 4,265 Events

Not Filtering

Label: Event

Groups: Source

Color: Same As Groups

```
A new process has been created  C:\WINDOWS\Micros
A process has exited  C:\WINDOWS\Microsoft.NET\Fra
A process has exited  C:\WINDOWS\Microsoft.NET\Fra
A new process has been created  C:\WINDOWS\syst
A process has exited  C:\WINDOWS\Temp\psx.exe
A process has exited  C:\WINDOWS\system32\net.exe
A new process has been created  C:\WINDOWS\system
A process has exited  C:\WINDOWS\system32\net.exe
User Logoff  SUPERadministrator
A new process has been created  C:\WINDOWS\system3
Successful Network Logon  SUPERadministrator from UI
A process has exited  C:\WINDOWS\Temp\wsus.exe
User Logoff  SUPERadministrator
A process has exited  C:\WINDOWS\system32\net.e
A new process has been created  C:\WINDOWS\syst
A process has exited  C:\WINDOWS\system32\net.e
A process has exited  C:\WINDOWS\system32\cmd.
A process has exited  C:\WINDOWS\Microsoft.NET\
A new process has been created  C:\WINDOWS\system
TESTFAC activity last
NDBNET ac
The System service was stopped
```

VISUALLY REVIEWING BINARY FILES

Visually Reviewing Binary Files

- Problem Domain
    - We treat files as (file names + arbitrary data)
    - But, what do files look like?
        - A step above hex encodings
    - Hashes, even SSDeep, have little meaning
- Once we start looking at files, can we compare them?

Motivating Example

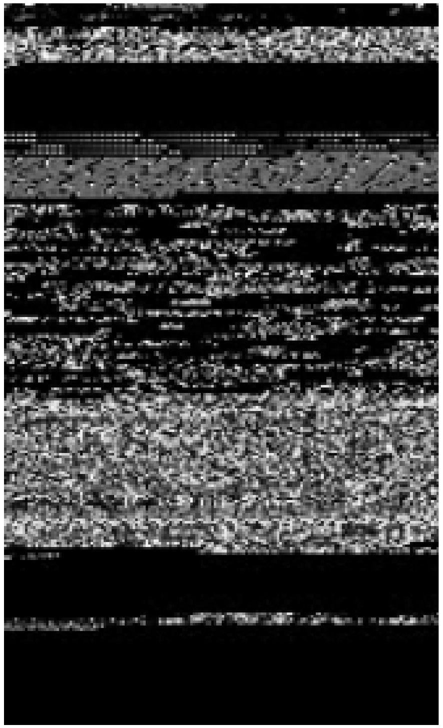
- We have two completely unknown files recovered during disk forensics
- Do they have a similar structure?
  - Sure, we can use traditional techniques, like `file`, but this doesn't capture embedded structures

Potential Solutions

- `file` - guess the file type based on headers and file structure
- `diff` - compare text and show differences
- Hex editor "compare files..."
- Distance function from part 3
- Domain-specific tools
    - e.g. `objdump` for executable files

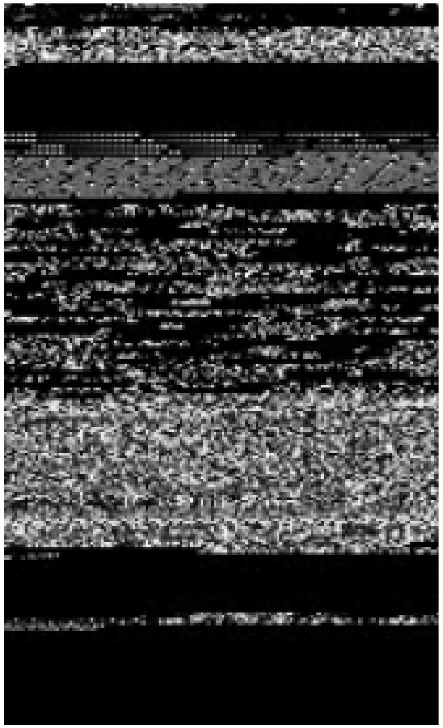
Let's try to draw the files

- *Malware images: Visualization and automatic classification*. L. Nataraj, S. Karthikeyan, G. Jacob, and B. Manjunath, 2011
  - Convert file to a vector of 8-bit values
  - Use this data as a bitmap
  - Ultimate goal: use image recognition techniques to identify malware
    - Turns out, this works



