International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

STORIES UNFOLD: ISSUES AND CHALLENGES ON DATA PRIVACY USING LATENT DIRICHLET ALLOCATION ALGORITHM

Jeffrey C. Cinco

Leyte Normal University, Philippines

ABSTRACT

An emerging trend in the cyberspace is called data privacy that is often unattended and ignored by Netizens. This paper looks into available documents found in journals, news articles and published reports to understand and uncover hidden issues and challenges about data privacy. These documents were extracted and processed using web mining technique and methods using topic modeling of Natural Language Processing (NLP). This study utilized Latent Dirichlet Allocation Algorithm (LDA) using R-programming. LDA is an unsupervised machine learning that has the capability to generate latent themes from the collected text documents or corpus. As a result, five latent themes were identified based on the result and a profound understanding of how data privacy affects the internet society nowadays.

Keywords: Data Privacy, Privacy, Unsupervised Machine Learning, Web Mining, Latent Dirichlet Allocation Algorithm

INTRODUCTION

Access to information has increased over time. Information on the cloud is becoming more public and accessible to netizens. Over time, the internet became a venue for expressing once thought, likewise became a venue for the exchange of information. The internet is one of the easiest and fastest ways of communication as internet access became a public information service (UNESCO, nd). However, as the means of communication improved, information privacy became a hot topic as well. The study examines the global views and perceptions about data privacy as a basis for policy redirection. Data privacy also defined as data protection is about protecting the fundamental right to privacy protected by international and local laws and conventions. It is a law designed to protect one's information that is collected, processed and stored by automated or part of a filing system (Privacy International, nd). Data protection may in happen in printed documents or an electronic one. In spite of the data protection law implemented by different countries, some issues were still encountered by netizens. The data privacy day (DPD) was set on every 28th day of January to raise awareness on this matter. DPD

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

empowers people to protect individual privacy, control digital footprint and escalate the protection of privacy and data as a priority. It aims to increase awareness of privacy and data protection issues among consumers, organizations, and government officials (Mississippi Department of Information Technology Services, nd). The government in different countries crafted data protection law that governs the access to information. In the Philippines, "it protects the privacy of individuals while ensuring the free flow of information to promote innovation and growth. It also regulates the collection, recording, organization, storage, updating or modification, retrieval, consultation, use, consolidation, blocking, erasure or destruction of personal data; and ensures that the Philippines complies with international standards set for data protection through National Privacy Commission (NPC)" (National Privacy Commission, nd).

In this paper, the author explored the different issues and challenges on data privacy using web mining technique and processes using topic modeling of Natural Language Processing. Content analysis of public documents was made using Latent Dirichlet Allocation Algorithm.

RESEARCH QUESTIONS

The study explored the issues and challenges of data privacy. Specifically, this it aims to answer the following questions:

1. What are the latent themes generated from online documents?
2. What countries that are actively talking on the web about this issue?
3. Base on the findings of the study what possible intervention and strategies to avoid data privacy.

THEORETICAL FRAMEWORK

The paper used theoretical underpinning that correlates to the research processes and procedures. A systematic research theory of Sandra Petronio known as communication privacy management (CPM) theory formerly known as communication boundary management, designed to develop an evidence-based understanding of the way people make decisions about revealing and concealing private information. The theory suggests that privacy boundaries need to maintain and coordinate individuals with various communication partners reliant on the supposed benefits and overheads of information disclosure. Petronio uses a border metaphor to explain the privacy management process. Privacy boundaries draw divisions between private information and public information. This theory argues that when people disclose private information, they depend on a rule-based management system to control the level of accessibility. An individual's privacy boundary governs his or her self-disclosures. Once disclosed, the negotiation of privacy rules between the two parties is required. An identified clashing of expectations for privacy management causes boundary turbulence. The conceptual image of Petronio's protective boundaries is essential to

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

understanding the five core principles of Petronio's CPM. First, people believe they own and have a right to control their private information. Second, people have control to their private information through the use of personal privacy rules. Third, other people became co-owner of one's individual information when are told or given access. Fourth, co-owners of private information need to negotiate agreeably social to privacy rules about saying others. Lastly, when co-owners of private information doesn't efficiently arrange and follow jointly held privacy rules, boundary turbulence is the likely result (Revolvy, nd).

