National Vulnerability Database Trend Analysis

Engineering Honors Thesis

by John Sullivan
Advisor: Professor Cybenko

Dartmouth College
Hanover, NH

May 2018

Table of Contents

Introduction

- This Honors Thesis is the product of three terms of research with the assistance of Professor George Cybenko.
- **The National Vulnerability Database (NVD)** is the U.S. government repository for vulnerability management data
- It was established by the National Institute of Standards and Technology (NIST), contains descriptions of 100,000 software vulnerabilities published since 1988.
- This thesis contributes to the understanding and value of the NVD with the goal of analyzing its data to improve computer security.

Definitions and NVD Statistics

Some Definitions

1. **Vulnerability** - A weakness in a system's security that could be exploited. For computers, this weakness may be a segment of code.
2. **Exploit** - Commands that allow a user to take advantage of a vulnerability and cause a computer to behave in an unintended way.

- Based on tables below, computer vulnerabilities are a growing issue and the NVD contains a wealth of information in this area.

NVD Overall Statistics (as of 12/10/17)

	No. Vulnerabilities
Total	98,183
Total (excluding missing entries)	93,295
Last 3 Months	4,037

Exploit DB Overall Statistics (as of 12/11/17)

	No. Exploits
Total	38,236
Mapped to NVD	9,267

Existing Research and Goal

Existing Research

- It looks at general trends in the NVD, but not at particular companies or at correlations between their vulnerability trends [1] [2].
- Other existing research also looks at applying topic modeling and machine learning techniques to data from the NVD [3] [4].

Goal

- By analyzing NVD vulnerability trends and descriptions, it is possible to determine which vulnerabilities are at high-risk of being exploited.

Honors Thesis Objectives

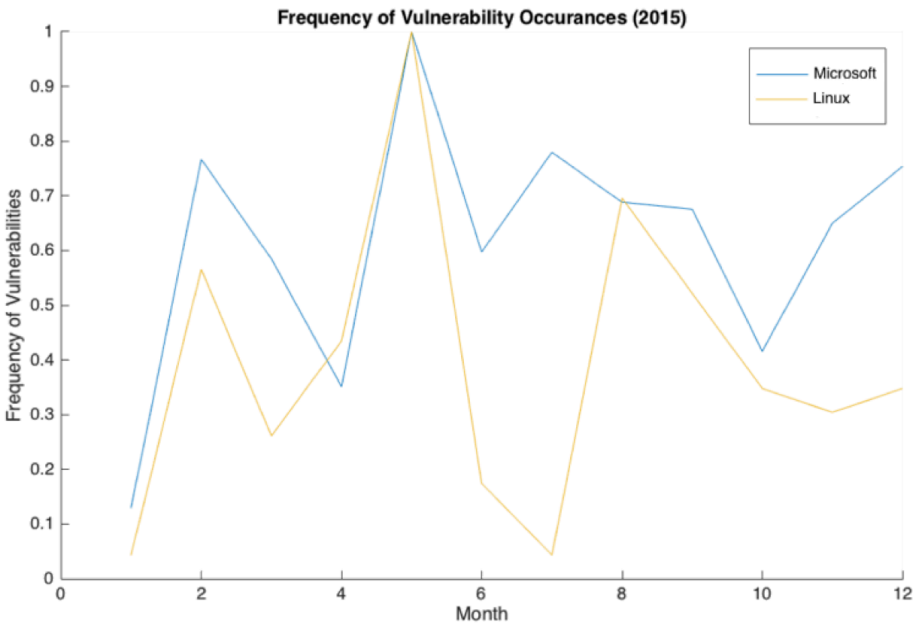
For this Honors Thesis, the findings are presented in the context of existing research. Accordingly, this thesis is divided into three sections:

