IJERSS | Volume 2 | Issue 2 | FEB 2015

Research Paper

BIG DATA REVOLUTIONS AND PRIVACY CONSIDERATIONS

COMPUTER SCIENCE

Keywords: information technology; IT; big data; BD; identity Fraud; IF; data set; DS

Dr C.Sunil Kumar	**Professor in CSE, Sreenidhi Institute of Science & Technology (An Autonomous Institution), Hyderabad**
Akshay Srirangam	**4th Year Student of ECM, SNIST, Hyderabad, Telangana**

ABSTRACT

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. Challenging issues in the data-driven model and also in the Big Data revolution are analyzed.

I. INTRODUCTION

The Information Technology (IT) revolution has formed a data revolution (DR); now commonly referred as big data (BD) in which enormous amounts of data can be gathered, stored and analyzed at comparatively low cost **[2, 3]**. This DR is based on the flood of new digital data, which has grown from an estimated 0.7 to 2.2 hexabytes in 2000 to 2,900 hexabytes in 2012, as shown in Fig. 1. About one-third of the data gathered internationally is estimated to originate in the USA. While one may be doubtful of the publicity surrounding the big DR, it obviously creates the potential for significant novelty in specific sectors as well as the overall financial system. Information by the World Economic Forum (WEF), McKinsey Global Institute (MGI) and others portray the potential advantages for healthcare, government services, fraud safety, trading, mechanized and other segments. MGI estimates that BD and analytics could yield advantages for healthcare alone of more than $350 billion yearly. Gains for the overall financial system could be up to $600 billion in annual production and cost savings. The emergence of BD also has raised privacy concerns on the part of activists and government certified **[1]**. Much of the concern relates to the gathering and use of data by governments for state security purposes, a problem which is not addressed here. Major concerns have also been expressed about the profit-making and other non-surveillance uses of BD. To address these problems, this paper addresses on the following queries that, while not new, have become more relevant in the world of BD:

- How should we assume about the reuse of data—i.e., the exploit of data for purposes not initially identified?
- Correspondingly, how should we assume about the collective use of data from diverse sources?
- What are the inferences of BD for data security—data breaches and self fraud?

It is concluded that there is no proof at this stage that the use of BD for profit-making and other non-surveillance purposes has caused privacy damages **[7]**.

A. Defining Big Data

Although the phrase is in general use, there is no exact definition of BD. MGI defines BD as referring to datasets (DSs) whose size is ahead of the ability of typical database (DB) software tools to confine, store, control and analyze. It refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, institutions, the relationship between nation and governments, and more. They focus on the ability of large DSs to give up associations between variables that can provide important community and private advantages. BDs prospective come from the detection of original patterns in behavior, and the development of analytical models, that would have been hard with smaller samples or less variables **[4]**. Data are now accessible in real time, at larger scale, with less composition, and on diverse types of variables than previously.

II. BIG DATA HAS NOT INCREASED IDENTITY FRAUD AND DATA VIOLATIONS

In hypothesis, BD could increase or decrease identity fraud (IF) and data violations (DVs). These security concerns might specify a market crash because of the complexity of daunting costs on the executors, who may be able to remain unspecified. Countervailing compels, yet, provide strong incentives for data owners (e.g., credit card corporations) to shield their data, while the data themselves are useful in preventing fraud. The BD increase the threats associated with identity fraud and data violations. It is useful, therefore, to

examine whether the propagation of data in recent years has shown up in greater incidence of IF and/or DVs.

A. Identity Fraud

Javelin Strategy and Research (JSR) gathers the only statistically representative series on IF of which we are attentive. These data are represented in Fig. 2. Despite worries expressed by the federal trade commission (FTC) and others, the overall incidence of IF has been horizontal since 2005. During the same phase, the total dollar amount of fraud has fallen—from an average of $29.2 billion for 2005-2009 to $19.7 billion for 2010-2012. To acquire a clear picture of what has happened to the risk of IF, we need to standardize the data on IF by some measure of experience. Fig. 3. shows that the cost of IF per $1,000 of USA gross domestic product (GDP) has been declining since 2005. If the IF cost data were discouraged by E-Commerce trade sales the downward tendency would be sharper, because E-Commerce has developed more rapidly than GDP. However, GDP is probably a more suitable deflator, since the greater part of IF is due to offline performance.

B. Data Violations

There are two basis of data on data violations—the Privacy Rights Clearinghouse (PRC) and the Identity Theft Resource Center (ITRC). Both of these form cumulative information on data violations from the media and public DBs from state administrations; yet, the annual totals vary slightly based on methodology and their individual definitions of a data breach. Fig. 4, which shows both series, suggests that the tendency is somewhat up since 2005. Data violations are purely an online occurrence, so it is suitable to reduce them by a measure of online activity. When reduced by the volume of E-Commerce, the threat of a data violation has been comparatively steady, as shown in Fig. 5.

