VISUAL HUNTING WITH LINKED DATA

sqrrl

Securely explore your data

ABOUT ME

Security Architect at Sqrrl. Research areas include threat intelligence, security analytics and the art & science of hunting.

15 years of detection & response experience in government, research, educational and corporate arenas.

A founding member of a Fortune 5's CIRT. Spent 5 years helping to build a global detection & response capability (500+ sensors, 5PB PCAP, 4TB logs/day).

AGENDA

What is Linked Data?

Why use Linked Data Analysis for Hunting?

Deriving Insights from Visualization

WHAT IS LINKED DATA?

Securely explore your data

WHAT IS LINKED DATA?

"[…] a method of publishing structured data so that it can be interlinked and become more useful through semantic queries."

Source: http://en.wikipedia.org/wiki/Linked_data, last checked May 2015

I DIDN'T QUITE CATCH THAT LAST SLIDE

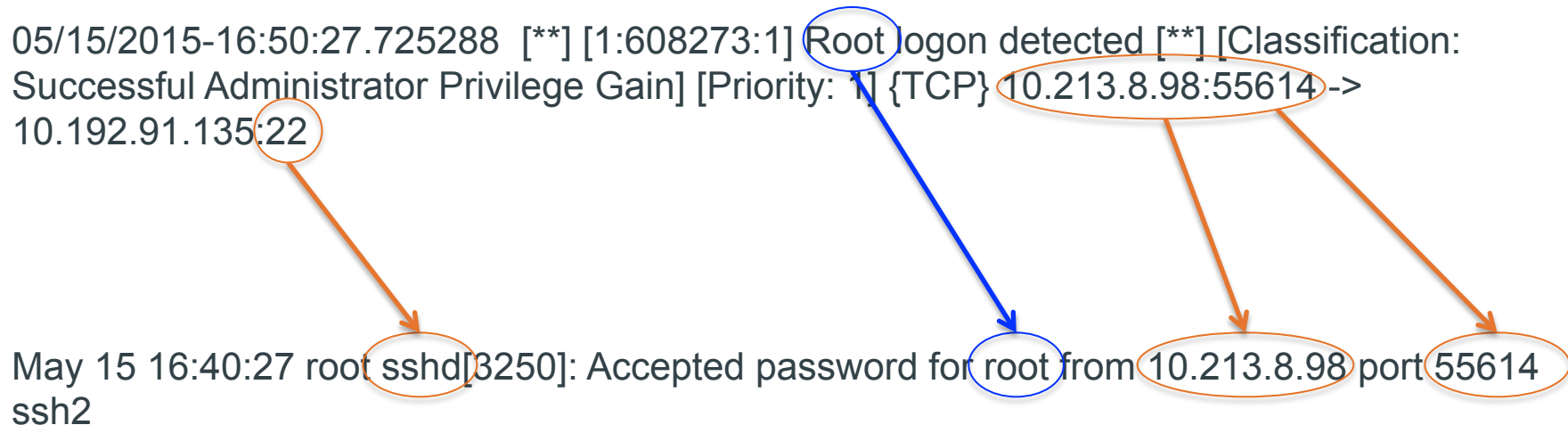**Can you say that in English this time, please?**

Data with connections to other data embedded in it, either implicitly or explicitly.
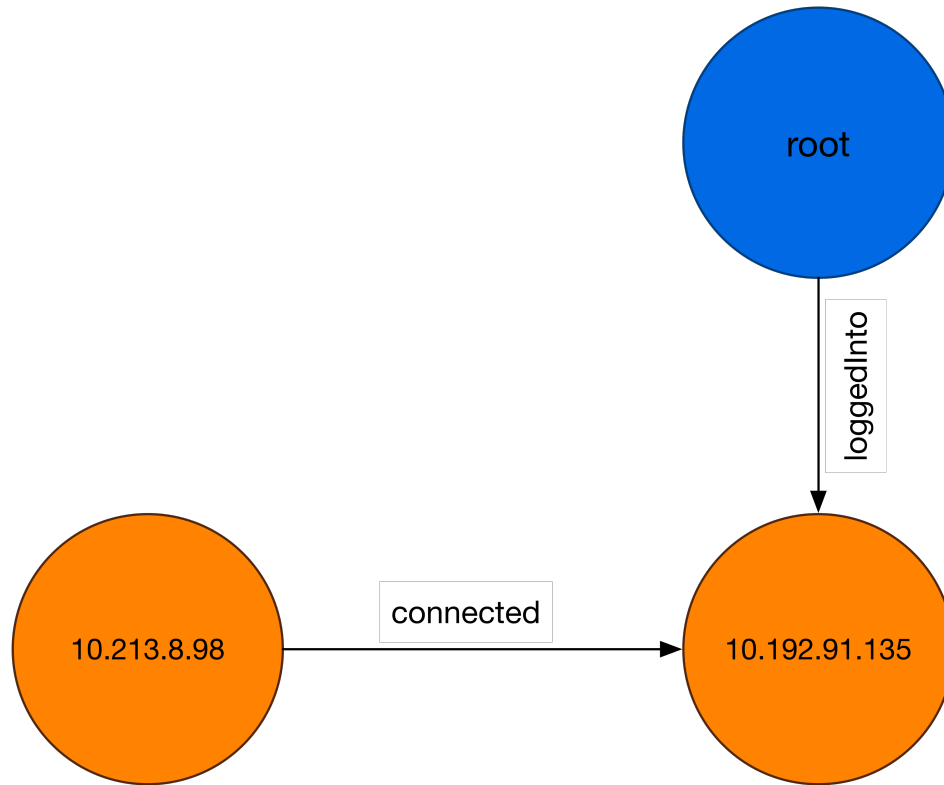
IMPLICIT LINKS ARE INFERRED

