Understanding Data and their Environment: Assessed Work (Essay)

Student number: 8520000

Introduction:

In this essay I discuss the release of data from the UK government's "Troubled Families Programme" (DGLG, 2017) to researchers willing to conduct in-depth analyses of the programme. Specifically, I evaluate the risks presented by such a release, and discuss ways in which said risk could be mitigated.

I will begin by outlining my approach, I will then conduct an analysis of the Troubled Families situation and, in a final section, I will lay out some recommendations to minimise the risk associated with allowing researchers to access the data set.

Description of approach:

The approach will draw heavily on the Anonymisation Decision-making Framework (ADF)...

Analysis of Troubled families case:

In this section I will analyse the Troubled Families case. I will draw heavily on the ADF, though I will not be going through all the components of the framework.

The data situation:

This is a complex data situation in which there are a number of data controllers. Indeed, the data set is created by combining administrative data collected by a number of government departments and local authorities. These are all joint data controllers, as is the Department for Communities and Local Government, who holds the pooled data.

Assessing risk:

There are a number of technical and non-technical reasons why the Troubled Families data set is risky. Risk is usually understood as some function of both the potential harms of an event and the probability of that event occurring. In this section, I will first discuss why the potential harms from a confidentiality breach are very high, then I will analyse the likelyhood of a data breach happening.

Potential harms:

The ADF proposes three questions for gauging whether elements of the data contribute to the sensitivity of the data situation:

1. Are some of the variables sensitive?
2. Are the data about a vulnerable population?
3. Are the data about a sensitive topic?

The more questions answered yes, the more sensitive the data situation (Elliot, et al. 2016: 64-65). I will consider each of these question in turn for the Troubled Families data.

1. Are some of the variables sensitive?

We do not have access to the exact variables that would be available for release, however we can to a large extent infer their presence from available reports (such as DGLG. 2017). I will first consider the variables considered sensitive by the Data Protection Act. These are data on:

"(a) The racial or ethnic origin of the data subject, (b) Their political opinions, (c) Their religious beliefs or other beliefs of a similar nature, (d) Whether they are a member of a trade union, (e) Their physical or mental health or conditions, (f) Their sexual life, (g) The commission or alleged commission by them of any offence, or (h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings" (Elliot, et al. 2016: 64-65).

There does not appear to be any information on the data subject's political opinions, their religious beliefs, or the mebership of trade unions.

However, the data set *does* contain detailed information on the ethnicity of the data subjects as can be seen in table 2 of the family outcomes document (DCLG.2017:6).

Furthermore, there is a wealth of information on the mentaland physical health conditions of the data subjects.

Another relevant technical aspect of this data is that it is hierarchical data. Specifically, it contains information on individuals within households. ()

Compounding this issue is the fact that the core use case of this data set is the study of troubled *families*, this means that potential approaches to mitigating the deanonymisation risks posed by hierarchical data, such as simplifying the data to a non-hierarchical structure, would run counter to the third core principle of the ADF: "Anonymisation is a process to produce safe data but it only makes sense if what you are producing is safe useful data" (Elliot,et al. 2016:4-5). The usefulness of a data set on families which does not contain families is clearly questionable.

Indeed, the stated motivation for the data being made available to researchers is for them to conduct deep analyses of issues that occur in families. However, as I will argue below section, it does not follow that hierarchical data is required for *all* analyses and it may be possible to seperately release a dataset that has been simplified to a non-hierarchical structure, with less restrictions on access than

One aspect of the data set that would *decrease* the risk associated with the data situation is the fact that it is unclear whether a given troubled family in actually part of teh data set i.e there is limited response knowledge.

Intruder motivations

There are a number of potential scenarios that should be considered. I present two possibilities here, though this is by no means exhaustive.

Nosy neighbor:

A nosy neighbor attack may not initially appear to be the most likely source of risk here, however, there is reason to believe that it should still be seriously considered. For example, 41.7% of families in the data set had been visited by the police in the year prior to starting on the programme (DCLG. 2017:13). This is typically a highly visible event for neighbors, with the police often leaving their emergency vehicle lightings on for the duration of the visit.

