Kyle Borah

Pr. Aleshunas

Math 3220 Data Mining Method Webster University

November 20, 2017

Social and Privacy Issues in Data Mining

Problem Description:

As the big tech companies like Apple, Facebook, and Google strive to make their technology better and try to differentiate themselves from their competition, they are adding more and more intelligence and smarts into the software they produce. This added intelligence is more commonly known as AI or artificial Intelligence. AI is added to many pieces of technology even if it doesn't deserve it. This is more of a marketing ploy though, and can confuse customers when asking questions about AI. Most true AI scenarios are backed up by data mining algorithms. From an end user perspective, a customer could potentially have anywhere from very little Personally Identifiable Information, PII, to most if not all of their PII might be in at least one of the major technology companies' databases. This creates a liability on the side of these companies to protect their customer's PII. PII can be anything that connects back to the customer: name, address, date of birth, likes or dislikes, and even a social security number or ID number. The likes ofApple, Google, and facebook are constantly under cyber attack, are more often petition both legally and extralegally by domestic and international agencies for information on users, and still trying to use collected data to provide better services and a better user experience to their customers.

According to a Washington Post article by Adrian Peterson on May 13, 2016, out of forty-one thousand American households surveyed fifty percent of those say they have changed

their internet usage habits because of security concerns. (Peterson) If Companies do not take privacy conscious users' concerns seriously, they risk loosing many potential customers. For example, Equifax was hacked in 2017 and leaked nearly half of United States citizens' PPI (Equifax), Microsoft had a bit of a kerfuffle with windows ten and how much PPI it was or was not collecting (Microsoft), and In the past several years Yahoo has said all three billion accounts have been hacked. (Exhibit) These are just a small subset of the hundreds of other stories every year. It is no surprise internet users are more reluctant and wary around the internet in the twenty-first century.

Companies must protect the data they hold. There are several things to do including encryption and data anonymization.

Background

A category where PII is commonly used is in big data. While no one definition of Big data has emerged, Microsoft puts it well saying, "Big data is the term increasingly used to describe the process of applying serious computing power—the latest in machine learning and artificial intelligence—to seriously massive and often highly complex sets of information." (ArXiv) Big data is used in industries like medical, population analysis, spending habits, and economic analysis. There can be many more, but this is just a small listing. Companies within these fields use these types of big data to train the algorithms they are producing. These algorithms can help predict diseases in segments of the population, specific stats of a type of people group, and predictions of economic market fluctuations. These big data sets can be made available at any time throughout the process. Specific care and sensitivity needs to be taken when dealing with big data because it can also contain PII. A set of medical data might have the names and other sensitive medical data of patients; sets of data collected for population analysis may

have names, addresses, and Social Security Numbers or ID card numbers; and sets of economic data might have personal income numbers and spending history. This kind of information should not be leaking out especially if these data sets are made available for anyone to use.

While the bar of protecting people's PPI is high, there are a few steps that can be taken to make it not be such a chore. Encryption is one of the first and most basic forms of anonymization and obfuscation. At its most basic level, encryption takes a message, it could be a string of text, but it could as easily be a stream of 1's and 0's from a computer, and it runs it through an algorithm with an encryption key. The output is known as cipherer text. The point of encryption is to hide the contents of the message so that someone who doesn't have the decryption key cannot read the message.

Secondly, Anonymization techniques must be taken in order to strip the customer or person from his or her data. A popular technique of data anonymization is Differential privacy. Differential privacy on its own isn't sufficient. Other techniques like injecting random noise into the day must also be taken into consideration. While differential privacy is out of the scope of this paper do to scope and limited time, it is noted here to provide for future research potential.

Methodology:

In order to show the effectiveness of encryption for anonymization, I will run an experiment on a small data set. I will then duplicate the set many times over and demonstrate how even though the set is many times larger, the encryption and decryption cost stays very minimal. I will use RSA encryption for both.

Assumptions:

I am assuming if you wish to replicate my results you are using comparable data sets and hardware. I am using a MacBookPro early 2015 with a 2.7 ghz Intel core I5 cpu run-ing MacOS

10,13,2 High Sierra. More powerful hardware and different software may get different results. This will be discuss more later.