"Malware Images" Technique

- This works well
  - Very intuitive
  - Fast
- However,
  - Color scale
  - File sizes / image dimensions
  - Feature locality

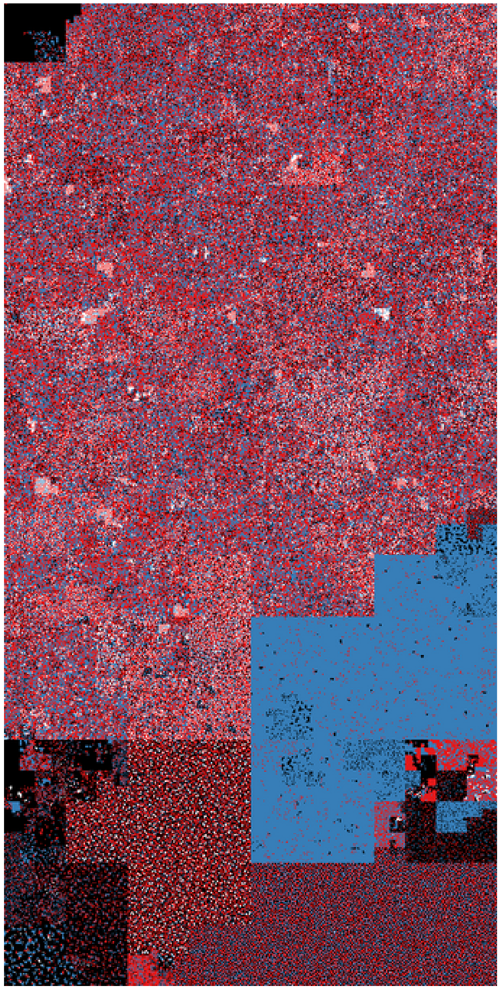


MIRcon 2012

Aldo Cortesi - binvis

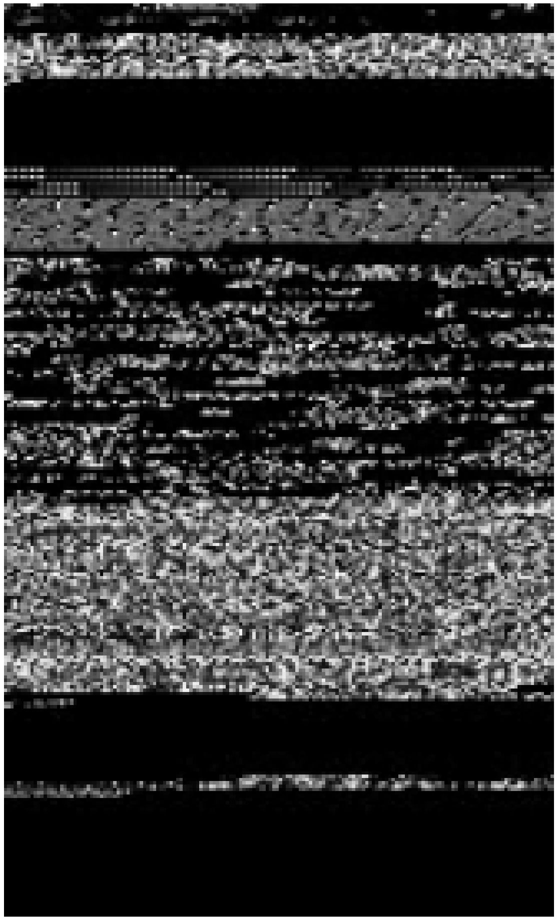
- Aldo of Nullcube suggests an improvement `binvis`
  - Meaningful colors
  - Better spatial clustering
  - Free, open-source, Python

