Mining Data Breach Records from ITRC

Shengzhe Qu, Hao Guo, Wen Chen Wu, Guang Yang, Shuchi Gan
Department of Computer Science
The University of Texas at Dallas
Richardson, TX 75080 U.S.A.
{sxq101020,hxg142230,wxw152030,gxy141230,sxg127030}@utdallas.edu

***Abstract*—Cyber-attacks are critical concerns across the world and data breach pose significant danger to the US's national security. This paper provides an analysis of the U.S. data breach list from recent five years to identify the characteristics of the data. Using data mining technique to analyze the data breach report from the year 2012-2015.**

Keywords—data breach; stolen data records; identity theft; cyber security; data mining

I. INTRODUCTION

Hacking activities increased tremendously during the last few years and cyber threats are among the gravest national security dangers to the US. Based on a recent study, there is 50 times more malware today than 10 years ago[1] Data breaches occurred in both consumer services and enterprise environments. Their goal is to obtain valuable information such as governments' classified data, confidential information of a business, and individuals' Social Security Number, DoB, name, and address. Attacks targeting enterprises, corporations, governments, and also individuals are on the rise. The data breach of Anthem Inc. exposed up to 80 million individuals losing protected health information with potential for identity theft[2]. Anthem's breach together with the Community Health System's breach in the year of 2014 has both shown the vulnerability and value US's organizations [3].

Based on the recent study, more than 75% of enterprise decision-makers list that incorporating cyber threat intelligence as their critical priority for the organization [4]. This paper builds upon the ITRC data by using data mining, machine learning and visualization techniques to extract and identify the patterns for the current enterprises, corporations, governments' data breach. By analyzing the characteristics of our extracted data, we can further organize and identify the pattern regarding what specific kind of data is taken and how these data can be utilized by the attacker.

II. PROBLEM DEFINITION

A. Task Definition

First, implementing a program to read the records from the recent 5 years (2012 ~ 2016) of ITRC's pdf formats and convert them into a more readable format like CSV handling and processing conveniently. We analyzed these 5 years' data breach to identify the pattern and using machine learning techniques to train the data in order to predict and the category and analyze the next attack. The breach data provided by ITRC was in PDF format, we then converted PDF files into plain text by using a Java program.

B. Dataset Descriptions

The data are collected from the Identity Theft Resource Center (IRTC), which is a non-profit organization established to support victims of identity theft in resolving their cases and to broaden public education and awareness in the understanding of identity theft, data breaches, cyber security, scams/fraud and privacy issues.

The ITRC currently includes a total of seven categories of data loss methods, which are Insider Theft, Hacking, Subcontractor/Third Party, Data on the Move, Employee Error/Negligence, Accidental Web/Internet Exposure, Physical Theft. Four types of information are Social Security number, Credit/Debit Card number, Email/Password/User Name and Protected Health Information.

Below are examples and a brief description of each attribute for one single record.

TABLE I. ATTRIBUTES OF A RECORD

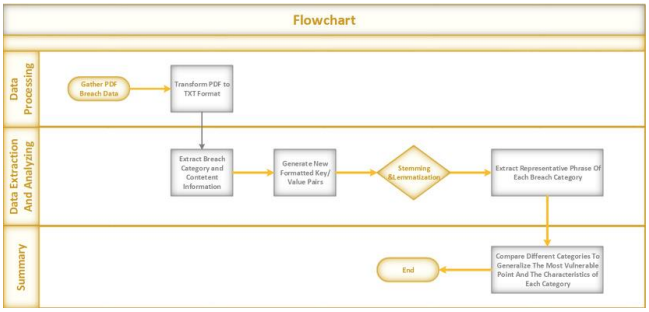
	Example	Description
ITRC Breach ID	ITRC20151229-07	This is the key for each record
Company or Agency	SAS Safety Corporation	Name of the agency
State	CA	Which state this data breach happens
Published Date	12/24/2015	When this data breach happens
Breach Type	Electronic	Type of this record belongs to
Breach Category	Business	Which category it belongs to
Records Exposed	Yes	Whether this record has been exposed
Records Reported	Unknown	Whether this record has been reported
Publication	NH AG's office	Where is the data breach published
Report Date	12/29/2015	When is the data breach published

III. Experimental Evaluation

A. Methodology

The procedures of the whole project are shown in Fig. 1.

