Is Bigger Safer? Analyzing Factors Related to Data Breaches Using Publicly Available Information

Ohud Alqahtani, Zhiyuan Chen, Qiong Huang, Karthik Gottipati
Department of Information Systems, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, USA
{*oh8, zhchen, qhuang1,karthi1*}*@umbc.edu*

Keywords: Data Breaches, Privacy, Security

Abstract: Data breaches have affected hundreds of millions of people. As consumers are exposed to constant risks of data breaches, it makes sense to ask what are the factors that contribute to data breaches such that consumers can make more conscious decisions to reduce risks. For example, suppose a consumer want to open a bank account, shall she use a bigger international bank or a smaller community bank considering risks of data breaches? Existing work on risk or vulnerability analysis typically requires detail internal information of an information system, which is not available to the public. Furthermore organizations typically do not want results of such analysis of their IT systems to be made public. This paper proposes a novel approach that analyzes publicly available information to identify factors contributing to higher data breach risks. This paper also presents an initial study that correlates data breaches in the US from 2005 to 2017 with publicly available information about affected organizations. We find that size and name recognition of these organizations are two factors contributing to higher data breach risks. This calls for further study in this direction.

1 INTRODUCTION

As more and more personal and sensitive information being stored in information systems, data breaches have become very common and affected hundreds of millions of people. In the US alone, the number of recorded data breaches is over 1,000 in 2016 (Identity Theft Resource Center , 2017) and over 7730 since 2005, with over one billion personal records being revealed (Privacy Rights Clearing House, 2017). Data breaches often lead to identity thefts and over 15 million people in US were hit by some kind of identity theft in 2016 (Pascual et al., 2017).

As consumers are exposed to constant risks of data breaches, it makes sense to ask what are the factors that contribute to data breaches such that consumers can make more conscious decisions to reduce risks. For example, suppose a consumer want to choose a bank to deposit her money, shall she use a bigger national or international bank or a smaller community bank considering risks of data breaches?

There has been a lot of work on vulnerability and risk analysis of a given information system (Peltier, 2005; Aven, 2007; Ammann et al., 2002; Swiler et al., 1998). However, such work requires detail information of internal of information systems being analyzed, which is typically not available for ordinary consumers. Furthermore, companies or organizations typically do not want the results of risk or vulnerability analysis being revealed to consumers because they may lose business.

This paper proposes a novel approach that analyzes publicly available information to identify factors contributing to higher data breach risks. There are two major challenges to the problem of analyzing risk factors using public information: 1) what factors shall be extracted from public information; 2) how to correlate these factors with data breaches.

To address these challenges, we conducted an initial study. The study used the Privacy Clearing House Data (Privacy Rights Clearing House, 2017) as well as public information about organizations where data breaches occurred. The Privacy Clearing House Data contains 7,730 reported data breaches in US from 2005 to 2017. The attributes in the data set include the name of the organization where a data breach occurred, time, and location of the incident, number of records being revealed, and a textual description of the incident. Initial results showed that organizations with larger sizes or better name recognition have higher number of data breaches.

The contribution of this paper can be summarized as follows:

1. This paper proposes a novel approach to analyze

factors related to higher risks of data breaches based on only public information and identifies size and name recognition as two factors that contribute to higher data breach risks.

2. Since these factors can not be captured by a single universal attribute, this paper proposes several attributes such as revenue, asset size, enrollment, and ranking that can be used to measure these two factors.
3. This paper studies the impact of these two factors for different types of organizations. The results showed that the impact varies with the type of organization. For example, for financial institutions size (in terms of asset) is the dominant factor. However for educational institutions such as universities name recognition is also important.

The results in this paper can be used to help consumers make more informed decisions with respect to risks of data breaches. For example, small community banks have lower risks than bigger international banks.

The rest of the paper is organized as follows. Section 2 summarizes related work. Section 3 describes the initial study using Privacy Clearing House and public data. Section 4 discusses the results and future work. Section 5 concludes the paper.

2 RELATED WORK

There is a rich literature on vulnerability and risk analysis of a given information system (Aven, 2007; Ammann et al., 2002; Swiler et al., 1998). An overview can be found at (Peltier, 2005). However, existing risk analysis methods require detail information of internal of information systems, which is typically not available for ordinary consumers. Furthermore, companies or organizations typically do not want the results of risk or vulnerability analysis being revealed to consumers because they may lose business.

There also has been a lot of work on protecting data privacy, including methods for anonymizing data before being shared (Zhou et al., 2008; Gkoulalas-Divanis and Loukides, 2015), methods for privacy-preserving data mining (Vaidya et al., 2005; Aggarwal and Yu, 2008; Dwork, 2008), and work on policy issues related to data privacy (Bennett and Raab, 2006; Flavián and Guinalíu, 2006). However such work only focus on protecting privacy during data collection and sharing. Although some data breaches happen when data is collected or shared, most data breaches are results of hacking or insider attacks where existing privacy protection techniques have limited use.

