Is Good Searchable Encryption Worth Finding?

Ashok Deb, ASM Rizvi, Rajat Tandon

November 5, 2017

Abstract

Data breaches have become a very hot topic for good reasons. While much of the news recently centers on the Equifax data breach which probably affected 143 million people, data breaches have been a long standing problem. Since 2013, there have been over 9 billion data records lost or stolen [19]. Additionally, approximately only 4% of breaches had a level of encryption to where the stolen data was not exploitable [19]. Most data breaches have occurred from either hacking, poor security or accidentally publishing. This widens the decision space of how companies and organizations should better protect their data given the various types of encryption currently deployed within a system. Both from a fiscal standpoint and a policy standpoint, decision makers are often left confused or misguided in their data security. The goal is to help decision makers understand the weakest link in their database security and to what degree searchable encryption could help.

1 Introduction

1.1 Motivation

Data breaches have become a frequent and alarming issue for computing and network security world. According to the breach level index [19], around 61 records are being stolen every second. These incidents of data breaches are increasing day-by-day. Verizon data breach report of 2017 [20] shows that

almost every sectors get affected by data breaches. This includes accommodation, education, finance, health-care, information, manufacturing and even public administration. Financial gain is the most common motive behind these data breaches. Around 24% of data breaches affects financial organizations, and 73% data breaches are financially motivated. Data breach is also common in sensitive sector like health-care. Around 15% of data breaches target health-care organizations.
According to Verizon report, 81% of hacking-related breaches are coming from stolen or weak passwords. Moreover, in many cases of health-care, manufacturing, and public administration, espionage attacks are common where insiders or administrators are involved in honest but curious attacks. Different searchable encryption techniques can handle honest but curious attacks. Different schemes have already been proposed to search over encrypted database, which might reduce the chance of breaching the database from honest but curious attacks. However, the number of breaching incident is not decreasing. There might be two reasons behind it. First, it is possible that the existing encryption mechanisms are not sufficient. They can leak information with the help of auxiliary data, or even with data inference [12], [3]. Secondly, it might be possible that organizations are not willing to use any encryption mechanism at all. Organizations have employees who are working for many years. Any change over the system will cost the organizations to train their employees. Moreover, there are performance and security trade-offs. The performance of the overall system normally gets degraded if we use encrypted mechanisms.

1.2 The Problem

We feel that there is a significant gap between the extensive research in security field, and their impacts in real systems. In this paper, we try to investigate different breaches and want to see whether the administrator used encryption technique or not in the incidents of data breaches. If the administrator used encryption or protected technique, then we will try to find why those techniques fail.

1.3 Our Contribution

In this paper, we evaluate and analyze recent data breaches to provide the research community of trends analysis of the root causes of these breaches.

This in turns allows for use to develop an evaluation framework for the security, performance and risk of various databases. Using this framework, a company or organization can conduct an assessment to see if protected search makes sense for their data.

1.4 Organization

This paper provides a significant background on recent data breaches to include their type of attackers, the type of vulnerability exploited, the industry and type of database targeted as well as any time of encryption that could be used. With that background, an analysis of significant data breaches is conducted in order to provide insight into a dataset experiment. The dataset experiment results are what allow us to develop the evaluation framework.

2 Background

2.1 Type of Attackers

In the cyber realm, there are a variety of the type of attackers which have to be defended against. There are four main categories which vary in their size, resources, abilities and motives. The first are nation state actors such as the United States, China and Russia. They each have a dedicated arm of the government whose focus is offensive cyber activities. For example The Tailored access Operations is part of the USs National Security Agency, while Russia has the GRU as part of their intelligence services and the Chinese have PLA 61398. They act on the behalf of and on the direction of their federal governments, each of which are well-resourced with talent and funding to affect their respective countrys national, political, military, economic and social interests. Additionally, there are organizations which may be independent of the government of a country, but work toward the same interest of their aligned country. Examples include the Equation Group for the US, Network Crack Program Hacker group for the Chinese, Fancy Bear or APT28 for the Russians, Syrian Electronic Army for Syria and Tarh Andishan for Iran. Their resources and talent are very similar to what their respective nation-state can provide as well as having similar interest. However, since they are not officially part of the government, they may have more latitude in the types of attacks and types of targets they can focus on. Both na-

tion states and their associated organizations can provide one-time targeted attacks as well as present an advanced persistent threat to others.

Lastly there are groups such as Anonymous, The Shadow Brokers, LulzSec and Lizard Squad. Their are completely independent of a nation state and while potentially have less resources, they can be just as effective. In addition to having some of the same motivations as nation states, they can also be motivated by financial gain, social justice or nefarious motives to disrupt the status quo. With the same motivations and potentially even less resources are individual threat actors who may or may not be a part of a hacking group. Inclusive of the individuals are those who are non-malicious, white hat hackers and honest but curious.

2.2 Type of Vulnerabilities

There are a variety of vulnerabilities that can be exploited in a cyber attack, specifically a data breach. The first is compromised credentials. Whether obtained by physical theft, social engineering, spyware or phishing, compromised credentials give an attacker to all of the data the compromised user would have. This is the hardest to detect because they system wouldnt recognize anything was wrong initially. Lost hardware presents another vulnerability where data may be stored on a laptop that is stolen. Since an adversary would have physical access to the device, they can take great time and use more resources to recover the information on the device. This because increasingly challenging with more remote workers an the increase of mobile devices in the workplace and personal lives. Software vulnerabilities in either an operating system, application, web interface or network software also presents an attack vector that can be studied and exploited. Either of these vulnerabilities could be presented to an attacker or created by an attacker. A laptop can be left at a coffee shop or be stolen from a house. Known software vulnerabilities can be exploited and Day Zero exploits engineered by attackers can not be known by others.

