Challenge 10 - Attack Visualization
The Honeynet Project / Forensic Challenge 2011 / 2011-12-18

Fabian Fischer  |  Data Analysis and Visualization Group  |  University of Konstanz
I wanted to visualize the log files with an visualization application written in Java developed for real-time event processing.

To make use of this tool, we need some preprocessing:

- Normalize the timestamps to ISO format using Python scripting.
- Merge all log files into one chronological file using Bash.

```
fischer@fischer-ubuntu:/mnt/fischer/sanitized_log$ ./convert.sh
Importing and normalizing timestamps auth.log... -> sanitized_events.tmp.auth.log
Importing and normalizing timestamps daemon.log... -> sanitized_events.tmp.daemon.log
Importing and normalizing timestamps debug... -> sanitized_events.tmp.debug
Importing and normalizing timestamps dpkg.log... -> sanitized_events.tmp.dpkg.log
Importing and normalizing timestamps kern.log... -> sanitized_events.tmp.kern.log
Importing and normalizing timestamps messages... -> sanitized_events.tmp.messages
Importing and normalizing timestamps secure... -> sanitized_events.tmp.secure
Importing and normalizing timestamps user.log... -> sanitized_events.tmp.user.log
Importing and normalizing timestamps www-access.log... -> sanitized_events.tmp.www-access.log
Importing and normalizing timestamps www-error.log... -> sanitized_events.tmp.www-error.log
Importing and normalizing timestamps www-media.log... -> sanitized_events.tmp.www-media.log
Sorting the output file based on timestamps... -> sanitized_events.log
Removing temporary files...
```
Analyzing and Importing the Events

- All log lines (events) start with a proper timestamp after preprocessing.
- Let’s write an analyzer module for our Java application:
  - Read events from text files.
  - Extract fields for hostname, process,...
  - Extract all IP addresses and make cached geo lookups.
  - Apply scoring* for well-known errors & risks based on regular expressions and weighted average for similar messages.
  - Create Java objects to represent events.
  - Forward the events to the visual analytics tool for interactive exploration.

```
fischer@fischer-ubuntu:/mnt/fischer/sanitized_log$ tail -5 sanitized_events.log

2010-05-02T23:11:01 app-1 CRON[5319]: pam_unix(cron:session): session closed for user root
2010-05-02T23:11:02 app-1 CRON[5316]: pam_unix(cron:session): session closed for user root
2010-05-02T23:11:13 app-1 sudo:   user1 : TTY=tty1 ; PWD=/var ; USER=root ; COMMAND=/bin/tar -cpzf log.tar.gz log
2010-05-02T23:11:13 app-1 sudo: pam_unix(sudo:session): session opened for user root by user1(uid=0)
2010-05-02T23:11:13 app-1 sudo: pam_unix(sudo:session): session closed for user root
```

* in a real scenario we would have more data for peak, trend and pattern detection on the event stream.
Event Visualizer

• The *Event Visualizer* is a visual analytics tool.

• It is a loosely coupled modular system for collecting, processing, analyzing and visualizing dynamic real-time event data streams.
Explanations – Real-Time System

• I was interested in the real-time use case, therefore I decided to stream the available log dumps into the system in real-time.

• Color does reflect the score, based on rules and simple "anomaly".

• Many observations can be made during visual exploration.
Visualization is updated in Real-Time.
(Log data is treated as "live stream")

Interesting Observation:
No log data for 2010-03-17.

Events on 2010-03-16

Events on 2010-03-18
Explanations – Layout Algorithm

• The space available for each day is fixed, which is important for real-time analysis to have a static non-changing layout.

• If the number of events per day exceeds the available space, less important (lower score) events will be removed in favor of more important ones. The number of removed events is shown as label (e.g., "+296 events).
Interaction with the Timeline

Selection of an event in timeline, will...

- enlarge the selected event.
- show the tooltip with the full message.
- highlight other occurrences of this event in the timeline.
- draw connecting lines to easily spot occurrences on long timelines located outside the visible region.
- show the Event Details with a line chart for this event type.
- will extract IP addresses and will search all events and will highlight them in black.

Interesting Observation: Configuration Problem? Binding problems of SSHD! (several occurrences)
Several Aligned Timeline Visualizations for Different Services

**Different Services/Programs:**
Each service has its own timeline. This reduces the number of changes and gives us more time for exploration. "Successful Logins" for example have their own timeline, because we are often interested in these.
Explanations – Panning & Zooming

• The visualization is drawn to a Zoomable User Interface, which provides zooming & panning capabilities.

