[ENGLISH BELOW]

SecHuman II Tandemprojekt in Zusammenarbeit mit GESIS

Schutz vor Re-Identifizierung bei der Verlinkung von Daten

PIs: Prof. Dr. Maike Buchin und Prof. Dr. Estrid Sørensen, Praxispartnerin: Dr. Katharina Kinder-Kurlanda, Gesis.

Heutzutage werden immer größere Mengen von digitalen Daten erzeugt. Um aktuelle Datenschutzstandards – z. B. die DSGVO – einzuhalten, wird ein Teil dieser Daten typischerweise de-identifiziert zugänglich gemacht. Die wachsenden Datenmengen, die umfassenderen Dateninhalte sowie die verbesserten Methoden zur Verlinkung von Daten haben auch zu einem erhöhten Wert von Daten (Big Data) geführt. Zunehmend werden durch die Verlinkung verschiedener Datensätze neue, noch aussagekräftigere Datensätze erstellt. Häufig handelt es sich dabei dann um vertrauliche Daten. Zum Beispiel steigt bei der Verlinkung von Umfragedaten und geographisch detaillierten Raumdaten die Wahrscheinlichkeit, dass einzelne Befragte re-identifiziert werden können, auch wenn in beiden Ursprungs-Datensätzen bereits Anonymisierungsmaßnahmen zur Anwendung kamen. Bei der Verlinkung von Datensätzen entsteht somit ein Risiko der Re-Identifizierung von Individuen. In einem Paper, das für viel Debatte gesorgt hat, haben Luc Rocher und Kolleg*innen (2019) kürzlich postuliert, dass durch die Verwendung von 15 demographische Merkmale 99.98% der US-Amerikaner in jeglichem Datensatz korrekt re-identifiziert werden können. Diese Zahl von Merkmalen entsteht leicht bei der Verlinkung von Datensätzen. Rocher et al. schlussfolgern, dass es dabei unwahrscheinlich ist, dass Datensätze die Forderungen der DSGVO und ähnlicher Datenschutz-Standards nach Anonymisierung erfüllen. Dies erweist sowohl eine rechtliche wie auch eine technische Herausforderung existierender De-Identifikations und release-and-forget-Modelle.

Das Re-Identifizierungsrisiko gilt insbesondere bei der Verlinkung von Daten, das heißt, wenn verschiedene Datensätze, wie Raumdaten, Zeitreihen, sowie Internetdaten (z. B. Social Media), miteinander verknüpft werden. In diesem Projekt sollen die Methoden zur De- und Re-Identifizierung von Datensätzen untersucht werden sowie auch das Risiko der De- und Re-Identifizierung von Datensätzen bewertet werden, die aus einer Verlinkung entstanden sind. Dabei soll einerseits das Risiko der Re-Identifizierung mit statistischen Methoden ermittelt werden und andererseits sollen verschiedene Methoden und Verfahren zur De- und Re-Identifizierung sowohl mathematisch-informatisch wie auch sozialanthropologisch betrachtet und bewertet werden. Das Secure Data Center des GESIS Leibniz-Instituts für Sozialwissenschaften, fungiert in diesem Tandem als Praxispartner. Vor allem soll mit dem Team „Data Linking & Data Security“ zusammengearbeitet werden. Das Secure Data Center bietet Zugang zu Forschungsdaten, die aus Datenschutzgründen besonderen Zugangsbeschränkungen unterliegen, und berät zur Forschung mit sensitiven Forschungsdaten. Die Verlinkung von Forschungsdaten besonders mit Social Media-Daten ist ein Schwerpunkt der Arbeiten.

Im mathematischen-informatischen Teil des Tandems soll untersucht werden, wie sich die Verknüpfung von Daten auf die Re- und De-identifikation auswirkt, und welche Methoden hier zum Einsatz kommen können. Ebenfalls soll untersucht werden, wie die Ergebnisse der Verknüpfung Nutzenden zur Verfügung gestellt werden können. Insbesondere interessieren wir uns für die Verknüpfung von Daten, durch die eine Bewegung einer Person in Raum und Zeit herleitbar und damit ihre Identifikation möglich ist. Wir betrachten Möglichkeiten, um für solche Daten bestehende Methoden der De-Identifikation zu erweitern und damit sicherer zu machen. Dazu betrachten wir insbesondere, welche räumlichen sowie attributbezogenen Vergröberungen und Manipulationen zu

verknüpfender bzw. verknüpfter Daten datenschutzrechtlich unbedenklich sind. Ebenfalls betrachten wir Konzepte zur Bereitstellung aggregierter oder synthetischer Daten. Das Ziel der De-Identifikation ist die Bereitstellung der Daten und wir untersuchen daher verschiedene Möglichkeiten und deren Wert für Nutzende der Daten.

