Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol 2, No.4, 2012

A Foundation for Breach Data Analysis

Adewale O Adebayo

School of Science and Technology, Babcock University, P.M.B.21244, Ikeja, Lagos, Nigeria

*E-mail of corresponding author: adebayoa@babcock.edu.ng

Abstract

Analysis of breach data would provide fresh insight into the subject of Information Security. There is, therefore the need to examine breach data repositories on which the analyses would be based. This work set to identify and describe breach data repositories, and highlight whatever is revealed. The use of Datalossdb.org data set as a base to be enriched for open verifiable analysis of the breach data was supported, and the necessity of a common vocabulary for describing security breach related issues was revealed.

Keywords: data breach, breach data, breach data database, breach data repository, breach data reports

1. Introduction

Information Security (IS) is ensuring that the information systems perform according to stipulation and retain optimum performance in the face of smart adversaries. Some generally visible failures of IS are spam and associated problems (Helbush, 2009), malicious codes (Eichin & Rochlis, 1989), bugs in software including operating systems (Keizer, 2010), and data breaches (Aitoro, 2007). It is, therefore, not particularly clear how effective and efficient the measures to IS issues are. Analysis of factual data is needed to provide fresh perspective to the subject of IS.

1.1 Problem Statement

Doing any survey in IS in a scientifically defensible manner is very challenging. Fourteen security surveys that were the most widely publicized from 1995 to 2000 were found to be replete with design errors in the areas of sample selection, the form of questions asked, and underlying methodologies (Ryan & Jefferson, 2003). On the surface, the breach data offers widely spread, unbiased, and easily accessible data for analysis that would provide varied insight into issues surrounding data breaches. A number of works have been done analyzing breach data (Lewis, 2003; Tehan, 2005; Acquisti, et al, 2006; Hasan & Yurcik, 2006; Culnan & Williams; 2009; Gordon, et al, 2010). No work yet examined breach data repositories on which analysis is based.

There is the need to examine the various repositories of the breach data before analyzing the breach data.

1.2. Goal and Objectives

The goal of this work is, therefore, to lay a foundation on which may be based a study of the breach data. In this light, the specific objectives are to identify and describe the main repositories of breach data, and to highlight what is revealed in examining the repositories.

Coincidental applicable research questions are: What is data breach and breach data? Where are and who are sponsors of the main repositories of breach data? What kinds of information do breach data repositories contain and from which sources? What is revealed in examining breach data repositories?

1.3. Significance of the Study

This work supports a more secure computing environment by laying a foundation on which could be based a study of the breach data in order to extract useful information to provide fresh perspective to the subject of IS. It is important

that the data on which analysis and conclusions would be drawn be authentic, reliable and relevant, and this study examines these aspects and sheds some light.

1.4. Methodology

An extensive literature review was performed. Search terms data breach, breach data database, breach data repository, breach data report, and breach data statistics, were used on search engines, and related results examined. These were complimented with snowballing non-probabilistic sampling. The relevant documents obtained were qualitatively analyzed for convergence, and relevant details were extracted, using inductive approach.

2. Outcomes

A data breach is an incident in which sensitive, protected or confidential data has potentially been viewed, stolen or used by an individual unauthorized to do so. Data breaches may involve personal health information (PHI), personally identifiable information (PII), trade secrets or intellectual property. A number of industry guidelines and government compliance regulations in United States of America (US) mandate strict governance of sensitive or personal data to avoid data breaches. ([Online] Available: http://searchsecurity.techtarget.com/definition/data-breach, 8/4/12). Incidents range from concerted attack by black hats with the backing of organized crime or national governments to careless disposal of used computer equipment or data storage media ([Online] Available: http://en.wikipedia.org/wiki/Data_breach, 8/4/12).

