Big Data And Anonymity

Paulo Leal
Faculdade de Ciências
Universidade do Porto
Email: up201101503@fc.up.pt

Inês Maia
Faculdade de Ciências
Universidade do Porto
Email: up201101593@fc.up.pt

***Abstract*—With the expansion of networking, devices with a great data collection capacity and data storage, Big Data works together with traditional data, integrating new types and processing of data. Big Data is now expanding and becoming a tremendous potential for different domains, including science and engineering, contributing for several researchers in significant studies.**

However, in researches relating to Human subjects (which seek patterns, associations, and trends) it's important to reflect on whether releasing the data used and the results obtained would be on the verge of breaking someones privacy due to a risk of re-identification; a problem that Data Anonymity (DA) can (try) to solve. This paper presents an explanation about Big Data and Anonymization, the defining characteristics and techniques of each, and an overview of two important studies in this subject.

I. Introduction

With the appearance and rapid development of Internet of Things (IoT) and Wireless Sensor Networks (WSN) the World is currently in a state where it is driven by Data and its collection is highly efficient; Data brings enormous value to different domains, providing information to governments, businesses, and individuals, improving research and possibly auxiliating in innovation and economic growth II.

With the advances in data mining and a continuous increase in Data storage capacity the ability to share, access, and generate Data have expanded, and with a growth in volume and size came the coining of the term Big Data (BD) (used to refer to the previous) which is currently becoming a key part in research.

Big Data is a term for large and complex datasets and use of predictive and advanced data analytics methods to extract value from data II. Due to this high amount of data here are some privacy concerns and data security requirements, it is important to keep a balance between using this data in a beneficial way and protecting each individuals privacy.

The purpose of data anonymization III is to protect privacy by obfuscating personal information (usually by removal, encryption, or substitution) in a given dataset with the porpuse of these techniques being to hamper the connection between the data and a particular individual (de-identification III-C).

Nowadays complete anonymization can not yet be guaranteed without severely hurting the quality of the dataset, and even excellent attempts of doing so may only last some time due to the constant improvemente of re-identification techniques IV. Some studies or articles even show that it is possible to re-identify an individual with very little information IV-A IV-B.

II. Big Data

A. What is Big Data

BD is usually used to refer to a higher volume of data or extremely large data sets, although the sheer volume or size is not what is of most importance but what it it is done with the data that is most relevant. BD can be analyzed in order to extrude patterns, associations, and trends, most of the time relating to Human relations or behavior.

BD brings with it a number of challenges regarding the collection and treatment of data, this, coupled with the usual size or complexion of BD sets, usually renders traditional methods of data processing inadequate to deal with them.

These data sets grow at an impressive rate due to data being collected from increasingly numerous information-sensing devices in which in many ways the IoT and WSN hold responsibility. As an example, WSN rapidly expand as mobile devices, cameras, microphones, and many more continue to be added to them as a means of gathering information.

B. Defining Characteristics

BD usually consists of the following characteristics [1]:

1) Volume: Besides determining whether it is considered BD or not the overall size of generated and stored data holds influence regarding its value and potential for analysis.

2) Variety: The type and whereabouts of the data; having this information allows for an effective analysis and use of results.

3) Velocity: The speed at which new data is generated and processed; this is highly important in relation to growth and development that uses these results.

4) Variability: How much inconsistency exists in the data set; this can slow down or make the process of analysis problematic.

5) Veracity: The quality of the data, which directly relates to the analysis accuracy overall value.

III. DA

A. What Is DA?

With the increase of data and its need to store and collect more and more information, an important need arises - create techniques to anonymize this information to guarantee privacy and prevent possible attacks.

Therefore, data anonymity is the use of anonymization techniques that allow information to be hidden, making it impossible or more difficult to establish a connection between store data to a particular individual.

This arises in a way that allows the privacy of the people to be protected, and thus, with the advances of the amount of information and the need to evaluate it, a data security of individuals is established.

With this, the transaction of datasets between entities is commonly used in a wide spectrum of applications, including recommendation systems, e-commerce, and biomedical studies [2] and for this sharing to be legal and not exceed the security limits, there were created some techniques for this purpose, called anonymization techniques, that will be explained below.