05/15/2015-16:50:27.725288 [**] [1:608273:1] Root logon detected [**] [Classification: Successful Administrator Privilege Gain] [Priority: 1] {TCP} 10.213.8.98:55614 -> 10.192.91.135:22

May 15 16:40:27 root sshd[3250]: Accepted password for root from 10.213.8.98 port 55614 ssh2

IMPLICIT LINKS ARE INFERRED

05/15/2015-16:50:27.725288 [**] [1:608273:1] Root logon detected [**] [Classification: Successful Administrator Privilege Gain] [Priority: 1] {TCP} 10.213.8.98:55614 -> 10.192.91.135:22

May 15 16:40:27 root sshd[3250]: Accepted password for root from 10.213.8.98 port 55614 ssh2

IMPLICIT LINKAGE

EXPLICIT LINKS ARE STATED

1999-03-29T13:01:38-0500   Fz892b2SFbpSayzLyl   172.16.113.204 194.7.248.153
Cr4RV91FD8iPXBuoT6   SMTP   1   MD5   text/x-c   -   0.000000   T   F   1522
-     0   0   F   -   6d01739d1d56c64209098747a5756443   -   -   -

1999-03-29T13:01:38-0500 Cr4RV91FD8iPXBuoT6   194.7.248.153   1027 172.16.113.204 25
     1   delta.peach.mil   <hamishs@delta.peach.mil> <tierneyr@goose.eyrie.af.mil>   Mon, 29
Mar 1999 08:01:38 -0400   -   tierneyr@goose.eyrie.af.mil -   <19990329080138.CAA2048>
     -     Phonetics software Tech,   -   (from mail@localhost) by delta.peach.mil (SMI-8.6/SMI-
SVR4)\x09id: CAA2048; Mon, 29 Mar 1999 08:01:38 -0400   -   250 Mail accepted
     172.16.113.204,194.7.248.153   -   F   Fz892b2SFbpSayzLyl F

1999-03-29T13:01:38-0500 Cr4RV91FD8iPXBuoT6   194.7.248.153   1027 172.16.113.204 25
     tcp   smtp 0.113325   1923 336 SF   ShAdDafF 13   2447 12   820   (empty)

EXPLICIT LINKS ARE STATED
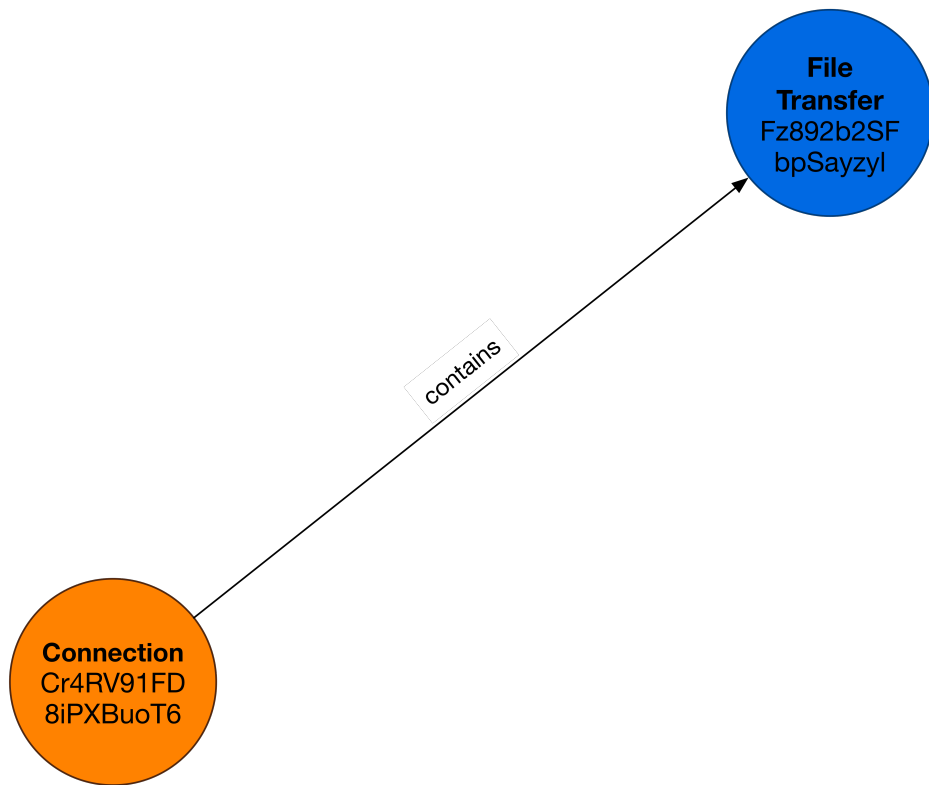
1999-03-29T13:01:38-0500   Fz892b2SFbpSayzLyl   172.16.113.204   194.7.248.153
Cr4RV91FD8iPXBuoT6   SMTP   1   MD5   text/x-c   -   0.000000   T   F   1522
-     0   0   F   -   6d01739d1d56c64209098747a5756443   -   -   -