• While the data is incrementally loaded, we can start to explore the different timelines, by panning & zooming around.

• New data will be added on the right.
Interesting Observation:
MySQL does complain several times about accounts without password.
Interesting Observation:
There is much log data missing. We should definitely check our remote syslog server (not mentioned by FC5 winners). Maybe the attackers tried to wipe their traces?
Interesting Observation:
High number of CMOS error messages on 2010-04-14.
Brute-Force Attacks from around the World
Where do successful logins come from?

Which are the most popular non-existing user names the attackers tried to login with?
Drill-Down and *Area Selection* loads all logins from the selected area...
Interesting Observation:
The IP 219.160.161.20 has successfully logged in, but it has hundreds of failed logins.
Investigation of successful brute-force attacks
Automatic Highlighting of events with same IP address helps...

Interesting Observation:
Events marked with black lines are from IP 219.150.161.20.
→ This IP has many brute-force attempts!
→ We have successful logins.
→ SUCCESSFUL COMPROMISED.
→ The IP 219.150.161.20 is a successful brute-force attacker.
Is 76.191.195.140 legitimate user or attacker?

Interesting Observation:
Only 1 failed login from this IP address. Now we see, that the connecting lines make sense! If there were other login attempts on the days before or after, a horizontal line would give us the hint!

→ The IP "76.191.195.140" is probably a legitimate user!
Interesting Observation:
222.66.204.246 is also a successful brute-force attacker!!!

How to improve the visualization?
The red rectangles are hard to distinguish.

→ Let's Apply shading to the visualization, to emphasize outlines (next slide)…
Interesting Observation:
222.66.204.246 is also a successful brute-force attacker!!!

Visual shading helps to distinguish events with the same scoring level (mapped to color).
Is 208.80.69.74 legitimate user or attacker?

Selected Event with IP 208.80.69.74

Interesting Observation:
No failed logins from this IP address, and we do have some legitimate web server requests.

➔ The IP is probably a legitimate user!
Is 121.11.66.70 legitimate user or attacker?

Selected Event having a high score and was a successful login.

→ IP 121.11.66.70 is a successful brute-force attacker!
Interesting Observation:
This IP address has some failed logins and does log in successfully to several different user accounts. We want to investigate these events further.

→ Load a word cloud and the raw messages (next slide).
Is 190.166.87.164 legitimate user or attacker?

Manual Inspection of events with this IP address.

Sadly it is still not clear, what is going on...
Is 190.166.87.164 legitimate user or attacker?

There was a new user generated which is called "dhg". If this was really done by an intruder (the responsible system administrator should know), the IP address 190.166.87.164 is probably an attacker, because he used that account to log in to the server by SSH.

Additionally in the user's home folder "psybnc" could be found. None of the FC5 winners mentions this IP address as intruder, so I could be wrong. However, it would actually make sense, that a real attacker does not use the IP address, which he uses for large brute-force attacks, to actually log in to the system!

⇒ 190.166.87.164 could still be a legitimate user.

⇒ 190.166.87.164 is an attacker if "dhg" is not an official user account.
Is 222.169.224.197 legitimate user or attacker?

Selected event has a high score and was a successful login.

→ The word cloud is obvious. This was a successful brute-force attack!
Is 122.226.202.12 legitimate user or attacker?

Selected Event having a high score and was a successful login.

"root" was added to ignore list.

→ Successful brute-force attacks. Most attempts to account "root", "oracle" and "guest".
Is 61.168.227.12 legitimate user or attacker?

Selected Event having a high score and was a successful login.

→ Successful brute-force attack from 61.168.227.12.
Changing the firewall settings is probably evil…

Several sudo commands changing the firewall settings (opening several ports).

Later it seems to be that someone configured a http proxy. The iptable changes might have been the preparation steps (not mentioned by FC5 winners).
Interesting patterns (orange, light green, green) occur after a successful break-in.

Log messages created by an installed IRC bot!
At this point you should switch to a textual search. Because visualization SHOULD NOT replace traditional workflows completely. It should support and enhance the analysis task!
Switching from Visual Representation to Textual Data

Searching the data reveals, when the bot has been installed.
More Interesting Patterns...

More Interesting patterns (orange, orange, green, green).

