THE UNIVERSITY OF MANCHESTER

PARTICULARS OF APPOINTMENT

FACULTY OF HUMANITIES

SCHOOL OF SOCIAL SCIENCES

CATHIE MARSH INSTITUTE FOR SOCIAL RESEARCH

RESEARCH ASSOCIATE FOR DIGITAL TRUST AND SECURITY

VACANCY REF: HUM-016025

Salary:	Grade 6 £32,816 to £40,322 per annum (according to relevant experience)
Hours:	1 FTE
Duration:	Fixed term, 3 Years between 01 January 2021 and 31 January 2024 dependent on start date.
Location:	Oxford Road, Manchester

Enquiries about the vacancy, shortlisting and interviews:
Manager: Name: Professor Mark Elliot
Email: mark.elliot@manchester.ac.uk
Or
Name: Professor Emma Barrett
Email: emma.barrett@manchester.ac.uk

We are looking for an outstanding Research Associate to be part of our exciting and fast-moving plans to develop our capability and capacity in Digital Trust and Security, as part of University of Manchester's ambitious Digital Futures programme.

The Digital Trust and Security (DTS) theme brings together expertise across the University on issues such as privacy, data protection, regulation, and governance; responsible innovation; human-centred cyber security in the workplace; cyber-dependent and cyber-enabled crimes, criminals, and victims; and cryptography, software verification, and secure hardware. We work with businesses, local and national government, law enforcement, third sector, and academic researchers from across the world to develop research that helps protect citizens and enhance prosperity.

The Research Associate will each be based in the *Privacy, Trust, and Data Protection* cluster, working closely with the cluster lead (Professor Mark Elliot), the strategic lead for DTS (Professor Emma Barrett), and other members of Digital Trust and Security theme. The successful candidate will work on one or more projects defined by Professor Elliot. Four potential projects are outlined at the end of this document.

To join us, you will need a relevant PhD or equivalent qualification. You'll bring to the role a relevant publication record commensurate with your career stage. You will demonstrate initiative and enthusiasm to develop a coherent vision for your project, contribute to the DTS theme more broadly, and have the ability to bring new concepts and ideas to extend intellectual understanding. The post is full-time, available immediately, and funded for three years. There is the possibility of extension, depending on success in obtaining funding, and post holders would be eligible for promotion to Research Fellow (Grade 7) on successful completion of probation.

BACKGROUND

The University of Manchester

The University of Manchester, formed in 2004 by bringing together The Victoria University of Manchester and UMIST, is Britain's first chartered university of the 21st century. With some of the highest quality teaching and research and the broadest spread of academic subjects, the University's vision for the future is its development as an international research powerhouse and a favoured destination for the best students, teachers, researchers, and scholars in the world. It is already the largest single-site higher education institution in the country, offering students a greater choice of degree programmes and options, and even better facilities and support services.

Faculty of Humanities

With a total income of over £130M pa, some 15,000 students, and some 900 academic staff, the Faculty of Humanities, at the University of Manchester, is equivalent to a medium-sized university in the UK. The Faculty encompasses academic areas as diverse as Arts, Education, Law, Development, Social Sciences, and Business and Management, and is addressing, with notable success, the aim to generate fresh synergies by overcoming traditional institutional barriers between Arts and Social Sciences. The Faculty's structure, scale, and academic range permit it to promote interdisciplinary research collaboration across its schools and with the other Faculties in the University, from a strong disciplinary base.

School of Social Sciences

The School of Social Sciences (SoSS) was set up in September 2004 within the Faculty of Humanities of the new University of Manchester. It is the second-largest School in the Faculty with around 200 academic staff. It embraces six discipline areas including Sociology, Social Anthropology, Politics, Philosophy, Economics, and Social Statistics.

The School has a highly developed research culture as demonstrated by its performance in the 2008 RAE where a joint Sociology/Social Statistics submission, including CCSR, was rated top in the UK and Economics re-entered the top 10 units in the country. Aggregating the results across the six discipline areas, the School of Social Sciences in Manchester is one of the top three (alongside Oxford and the LSE) in the UK. It is a leading international research school in the Social Sciences with aspirations to enhance its standing even more in the future. The School's international and national reputations in its constituent disciplines are reflected in the substantial external research income that it generates its involvement in ESRC Centres and the presence of an ESRC Doctoral Training Centre. The School embraces the full range of quantitative and qualitative research methods.