Fig. 1. Flowchart of the methodology in this paper



1) Select data files form of which is pdf from ITRC.

2) Convert the PDF files to txt file.

3) Use a Java program to extract the breach categories and corresponding information from txt files, and generate files in which each line contains breach_category::state::text_content".

4) Use Scala and Python to get the statistical data (such as total words in each category, top frequency words and so on) from extracted information and analysis it.

5) Create a model which can predict the category of a file based on the abstract of the file.

6) Based on the obtained results, complete project report and presentation slides.

B. Results

In order to analyze the data, we first converted the original data files (pdf format) to .txt format using Java code. Take the information of Fig. 2 as an example, after we have successfully processed the original data file to the .txt format file. We wrote the program to extract the information based on the main category, like (Banking/Credit/Financial, Government/Military, Medical/Healthcare, Educational, Business). We use this way for the graph in order to show different category and their percentage intuitively.

Next part, we wrote a word count program to generate word count frequency based on the top 15 occurred words, and change the parameter of the program in order to generate both two and three consecutive words, after collecting the top frequency words, we screen the choose the valid word manually.

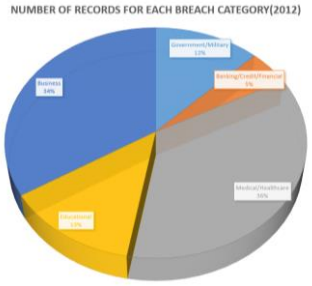
Take social security number as an example, most of the times "social security", "security number" and "social security number" are among the most frequent words, however, these three belong to one valid word record, and choosing one of them is a valid screen operation. By using this method, we are only interested in those word record with most distinct characteristic, and finally, we take each category into account and generate the table to show the detailed analysis of the breach category in each year between 2012- 2016.

We take a deeper analysis for each specific category and generate the figures, by comparing every top frequent word record in each main category, it is very visually clear to distinguish each word and also compare with a different category. Thus after generating the table from different year's result, we can compare and visualize the trend in each category and their top occurred words. The following graph (2012-2016) shows the same methodology of analysis, thus we will not discuss each one individually.

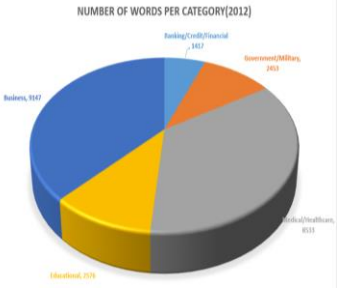
TABLE II. DETAILED ANALYSIS OF BREACH CATEGORY IN THE YEAR OF 2012

Year	Breach Category	Total Number of Words	Word	Count
2012	Banking/Credit/Financial	1417	personal information	10
			account	10
			federal credit union	4
			social security numbers	3
	Government/Military	2453	personal information	22
			social security numbers	17
			confidential information	5
			family services	4
			department revenue	4
	Medical/Healthcare	8533	personal information	30
			health information	30
			social security numbers	25
			medical center	23
	Educational	2576	social security numbers	21
			personal information	23
			university	46
			state university	5
			computer server	5
	Business	9147	personal information	69
			credit card	48
			credit card information	20
			social security numbers	19