→ This pattern does occur for a while, but is not present on all days. Based on the available data the reason cannot be identified. This was not mentioned by FC5 winners...
Many segfaults for SSHD. Reason is unclear...

There are much more interesting messages which can be interactively and visually explored.
Visualizing the Category for "User Changes" does reveal newly generated user accounts.

Those accounts do have UID=0 and have root permissions (not mentioned by FC5 winners).

2010-04-19 22:38:00 - app-1-user - useradd - new user: name=packet, UID=0, GID=0, home=/home/packet, shell=/bin/sh
2010-04-25 10:41:44 - app-1-user - useradd - new user: name=fido, UID=0, GID=1004, home=/home/fido, shell=/bin/sh
The Challenge

Design and build a visualization that describes the attacks that were analyzed in FC5.

As an example, the visualization may have a geographic element, represented as a map, link graphs, histogram, or parallel coordinates, that sheds light on the following:

(1) Where the attacks came from?
(2) The volumes of attacks originating from various locations?
(3) The success or failure of these attacks?
(4) The nature of the attacks (“primary” and “secondary” phases)?
(5) Can the attacks be color coded to describe groups of attacks/attackers?

Use external data sources such as the many freely available geomapping databases.

The output can be anything that you like - from a still image, to interactive flash/java, dynamically updating, dashboard style, magazine infographic, holograms are also accepted.
Where do attacks come from? (10 points)

Does only make sense in the interactive prototype, because the map can be explored.
What is the most prolific attack? (5 points)

- Shown in the previous slides.
- All successful brute-force attacks are critical, because programs have been installed! The machine is compromised!
Which attacks were successful and which failed? (5 points)

• Shown in the previous slides.
What assumptions were made and what was the reasoning? Don't be afraid to make assumptions! (5 points)

• Shown in the previous slides.
What are the limitations of the visualization? (5 points)

• Already slightly discussed in the previous slides.
• The visualization will not show all data in time intervals with too many items.
• The time intervals can be changed from "days" to "hours", if there is too much incoming data.
How could you improve the visualization if given more time and resources - e.g. on a future GSOC project? (2 points)

- First of all, I would provide a downloadable prototype. I had not enough time to modify the software, so that it could be published as open-source for public use.
- The intended setup is to make all analytics on the server side and not on the client machine running the visualizer. To solve the challenge, I ignored this recommended setup and included everything into the visualizer tool. This lead to scalability issues.
- To solve the challenge I kept the data in memory and haven't used MongoDB, which I use for real scenarios with the tool.
- The implementation of the visualizations should be heavily improved for larger amounts of data. Integration of OpenGL based visualization would increase scalability.
- If this would be a long-term project I would try to move the visual interface using the time line and scoring to a web-based system.
- I would probably try to connect to data from data sources stored in Loggly or Graylog2.
Provide a description of the toolsets and scripts used (10 points)

• Many things have already been discussed.
• Custom scripts and a Java program used.
• The prototype is not available for public-use right now, because the framework is submitted to an event/data streaming conference for next year.
Aesthetic appeal and ability to hold the subject's attention (5 points)

- The color scale can be changed.
- The used default color scale ("green – yellow – red") is not the best choice for all users.
- But it does represent the intuitive "importance" of an "red" event.
- The timeline can be explored, even many items are still incoming. Therefore, the analyst has time to smoothly switch between exploration and monitoring.
- Visual shadows are used as improvements.
- Many improvements could be made, of course.
Interactivity, e.g. the ability to drill down, explore, or zoom in on events. (10 points)

- Fully functional.
- The main visualization is actually based on a zoomable user interface.
- Sadly, there was not enough time left to provide a video demonstration.
Animation, particularly based on a timeline. (10 points)

• An interactive real-time timeline is integrated, which preserves the ordering up to a certain degree.

• The intended usage is real-time monitoring. Therefore depending on the amount of data, it is not possible to show everything.

• At one point the analyst can only take care of the most important events and has to trust the underlying analytics.
Creating a visualization which uncovers any trends, observations or artifacts which were not described in the FC5 prize winning solutions. (20 points)

• Mentioned in the previous slides...
Creating a visualization that tells a story about the data set, threat environment, and the attack. (20 points)

• This visualizations do focus on visual exploration and not on visual presenting the data / attack to others.

• I had plans to include another type of visualization to visually "present" the attacks, but I started just a couple of days before the deadline, so time went out. ;-)
Thank you very much for your attention!