III. DATA DETERMINISM: THE ADVANTAGES OF ALGORITHMS

The methodical use of individuals' data for an extensive range of reasons is not new. The direct advertising industry, for example, has for decades assembled mailing lists of customers interested in definite products and services. Credit agencies use formulas that verify individuals' eligibility for loans and the rates they may be obtainable. Similarly, the insurance industry uses key variables that specify threat to decide whether and at what rates to recommend insurance policies. A subject running through the privacy-centric BD literature is that the use of data to develop analytical models integrated apparently distinct variables is harmful to customers. As per FTC, there is another threat that is a by-product of BD analytics, namely, that BD will be used to make strength of mind about individuals, not based on tangible facts, but on assumptions that may be unnecessary. It
is noted that persons may be judged not because of what they've done, or what they will perform in the future, but because associations drawn by algorithms suggest they may behave in ways that make them poor recognition or insurance threats, inappropriate candidates for service or improbable to carry out certain functions. All of these conclusions are based on small data sometimes, one test gain or one data point (DP). BD can only develop this procedure. If more DPs are used in building decisions, and then it is less likely that any single DP will be determinative, and more likely an accurate decision will be reached. Organizations that dedicate sources to gather data and undertake complex examination do so because it is in their interest to make more exact decisions. Thus, the use of BD should lead to fewer customers being classified and less randomness in decision-making process. It is uncertain what associations may be unnecessary. For e.g., Insurance companies normally give a discount on auto insurance to students with good grades. They also distinguish on the basis of the gender of young drivers. This is most likely because the data show that there is an association among these variables—gender—and mishap costs. The use of additional variables made possible by BD should direct to more precise decisions that also might be reasonable. For example is the greater use of data by state parole boards to help inform parole decisions. Whether this reasonable is vague, but supporters believe the use of BD in this style provides more perfect forecasts of the threat of previous behavior and therefore can help decide which criminals should be released and thereby amplify public safety and perhaps also reduce top-security prison costs.

A. Big Data Do Not Distinguish Against the Poor

Some authors disagree that the uses of BD in advertising decisions support the rich over the poor. A few particularly stirring quotes from critics include ever growing data collection and analysis have the possible to intensify class differences and BD bias, profiling, tracking, barring—warn the autonomy and personal independence of the poor more than any other class. The dispute that data collection favors the rich over the poor is usually offered without proof. Price bias transfers some excess from customers to producers. However, price inequity can be cheaply efficient if it increases total output in a marketplace. Mainly in the case of products with high fixed and low insignificant costs—such as airline tickets—price inequity may be essential for the good to be produced at all. There would be smaller number of flights if airlines were not able to charge unstable prices **[5]**. Many practical goods, such as apps and software, also have high fixed and low or even zero trivial costs, and price inequity may be necessary to the production of these goods. Price bias involves charging prices based on a customer's readiness to pay, which in general is positively related to a customer's ability to pay. This involves that a price perceptive firm will usually charge lower prices to lower-income customers. Certainly, in the lack of price inequity, some lower-income customers would be unable to purchase some products at all. So, dissimilar to arguments above, the use of BD, to the level it helps price bias, should usually work to the advantage of lower-income customers.

B. ***Big Data and Consumer Choice***

Two supplementary topics running through some of the current privacy literature propose that the utilize of data and algorithms may produce "threats" quite diverse from what we usually think of as privacy and safety damages. Some authors disagree that BD will make possible to manipulate customers to buy things they don't "actually" want. Others are worried that customers will get too much of what they want—that they will exist in a filter bubble agreed on by BD. The literature on exploitation of algorithms does not present any proof that is not consistent with our conclusion that there is little verifiable threat from the lawful use of profitable information. For example, use of blood test information for drunk driving; data used for a different things; police utilize the information from a psychologist. None of these occupy profitable information. They make use of commercial purposes i.e. from Google ads. But in this case, the customer willingly uses the service with full familiarity that receives targeted ads **[6]**. Moreover, the "threat" recognized is tentative and quite oblique— customers using the service are not typically aware of any threat.

CONCLUSIONS

The budding of BD may be dissimilar from small data in terms of their structure effects on the financial system and specific segments that remains to be seen. Yet, there is no understandable reason to approach privacy policy queries arising from BD in a different way than we approach problems involving smaller amounts of data. The same queries are appropriate:

- Is there a market crash and facts of harm to customers? The recent literature on BD does not provide such proof, at least as far as the legal use of data for commercial purposes are concerned. Furthermore, no proof of an increase in threat to customers from IF or data violations has been found.
- If proof of market crash or threat is found, is there an existing cure that can logically be expected to yield advantages larger than costs and consequently surrender net benefits to customers.

ACKNOWLEDGMENT

The author would like to express their sincere gratitude to the Management of SNIST, Hyderabad for their constant encouragement and co-operation.

