Privacy Inspired Architecture: Social Mining & Big Data Analytics

Pratik Gupta[1], Prithvi Chouhan[2], Khushi Dhingra[3]

[1,2,3]Computer Science Engineer,
[1,2]Acropolis Institute of Technology & Research, Mangliya Sadak, Indore-453771, (M.P), India.
[3]Shri Vaishnav Vidyapeeth Vishwavidyalaya, Gram Baroli, Sanwer Road, Indore-452002, (M.P), India.

Abstract: *Privacy is a growing phenomenon in our society and is becoming an important consideration when one wants to use, publish and analyse data that includes sensitive personal information. Unfortunately, it is becoming increasingly difficult to convert data in a way that protects sensitive data: we live in an age of big data characterized by unprecedented opportunities to hear, store and analyse social data that describes human activities in detail and with great resolution. For this reason, confidentiality should not be achieved by private identification. In this paper, we propose privacy through architectural drawings to create technological frameworks to combat the threats of unintended, illegal consequences of breach of privacy, without compromising access to social mining information and major data analysis technologies. Our main idea is to incorporate privacy protection in architectural information technology, so that analysis also includes appropriate privacy requirements from the outset.*

Keywords: privacy by architecture; privacy-by-design (PbD), big data analytics; social mining; data mining; data analytics; hadoop; rapid miner, kaggle.

I. INTRODUCTION

The most notable Big Data is called rocket fuel for economic growth. As the big data field continues, growth will grow as the focus from the first joy that we can process big data and understand the acquisition, management, and sharing of our data. Switching to the world's largest collection of integrated data, Google's definition of "Big Data" states that "extremely large data sets can be analysed electronically to reflect patterns, styles, and organizations, particularly those related to human behaviour and interaction." Therefore, at least on the basis of the reason why we value Big Data, the context of this paper focuses on using this framework to control this valuable asset of Big Data in its collection, storage, and use. Over the past few years, many strategies have been proposed to develop technological frameworks to combat breach of privacy, without losing the benefits of high-tech data analytics. Without these efforts, there is no standard method available for managing general personal information and maintaining standard analytical results. Anonymity in the general sense is considered a chimera and concerns about entering the private sector through big data are now in the headlines of major media outlets. However, big data analysis and privacy are not the necessary enemies. The purpose of this paper is to demonstrate precisely how many efficient and effective services based on big data analytics can be constructed in such a way that the quality of the results can remain with the highest protection of personal data. The magic name is privacy-constructed. The magic word is privacy-by-design. We propose here a methodology for purpose-driven privacy protection, where the purpose is a target knowledge service to be deployed on top of data analysis. The basic observation is that providing a reasonable trade-off between a measurable protection of individual privacy together with a measurable quality of service is unfeasible in general, but it becomes feasible in context, i.e., in reference to the kind of the analytical goal desired and the reasonable level of privacy expected. service quality is not usually possible, but it does happen in context, that is, based on the type of analytical goal required and the appropriate level of privacy expected.

II. PRIVACY BY DESIGN

Privacy-by-design is a paradigm developed by Ontario Information and Privacy Commissioner, Dr. Ann Cavoukian, in the 1990s, to address the emerging and growing threats to online privacy. The main idea is to incorporate privacy protection in the construction of information technology from the outset. This paradigm represents a major innovation with regard to traditional methods of privacy protection because it requires a major shift from a functional model to an active one. In other words, the concept prevents privacy issues instead of fixing them. Given the ever-growing

diffusion and availability of big data and given the great impact of the big data analytics on both human privacy risks and the possibility of understanding important phenomena many companies are realizing the necessity to consider privacy at every stage of their business and thus, to integrate privacy requirements 'by design' into their business model Unfortunately, in many cases it is not entirely clear what are the ways to instil privacy through construction.

III. PbD: APPLIED TO BIG DATA

The Big Data environment creates unique challenges that help solve the basic goals of PbD. What we see as Big Data's evolving structure is that traditional theories of privacy protection with basic commands such as the identification and removal of PII data objects, while basic, are not self-sufficient. To establish an effective Big Data system, the basic principles of PbD allow us to go back and consider the overall health of the Big Data environments and the data usage we use. As we look to explore the possibilities of PbD in Big Data, too, the basic principles of PbD include:

a. Proactive not reactive, preventative not remedial
b. Privacy as the default setting
c. Privacy embedded into design
d. Full functionality—positive-sum, not zero-sum
e. End-to-end security—full lifecycle protection
f. Visibility and transparency—keep it open
g. Respect for user privacy—keep it user-centric.