- http://corte.si/posts/visualisation/binvis/index.html

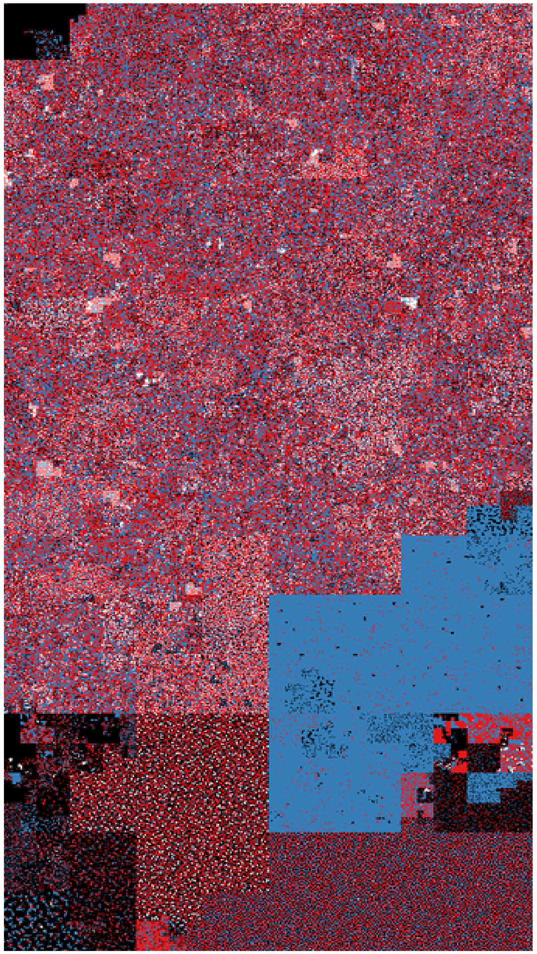


MIRcon 2012

"Malware Images"



"binvis"