Breach data is generated as a result of reports of data breaches. The catalyst for reporting data breaches to the affected individuals has been the US California law that requires notice of security breaches implemented July 2003 ([Online] Available: http://www.ncsl.org/default.aspx?tabid=13489, 8/4/12). More than forty of US states have since passed laws requiring that individuals be notified of security breaches (Attrition, 2011; PrivacyRights, 2011). There is typically no public disclosure requirements in the US state laws (except Massachusetts) and disclosure laws have not been actively and uniformly enforced. Nearly all state laws provide an exemption for breach disclosure if the personal data was encrypted, with a very small chance of it being broken, at the time of the compromise (Mueller, 2006). A wide range of organisations have disclosed their storage security breaches in the mass media.

No comprehensive data source on storage security breaches exists because there is no requirement for public reporting in the US state laws. Strong economic reasons for organisations not to publicly report storage breach include damage to reputation, loss of current/future customers, liability from other state's laws, and possible lawsuits from shareholders/customers (Hasan and Yurcik, 2006). However, there are lengthy lists of breach incidents maintained on a growing number of websites (Privacyrights, 2011). A major repository must necessarily be visible, popular or well known, and easily accessible or web-based, or law required or enforced. The relatively visible ones are presented in succeeding paragraphs in alphabetic order.

2.1 Attrition.org

Attrition.org is a computer security web site dedicated to the collection, dissemination and distribution of information about the security industry. Attrition.org maintains the largest mirror of Web site defacements and was party to the creation of the Data Loss Database (Open Source), which eventually became DatalossDB. Attrition.org is a privately owned and operated system ([Online] Available: http://attrition.org/attrition/, 8/4/12).

2.2 Computer Emergency Response Team

Computer Emergency Response Team (CERT-Carnegie), a US federally funded research and development centre at Carnegie Mellon University (CMU) has the mission to enable the survival of critical networked systems against contemporary threats and attacks by removing technical, maturity, information, and capacity barriers in cyber security and incident response. The overall goal of CERT program is improved practices and technologies that are

widely understood and routinely used to protect, detect, and respond to attacks, accidents, and failures on networked systems (Shannon, 2011). Its breach data set is not readily available. CERT Research by the Numbers, however, indicates over five hundred cases in its Insider Threat database (2010 CERT report [Online] Available: http://www.cert.org/research/researchreport.html, 8/4/12).

2.3 Databreaches.net

Databreaches.net site ([Online] Available: http://www.databreaches.net, 8/4/12) began in 2009 as a spinoff from PogoWasRight.org. It receives no funding or financial support. This site keeps data breach blogs, and stores news about breach incidents using as file index a combination of organisation type, (business, education, financial, government, health care, miscellaneous), breach type (exposure, hack, insider, lost/missing, malware, other, paper, skimmers, subcontractor, theft, unauthorized access), whether of note and whether US or non-US data breach.

2.4 DataLossDB

The Open Security Foundation DataLossDB is a research project aimed at documenting known and reported data loss incidents world-wide. The effort is now a community one accepting contributions of new incidents and new data for existing incidents ([Online] Available: http://**data**lossdb.org/, 8/4/2012*)*. This repository keeps information about incidents involving the breach of PII, when an organisation is responsible for the mishandling of PII that result in a breach. The breach data set includes the who, the what and the where, Breach Types, Data Type, Data Family, Address where the breach occurred, Whether there was an arrest, Whether the Data were recovered, Whether there is /was a lawsuit, The When, References, and other information ([Online] Available: http://datalossdb.org/submissions/new, 8/4/12).

2.5 Identity Theft Resource Centre

Identity Theft Resource Centre has a mission is to provide best in class victim assistance at no charge to consumers in US, to educate on best practices for fraud and identity theft detection, reduction and mitigation ([Online] Available: http://www.idtheftcenter.org/about/mission.html, 15/4/12). This site keeps data breaches information, which includes Company Name/Agency, State in US, Date incident was established, breach Category, number of records exposed, brief description of the incidence, and source(s) (or hyperlink to source(s)) of the information. This centre has a record of 3 248 breach incidents spanning 2005 to March 2012. ([Online] Available: http://www.idtheftcenter.org/artman2/publish/lib_survey, 30/3/12). In 2011, the project was supported by a grant of the Office for Victims of Crime, Office of Justice Programs, US Department of Justice ([Online] Available: http://www.idtheftcenter.org/, 30/3/12).