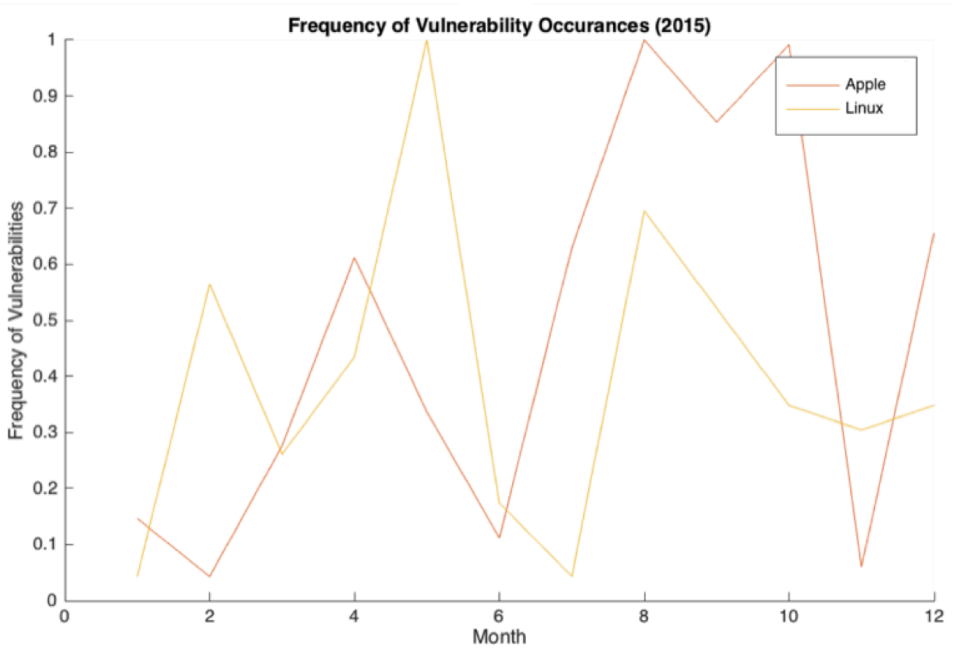
1. **Background** - The relationships between the trends in vulnerabilities for many commercial organizations and their various products, rather than examining such organizations on their own.
2. **Topic Modeling** - Correlations of trends between vulnerabilities are explored through analysis of NVD descriptions and topic modeling.
3. **Machine Learning Models** - Machine learning models and techniques are used to analyze the predictive power of textual descriptions to determine whether a vulnerability has been exploited.

Table of Contents

Trends Between Companies

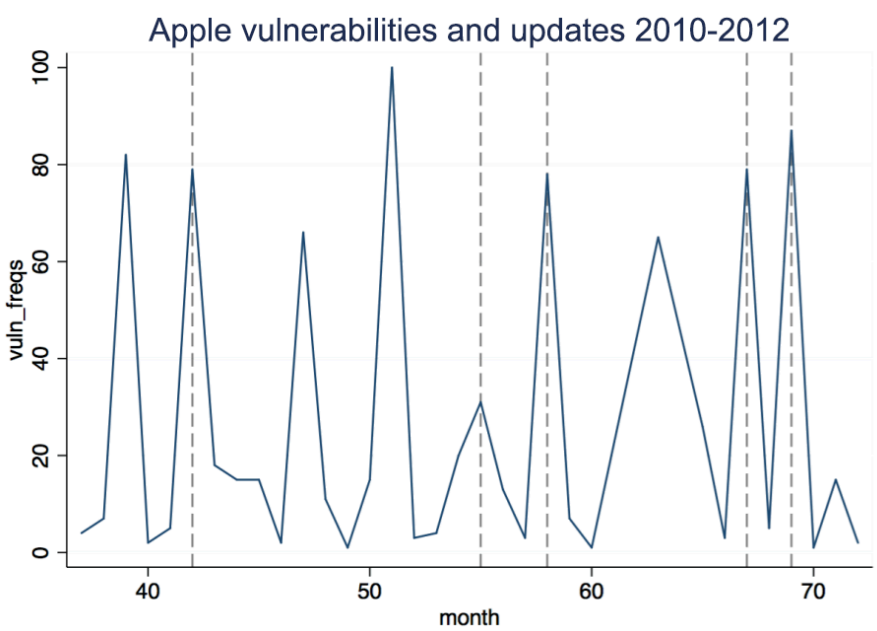
- Comparing specific companies, some had a statistically significant relationship in vulnerability frequency (Microsoft and Linux on left).
- Others did not show these patterns (Apple and Linux on right).
- One goal of this research was to understand such trends.