The School of Social Sciences is committed to research-led teaching. It has an extensive portfolio of undergraduate teaching programmes: single and joint honours programmes in its constituent disciplines; joint programmes with other Schools in the Faculty of Humanities; and large interdisciplinary programmes. At Masters Level, there are currently over 300 students following in-house programmes. There are over 150 PhD students registered in the School.

The Cathie Marsh Institute for Social Research (CMI)

The Cathie Marsh Institute for Social Research (CMI) is directed by Professor Martin Everett. CMI is an internationally renowned centre of research excellence specialising in the application of advanced quantitative methods in an interdisciplinary social science context. CMI is co-located with the Social Statistics discipline area, headed by Professor Tarani Chandola, and which was launched in January 2009. It is committed to high-quality research, innovative teaching methods, and collaboration with other disciplines within the University, to improve the methodological rigour and range of quantitative enquiries in social science. The research programme of CMI and Social Statistics is orientated around the themes of advanced quantitative methods, inequalities, and social dynamics.

Key Responsibilities, Accountabilities or Duties:

This research role will involve the following set of activities:

(a) Working on and contributing to the development of one or more of the research projects listed below;
(b) Participation in the development of grant applications;
(c) Developing collaborations with other DTS clusters, and between DTS and other Digital Futures themes;
(d) Publishing in high-quality journals and international conferences;
(e) Presenting at academic and non-academic events;
(f) Engaging stakeholders in government, law enforcement, industry, and broader academia in the UK and overseas;
(g) Participating in cluster team meetings, and in broader DTS meetings;
(h) Presenting at institute and group seminars aimed at sharing research outcomes and building interdisciplinary collaboration within and outside the department;
(i) Maintaining and continuing own professional development;
(j) Following and promoting University of Manchester's policies, including Equal Opportunities;
(k) Maintaining an awareness and observation of fire, and health and safety regulations;
(l) Carrying out any other duties commensurate with the grade and purpose of the post.

This job description reflects the present requirements of the post and, as duties and responsibilities change/develop, the job description will be reviewed and be subject to amendment in consultation with the post-holder.

PERSON SPECIFICATION

Essential Knowledge, Skills and Experience*:*

- A PhD (or equivalent) in a relevant topic and strong relevant methodological expertise;

- Comfortable working with scholars across disciplines;
- A track record of publication commensurate with career stage;
- The ability to contribute to developing grant applications;
- A demonstrable understanding of at least one or more project topics;
- Comfortable working with minimal supervision;
- Project management skills, including ability to organise resources and deliver results to deadlines;
- The ability to communicate effectively in writing and in person, in both academic and non-academic contexts;
- Strong interpersonal skills and the ability to develop constructive and productive relationships with a wide range of stakeholders, including from academia, business, government, law enforcement, and the third sector.

Desirable Knowledge, Skills, Experience and Qualifications:

- Experience directly relevant to one or more of the project topics;
- Track record of applying for funding.

THE POTENTIAL PROJECTS

The Research Associate will focus on one or two projects. The choice will be based on the selected candidate's specific skills and interests.

Project 1: Measuring the trade-off between data utility and disclosure risk in synthetic population data

UoM Collaborators: Richard Allmendinger (AMBS)
External Collaborators: ONS, Replica (Canada)
Background

Synthetic data are artificial data, which have been generated to represent real data but contain no data that directly corresponds to real population units. The main function is to allow the sharing of data, which for confidentiality reasons would not be possible for the real data (usually referred to as the original data). The idea of synthetic data was first introduced by Rubin (1993), who proposed treating each observed data point as if it were missing and imputing it conditional on the other observed data points to produce a fully synthetic dataset.
Description

One of the key issues with synthetic data is utility. Data users are understandably sceptical – if the data are not real, how can we trust them? On the other hand, measuring utility is difficult – how can the data owner know what the user can do with the data? A variety of approaches have been tried (see Taub et al 2017 for a discussion).
One of the possible uses of synthetic data is with censuses. Census data cannot be released in full – the confidentiality risk is too high. Therefore, the traditional approach is to release small samples typically between 1% and 5%. There is an intriguing but untested possibility that the synthetic version of the full census might have better utility at lower confidential risk than samples of the real data. A generalisation of this is to consider at what level sampling is required to achieve

the same level of accuracy of estimates as a full population synthetic dataset. The potential here is for synthetic datasets to be labelled with “sample equivalent utility: x.y%”, which would give analysts confidence in the level of data quality.