2.3 Industry

In additional to personally held data which may be subject to ransomware attacks, there are a few industries which typically has exploitable data sets. Government organizations will maintain data that would be harmful to national security if leaked, to include employee data, counter-intelligence files,

operational or contingency plans, classified communication, weapons design and characteristics. The weapons and research information would also be held by those in the defense industry to include those that provide contractors for augmented personal to government agencies. The medical community as a hold has very sensitive data as it concerns patient files which are more commonly held in electronic medical records. The financial industry has personal information on consumers and corporate information for companies that can be exploited for ransomware or to illegally operate in stock markets with insider information. Entertainment companies would have media such as film, photos or music that is copyrighted or unrelated which would be a financial loss if leaked. Educational institutions would have personal files as well as research material that is sensitive. This is especially the case as more universities conduct basic and applied research on behalf of governments. The retail industry often has customer data to include username, login, address and credit card information that is highly exploitable. Lastly, social media organizations often have private communication and photos that can be exploited, especially for high-network individuals or high-profile people.

2.4 Database Types

They type of content that is stored in databases vary by industry and purpose. The main types of data that need to be protected or human resource or employee data, banking information such as credit card data, media to include video, photographic or music files, documents to include banking, legal or government, and emails. These each have unique characteristics in their exploitability and in which methods would best protect them.

2.5 Encryption

Encryption in general is converting data into a coded message so that only authorized people can have access. Encryption is widely used for system access, network access and to protect data at rest. The research considers the potential functionality of searchable encryption on databases in light of the overall security platform. Searchable encryption is defined to be any scheme that allows the data in a database to be encoded so that an unauthorized person would not be able to make sense of it while authorized users would be able to perform some set of queries on the data while it is still encrypted. The

overall added value of searchable encryption has to take into consideration recent work on active attacks, leakage attacks and inference attacks.

2.6 Communications Encryption

One type of encryption method deals with securing the line of communication between a client and a server or a customer and the company. While this manner of encryption is not within the scope of the analysis of this paper, it is worth noting that this is an important aspect of overall data security. The two main types of algorithms that address data communications across a network are either symmetrical or asymmetrical. Wang and Wu provide [24] an analysis of popular schemes such as Data Encryption Standard, Advanced Encryption Standard, RSA and Elliptic Curve Cryptography. Not just for data breach prevention, but for the entirety of network integrity, all stakeholders on the network should ensure the proper communications encryption is being used.

2.7 Disk Encryption

Disk encryption is to encrypt data at rest, such as stored on a hard drive. Authentication is needed to decrypt the cipher text into plaintext and this method helps protect against a compromised hard drive or database from being exploited. Hars from Seagate Research presents many of the facets and challenges of hard-disk encryption for secure storage [8]. While disk encryption should be used for remote devices, it is often difficult to use on a server because the data needs to be accessed and updated by mainly people. This is what leads to the current research efforts of searchable encryption.

2.8 Searchable Encryption

2.8.1 CryptDB [15]

Researchers from MIT CSAIL present CryptDB which a software system that provides increased security and confidentiality for database-backed applications. The solution addresses two main concerns which are database administrators (DBAs) that might exploit the data and compromises to the application side server or database system. In order to address their two

named threats: DBMS Server Compromise and Arbitrary Threats, they device a layered, SQL-aware encryption scheme. The SQL-aware scheme allows for the type of encryption on data elements to be determined by the anticipated SQL queries to be processed. This enables them to have a more robust encryption placed on data columns that do not need to be ordered or word searched. It does all for sorts and searches by using less secure encryption methods but allowing all access keys to be generated at user login and deleted at user logout. This reduces and DBA from making sense of the data they are only suppose to be administrating and it limits an application side breech to only the data accessible by currently logged-in users. CryptDB was tested using phpBB, HotCRP and grad-apply for application case studies and had moderate success. While the vast majority of SQL queries ran were well supported by CryptDB, the software struggled with WHERE, GROUP BY an ORDER BY statements. In all, the software addresses the two specific threats that they had outlined. They indicated the Insider Threat as their first main threat which is a growing concern for organizations and often overlooked. They state that the goal is confidentiality (data secrecy), not integrity or availability. Working definitions of each would be good.

2.8.2 Searchable Symmetric Encryption [4]

This paper presents the details of proposed protocols that can be used for conjunctive search on encrypted database while providing a fair and documented tradeoff between security and efficiency. The authors provide a great review of the current field and incrementally build to the Oblivious Cross Tags protocol which enables conductive keyword search while minimizing leakage. They provide a proof of the upper bound of the protocols leakage recognizing, there is not efficient or scalable way to search on an encrypted database without some leakage. In addition to the proofs of their maximum leakage, they implement their protocol on three databases of various sizes to show that their method is scalable, especially to datasets that cannot be wholly contained in RAM. They build on work using searchable symmetric encryption where one key term can be search and their work can be incorporated into other systems such as CryptDB in order to make them more secure. The novelty in their design is seeking to do the conjunctive search is order of most restrictive term first and by reducing the communication required with the server to minimize leakage from search results by pre-computing the blinding part of the oblivious computation and storing it in encrypted form

at the server. This protocol would prevent the typical inference attack easily done on encrypted data columns that contain gender. The could have made a better highlight into countermeasures against an inference attack with the attacker has prior (or external) information on the database.

2.8.3 Seabed [14]

This paper presents the the underpinnings of a system called Seabed which allows for encrypted search on large databases. The threat model that they use is the honest but curious database administrator on a third-party cloud storage solution. Not unlike CryptDB or Monomi, their claim is to have a system that is efficient and scalable to big data for multiple terabytes of data. Their novelty is to leverage asymmetric symmetry to increase speed and to randomize the the dummy coding of low dimensional data in order to prevent frequency attacks. They implement and perform some low-level tests of the Seabed system in a case study. It is noteworthy that they seek to address a balance between confidentiality, performance and functionality because as of now, it is not possible to have all three. This is sort of like in consulting where you say, fast, cheap, good, pick two of the three. It is still misleading to make a statement that DET and OPE schemes leak a small amount of information.