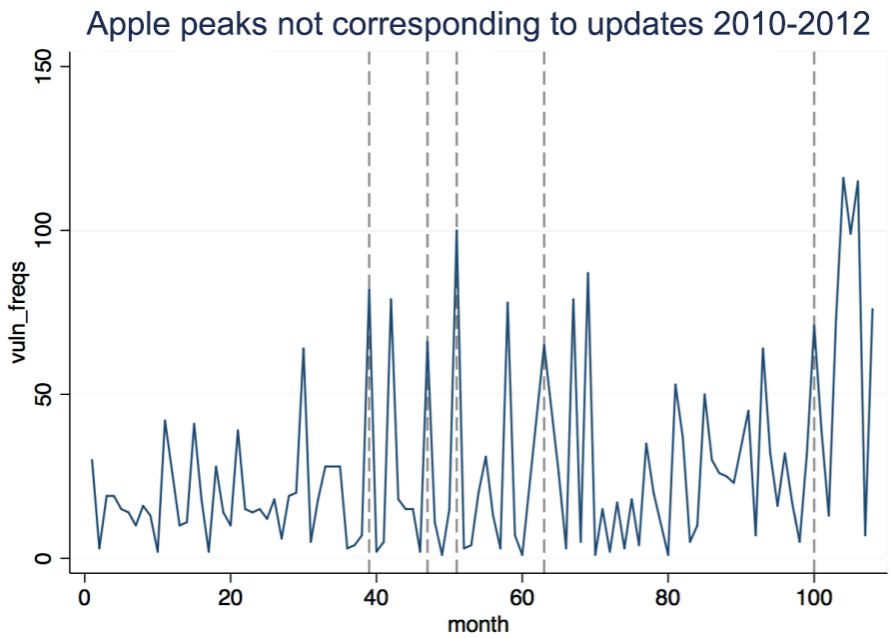




Apple Updates

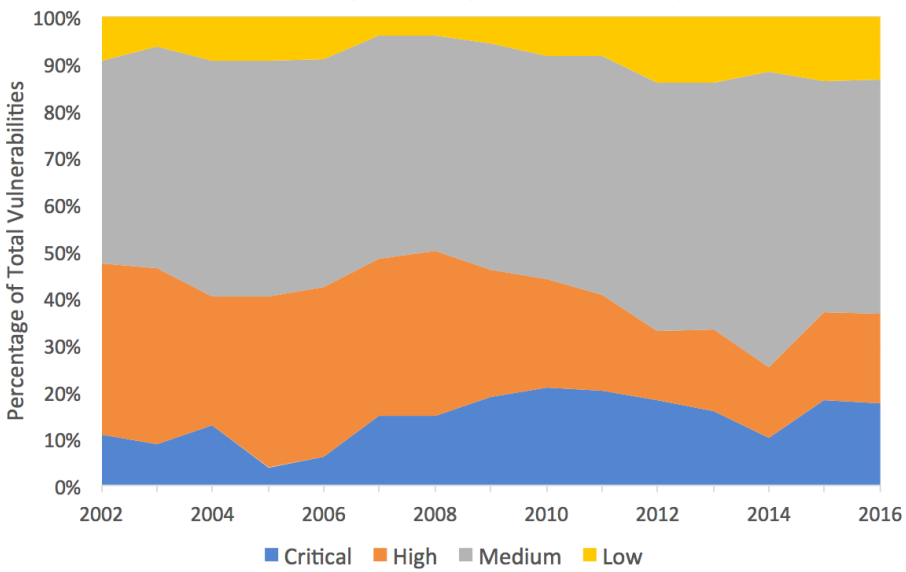
- Apple exhibits an increased number of vulnerabilities around the time of software updates.
- Those peaks that do not correspond to software updates are generally due to new product releases.





Severity and Operating System Distributions

- This data analysis was used as part of a paper with Kate Farris [5].
- For vulnerability survival analysis, it is important to understand the vulnerability distribution using different criteria.
- Over the years, despite the number of vulnerabilities increasing, the distribution by severity has been consistent.
- The distribution by operating system has fluctuated.