A researcher with access to the data set may seek to use their access in combination with their knowledge of their neighbors to gain sensitive information about the nature of the disorder.

Furthermore they may have access to information on their neighbors via social media, for example as direct 'friends' or as co-members of a neighborhood Facebook group. This extra information can increase the ease of identifying their neighbor in the data set, though it could also potentially *decrease* the motivation, for example if some of the information that we wanted to protect was published by the neighbor themselves on social media (Elliot,et al. 2018:2).

Political motivations

The key data controller in this situation is the Department for Communities and Local Government (now the Ministry of Housing, Communities and Local Government), which is associated with five ministers including a secreatary of state i.e. very senior politicians (MHCLG, 2020). It is conceivable that a serious breach of such a sensative data set would be a major embarrasment for the government, and could trigger the resignation of one of these senior politicians. This is important as it provides plausible motivations for data intruders.

A number of activists, with a wide spectrum of capabilities, could be interested in causing a major upset in the ministry. This could include opposition activists seeking to discredit the government, members of the government eager for their own cabinet position, disgruntled former employees, or even a foreign state seeking to cause disruption.

Implications/recommendations:

There are a number of data controllers in this situation, an important reccomendation is that they establish in writing who takes responsibility for various elements of the release, including ensuring DPA complience, and how to minimise harm in the event of a breach (One advantage we have in this data situation is that the most important data controller, effectively the UK government, is *exceptionally* ressourcefull).

As I have argued above, the data set is extremely detailed and sensitive. Furthermore, the use case generally does not allow for the data be modified to mitigate these features. Whilst there may be some exceptions where heavily modified data could be usefull to researchers, in most cases the most appropriate solution would be to leave the data set largely intact and to ensure that access is heavily restricted, and outputs checked for disclosivness.

There are two potential ways of acheiving this: only allowing access to the data at secure data centres, or virtual access via an analysis server.

Both of these solutions can be inconveniant to the user, either due to the cost of travelling to the secure data centre, or due to the difficulty of exploring data you cannot see and restictions on allowed quieres if accessing via an analysis server (O'Keefe, et al. 2014: 308).

Furthermore both solutions may prevent the user from using their software of choice, for example if the data centre only permits access on certain computers (with USB ports and external disk drives disabled) or if the analysis server is only accesible via a custom interface.

However, despite these drawbacks, both approaches allow for the disclosure risk to be reduced to a negligable level whilst preserving the usefullness of the data set.

There may well be further steps that could be taken to decrease the risk of releasing the data set to researchers, for example removing unique combinations of variables or simplifying continous variables into categories, it is hard to say without more information on the data set (such as the variables).

Conclusion:

In this essay I have discussed issues surrounding the release of data from the Troubled Families programme. Drawing heavily on the ADF, I have argued that the data set is *exceptionally* sensitive and that there are a number of conceivable motivations for attempting a data breach.

Bibliography:

Department for Communities and Local Government. (2017). 'National evaluation of the Troubled Families Programme 2015 - 2020: family outcomes – national and local datasets: part 1' Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605185/Family_outcomes_-_national_and_local_datasets.pdf

Elliot, M. J., Mackey, E., O'Hara, K., & Tudor, C. (2016). 'The Anonymisation Decision-Making Framework'. UKAN publications: Manchester.

Elliot, M., O'Hara, K., Raab, C., O'Keefe, C.M., Mackey, E., Dibben, C., Gowans, H., Purdam, K. and McCullagh, K., (2018). Functional anonymisation: personal data and the data environment. *Computer Law & Security Review*, 34(2), pp.204-221.

Ministry of Housing, Communities and Local Government. (2020) "Our governance". Available online: https://www.gov.uk/government/organisations/ministry-of-housing-communities-and-local-government/about/our-governance