AVOIDING THE JIGSAW EFFECT: EXPERIENCES WITH MINISTRY OF JUSTICE REOFFENDING DATA

Kieron O'Hara
University of Southampton

Edgar Whitley
London School of Economics

Philip Whittall
Detica
12th Dec, 2011

The Prime Minister's latest transparency commitments, set out in a letter of 7th July, 2011, include the following:

Opening up access to anonymised data from the National Pupil Database to help parents and pupils to monitor the performance of their schools in depth, from June 2012. This will enable better comparisons of school performance and we will look to strengthen datasets in due course

Sentencing data by court will be published by November 2011, enabling the public to see exactly what sentences are being handed down in their local courts, and compare different courts on a wide range of measures. The data, anonymised, will include the age, gender and ethnicity of those sentenced, the sentence given, and the time taken at each stage from offence to completion of the case in court

Data on performance of probation services and prisons including re-offending rates by offender and institution, to be published from October 2011

The publication of these data is likely to be very important for the development of innovative services in areas such as education and criminal justice, and the government's strong and abiding commitment to transparency should be applauded.

However, as Kieron O'Hara advised in a report for the Cabinet Office earlier this year, *Transparent Government, Not Transparent Citizens*, there is an issue with anonymised databases. Databases are anonymised by removing identifiers (such as names, addresses, etc) from the data. However, it is sometimes possible, using auxiliary information, to piece together enough information from other sources to re-identify the data subjects (hence the term 'jigsaw identification'). US legal scholar Paul Ohm has described the science and set out a worst-case legal scenario in a well-known paper; even if one does not endorse his vision, it is clear that there is a probably small but not well-understood risk of jigsaw identification with the inclusion of anonymised databases in a transparency programme, for three reasons.

1. The governing principles of the transparency programme preclude the possibility of controlling or withdrawing access to the data.
2. The risks of jigsaw identification grow with the amount of auxiliary information available to the identifier. Of course the amount of data on the World Wide Web grows annually, particularly thanks to information on social networking sites and local press coverage. The transparency programme is also likely to be a major contributor over time.

3. Jigsaw identification is computationally complex. However, dramatic increases in computer power have lowered the barriers, and we can expect them to continue falling, so that soon even small, personal devices might be sufficient for someone to identify people, or disclose sensitive information about them, from anonymised data.

Given the uncertainty in this area, it is important to develop and archive best practice in anonymisation. To that end, a group of academics, students and professional data security consultants worked with statisticians at the Ministry of Justice in order to ensure that the anonymity of the individual level reoffending and sentencing data would be as secure as possible.

The risk

The data would not include explicitly identifying data (such as names or ages). However, offenders could be identified, or disclosures made about them, if they had a unique combination of characteristics which could be matched against outside information. Groups of offenders sharing characteristics could also provide a route to disclosure; for example, if each member of a group has characteristics A, B, C and D, and a named individual with characteristics A, B and C can be placed within the group, then further facts about the individual could be deduced (viz, that he or she had characteristic D).

The Ministry of Justice were of course aware of these risks, and were keen to minimise them.

The method

It was decided to test the risk of identification and disclosure, but that the tests would not be performed within or by the Ministry of Justice, in order to ensure a disinterested testing regime. Furthermore, the tests would include elements of creative thinking as a complement to computing power; this would help model real-world, less predictable risks.

Three stages were involved.

1. Within the Ministry of Justice, statisticians aggregated the data into ranges for each characteristic. Within the reoffending data, the characteristics were gender, age, offence, establishment/trust, previous offences, whether reoffended and number of reoffences. The ranges were widened in order to reduce the number of unique cases (with the reoffending data, this was down to a few dozen out of over 200,000 cases).

2. The resulting reoffending dataset was passed to the academics, including the authors and Chris Skinner of the London School of Economics. Postgraduate students with relevant experience from the LSE, Royal Holloway and Southampton were recruited (at short notice, with little or no preparation) to try to identify anonymised subjects, or disclose further information about them. Guidance was issued in the form of a link to a technical paper about jigsaw identification, but the students were told they could use any method that they chose, and could ignore the methods in the guidance paper if they so wished. The students were required to sign confidentiality agreements.

There was disclosure – not in one of the unique cases – as a result of matching the profile of an offender named on a local news website. Although the offender was not unique, because there was a match certain deductions about his/her reoffending data could be made.

The students' work had shown the possibility of disclosure, although not necessarily in high volumes. Certain details had proved important for the jigsaw effect, and so the MoJ statisticians further aggregated the data (removing information about the offence committed), and also removed all of the remaining unique cases.

3. This newly-adjusted dataset was then passed to the data security specialists Detica, where a team led by Philip Whittall was able to determine that the finally released data were secure against this threat, and also to further specify the risk profile of the release.

With regard to the sentencing data the approach was different, in accord with the different risk profile. The overwhelming majority of the details to be released were given in open court so disclosure or identification was less of a risk. However, it was absolutely essential to avoid (i) identification of any victim, and (ii) breach of a reporting restriction. The data were aggregated and anonymised according to risk – in particular grouping violent and sexual offences together to minimise the chance of a victim of a sex offence being identifiable. The test for the students with respect to this data was to see if they could identify someone where the details of the case were not already available on the Web or in newspapers. No student managed to do this.

Conclusions

The testing method, involving three diverse sources of testers, resulted in a greater understanding of the anonymised data and its properties. Groupthink was avoided, with the combination of MoJ statisticians, professional IT consultants and the students. The costs of the exercise were relatively low.

The involvement of the students was important for three reasons. First, they brought some unorthodox methods into the tests. Second, their involvement mimicked that of 'hackers' in the real world, lacking professionalism and experience in large-scale data handling, but driven by the nature of the problem. Third, the short time given to them set a fairly high barrier, which at least one of them was able to cross. This indicated that the dataset in its original form contained risks which warranted detailed investigation.

The exercise resulted in the MoJ statisticians having a much greater understanding of the properties, and the potential break points, of the data. They were able to make finer judgments about the level of aggregation consistent with preserving anonymity. This will be incorporated into the Privacy Impact Assessment covering these data releases. It will also of course feed into future data releases, which will allow the statisticians to try to balance data utility with the risk of jigsaw identification.

As O'Hara wrote in his report, the transparency programme must retain public confidence if it is to succeed. If the public feels that its privacy will be compromised by transparency, then the programme itself will be at risk. This series of tests of MoJ data has demonstrated:

i. That privacy and transparency are compatible
ii. That methods for dealing with anonymised datasets can go beyond what Ohm has characterised as 'anonymise-and-forget'
iii. The value of precautionary testing for understanding the weak points of an anonymised dataset
iv. The value of creative thinking as a complement to computing power in this field

v. The respect for privacy amongst the officials tasked with delivering transparency

The tests here were devised and carried out with very short notice. However, there is clearly a case for adopting similar precautionary testing procedures across government.