There has been some studies on data breaches (Verizon Enterprise Solutions, 2017; IBM, 2017; Romanosky et al., 2014). The Verizon Data Breach Report (Verizon Enterprise Solutions, 2017) gave some statistics on the types of attackers, victims, means of attack, etc. The IBM Cost of Data Breach Report (IBM, 2017) estimated the cost of data breaches. Romanosky et al. (Romanosky et al., 2014) analyzed statistics of lawsuits filed by victims of data breaches. However, none of these work studies risk factors for data breaches.

3 INITIAL STUDY

Our initial study was conducted on Privacy Clearing House data. Section 3.1 describes the data set and presents some statistics. Section 3.2 described our study on size factor and Section 3.3 described our study on name recognition.

3.1 Privacy Clearing House Data

The Privacy Clearing House Data has 9 attributes: date made public, name of organization, location, type of breach, type of organization, records breached, total records, description, and information source. We also computed number of breaches as the number of times an organization has a data breach. The data set stores privacy data breaches from 2005 to present and contains a total of 7730 records. There are some missing values in the data set so we excluded all records with missing values and used the remaining 5573 records. Some organizations have multiple data breaches. So we also grouped records according to name of organization and computed total number of data breaches for each organization.

There are following types of breaches: payment card fraud, hacking or malware, insider, physical loss, portable device loss, stationary device loss, unintended disclosure, and unknown. Table 1 shows breakdown of types of breaches. In terms of number of breaches, the most common type is hacking or malware (26%) followed by portable device loss (24%). Hacking or malware also accounts for about 70% of the records breached and portable device loss accounts for about 20%.

Types of organizations include financial services, retail, other business, educational institutions, government, medical institutions, and NGO. Table 2 reports breakdown of organizations. In terms of number of breaches, medical organizations have the high-

Table 1: Breakdown of Types of Breaches.

Type of breaches	Percentage (# Breaches)	Percentage (# Records)
Disclosure	17%	3%
Hacking	26%	70%
Payment	1%	1%
Insider	12%	4%
Physical	12%	$< 0.1\%$
Portable	24%	20%
Stationary	5%	1%
Unknown	3%	1%

est percentage (around 27%) followed by educational, government, and others. In terms of number of breached records, financial accounts for 41%, followed by retail (30%) and government (20%).

Table 2: Breakdown of Types of Organizations.

Type of organizations	Percentage (# Breaches)	Percentage (# Records)
Other	14%	2%
Financial	13%	41%
Retail	12%	30%
Educational	17%	2%
Government	15%	20%
Medical	27%	5%
NGO	2%	$<$0.1%

3.2 Initial Study on Size Factor

Next we study the impact of size and name recognition on data breach risks. We choose these two factors because they are relatively easy to measure and are often made publicly available. We also find they are both correlated to data breach risks. We will explore other factors such as types of business and location of the business in future.

The main challenge is how to measure size. There are many possible ways to measure size, e.g., by asset size for banks, by enrollment number for universities, by number of beds for hospitals. However it is difficult to find a universal measure for all types of organizations. In addition, it is often necessary to find values of these measures manually for each organization, which could take a lot of time.

To address this challenge, we conducted two sets of experiments.

Size as binary variable: In the first set of experiments, we divide each type of organizations into a "big" group and a "small" group. This basically treats the size factor with a binary variable (big or small) and greatly reduces the amount of information we need to collect. We then conducted t-test to compare average number of breaches in 2005-2017 between the big and small group for each type of organizations. Table 4 reported the results.

To divide organizations into big and small, we used existing rankings in terms of size for different types of organizations. For example, banks and insurance companies are typically ranked by asset or revenue, universities are ranked by enrollment, hospitals are ranked by number of beds, So we use these ranking and select a cut-off between big and small organizations. The cut-off is selected such that the big group contains at least a certain number of organizations such that any analysis we conduct will be statistically meaningful but at the same time organizations in the big group has significantly larger size than those in the small group. For example, we selected banks with at least 50 billion asset as big banks because there are about 24 of them in the data set. Similarly, we selected 100 universities with highest enrollment in 2016 (the most recent number available) as big universities.

Table 3 lists the criteria we used to define the big group. We also merge the type retail and other business into business (non-financial).

