Hadoop Success Stories in Trend Micro SPN

Yun-Chian Cheng

October 2012

Agenda

- SPN Brief Introduction.
- Big Data In SPN.
- Solved Problems using Hadoop.
- Major Approaches.
- Summary.

SPN: Brief Introduction

- SPN: Smart Protection Network is a cloud-client Internet-based infrastructure designed to protect customers from all kinds of Internet related threats.
- The Services provided by SPN:

  - ERS: E-mail Reputation Service

  - FRS: File Reputation Service

  - WRS: Web Reputation Service
- Client/Customers also can send suspicious web pages, e-mails or files back to SPN backend cloud.
- Customers can be thought of as SPN’s sensors.
- Complicated things are done in the SPN cloud backbone.

SPN's Major Services

- 3 major services:
- ERS (Email Reputation Service).
  - Blocks unwanted spam e-mails
  - Identify dangerous phishing e-mails
- FRS (File Reputation Service).
  - detect malicious files, virus, trojans, worms,
  - cleanup or recovery.
- WRS (Web Reputation Service).
  - determine web page is safe to visit or not,
  - determine the category of the web pages.

Smart Protection Network

Sourcing | Processing & Analysis | Validate & Create Solution | Quality Assurance | Solution Distribution | Solution Adoption

File

Web / URL

Email

Domain

IP

File Reputation Service

Web Reputation Service

Email Reputation Service

Smart Protection

Customer

SPN Correlation

Community Intelligence (Feedback loop)

Big Data in E-Mail Security

- Spam Honeypot Networks
  - 50M e-mails per day,
  - Discards e-mails that can be detected
- Internet Hosted Message Service
  - It is a SaaS model,
  - 20M e-mails per day, 30GB logs per day
- ERS Mail Server Address Checking Service: 320M queries per day.

Big Data From Spam Filtering in The Cloud

- Customers redirect in-coming e-mails to Trend Micro's mail servers around the world, performs scanning to drop spams and then redirect normal e-mails to customers' mail servers.
- Each mail server maintains log files that records each smtp connection with some important data:

  - sending server's IP address
  - helo/ehlo string identifying the mail server
  - Unix time-stamp
  - number of recipients, etc
- This log data is really a gold mine to be dug.

Big Data in FRS

- Customers feed back suspicious files.
- Files extracted from e-mails or web pages collected in SPN backend.
- 3rd party exchanges.
- About 200 to 300K samples per day.

Big Data in WRS

- Crawled web pages that are requested by customers.
- Customer clicked logs that records each customer's click of a URL:
  - customer's id, IP address
  - the clicked page's complete URL
  - the resolved IP address that contains the web page
  - the reputation/category of the page
  - the access is blocked or not
  - 4B+ entries per day, 1TB data collected per day,

Problems Encountered in SPN page 1

- How to obtain simple but useful information from huge log data:

  - How many customers visited a newly identified malicious web pages and got infected? Who are they?

  - How many customers are protected by blocking the access to a malicious web page?

  - What are the ip addresses of the web or mail servers used by some legitimate companies like IBM, Citi Bank?

Problems Encountered in SPN page 2

- Can we identify the moment of spam attacks or when a known web site got compromised?
- How to determine the files sent back by customers are similar?
- How to determine the e-mails are similar?
- How can we reduce False-Positive rates for ERS and WRS?
- How do we determine the reputation of a mail server or a web server along with its associated IP address?

First Success: Using Hadoop For Basic Statistics

- How many customers are infected by visiting a malicious web page? How many customers are protected?
- A very simple one-iteration of Map-Reduce cycle solves such problem.
- This is the first success that we realize the power of Hadoop.

Second Success: Using Hadoop To Create Histograms

- Initial Goal: Create the histograms for the number of hits in a minutes of some companies web sites, so that we can determine the sudden surge point to identify there is a web site compromise.
- Similar goal for Email case: Create the histograms of hits in a minute from certain suspicious sources, like service providers. Sudden surge can be caused by newsletters or spams.
- The histograms can be computed from the web users clicked logs or e-mail servers logs using very simple Map-Reduce mechanism in Hadoop.

Problem: How to identify sudden surge on a Histogram?

- **Solution**: **Wavelet.**
- Histograms can be treated as time series in statistics or signals in EE.
- Wavelet can be used to identify spikes/surges or change points.
- Wavelet method gives better results than using the original histograms.
- Use same idea to collect histograms for e-mails sent from suspicious service providers
- Usually one single histogram can be handled easily using one CPU. However, if we have many histograms to handle, Hadoop is a good choice.