2.8.4 MiniCrypt [25]

MiniCrypt is a key-value store that addresses both encryption and compression in an unique way. They do this be provided a simple mapping scheme that allows the server to identify the pack (or grouped rows) that contain the needed keys. This is because encrypting packed rows removes the servers ability to manage key-value pairs and to maintain correct semantics. Therefore, MiniCrypt provides end-to-end encryption for the values in the key-value store while providing significant compression. As an added benefit, MiniCrypt works as a layer on top of unmodified key-value stores. The consistency guarantees which are needed for big data key-value stores are engineered to be there as well. It is good that they are addressing the protection of confidentiality and the preservation of performance by incorporating encryption and compression in the same design. While the authors consider the curious system administer as the threat, they do not speak much to inference attacks that might be possible on their system.

3 Methodology

We seek to understand the nature of recent data breaches in order to better protect and defend against future breaches. We do this by conducting a statistical analysis of all data breaches going back to as early as 2004. The four main sources of our data are Vigilante.pw, Wikipedia, Privacy Rights Clearing House and Beauty of Information. With a combined database of data breaches which are the union of those sets, we can estimate the likelihood of data breaches occurring under various scenarios. Furthermore, within those scenarios, we can determine which countermeasures or mitigating factors would be the most effective. With this combined information we will leverage data mining techniques to cluster the breaches and identify key factors associated with breaches. From this analysis, we will be able to make a decision tree which could be used by decision makers and security professionals to predict data breaches and develop a plan of action to mitigate them.

3.1 Scrupulous Classification of Data Breaches

There are multifarious ways of classifying data breaches encompassing categories like whether the data revealed was encrypted or plaintext, internal or external actor involvement, hack or malware or weak passwords, client side hack or server side hack, financial gains or privacy breaches, known or hidden breaches.

3.1.1 Breaches on encrypted vs unencrypted data

1. **Breaches on encrypted data [1]** - There has been quite a number of incidents wherein data breaches occur in spite of the presence of encryption. Either vulnerabilities in the system as a whole or human errors allow attackers access to encrypted channels. Zero-day attack vulnerabilities are more than enough for attackers to break into the system.

   A few examples of this variant of data breaches are:

   - **e-Bay data breach** - A group of attackers leveraged phishing attacks and stole credentials of 100 eBay employees. They used that information to gain access to eBay's internal network, where

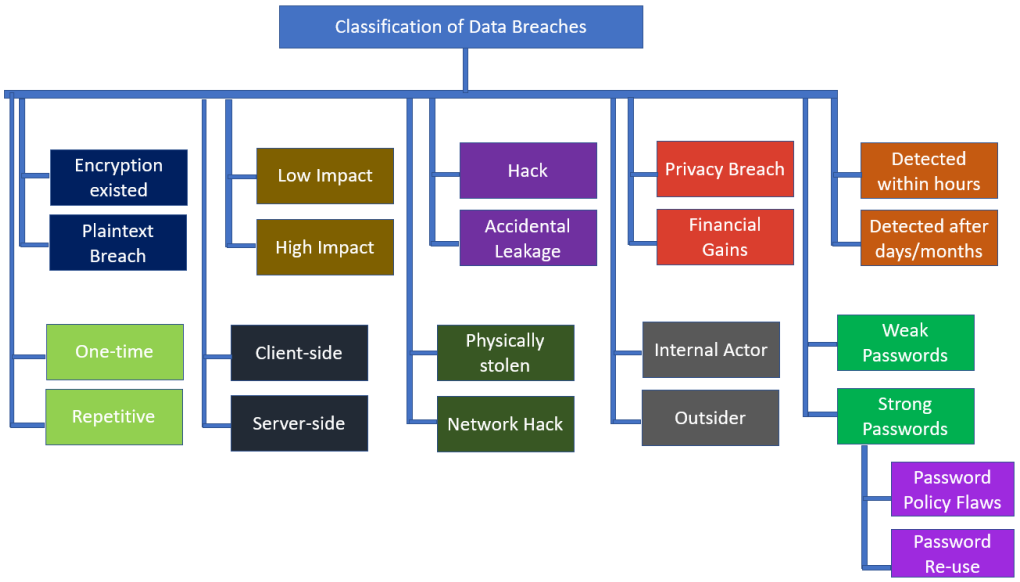


Figure 1: Classification of data breaches

they then exfiltrated the names, passwords, email addresses, physical addresses, and other personal information of 145 million customers. The attackers allegedly had access to eBay's systems for 229 days. Because the attackers have access to a network, they could install fake or stolen certificates that allowed them to hide exfiltration in the encrypted traffic.

- **Pentagon data breach** - In July 2015, the attackers used a spear-phishing attack to hack the Pentagon's Joint Staff unclassified email system. The attack consisted of encrypted social media accounts for coordination. The key reason behind the data breach was if organizations do not monitor keys usage and certificates that are used in encryption, then attackers could fake or steal keys to create illegitimate encrypted tunnels.

2. Breaches on unencrypted data When safe is unsafe then how can unsafe be safe. A few examples of this variant of data breaches are [11]:

    - **Incident reported to HHS on June 1 by Oregon Health Co-Op** - On April 3, 2015 a laptop was stolen. The device con-

Database Source	Entity	Description of Incident	Location	Year	Records Lost	Source of Information	Method of Leak	Organization	Data Sensitivity
Information is Beautiful	X	X		2004-2017	X	X	Accidentally Published	Academic	Email Address
							Hacked	App	SSN/Personal Detail
							Inside Job	Energy	Credit Card Information
							Lost/Stolen Device or Media	Financial	Email Password/Healh Info
							Poor Security	Gaming	Full bank Account
								Govern,ent	
								Helatcare	
								Legal	
								Media	
								Military	
								Retail	
								Tech	
								Telecoms	
								Transport	
								Web	
Privacy Rights Clearning House	X	X	X	2005-2017	X	X	Payment Card Fraud	Businesses- Financial and Insurance	
							Hacking or Malware	Businesses-Other	
							Insider	Businesses-Retail/Merchant	
							Physical Loss	Educational Institutions	
							Portable Device	Government and Military	
							Stationary Device	Healthcare, Medial Providers & Medical Insurance Services	
							Unintended Disclousre	NGOs and Nonprofits	
							Unknown		
Wikipedia	X			2004-2017	X	X	poor security	consulting, accounting	
							hacked	financial, credit reporting	
							lost / stolen computer	healthcare	
							poor security / hacked	web	
							inside job	financial	
							accidentally published	government	
							lost / stolen media	telecoms	
							unknown	political	
							inside job, hacked	tech	
								gaming	
								telecom	
								hotel	
								retail	
								academic	
								restaurant, retail	
								restaurant	
Vigilante.pw	X			2012-2017	X				