1999-03-29T13:01:38-0500   Cr4RV91FD8iPXBuoT6   194.7.248.153   1027 172.16.113.204   25
    1   delta.peach.mil   <hamishs@delta.peach.mil> <tierneyr@goose.eyrie.af.mil>   Mon, 29
Mar 1999 08:01:38 -0400   -   tierneyr@goose.eyrie.af.mil   -   <19990329080138.CAA2048>
    -     Phonetics software Tech,   -   (from mail@localhost) by delta.peach.mil (SMI-8.6/SMI-
SVR4)\x09id: CAA2048; Mon, 29 Mar 1999 08:01:38 -0400   -   250 Mail accepted
     172.16.113.204,194.7.248.153   -   F   Fz892b2SFbpSayzLyl   F

1999-03-29T13:01:38-0500   Cr4RV91FD8iPXBuoT6   194.7.248.153   1027 172.16.113.204   25
    tcp   smtp 0.113325   1923 336   SF   ShAdDafF 13   2447 12   820   (empty)
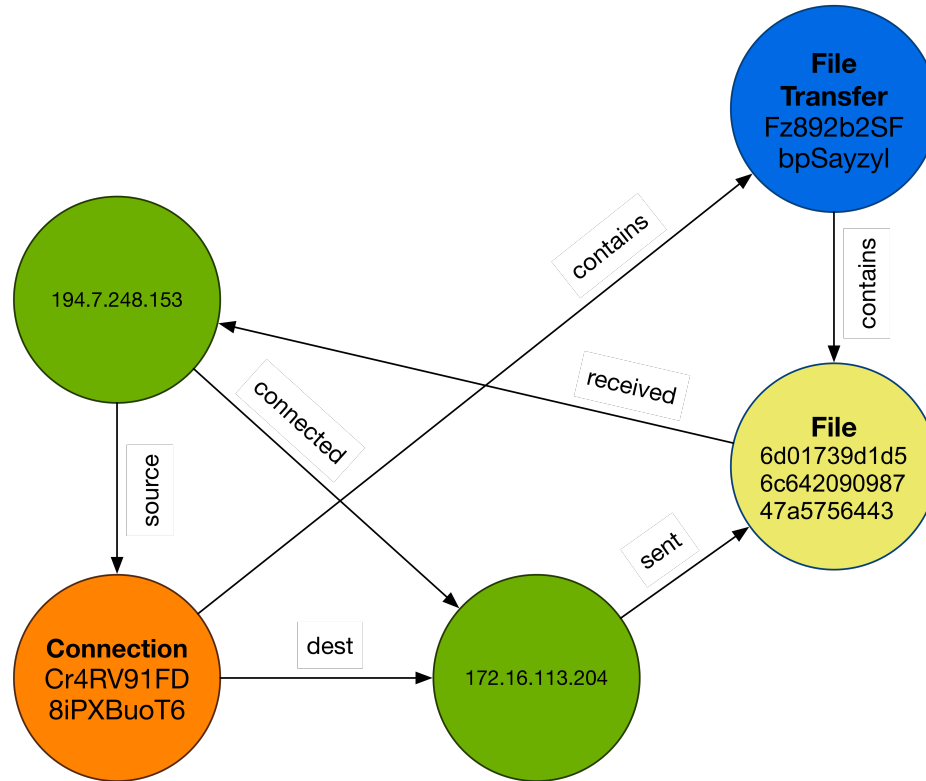
EXPLICIT LINKS ARE STATED

1999-03-29T13:01:38-0500   Fz892b2SFbpSayzLyl   172.16.113.204   194.7.248.153
Cr4RV91FD8iPXBuoT6   SMTP   1   MD5   text/x-c   -   0.000000   T   F   1522
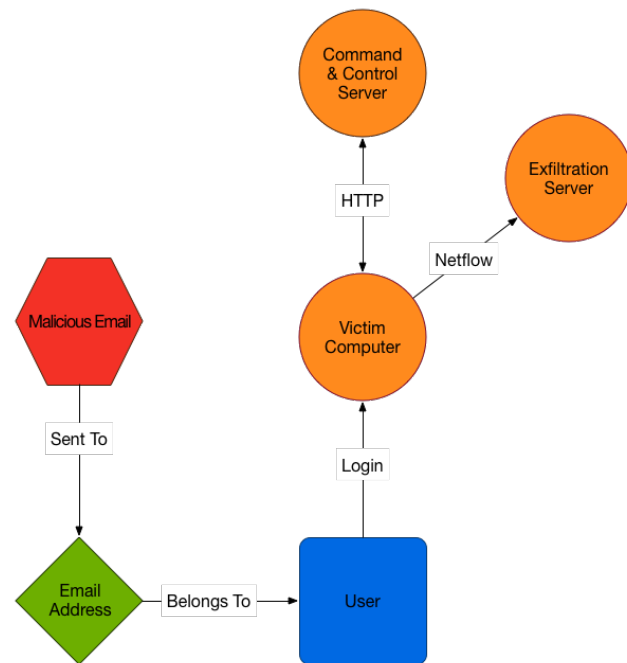-   0   0   F   -   6d01739d1d56c64209098747a5756443   -   -   -