Problem: How Do We Group Similar Things Together?

- **Solution: Clustering.**
- Clustering is active topic for the last 50 years and will remain so for the next 50 years, because it is hard and important.
- Clustering is a crucial method to reduce the efforts of cpu and engineers efforts.
- Clusters can indicate occurrence of **unusual event**.
- There are open source code, like Apache's Mahout for some popular cloud based clustering algorithms, k-means, canopy, mean shift, etc.

Clustering Problem

- Trend Micro uses suffix tree clustering heavily to identify common substrings in many projects.
- The problem with suffix tree clustering is that it uses a lot of computer memory and can only be done on a single server. It is not easy to do it on Hadoop.
- Many clustering algorithms like k-means or canopy requires a lot of computing of Euclidean distance, which can be time consuming.
- K-means or canopy algorithms need to specify some parameters that can drastically alter the results.

3rd Success: Combine Wavelets and Clustering

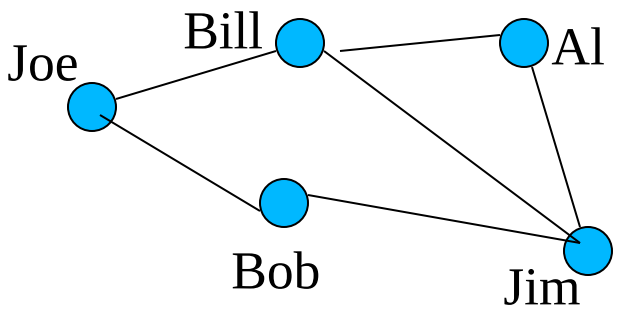
- Suffix tree clustering cannot handle the whole days' mail samples.
- We use wavelet to identify the starting point of sudden surge of histograms using wavelets.
- We can only do the clustering of samples received within X minutes after the surge.
- The clusters can then be determined for new spams.

Trend Micro And Graph Mining

- Graph Mining.
- Bipartite Graph.
- Large/Maximal cliques in bipartite graphs.
- Community Detection, Centrality in Social Networks.

Graphs and Graph Mining

- Graph is a very simple mathematical object that can be used to represent many real world problems.
- Mathematically, a **graph** G = (V,E) consists of 2 sets V and E where V is a finite set of something and E is a set of unordered pairs $\{ (v_1,v_2) \mid v_1 \neq v_2 \text{ in } V\}$. Here V is called the **Vertex** or **Node** Set and E is called the **Edge** or **Link** Set. Here an edge (v_1,v_2) has 2 end points v_1,v_2; v_1, and v_2 are connected.
- Examples: Friendship graph



V= {Joe, Jim, Bill, Al, Bob}

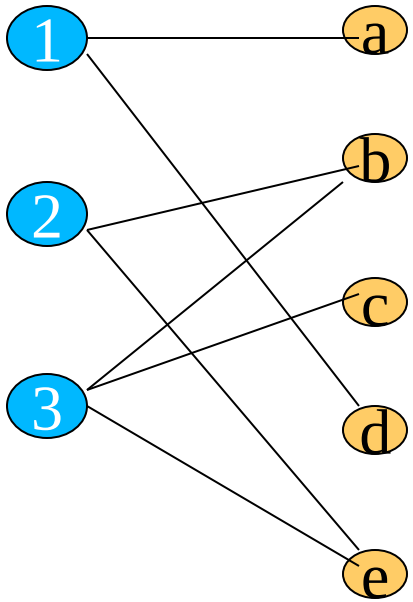
E = {(Joe,Bill), (Joe,Bob), (Bob,Jim), (Bill,Jim), (Al,Jim)}

Graph Mining

- Graph mining is a special type of data mining with graphs as objects of interests.
- We are interested in finding some frequent graph patterns like dense subgraphs, or finding a subgraph that is isomorphic to a given graph.
- Although typical data mining algorithms can be used, more efficient algorithms may be available in graph mining.
- Large scale graph mining should be credited to Google that developed its proprietary Map Reduce programming model with Google File System and Big Table technology to give web page ranking on huge web graph with billions of vertices.
- Graph mining is already widely used in Social Networks services such as Facebook, LinkedIn, Twitter.