An MSc project last year piloted this concept using a teaching dataset to simulate a population; Smith (2019). The results were promising. This project will extend this by bringing into the mix the trade-off of (a battery of) utility measures with disclosure risk – using a variant of the differential correct attribution probability (see Taub et al 2018) extended to allow comparisons between synthetic and non-synthetic data.

Deliverables

1) A test of the accuracy of a set of estimates from the synthetic data and samples of the original datasets;
2) A composite utility measure comparing the samples with synthetic datasets;
3) A composite measure of risk comparing the samples with synthetic datasets.

References

Raab, G., Nowok, B. and Dibben, C. (2016). Practical data synthesis for large samples. Journal of Privacy and confidentiality. Available at: http://arxiv.org/abs/1409.0217 [Accessed 15 Mar. 2017].

Rubin, D. B.: Statistical Disclosure Limitation. Journal of Official Statistics, 9(2), 461-468. (1993)

Smith, C. (2019). Measuring the trade-off between data utility and disclosure risk in synthetic population data. Dissertation submitted as part of an MSc in data science at the University of Manchester, September 2019.

Taub, J. Elliot, M. and Sakshaug, J. (2017) A Study of the Impact of Synthetic Data Generation Techniques on Data Utility using the 1991 UK Samples of Anonymised Records. UNECE Work Session on Statistical Confidentiality. Skopje, September 2017.https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/4_utility_paper.pdf

Taub, J., Elliot, M., Pampaka, M., & Smith, D. (2018). Differential correct attribution probability for synthetic data: An Exploration. In International Conference on Privacy in Statistical Databases (pp. 122-137). Springer, Cham.

Project 2: Exploring the Use of Machine Learning For Synthetic Data Generation

UoM Collaborators: Richard Allmendinger (AMBS)
External Collaborators: ONS, Replica (Canada)

Background

Synthetic data are artificial data, which have been generated to represent real data but contain no data that directly corresponds to real population units. The main function is to allow the sharing of data which for confidentiality reasons would not be possible for the real data (usually referred to as the original data). The idea of synthetic data was first introduced by Rubin (1993), who proposed treating each observed data point as if it were missing and imputing it conditional on the other observed data points to produce a fully synthetic dataset.

Orthodox data synthesis involves building a statistical model of the original data and then using that model to generate the synthetic version (see for example Little 1993, Rubin 1993, Dreschler and Reiter 2011, Raab et al 2016). More recently machine-learning approaches have been applied to the problem (see for example Dreschler 2010, Chen et al 2017, 2018 and Park et al 2018).

Description

Much of the work on data synthesis concerns relatively simply structured cross-sectional census and survey datasets. Obviously, the range of data in the data environment now extends much beyond these simple datasets: for example, longitudinal and hierarchical datasets, social network data, text and trajectory data. Work on synthesising such data is relatively undeveloped. The project will examine the potential for using machine learning data synthesis algorithms with datasets other than those in orthodox formats.

Deliverables

1) A review of the state of art in data synthesis with a focus on machine learning methods.
2) An application of machine learning approaches to synthesising one or two alternative data forms.
3) An evaluation of the suitability of the method (for the data).

References

Chen, Y., Elliot, M., & Smith, D. (2018). The Application of Genetic Algorithms to Data Synthesis: A Comparison of Three Crossover Methods. In International Conference on Privacy in Statistical Databases (pp. 160-171). Springer, Cham.
Chen, Y., Elliot, M., & Sakshaug, J. Genetic Algorithms in Matrix Representation and Its Application in Synthetic Data. Paper presented to UNECE work session on statistical data confidentiality. Skopje September 2017. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/2_Genetic_algorithms.pdf
Drechsler, J. and Reiter, J.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic data. Computational Statistics and Data Analysis, 55, pp.3232-3243. (2011)
Drechsler, J. (2010). Using Support Vector Machines for Generating Synthetic Datasets. Lecture Notes in Computer Science, 6344, pp.148-161.
Little, R. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9 (2), pp.407-426.
Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. Proceedings of the VLDB Endowment, 11(10), 1071-1083. DOI: https://doi.org/10.14778/3231751.3231757
Raab, G., Nowok, B. and Dibben, C. (2016). Practical data synthesis for large samples. Journal of Privacy and confidentiality. Available at: http://arxiv.org/abs/1409.0217 [Accessed 15 Mar. 2017].
Rubin, D. B.: (1993). Statistical Disclosure Limitation. Journal of Official Statistics, 9(2), 461-468.