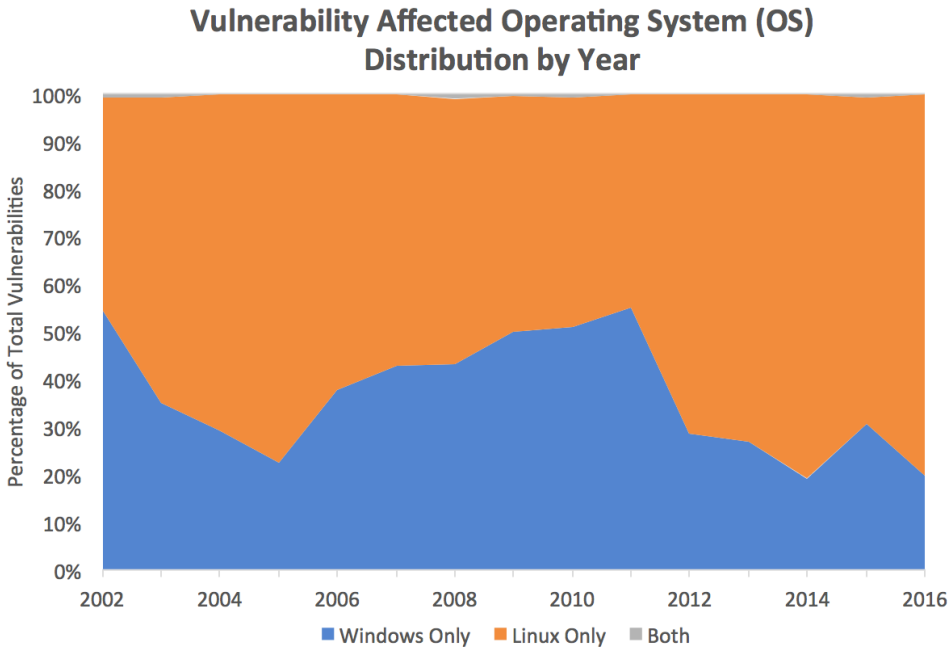


Table of Contents

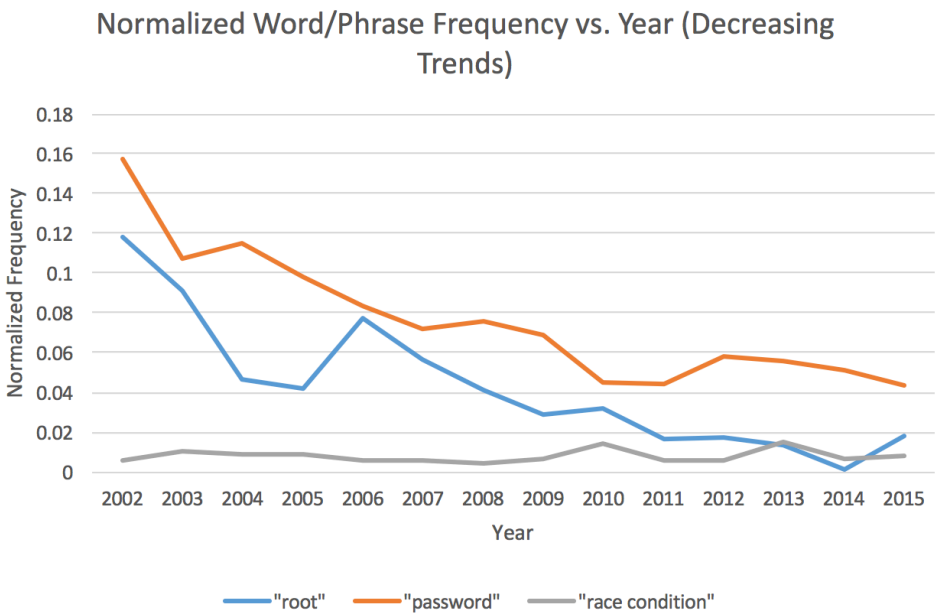
Vulnerability Descriptions

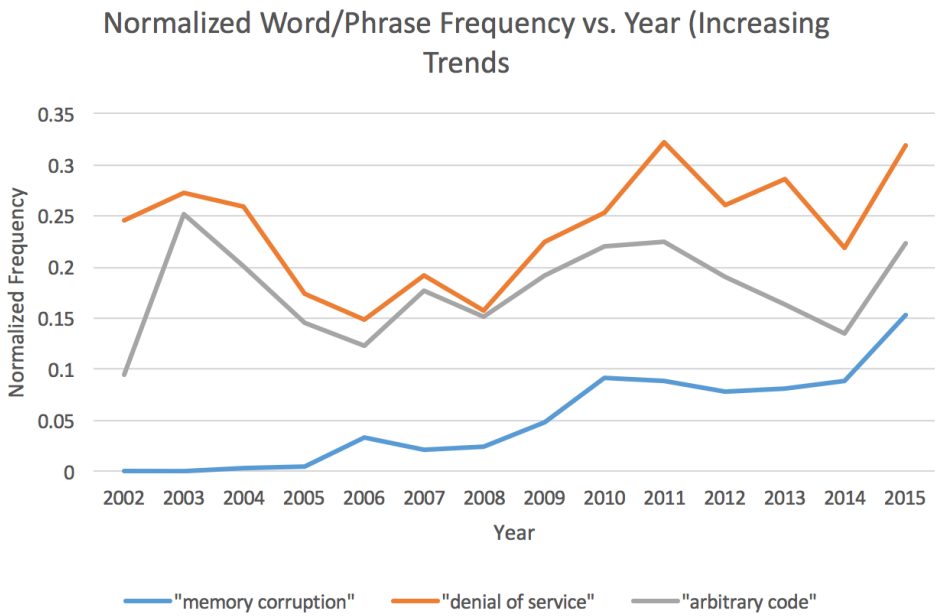
- The descriptions of vulnerabilities in the NVD are analyzed in detail to understand these trends.
- Previous research has looked at these descriptions and shown that they can be used to understand vulnerability trends [3].
- A goal of this research is to use these descriptions to help determine which vulnerabilities are a high priority to remediate.

CVE-2014-0160
The (1) TLS and (2) DTLS implementations in OpenSSL 1.0.1 before 1.0.1g do not properly handle Heartbeat Extension packets, which allows remote attackers to obtain sensitive information from process memory via crafted packets that trigger a buffer over-read, as demonstrated by reading private keys, related to d1_both.c and t1_lib.c, aka the Heartbleed bug.