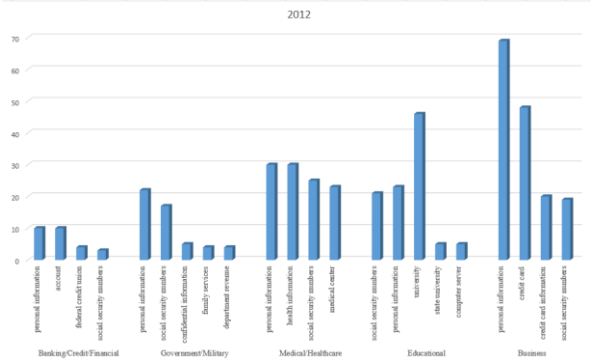
Fig2. Graph for 2012



a. Breach category analysis from Breach Report in the year of 2012



b. Breach category analysis from Breach Report in the year of 2012



c. Breach Deeper analysis for each specific category in the year of 2012

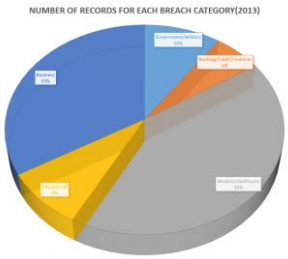
TABLE III. DETAILED ANALYSIS OF BREACH CATEGORY IN THE YEAR OF 2013

Year	Breach Category	Total number	Words	No. of Words
2013	Government/Military	2673	social security numbers	20
			personal information	20
			security breach	5
			personal data	5
	Banking/Credit/Financial	2179	personal information	24
			third party	7
			social security numbers	5
			account number	5
			new hampshire residents	4
	Medical/Healthcare	12557	personal information	52
			health information	48
			medical center	30
			protected health information	28
			social security numbers	24
	Educational	2573	personal information	23
			social security numbers	20
			state university	8
			personal data	6
			community college	6
	Business	11793	personal information	107
			credit card	40
			social security numbers	45
			new hampshire	31
			credit card information	16

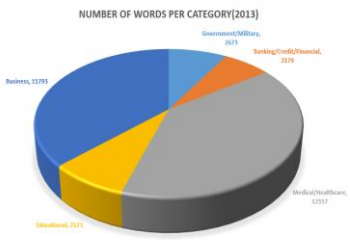
TABLE IV. DETAILED ANALYSIS OF BREACH CATEGORY IN THE YEAR OF 2014

Year	Breach Category	Total number	Words	No. of Words
2014	Banking/Credit/Financial	2876	personal information	25
			law enforcement	7
			social security	7
			credit union	7
			new hampshire	6
			american express card	5
	Government/Military	5136	personal information	27
			social security	18
			human services	10
			health information	7
			dentity theft	7
			copyright identity theft	6
	Medical/Healthcare	14911	social security numbers	53
			medical center	49
			health information	47
			healthcare provider	45
			identity theft	33
			patient information	31
			protected health information	29
	Educational	3306	personal information	34
			social security	27
			unauthorized access	11
			state university	8
	Business	15854	personal information	111
			credit card	46
			social security	42
			payment card	30
			security incident	28
			copyright identity theft	19
			page report details	19

