Challenges of Privacy Protection in Big Data Analytics

Meiko Jensen*
Independent Centre for Privacy Protection Schleswig-Holstein (ULD)
Kiel, Germany
Email: Meiko.Jensen@rub.de

***Abstract*—The big data paradigm implies that almost every type of information eventually can be derived from sufficiently large datasets. However, in such terms, linkage of personal data of individuals poses a severe threat to privacy and civil rights.**

In this position paper, we propose a set of challenges that have to be addressed in order to perform big data analytics in a privacy-compliant way.

***Keywords*-privacy; big data; challenges**

I. Introduction

The ongoing move towards ever bigger reservoirs of data collected from mostly digital sources implies tremendous changes to the way we interact with online services. Given a sufficient dataset as basis, almost every single user of online services can be identified, leading to neatly tailored online services being provided to each individual. But, while hyped as a fortune by many, the upcoming focus on big data and the implications derived thereof also has its pitfalls.

In this position paper, we argue that the upcoming trend towards big data analytics, and the use of the results thereof, leads to an erosion in terms of privacy and user's rights. Based on a very generic model of big data analytics, we derive a set of key challenges of big data with respect to privacy and civil rights of individuals. Though being far from complete, we outline the core fields of research to be fostered in order to reach a privacy-compliant level of big data analytics.

II. Big Data

In our model, the general term *big data* refers to a huge collection of information (the *dataset*), typically stemming from more than one *data source*, and being processed by a *data analyst* or *data processor*. We assume two modes of big data analytics, depending on the intention of its use.

The first mode consists in verification of a pre-existing hypothesis. Here, the data analyst already has an assumption, e.g. about a certain property of a service's users, that he wants to verify by means of data analytics. Examples range from rather trivial issues like *"are most users of online pharmacy services over 50 years of age?"* to complex questions like *"do we achieve most of our profits from users that saw our advertisements on external websites?"*. In such scenarios, the data analyst will intentionally focus on shaping the dataset to a representation that best-possibly provides the requested answer. Hence, this type of big data analytics is about *verification* of assumptions.

In contrast, the second mode is that of *identification*. Here, the data analyst gathered a large dataset, potentially from multiple sources, and tries to identify interesting facts hidden within the dataset. *A priori*, the data analyst has few to no information about interesting aspects of the dataset, but hopes to stumble over interesting statistical outliers hidden in the overall bunch of information. An example for such type of big data analytics would be *"Hey, look at this! Most players of online game X also bought our new mouse!"*

Here, two key concepts of aggregation of datasets within big data context must be defined. The first is the aggregation of schematically identical datasets, e.g. joining the service access logs of two different online services that are based on the same web server implementation. This is mostly used to gain more information within an existing context.

The second type of aggregation is about linkage created from joining two datasets from disjunct contexts, based on some key information shared in both datasets to be aggregated. A key aspect here is that it must be feasible for the data analyst to identify *links* in the dataset. Such links are data fields that have identical, similar, or otherwise sufficiently related values in different datasets, such as user email addresses, postal codes, or combinations of IP addresses and timestamps. As datasets may stem from different sources, or may contain information from different contexts, a key challenge of big data analytics consists in identifying such linkage.

In this context, the *identity of service users* plays a major role. Given that many types of data in big data contexts are generated by human individuals, using their identity as the linking element of otherwise disjoint datasets is a tempting approach. Unfortunately, as we will discuss in the next sections, this linkage via user identity bears some very challenging pitfalls in the field of privacy.

III. Challenges to Privacy

In this section we iterate through a set of challenges that may threaten privacy of individuals in the context of big data analytics.

*This work was partially funded by the European Commission, FP7 ICT program, under contract No. 257243 (TClouds project) and 318424 (FutureID project).

978-0-7695-5006-0/13 $26.00 © 2013 IEEE
DOI 10.1109/BigData.Congress.2013.39

A. Interaction with Individuals

1) Providing Transparency: Probably the most challenging issue with respect to privacy enforcement in big data contexts is the involvement of individuals. On the one hand, in most big data contexts it is absolutely neccessary to collect and process information that is bound to specific individuals (i.e. is *personally identifiable information*, cf. [1]). On the other hand, including such to the scope of big data analytics automatically entitles each individual to be informed about every type of data processing that involves its data. This right, which for European countries is fixed in the *European Data Protection Directive* [2], must be granted on a per-request basis. However, with huge data amounts as required by big data, it is highly challenging to even identify which information hidden in such piles of data actually are bound to an individual's identity.

