(Longer version of Report)

Data Linkage and Anonymisation (4 July-21 December 2016)

Background and Aims

Data about people play an increasingly important role in research in the social and health sciences, as well as in government and commerce. This programme originated from discussions between the Institute and the Economic and Social Research Council (ESRC) about how mathematical sciences research could benefit the social sciences. The twin themes of linkage and anonymization were identified as urgent methodological challenges. Linking databases can considerably enrich data sources; protecting confidentiality and privacy is often critical to data access and requires new approaches to data anonymisation. A subsequent exercise identified common data challenges in the health sciences, government and commerce and broadened the scope of the programme accordingly. The potential for connections between the two themes was adopted as a distinctive aim of the programme. A key way in which the programme sought to stimulate methodological development was by promoting exploration of the underlying theory of these techniques and bridging the associated disciplinary contexts, particularly between statistics and computer science.

Programme outline

The programme kicked off with an opening workshop bridging the two main themes and bringing together leading researchers in both fields to identify the frontiers of research. Two further workshops focused separately on the linkage and anonymization/privacy areas. Cross-disciplinary discussion was promoted by including speakers from both statistics and computer science in all three workshops and, for example, through 'speed-networking' sessions, organised to stimulate conversations between disciplines and welcomed particularly by junior participants. A fourth workshop (co-sponsored by the ESRC Centre for Microdata Methods and Practice) addressed the interface between economics and computer science regarding privacy. The programme themes fall within the field of 'data science' and the programme also hosted a well-attended Women in Data Science event.

In both thematic areas, the practitioner community involved in the application of techniques is larger than the community of researchers undertaking mathematical sciences research. Three very well-received Open for Business days were organised, supported by the Turing Gateway to Mathematics, to promote interchange between these two communities. These benefitted from collaboration with the UK Anonymisation Network (UKAN), which represents the practitioner community.

Scientific Outcomes

The programme stimulated both theoretical and applied developments. At the theoretical level, the interdisciplinary setting, rich with research challenges, opened up many new lines of work. One participant reported that “this has been the most stimulating environment I have ever been in. The ideas of this programme will keep my group busy for one or two years. I consider this programme as one of the best academic projects I have seen so far.” One example in a field that crosses the two themes was in privacy-preserving record linkage where new kinds of cryptanalysis attacks on Bloom filter encoding were developed. Another example was the development of new ways to assess re-identification risk in anonymised data using developments in record linkage theory. Many of these lines of research arose from new collaborations between participants initiated on the programme.

In terms of applied developments, the programme was planned with relevant needs of major UK scientific research groups, such as the ESRC Administrative Data Research Network, in mind and with significant participation of members of such groups. Many of the applied outcomes of the programme were orientated to the needs of such constituencies. For example, plans for a dataset repository for use in practice as a ‘testbed’ for evaluating data linkage techniques were developed.

Regarding applications of anonymization methods, three particular lines of developments emerged. One was to explore the application of ideas of differential privacy to a government statistics setting. The programme benefitted from participation by John Abowd, Chief Scientist at the US Census Bureau, who is seeking to apply such ideas in their work, and by the Rothschild Distinguished Visiting Fellow, Cynthia Dwork, who delivered a lecture alongside the National Statistician, John Pullinger. He was open to a suggestion that opportunities to apply some of the methods discussed in a government statistics setting be explored. The idea of a joint discussion/workshop with the Office for National Statistics later in 2017 is being explored. A second area of potential follow-up is in synthetic data methods, for which a challenge-based collaboration was piloted during the programme and it is hoped that a formal open challenge will be issued later in 2017. A third line of development related to the constituency of practitioners for which UKAN has supported the development of a decision-making framework for implementing anonymization techniques in practice. Ways of expanding this framework were discussed on the programme and ideas for cooperation with other organisations internationally were developed.