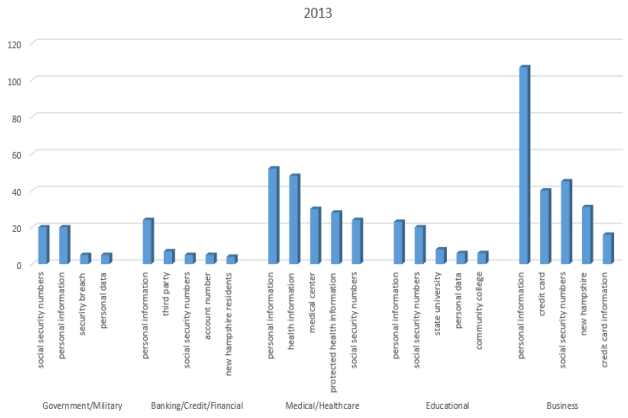
Fig3. Graph for 2013



a. Breach category analysis from Breach Report in the year of 2013

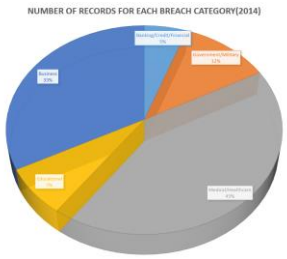


b. Breach category analysis from Breach Report in the year of 2013

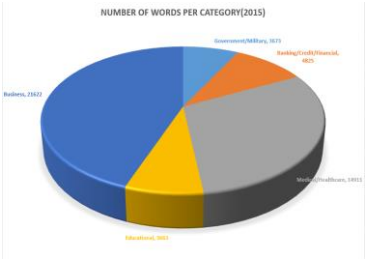


c. Breach Deeper analysis for each specific category in the year of 2013

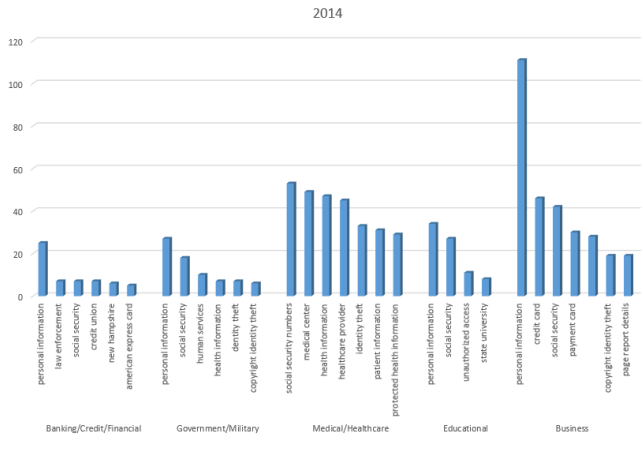
Fig4. Graph for 2014



a. Breach category analysis from Breach Report in the year of 2014



b. Breach category analysis from Breach Report in the year of 2014



c. Breach Deeper analysis for each specific category in the year of 2014

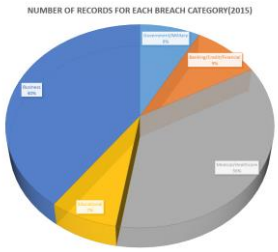
TABLE V. DETAILED ANALYSIS OF BREACH CATEGORY IN THE YEAR OF 2015

Year	Breach Category	Total number	Words	No. of Words
2015	Government/Military	3673	social security	28
			security numbers	22
			personal information	13
			health information	9
			protected health information	7
			personally identifiable	5
	Banking/Credit/Financial	4825	personal information	38
			social security	28
			security number	15
			credit union	12
			name address	11
			vestibule door	11
	Medical/Healthcare	14911	healthcare provider	45
			health information	45
			social security numbers	42
			medical center	30
			identity theft	29
			protected health information	38
			page report details	19
	Educational	3663	social security	30
			personal information	25
			high school	8
			community college	6
			copyright identity theft	3
	Business	21622	personal information	115
			social security	96
			credit card	62
			payment card	46
			name address	37
			copyright identity theft	23

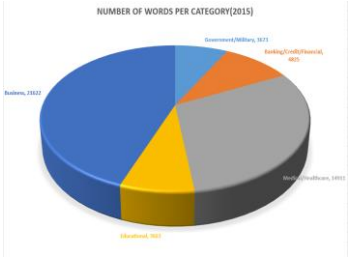
TABLE VI. DETAILED ANALYSIS OF BREACH CATEGORY IN THE YEAR OF 2016

Year	Breach Category	Total number	Words	No. of Words
2016	Banking/Credit/Financial	2608	social security	21
			personal information	21
			social security number	15
			date birth	10
			account number	9
			email account	8
	Government/Military	3991	Social Security	23
			personal information	23
			Social Security numbers	18
			security numbers	8
			phone numbers	5
			medical record numbers	4
	Medical/Healthcare	18647	social security	78
			health information	77
			social security numbers	64
			protected health information	63
			personal information	42
			healthcare provider	39
			unauthorized access disclosure	27
	Educational	4956	social security	44
			personal information	25
			school district	16
			identity theft	7
			phishing scam	7
	Business	27213	social security	112
			personal information	102
			third party	82
			credit card	58
			ag's office	57