Experimental design:

My first small data set is the build-in iris set in the base R packages. This a data set of 150 items. It is encrypted and decrypted in fractions of a second. The second larger set is iris duplicated 10,000 times over. This forms 1.5 million items; this is much more on the level of a big data set a company might hold on it's customers. It is also encrypted and decrypted in fractions of a second. The obsessive duplication of the iris set doubles as an example of data processing that someone might run on a data set.

Results:

as seen in the output section below, both encryptions and decryptions take fractions of a second. Even the 1.5 million item data set took just under 0.2 seconds to encrypt. While many of these times do very depending on over all system load and other processes, encryption and decryption times are minuscule as compared to over all datelining costs and times. Because this shows that encryption and decryption takes only a small effort to implement. It also should not effect over all system performance.

Issues:

The reader must note that the code used here should be taken as only an example. I do not recommend using this exact code for any real-world applications. It really doesn't represent an actual workflow. Furthermore, a company in the Data Mining field would have access to better hardware than me being a student. Different results may also be achieved if you ran these processes in parallel on a dedicated GPU array.

Conclusion:

In light of many risks to companies with big sets of data containing person's PPI, Encryption is a quick and easy way to keep it all that much more safe. This is crucial for companies to do in the current situation we find ourselves in.

Works cited

ArXiv, E. T. (2014, April 09). The Big Data Conundrum: How to Define It? Retrieved November 20, 2017, from https://www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/

Equifax. Consumer Notice - Cybersecurity Incident & Important Consumer Information | Equifax. (n.d.). Retrieved November 20, 2017, from https://www.equifaxsecurity2017.com/consumer-notice/

Exhibit. (n.d.). Retrieved November 20, 2017, from https://www.sec.gov/Archives/edgar/data/732712/000073271217000003/a2017_10x3xoathxexhibitx991.htm

MicroSoft. Windows 10 and your online services – Microsoft privacy. (n.d.). Retrieved November 20, 2017, from https://privacy.microsoft.com/en-US/windows10privacy

Peterson, A. (2016, May 13). Why a staggering number of Americans have stopped using the Internet the way they used to. Retrieved November 20, 2017, from https://www.washingtonpost.com/news/the-switch/wp/2016/05/13/new-government-data-shows-a-staggering-number-of-americans-have-stopped-basic-online-activities/

Smalldata.r

```
#!/usr/bin/env Rscript
#Kyle Borah
#Math 3220 Data Mining Methods Webster University

library("openssl")
key = rsa_keygen()
pubkey <- key$pubkey

data = iris
file = serialize(data, NULL)

print("encrypting")
start.time =Sys.time()
encrypted_file = encrypt_envelope(file, pubkey)
end.time = Sys.time()
time.taken = end.time - start.time
time.taken

print("decrypting")
start.time = Sys.time()
unencrypted_file = decrypt_envelope(encrypted_file$data, encrypted_file$iv,
encrypted_file$session, key)
end.time = Sys.time()
time.taken = end.time - start.time
time.taken

out = unserialize(unencrypted_file, NULL)
```

Bigset.r

```
#!/usr/bin/env Rscript
#Kyle Borah
#Math 3220 Data Mining Methods Webster University

library("openssl")
key = rsa_keygen()
pubkey <- key$pubkey

start.time = Sys.time()
data = do.call("rbind", replicate(10000, iris, simplify = FALSE))
end.time = Sys.time()
time.taken = end.time - start.time
time.taken

file = serialize(data, NULL)

print("encrypting")
start.time = Sys.time()
encrypted_file = encrypt_envelope(file, pubkey)
end.time = Sys.time()
time.taken = end.time - start.time
time.taken

print("decrypting")
start.time = Sys.time()
unencrypted_file = decrypt_envelope(encrypted_file$data, encrypted_file$iv,
encrypted_file$session, key)
end.time = Sys.time()
time.taken = end.time - start.time
time.taken

out = unserialize(unencrypted_file, NULL)
```

Output

```
Kyles-MBP:course project kyle$ time ./smallset.r
[1] "encrypting"
Time difference of 0.0006859303 secs
[1] "decrypting"
Time difference of 0.001921177 secs

real    0m0.300s
user    0m0.249s
sys     0m0.044s
Kyles-MBP:course project kyle$ time ./bigset.r
Time difference of 41.03269 secs
[1] "encrypting"
Time difference of 0.1990039 secs
[1] "decrypting"
Time difference of 0.07917595 secs

real    0m42.007s
user    0m26.490s
sys     0m15.415s
```