Word Frequency Modeling

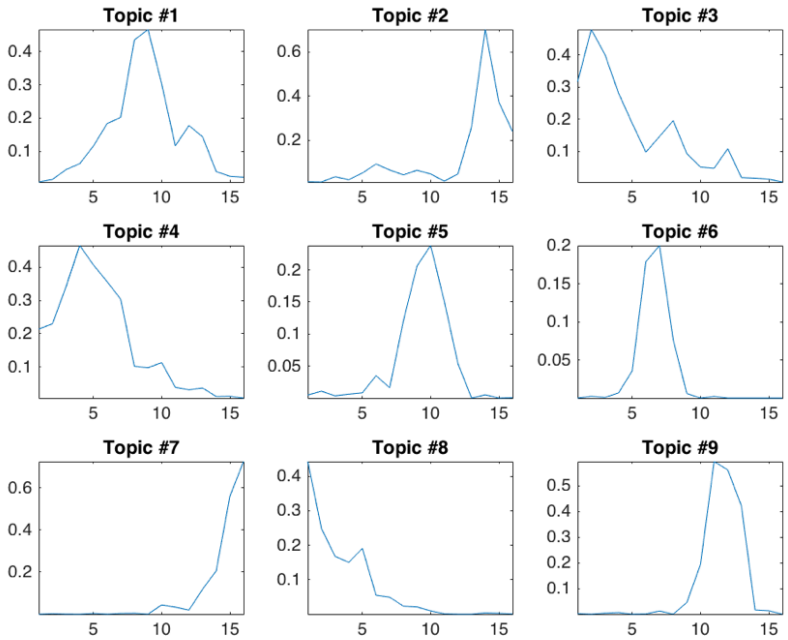
- Vulnerability descriptions were analyzed to understand some of the causes of underlying trends.
- First, the trends in some common words and phrases were analyzed: “password,” “memory corruption,” “arbitrary code,” etc.
- As shown, certain words and phrases have been decreasing over time (left graph), and others have been increasing (right graph).





Forming Topic Models

- Using Mallet and LDA, topics are formed using NVD descriptions.



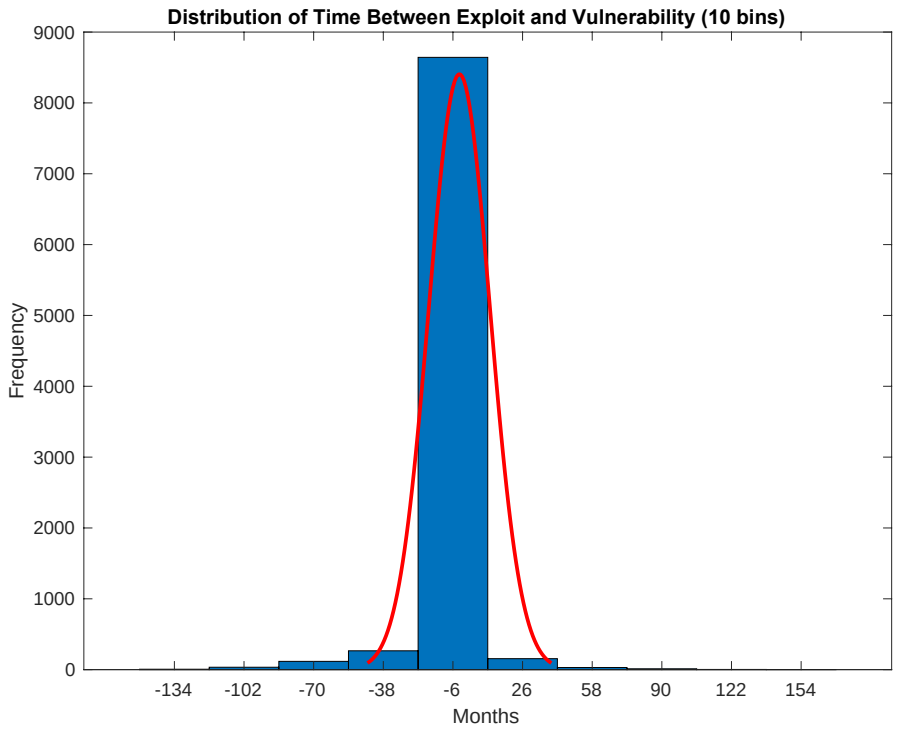
Microsoft Topics with weight given each topic:

1. 0.80832 sp microsoft vulnerability **server** windows code office **crafted remote** execute
2. 0.66163 cve explorer internet **memory corruption** vulnerability aka attackers microsoft service
3. 0.75498 remote attackers microsoft internet execute **arbitrary** explorer **files file web**
4. 0.80438 windows attackers microsoft remote **arbitrary code** xp **buffer** execute internet
5. 0.30644 aka excel **gold** properly office **mac powerpoint corruption** unspecified converter
6. 0.13328 user assisted unspecified note issue **activex** crash vectors information related
7. 0.23908 windows sp vulnerability microsoft **server gold** aka crafted **remote office**
8. 0.31534 server service vulnerability windows **denial** user users properly nt earlier
9. 0.27023 sp windows vulnerability aka **server crafted** microsoft attackers **remote** win

Table of Contents

Predicting Exploits with the NVD

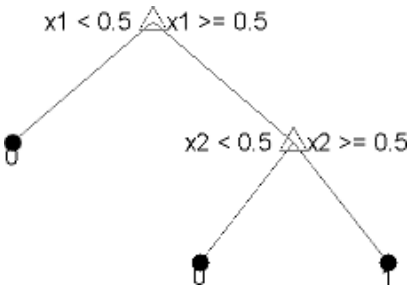
- Using data from the vulnerability descriptions and scores in the NVD, it would be useful to predict which vulnerabilities will be exploited.
- On average, an exploit is reported 3 months prior to an associated vulnerability being reported with a standard deviation of 1 year.

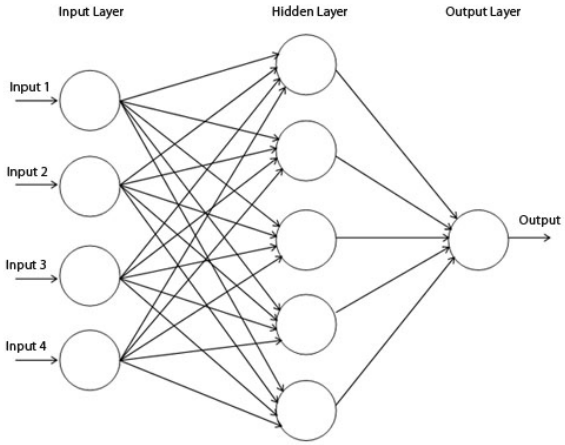


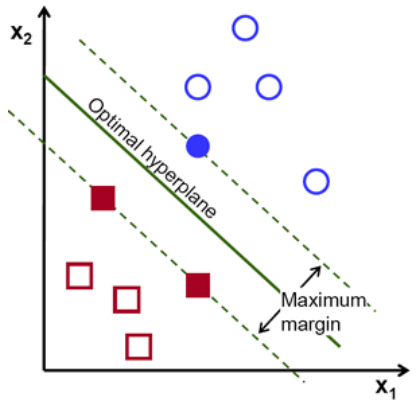
Variable	Mean	Std. Dev.	Min.	Max.	N
Month_Difference	-3.026	14.074	-145	167	9268

Machine Learning Models

- Using four different machine learning models, I replicated the tests and results from a past study by RecordedFuture [4].
- Not all exploits from the Exploit DB are mapped to the NVD.
- It would be useful to classify vulnerabilities with similar characteristics in the NVD as at risk of being in the Exploit DB.