Figure 2: Different sources of data breaches

tained data related to the member and dependent names, addresses, health plan and identification numbers, dates of birth and Social Security numbers which impacted around 14,000 individuals.

- **Theft reported by Nevada healthcare** - On June 10, 2015, Nevada Healthcare reported that there was a theft involving electronic medical records, a laptop, a network server and other portable electronic devices. Approximately 12,000 individuals were affected.

In Fig. 1, we can see the classification of data breaches.

3.2 Data breach Dataset

Here we will document the sources of our dataset, summary statistics as well as visual charts to illustrate the delineation of data breaches.

We have four main sources for the data we used to analyze databreaches. The first main one is from Privacy Rights Clearinghouse [16] which has

a dataset of 7,730 public data breaches that comprise of over one billion breached records.

There is a Breached Database Directory maintained by vigilante.pw [21] that lists of over 3,200 websites whose databases were breached with over 3.6 billion records in total.

The third data source is Information is beautiful [9] which is a small team of independent data journalist an information designers who maintain a data set of the worlds biggest data breaches (with losses of greater than 30,000 records).

Additionally, we cross-referenced with he list of data breaches available publicly on Wikipedia [23]. The Wikipedia data set mainly consist of breaches where the number of records breached is over 30,000 or a breach of a major company or organization and the number of records breached is unknown. The majority of the breaches listed there are from North America.

Fig. 2 gives a view of the data structure of each data source.

The Organization category initially contained 15 various organizations which we have consolidated down to 7 which are listed and described below:

Government - We have combined Government and Military into one category since their is a push, especially in America for all government networks to be behind one common firewall that can be better protected. This applies mainly to the United States, but many other countries are already doing this. This need was highlighted in recent attacks [6].

Academic - The academic environment includes both primary and secondary education systems. These are important as they contain vast amount of information on a significant portion of the US population. There are finical and social risks associated with grades and student records. An additional aspect is the research done by colleges and universities that has become more intertwined with national security and government interests.

Healthcare - This sector will always be of great concern as we move to electronic health records and create a data repository that will always be attractive to advisories. There are additional government regulations in this field and the application of those regulations to prevent data breaches in this sector is of a primary concern.

Retail - The retail industry has faced numerous data breaches particularly aimed at customer financial data to include credit card information, banking information and personal information that could maliciously be used to create fake accounts.

Financial - The financial sector like the healthcare sector has numerous gov-

ernment regulations that apply to this field. Given the amount of money that could be obtained by threat actors in this domain, extra precautions need to be made in this area.
Technology - The sector is a merger of the App, Media, Gaming, Tech, Telecoms, and Web classifications from Information is Beautiful as these data types and they parent companies share a lot of similarity.
Infrastructure - This sector is a merger of the Energy and Transport classifications from Information is Beautiful and primary deals with any entity that has a focus on utilities, national commodes or infrastructure management at any government-sized level.

3.3 Method of Leak

Accidentally Published (Internal) - Is the case when any data records that were disclosed or made available to unauthorized users without any threat actor needing to be involved. This includes companies that did not password protect data or if a company posted data to an open source location. In either case, confidential data was made available to unauthorized users and not threat actor created the situation.

Inside Job (Internal) - Is the case when any data records were disclosed or made available to unauthorized users an this was done with malicious intent by a threat actor that is a part of the organization. This could be anyone from a disgruntled employee with credentials or an employed that was paid/blackmailed to compromise an internal network in order to disclose the data.
Lost/Stolen Device or Media (Internal) - This is the case the data breach was as a result of a physical loss of property that contained an instance of the data. Clearly, this is a scenario where disk encryption or encryption of the data at rest becomes very important.
Poor Security (Internal) - While a little more vague, poor security is where the organization failed to provide reasonable and expected countermeasures to protect the data. While not having a password would be accidentally published, having a very weak password would be poor security. Also in the category is failing to update and patch software as vulnerabilities are made known.
Hacked (External) - This term refers to all the external and subsequently, malicious breaches of data. There are a number of subcategories that this

term refers to, including malware, phishing, SQL injection, password cracking, social engineering and others.

3.4 Statistical Analysis

Here we will conduct statistical analysis to include clustering, principle component analysis and k-nearest neighbors

3.5 Decision Tree

Here we will present a decision tree model to show the final clustering with associated probabilities

3.6 Predictive Model

Here will will present a predictive model that will forecast a lower bound on the type and number of data breaches expected in the next year

3.7 Recommendations

Here we will itemize a list of recommendations that were a result of our analysis.

