Delineating Data Security and Privacy Boundary in Bioinformatics

Mukti Routray
Silicon Institute of Technology, Bhubaneswar, Odisha, India
mukti@silicon.ac.in

Abstract: This work draws a visible boundary for data privacy with respect to Big Data in bioinformatics. Defining Big Data boundary is a growing field where researchers have found an enormous repository of data which has huge business attraction leading to inappropriate usage of such data. This work provides a visibility to that thin line of ethical frontier which needs to be identified by data contributors and preservers so that the data fetchers are limited to data analysis or pattern analysis and the source remains encrypted and protected from invasion. Genomic data as well as bio-medical data have a true potential in their exploitation by commercial sectors which should be clogged with more privacy and security.

Keywords: big data; bioinformatics; security; privacy; ethics.

Introduction

Many areas of big data security have been observed among which underutilization of data, attacks with malefic intensions, erroneous data reporting, and passive interpretation of the scope of big data are the potential and significant threats. The security feature not only is multi-dimensional by itself but also generate an immediate requirement for policies [1] on every other feature of data stated above.

In bioinformatics the growth of data size is enormous in the recent years. The European Bioinformatics Institute (EBI) has almost 40 PB of data about genes, proteins, and small molecules in 2014, in comparison to 18 petabytes in 2013 [2]. Now this is a huge task as the data itself is not confined to general perceivable boundaries. As in case of underutilization of data, somewhere, the retention period [3] is questionable. Voices have been raised by privacy organizations that have monitored massive data retained for five years at the National ANPR (Automatic Number Plate Recognition) Data Center in Hendon, North London. Also a major question is unanswered when the data itself finds no single user right from its time of origination. Such zero-utility data in full or fragments occupy significant storage units. Multimedia data from CCTV cameras produce data for years and stored for another four to five years out of which the data may not find a single user. The data is stale from the very beginning. The data remain there with a negligible probability of utilization.

Big data is indeed enormous. It is a well-known fact now that Google has more than three billion searches stored daily. [4] From way back time in April 1994, the World Wide Web Worm [5] received an average of almost 1500 queries daily which increased to almost roughly 20 million queries per day with Altavista in 1997. Google from 9800 requests daily in 1999 to 60 million in 2000 to 200 million in 2004 and to 4.7 billion in 2011 [6] . Twitter generates over 340 million tweets daily [7].

Medical bioinformatics is concerned with sensitive and expensive data mostly contributing projects for drug design or in environments like hospitals. The distribution of data increases the complexity and involves data transfer through many network devices. Thus, data loss or corruption can occur. In absence of well-defined policies to classify the usage of such bursts of data, they tend to be misused or overused among unintended groups who bestow interest in such data.

An aspect of security of big data is in its significance in real time applications. Such applications expect the arrival of streams within strict time constraints [8]. A bunch of security attacks over transmission and related security in multimedia big data in absence of guided rules or policies may result in its unethical usage in regular online activities.

Functional genomics information is very much separated with genome-sequencing data for privacy aspect. Specifically, release of low level data sequencing leads to extraction of variation in genomes as in genome sequencing. However proteins annotated may be recovered in very less amount, at most with 5% of typed human genome. [9]

Open source multi-lingual intelligence has gained milestones with predicting criminal behavior and has tracked criminal transits in countries with malicious intentions. However this intelligence itself is exposed if the criminal faction gains superiority in locating the same data and generates a different outcome in pattern. Not only policies are to be framed, they need to be guided by several layers to address the importance of big data security. Protecting critical cyberspace infrastructure is a form of defense against catastrophic terrorism in security informatics [10] especially bioinformatics. There are several other areas of importance as global economy rests on the heap of big data exploration. The need of the hour is a policy directed definition of cyber security over big data usage, justifying the privacy of the data donor and building a strong

economy as a whole. Financial systems are the nerves of global economy. On a long run the resilience of the global financial system nurtured on big data will also depend on cyber security [11].

Privacy in Big Data and Application Level Security

Generally genomic types are accessed over web-servers through the inter-network. Most applications having a GUI based user interactive module which responds to queries from the users or sometimes perform a method of identifying types and sequences based on some machine learning or deep learning approaches.

This kind of data submissions must be related to generic use of web infrastructure in data gathering.

Facebook did an experiment [12] with human data through its newsfeed content sampled across its billions of users and later publishing the observations from the big data in PNAS.