2.6 InfosecurityAnalysis.com

InfosecurityAnalysis.com keeps summary of data breach incidences showing name of company/agency, number of records lost/exposed, and a summary of how the incidence occurred. It has about 1 229 incidence records spanning years 2000 to 2008. This site allows data breach incidence sorting by year and by industry, separately. Its sources of incidents are DatalossDB statistics and summary information, Privacy Rights Clearinghouse and attrition.org ([Online] Available: http://www.inforsecurityanalysis.com, 15/4/12).

2.7 MyID.com

MyID.com is an online privacy and identity monitoring, protection and alerts tool designed to help effectively protect privacy, finances, reputation and safety online ([Online] Available: http://www.myid.com/about, 15/4/12). The site has in archive data breach blogs spanning March 2011 to March 2012 ([Online] Available: http://www.myid.com/blog/2011/03/ & http://www.myid.com/blog/2012/03/, 8/4/12).

2.8 National Association for Information Destruction

National Association for Information Destruction, Incorporated (NAID) is the international trade association for companies providing information destruction services. NAID's mission is to promote the information destruction industry and the standards and ethics of its member companies ([Online] Available: http://www.naidonline.org/nitl/en/, 1/4/12). NAID provides monthly newsletters that include a number of data breaches largely due to improper document destruction.

2.9 Personal Health Information Privacy

Personal Health Information Privacy (PHI Privacy**)** in affiliation with Databreaches.net, compiles and stores healthcare-related breaches or data loss incidents (2003 to 2008 incidents were available). Many of these were obtained from the US Department of Health and Human Services' medical data breach list ([Online] Available: http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/postedbreaches.html, 1/4/12), which provides only minimal information. PHIprivacy.net allows site visitors to comment about news articles and to submit health-related privacy breaches or security breaches that have not been reported in the media ([Online] Available: http://www.phiprivacy.net/?page_id=2, 1/4/12).

2.10 PogoWasRight.org

PogoWasRight.org is a private attempt to increase awareness and expand the privacy news and issues coverage. This site does not accept any advertising or political sponsorship ([Online] Available: http://www.PogoWasRight.org/?page_id=5661, 29/3/12). This site keeps privacy news from around the world. A number of data breach news from April 2006 to January, 2009 can be found on archive.PogoWasRight.org.

2.11 Privacy Rights Clearinghouse

Privacy Rights Clearinghouse (PRC) is a non-profit consumer organization with consumer information and consumer advocacy mission. It is primarily grant-supported and serves individuals in US ([Online] Available: https://www.privacyrights.org/about_us.htm#goals, 29/3/12). This repository keeps records of data breaches that expose individuals to identity theft as well as breaches that qualify for disclosure under US state laws. PRC data set includes the Type of Breaches, Organization Type(s), the Size of Breach, Date of Incidence, and verifiable Source of Report ([Online] Available: http://www.privacyrights.org/data-breach, 29/3/12). Only reported incidents affecting more than nine individuals from an identifiable entity are included except there is a compelling reason to alert consumers. Most of the information is derived from the Open Security Foundation list-serve. Other sources, beginning in January 2010, include Databreaches**.**net, PHI Privacy and NAID ([Online] Available: http://www.privacyrights.org/about_us.htm, 29/3/12).

2.12 United States Computer Emergency Response Team

United States Computer Emergency Response Team's (US-CERT's) mission is to improve US cyber-security posture, coordinate cyber information sharing and proactively manage cyber risks to the nation while protecting the constitutional rights of Americans. This repository has information about computer incidents which include but not limited to attempts (either failed or successful) to gain unauthorised access to a system or its data. US-CERT breach data Set includes Agency name, Point of Contact Information, Incident Category, Incident Date and Time, Source Internet Protocol (IP), Port, and protocol, Destination IP, Port, and Protocol, Operating System, Location of the system involved in the incident, methods used to identify the incidents, Impact to agency and Resolution ([Online] Available: http://forms.us-cert.gov/report & http://www.us-cert.gov/federal/reportingrequirements.html, 22/3/12). This data is not readily available.