3.8 Data Breach Events

——— Some description how we organize this subsection———

Table 1: Breaches at a glance

Breach event	Data amount	Cause	Who was behind it?	Encryption technique?	Time to identify	Financial loss (%)	One time or repititive?	Mitigation technique
E-bay	145 million	Phishing	–	–	229 days	$41 million	One-time	–
Ashley Madison	15.26 millions	Flaw in encryption and no security concern	The Impact Team	MD5 and BCrypt	—	$200 million	One-time	–

Equifax	143 millions	Flaw in Apache Struts and Lack of encryption	–	Lack of encryption in personal data	Around 60 days	$20 billion	One-time	–
Deloitte Attack	Unknown to company	Single password for a large pool of data	Financially motivated Russian hackers	Lack of encryption	Long period	—	One-time	–
Sonic Attack	1 million credit card info	Yet to investigate	–	Yet to investigate	—	—	One-time	–
Hyatt Attack	250 locations	Malware	–	–	–	–	Repetitive	–
Inuvik Attack	6700	Employees	Internal employees breach	–	–	–	One-time	–
JP Morgan Chase	76 million	Application vulnerabilities	Overseas hackers	–	–	–	Repetitive	–
Friend Finder	412 million	Local file inclusion exploit	–	SHA1 + plaintext	–	–	One-time	–
UC Berkeley	80,000	Security flaw	–	–	–	–	Repetitive	–
University of Central Florida	63,000	Security flaw	–	–	–	–	One-time	–
Ubuntu	2 million	SQL injection attack	Passwords were hashed with MD5	–	–	–	One-time	–
MongoDB attack	28,000	Mis configuration	3 criminals	–	14+ days	–	One-time	–
AT&T attack	280,000	Employees	40 company employees	–	–	$25 million	One-time	–
Central Hudson Gas & Electric	110,000	Hack	–	–	–	–	One-time	–
Yale University	43,000	Unaware of change in FTP server	–	Not encrypted	10+ months	–	One-time	–
Yahoo Attack	3 billion	Cookie based attack	State sponsored attackers	BCrypt and MD5	Yahoo did not publish for 2 years	3-4 billion	–	–

3.8.1 eBay Data Breach

In 2014, hackers had infiltrated eBay systems and stolen the passwords of 145 million users. In addition to account passwords, hackers obtained names, email addresses, birthdates, physical addresses and phone numbers. eBay claimed that no credit card information nor social security numbers were lost in the compromise. The company discovered the breach after noticing several unusual behaviors on the company network. Essentially, eBay detected anomalies in their network usage statistics. The data was collected in late February and early March, 2014. EBay's mistakes also included taking days to post a notice about the breach on eBay.com and confusing users as to whether their PayPal accounts had also been affected. Cyber-attackers compromised a small number of employee log-in credentials, allowing unauthorized access to eBays corporate network. The fact that the hacking was detected so late has also enabled the hackers to check for cross-platform log-in opportunities and also sell the stolen information online.
The root cause behind the breach was phishing attack. Attackers tricked eBay employees into giving up important security credentials and then used them to infiltrate the site. Reports and forensics revealed that the attackers must have gone to sites like LinkedIn, for example, and searched for employees of eBay. Using LinkedIn they could then get important names and correlate that data with social media posts, accounts, and other sites. The employee in question would then be sent an email with an embedded link to click on. When the link was executed, malware would be installed on the computer and the attacker would gain control of the machine in question.

3.8.2 Ashley Madison

Ashley Madison is an online commercial dating website for people who are married or in a relationship [22]. In July 2015, this website was hacked. It takes around 15.26 millions of user passwords [7]. This hack had become the cause of public humiliation for many individuals. For example, around 1200 Saudi Arabian people were registered in that site with .sa email address. In Saudi Arabia, adultery can be punished with death. Several thousands US .mil or .gov emails were also used in that site. Toronto police also stated two unconfirmed suicides that had relation of this data breach.
What are the causes behind this hack? A team name "CynoSure Prime" identified the weaknesses in Ashley Madison website. The data was encrypted

using MD5 and BCrypt algorithm. However, there was programming flaws that causes the leakage. The programmer used small case letters before making the encryption which shortened the sample key-space. MD5 is usually very fast, and it was easy for the hackers to try billions of guesses within a short period of time. Also, 90% of users did not use passwords which were strong enough (only with small case characters). The details of this hack can be found in [7].

3.8.3 Equifax Breach:

Equifax is a consumer credit reporting agency which was founded in 1899. Within May to July, Equifax announced a data breach event which has impact over 143 million people. This data breach is considered to be the worst leak of personal information ever. Though the attack was started in mid-May, the breach was not observed until July 29. The attackers could able to breach personal information like first names, last names, Social Security numbers, birth dates, addresses and, in some instances, driver's license numbers.
Equifax hired Mandiant to investigate the intrusion. Equifax revealed that the primary cause of the breach is a flaw in Apache Struts. The vulnerability of Apache Struts was known from early March, however, Equifax completely failed to take necessary security steps against this vulnerability. There are other contributing factors too. Equifax had insecure network design which lacked sufficient segmentation. Moreover, there are insufficient encryption for personal information. Equifax did not have an effective breach detection system as well which costs them to identify the breach after a long time.

3.8.4 Deloitte Attack [17]:

In this recent data breach, the attackers compromised the user email data, client plans, and administrative accounts. Most importantly, a significant amount of data has been transferred or copied by the hackers. The hackers had a free reign into the system for a long time, hence, the company does not properly know how many data has been stolen.
Reason behind the attack: Deloitte did not deploy elementary security measure like two-factor verification. Deloitte also guarded a large pool of data with a single password.
Who are the attackers?: No one has claimed the responsibility, however,

based on the pattern, it seems the attack was commercially motivated and seeking confidential information to sell. This activity pattern is close to a Russian hacker group's activity.

3.8.5 Sonic Attack:

Krebs on security publishes that about one million credit card information of Sonic customers has been breached [13]. In a dark web market site, these credit card numbers are in sale with only $25 to $50. Some other credit card fraud events are also based on this Sonic attack. This attack is yet to be investigated.

3.8.6 AT&T Attack:

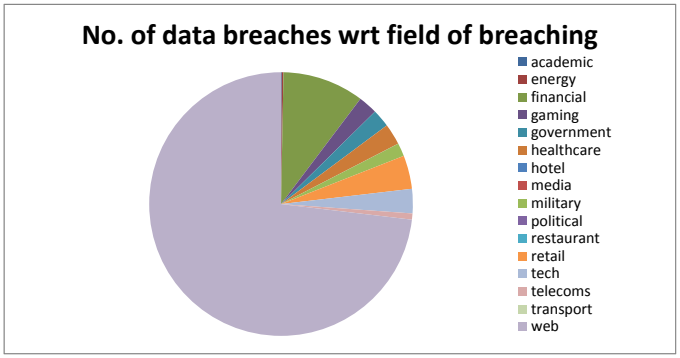
Company employees were involved in publishing sensitive data like customer name or social security number of 280,000 US customers. These data were used to unlock stolen cell phone. The company had to pay $25 million to settle an investigation with consumer privacy violation. About 40 company employees were involved in this data breach.