Susan T. Fiske and Robert M. Hauser have voiced for regulation of human participation [13] in such experiments. They had expressed concern about the differences between academic and commercial research with respect to such experiments.

An ethical perspective allowing the forward movement of social-computing research [14] needs a policy framework for upholding the ethics.

The application layer is found to be more vulnerable in big data mostly where the computing takes place in the cloud. Cloud provides Software as a Service (SaaS) where the deployment is either over the internet or at a convenient location of the user. SaaS throughout the current years have become vulnerable to security threats and mostly the threats results with the application layer breakdown. Security breaches were almost reported with a higher level of 39 % [15]. Attacks were targeted through a popular SQL injection [16] which exploited 18% known and 5% unknown vulnerabilities [17].

Other most common application layer security threats faced by web application were OS and LDAP injections. Scripts are uploaded through left open doors in the web and users unknowingly execute the automated script without any knowledge of exposing their vital data to the attackers.

Access to the cloud environment exposes all data of users of the space and the cloud mostly becomes a high valued target [18][19].

These security threats are tightly coupled with privacy concerns in the application layer with unethical human tendencies to use data maliciously.

As the employees of the SaaS providers have access to enormous data, inherent policy needs to be framed for an infrequent detailed check on their behavior. A concern similar to this has been given importance by the US government which initiated the ADAMS [20] analysis for detailed understanding of human behavior. A project is floated to detect and prevent insider threats where the potential threat is more harmful as in case of government employee may abuse access privileges and share classified information.

Erroneous data reporting with the increase use of mobile applications generating real time data may be checked with higher precision in the application layer. As for crisis data management [21] multiple communities are likely to interact using the mobile applications where decisions are taken very quickly in response to vital information. Community interactions are highly flawed due to inaccurate reporting by the applications.

Scalable data analysis and management also stresses appropriate systems [22] for specific set of application requirements.

Data privacy problems can be solved by several restrictions that are provided by Data as a Service (DaaS) platform [23] where the processing of critical data may be limited to private cloud infrastructures. Public cloud resources should utilize the publicly available data.

Role based access control [24] in the application layer is needed by business organizations in order to secure its data being threatened by several external attacks.

A new multilevel [25] security model is proposed which controls data flowing in multiapplicative smart cards. The model is also effective in detecting illegal data sharing from the smart cards.

A data center infrastructure dealing with cloud platform for computing needs both outside and inside securities. Deep packet inspections and enhanced security is assured through content aware networking paradigm [26]. A growth of new type of consumers in this big data era could be seen as people might tend to buy personalized services and may lend not only their “likes” but also their profile data which may create more and more organizations to come up with near to exact, need based solutions to a margin of society. Privacy alone can be bargained for gaining an extra edge. Therefore anonymity in big data still remains a challenge in the future evolution of business related to big data and ever evolving human needs catering for more unified solutions.

Privacy and security in Big Data Usage in Biomedical Data and Bioinformatics

Schadt had pointed out [27] that it is very difficult to protect medical and genomics data when data have grown tremendously where privacy is the biggest challenge. Commercial Bio-medical sector have shown a lot of interest in widespread adoption of cloud computing which is there to handle this big volume. Behind this interest lies the perceived exploitation of the security of handling such sets. Clinical sequencing meets several regulations, mostly by Health Insurance Portability and Accounting Act of 1996 (HIPAA) in US, though presently there are no specific regulations pertaining to molecular/ genetic

diagnostics in India and even though cloud computing is being cautiously considered in such cases in US, in the developing countries like India where such regulatory mechanism is missing might be a favorable ground for illegal practices which might lead to severe exploitation of Big Data through cloud computing. In US HIPAA standards must be complied by several services as well as platform and infrastructure layers. Amazon released a paper that complies with a bunch of the standard regulatory body. [28]. Hybrid Clouds undertaken by Dell along with TGen to support first personalized medicine trial for pediatric cancer [29]. Key issues in sequencing should have detailed encryption mechanisms mostly key based. Issues such as incorrect data deletion and access to deleted data of a customer by another should be addressed.

Degree of Recognition of Actual Threats

Data security has been tougher with ever increasing dimensions of big data.