1999-03-29T13:01:38-0500   Cr4RV91FD8iPXBuoT6   194.7.248.153   1027 172.16.113.204   25
        1   delta.peach.mil   <hamishs@delta.peach.mil> <tierneyr@goose.eyrie.af.mil>   Mon, 29
Mar 1999 08:01:38 -0400   -   tierneyr@goose.eyrie.af.mil   -   <19990329080138.CAA2048>
        -         Phonetics software Tech,   -   (from mail@localhost) by delta.peach.mil (SMI-8.6/SMI-
SVR4)\x09id: CAA2048; Mon, 29 Mar 1999 08:01:38 -0400   -   250 Mail accepted
        172.16.113.204,194.7.248.153   -   F   Fz892b2SFbpSayzLyl   F

1999-03-29T13:01:38-0500   Cr4RV91FD8iPXBuoT6   194.7.248.153   1027 172.16.113.204   25
        tcp   smtp 0.113325   1923 336   SF   ShAdDafF 13   2447 12   820   (empty)

MODELING THE DATA

MODELING THE DATA
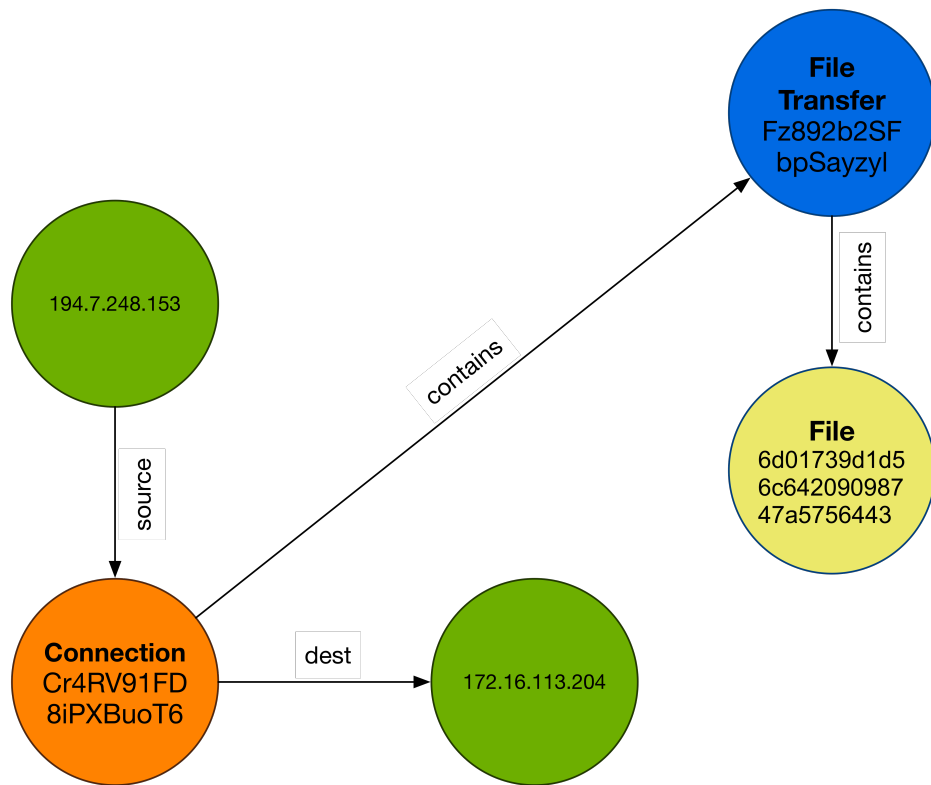
TRANSITIVE CLOSURE

WHY USE LDA FOR HUNTING?

Securely explore your data

HOW YOU'RE PROBABLY DOING IT NOW

Row-oriented techniques can only get you so far

Davids-MacBook-Pro-2:/Users/bianco/temp> grep 6d01739d1d56c64209098747a5756443 *.log

files.log:922712498.188977   Fz892b2SFbpSayzLyl   172.16.113.204   194.7.248.153

    Cr4RV91FD8iPXBuoT6   SMTP 1   MD5,SHA1   text/x-c   0.000000   T   F   1522 -   0

    0   F   -   6d01739d1d56c64209098747a5756443 0d1c6b7dcc82b05c719d4cc9dd8d8577e8cb36cb

    -

Davids-MacBook-Pro-2:/Users/bianco/temp> grep Cr4RV91FD8iPXBuoT6 *.log

conn.log:922712498.086765   Cr4RV91FD8iPXBuoT6   194.7.248.153   1027   172.16.113.204   25   tcp

    smtp   0.113325   1923   336   SF   ShAdDafF   13   2447   12   820   (empty)

files.log:922712498.188977   Fz892b2SFbpSayzLyl   172.16.113.204   194.7.248.153

    Cr4RV91FD8iPXBuoT6   SMTP 1   MD5,SHA1   text/x-c   0.000000   T   F   1522 -   0

    0   F   -   6d01739d1d56c64209098747a5756443 0d1c6b7dcc82b05c719d4cc9dd8d8577e8cb36cb

    -

smtp.log:922712498.119932   Cr4RV91FD8iPXBuoT6   194.7.248.153   1027   172.16.113.204   25   1

    delta.peach.mil   <hamishs@delta.peach.mil>   <tierneyr@goose.eyrie.af.mil>   Mon, 29 Mar 1999