It is through the examination of each of these core components that the applicability of PbD to this rising area is best considered.

IV. PbD APPLIED TO SOCIAL MINING

Our main idea is to incorporate privacy protection in any design analysis process, so that analysis including appropriate privacy requirements from the outset, evokes the concept of privacy-by-construction mentioned above. The clarification of the general principle of 'construction' on a large data base is that high security and quality can be better achieved in a way that focuses on goals.

In this way, the data analysis process is built on thoughts about:

a) The subject of sensitive personal data for analysis;
b) An attack model, i.e., the knowledge and purpose of an enemy interested in obtaining sensitive data of specific individuals;
c) A section of analytical questions to be answered in detail.

This thinking is fundamental to the development of technology that knows privacy. First, privacy protection strategies are highly dependent on the type of data protection. For example, the ways in which social data is appropriate were not appropriate for data. Second, the legal framework must define an attack model, which can be a reliable but curious enemy model or a vicious enemy model, with sufficient scale. These two species require different actions because of their characteristics. The first one makes the agreements correctly but tries to learn as much as possible. For example, by reading offline the standard output of an algorithm can try to get information from another group. This is in contrast to the dangerous rival who since then has also illegally deviated from the protocol is difficult to disprove. Typically, attacks are based on the background information of an enemy and different thinking background information includes different defensive tactics. For example, an attacker can have a close understanding of a person's movement behaviour and use it to conduct all his movements. In some cases, an enemy might block a person and gain access to certain areas visited by him. It is clear that a defence strategy designed to counter opposition with limited information can be very weak in the event of detailed and vice versa. Ultimately, the privacy policy should find an acceptable trade between data privacy and data usage. In the meantime, it is important to look at the analysis questionnaire that will be answered to understand which data structures need to be maintained.

For example, the development of a self-defence data protection strategy should be considered so that this data can be used to analyse behavioural interactions in an urban environment.

Under the above assumptions, we state that it is possible to consider a privacy-sensitive analysis process that could:

a) Turn data into an anonymous version with an unparalleled privacy guarantee - that is, the chances of malicious attacks failing;
b) Make sure the segment of the analytical questions can be answered correctly, within the unambiguous speculation that specifies the use of the data, using modified data instead of the original.

Trading between privacy protection and data quality should be a major goal in the development of technology that knows the privacy of large data analytics. If in creating such a framework only one of these two factors is considered, the result is that either we guarantee high levels of privacy but the data cannot be used for analysis, or we guarantee the best quality of data by compromising individual privacy protection on data. Note that, in large data analytics and social mines, a person is often interested in extracting integrated information and this would not involve the use of personally identifiable information.

The basis should be considered, because it allows for the management of data privacy risks, by effectively eliminating the risk at the first level of the information life cycle. This principle requires that in the design of big data analytical frameworks we should consider that we need no collection of personally identifiable information, unless a specific purpose is defined.

The above privacy-by-design methodology can help to understand which is the minimal information that enables a good analysis and protection. As we can see in the scenario presented, we are able to find the minimal information for mining data with perfect quality and, we show how the level of data aggregation useful for the analysis already provides very low privacy risks. In the following, we show how we apply the privacy-by-design paradigm for the design of analytical frameworks "**Outsourcing of data mining tasks (V-IX)**" In these scenarios we first analyse the privacy issues related to this kind of data, second, we identify the attack model and third, we provide a method for assuring data privacy taking into consideration the data analysis to be maintained valid. However, these are not the unique privacy preserving frameworks adopting the privacy-by-design principle, many approaches proposed here can be seen as instances of this promising paradigm.

V. OUTSOURCING OF DATA MINING TASKS

Privacy-by-design paradigm can also be applied with success to distributed analytical systems where we have an untrusted central station that collects some aggregate statistics computed by each individual node that observes a stream of data. In this section we discuss an instance of this case; in particular, we show as the privacy-by-design methodology can help in the design of a privacy-aware distributed analytical processing framework for the aggregation of movement data. We consider the data collector nodes as on-board location devices in vehicles that continuously trace the positions of vehicles and periodically send statistical information about their movements to a central station.

The central station, which we call *coordinator*, will store the received statistical information and compute a summary of the traffic conditions of the whole territory, based on the information collected from data collectors. We show how privacy can be obtained before data leaves users, ensuring the utility of some data analysis performed at collective level, also after the transformation. This example brings evidence to the fact that the privacy-by-design model has the potential of delivering high data protection combined with high quality even in massively distributed techno-social systems. The aim of this framework is to provide both *individual* privacy protection by the differential privacy model and acceptable *collective* data utility.