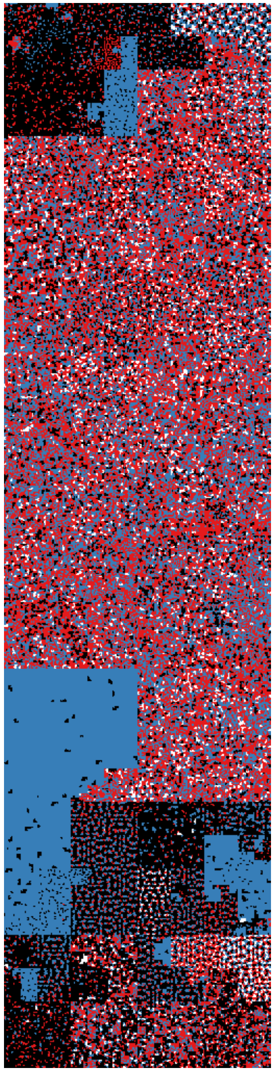
"binvis" Color Schema

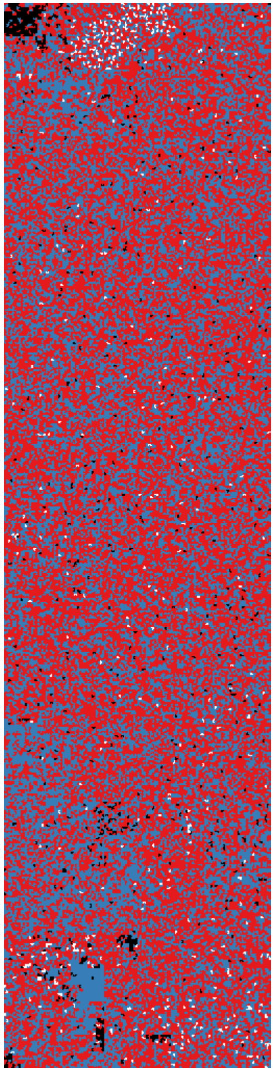
⬛ Black - `0x00`

⬜ White - `0xFF`

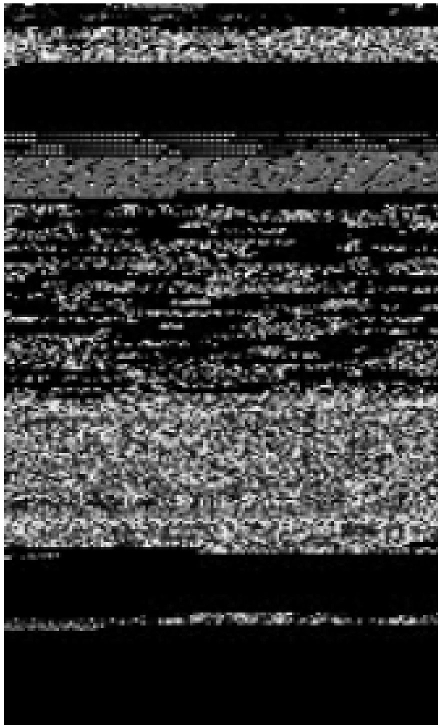
🟦 Blue - Printable

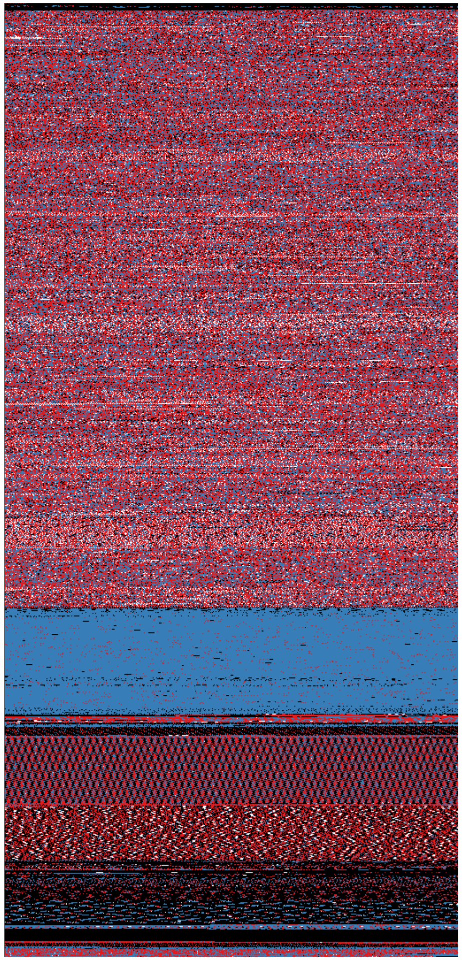
🟥 Red - Else

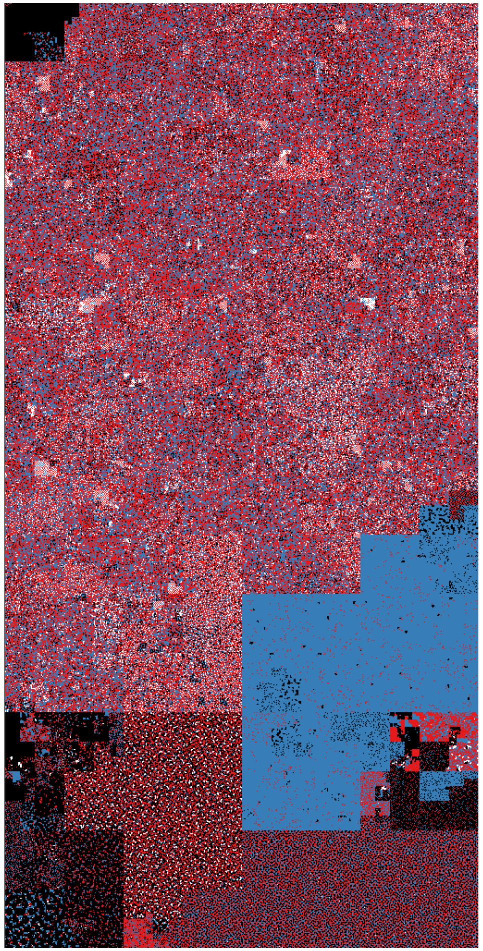




Coloring is a start...





