A Review of the Taxonomy of Private Data

-*Abhishek Gupta*

"Data today is not just information, but an asset" says Irving Wladawsky-Berger. Based on the World Economic Forum (WEF) reports on "Re-thinking data", Irving states that data kept at a place is like money hiding under a mattress. Data needs to move to create value, much like money. But Personal data is as dangerous as useful and presently, it lacks the trading rules and policy frameworks which are required for this "movement of data". It can be essentially considered as a physical asset, but if questions like protection, security rights and responsibilities are not answered, there will continue to be breaches, intentionally and un-intentionally. There can never be trust amongst stakeholders unless there is sufficient accountability for the information. [1]

Before any of the solutions to the above problems can be devised, it is important to frame the questions correctly. Bruce Schneier in his work [2] presented in the Internet Governance Forum discusses this exact problem of formulation, the questions to which solutions need to be devised. He believes that it is more important to classify the data first. It is however important to identify the nature of data and how much it influences the society i.e. its impact factor.

Over the last decade, Social Networks have taken an important place in the society. Data can now be classified into Conventional and Networked data where, the latter represents the data that show relation between the sources of the data [3]. In fact, social networks are the biggest sources of data movement and hence the biggest sources of potential data crimes and security breaches. To be able to address these problems in a more quantized manner, Bruce in [2] presents a classification of social networking private-data. This is a short compilation of his taxonomy:

a. Service Data : Basic Information like legal name, credit card number etc that we provide.
b. Disclosed Data : Stuff we post - Articles, videos, photos.
c. Entrusted data : Data we post on others pages. Now they have control over this information.
d. Incidental data : Posts in which others tag you in; you have no control.
e. Behavioral data : The data that the sites collect about our behavior to post relevant ads.
f. Derived Data : Conclusions that can be drawn after seeing ones profile.

Social networks are not the only place for data exchange. In fact, any environment that facilitates the exchange of information amongst people, i.e. a collaborative environment, falls into the category of maximum data flow. Geoff Skinner, Song Han and Elizabeth Chang in [4] have performed their research in this 'Collaborative Environment' and they present a classification for private information.

They define privacy as the interest an individual has in controlling, or at least significantly influencing, the handling of data about themselves. They classify privacy and private information in three dimensions: space, time and matter. Each dimension has multiple views like computational, structural and content that further describe that class. Its time relevance relates to the amount of time and resources required to compromise the stated level of privacy protection. There are further three categories identified: ideal privacy, computational privacy and fragile privacy. The matter dimension, and therefore the content view, reflects the privacy of collaborative environment objects. Its relevance relates to the different types of data that require privacy. Three categories of objects have been defined and each of them has been classified accordingly. The three include -

Data Privacy, Identity Privacy, and Meta Privacy. The space dimension, which presents the structural view, reflects the privacy of collaborative environment entities. Its space relevance relates to the different types of privacy applied to various entities and relationships within a Collaborative Environment. This form of classification helps us to look at the data from three dimensions in a particular environment.

However, a more comprehensive classification that includes both collaborative and non-collaborative environments, is essential and hence the Government of Alberta (GoA) defined privacy in [5] as an attempt to devise an architecture for privacy. The author initially describes the use having a taxonomy for private information and what purpose it serves. The taxonomy basically has three levels:

- Root Level – contains "universal" dimensions that reference outside standards wherever possible.
- GoA Level – contains GoA specific dimensions that will be common across GoA.
- Ministry Level – contains Ministry-unique dimensions common within a Ministry.

Each of these three levels is discussed in depth and includes categorization as to where each type of data falls in.

It is important to note that the classification scheme followed by GoA is based on the web standards as defined by the Platform for Privacy Preferences (P3P). This makes it easy for user agents to automatically retrieve and interpret the format. GoA has crafted their model on top of this P3P standard with the Root layer as the base. The GoA and Ministry levels are built above this layer to add more description to data, categories, purposes and recipients. Hence if at any point of time, all these dimensions have to be resolved back, they would all trace down to the root level.

These approaches to classification focus on understanding and clearly stating the problem and thus help by classifying them. However, it is important to also understand the risks involved in-case security breaches occur. In article [6] by Daniel J. Solove in the University of Pennsylvania Law Review brings out this aspect in a creative way.