VI. STATE-OF-THE-ART ON PRIVACY-PRESERVING DISTRIBUTED DATA ANALYTICS

The privacy model is particularly relevant in ensuring individual privacy while answering questions covered by the privacy of diversity. Recently, much attention has been paid to using the privacy of private data analysis. In this setting n groups, each with some sensitive data, wish to calculate aggregated statistics in addition to data for all groups with or without a central link. Prove that when calculating the sum of all group entries except the central link, any multi-group protocol with a small number of rounds and a small number of messages should have a large error, consider the problem of secretly merging amounts in most cases. They both look for a malicious link and use encryption and privacy encryption in the construction of privacy data systems. Compared to their work, we focus on a trusted consultant, with the goal of building privacy-saving strategies by adding sensible sounds to improve data usage. In addition, both consider aggregated queries as a primary function of use, while considering network-based analysis of collected data. Different types of use lead to the development of different privacy strategies. We agree that our method can be

further enforced to against the malicious coordinator by applying the encryption methods.

VII. ATTACK AND PRIVACY MODEL

As in the case analysed, we consider as sensitive information any data from which the typical mobility behaviour of a user may be inferred. This information is considered sensitive for two main reasons:

a. Typical movements can be used to identify the drivers who drive specific vehicles even when a simple de-identification of the individual in the system is applied;
b. The places visited by a driver could identify peculiar sensitive areas such as clinics, hospitals and routine locations such as the user's home and workplace.

The assumption is that each node in the system is honest; in other words, attacks at the node level are not considered. Instead, potential attacks are from any intruder between the node and the coordinator (i.e., attacks during the communications), and from any intruder at coordinator site, so this privacy preserving technique has to guarantee privacy even against a malicious behaviour of the coordinator. For example, a consultant may be able to obtain real-time traffic information from other sources, such as the public data sets on the web, or with personal information about a particular participant, as previously (and separately) discussed with the link attack.

The proposed solution is based on Alternative Privacy, the latest version of random action introduced by Dwork. The general idea of this paradigm is that privacy risks should not increase the respondent due to appearing in the statistical database; Differential confidentiality ensures, in fact, that the enemy's ability to harm should be the same, regardless of whether anyone enters, or leaves the database. This privacy model is called ε-differential confidentiality, due to the ε-guaranteed level of privacy. Note that when ε is prone to having very little interference is introduced and this results in lower privacy protection; on the contrary, better privacy guarantees are available when ε tends to zero. A separate privacy assures the record holder that any breach of privacy will not result in participation in the database because anything, or almost anything, that can be read from a database with his or her own record is also readable to that person without his or her data. Moreover, it is formally proved that ε-differential privacy can provide a guarantee against adversaries with arbitrary background knowledge, thus, in this case we do not need to define any explicit background knowledge for attackers.

VIII. PRIVACY PRESERVING TECHNIQUE

First of all, each participant must share a common partition of the examined territory; for this purpose, it is possible to use an existing division of the territory (e.g., census sectors, road segments, etc.) or to determine a data-driven partition as the Voronoi tessellation introduced. When a subdivision is distributed, each route is grouped as a sequence of cross-sections (i.e., trajectory sequences). For simplicity, this information is mapped to a frequency vector, linked to a subdivision. Unfortunately, removing the frequency of traffic instead of raw trajectory data from the link does not maintain privacy, as the culprit may still be transmitting sensitive driver information. For example, an attacker can learn a lot from the driver; this information can be very sensitive because such a move is usually accompanied by the user's movement between home and work. Therefore, the proposed solution relies on a divisive privacy process, the distribution of Laplace. At the end of the predefined interval τ, before sending the vector frequency to the connector, for each object in the vector the node emits the sound transmitted by Laplace and adds to the actual value in that vector position. At the end of this step the node Vj changed its frequency vector fVj to its private version ˜ fVj. This ensures respect for the privacy of ε-differences. This simple general strategy has some drawbacks: firstly, it can lead to a large amount of noise which, even in the slightest, becomes unimaginably large; secondly, adding a sound drawn from the Laplace distribution could expose the calculation of negative, unintentional frequency in travel conditions. To fix these two problems, it is possible to bind the sound drawn from the distribution of the Laplace, lowering it into a separate secretive scheme (ε, δ). Specifically, for each x number of vectors fVj, it is possible to draw a sound that binds to the space [–x, x]. In other words, for any original frequency fVj [i] = x, its disturbed version after adding sound should be in the space [0,2x]. This method satisfies a separate privacy (ε, -), where δ measures the loss of privacy. Note that, as in a distributed area the critical problem is greater than in communication, it is possible to reduce the amount of data transmitted, i.e., the size of the frequency vectors.