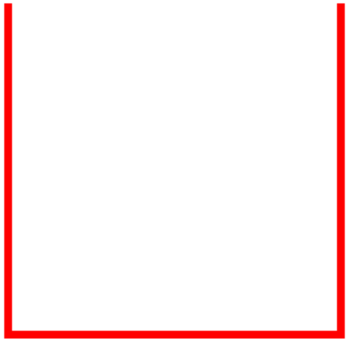


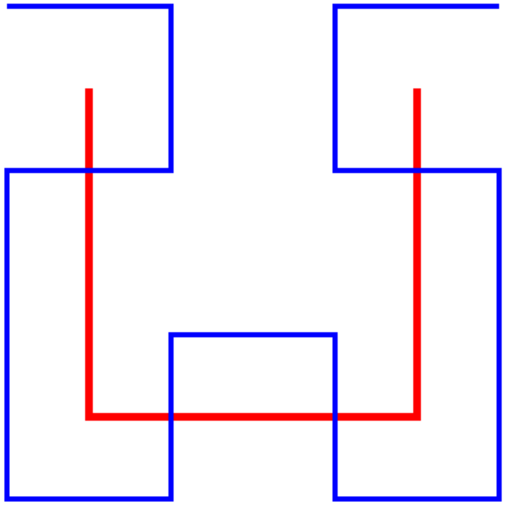
Some mathematics: Hilbert Curves

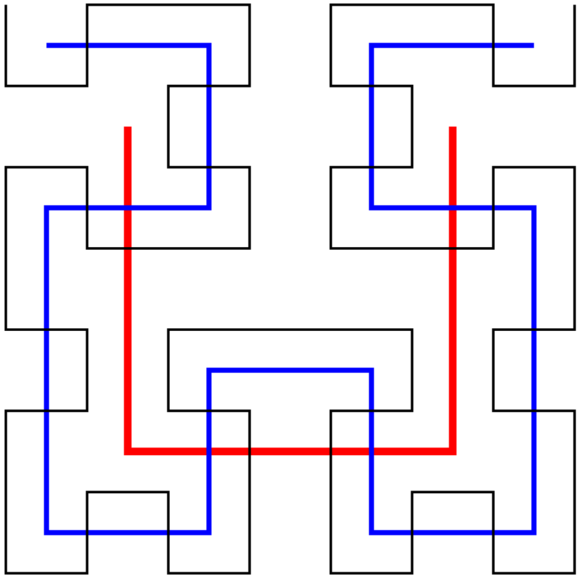
- Space filling curves
  - Intuitively, draw line along all points in a region without crossing
  - Why? Georg Cantor: the infinite points on a unit line has the same cardinality as the infinite points in the unit square

- Hilbert curve
  - David Hilbert in 1891
  - Mapping preserves (some) locality from 1D to 2D
  - Close association with fractals, so plots are approximations

Building Hilbert Curves



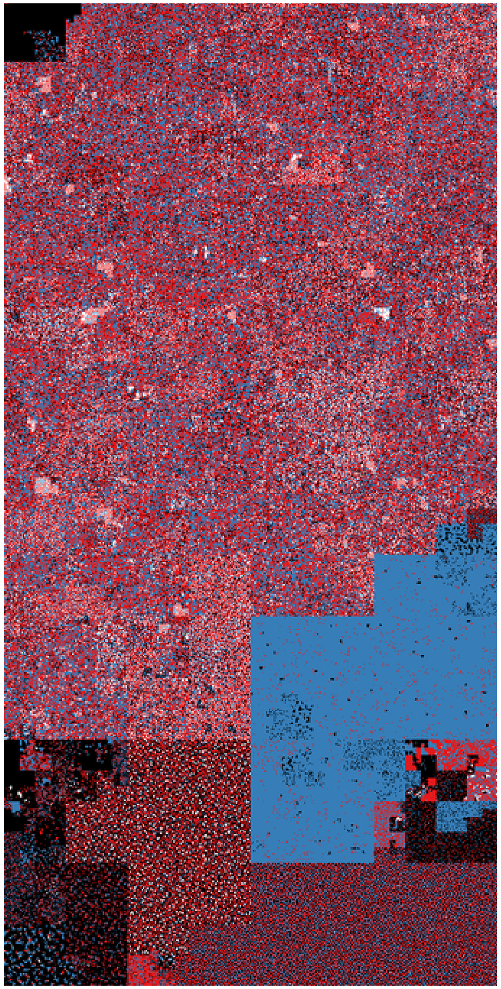




"binvis" Technique

- This works well
  - Colors are meaningful
  - Features are obvious

- However,
  - Slow (Hilbert curve calcs)
  - Feature shapes inconsistent
  - Feature locations unintuitive



