A text mining analysis of the National Vulnerability Database

Hen Su Choi Ortiz

April 11, 2015

1 Introduction

The National Vulnerability Database is the U.S. government repository of standards-based vulnerability management data represented using the Security Content Automation Protocol (SCAP).

The Common Vulnerabilities and Exposures (CVE) system provides a reference-method for publicly known information-security vulnerabilities and exposures. It is funded by the National Cyber Security Division of the United States Department of Homeland Security. A public CVE database is available online.

2 Methodology

There are several fields within the CVE database. One of the fields is the "description" that contains is a standardized text description of the issue(s). Some examples of descriptions are:
"Adobe Digital Editions 2.0.1 allows attackers to execute arbitrary code or cause a denial of service (memory corruption and application crash) via unspecified vectors."
"Cisco 9900 Unified IP phones allow remote attackers to cause a denial of service (unregistration) via a crafted SIP header, aka Bug ID CSCul24898"
The data is messy, so it was first cleaned using Google refine. Once the data is clean, an extensive analysis is performed, including data mining and machine learning using R. Tableau and R are the tools used for the visualizations.

3 General Data Analysis

The dataset contains reports of vulnerabilities from 1999 to 2015. A histogram of the data, shows the frequency density of reports over the years (fig. 1)

In this work, we are going to consider the previous year (2014) for the analysis. Figure 2 shows a histogram of most relevant terms for the aforementioned year.