METHODOLOGY

Research Design

This study utilized sequential exploratory design using content analysis on online documents. The study imposes quantitative and qualitative which allows the theoretical perspective of the researcher to guide the research and determine the order of data collection. Data mining was the critical component in the collection of data. It was used to generate new information from a significant amount of database (**Priyadharsini and Thanamani, 2014**). Data mining is the crucial part of the Knowledge Discovery Database(KDD) process. KDD is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data, where data is a set of facts. KDD process involves numerous interactive and iterative steps with many decisions made by the user (Fayyad, Shapiro, and Smyth, 1996).

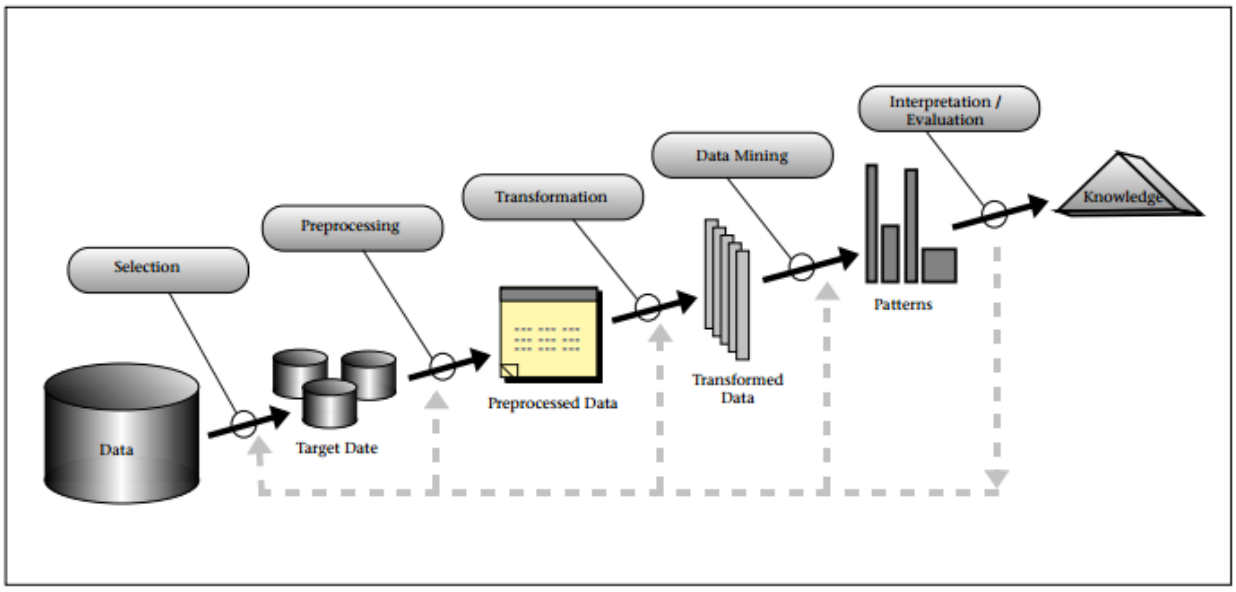


Figure 1. An Overview of the Steps That Compose the KDD Process.

Source: https://ww.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

KDD includes the following process: **Selection**, google trends was used to determine the pattern and the countries actively talking data privacy. Data from online journals, news article or published documents were retrieved from the top ten nations; **Preprocessing** cover the removal of stop words and text formatting which includes the cleaning by removing unnecessary texts; **Transformation** of documents collected where each document collected was saved in text format using notepad or another text editor. **Data Mining** techniques were utilized to discover patterns based on the frequency of terms in the data set. Latent Dirichlet Allocation Algorithm (LDA) was employed using R-programming to generate latent terms from the collected text documents. Each word was classified into one of the emerging topics. Terms were clustered into groups whose members are similar in the same way; **Interpretation/Evaluation** of patterns from the documents collected was employed to generate latent themes. Gibbs sampling was used to check the result reliability.