08:01:38 -0400   -   tierneyr@goose.eyrie.af.mil   -   <19990329080138.CAA2048>   -   Phonetics

software Tech,   -   (from mail@localhost) by delta.peach.mil (SMI-8.6/SMI-SVR4)\x09id: CAA2048; Mon, 29

Mar 1999 08:01:38 -0400   -   250 Mail accepted   172.16.113.204,194.7.248.153   -   F

    Fz892b2SFbpSayzLyl   F

LINKED DATA ANALYSIS (LDA)

Different techniques, different perspectives

TRADITIONAL TECHNIQUES VS. LDA

| Row Oriented Analysis	Linked Data Analysis
Operates on individual events Many existing toolsets (grep/awk, ELSA, Splunk, ELK stack, etc) Hard to see the big picture Limited pivoting ability Best for searching, counting and extracting detailed proof of events	Aggregates data into entities and relationships Visual representation promotes understanding of the data Apply specialized graph algorithms: • Search for "patterns" in a graph • Identify important nodes with betweenness, page rank, etc. • Path finding ("auto-pivot++")

A TYPICAL IDS/SIEM ALERT
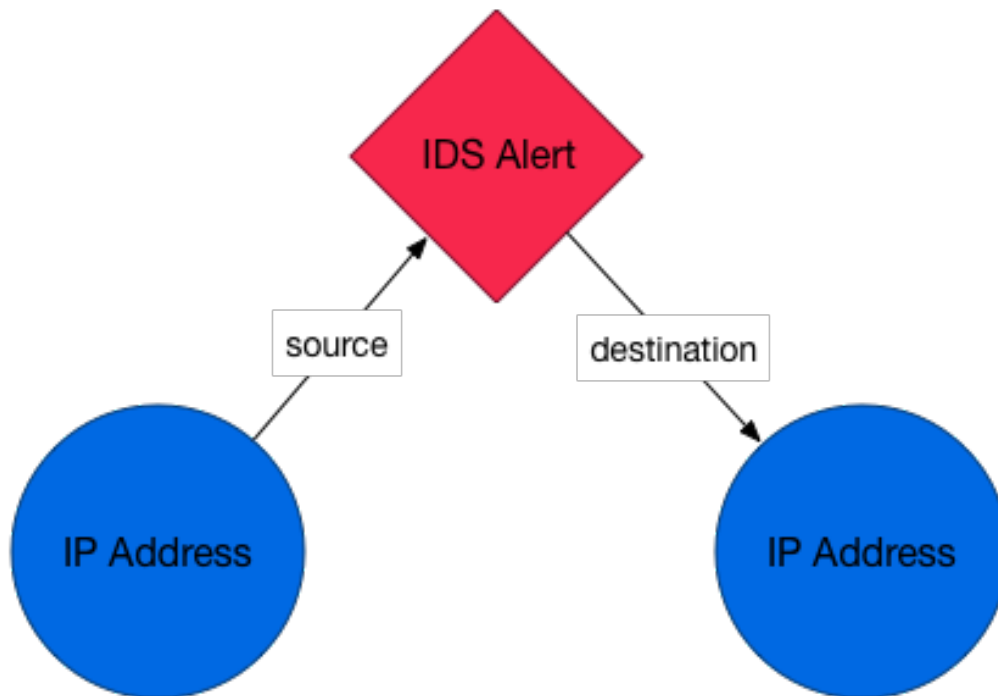
EXISTING LINKS SHOW CONTEXT