He provides a framework within which the legal system could come to a better understanding of privacy. He aims to develop a taxonomy that focuses more specifically on the different kinds of activities that impinge upon privacy. He endeavors to shift focus away from the vague term "privacy" and towards the specific activities that pose privacy problems. Although various attempts at explicating the meaning of "privacy" have been made, few have attempted to identify privacy problems in a comprehensive and concrete manner.

The purpose of his taxonomy is not to argue that the law should or should not protect against certain activities that affect privacy. Rather, the goal is simply to define the activities and explain why and how they can cause trouble. There are four basic groups of harmful activities: (1) information collection, (2) information processing, (3) information dissemination, and (4) invasion. Each of these groups consists of different related subgroups of harmful activities. This is shown in the figure 1 below:

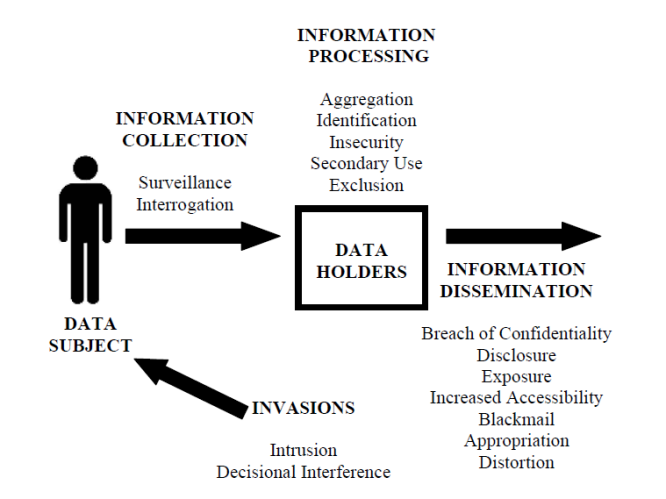

Figure 1: Groups of Harmful Activities

He discusses in detail about each of the stages mentioned in the above data flow and how each of these can lead to privacy issues and cause problems. Even in stages where information is just collected, like surveillance and interrogation or its processing, there can be cases which can lead to privacy breaches.

While classifying information it is important to note who precisely owns the data and who has the right to view and/or edit that information. Most of the commercial organizations, in any sector, tend to store private information. The concerned firm then runs statistical analysis on this private data to form consumer groups and provides services to these groups based on the results of this analysis. Individual people, households, communities and businesses are the subjects of this data collection and the corresponding analysis. People as consumers, have very little knowledge about the what-and-how of this data.

Like in any of the environments discussed above, the focus is on the subject and hence data can be classified as information on financial, commercial, health, employment, travel and relationship of a person. Such "People-data" can be either permanently stored (like health records) or temporarily stored (like closed-circuit television recordings). Such an approach to classification is termed as Statistical Approach. [7]

Amongst the members of these groups, data can be either partly or fully shared. For example, all health and financial data might be shared amongst the (senior) members of the household. However, in businesses, the employer holds all information about its employees and not vice versa. In certain scenarios, it is important to understand how the sharing entities view each other. One such example is the data from surveillance. The perception of the legitimacy of holding and using this data depends on how the people view the government and vice versa.

The above approaches together encompass five approaches to classifying data – Social Networks, Collaborative Environment, Legal Framework, Government Data and Statistical Data. These approaches, together, aim to provide a comprehensive taxonomy to private data. However, there are still many issues that need to be addressed and more importantly, different entities have to be able to come to a consensus to follow a developed standard.

References

[1] http://blog.irvingwb.com/blog/2012/07/rethinking-personal-data.html
[2] http://www.schneier.com/blog/archives/2010/08/a_taxonomy_of_s_1.html
[3] http://faculty.ucr.edu/~hanneman/nettext/C1_Social_Network_Data.html
[4] http://www.wseas.us/e-library/conferences/2006hangzhou/papers/531-250.pdf
[5] http://www.ipc.on.ca/images/Resources/up-PPPP062.pdf
[6] https://www.law.upenn.edu/journals/lawreview/articles/volume154/issue3/Solove154U.Pa.L.Rev.477(2006).pdf
[7] http://www.isi-web.org/about-isi/professional-ethics
[8] http://www.cmu.edu/iso/governance/guidelines/data-classification.html
[9] http://www.ico.gov.uk/upload/documents/library/corporate/research_and_reports/executive_summary.pdf