MAKING SENSE OF MALWARE VARIANTS

Making Sense of Malware Variants

- Problem Domain
  - Malware is not unique
  - Variants are grouped into families
    - zbot/Zeus Trojan
    - Poison Ivy RAT
    - Gh0st RAT
- How do we identify families?
  - Differences in settings
    - C2 domains or IPs
  - Differences in capabilities
    - Gh0st extended to inject shellcode
  - Differences in bugs
    - New versions of Poison Ivy

Motivating Example

- A client gives us 500 malwarez and asks for a report on each one
  - We know many share the same author, intent
  - Let's just find the families, pick representative samples, and reverse those, instead
- Result
  - Client is happy and richer
  - We spend less time in front of IDA

Data Sources

- Binary file similarities (static)
  - Entropy
  - Fuzzy hashing - ssdeep
- Malware analysis sandboxes (dynamic)
  - Cuckoo sandbox, Mandiant Threat Analyzer
- PE file similarities (static)
  - objdump
- Disassembly-based graph theory comparisons (static)
  - bindiff
- Anti-virus signatures
- Malware analyst brains (expensive)

Clustering

- Explorative data mining
- From a bunch of samples, produce groups of similar things
- Here, require only a distance function to identify nearest neighbors
    - Distance function: a metric between two samples that describes how similar (or different) they are
    - Compose a distance function from a set of weighted metrics

```
D(x,y) = a0 * d0(x,y) + a1 * d1(x,y) + ... aN * dN(x,y)
```

Distance Function Ideas - Static Analysis

- Find the range of the function and normalize
  - e.g. Entropy, scale to 1.0 by dividing by 8.0
  - Other numeric functions, you may scale by the standard deviation
  - Categorical distance metric - use a points-based function
    - 10 points * number of shared imports, max. 10
    - 20 points if both are a DLL
    - etc.