Genomic and bio-medical Big Data should include encryption for related information. Drug development should involve queries from regulatory authorities to the users of the cloud service which should be mandatory in drug development. Moreover when a person gives his or her consent to use data in a particular way, the researchers have no right to deviate. [30]

This work finds the significant areas of concern for big data security which may be traced with the composition of the data source. Generally a voluminous data source is polluted with data underutilization and data with errors. This work proposes the following categories of data underutilization and erroneous data submissions more of a generic kind however well applied in bio-medicine and genomics:

Underutilization of Data

1) Careless Submissions:
   Data is carelessly submitted through web forms or scripts overcrowding the database.
2) Data Littering:
   Less knowledge of prevailing security threats leading to unacknowledged data donations.
3) Left-over:
   Stripping of data during extraction and cleansing leads to such underutilizations.
4) Less resources available for processing:
   Utilization drastically reduces with reduced resources for processing big data.
5) Zero withdrawals:
   No available procedure for data withdrawals for voluntary donated data.
6) Data redundancy:
   There is typical growth of redundant data across and within distinct servers.
7) Data archives:
   Archives leads to huge data storages with significant reduction of users over time.
8) Careless AI:
   Mechanical interpretation often leads to chaos with a useful data.

Erroneous Data

1) Inaccurate readings from sensors/devices:
   Data generated from inaccurate readings from machines are accessed in parallel in absence of proper identification.
2) Faulty machines:
   Machines or devices with fault generating inaccurate readings.
3) Human data reporting:
   In absence of mechanized data reports transfer, human resource is highly engaged which contributes in error prone reporting.

And there are many other similar causes in overall data source pollution. This leads to a security trap for big data where the data underutilization invites malefic attacks and erroneous data leads to misrepresentation of facts. Additionally data underutilized often leads to a metadata of errors in representation. The complexity typically grows with the growth of voluminous data. There is a set intersection represented for misrepresentation of facts and malefic attacks. As attacks with intentional misrepresentation often becomes indistinguishable.

Thus the security threats can now be broadly classified as actual security threats and misrepresentation of facts. Security threats needs to evaluated in actual as most of the time the attackers disguises behind misrepresentation of facts. Such misrepresentation is widely fuelled by both data under-utilization and error in the data. There are two broad types of misrepresentation of data, actual representation of erroneous data and forced representation of unutilized data.

This work also proposes two broad types of misrepresentation of data, actual representation of erroneous data and forced representation of unutilized data.

Here degree of recognition of actual threats is directly proportional to the probability of countering underutilized data given they are error free.

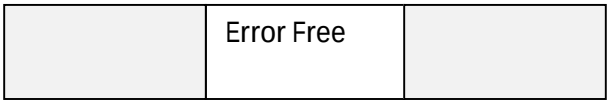

Figure 1. The whole region indicates data source with both underutilization and data with errors. Typically a portion may be error free

$$degree(R) = P(U/E_f)$$

Where R is recognition of actual threats, U is underutilized data and E_f is error free data.
The error-prone underutilized data is ignored as mostly harmless though there is a need for a separate study of such false positives where such data may be processed for any critical system.

Conclusion and Future Work

On demand cloud infrastructure which is a must in the near future in bio-medical data access may be promoted to ensure better privacy and security in the future. All of the above security breaches or threats in the application layer affecting privacy may be reduced with developing security based architecture with advanced cryptography. It is all about the knowledge of distributing the key to access private data.

Acknowledgment

The author wish to thank Dr. Swati Vipsita. Assistant Professor, IIIT, Bhubaneswar, for her support in acquiring information related to this area.