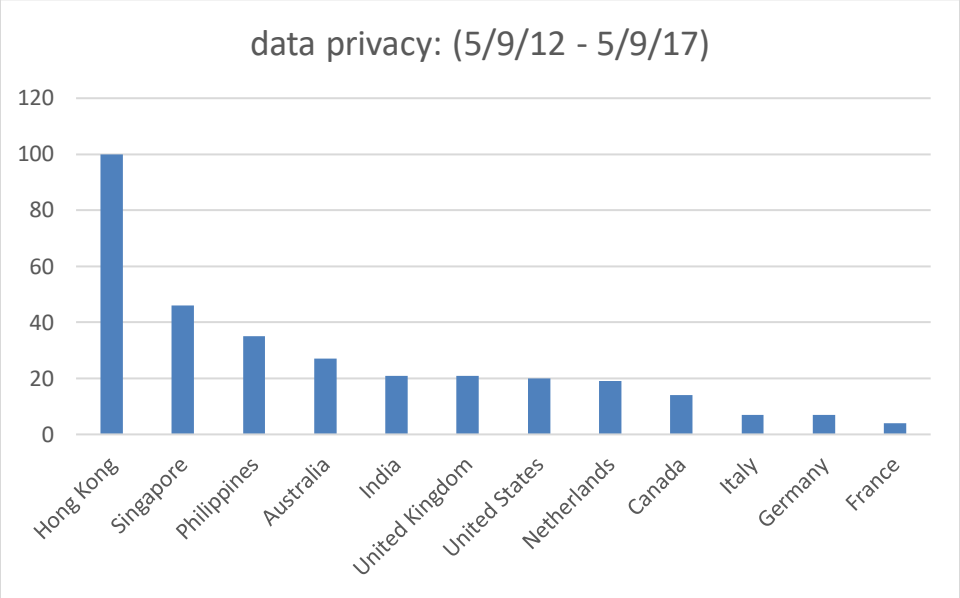
Research Method

The study anchored the approaches and processes on web mining: information and pattern discovery on the world wide web of Cooley, Mobasher, & Srivastava (1997).

Document from the top Ten (10) countries talking about data privacy were collected and processed using web mining technique and methods using topic modeling of Natural Language Processing (NLP). This study utilized Latent Dirichlet Allocation Algorithm (LDA) using R-programming. LDA is an unsupervised machine learning that can generate underlying themes from the collected text documents or corpus. The figure below shows the top ten (10) countries talking about data privacy.

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"



Source: Google Trends

Figure 1

After collection of documents from online sources, Latent Dirichlet Allocation Algorithm was used to generate latent topics for creating latent themes. In this study, five (5) topics have been set up with Ten (10) latent terms in each group which then analyzed to create latent themes. Each Topic will be checked by latent distribution probability using Gibbs sampling. Further, this paper utilized various software to generate the output using Latent Dirichlet Allocation Algorithm such as:

- Google Trend, for identifying the top ten (10) countries that contribute to data privacy for the past five years.
- Rstudio and R-programming, for LDA algorithm application using Python programs like gibbs sampler and lda.

Ethical Consideration

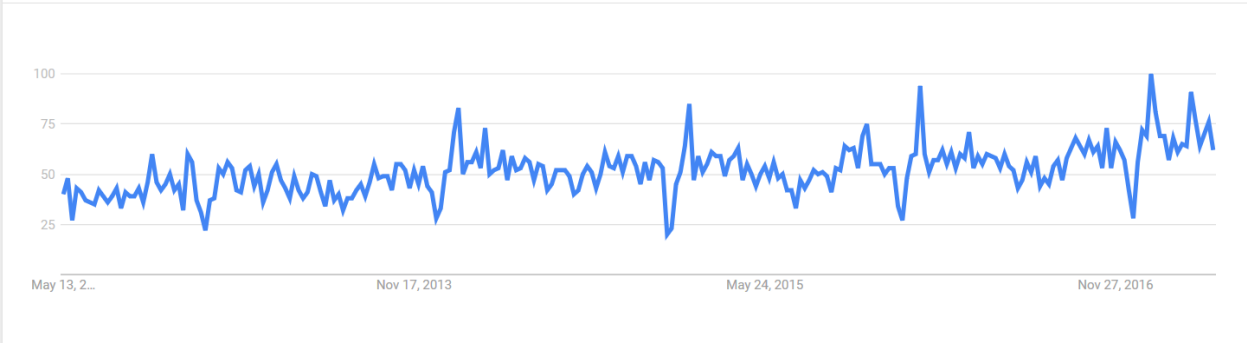
The information used in this paper were raw data taken from published articles, news networks website, and online journals from the top ten (10) countries that contribute most to data privacy. The researcher also wishes to remind the readers that the result of this study would need further evaluation.