Distance Function Ideas - Dynamic Analysis

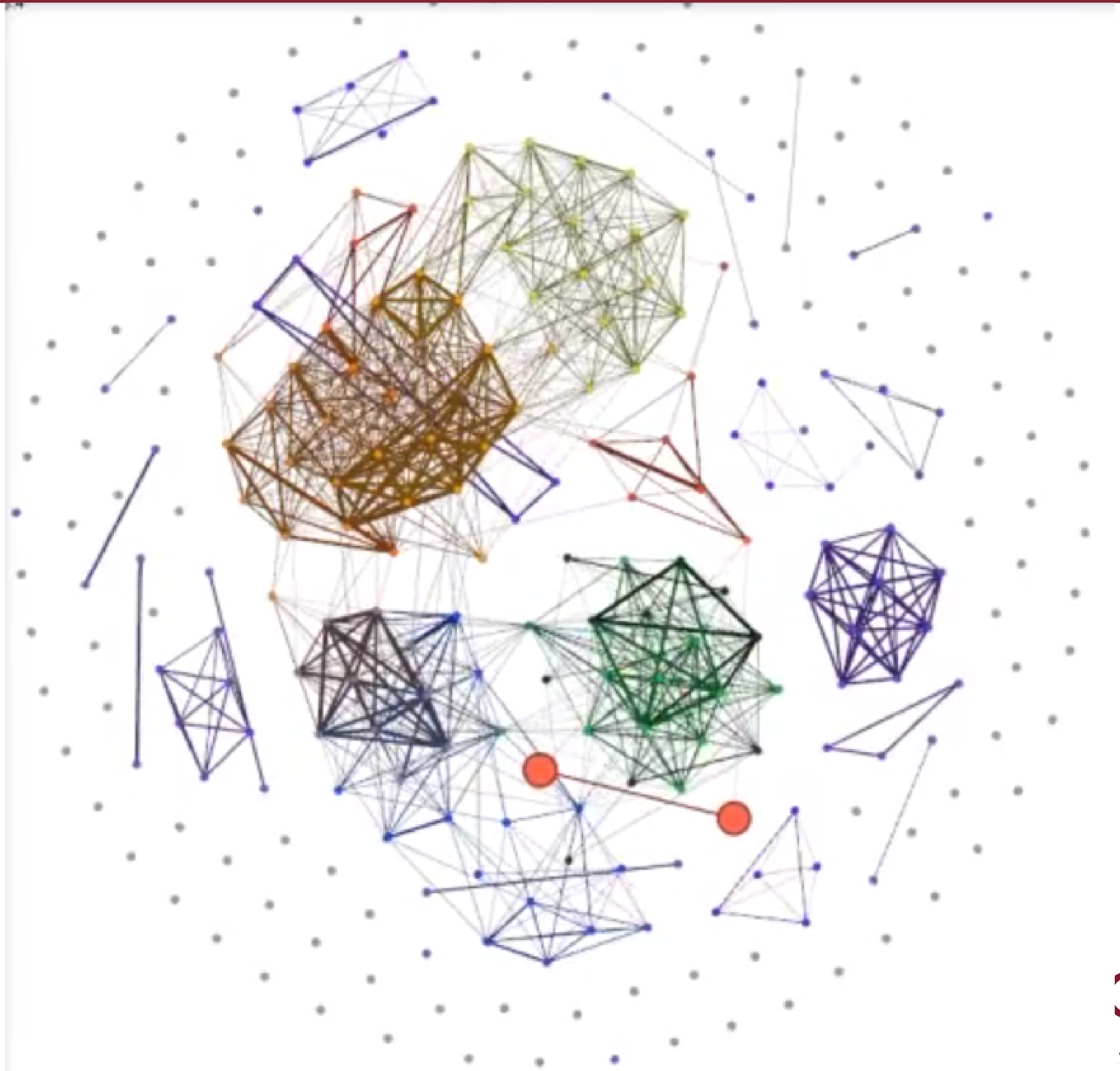
- Record API calls and use the Levenshtein edit distance
  - "the number of single-character edits required to change one word into the other"
  - `s/character/api call/g` and `s/word/call history/g`

- Record file system/Registry/etc. activity and define a categorical composite distance metric
  - 10 points if it writes to the same directory
  - 50 points if it changes the same Registry key

Let's find some families

- We'll use a force-directed layout when graphing nodes
  - aka. minimize a global energy function
  - akka. pretend each spring is a bowling ball and there's springs among all the balls
  - *Graphviz*
    - http://www.graphviz.org/
    - 'neato', 'fdp', and 'sfdp' layout algorithms
  - *Gephi*
    - https://gephi.org/
    - "an interactive visualization and exploration platform for networks and complex systems"