2.13 U.S. Securities and Exchange Commission

U.S. Securities and Exchange Commission (US-SEC) is responsible for implementing a series of regulatory initiatives required under the Dodd-Frank Wall Street Reform and Consumer Protection Act. The Commission has proposed or adopted a number of rules in connection with the Dodd-Frank Act that makes it a repository of certain breach data ([Online] Available: http://www.sec.gov/spotlight/dodd-frank.shtml & http://www.sec.gov/spotlight/dodd-frank/accomplishments.shtml, 15/4/12). The breach data kept by this repository is not readily available.

2.14 Verizon

Verizon keeps data of Verizon confirmed compromise involving organisational data breaches as revealed in Verizon Data Breach Investigation Reports (DBIRs). The data used in the 2012 DBIR includes breach data contributions from US Secret Service, Dutch National High Tech Crime Unit, Australian Federal Police, Irish Reporting and Information Security Service CERT, and Police Central e-Crime Unit of the London Metropolitan Police. The 2012 DBIR caseload is over two thousand (2 000) spanning eight years, 2004 to 2012 ([Online] Available: http://www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2012_en_xg.pdf, 15/4/12). This data set is not readily available to the public for scrutiny.

2.15 Review and Evaluation of Breach Data Repositories

Privacy Rights Clearinghouse, Identity Theft Resource Centre, and DatalossDB.org data sets have large number of data breach incidents and cover wide details of each breach. They are also open and available to researchers. DatalossDB.org in particular outstandingly offers the largest number of reported incidents and wider data types.

The absence of a framework for collecting and classifying security incident information in a common language and structure is obvious.

Table 1 below sheds more light on the repositories discussed in Sections 2.1 to 2.14.

3. Related Works

No existing work examining data breach repositories was found. Nevertheless, the following works are worth noting.

A summary of selected storage security incidents reported in the press between 2000 and 2005 was conducted (Tehan, 2005). In this study, a small data set of incidents was used and biased sampling could have occurred. At this time, the repositories were just growing.

A study that claimed to be the first valid statistical analysis of disclosed storage security breaches used combined data set spanning January 1, 2005 to June 5, 2006, from DatalossDB.org and Privacy Right Clearinghouse (Hasan and Yurcik, 2006). It rightly claimed that DatalossDB.org and Privacy Right Clearinghouse were leading repositories as this work supports.

2011 Data Breach Investigation Report by Verizon Risk Team, US secret service and Dutch High Tech Crime Unit (April, 2010) data set included only Verizon confirmed incidents of data compromise involving deliberate breach and compromise situations ([Online] Available: http://www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2011_en_xg.pdf, 15/3/12). The setback is that the data set is secretive and the work, therefore, not repeatable.

Verizon offers a VERIS framework designed to provide a common language for describing security incidents in a structured and repeatable manner ([Online] Available: http://www.verizonbusiness.com/resources/whitepapers/wp_verizon-incident-sharing-metrics-framework_en_xg.pdf, 23/4/12). The complete framework can be obtained from the VERIS community wiki ([Online] Available: https://verisframework.wiki.zoho.com/VERIS-framework.html, 23/4/12). Its refinement would be worthwhile toward the common language.

4. Conclusions

A data breach is an incident in which sensitive, protected or confidential data has potentially been viewed, stolen or used by an individual unauthorized to do so. Breach data is generated as a result of reports of data breaches. No comprehensive data source on storage security breaches exists because there is no requirement for public reporting in the US state laws.

This work supports the use of Datalossdb.org data set as a base to be enriched with the others (notably Privacy Rights Clearing House) by the addition of founded incidents possibly omitted and by expanding its reporting details with offered additional information, for open verifiable analysis of the breach data.

The need for a common vocabulary for describing security breach related issues is vivid. A US federal data breach notification and reporting law could achieve this with attendant benefits, though many argue that it is untenable ([Online] Available: http://www.myid.com/blog/the-debate-over-data-notification-laws-returns/, 23/4/12). Welcome other efforts towards the common vocabulary are also on the way.