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

RESULTS AND DISCUSSION

Topic: Data Privacy (Worldwide)



Source: Google Trends

Figure 2

The figure above shows the trends of documents and conversations talking about data privacy over the last five years. This occurrence happens in every part of the globe by writing their views and opinions on the web. The figure shows the increasing rate of discussion on data privacy. An increased percentage of discussion during last week of December up to the first week of February can be observed; these dates are close to data privacy day. January 28 is data privacy day which is one contributor to this occurrence. In January 2008, data protection day started in the United States and Canada as an extension of the data protection day in Europe. In every 28th day of January, an annual effort is made to raise awareness about the importance of privacy and to protect personal information (energy.gov, nd).

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

Table 1: Document to Topics

#	Document	Topic
1	Protection of big data privacy.txt	2
2	Security and Privacy Concerns Raised Around Internet of Things.txt	1
3	Comparative Study on Different Approaches to Privacy Challenges in Particular in the Light of Technological Developments.txt	5
4	Big data privacy a technological perspective and review.txt	2
5	Data Security and Cybercrime in Italy.txt	4
6	Big Data Privacy and Security.txt	1
7	Privacy Body Probes new Comelec Data Breach.txt	1
8	Singapore Issues New Regulations in Advance of Data Protection Law Entering Into Force.txt	4
9	Online Privacy Law United Kingdom.txt	4
10	How not to protect genomic data privacy in a distributed network using trail reidentification to evaluate and design anonymity protection systems.txt	3

Table 1 presents the documents collected from the web from the top 10 countries talking about data privacy and were transcribed using LDA. There were ten (10) documents gathered from the top 10 countries that contribute to data privacy. Each document is talking about a particular topic. Example, *"Security and Privacy Concerns Raised Around Internet of Things"* text file is talking about topic 1. The same theme is discussed in "*Big Data Privacy and Security"* and "*Privacy Body Probes new COMELEC Data Breach"* text files; Topic 2 is discussed in *"Big data privacy a technological perspective and review"* text file while topic 3 is the main theme in *"How not to protect genomic data privacy in a distributed network using trail reidentification to evaluate and design anonymity protection systems"* text file. "*Data Security and Cybercrime in Italy"*, "*Singapore Issues New Regulations in Advance of Data Protection Law Entering Into Force"*, and "*Online Privacy Law United Kingdom"* documents has another topic discussed labeled as topic 4 and another topic which is topic 5 is the main point of discussion in *"Comparative Study on Different Approaches to Privacy Challenges in Particular in the Light of Technological Developments"* document.

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

Table 2: Word Frequency

Word	Freq		Word	Freq		Word	Freq
data	1919		european	62		integrity	47
privacy	514		new	60		owner	47
information	292		research	60		systems	47
big	270		methods	59		legislation	47
personal	260		proposed	59		country	46
protection	217		required	58		without	45
cloud	138		genomic	55		authority	45
processing	138		service	54		electronic	45
security	134		table	54		enforcement	45
public	125		third	54		algorithm	44
access	112		rights	54		preserving	44
commissioner	111		collection	53		section	44
number	104		protect	53		analytics	43
storage	104		provide	53		identity	43
subject	93		commission	53		need	43
law	92		consent	53		system	43
reidentification	83		collected	52		report	43
ordinance	80		communications	52		available	42
techniques	76		case	50		card	42
individuals	75		encryption	50		credit	42
different	71		model	50		direct	42
users	71		records	50		image	42
code	70		order	49		legal	42
trail	68		party	49		location	42
private	66		government	49		identified	42
breach	64		provided	49		given	41
individual	63		general	48		large	40
sensitive	63		process	48		provides	40
however	62		study	48		stored	40
purposes	62		anonymization	47		values	40