As an example, consider one of the most common types of big data sources: network traffic. Here, in order to answer a request for transparency, it becomes necessary to crawl huge lists of IP addresses and timestamps, accompanied with other types of data, such as urls accessed, session cookies used, unique identifiers for certain scopes, etc., always looking for information that might be—or might be not—correlated to the requesting individual. Obviously, this is not feasible in a trivial way, so simplifications have to be made.

Moreover, a request for transparency according to European law does not only cover the personal data of an individual, but also involves detailed documentation of the processes by which such data is processed. Hence, it is not only necessary to disclose the personal data of an individual, but also provide details on the algorithms and processes that are involved in the big data analytics performed with them. Given that most of these processes will contain quite complex data mining algorithms, which moreover may be considered as business secrets of the big data analyst, this legal obligation turns out to pose a major challenge to big data processing on a broader scale.

All in all, providing transparency towards individuals with respect to the type of processing and set of personal data used therein must be seen as the most challenging part of privacy-compliant big data analytics.

2) Getting Consent: Beyond sheer transparency, individuals under certain circumstances may have the right to refuse that their data is being processed in a certain scope or by a certain data processor. Thus, many privacy laws and regulations entitle individuals with the right to be asked *prior* to processing of their data. This aspect, gathering consent from individuals prior to processing of their personal information, suffers from the same complexity isues as listed in the previous section. Moreover, most regulations do not only require consent, but even *informed consent*, meaning that the individual must be able to understand what sort of processing is performed, and what may happen with their data as a result. Given that many types of big data analytics are based on highly complex data mining algorithms, *informed consent* then implies that each individual must be provided with an explanation of all of these algorithms such that they can *understand* what happens there. This, again, must be considered a tremendously challenging issue with respect to big data analyses to come.

3) Revocation of Consent and Deletion of Personal Data: Similar to the issue of getting consent, an individual may decide to revoke its given consent for processing its personal data at a later stage. For instance, an individual may decide no longer to trust a data collector, e.g. due to assumed leakage of sensitive data.

In that case, the legal obligations—at least according to European privacy laws—grant such individuals the general right to revoke their consent, implying that all processing of personal data of such individuals has to be stopped as soon as possible, and that furthermore all personal data of those individuals have to be deleted. Given that the personal data of a particular individual may have been spread widely among data collectors and data analysts, implementing such revocation requests turns out to become a highly challenging issue.

B. Re-Identification Attacks

Another major threat with respect to privacy in big data analytics is the ability to perform "re-identification attacks", meaning that a huge dataset available is explicitly scanned for correlations that lead to a unique fingerprint of a single individual. More precisely, by linking different types of datasets together, the uniqueness of each entry is increased, up to the point that a link back to an individual's identity can be established.

As an example, consider the following. As discussed in [3], the Internet Search Provider AOL in 2006 published a set of search terms of their users. The assumption was that the search terms themselves, correlated only by means of a number as a user identifier, would be anonymized sufficiently to prevent identification of individuals among the search query authors. Unfortunately, as it turned out, re-identification of individuals was indeed feasible, and Thelma Arnold was re-identified by her search terms only—which were assumed to have been properly anonymized. This analysis, which linked the search term database to other publicly available databases of U.S. citizens, impressively illustrates the power of re-identification attacks. However, it is worth noticing that a re-identification attack is an intentional act, which starts with a dataset (anonymized, pseudonymized, or plain), and ends with the identification of an individual.

There are three sub-categories for this type of attack, namely *correlation attacks*, *arbitrary identification attacks*, and *targeted identification attacks*. These are to be described next.

1) Correlation Attacks: This type of attack consists in linking a dataset of mostly uniform data values to other sources in order to create more unique database entries. For instance, linking pseudonymized customer data of pharmacies to equivalently pseudonymized data of medications obtained from a hospital leads to more fine-grained data per entry. More precisely, if one database lists userIDs and pharmacies visited, and the other lists the same userIDs correlated to medication prescriptions, the correlated database consists of entries that indicate which hospital patient bought its medication at which pharmacies. Thus, the correlated dataset has more information per userID, allowing for a more precise analysis on the individuals behind the data. A correlation attack, in this scenario, consists in linking additional datasets in up to the point that there is either at least one entry in the correlated dataset that is unique in its combination of data fields (despite the userID), or up to the point that no two entries in the database have all data field values identical (again despite the userID). Both variants are possible, and both can serve as a basis for a subsequent other type of identification attack, as discussed next.