B. How Does It Work?

Coupled with the advantage that comes with the increasing of data, it is necessary to question how this information can be safe, without violating the right to privacy of any individual.

Thus, to prevent this published data from being attacked, data anonymization techniques emerge that can be classified into several dimensions: [3]

1) Nature of Data: The present technique is concerned with the nature of the information and there are three ways to evaluate it:

- Tabular Data - is based on entities' information and thus this entity may be people having their identifiers like age, gender and may contain sensitive information, ie information that should be kept anonymous such as salary and diseases;
- Item Set Data - it is basically the transactional data of the entity, such as purchases made by a customer (their "market basket");
- Graph Data - it is based on the associations between entities, being inserted here the social networks and this relationship between people.

2) Anonymization Approaches: Depending on the type of data being hidden, there are several ways to create anonymization techniques, including:

- Data Suppression- it's based on the removal of information from the data. This removed information may be the identifiers of an entity, such as personal information about someone, like address, gender;
- Disturbance of Data - here the data is altered in order to hide its real value, adding noise (meaningless data) to corrupt the real values. This form may be important to sensitive data such as a person's salary;
- Generalization of Data - maintaining the data in a general way, not providing its exact value. For example, creating intervals of age instead of giving the exact age of a person;
- Data Exchange - it is based on the exchange/ swap of sensitive data in entities, like exchanging a list of medication purchased by a person.

3) Anonymization Objectives: For the data to actually be kept hidden, it is necessary that they meet certain requirements. Thus, after applying anonymization techniques, the data have important goals and objectives:

- K-Anonymity - given a database, each individual should be different in some parameters from the rest. In this way, it is verified if each one of them is indistinguishable from the remaining k-1;
- Diversity - for sensitive data, a differentiation of information between individuals must be maintained;
- Other methods - there are other methods that must be taken into account in such a way that an attacker can not assume an entity's value from its knowledge of the shared dataset information.

C. De-identification

De-identification simply put is preventing that an identity can be connected to information, which is commonly done (to name a few strategies) by the removal of Personal Identifiers (PI) and Quasi-Identifiers (QI). The reversal of this process is appropriately named re-identification.

However, due to the strength of modern re-identification techniques the absence of PI alone does not guarantee that the remaining data does not identify individuals. [4][5][6][7][8]

1) PI: A subset of information that can be used to, by itself or in use with other information, identify, contact, or locate a single person or individual is defined as a PI.

2) QI: A piece of information that by itself cannot by a PI but is sufficiently correlated to an entity that in combination with another can create a unique identifier is a QI. Thus QI can when combined become PI, and this is the basis of re-identification.

D. Simply Anonymized Datasets

A Simply Anonymized Datasets (SAD) is a dataset that has had the data that the National Institute of Standards and Technology (NIST) has classified under the full definition of PI removed, which follow [9]:

- Full name (if not common)
- Home address
- Email address (if private from an association/club membership, etc.)
- National identification number
- Passport number
- IP address (when linked, but not PII by itself in US)
- Vehicle registration plate number
- Driver's license number
- Face, fingerprints, or handwriting
- Credit card numbers
- Digital identity
- Date of birth
- Birthplace
- Genetic information
- Telephone number
- Login name, screen name, nickname, or handle

IV. De-anonymization/Re-identification

Re-identification (RI) is a process of data mining that consists in cross-referencing anonymous data from one source with others in order to re-identify the anonymous data source. In its basis QI are used and always have been the foundation of RI. As RI techniques improve the usage of SAD III-D is no longer enough to successfully anonymize data as mentioned in III-C. An example of this is when health records were linked to public information and used to locate the then-governor of Massachusetts' governor hospital records [10].

A. Unique In The Crowd

1) The Work: Given a set of spatio-temporal points and a simply anonymized mobility dataset the study evaluates the uniqueness of traces by extracting a subset of trajectories from the given mobility dataset that match the points that compose the starting set.