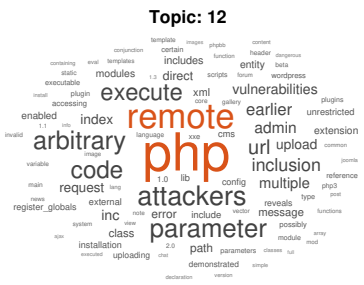


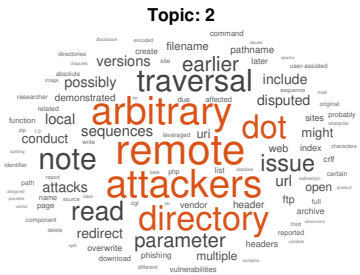


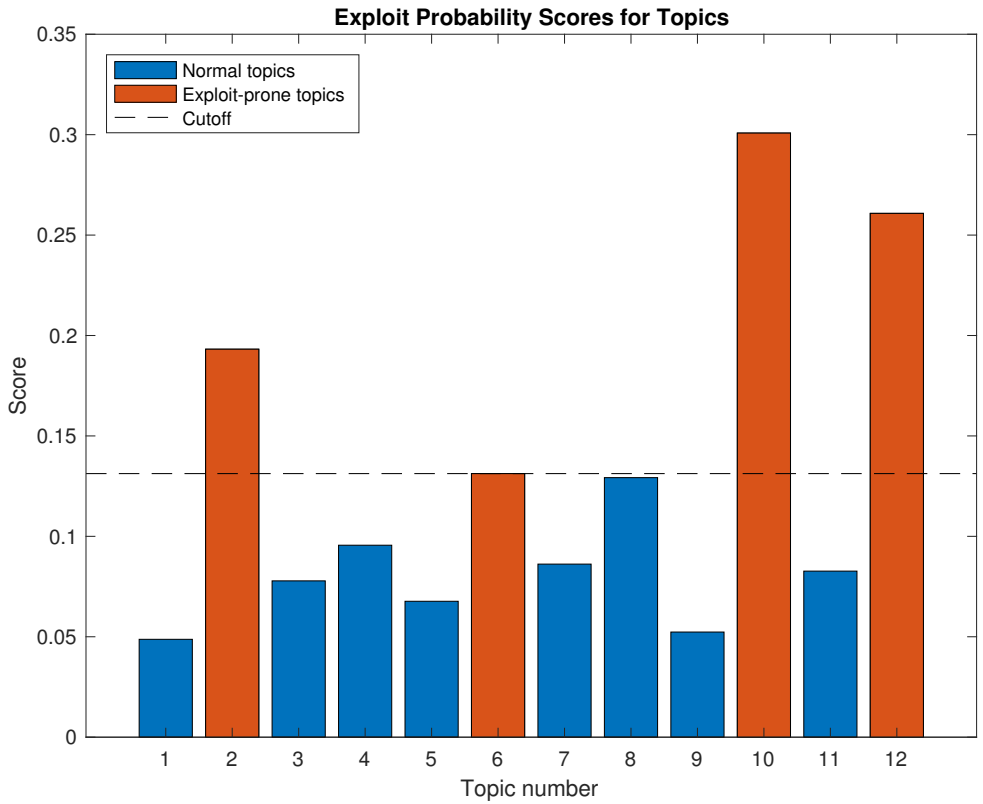
Model Name	Info	Accuracy	Precision	Recall
Fine Tree	Cost = 2	92.2%	28.4%	6.1%
Bagged Trees	Cost = 2	92.7%	40.6%	4.1%
Neural Net	Hidden Neurons = 20	92.7%	42.3%	3.5%
LibLinear	Cost = 21	90.4%	27.8%	21.3%

Topic Model Classification

- As an alternative to the machine learning models, a model is created here that uses the topics found previously.
- The idea is that if a vulnerability belongs to a topic that is at high-risk of being exploited, it should be a high priority to remediate (patch).







Topic Model Classification Results

- This model can predict if a vulnerability has been exploited 85-90% of the time and be adjusted to predict true positives with 50-60% recall.

	Accuracy	Precision	Recall	No. Samples
p=1/12				
Training Data	86.4%	46.3%	31.8%	59,121
Test Data	86.0%	46.6%	32.0%	10,433
2017 (New) Data	91.5%	27.5%	11.1%	13,336
p=1/6				
Training Data	85.4%	44.3%	51.1%	59,121
Test Data	85.6%	46.6%	52.8%	10,433
2017 (New) Data	89.4%	20.5%	16.6%	13,336
p=1/4				
Training Data	83.2%	40.2%	62.2%	59,121
Test Data	83.3%	41.9%	64.0%	10,433
2017 (New) Data	86.2%	16.4%	22.5%	13,336
p=1/3				
Training Data	71.7%	27.0%	69.9%	59,121
Test Data	71.8%	28.1%	70.9%	10,433
2017 (New) Data	81.5%	12.7%	27.0%	13,336
p=1/2				
Training Data	55.6%	20.6%	85.2%	59,121
Test Data	55.9%	21.6%	87.1%	10,433
2017 (New) Data	68.5%	10.5%	45.2%	13,336