sqrrl
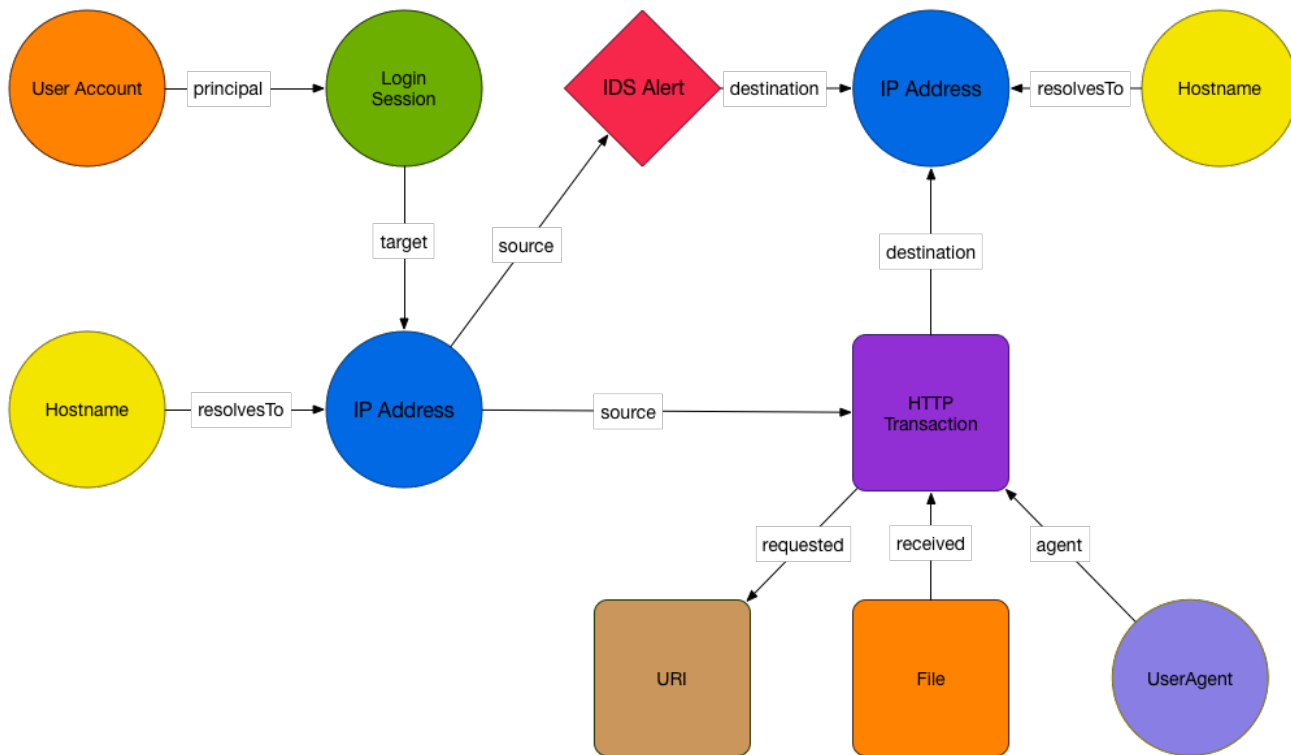
Securely explore your data

DERIVING INSIGHTS FROM VISUALIZATION

A WORD ABOUT PROCESS

To replicate this at home, you will need…

|	
DARPA99 Challenge Data	http://www.ll.mit.edu/ideval/data/1999data.html
Bro Network Analysis Platform	https://www.bro.org
Bro2Graph Scripts	https://github.com/DavidJBianco/Bro2Graph
Rexster Graph DB	https://github.com/tinkerpop/rexster/wiki
Bulbflow Python API	http://bulbflow.com/ pip install bulbs
Gephi	https://gephi.github.io/ Be sure to get the "Give Colors To Nodes" and "Graph Streaming" Plugins!

FIRST TRY: GRAPH ALL THE THINGZ!!
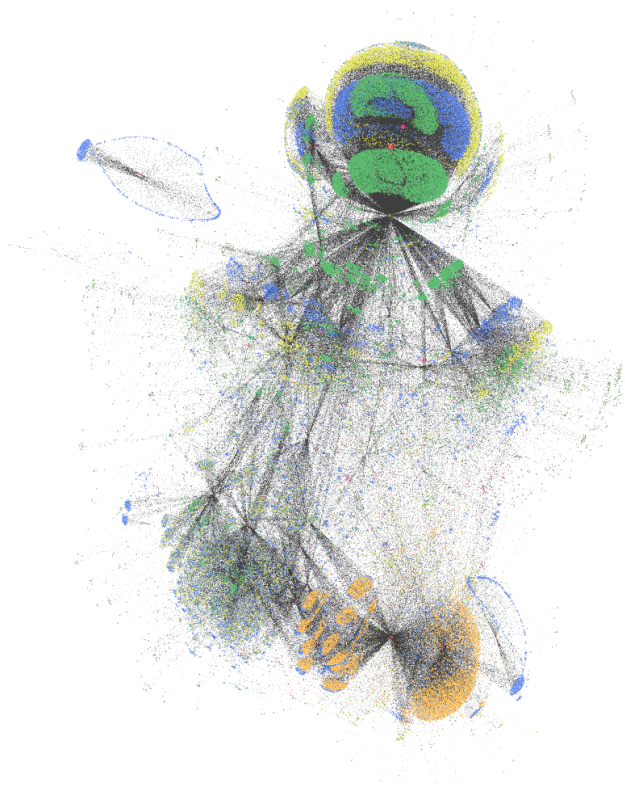
Nodes are color coded, so you can begin to see a few hints based on colors and structures.

There are some obvious hubs of activities, some strongly associated with certain colors.

This gets messy quickly! Best to restrict it to a specific network sensor, subnet, types of nodes, etc.