Using LDA, contents from the documents in Table 1 were extracted, and frequencies of each term were determined. Table 2 shows the frequency of words found in the documents. Out of 5,847 words excluding words like “fig, made, much, also, may, and other words that have no

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

meaning in the context of data privacy" found in the documents. Top 90 words were presented from most frequent to the least frequent word. It is significantly observed that the frequency of words contributes to the topic that can be derived.

Table 3: Topics to Terms

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	privacy	data	data	data	Privacy
2	security	privacy	reidentification	personal	Commissioner
3	provide	Big	trail	protection	Ordinance
4	breach	cloud	information	information	Public
5	issue	information	number	subject	Law
6	internet	storage	genomic	processing	Government
7	address	techniques	image	access	Card
8	business	security	research	code	Credit
9	management	different	location	purposes	Personal
10	protect	sensitive	patients	rights	European
11	case	anonymization	model	consent	Study
12	issues	integrity	identified	communications	Commission
13	number	processing	hospital	provided	Data
14	year	proposed	table	legislation	Complaints
15	disclosure	encryption	track	authority	Country
Latent Themes	**Provision of Data Security and Privacy Against Business Breach**	**Increased Productivity on Big Data Security and Integrity Innovation**	**Health Issues on Data Breach**	**Laws on Data Privacy and Personal Protection**	**Issues on the Implementation of Data Privacy Law Against Identity Theft and Security**

Table 3 shows the latent terms defined from the documents collected from the web. It is also noticeable that the latent terms specified in table 3 can be found in the most frequent terms in table 2. LDA algorithm was set to determine five latent topics. Each latent topic consists of fifteen (15) words to fall under each topic. Latent themes were identified from the five latent topics using the latent terms in above table and the philosophical perception of the researcher. Topic 1 has the theme *"Provisions of Data Security and Privacy Against Business Breach."*

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

Businesses nowadays collect data from customers to meet their needs and to further improve products for maximum customer satisfaction. The protection of proprietary data and personal information is at the forefront of every organization's planning process (Data Privacy and Cybersecurity, nd). The second group of terms talks about "*Increased Productivity on Big Data Security and Integrity Innovation."* Big data is a massive and complex data sets that a common software cannot manage. It may be classified regarding its data type, data format, data source, data consumer, data usage, analysis type, processing purpose, processing method, data store and data frequency (Terzi, DS, Terzi, R, Sagiroglu S, 2015). Traditional solutions to ensure data security and privacy are insufficient when dealing with big data. Advanced technologies and techniques are needed to protect big data. *"Health Issues on Data Breach"* is one of the major problems in the medical field. Massive data breach is occurring with alarming frequency. Healthcare breaches have increased steadily from fourth place in 2007 to 2009 to second place in 2010 to 2011. (Thomson, 2013). Another theme derived is *"Laws on Data Privacy and Personal Protection."* Cloud has become a model for enabling convenient, on-demand network access to a shared computing resources. In spite of the advantages that cloud computing has brought to us there are several issues that needs to be resolve, one of this is the customers' risk of losing data having them locked into patented formats and may lose control over the data since the tools for monitoring the data is not always provided to the customers (Sen, 2013). The government in different countries has implemented policies on data protection. In the Philippines, a commission is ordered to protect the privacy of individuals while safeguarding the free flow of idea to promote innovation and growth; regulate the collection, recording, organization, storage, updating or modification, retrieval, consultation, use, consolidation, blocking, erasure or destruction of personal data; and ensure that the Philippines abides to the international standards for data protection through National Privacy Commission(National Privacy Commission, nd). In spite of the implementation of data protection law, there are still problems that arise. The last theme derived is the "*Issues on the Implementation of Data Privacy Law against Identity Theft and Security."* Koops describe data protection law as dead. The direction of the data protection reform is basically flawed. It focuses on narrowing on solving too many ICT problems to legal protection within a single general framework of the data protection law which diverges the 21st century data protection law (Koops, 2014).