Table of Contents

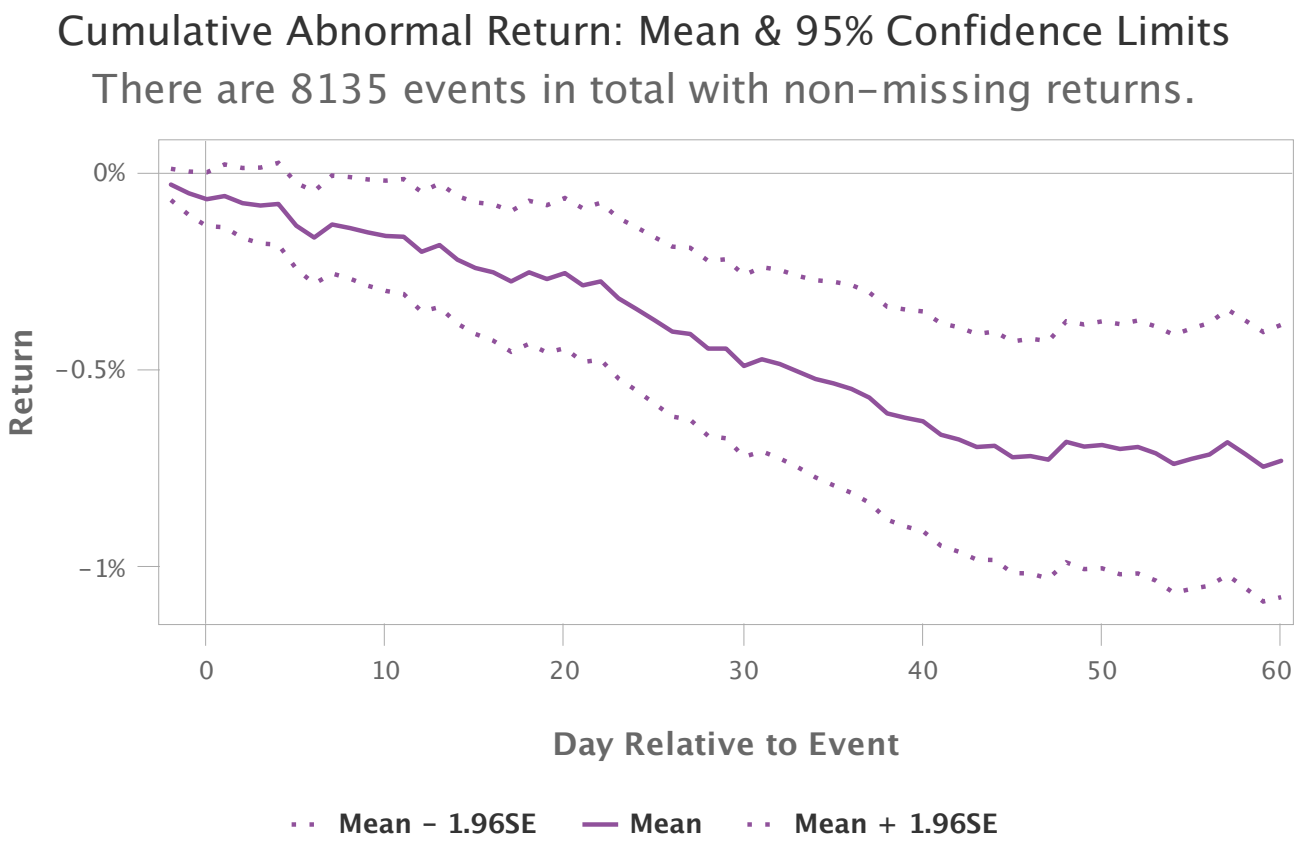
Conclusion

Main findings include:

1. **Correlation with updates** - There was a strong correlation between vulnerability incidence and when updates were released (i.e. Apple).
2. **Feature consistency** - The distribution of vulnerabilities by severity level remained consistent over time, but varied for operating systems.
3. **Topics are present in the data** - It is clear that the NVD presents different topics over time that may support analysis of vulnerabilities.
4. **Negative time between exploits and vulnerabilities** - Many exploits are reported before they are published in the NVD.
5. **Topic modeling classification model** - It is possible to form a classification model based on topics in the data. This approach may be preferable because it would support classification of not yet exploited vulnerabilities that fall within topics that are associated with exploits.

NVD Trends Application: Stock Market

1. Stocks experience about -0.75% in Cumulative Abnormal Returns over the 45 days after a vulnerability is reported in the NVD.
2. It is possible to form a trading portfolio based on this strategy and earn excess returns over what the stock market would return.



Acknowledgements

I would like the following people for all of their support while I was writing my thesis and during my studies:

- Professor Cybenko, my advisor
- Kate Farris, Ph.D.
- My engineering professors
- My family
- My friends

- I would also like to thank the James O. Freedman Presidential Scholars Program for its funding and support.

Table of Contents

Bibliography I

[1] R. Kuhn and C. Johnson, "Vulnerability trends: Measuring progress," *IT professional*, vol. 12, no. 4, pp. 51–53, 2010.

[2] Y. Y. Chang, P. Zavarsky, R. Ruhl, and D. Lindskog, "Trend analysis of the CVE for software vulnerability management," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Oct 2011, pp. 1290–1293.

[3] S. Neuhaus and T. Zimmermann, "Security trend analysis with CVE topic models," in *2010 IEEE 21st International Symposium on Software Reliability Engineering*, Nov 2010, pp. 111–120.

[4] "Anticipating cyber vulnerability exploits using machine learning," Recorded Future, Somerville, MA, Tech. Rep., July 2015.

[5] K. Farris, J. Sullivan, and G. Cybenko, "Vulnerability survival analysis: A novel approach to vulnerability management," pp. 10 184 – 10 184 – 14, 2017. [Online]. Available: http://dx.doi.org/10.1117/12.2266378