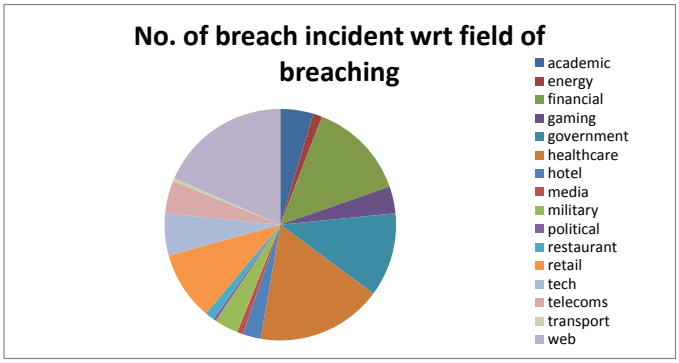
3.8.7 Yahoo Attack:

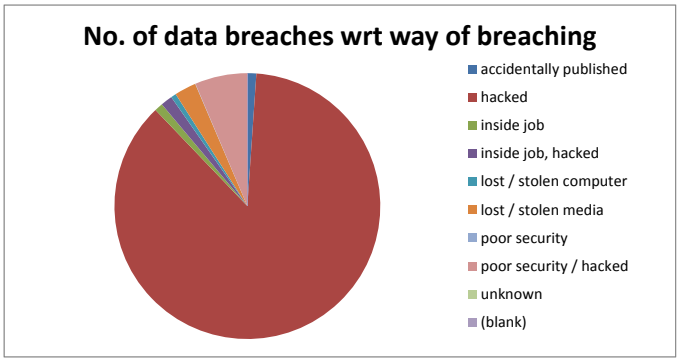
Yahoo data was breached during 2013-2014, and Yahoo reported these breaches during 2016. Yahoo claimed a state sponsor hacker group was behind this breach. Some people think the hacker group is from Russia or China. Some other people think Yahoo claims state sponsored hacker group to hide their embarrassment. The hack was done by falsify web cookies, and the hackers get access to account without even having the user-name and password. The hackers could obtain 500 million people data regarding names, email address, date of birth, security questions, and even hashed password. Security experts say some Yahoo password used bcrypt hashing algorithm which is considered to be difficult to crack, however, some other passwords use MD5, which is sometimes easy to crack. Yahoo lost its price from $4.8 billion to $350 million for this security breach.

3.8.8 Other Threats [5]

While there is a potential for searchable encryption to assist in data breaches, Chiesa and De Luca Saggese argue that there may be other issues to address







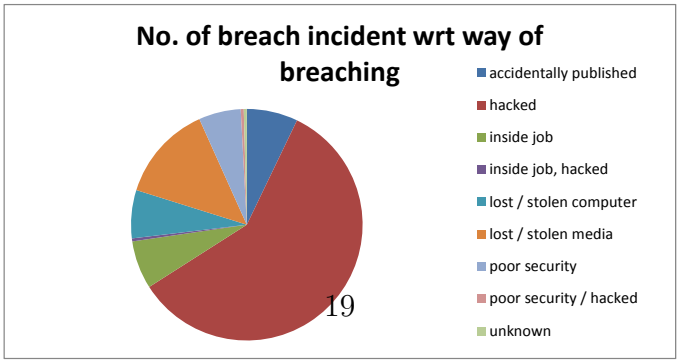


Figure 3: Breach statistics from Wikipedia page

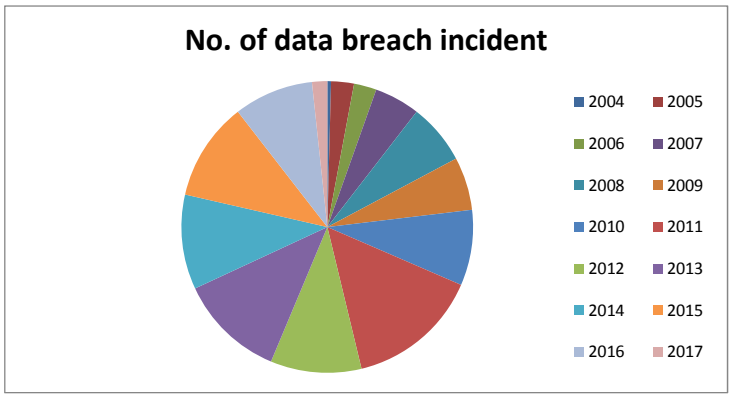
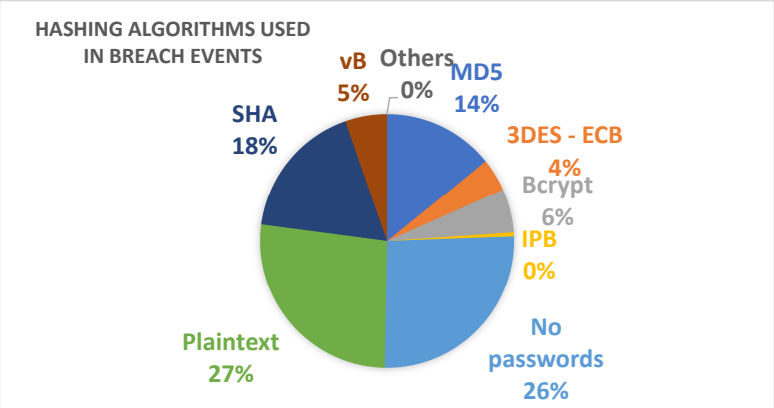