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

Table 4: Topic Probabilities

#	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	0.048567634	0.858091444	0.044963005	0.042306963	0.006070954
2	0.62195122	0.092682927	0.082926829	0.090243902	0.112195122
3	0.125093794	0.007289099	0.01061207	0.256940722	0.600064316
4	0.0997151	0.781339031	0.075854701	0.02991453	0.013176638
5	0.035421327	0.023117077	0.011931394	0.889634601	0.0398956
6	0.419642857	0.282738095	0.06547619	0.154761905	0.077380952
7	0.645333333	0.08	0.064	0.149333333	0.061333333
8	0.141914191	0.069306931	0.04620462	0.676567657	0.066006601
9	0.105596269	0.018987342	0.016988674	0.783810793	0.074616922
10	0.025628856	0.044138586	0.898671096	0.017560513	0.014000949
Latent Themes	**Provision of Data Security and Privacy Against Business Breach**	**Increased Productivity on Big Data Security and Integrity Innovation**	**Health Issues on Data Breach**	**Laws on Data Privacy and Personal Protection**	**Issues on the Implementation of Data Privacy Law Against Identity Theft and Security**

Documents talking about "Protection of big data privacy" and "Big data privacy a technological perspective and review" has a probability of 85.80% and 78.13% respectively are presenting a hidden topic on the increased productivity on big data security and Integrity Innovation. "Security and Privacy Concerns Raised Around Internet of Things," "Big Data Privacy and Security," and "Privacy Body Probes new Comelec Data Breach" having the probability of 62.19%, 41.96%, 64.53% respectively presents the topic on the provision of data security and privacy against a business breach. "How not to protect genomic data privacy in a distributed network using trail re-identification to evaluate and design anonymity protection systems" having the topic on Health issues on data breach has the probability of 89.87%. Law on data privacy and personal protection theme was the topic presented in "Data Security and Cybercrime in Italy", "Singapore Issues New Regulations in Advance of Data Protection Law Entering Into

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

Force", and "Online Privacy Law United Kingdom" with the probability of 88.96%, 67.66%, and 78.38% respectively. "Comparative Study on Different Approaches to Privacy Challenges in Particular in the Light of Technological Developments" on the other hand with the probability of 60.00% had shown the issues on the implementation of data privacy law against identity theft and security.

CONCLUSION

Based on the findings of the study, it is imperative to say that Data Privacy is a general concern. Data breach may happen by any means and source of the breach. Data security provisions are being formulated and implemented by different organizations as well as the government of various nations. Issues, in particular on the business industries and medical field are prawn to data privacy concerns. The result also shows that privacy issues still exist due to problems with the implementation of data privacy law against identity theft and security. Access to information is inevitable. Therefore it needs to be secured and protected. Most of the security threats and risk to an organization are the result of inadequate and improper access control. Poor access control can expose the organization to illegal access of data and programs, fraud, or the shutdown of computer services (Hau, 2003). Further, it is the role of the information source to protect the data from the access by unauthorized users.

ACKNOWLEDGEMENT

The researcher would like to thank Dr. Las Johansen B. Caluza a professor from Leyte Normal University and our resource speaker during our training for his generous sharing of expertise in research. For his constant motivation to engage me in research. This paper is a product of our training. To the IT Faculty of Leyte Normal University for helping and guiding me in writing this article. Thank you so much!

About the Author

Mr. Jeffrey Costiniano Cinco is a faculty of the Information Technology and Computer Education Unit of Leyte Normal University. He graduated his Bachelor in Information Technology at Leyte Normal University and continued his Master in Information Technology at Eastern Visayas State University.

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

REFERENCES

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*(1), 25.

Malin, B., & Sweeney, L. (2004). How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of biomedical informatics*, *37*(3), 179-192.

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. *IEEE access*, *4*, 1821-1834.