Project 3: The Special Uniques Problem

External Collaborator: Ann-Sophie Charest (Laval University, Canada)

Background

The notion that disclosure risk within microdata is not distributed evenly across the file has been understood since the seminal work of Bethlehem et al (1990).

In 1998, Elliot et al - observing non-monotonicity of the standard risk measure of the level of population uniqueness given sample uniqueness with respect of geographical granularity -

proposed the notion of *special uniqueness*. They observed that a sample unique on a given set of variables *X* that remained unique despite geographical aggregation had a higher probability of being a population unique (on *X*) than a sample unique which did not retain its uniqueness over the same aggregation. They coined the phrase *special uniques* for the former and *random uniques* for the latter.

In subsequent work, Elliot et al (2002) found that the property held for aggregations across any variable and that the higher the degree of aggregation the larger the effect.

Description

More generally, consider a dataset *D* that is a sample of population units cross-classified on at least two variables. Employ the following notation:

- *X* is a set of key variables
- *Y* is an arbitrary subset of *X*
- For a given data unit let *x* be its values for *X* and *y* be its values for *Y*
    - *f(x)* = number of data units in *D* with *X=x*
    - *f(y)* = number of data units in *D* with *Y=y*
- A data unit is called special unique with respect of *X* if for any *Y*, *f(x)=f(y)=1*.

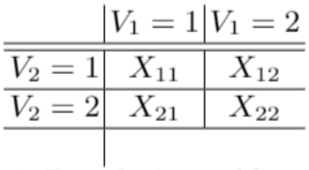
This in turn led to the definition of the *minimal sample uniques (*MSUs), one which contains no unique subsets. A special unique can contain any number of MSUs.

Elliot et al (2002) show in numerical studies of census data that, for a given data unit, as the size of the MSUs decreases and/or the number of MSUs increases, the probability of that data unit being population unique increases.

It follows from this observation that it should be possible to obtain a metric that collates - for a given data unit and set of key variables - the size and number of MSUs and through this to calculate the risk for each data unit. Implementing this observation in a principled way has proved difficult. The method that Elliot et al implemented into SUDA (the Special Uniques Detection Algorithm), which they term the "SUDA score", is based on the principle that the key variable values form a lattice structure. The SUDA score exploits the fact that the smaller an MSU is, the larger the number of paths through the lattice that are unique, and therefore the larger the proportion of the lattice that is unique. In essence, the SUDA score is the sum of the number of paths through the lattice that are unique. Experiments show respectable correlations between this score and the underlying risk measure 1/F(x) – the reciprocal of the frequency x in the population. Elliot et al combined the SUDA score with a file-level measure described in Skinner and Elliot (2002), to obtain a pseudo-probabilistic measure of data snooper confidence in a given match. But again this approach is heuristic, and like many such techniques, it is impossible to completely specify (with any precision) the conditions under which it will break down. However, Elliot et al were able to demonstrate sufficient robustness that the technique was used with 2001 UK census microdata.

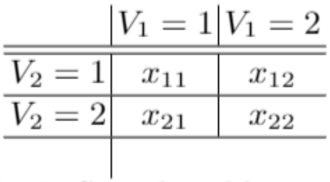
Since then, however, little progress has been made and critically we still lack a principled statistical underpinning to the special uniques proposition - that special uniques are more likely to be population uniques than random uniques. The approach we adopt here is to strip the concept down to a minimalist framework of two by two tables of counts.

Consider Tables 1 and 2. Table 1 is a population table of counts and Table 2 is a sample that is drawn without replacement from Table 1.



	$V_1 = 1$	$V_1 = 2$
$V_2 = 1$	X_{11}	X_{12}
$V_2 = 2$	X_{21}	X_{22}

Table 1. Population table notation



	$V_1 = 1$	$V_1 = 2$
$V_2 = 1$	x_{11}	x_{12}
$V_2 = 2$	x_{21}	x_{22}

Table 2. Sample table notation

What has been observed empirically is that special uniques tend to be population uniques more often than random uniques. If this is true, then we would expect:

$P(X_{11}=1 \mid x_{11} = 1, x_{12} = 0) > P(X_{11}=1 \mid x_{11} = 1, x_{12} > 0)$ [1]

From this, key research questions are:

1. Does equation 1 always hold?
2. If not, then under what conditions does it hold?

Pilot work by Charest and Elliot (2020) indicates that it does not always hold and is dependent on the data generating process for the population data. This needs further investigation. In particular we need to (i) develop a more robust mathematical representation of the effect; (ii) extend out the above work to dimensionality above 2; and, (iii) establish a more principled way of tying the special uniques property to the underlying risk measure 1/Fj.