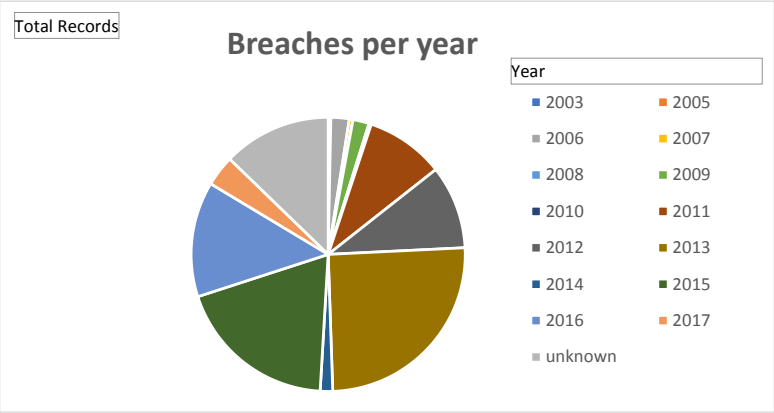
Figure 4: Wikipedia breaches by year

first. Primarily, using open and closed sources to gather Cyber Intelligence on your particular company of industry. This intelligence could provide predictive information concerning who might try to breach your data and when. Additionally, they contend that secure coding, such as applying a secure Software Life Development Cycle, could better address the SQL injection vulnerabilities and even the excessive database user grants. Shoring up these vulnerabilities may prevent data breaches in the first place.
In Table 1, we can see a set of security breaches.

3.9 Deep Dive into The Attacks

In this subsection, we will discuss about the attacks in more detail.





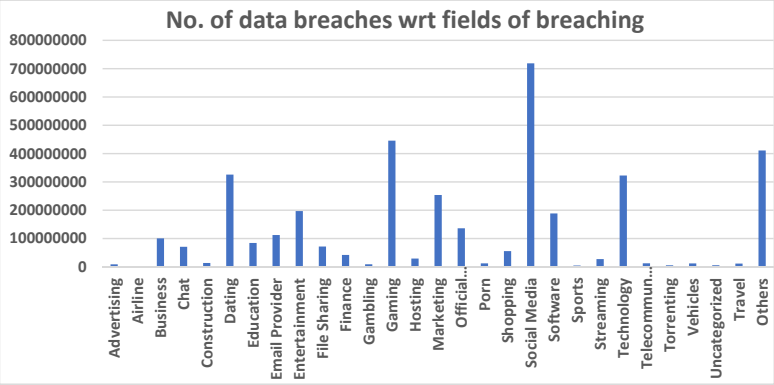


Figure 5: Breach statistics from Vigilante page

Table 2: Deep inspection over the breaches

Breach name	Phishing	Malware	Encryption	Using client's credentials?	Server side breach	Repetitive	Weak password	Internal actor	Mitigation technique
E-bay	✓	×	✓	✓	×	×	×	×	Better pass word and Dual auth
Yahoo	×	×	Password was Encryp ted, Security questions are not	×	×	✓	✓ MD5	×	Better encryp. and encryp ted security question
Equifax	×	×	✓	×	×	×	×	×	By updating Apache Struts
Inuvik hospital	×	×	×	✓	×	×	✓	×	Better internal security policy
JP Morgan Chase	×	×	×	✓	×	✓	×	×	Dual factor auth
Hyatt	×	✓	×	×	✓	✓	×	×	Better Anti-virus softwares
Friend Finder	×	×	✓ SHA1 + Plaintext	×	✓ Local file inclusion exploit	×	×	×	Encrypted Database
UC Berkely	×	×	×	×	✓ Security flaw being patched	✓	×	×	Earlier Patch release and security practices
Univ. of Central Florida	×	×	×	×	✓ Security flaw	×	×	×	Better security practices
Ubuntu	×	×	✓ MD5	×	✓ SQL injection attack	×	×	×	Better Firewalls and better access privileges
MongoDB Attack	×	×	×	×	✓ Misconfig	×	×	×	Auto mation in best known config

AT&T Attack	×	×	×	×	×	×	×	✓	Better internal security policy
Taringa	×	×	✓ MD5	×	×	×	✓	×	Better passwords
Erie County Medical Center	✓	✓	×	✓	×	×	×	×	Better Anti-virus softwares

In Table 2, we can see the details of the attacks. Here, ✓means affirmative, and × sign means information not available or negative.
We further divided the dataset into server-side and client-side datasets.
Client-side data breach: In client-side data breaches, the vulnerabilities come from the fault made by the clients who use the system.
Server-side data breach: In server-side data breaches the attackers exploit the vulnerabilities that remain in the server. We differentiate these differences in http://bit.ly/2lUUKlF. Out of 23 breaches, we found 48% were coming from client-side, and 52% data breaches were coming from server-side.

3.10 Security Breach Report

In Table 1, we describe some of the recent breaches. However, in web we can find some existing tables describing data breaches. We also consider those tables, and make statistical results over them. Fig. 3 shows results that we get from Wikipedia page [23]. From the first figure we can see that web data is breached more than the other data. From the second figure, we can see that most of the breach incidents were done with web and health care information. From the third and fourth figure, we can see that hacking is the most common reason behind data breach. Wikipedia page covers breaches from 2004 to 2017 4.
In Fig. 5 we can see the data breaches from vigilante site [21]. This site covers data breaches from 2003 to 2017. From the first figure, we can see that in 26% breaches, there was no password. Also, in 27% times the data was kept in plaintext form.

