Big Data characteristics and their implications for data privacy

General remarks

What are the Characteristics of Big Data and How do they Affect the Risk of Identification?

Below is a list of Big Data specificities in terms of privacy, identified by the Big Data Privacy Group. It is written from the perspective of an official agency releasing Big Data or statistics based on it.

o Velocity. Data get “younger and younger”. The time of day (perhaps within seconds) associated with data could make it identifiable. Related to this, if a constant stream of new data are released in real time, data that were previously considered non-identifiable, may quickly become identifiable. This may mean assessments about identification risks may need to be ongoing.

o Variety. The need for the creation of multiple solutions for the protection of privacy in the context of multiple data sources which contain different types of data available at different levels of aggregation. Given the variety of information, all people or businesses may be identifiable in terms of at least some variables.

o Size. Virtually unlimited. Use of current risk assessment techniques on such large data may be computationally difficult and involved.

o Veracity (or accuracy). Reporting errors, poorly defined metadata, data that are masked for the purposes of statistical disclosure control or data that are out-of-date, will reduce the likelihood of identification. For example, the data obtained directly from people, especially through social media, may be less accurate and so less likely to lead to identification compared with information obtained from an institution, such as a bank.

o Value. The value could be high if there is the potential for substantial monetary reward from identification (e.g., commercially sensitive data) or if the information is highly private or personal (e.g. health conditions, political opinions or personal behavior). The greater the value, the more likely attempts at disclosure will be made and the greater the disclosure risk.

o Availability. Very different entities can be data owners. The type of cooperation with them depends on whether we get the raw data or already aggregated, and therefore affects the type of confidentiality protection required.

o Aggregate. A significant part of the data will be aggregated at the stage of obtaining them from the owners.

o Awareness of society. The need to educate and reassure society about the use of Big Data will affect the public view on the level of confidentiality protection required

o Geographical differences. Limitations in access to new technologies in some societies cause the need to ensure adequate privacy tools, regardless of the available technology.

Data access strategies for big data

Over the years statistical organizations established three different settings for providing external researchers access to their data: dissemination of microdata to the public, onsite access in research data centers (RDC), and remote access (see also Deliverable 1 for a detailed discussion of the three access strategies). In the context of Big Data, microdata dissemination is no longer an option in most cases. One feature of Big Data is that the size of the data implies that transferring the data is cumbersome. It would not be possible to send the data to the researcher on CDs or provide a link for a download as is the current practice with microdata dissemination. Thus, onsite and remote access will be the only viable solutions. Onsite access has the advantage that the agency has a better control over who accesses the data and what is done with the data. Researchers usually are not allowed to bring any own devices such as laptops or cameras to RDCs, all their activities are monitored, and no analysis result leaves the RDC before it has been carefully checked by RDC staff for potential confidentiality violations. However, this manual output checking is a labor intensive task and most RDCs are already working at their limits resulting in increasing waiting times for the researchers before they can obtain the cleared outputs. If the standard set of surveys and administrative data that is already offered at the RDC will be enriched by new data sources it is likely that offering manual output checking for all these databases will no longer be feasible. It is thus more relevant than ever that general strategies for ensuring the confidentiality of the generated outputs without manual intervention are developed. These strategies will likely consist of a combination of output checking to suppress risky outputs and output perturbation to allow the release of outputs that are considered safe in most situations but could be misused by an ill-intentioned user, to learn some private information about specific individuals in the database. If general strategies to automatically ensure confidentiality of the outputs can be developed and a secure connection to the server hoisting the data can be guaranteed it will only be a question of convenience whether the data will be accessed on site or remotely. However, whether this goal can ever be achieved is currently an open question and an area of active research.

Approaches to privacy in the context of Big Data

In this section, current and future approaches to privacy in the Big Data context are presented.

Possible Use of Mobile Data

Potential information from Big Data about mobile phone users include the GPS location, time of day, and duration of phone call, and the GPS location and time of day of text messages. The GPS location could be the location of the mast that routes the communication or the location of the phone user at the time of communication. There are many important questions that mobile device data could help answer, such as "what are the daily and weekly changes in regional population counts", "are there any seasonal patterns in population counts", "are there periods of un-seasonal movements and can these be attributed to specific events, such as severe drought or storms", to name a few. The benefit to society of answering such questions is clear. For example, it could assist governments target transportation infrastructure spending where it is most needed. There is also the potential to link people who are selected in surveys that are conducted by statistical organisations to their mobile phone data (e.g. when selecting a person in a survey, ask the person for their mobile phone number). This would allow analysis of how personal characteristics (e.g. employment, education) are related to their mobility.