References

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of Microdata. Journal of the American Statistical Association, 85(409), pp 38-45.

Charest, A-S., and Elliot M. J. (2020) Special Uniques: A simulation study, *Proceedings of Privacy in Statistical Databases;* paper 48 September 2020.

Elliot, M. J., Skinner, C. J., and Dale, A. (1998). Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. Research in Official Statistics, 1, pp 53-67.

Elliot, M., Manning, A. and Ford, R. (2002). A Computational Algorithm for Handling the Special Uniques Problem. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems 10(5), pp.493-509.

Skinner, C. J., and Elliot, M. J. (2002). A measure of disclosure risk for microdata. Journal of the Royal Statistical Society: Series B (statistical methodology), 64(4), pp 855-867.

Project 4: Automated Data Environment Analysis

UoM Collaborators: Robert Stevens (CS), Goran Nenadic (CS), John Keane (CS)
Potential External Collaborators: ICO, ONS

Background

The knowledge economy needs data, datasets, and data processes (streaming, sharing, and linking). But the dissemination and use of data carry privacy and confidentiality risks through the inadvertent disclosure of personal data. Measuring that risk has, to date, been based on uncertain guesstimates about the data environment. In the face of such uncertainty data stewardship organisations tend to be cautious and so the quality and quantity of data made available for re-use are constrained.

Through research and service provision, Manchester is the world leader in Data Environment Analysis (see for example Elliot et al 2011, Mackey and Elliot 2013). The methods of data environment analysis provide the means for properly grounded disclosure risk analyses. If this were scaled and automated, it would enable organisations to directly identify where the risks were located in their intended data and to be confident in their data dissemination process. This would square the circle of privacy protection in an era of open data. It is this that the Augmented Data Environment Analysis System (ADEAS) project aims to achieve.

Description

Vital to understanding the data environment is mapping the data available to a data intruder (an agent wishing to identify individuals within a released anonymised dataset). Manchester has developed a method for doing this but this is resource-intensive and cannot be done at scale. The proposed ADEAS will solve this problem by inferring at scale the data contained in external databases by deriving metadata from data collection instruments and thereby building a representation of the data environment. ADEAS combines web crawling, text mining, and ontology generation techniques to populate a *Key Variable Mapping System* (see Smith and Elliot 2014), which is already developed but in need of software hardening to generate key variable specifications that can then be used by disclosure control practitioners to make grounded risk assessments based on a realistic representation of the data environment.

Thus, ADEAS will provide a critical component of the infrastructure necessary to deliver initiatives such as the transparency agenda and the digital economy. Data holding organisations are hungry for software that can deliver the functionality that ADEAS promises. Forming a community of those will be is a second thread of the proposal, which will also build on the UK Anonymisation Network (www.ukanon.net) lead by Manchester and in particular the Anonymisation Decision Making Framework (Elliot et al 2016, 2020). The community formation is as important as the delivery of the software itself as part of the point of ADEAS is to reduce the expertise gap so that more disclosure risk analysis can be done by DSO's in-house rather than relying on external expertise that is in short supply.

Deliverables

ADEAS is an ambitious undertaking and the goal here would be to develop a proof of concept, which could be developed into a large grant proposal.

References

Elliot, M., Lomax, S., Mackey, E., & Purdam, K. (2010, September). Data environment analysis and the key variable mapping system. In *International Conference on Privacy in Statistical Databases* (pp. 138-147). Springer, Berlin, Heidelberg.

Elliot, M., Mackey, E., O'Hara, K., & Tudor, C. (2016). *The anonymisation decision-making framework*. Manchester: UKAN.

Mackey, E., & Elliot, M. (2013). Understanding the data environment. *XRDS: Crossroads, The ACM Magazine for Students*, *20*(1), 36-39.

Smith, D., & Elliot, M. (2014). A graph-based approach to key variable mapping. *Journal of Privacy and Confidentiality*, *6*(2).