Fig 5. Graph for 2015

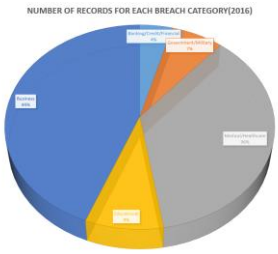


a. Breach category analysis from Breach Report in the year of 2015

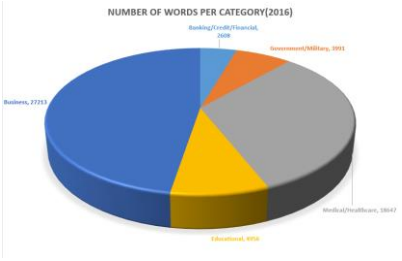


b. Breach category analysis from Breach Report in the year of 2015

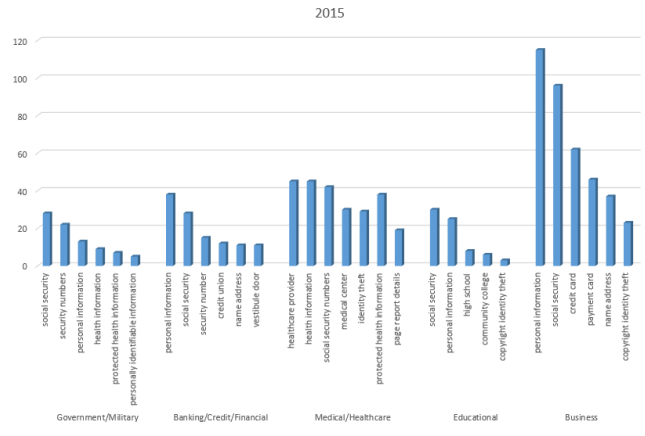
FIG. Graph for 2016



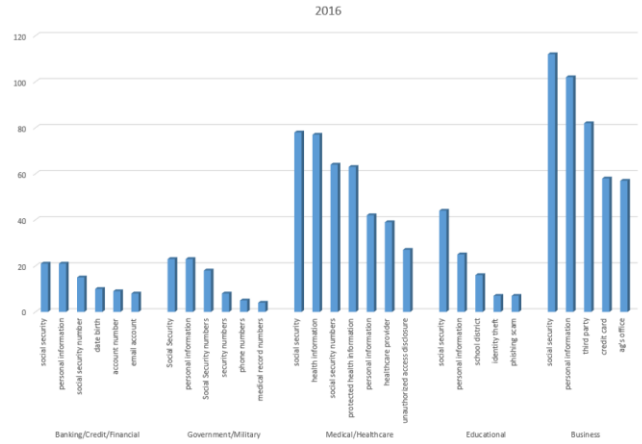
a. Breach category analysis from Breach Report in the year of 2016



b. Breach category analysis from Breach Report in the year of 2016



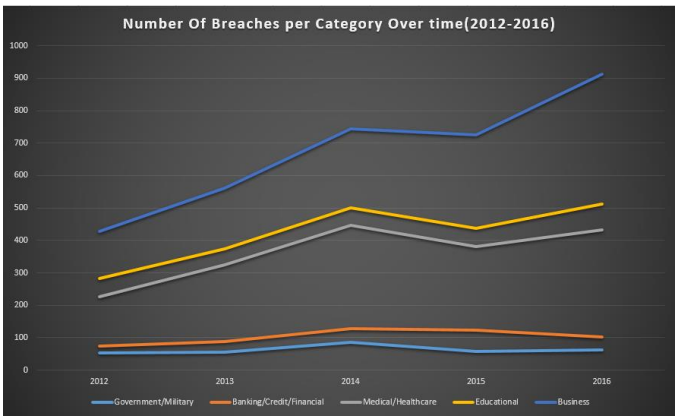
c. Breach Deeper analysis for each specific category in the year of 2015



c. Breach Deeper analysis for each specific category in the year of 2016

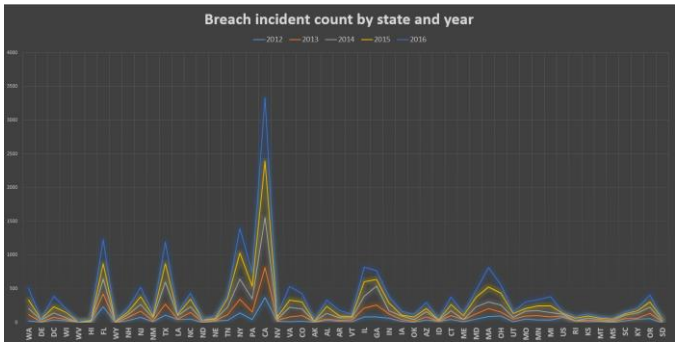
Moreover, we summarized the number of breach in each year (in Figure 7). Based on this figure, we could know the trend of breach in recent five years. The number of breach is increasing. In business, this increasing trend is particularly significant.

Fig. 7. Number of Breaches in recent five years.



Meanwhile, we also summarized the number of breaches in each state (in Figure 8). From this figure, we could clearly find that there is a significant difference among states. In CA, it is the most. We guessed that it may be caused by the well-developed economy in CA. It requires more experiment to determine the relationship between number of breach and the degree of economic development.

Fig. 8. Number of data breach counts in US states for different years



Finally, we used the transformed data to create a model which can predict the category of a report by Naïve Bayes classifier. In creating the model, we used Term frequency-inverse document frequency method (TF-IDF) to analyze the data, used the HashingTF class (containing 60,000 features) from pyspark mllib, used IDF to measure the importance of each term, and finally used Naïve Bayes model to train the data. The accuracy of this model we get from the data is 81.36%.

C. Discussions

As shown in the result, we found that the number of breach reports are increasing in recent years, especially for the "business" category. The reason behind this is relatively ease of the attacks and greater profit from a successful attack at the business corporation. On the other hand, Government/ Military and Banking/Credit/Financial category data breach incident numbers are relatively steady compared to the business category. The main reason is the higher sensitivity and security makes these industries focus more on the prevention of data breach.