Table 3: Criteria for Big Group

Type	Definition of Bigger Group
Financial	Asset$>=$50 billion
Educational	Top 100 with highest enrollment in 2016
Government	Federal, state level and city with over 1 million
Medical	Top 100 hospitals with largest # of beds, Top 25 insurance companies by revenue
Business	Fortune 200 companies

Table 4: T-test between big and small group

Type	Avg. # of breaches big group	Avg. # of breaches small group	p value
Financial	2.65	1.25	0.03
Educational	2.29	1.28	0.002
Government	1.13	1.02	$<$0.001
Medical	1.41	1.05	0.008
Business	1.99	1.06	$<$0.001

The results show that for all types of organizations the big group has higher average number of data breaches and the p-value of t-test is below 0.05 so the difference is significant. This means that larger organizations tend to have higher data breach risks than smaller organizations on average.

We also examined distribution of number of data breaches. Figure 1 shows histogram of big banks and Figure 2 shows histogram of small banks. The results show that most banks have just one data breach. However, there are many more bigger banks than smaller banks that have higher number of data breaches. This suggests that bigger banks are more likely than small banks to have multiple data breaches. The results for other types of organizations are similar and are not shown due to space.

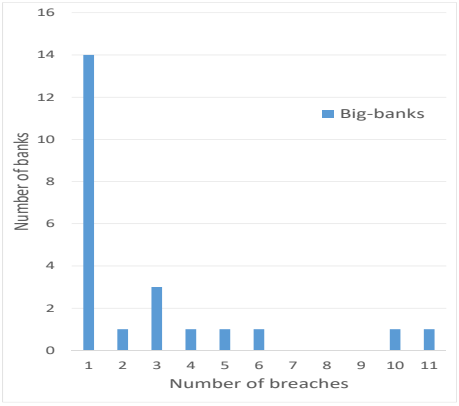

Figure 1: Histogram for big banks

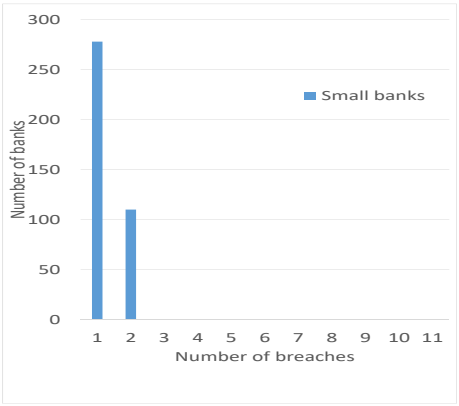

Figure 2: Histogram for small banks

Size as continuous variable: In the second set of experiments, we treat size as a continuous variable and selected a random sample of organizations for each type and collected values of the size measure for each organization in the sample.

Table 5 shows the size measure for each type of organization. For example, we used revenue for medical and business organizations, and the size of population served for government agencies. We then computed Pearson product-moment correlation coefficient between the size measure and number of breaches. Table 6 reports the type of organizations, the sample size, correlation coefficient, and p-value.

Table 5: Size Measure

Type	Size Measure
Financial	Asset
Educational	Enrollment in 2016
Government	Size of served population
Medical	Revenue
Business	Revenue

Table 6: Pearson Correlation between Size and Number of Breaches

Type	Sample Size	Correlation	p value
Financial	23	0.82	$<$0.001
Educational	149	0.24	0.004
Government	21	0.71	$<$0.001
Medical	36	0.31	0.069
Business	21	0.79	$<$0.001

The results show that for financial and business there is strong correlation between number of breaches and revenue size. The correlation is also quite strong between served population size and number of breaches for government. The correlation for these three types of organizations is also quite significant (with p-values < 0.001). We also built linear regression models for financial, business, and government data sets, using the size measure as independent variable and number of data breaches as dependent variable. The R-square value is 0.65 for business organizations, 0.6 for business organizations, and 0.48 for government agencies. This shows that size alone can explain about 48% to 65% of variance of the number of breaches for these three types of organizations.

However for educational and medical organizations the correlation is weaker. The correlation for educational organizations is 0.24 with p value of 0.004 and the correlation for medical organizations is 0.31 with p value of 0.069. We built linear regression models for these two types of organizations and R-square value is 0.09 for medical organizations and 0.06 for educational organizations. This is possibly due to limited sample size as well as other factors contributing to the risks of data breaches for these two types of organizations. In next section we do find that name recognition is another contributing factor for eduationcal organizations.

3.3 Initial Study on Name Recognition

We also studied the correlation between name recognition and number of data breaches. We selected a sample of 149 universities in US and used their US News National University Ranking in 2016 as measure of name recognition. The Pearson correlation is -0.47 with a p-value less than 0.001. The correlation is negative because top ranked school has smaller ranking values. This shows that there is correlation between name recognition and number of data breaches for educational institutions. We also built a linear regression model using ranking as independent variable and number of breaches as dependent variable and the R-square was 0.21.

4 DISCUSSION AND FUTURE WORK

Our initial study shows that size is a significant factor for data breach risks for all types of organizations. The correlation is quite strong for financial, government, and business organizations. One possible explanation is that hackers' goal is financial returns and they can get higher returns by attacking larger organizations.