References

Acquisti, A., Friedman, A., and Telang, R.(2006). Is there a cost to privacy breaches? An event study. In Workshop on the Economics of Information Security, 2006.

Aitoro, J. (2007). Reports of federal security breaches double in four months. Government Executive.com, October 23, 2007, [Online] Available: http://www.govexec.com/dailyfed/1007/102307;1.htm, 11/11/10.

Attrition. (2011). Entities that suffer large personal data incidents (list). [Online] Available: http://attrition.org/errata/dataloss, 16/3/12

Culnan, M J, and Williams, C C. (2009). How Ethics Can Enhance Organizational Privacy: Lessons from the ChoicePoint and TJX Data Breaches. MIS Quarterly December 2009, Vol. 33, Issue 4 (pp. 673-687)

Eichin, M and Rochlis, J. (1989). With Microscope and Tweezers: An analysis of the Internet virus of November 1998. 1989 IEEE Symposium on Research in Security and Privacy, [Online] Available: http://www.mit.edu/people/eichin/virus/main.html, 15/11/10.

Gordon, L A, Loeb, M P, and Sohail, T. (2010). "Market Value of Voluntary Disclosures Concerning Information Security." MIS quarterly Vol. 34, No. 3

Hasan, R., and Yurcik, W. (2006). A Statistical Analysis of Disclosed Storage Security Breaches. International Workshop on Storage Security and Survivability: in conjuction with 12th ACM Conference on Computer and Communications Security, October, 2006.

Helbush, A. (2009). Phishing Attacks Still on the Rise. Where to Start Technology Solutions Blog, [Online] Available: http://www.wtsci.com/2009/11/Phishing-attacks-still-on the rise/, 11/11/10.

Keizer, G. (2010). Apple Smashes Patch Record with gigantic Update. [Online] Available: http://Computer World.com/s/article/9196118/Apple_smashes_patch_record_with_gigantic_update, 5/11/10.

Lewis, M. (2003). Moneyball: The Art of Winning an Unfair game. New York – W.W. Norton & Company, Inc

Mueller, P. (2006). How to survive data breach laws. Network Computing, June 8, 2006.

Privacyrights.(2011). A chrology of data breaches reported since the Choicepoint incidence (list). Privacy Rights Clearing house. [Online] Available: http://www.privacyrights.org/ar/ChronDataBreaches.htm, 30/3/12.

Ryan, J C H, and Jefferson, T I. (2003). The Use, Misuse and Abuse of Statistics in Information Security Research. Proceedings of the 2003 ASEM National Conference, St. Louis, Missouri.

Shannon, G. (2011). CERT Research Vision, 2010 CERT Research Report. ([Online] Available: https://forms.us-cert.gov/report/, 29/3/12).

Tehan, R. (2005). Personal Data Security Breaches: content and incident summaries. In Congressional research

Service Report for Congress, December 16, 2005.

Table 1 - Summary of Breach Data Repositories

REPOSITORY	DATA SET TYPE	NO. OF INCIDENTS	YEARS COVERED	SPONSORS	OPEN & VERIFIABLE
Attrition.org		not available		Private	NO
CERT-Carnegie	General	not available	2006 - date	Part US Govt	NO
Databreaches.net	Blogs	not available	2009 - date	Private	YES
DatalossDB	General	5294	2003 - date	Private & corporate	YES
Identity Theft Resource	General	3248	2005 - date	Foundations & Corp.	YES
Infosecurityanalysis.com	summary	1229	2000 - 2008		
MyID.com	Blogs	not available	2011 - 2012	Private	YES
NAID					
PHI Privacy	Health-care	488	2003 - 2008	Private	YES
PogoWasRight.org	Privacy news	not available	2009 - date	Private	YES
Privacy Rights ClearingHo.	General	3009	2005 - date	Found., Corp. & Private	YES
US-CERT	General	not available	2007 - date	US Govt. Dept.	NO
US-SEC	General	not available		US Govt. Dept.	NO
Verison	General	>2000	2004 - 2012	Private & Govt	NO

(Corp .= corporation; found. = foundation; Govt. = Government)