15 months of mobility data (pertaining 1.5 Million people), which was simply anonymized, was used for analysis. Results show that four randomly chosen points are enough to characterize, uniquely, 95% of the users, and that only two points are enough for over 50% of the users. The percentage of identified users depends on the resolution of the dataset which can be manipulated by changing a few parameters (which includes the size of a region, clusters, and more).

Results are directly influenced by population density, being that a higher one will tend to decrease the percentage of unique re-identifications, however, locations with a higher count in population will tend to have an increased number of antennas or WiFi spots. These will run in an opposite effect regarding results meaning that they should generalize and scale regardless, thus all that a higher density will bring is a higher raw number of re-identifications, assuming that the overall percentage indeed suffers little change.

2) Second Thoughts: While this study successfully shows that mobility data is highly unique for each individual it doesn't, in reality, re-identify particular individuals from their traces. It is suggested that re-identification can be done by linking results to outside information like addresses or any sort of geo-located information shared by individuals; this process is as was mentioned in IV done by using QI III-C2 to create new PI III-C1, however it was not done or demonstrated.

In order for someone to do connect an individual to the correct trace, information about each individual that pertains to the original sample would have to be collected and doing so would be complex both in terms of having access to such information and in terms of gathering, being that four spatio-temporal pieces of information would be necessary (in a best case scenario) to match with the four resulting points from the study.

The data set used was also a SAD which is as mentioned in III-C not enough to guarantee anonymity, and while the study produces strong arguments concerning mobility data uniqueness we cannot, for this reason, generalize its results to many other types of data that could be properly anonymized and used in research. We can conclude then that while we can with very little information determine a mobile identity it is in a limited scope regarding overall de-identification and user anonymity.

B. Breaking The Netflix Dataset

1) The Work: In this paper, the authors present a class of statistical de-anonymization algorithms and then show how these methods can be used in practice to de-anonymize the Netflix Prize dataset.

Firstly, this algorithm assumes that even if only a small part of the dataset is provided, an adversary can recognize an individual's record in the anonymized dataset and get even more information, such as sensitive attributes, establishing the idea that the adversary can identify with great probability the anonymized information disclosed. They also point out that, for sparse data (like transactions and preferences), it is sufficient to know only 5 to 10 attributes.

Netflix, the world's largest online movie rental service, published a dataset containing anonymized movie rating of 500,000 subscribers, to support the Netflix Prize data mining contest - a contest created to improve their movie recommendation service.

The published dataset consists of only 1/8 records of all Netflix subscribers and then the authors needed to be concerned with false positives. However, if a record is identify in the dataset, it will very likely not be a false positive and there is a lot of information that can be used to remove these false positives.

Therefore, the authors demonstrated the applicability of this algorithm, by de-anonymize the movie viewing records by combining the published dataset with a public database of ratings. So, they created an algorithm that compares a person's public movie ratings, from the Internet Movie Database (IMDb), to the anonymous identifier movie ratings in Netfix's data to find if the subscribers rated movies in the Netflix.

The results show that with 8 movie ratings, where 2 of them may be incorrect, and dates that may have 14-day error, 99 percent of the records could be uniquely identified in the Netflix data set, to detect if a person is in that database. Since they used a small sample of IMDb users (due to the IMDb's terms of service), the algorithm was able to identify the records of two users in the Netflix Prize dataset.

To conclude, the authors assume that, as these algorithms are robust to perturbation and sanitization, these are capable of being used and breaking any dataset containing anonymous multi-dimensional records, such as preferences and individual transactions. [5]

2) Second Thoughts: With this work, it was verified that when publishing movie ratings of subscribers, in which they thought that their identification would be kept private, it is possible to find out who they are and their personal feeling and opinions.

The author's gave an example of an user's public IMDb ratings in which, through his likes and dislikes in movies, are identified all his rankings on the Netflix system, which he would probably think that would remain private. By exposing

this information, the authors deduce some traces of the personality of this subscriber - they found his political orientation based on his opinions about some movies like “Fahrenheit 9/11”, his religious views and understood that he had strong opinions about liberal and gay-themed films.

Vitaly Shmatikov, professor of Computer Science and one of the authors of this paper, clarified: ”Releasing the data and just removing the names does nothing for privacy. If you know their name and a few records, then you can identify that person in the other (private) database.”