Graphing multiple node types against each other is often interesting.
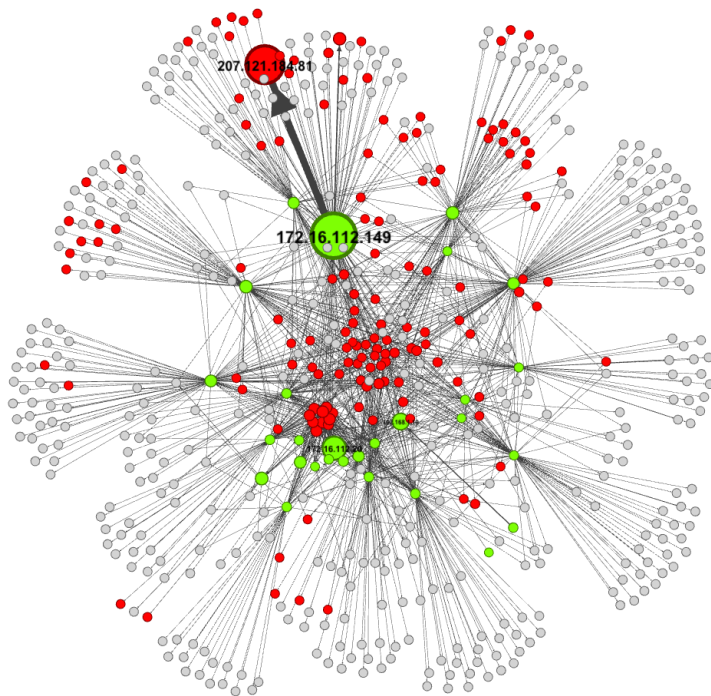
JUST THE HOSTS
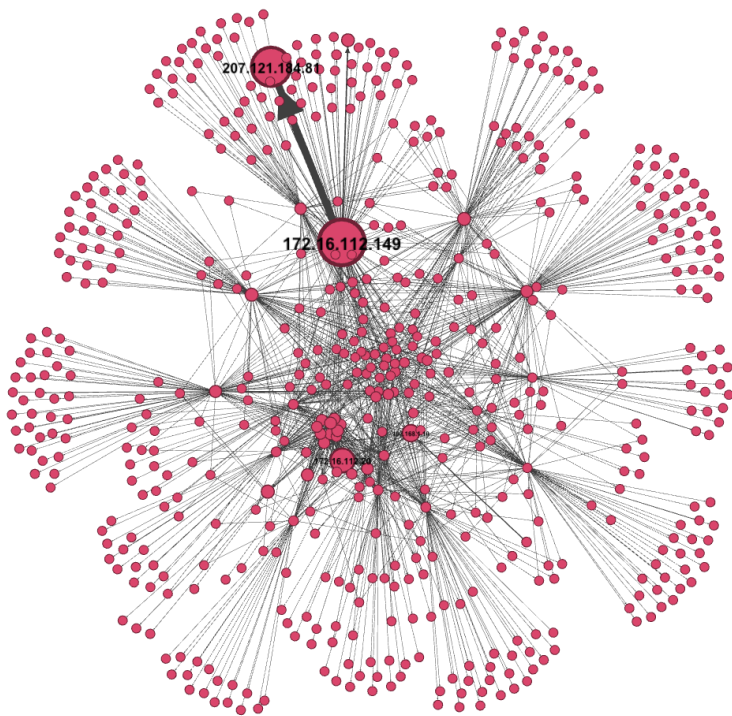
Interesting features start to appear!

Nodes are hosts present in your logs. Edges denote some sort of connection. Sizes denote rank.

See those two big hosts with the fat edge between them? What's that about?

All the hosts are the same color, though. Can we show the local vs. the remote hosts?

STILL HOSTS, BUT MORE CONTEXT

Bro tells us which hosts it knows are local (green), which it knows are not (red). Anything else is unknown (grey) but mostly not local.

Those big two hosts? They tell a bit more of a story now, don't they?

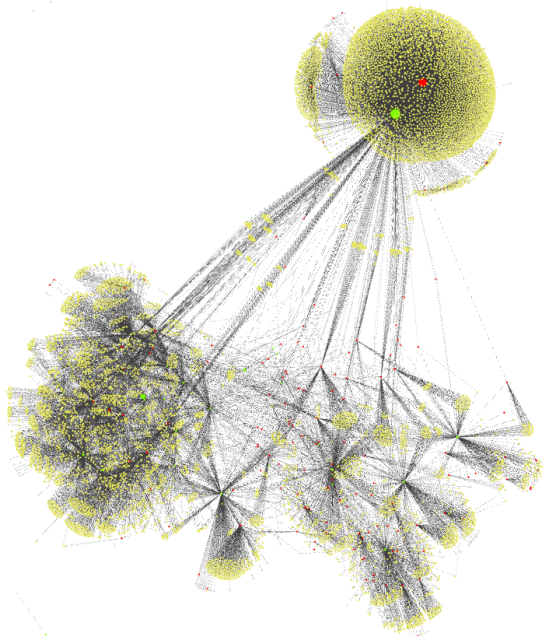There are a *lot* of connections from the 172.16.112.149 system to that 207.121.184.81 Internet host.

Maybe check that one out first.