Bipartite Graph

- Bipartite graph is a special type of graph whose Vertex Set consists of two disjoint subset and each edge has two endpoints must be on different subsets.
- Example: Set-Element, 3 sets $s_1 = \{a, d\}$, $s_2 = \{b, e\}$, $s_3 = \{b, c, e\}$,



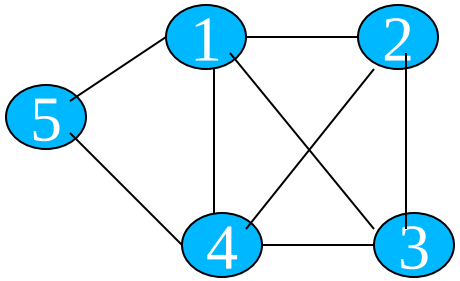
Why Bipartite Graph Is Useful?

- Many real world problems can be modeled.
- Frequent Itemset Mining: Transaction Set vs Item Set.
- Information Retrieval: Documents vs words.
- Anti-Malware: Suspicious files vs features.
- Domain Reputation: Domains vs IP addresses,
- And More!

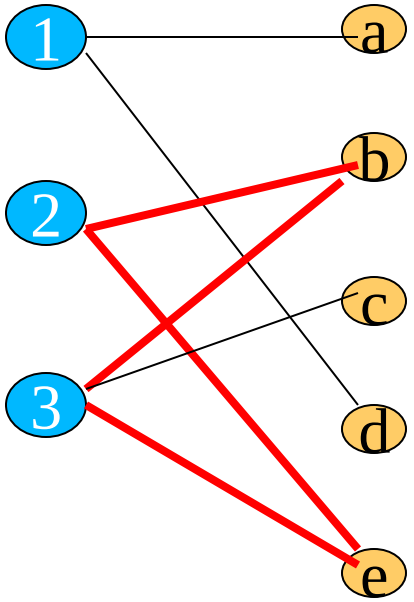
- Bipartite Graph data can be manipulated easily in Hadoop.

Cliques and Bicliques

- A **clique** is a **complete** subgraph whose vertices are connected to each other.
- A **biclique** is a **complete** subgraph in a bipartite graph in which each vertex is connected to all vertices on the other side.



4-clique: {1,2,3,4}
3-clique: {1,2,3}, {1,2,4},
{1,3,4}, {2,3,4}
(1,4,5}



Why Biclique Is Useful?

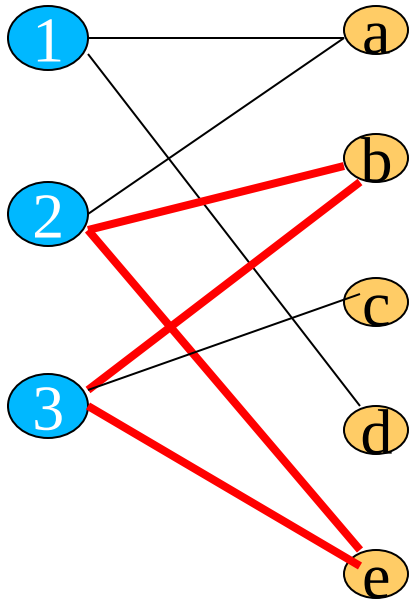
- Biclique can be used to find a set of transactions that contain common items.
- Similar ideas are found in anti-malware case that we can use bicliques to get similar clusters of files that contain common features.
- Divide-And-Conquer: Each biclique may generate a cluster with size suitable for a single machine.

Why Biclique Is Useful?

In this Set-Element graph,

nodes {2, 3, b, e} form a 2 by 2 clique.