IX. ANALYTICAL QUALITY

To date, we have issued official guarantees on individual privacy, but we must demonstrate at this time whether the individually converted values still apply once they have been collected and compiled by the consultant, that is, if they are appropriate at the combined level of analysis. In the proposed framework, the coordinator collects disruptive frequency veins for all vehicles over time τ and balances them with movement. This allows you to obtain a guided global vector, representing the flow values in each local tessellation link. Since the privacy transformation operates on the entries of the frequency vectors, and hence on the flows, we present the comparison (before and after the transformation) of two measures: (1) the *Flow per Link*, i.e., the directed volume of traffic between two adjacent zones; (2) the *Flow per Zone*, i.e., the sum of the incoming and outgoing flows in a zone. The following results refer to the application of this technique on a large dataset of GPS vehicles traces, collected in a period from 1st May to 31st May 2011, in the geographical areas around Pisa, in central Italy. It counts for around 4,200 vehicles, generating around 15,700 trips. The τ interval is one day, so the global frequency vector represents the sum all the trajectories crossing any link, at the end of each day.

Figure 1 shows the resulting Complementary Cumulative Distribution Functions (CCDFs) of different privacy transformation varying ε from 0.9 to 0.01. Figure 1 shows the reconstructed flows per link: fixed a value of flow (x) we count the number of links (y) that have that flow. Figure 2 shows the distribution of sum of flows passing for each zone: given a flow value (x) it shows how many zones (y) present that total flow. From the distributions we can notice how the privacy transformation preserves very well the distribution of the original flows, even for more restrictive values of the parameter ε. Also considering several flows together, like those incidents to a given zone (Figure 2), the distributions are well preserved for all the privacy transformations. These results reveal how a method which *locally* perturbs values, at a *collective* level permits to obtain a very high utility.

Qualitatively, Figure 3 shows a visually comparison of results of the privacy transformation with the original ones. This is an example of two kind of visual analyses that can be performed using mobility data. Since the global complementary cumulative distribution functions are comparable, we can choose a very low epsilon (ε = 0.01) with the aim to emphasize the very good quality of mobility analysis that an analyst can obtain even if the data are transformed by using a very low ε value, i.e., obtaining a better privacy protection.

In Figure 3(A) & (B) each flow is drawn with arrows with thickness proportional to the volume of trajectories observed on a link. From the figure it is evident how the relevant flows are preserved in the transformed global frequency vector, revealing the major highways and urban centres. Similarly, the *Flow per Zone* is also preserved, as it is shown in Figure 3(C) and (D), where the flow per each cell is rendered with a circle of radius proportional to the difference from the median value of each global frequency vector.

