Data Science Project

Ryan Heffernan

Instructor: Jim Byers

8/15/2016

Cyber security has become a significant problem for our industry, and our nation. Security researchers have demonstrated proof-of-concept hacks of ATMs, cars, and planes, in addition to the ubiquitous breaches of computers and cell phones. The NSA has reported that the Chinese military has the capability to shut down the U.S. power grid, and even classified networks, such as the Pentagon’s, have been breached by malicious hackers.

Each successful hack depends upon one or more vulnerabilities, which are security bugs in software that allow attackers to make systems behave in an unauthorized manner. Due to the size and complexity of modern software, these vulnerabilities are nearly impossible to completely eradicate. For example, the codebase for Windows 7 is composed of almost 40 million lines, and the Debian 5.0 codebase is over 65 million.

Because the elimination of vulnerabilities is impractical, reducing the risk of a security breach depends upon identifying the minority of vulnerabilities that most useful to attackers, and fixing those before the attackers can exploit them. While cyber security professionals have many rules of thumb for identifying the most dangerous vulnerabilities, surprisingly little rigorous analysis has been perform to ground these ideas in actual data.

For this project, I plan to utilize two data sets. First, the National Vulnerability Database, which is the most comprehensive public registry of vulnerabilities. Second, the VERIS Community Database, which is the most comprehensive public record of security incidents. By comparing the known universe of vulnerabilities with the known universe of successful attacks, I plan to shed light on what characteristics make a vulnerability more likely to be used by attackers. This should result in useful guidance for cyber security defenders for allocating their limited resources so they can fix the most critical vulnerabilities first.

My hypothesis is that attackers will favor vulnerabilities that offer them the most return on their investment. In other words, attackers will favor vulnerabilities where identification and exploitation are inexpensive, and where the access obtained is of high value relative to the attacker’s objectives.

The features I plan to use include Attack Vector, Attack Complexity, Privileges Required, Authentication Required, and User Interaction. These features are a proxy for the attacker’s cost. The lower their ratings, the easier for the attacker to exploit the vulnerability. To approximate value relative to the attacker’s objectives, I plan use the total number of records stolen.

It will be unsurprising if the evidence bolsters my hypothesis, since it merely presumes that attackers responder somewhat rationally to incentives. However, I also hope to determine what specific vulnerability characteristics correlate most directly with use by attackers. This is currently a matter of speculation and debate among cyber security professionals, and data on correlations, along with existing knowledge of cause and effect, may help tease out causal relationships. I plan to look at the data through several different models, including linear regression, logistic regression, and KNN, or order to find these relationships.

After performing this analysis on public data sets, I hope to turn it to the private data sets of my employer, Microsoft. Doing this may allow my employer to more effectively allocate it cyber security resources based on data from previous incidents.

So far, by biggest challenge has been cleaning the data. I was able to download both data sets in the form of CSV files using a web browser. I tried to examine and clean the data sets in Excel, but found them too large and messy to do so. I am now working on a Python script to clean the data.

If all goes as planned, this project should yield some useful analysis that will help cyber security defenders minimize risk to the users and stakeholders they aim to protect.