In between of the two categories we mentioned above are Medical/Healthcare and Educational, even though the figure of these two categories is not as significant as business category, we can tell from 2012 to 2016 the total number for both categories doubled, which indicates these categories also attract increasing number of attacks due to their valuable and confidential information content.

By analyzing the results generated by our model, we have summarized the top frequent words across categories as well as in each specific category. Top frequent words in each category are "social security" and "personal information", this is obvious since most data breach incidents are targeting the confidential personal information like social security numbers.

As we turn our attention to the specific category, we can conclude most words are close related to its corresponding category. For instance, "account number" and "email account" are the highest frequent words in the category of Banking/Credit/Financial. "Phone numbers" and "medical record numbers" are special ones in the category of Government/Military. "Healthcare provider", "unauthorized access disclosure" and "protected health information" are the characteristic words in Medical/Healthcare category. And " school district" is a unique and one of the highest frequent words in Educational category. Last of all, "credit card" and "third party" are closely related to the Business category since in the field of business, credit card information is the key target in a data breach incident because of its high profit when acquiring the credit card information.

We wrote another program to generate a model in order to predict the category of an unclassified report, as no model is perfect, our model does have some limitation because of two reasons. Firstly because of the training examples are extremely limited especially in some particular years, there is not enough data to build a robust model. Secondly, different words' weight can change over time and some words do take a lot of percentage during a certain amount of time, thus the model needs to take more training data to make it more versatile and relatively accurate when predicting a new record.

As what we have discussed above, the whole paper is built upon our three programs, one for data transformation and processing, one to accurately categorized and organized each category based on our certain requirement and last program to using our previous transformed data to make future predictions. If the limitation mentioned above could be solved, our model could be improved to a more robust and accurate tool which could be used to predict one report basing on only the abstract of the report which is a small part of the report. That will significantly reduce the number of data files required analysis and improve the efficiency of analysis.

IV. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we wrote our own program for the original data transformation and processing, one program for generating the count of each top frequent words, and one program for creating the model to predict the category of one report. These programs are high versatile and are suitable for data files from ITRC;

With the trend of the recent five years detailed analysis, it is highly likely that the threats to every main category in our paper will increase in the future by analyzing the comparison and trend for each category. We believe that the relative ease of the attacks and great profit from a successful attack will continue to lead more data breach in these categories;

Banking/Credit/Financial category has a steady trend, with a slight peek at the year of 2014 and stays the same ever since after that, we believe because of this category belongs to the sensitive area with much focus has been put on the prevention of data breach, therefore Banking/Credit/Financial category will stay the same and might even drop in the future;

To improve the efficiency of analysis of numbers of reports, we created a model, which could predict the category of a report just based on the abstract of the report, by using the transformed data. It works well on current data and may be a good tool for analysis in future;

Due to the nature of the data, we only have grasp of the main category for the data breach, in the near future, if we have the access to the more detailed data to cover a full category list, we can provide an all-around and more comprehensive analysis;

Next, we may combine our dataset with more machine learning techniques, like SVM, Bagging, Random Forest, Ada-boost and Gradient Boosting to predict the category of one report more efficiently, accurately and fast.

ACKNOWLEDGMENT

The paper's data is based upon Identity Theft Resource Center's breach report from the year of 2012 to 2016. The authors would like to thanks Dr. Latifur Khan, for his technical assistance during his Big Data Management and Analytics course.

REFERENCES

[1] AV-TEST, "Malware Statistics & Trends Report," 2016. [Online]. Available: https://www.av-test.org/en/statistics/malware/. [Accessed: 28-Apr-2016].

[2] Nakashima, Ellen. "Security Firm Finds Link between China and Anthem Hack." The Washington Post, February 27 2015. Web. February 7, 2016.

[3] Pagliery, Jose. "Hospital network hacked, 4.5 million records stolen - Aug. 18, 2014." CNN Money. August 18 2014. Web. December 7, 2015.

[4] M. Hartley, "Cyber Threat Intelligence – Thoughts on Recent NetworkWorld Article," 2014.[Online].Available:http://www.isightpartners.com/2014/07/cyber-threatintelligence-thoughts-recent-networkworld-article/. [Accessed: 28-Apr- 2016].

[5] http://spark.apache.org/docs/latest/mllib-feature-extraction.html