2) Arbitrary Identification Attacks: The primary goal of this type of re-identification attack is to link at least one entry in the given dataset to the identity of a human individual. Typically, the attack is completed when a link between one entry in the aggregated dataset and one identity of a human being can be drawn, with a sufficient level of probability.

This type of attack neatly illustrates the failure of anonymization for a certain set of anonymized data. The AOL example given above is a perfect example, as it illustrates that the anonymized dataset—in correlation with other datasets—was sufficient to identify a *few* individuals (besides Thelma Arnold) among the huge set of users by name. A key aspect here is that is was not feasible to link *all* search terms to real-world individuals. Hence, given the identity of an arbitrary human individual a priori, it is very unlikely to be able to find an entry in the correlated dataset that can clearly be attributed to that identity. In that point, the arbitrary identification attack differs from the targeted identification attack.

3) Targeted Identification Attacks: With the targeted identification attack, an adversary tries to find out more details for a given human being. Hence, the targeted identification attack is only successful if it is possible to link some entries in the database to the identity at hand, with a sufficient level of probability. In that sense, the AOL search term issue was not a targeted identification attack, as it was not the adversary's intent to find out information on Thelma Arnold, but to identify any person within the dataset.

Targeted identification attacks can be considered the most threatening type of re-identification attacks, as they are likely to have the largest impact to an individual's privacy. Depending on the relation of adversary and searched individual, privacy issues may rapidly come up, being performed intentional or by accident. For instance, if an employer searches a huge dataset of pharmacy customers for occurrences of its employees, this may disclose information on medical treatments—and hence illnesses—of its employees, which clearly is a violation of those individual's privacy. The same obviously holds if the employer was actually looking for information on, say, himself, but happened to stumble over the name of an employee. Though not being intentional in this case, the privacy violation remains.

C. Probable vs. Provable Results

A key threat to big data analytics is the validity of the results gathered. Depending on the query a big data analyst performs on the dataset available, different types of results can be produced. For instance, if the dataset was formed by correlation of email addresses contained in all of the correlated datasets, it may be assumed that two entries that held the same email address before correlation will lead to a single database entry that reflects all information on the same identity, i.e. the holder of the email address.

However, in some cases, this assumption may not hold. For instance, it might be possible that an individual managed to spoof its email address within one of the datasets. This would cause a false correlation of dataset entries, and hence result in incorrect correlated data. Additonally, some email addresses may be shared by two or more individuals, causing false data linkage if used for correlation to individuals. Either way, the use of email addresses is not 100% reliable, but gives a very precise notion of "identity" in the sense of uniqueness and correlation within a dataset.

The situation gets more challenging if the type of linkage is not that reliable. For instance, correlating dataset entries by means of IP addresses and timestamps, e.g. within online service access logs, can never give a reliable, 100% trustworthy correlation. Depending on the distribution of access events, and the temporal distance between two events originating from the same IP address, a correlation is possible, maybe even likely, but never guaranteed. Given the high degree of flexibility in terms of using IP addresses, it is perfectly feasible that the same IP address is used by two completely different individuals, with an almost negligible delay between the two resulting events. Hence, if using such "probability-based" linkage of datasets, the general probability of false correlations is increased. In contrast to linkage via email addresses (that are either identical or not identical), the linkage of IP addresses and timestamps is rather fuzzy. Hence, instead of a *provable* link (in the sense discussed above), this type of link merely is a *probable* one.

The threat to privacy in this terminology obviously stems from big data query results that originate in such probable links. For instance, if the big data query was to find the exact number of users that accessed both system A and B wihtin the last five weeks, that number may become falsified by considering such false correlations. Depending on the

subsequent use of such a result (as part of another big data query), this effect may cascade, resulting in a completely wrong final statement.