REFERENCES

[1] Edith Ramirez, "The Privacy Challenges of Big Data: A View from the Lifeguard's Chair", Speech at Technology Policy Institute's Aspen Forum, August, 2013, accessed at http://ftc.gov/speeches/ramirez.shtm.

[2] Mayer-Schönberger and Cukier, "Big Data: a revolution that will transform how we live, work and think", Houghton Mifflin Harcourt, 2013, p. 6.

[3] Liran Einav and Jonathan Levin, "The Data Revolution and Economic Analysis", Set up for the NBER Innovation Policy and the Economy Conference, April, 2013, p. 2.

[4] Jeremy Ginsburg et al., "Detecting Influenza Epidemics Using Search Engine Query Data," Nature, Vol. 457, February 2009, pp. 1012-14, http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html.

[5] Lynn Wu and Erik Brynjolfsson, "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities", ICIS 2009 Proceedings, Paper 147, 2009, http://aisel.aisnet.org/icis2009/147.

[6] Federal Trade Commission, Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers, March, 2012, pp. 44, available at http://ftc.gov/os/2012/03/120326privacyreport.pdf.

[7] Joseph Jerome, "Buying and Selling Privacy: Big Data's Different Burdens and Benefits", 66 Stanford Law Review Online 47, 2013, p. 50.

FIGURES:

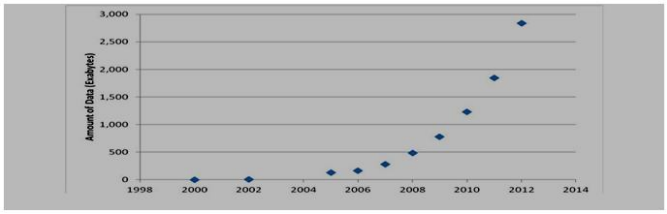

Fig. 1. Digital Data Created Yearly Worldwide

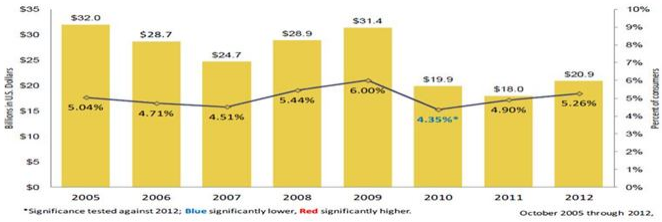

Fig. 2. Overall Identity Fraud Incidence Rate and total Fraud Amount per Year

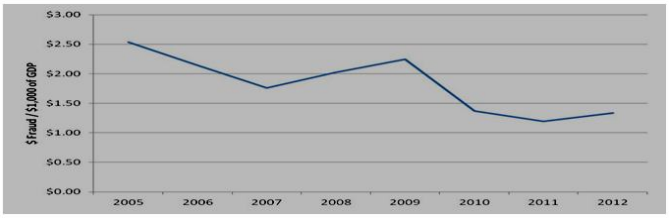

Fig. 3. Annual Cost of Identity Fraud (in dollars) Deflated by GDP

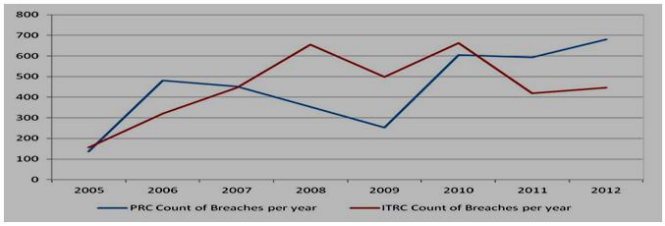

Fig. 4. Number of US Data Violations per Year

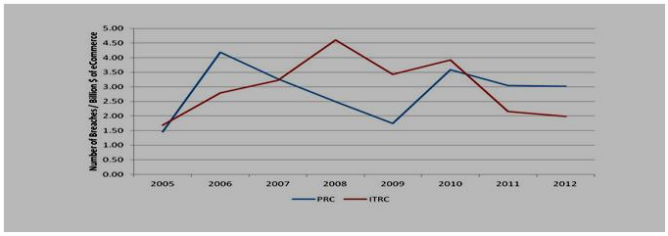

Fig. 5. Number of US Data Violations per Year Deflated by US Ecommerce

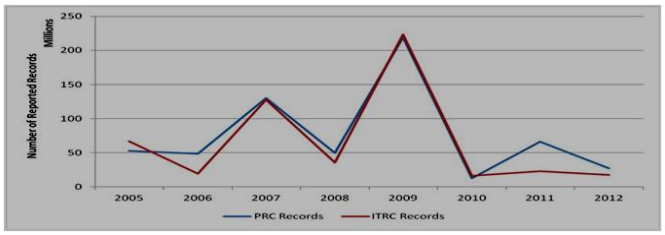

Fig. 6. Number of Reported Individual Records Compromised by Data Violations