This raises questions about privacy breaches when a large set of data is released. Even if Netflix said that this published dataset does not have any information that could identify the subscribers, this paper demonstrates that it is a lie. However, the data provided by the Netflix does not give any information about the users, it is only when this information is combined with the IMDb data that a leak of information is perceived.

V. Conclusion

The term BD refers to data sets of a large volume or size which rapidly expand due today’s usage of the IoT and WSN; their size can mean that they are complex and both of these characteristics require new non-traditional methods of analytics and processing. BD is defined by its volume, variety, velocity, variability, and veracity II.

DA is the use of anonymization techniques in an attempt to make the connection between Data and an Individual difficult or impossible III, it is usually done by obfuscating the Dataset by removal, substitution, or addition (noise) of PI III-C1 and QI III-C2. The bare basics of DA create a SAD which is no longer enough to guarantee a safe Dataset III-D. RI utilizes the remainder of PI and QI between different sources of information in order to connect them to an identity.

Some articles or researches claim that with very little information RI is possible IV-A IV-B, after analyzing and reconsidering the results (and our research) we have found them to be correct yet overstated in their conclusion, we feel the answer is more complex than what some headlines make the appear to be.

While RI is indeed possible, in some cases it is not as simple as it as seems as the having the required data to cross-reference with what the studies we reviewed offer us is complex both in terms of collection and access to it. Fully complete DA is possible but not without massively undermining the quality of the Dataset available for analysis and in so hurting the results, possibly even making the study worthless in its findings. This is something that in the corporate world is not wished for, we can infer then that business prone Datasets could be ’more’ anonymized but aren’t in favor of more noticeable results, which may be a problem regarding privacy.

At the time of writing, new laws are being passed in regards to these Datasets in which they must comply to an European Union determined set of characteristics in face of a large fine; while we feel that this is a step in the right direction the legality of information in regards to privacy is something that, we hope, may be enough to help bring new standards in DA.

DA is a complex subject where we must weigh the benefits of quality Data versus the guarantee of user anonymity, where having both is some that is to be considered (currently) a happy medium instead of a certainty, especially in BD where anonymization techniques suffer over the complexity and variety of information.

References

[1] M. Hilbert, “Big data for development: A review of promises and challenges.” *Development Policy Review*, vol. 34, 2016. [Online]. Available: http://doi.org/10.1111/dpr.12142

[2] A. Gkoulalas-Divanis and G. Loukides, “Utility-guided clustering-based transaction data anonymization.” [Online]. Available: http://www.tdp.cat/issues11/tdp.a083a11.pdf

[3] G. Cormode and D. Srivastava, “Anonymized data: Generation, models, usage.” [Online]. Available: http://www.dimacs.rutgers.edu/~graham/pubs/papers/anontut.pdf

[4] M. V. Yves-Alexandre de Montjoye, César A. Hidalgo and V. D. Blondel, “Unique in the crowd: The privacy bounds of human mobility,” *Scientific Reports*, no. 3, 2013. [Online]. Available: http://www.nature.com/articles/srep01376

[5] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset),” 2008.

[6] ——, “De-anonymizing social networks.” 2009. [Online]. Available: 10.1109/SP.2009.22

[7] ——, “Robust de-anonymization of large sparse datasets.” *Security and Privacy 2008*, 2008. [Online]. Available: 10.1109/SP.2008.33

[8] D. 95/46/EC, “Opinion 05/2014 on anonymisation techniques,” 2014. [Online]. Available: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[9] T. G. Erika McCallister and K. Scarfone, “Guide to protecting the confidentiality of personally identifiable information (pii).” [Online]. Available: https://doi.org/10.6028/NIST.SP.800-122

[10] D. C. and Barth-Jones, “The ”re-identification” of governor william weld’s medical information: A critical re-examination of health data identification risks and privacy protections, then and now.” 2012. [Online]. Available: https://fpf.org/wp-content/uploads/The-Re-identification-of-Governor-Welds-Medical-Information-Daniel-Barth-Jones.pdf