D. Economic Effects

Despite the threats caused by interaction with concerned individuals, intentional attacks, or issues of false data processing methodology, a fourth category of threats covers the economic issues of the big data paradigm. Given that most types of big data analytics require a huge set of different datasets in advance, it becomes necessary to exchange such datasets among business partners. This can e.g. be based on mutual agreement, legal obligations, or by means of an economic *data market* where data providers sell data of their users to their customers. In that latter case, a lot of threats to privacy may arise from economic considerations in such data trading. Out of these, in the following the two examples of *confusion and distraction* and *context faults* will be discussed. Note, however, that there are many more threats arising in this category, ranging from fraud to censorship to surveillance, which are not discussed here.

1) Confusion and Distraction: The idea behind the approaches of confusion and distraction is to make a sold dataset less useful for a customer, e.g. in order to influence a competitor's business strategies. By means of slightly altering the original dataset in some way, it becomes possible to push subsequent big data analytics towards a certain direction. For instance, a company A may decide to sell their user's data to one of their competitors, company B, but does not want to reveal the number of users that actually are using both A and B (as this may be a critical information w.r.t. the business strategies of both A and B). Thus, A decides to preprocess the dataset sold to B in a way that all users that also use services of B are removed. In this rather trivial case, B will certainly notice the lack of an intersection, and will detect the alteration.

However, more subtle modifications applied to the dataset by A may be less obvious, but still impact on the results gathered out of that data at B. Especially in situations where B regularly buys and processes datasets from A, the opportunities of A to shape its sold dataset in such a way are multifold.

Either way, as such alteration of datasets inevitably implies modification of entries of individual users (by means of deletion, alteration, or duplication), the impact to the privacy of individual users of A becomes evident. The economic environment induces a falsification of big data analytics results here, with all implications w.r.t. privacy, as discussed for probability-based linkage in Section III-C.

2) Context Faults: Another threat arising from the economic scenario of selling big data datasets consists in faults of context correlation. If company A sells its user's data to company B, this is not necessarily accompanied with an in-depth description of each data field's semantics. Though that information might be obvious (e.g. for email addresses) or self-descriptive (for XML data), there nevertheless remains the threat that B interprets some of the data coming from A different that they were contexted at A. For instance, if A accompanied all their user's data with a self-calculated field *activity index in %*, the calculation of such an activity index may be different from the calculations B performs for its own users. If, in that case, B blindly correlates A's activity index fields and B's activity index fields, this might e.g. result in the interesting observation that A's users are a little more active than B's users—though in reality it was just a different type of calculation that was not spotted in the big data analytics.

Though the given example might be a little simplistic, more complex data, stemming from advanced aggregation algorithms, may induce such context-related misinterpretations, which ultimately result in false results to big data queries, as discussed above. Again, as big data queries often are about individual's properties, this may lead to false statements w.r.t. individual users (e.g. "*none of our users uses product X*"), affecting their privacy again.

IV. Conclusions and Research Indications

As can be seen, the field of privacy in big data contexts contains a bunch of key challenges that must be addressed by research. Many of these challenges do not stem from technical issues, but merely are based on legislation and organizational matters. Nevertheless, it can be anticipated that it is feasible to meet each of the callenges discussed here by means of appropriate technical measures. For instance, keeping track of a particular individual's data throughout big data analytics contexts is merely an organizational requirement that can e.g. be met by means of logfiles. Linkage of disjoint datasets can often be performed without relying on linkage via user's identities, but based on other types of data. In the same direction, many types of data can be preprocessed with proper *anonymization* or *pseudonymization* prior to sharing, such that linkage of datasets remains feasible, but linkage to an individual's identity becomes hard.

Each of these fields already has a lot of pre-existing research that has already been performed. Hence, the future directions for research in this red hot topic are obvious: applying these existing techniques to the context of big data.

References

[1] E. McCallister, T. Grance, and K. Scarfone, "Guide to protecting the confidentiality of personally identifiable information (pii)," 2010.

[2] The European Parliament and the Council, "Directive 95/46/ec on the protection of individuals with regard to the processing of personal data and on the free movement of such data," 1995.

[3] M. Barbaro and T. Zeller, "A face is exposed for aol searcher no. 4417749," 2006. [Online]. Available: http://select.nytimes.com/gst/abstract.html?res=F10612FC345B0C7A8CDDA10894DE404482