References

[1] US survey 2015-Survey WhiteHouse FAQ-Security Release-Whitepaper.
[2] EMBL-European Bioinformatics Institute, “EMBL-EBI annual scientific report 2013,” 2014.
[3] Reported by Paul Lewis on Monday, http://www.theguardian.com/uk/2008/sep/15/civilliberties.police, 15th September, 2008,
[4] Genta, Robert M., and Amnon Sonnenberg. "Big data in gastroenterology research." Nature Reviews Gastroenterology & Hepatology 11, no. 6 (2014): 386-390.
[5] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." Computer networks and ISDN systems 30, no. 1 (1998): 107-117.
[6] Coté, Mark. "Data motility: The materiality of big social data." Cultural Studies Review 20.1 (2014): 121.
[7] Das, T. K., D. P. Acharjya, and M. R. Patra. "Opinion mining about a product by analyzing public tweets in Twitter." In Computer Communication and Informatics (ICCCI), 2014 International Conference on, pp. 1-4. IEEE, 2014.
[8] Singh, Dilpreet, and Chandan K. Reddy. "A survey on platforms for big data analytics." Journal of Big Data 2.1 (2014): 1-20.
[9] Dov Greenbaum1,2,3,4,5, Andrea Sboner1,2¤, Xinmeng Jasmine Mu1, Mark Gerstein. PLoS Computational Biology December 2011, Volume 7, Issue 12, e1002278
[10] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4 (2012): 1165-1188.
[11] http://www.business-standard.com/article/news-ians/at-g20-modi-pushes-for-coordination-to-tackle-black-money-114111600146_1.html, 2014.
[12] Kramer, Adam D. I. and Guillory, Jamie E. and Hancock, Jeffrey T. "Experimental evidence of massive-scale emotional contagion through social networks." Proceedings of the National Academy of Sciences 111, no. 24 (2014), 8788-8790.
[13] Fiske, Susan T. and Hauser, Robert M.. "Protecting human research participants in the age of big data." Proceedings of the National Academy of Sciences 111,no. 38 (2014), 13675-13676.
[14] Kahn, Jeffrey P. and Vayena, Effy and Mastroianni, Anna C."Opinion: Learning as we go: Lessons from the publication of Facebook’s social-computing research." Proceedings of the National Academy of Sciences 111, no. 38 (2014), 13677-13679.
[15] Wade HB, David Hylender C, Andrew Valentine J. Verizon Business 2008 data breach investigation report, 2008 http://www.verizonbusiness.com/resources/security/databreachreport.pdf [accessed on 23 October 2014].
[16] Auger R. SQL Injection, 2009 /http://projects.webappsec.org/SQL-Injection [accessed on: 12 July 2014].
[17] Subashini, Subashini, and V. Kavitha. "A survey on security issues in service delivery models of cloud computing." Journal of Network and Computer Applications 34, no. 1 (2011): 1-11
[18] Bernard Golden. Defining private clouds, 2009 /http://www.cio.com/article/492695/Defining_Private_Clouds_Part_One [accessed on: 10 August 2014].
[19] Kaufman LM. Data security in the world of cloud computing, security and privacy. IEEE 2009;7(4):61–4.
[20] Fact sheet: Big data across the federal government (2012). http:// www.whitehouse.gov/sites/default/files/microsites/ostp/big data fact sheet 3 29 2012.pdf

[21] Herbert, Katherine G., Emily Hill, Jerry Alan Fails, Joseph O. Ajala, Richard T. Boniface, and Paul W. Cushman. "Scientific Data Infrastructure for Sustainability Science Mobile Applications." In Big Data (BigData Congress), 2014 IEEE International Congress on, pp. 804-805. IEEE, 2014.
[22] Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi. "Big data and cloud computing: current state and future opportunities." In Proceedings of the 14th International Conference on Extending Database Technology, pp. 530-533. ACM, 2011.
[23] Cheptsov, Alexey, Axel Tenschert, Paul Schmidt, Birte Glimm, Mauricio Matthesius, and Thorsten Liebig. "Introducing a New Scalable Data-as-a-Service Cloud Platform for Enriching Traditional Text Mining Techniques by Integrating Ontology Modelling and Natural Language Processing." In Web Information Systems Engineering–WISE 2013 Workshops, pp. 62-74. Springer Berlin Heidelberg, 2014.
[24] Schumacher, Markus, Eduardo Fernandez-Buglioni, Duane Hybertson, Frank Buschmann, and Peter Sommerlad. "Security patterns." Integrating Security and Systems Engineering. Hoboken (2006).
[25] Girard, Pierre. "Which security policy for multiapplication smart cards." InUSENIX workshop on smartcard technology. 1999.
[26] Bolsens, Ivo, Georges GE Gielen, Kaushik Roy, and Ulf Schneider. "" All Programmable SOC FPGA for networking and computing in big data infrastructure"." In ASP-DAC, pp. 1-3. 2014.
[27] Schadt EE. The changing privacy landscape in the era of big data. Mol Syst Biol 2012;8:612.
[28] Creating HIPAA-Compliant Medical Data Applications With AWS. <http://aws.amazon.com/about-aws/whats-new/2009/04/06/whitepaper-hipaa/>.
[29] Managing data in the Cloud Age. <http://www.dddmag.com/articles/2012/10/managing-data-cloud-age>.
[30] Vivien Marx. "Biology: The big challenges of big data", In Nature 498, 255–260 (13 June 2013)