Sets s_2 = {a, b, e} and s_3 = {b, c, e} has 2 elements in common {b, e}



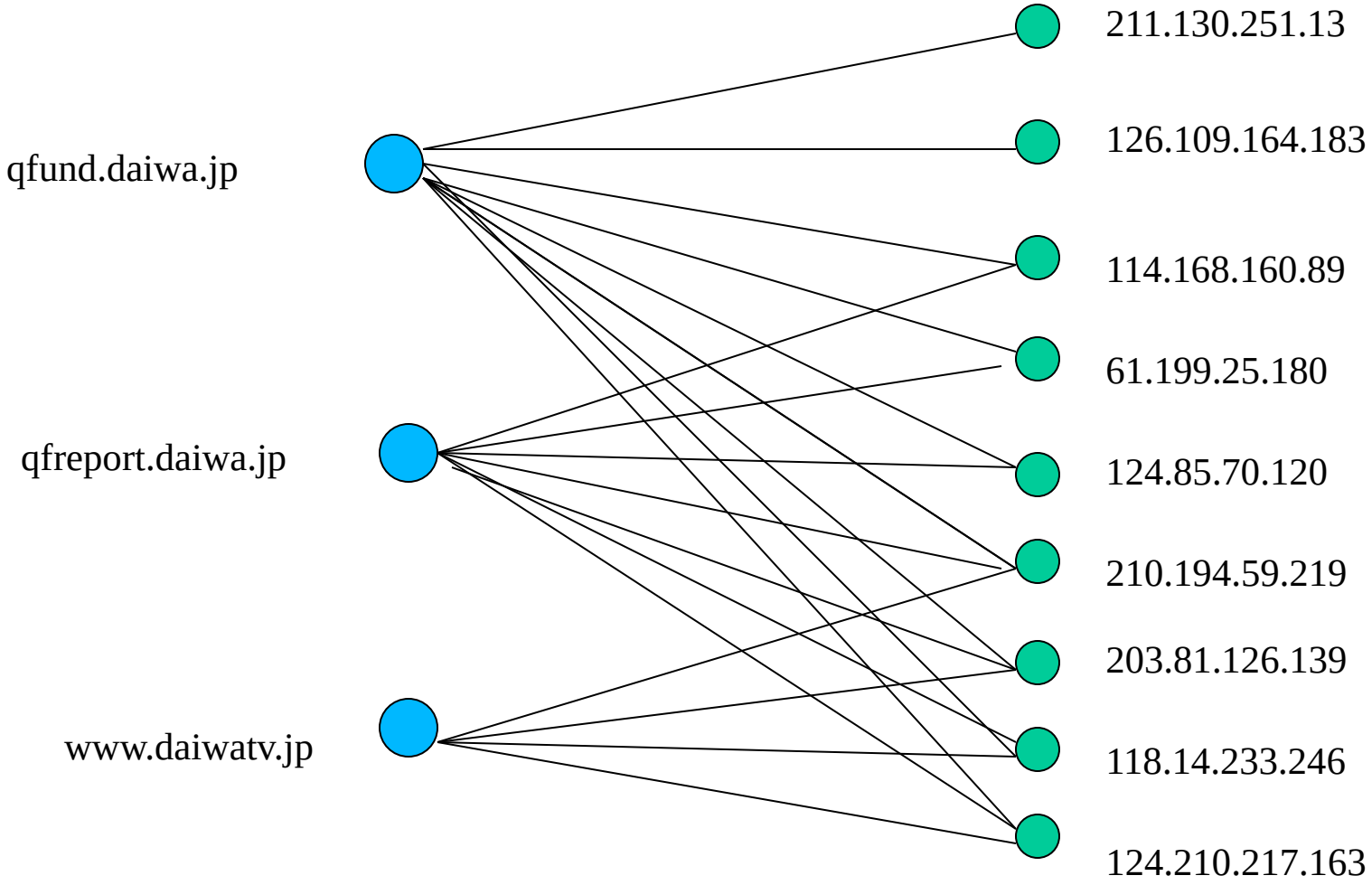
4th Success: Using Hadoop To Build White List

- Simple Map-Reduce can construct a pair of domain and the list IP addresses of web or mail servers resolved.
- Do this every hour or everyday and store the results.
- For legitimate companies the association is very stable. The white list can be obtained easily.

Web Servers of Google Cache Servers

Domain	IP address
Webcache.googleusercontent.com Data collected for 4 days	72.14.203.132; 74.14.213.132; 74.125.67.132; 74.125.71.132; 74.125.91.132; 74.125.93.132; 74.125.65.132; 74.125.19.132; 74.125.157.132; 64.223.169.132; 64.233.189.132 66.102.11.132 209.85.129.132; 209.85.225.132; 209.85.227.132; 209.85.229.132 74.125.153.132 (3 days)

Example: Daiwa



5th Success: Using Bicliques To Get Reputation Score

- Look at bicliques of domains and server IP addresses over time.
- Look at these bicliques associated with good domains have very stable server IP address pool.
- Internet Service Providers, and Content Distribution Network providers have large, but stable bicliques.
- Bad domains using fast flux have fast changing bicliques.
- We can use the stability of the bicliques over time to determine the reputation of either domain or server IP address.

Summary

- Hadoop is good for computing statistics in huge log files.
- Bipartite cliques can be used in find white-list, black-list.
- Bipartite cliques can be used to find frequent item sets.
- More graph mining can be used to SPN's big data.
- New platforms such as Giraph may be explored.
- **Data Mining/Graph Mining is a continuous effort!**