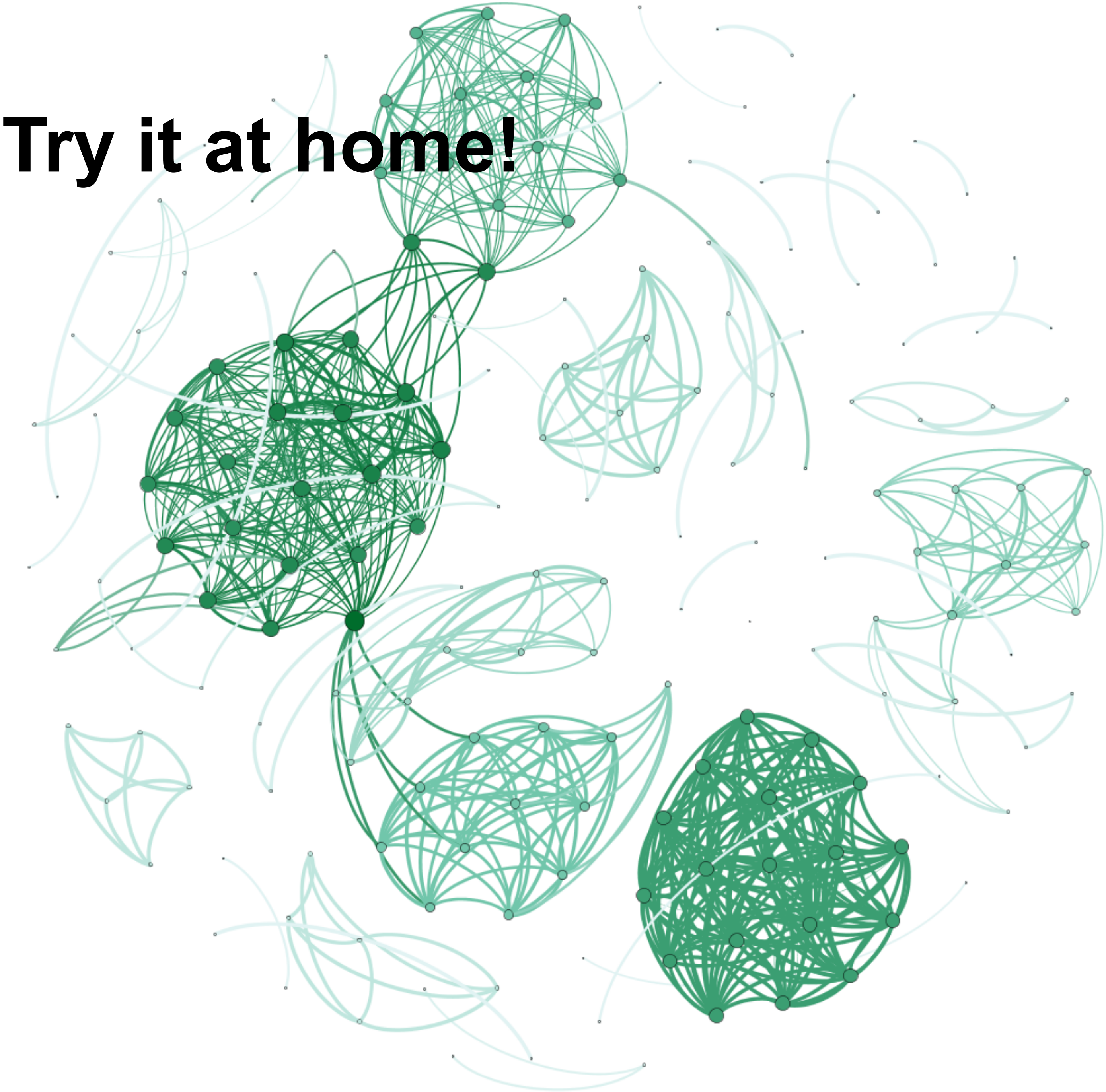
Motivating Example: Results



Try it at home!

```
ssdeep -r -p .   |
  grep "matches" |
  sed
    -e "s/.*\/\([^\/]*\) matches/\1,matches/g"
    -e "s/matches.*\/\([^\/]*\)/\1,/g"
    -e "s/ (\\([0-9]*\\))/,0.\1/g" |
  awk '
    BEGIN{print "Source,Target,Weight,Type"}
         {print $0",Undirected"}'
> /tmp/clusters.csv
```

Try it at home!



Try it at home!

- With Gephi
  - New Project...
  - Data Labratory
  - Import Spreadsheet
  - As Table... Edges table
  - Finish
  - Overview
  - Choose a layout... "Fruchterman Rheingold"
  - Run
  - ???
  - Profit

Q&A

Citations

- Malware Images: Visualization and Automatic Classification
- A Comparative Assessment of Malware Classification using Binary Texture Analysis and Dynamic Analysis
- Wikipedia
- http://corte.si/posts/visualisation/hilbert-snake/index.html and others
- http://flowingmedia.com/timeflow.html
- http://www.simile-widgets.org/
- https://gephi.org/