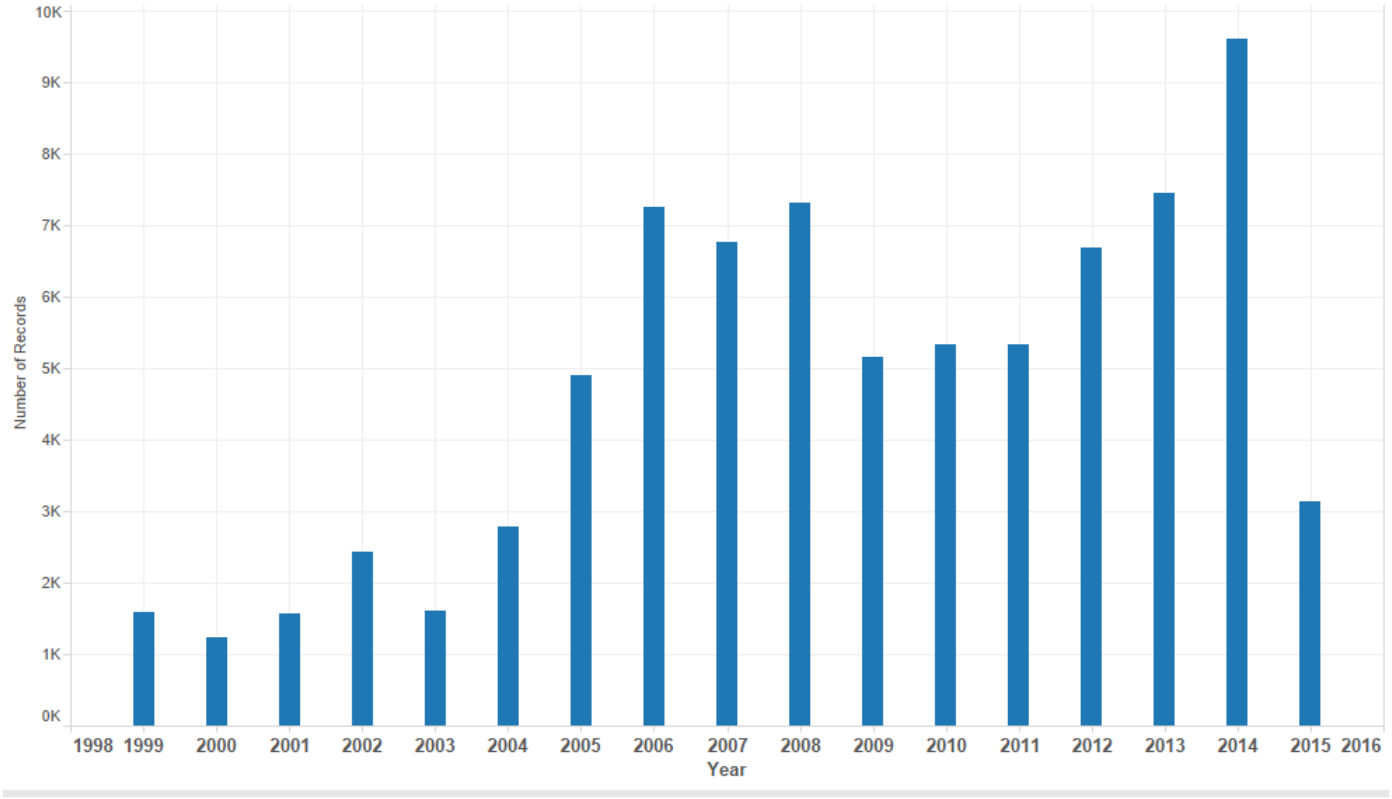

Figure 1: Vulnerabilities by year

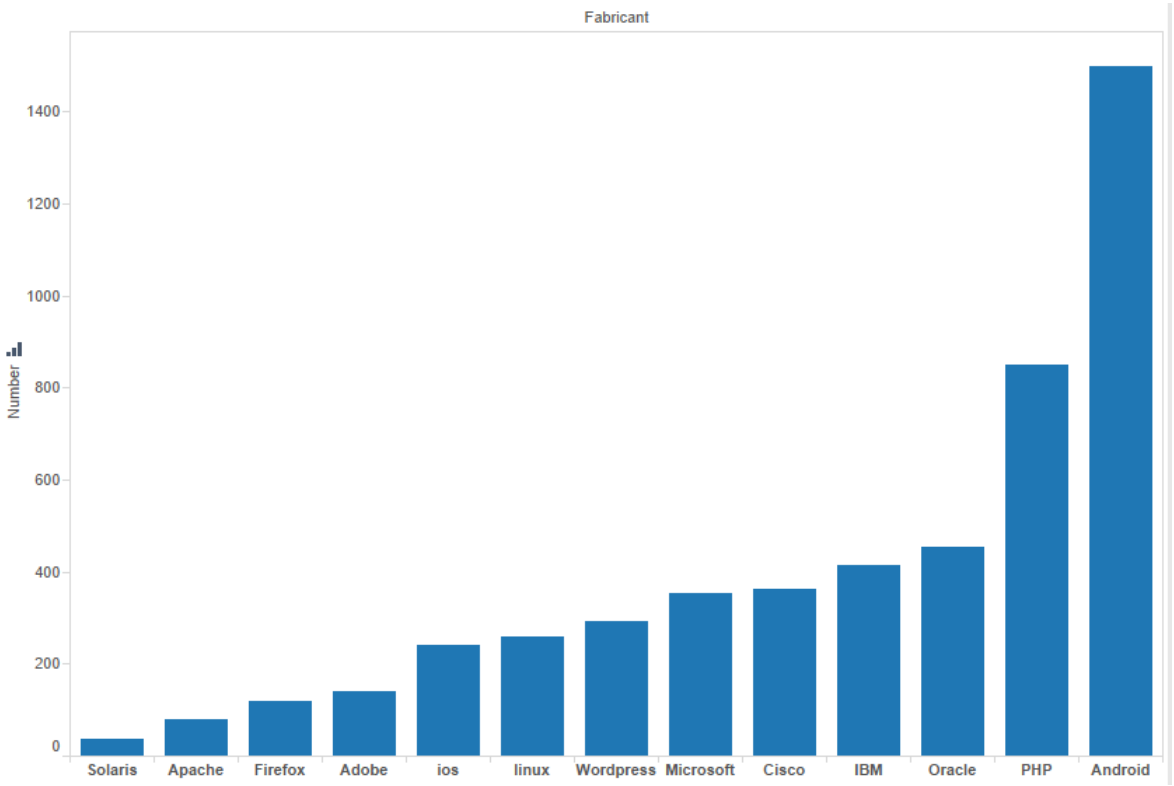

Figure 2: Analysis by platform in 2014

4 Machine learning

4.1 Clustering

A correlation analysis is performed trying to find clusters of data. The **hclust** function in R uses a method to cluster the data in a hierarchical way . This method defines the cluster distance between two clusters to be the maximum distance between their individual components. At every stage of the algortihm process, the two nearest clusters form a new cluster. The process is repeated in loop until the whole data set is agglomerated into one single cluster. Using R we created a dendrogram in order to illustrate how clusters of data are correlated. In this case, we limited our clustering algorithm to eight clusters(k=8), what can be used to understand how some vulnerabilities are related, and to predict future types of attacks.

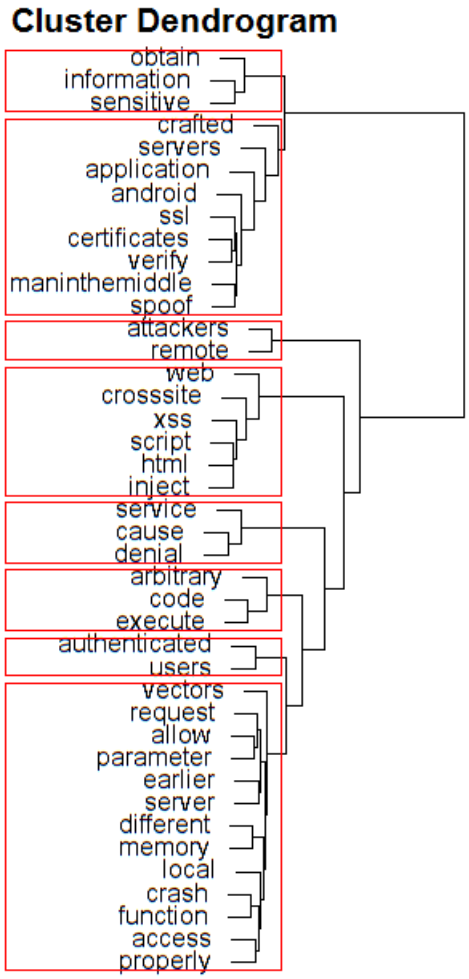

Figure 3: Word clustering

5 Text Mining by type of platform

Once we have an idea of how the whole data is related, we proceed to create subsets of the dataset, isolating relevant terms and visualizing the word frequency for each term. Word clouds are tools of visualization that can help to understand and see the differences between platforms, types of attacks and vulnerabilities.