From a privacy perspective, it is interesting to ask whether a person's GPS locations are sensitive and, if so, at what level of geographic detail (e.g. suburb, state) can GPS locations be aggregated so that they are no longer sensitive? Furthermore, how identifiable is GPS location? Clearly a mobile phone's GPS location centred on a particular house would likely mean that the mobile phone belongs to a person who lives at the house. This would clearly be a disclosure of the person's GPS locations. Again, at what level of geographic detail would GPS locations need to be aggregated so that they are no longer identifiable? The answer to this question is related to the frequency with which GPS locations are taken- the more frequent the GPS locations are taken, the more likely that they will lead to idenfication. If a mobile phone user travels to a particular sequence of suburbs on a routine basis, this could be potentially identifiable.

The PARAT software

Protected dissemination of patient-level data is critical for healthcare organizations to increase the quality of care, improve patient safety and reduce costs. The PARAT software, which automates de-identification and masking of data for secondary use, has been developed and utilized in Canada. PARAT works based on peer-reviewed algorithms and technology, and it is in compliance with the United States Health Insurance Portability and Accountability Act, and other legal requirements for sharing data. The Ontario Cancer Data Linkage program, which aims at linking Ontario's rich cancer data resources and providing the de-identified data directly to health services researchers, used PARAT to assess the risks of re-identification and to de-identify the data. Key government officials expressed their approval of using PARAT for this program. To know more about PARAT: http://www.privacyanalytics.ca/software/

Hadoop (HDFS and MapReduce)

Hadoop is an open- source software project, developed by the Apache Software Foundation, targeted at supporting the execution of data--oriented applications on clusters of generic hardware. The core component of the Hadoop architecture is a distributed file system implementation, namely HDFS. HDFS realizes a file system abstraction on top of a cluster of machines. Its main feature is the ability to scale to a virtually unlimited storage capacity by simply adding new machines to the cluster at any time. The framework transparently handles data distribution among machines and fault--tolerance (through replication of files on multiple machines).

The basic idea behind the processing of Big Data consists in distributing the processing load among different machines in a cluster, so that the complexity of the computation can be shared. The number of the machines can be adapted in order to fit to the complexity of the task, the input size and the expected processing time. Besides managing distributed storage, a technological platform such as Hadoop also provide the abstraction of distribution of processing in a cluster.

The MapReduce paradigm is a programming model specifically designed for the purpose of writing programs whose execution can be parallelized. A MapReduce program is composed by two functions, "map", specifying a criteria to partition the input into categories and "reduce" where the actual processing is performed on all the input records that belong to a same category. The distributed computation framework (e.g. Hadoop) takes care of splitting the input and assigning it to a different node in the cluster that takes in turn the role of "mapper" and "reducer". Hence, the general processing task is split into separate sub-task and the result of the computation of each node is independent from all the others. The physical distribution of data through the cluster is transparently handled by the platform and the programmer must only write the processing algorithm according to the MapReduce rules.

MapReduce is a general paradigm not tied to a specific programming language although the core implementation in Hadoop requires MapReduce programs to be written in Java.

Pig and Have, from the SandBox

Pig is a high level interface to MapReduce. Writing MapReduce programs in Java can be complex and also common operations on data, like joins and aggregations, may require a significant amount of code. This requires trained IT developers, slowing down the analysis of data. Pig is based on a high-level language, namely PigLatin, oriented to data transformation and aggregations. Complex operations on data can be performed with short scripts, that can be simply read and modified also by business analysts.

Hive is a SQL interface to MapReduce. It allows to structure data in HDFS in tables, like in a relational database, and to query data using familiar SQL constructs such as selections, joins and filters. Hive and Pig have different purposes and complement themselves in the Hadoop ecosystem: Hive is oriented to interactive querying of data, while Pig allows to build complex transformation flows.

RHadoop

RHadoop is a R library that allows to write MapReduce programs in R. Once installed and configured, it integrates with the Hadoop cluster, allowing to read and write files from/to HDFS and to stream MapReduce jobs over the cluster. The advantage it provides consists in the possibility of using a language which is already familiar to statistical users, allowing to work on Big Data with a graceful learning curve and exploiting established know-how.