Our study shows that for educational and medical organizations, very big organizations tend to have higher data breach risks than very small organizations. However the correlation is weaker than the correlation for other types of organizations. This may be due to the small sample size as well as other factors that may contribute to risks of data breaches.

Our study also finds some correlation between name recognition and data breach risks for educational institutions. One possible explanation is that hackers need to know the name of the organization to launch an attack, so a lesser known university may not attract many hackers.

There are many possible future research directions and we encourage more research to be done in this area. Below is a non exhaustive list of possible research directions.

First, there is need for more comprehensive study of risk factors associated with data breaches. For example, one can increase sample size of studies in this paper. One can also study other factors such as locations of organizations (e.g., local vs. national vs. multinational). It is possible that a more global organization may attract more hackers because many hackers are from foreign countries.

Second, there is need to study better measurement of risk factors. For example, it is difficult to find rankings based on name recognition for non educational organizations. One possible universal measure of name recognition is the number of Google search results when the search contains the name of an organization. Organizations with better name recognition usually have more search results than organizations with less name recognition. For example, a search for "Stanford university" returns over 38 million results, but a search for "UMBC" only returns 3 million results.

Third, our study focuses mostly on number of data breaches. However each data breach is not equal. Some may lead to more severe consequences in terms of number of records compromised as well as the type of information being compromised (e.g., whether social security number is revealed). So another direction is to take into account the consequences of data breaches.

Finally, once all risk factors are identified, one can create a prediction model that estimates the risks of data breaches based on these factors. Note that we do not have to estimate the absolute probability of data breaches for an organization because that may require more internal information about the organization. Instead, we are more interested in comparing relative risks due to several risk factors so consumers can use that for better decision making. For example, it will be helpful to know that using a big bank rather than a community bank will increase the risks of data breaches by a factor of x. One possible model is Cox proportional hazards regression model because it has been widely used for studying impact of risk factors.

5 CONCLUSION

This paper analyzes factors that may increase risks of data breaches using publicly available information such as sizes and name recognition. Our initial studies using the Privacy Clearing House data show that there is strong correlation between data breach risks and size of financial, government, and business organizations. Name recognition is also correlated with data breach risks for educational institutions. The results of this paper can be used to help consumers make better decisions. For example, consumers should use smaller community banks over large banks when considering data breach risks. We also discussed a few future research directions and encourage more research to be done in this area.

REFERENCES

Aggarwal, C. C. and Yu, P. S. (2008). *Privacy-Preserving Data Mining: Models and Algorithms*. Springer Publishing Company, Incorporated.

Ammann, P., Wijesekera, D., and Kaushik, S. (2002). Scalable, graph-based network vulnerability analysis. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, pages 217–224. ACM.

Aven, T. (2007). A unified framework for risk and vulnerability analysis covering both safety and security. *Reliability engineering & System safety*, 92(6):745–754.

Bennett, C. J. and Raab, C. D. (2006). *The governance of privacy: Policy instruments in global perspective*.

Dwork, C. (2008). Differential privacy: a survey of results. In *Proceedings of the 5th international conference on Theory and applications of models of computation*, TAMC'08, pages 1–19, Berlin, Heidelberg. Springer-Verlag.

Flavián, C. and Guinalíu, M. (2006). Consumer trust, perceived security and privacy policy: three basic elements of loyalty to a web site. *Industrial Management & Data Systems*, 106(5):601–620.

Gkoulalas-Divanis, A. and Loukides, G. (2015). A survey of anonymization algorithms for electronic health records. In *Medical Data Privacy Handbook*, pages 17–34. Springer.

IBM (2017). IBM 2017 cost of data breach study.

Identity Theft Resource Center (2017). Data breaches increase 40 percent in 2016, finds new report from identity theft resource center and cyberscout.

Pascual, A., Marchini, K., and Miller, S. (2017). 2016 identity fraud: Fraud hits an inflection point by javelin strategy and research.

Peltier, T. R. (2005). *Information security risk analysis*. CRC press.

Privacy Rights Clearing House (2017). Data breaches since 2005.

Romanosky, S., Hoffman, D., and Acquisti, A. (2014). Empirical analysis of data breach litigation. *Journal of Empirical Legal Studies*, 11(1):74–104.

Swiler, L. P., Phillips, C., and Gaylor, T. (1998). A graph-based network-vulnerability analysis system. Technical report, Sandia National Labs., Albuquerque, NM (United States).

Vaidya, J., Zhu, Y. M., and Clifton, C. W. (2005). *Privacy Preserving Data Mining (Advances in Information Security)*. Springer-Verlag New York, Inc.

Verizon Enterprise Solutions (2017). 2017 data breach investigations report.

Zhou, B., Pei, J., and Luk, W. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter*, 10(2):12–22.