5.1 Android

Figure 4: Word cloud: Android

5.2 PHP

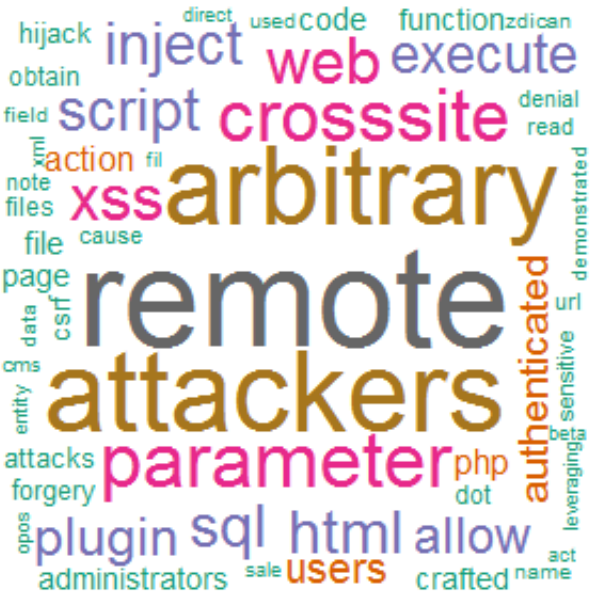

Figure 5: Word cloud: PHP

5.3 Adobe

Figure 6: Word cloud: Adobe

6 Text Mining by vulnerability

6.1 "Inject"

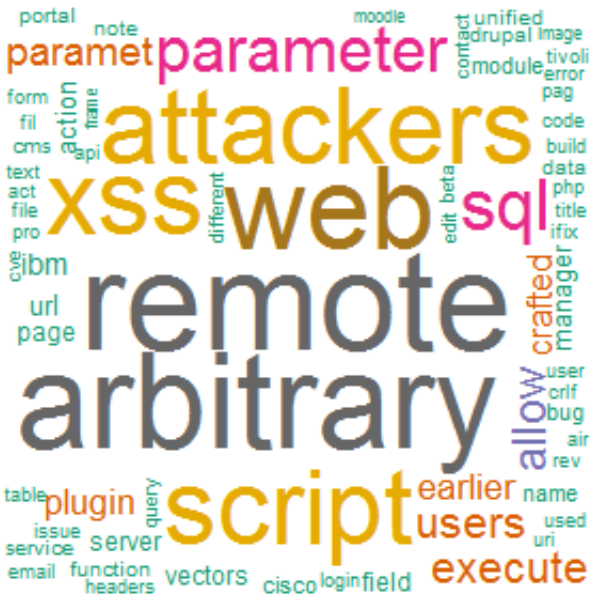

Figure 7: Word cloud: Inject

6.2 "Cross-site"

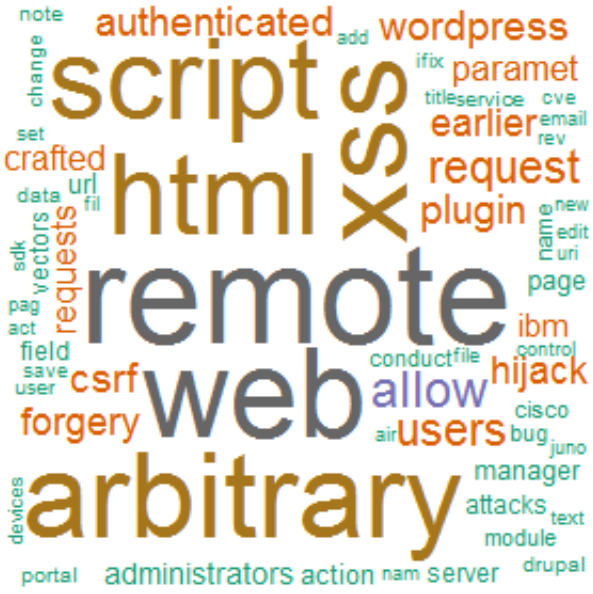

Figure 8: Word cloud: Cross-site

7 Conclusion

Several conclusions can be inferred from this analysis. We can expect a report of more than 10k vulnerabilities in 2015, being Android the platform with more incidents. From the cluster analysis, we can see that Android problems are strongly correlated with ssl certificates and man-in-the-middle attacks. Word cloud visualization is a powerful tool that allow users, industry and government to take appropriate measures analysing an specific platform or type or vulnerability.