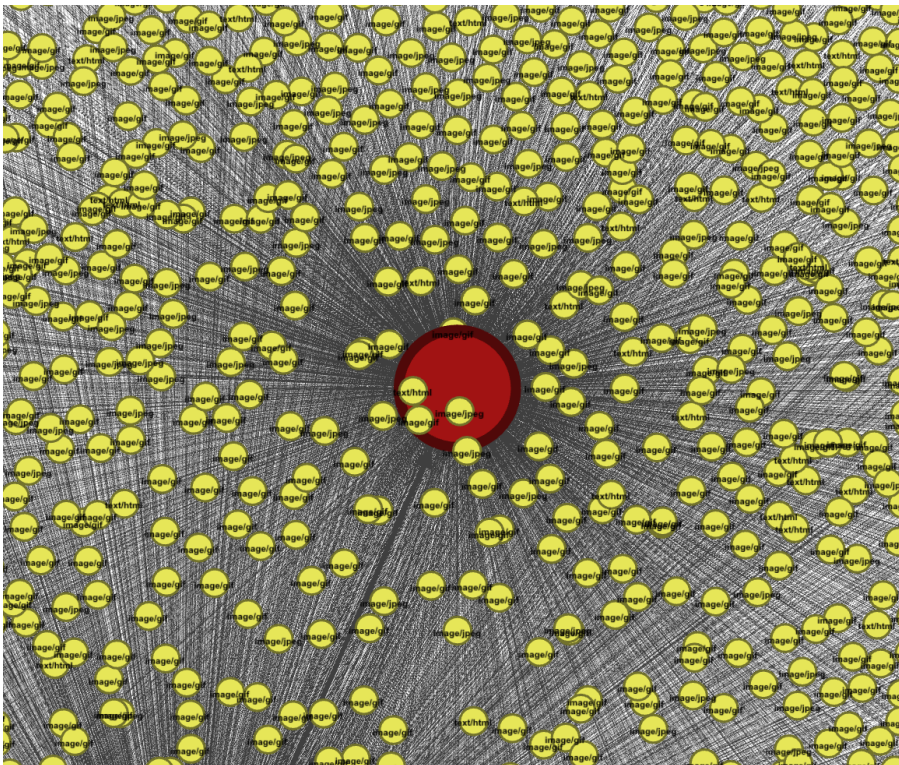# EXAMINING HOSTS & FILES

Adding file nodes to the graph also shows some interesting relationships.

Those same two hosts now make a dandelion shape.

What are those files?

EXTREME CLOSEUP

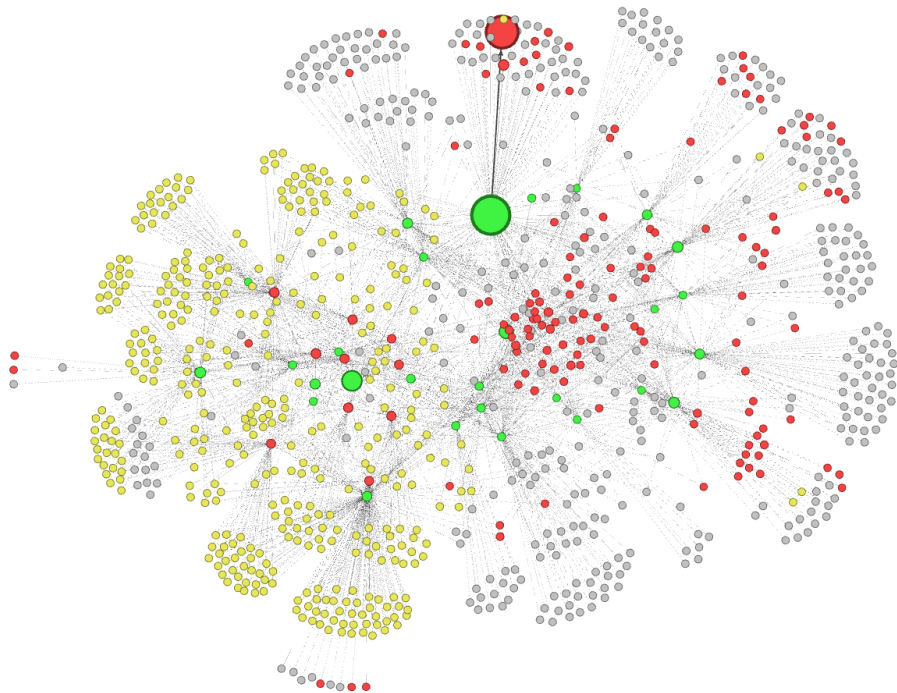Zooming in starts to make things more clear.

Lots of images, a few HTML pages…

This is probably all web traffic!

The thick "connectedTo" edge shows lots of HTTP transactions initiated by the internal node.

Directions on the files show they are responses from the server.

FILTER TO SIMPLIFY THE GRAPH

Normal web traffic in this graph is highly likely to be legitimate, so filter it out.
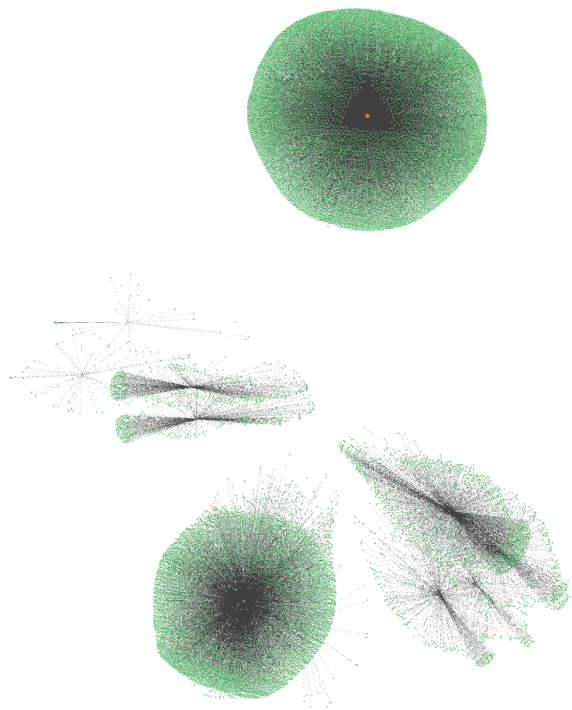
What's left is **much** simpler.

We don't have time for a full investigation here, but follow the same process:

>Dig into some of those clusters
>Filter out the known good

If there's anything left, it's pretty suspicious!

BONUS: USER AGENT EXPLORATION

Green nodes are individual HTTP transactions. Brown ones are specific HTTP User-Agent strings.

In theory, most users have similar computers & software, so most will have similar UAs.

We expect to see a few big groups. It's the small groups you want to focus most on (unless you think you have a big malware problem).

CONCLUSION

Securely explore your data

QUESTIONS?

David J. Bianco

dbianco@sqrrl.com

@DavidJBianco