4 Related Work

4.1 Alternate View of Security

In addition to perimeter security by way of firewalls, networks need to be internally defended as well. While searchable encryption is one of the many internal security tools, that a company could use, it may not be the most pressing. ArcSight is a software that allows network security professionals to be alerted by activity inside the network that seems out of the ordinary. Whether it is too many login attempts or unusable file access from a particular device [2]. The program which is used extensively by the U.S. Department of Defense has come under scrutiny due to the company allowing Echelon, which has ties to the Russian military, from analyzing ArcSights source code [10]. Additionally, there are companies such as Darktrace [18] who are trying to leverage artificial intelligence to detect an intruder into the network because they can exploit the breach. Given that an attacker is in the network for an average of 200 days, they seek to develop a computer security immune system to monitor typical behavior and alert the security team when something seems wrong. Both of these efforts show that there is added value in defense in depth and that having a really big wall is not all that is needed for effective security. The same analogy could be made for searchable encryption, especially if it is at the sake of other efforts that will prevent a data breach in the first place. The immune system analogy to computer security is a good one as it highlights the need for external and internal controls. This is typically referred to as a defense in depth approach. After reviewing the vulnerabilities that lead to many of the recent high profile data breaches, there were some preventable measures that could have prevented the breach in the first place. In this way, searchable encryption is akin to the airbag of a car, which is to help save your life, once you are an action. However, secure coding to prevent SQL injection attacks and software updates are similar to anti-lock breaks which will help your prevent getting into a car accident in the first place.

5 Conclusion

References

[1] David Bisson. 7 Data Breaches Caused by Human Error: Did Encryption Play a Role? https://www.venafi.com/blog/7-data-breaches-caused-by-human-error-did-encryption-play-a-role, 2017. [Online; accessed 1-Oct-2017].

[2] Product Brief. Arcsight logger, simplifying log collection, storage and analysis, 2008.

[3] David Cash, Paul Grubbs, Jason Perry, and Thomas Ristenpart. Leakage-abuse attacks against searchable encryption. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 668–679. ACM, 2015.

[4] David Cash, Stanislaw Jarecki, Charanjit Jutla, Hugo Krawczyk, Marcel-Cătălin Roşu, and Michael Steiner. Highly-scalable searchable symmetric encryption with support for boolean queries. In *Advances in cryptology–CRYPTO 2013*, pages 353–373. Springer, 2013.

[5] Paolo Ciancarini, Alberto Sillitti, Giancarlo Succi, and Angelo Messina. Proceedings of 4th international conference in software engineering for defence applications: Seda 2015. volume 422, pages 261–271. Springer, 2016.

[6] CNN. First on CNN: Newly discovered hack has U.S. fearing foreign infiltration. http://www.cnn.com/2015/12/18/politics/juniper-networks-us-government-security-hack/index.html, 2015. [Online accessed on November 5, 2017].

[7] Dan Goodin. Once seen as bulletproof, 11 million+ Ashley Madison passwords already cracked. http://bit.ly/2fJxFMX, 2015. [Online; accessed 24-Sept-2017].

[8] Laszlo Hars. Discryption: Internal hard-disk encryption for secure storage. *Computer*, 40(6), 2007.

[9] Information is beautiful. List of breaches. http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/, 2017. [Online accessed on October 29, 2017].

[10] J. Stubbs J. Schectman, D. Volz. Special Report: HP Enterprise let Russia scrutinize cyberdefense system used by Pentagon. http://reut.rs/2ySna1N, 2017. [Online; accessed 8-Oct-2017].

[11] Marianne Kolbasuk McGee. Unencrypted Device Breaches Persist. https://www.databreachtoday.com/unencrypted-device-breaches-persist-a-8339, 2015. [Online; accessed 1-Oct-2017].

[12] Muhammad Naveed, Seny Kamara, and Charles V Wright. Inference attacks on property-preserving encrypted databases. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 644–655. ACM, 2015.

[13] Krebs on Security. Breach at Sonic Drive-In May Have Impacted Millions of Credit, Debit Cards. https://krebsonsecurity.com/2017/09/breach-at-sonic-drive-in-may-have-impacted-millions-of-credit-debit-cards/, 2017. [Online; accessed 1-Oct-2017].

[14] Antonis Papadimitriou, Ranjita Bhagwan, Nishanth Chandran, Ramachandran Ramjee, Andreas Haeberlen, Harmeet Singh, Abhishek Modi, and Saikrishna Badrinarayanan. Big data analytics over encrypted datasets with seabed. In *OSDI*, pages 587–602, 2016.

[15] Raluca Ada Popa, Catherine Redfield, Nickolai Zeldovich, and Hari Balakrishnan. Cryptdb: protecting confidentiality with encrypted query processing. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, pages 85–100. ACM, 2011.

[16] Privacy rights. List of breaches. https://www.privacyrights.org/data-breaches, 2017. [Online accessed on October 29, 2017].

[17] Jeff John Roberts. Deloitte attacks. http://fortune.com/2017/09/25/deloitte-hack/, 2017. [Online; accessed 25-Sept-2017].

[18] S. Rossenberg. Firewalls Don't Stop Hackers. AI Might. https://www.wired.com/story/firewalls-dont-stop-hackers-ai-might/, 2017. [Online; accessed 8-Oct-2017].

[19] Crowd sourced data. Breach level index. http://breachlevelindex.com/, 2017. [Online; accessed 24-Sept-2017].

[20] Verizon. Verizon annual report, 2017. rp_DBIR_2017_Report_en_xg.pdf, 2017. [Online; accessed 24-Sept-2017].

[21] Vigilante. List of breaches. https://vigilante.pw/, 2017. [Online accessed on October 24, 2017].

[22] Wiki. Ashley Madison. https://en.wikipedia.org/wiki/Ashley_Madison, 2017. [Online; accessed 24-Sept-2017].

[23] Wikipedia. List of breaches. https://en.wikipedia.org/wiki/List_of_data_breaches, 2017. [Online accessed on October 24, 2017].

[24] Ying Zhang. *Future wireless networks and information systems*. Springer, 2012.

[25] Wenting Zheng, Frank Li, Raluca Ada Popa, Ion Stoica, and Rachit Agarwal. Minicrypt: Reconciling encryption and compression for big data stores. In *Proceedings of the Twelfth European Conference on Computer Systems*, pages 191–204. ACM, 2017.