Grahan Greenleaf. (2010). Comparative study on different approaches to privacy challenges, in particular in the light of technological developments. Retrieved from http://ec.europa.eu/justice/data-protection/document/studies/files/new_privacy_challenges/final_report_country_report_b3_hong_kong.pdf. Retrieved on May 9, 2017.

Jeremy M. Mittman. (2014). Singapore issues new regulations in advance of data protection law entering into force. Retrieved from http://privacylaw.proskauer.com/2014/06/articles/international/singapore-issues-new-regulations-in-advance-of-data-protection-law-entering-into-force/. Retrieved on May 10, 2017.

Data security and cybercrime in Italy. *Lexology*. Retrieved from http://www.lexology.com/library/detail.aspx?g=ed4fa22f-e1de-4333-9ba7-15698e038aff. Retrieved on May 10, 2017

Online privacy law: United Kingdom. *Library of Congress*. Retrieved from https://www.loc.gov/law/help/online-privacy-law/uk.php. Retrieved on May 10, 2017.

Danny Bradbury. (2015). Security and privacy concerns raised around internet of things. Retrieved from http://www.itworldcanada.com/article/what-canadians-can-learn-from-ftcs-internet-of-things-report/101687. Retrieved on May 10, 2017.

Big data, privacy and security. (nd). *The Netherlands Scientific Council for Government Policy*. Retrieved from https://english.wrr.nl/topics/big-data-privacy-and-security. Retrieved on May 10, 2017.

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

Rainier, A.R. & Crisostomo, S. (2017). Privacy body probes new COMELEC data breach. Retrieved from http://www.philstar.com/headlines/2017/02/21/1674297/privacy-body-probes-new-comelec-data-breach. Retrieved on May 10, 2017.

Data protection. (nd). *Privacy International*. Retrieved from https://www.privacyinternational.org/node/44. Retrieved on May 10, 2017.

Data privacy day – January 28, 2017. (nd). *Mississippi Department of Information Technology Services*. Retrieved from http://www.its.ms.gov/services/pages/january-28th-is-national-data-privacy-day.aspx. Retrieved on May 10, 2017

Data privacy act of 2012. (nd). *National Privacy Commission*. Retrieved from https://privacy.gov.ph/wp-content/uploads/DPA-of-2012.pdf. Retrieved on May 11, 2017.

Information. (nd). *YOUR Dictionary*. Retrieved from http://www.yourdictionary.com/information. Retrieved on May 13, 2017

Communication privacy management theory. *Revolvy*. Retrieved from https://www.revolvy.com/main/index.php?s=Communication%20privacy%20management%20theory. Retrieved on May 16, 2017.

Data privacy Day. *Energy.gov*. Retrieved from https://energy.gov/cio/data-privacy-day. Retrieved on June 5, 2017.

Data Privacy Act of 2012. *National Privacy Commission*. Retrieved from https://privacy.gov.ph/data-privacy-act-primer/. Retrieved on June 7, 2017.

Data Privacy and Cybersecurity. (n.d.) *Andrews Kurth*. Retrieved from https://www.andrewskurth.com/assets/pdf/marea_240.pdf. Retrieved on June 10, 2017.

Thomson, L. L. (2013). Health Care Data Breaches and Information Security. In *American Bar Association* (pp. 253-267).

Sen, J. (2013). Security and privacy issues in cloud computing. *Architectures and Protocols for Secure Information Technology Infrastructures*, 1-45.

ISSN: 2455-8834

Volume:03, Issue:01 "January 2018"

Koops, B. J. (2014). The trouble with European data protection law. *International Data Privacy Law*, *4*(4), 250-261.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. (1996). From Data Mining to Knowledge Discovery in Databases. Retrieved from https://ww.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf Retrieved on June 14, 2017.

Hau, D. (2003). Unauthorized Access – Threats, Risk, and Control. Retrieved from https://www.giac.org/paper/gsec/3161/unauthorized-access-threats-risk-control/105264. Retrieved on June 20, 2017

Priyadharsini, C and Thanamani, A.S. (2014). An Overview of Knowledge Discovery Database and Data mining Techniques

UNESCO. (nd). *Access to Information.* Retrieved from http://en.unesco.org/themes/access-information. Retrieved on July 24, 2017.