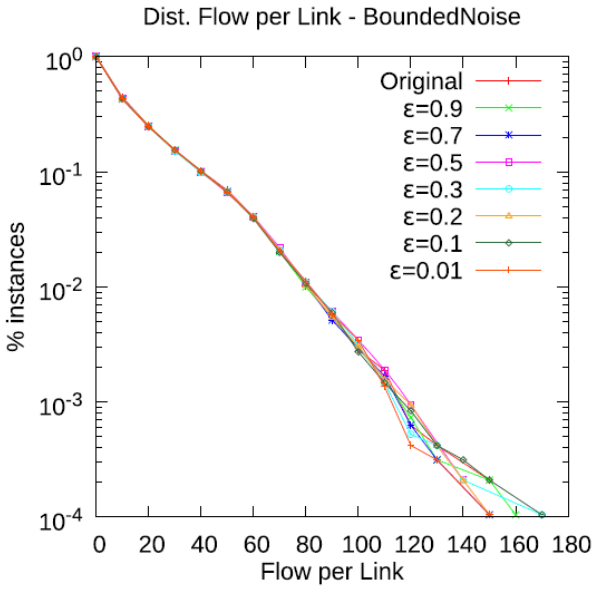

Figure 1: Shows CCDFs of *Flow per Link* for different levels of protection ε

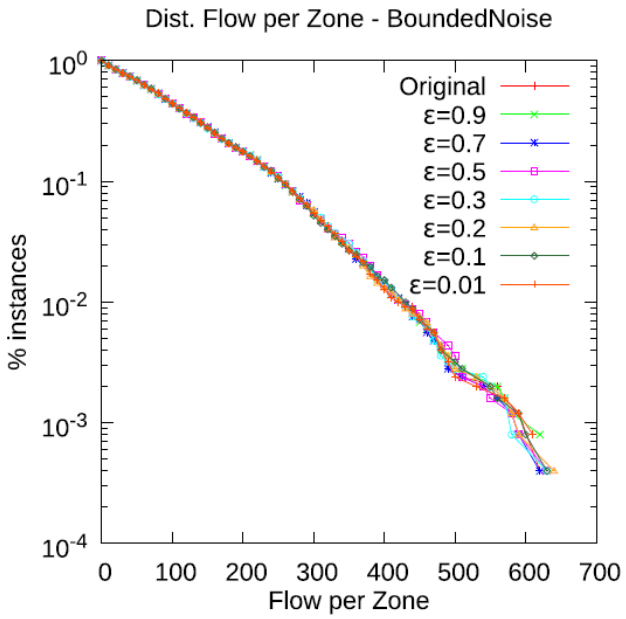

Figure 2: CCDFs of *Flow per Zone* for different levels of protection ε

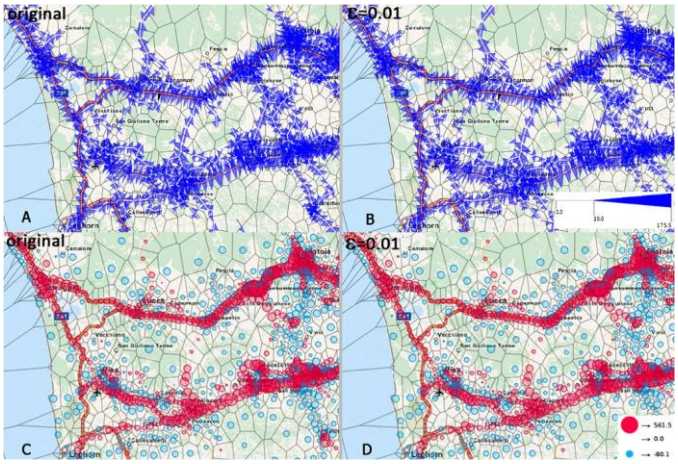

Figure 3: Visual comparison of *Flow per Link* and *Flow per Zone* measures: the resulting overview after the privacy transformation preserves relevant information and properties. Top: flows are rendered as arrows with width proportional to the flow between two regions. Bottom: flow in a region is represented as a circle whose radius is proportional to deviation from median volume in all zones.

The maps allow us to recognize the dense areas (red circles, above the median) separated by sparse areas (blue circle below the median). The high-density traffic zones follow the highways and the major city centres' along their routes. The two comparisons proposed above give the intuition that, while the transformations protect individual sensitive information, the utility of data is preserved.

X. CONCLUSION

The probable impact of the big data analytics and social mining is quite high as it could generate enormous value to our society. Unfortunately, often big data describes sensitive human activities and the privacy of people is always more at risk. The threat is on the rise also thanks to the growing capability to integrate diversified data. In this paper, we have instigated the articulation of the privacy-by-architecture and privacy-by-design in big data analytics and social mining for enabling the design of analytical processes that minimize the privacy harm, or even prevent the privacy harm. We have discussed how applying the privacy-by-design principle to four different scenarios showing that under suitable conditions is feasible to reach a good trade-off between data privacy and good quality of the data. We believe with the privacy-by-design principle social mining has the potential to provide a privacy-respectful social microscope, or socio-scope, needed to observe the hidden mechanisms of socio-economic complexity.

REFERENCES

[1] Monreale et al. EPJ Data Science 2014, 2014:10 http://www.epjdatascience.com/content/2014/1/10

[2] Batty M, Axhausen KW, Giannotti F, Pozdnoukhov A, Bazzani A, Wachowicz M, Ouzounis G, Portugali Y (2012) Smart cities of the future. Eur Phys J Spec Top 214(1):481-518

[3] Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439(7075):462-465

[4] Giannotti F, Nanni M, Pedreschi D, Pinelli F, Renso C, Rinzivillo S, Trasarti R (2011) Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J 20(5):695-719

[5] Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779-782

[6] Song C, Koren T, Wang P, Barabasi A-L (2010) Modelling the scaling properties of human mobility. Nat Phys 6(10):818-823

[7] Cavoukian A (2000) Privacy design principles for an integrated justice system. Working paper. www.ipc.on.ca/index.asp?layid=86&fid1=318
[8] Monreale A, Andrienko GL, Andrienko NV, Giannotti F, Pedreschi D, Rinzivillo S, Wrobel S (2010) Movement data anonymity through generalization. Trans Data Privacy 3(2):91-121
[9] Giannotti F, Lakshmanan LVS, Monreale A, Pedreschi D, Wang WH (2013) Privacy-preserving mining of association rules from outsourced transaction databases. IEEE Syst J 7(3):385-395