Im sozialanthropologischen Teil des Projekts werden einerseits die Bedeutung von wissenschaftlichen Datenpraktiken sowie der Konfiguration der soziomateriellen Infrastruktur am Daten-Zugriffspunkt für den Schutz vor Re-Identifizierung untersucht. Andererseits soll beobachtet und analysiert werden, wie sozialwissenschaftliche Datenpraktiken und dadurch auch epistemologische Praktiken der Sozialwissenschaft sich transformieren durch die Risikoeinschätzungen der Re-Identifizierung sowie ihren Schutz. Durch teilnehmende Beobachtungen, Interviews und Dokumentenanalyse soll die Bedeutung von Datenpraktiken für den Schutz vor Re-Identizierung untersucht werden. Dabei stehen die Begrenzung der Datennutzung auf kontrollierte, wissenschaftliche Kreise, sogenannte „trusted communities“, im Fokus der Untersuchung, wie auch die formellen und informellen Normen, Belohnungs- und Kontrollsysteme der Wissenschaft. Darüber hinaus soll erforscht werden, wie epistemische Praktiken durch veränderte Datenpraktiken beeinflusst werden, z. B. ob Änderungen in Fragestellungen und Themen sowie auch in der Kategorisierung und Theoretisierung von untersuchten Phänomenen beobachtet werden können.

Eine enge Zusammenarbeit der zwei Tandemprojekte wird erwartet, indem beide an Datensätzen, Verfahren zum Schutz vor Re-Identifizierung sowie Datenpraktiken des GESIS Secure Data Center forschen werden. Einerseits werden Verfahren zum Schutz vor Re-Identifizierung, die durch das mathematisch-informatische Projekt entwickelt werden in Datensätzen implementiert, an denen Sozialwissenschaftler*innen arbeiten und ihren Datenpraktiken anpassen. Diese sind Gegenstand der sozialanthropologischen Forschung. Andererseits weisen die Erkenntnisse des sozialanthropologischen Projekts über wissenschaftliche Datenpraktiken auf die Arten von Verlinkungen hin, die besonders schutzbedürftig sind. Die Zusammenarbeit der zwei Tandemprojekte erhöht sowohl die Breite wie auch die Relevanz der Ergebnisse zum Schutz vor Re-Identifizierung bei der Verlinkung von Datensätzen.

Für Fragen zur sozialanthropologischen Teil steht Estrid Sørensen (estrid.sorensen@rub.de) sehr gerne zur Verfügung. Fragen zum mathematisch-informatischen Teil können an Maike Buchin gerichtet werden (maike.buchin@rub.de).

Bitte beachten, dass Bewerbungen so lange in Erwägung gezogen werden, bis die geeigneten Kandidat*innen gefunden sind. Projektbeginn soll voraussichtlich zum 1. April 2021 stattfinden.

SecHuman II tandem project in cooperation with GESIS

Protection against re-identification in data linkage

PIs: Prof. Dr. Maike Buchin and Prof. Dr. Estrid Sørensen, practice partner: Dr. Katharina Kinder-Kurlanda, GESIS.

Increasingly larger amounts of digital data are produced today. In order to comply with current data protection standards - e.g. the GDPR - data are often made available anonymously. The growing amounts of data, the more extensive data content and the improved methods of linking data have also increased the value of data (big data). By linking different data sets, new and more valuable data sets are created. Often private or confidential data are used. However, when data sets are linked – such as survey data and specified geo-location data – the probability increases that individual

informants can be re-identified, even if both data sets are anonymised. In other words, a risk of re-identification of individuals arise in the linking of data sets. An article by Luc Rocher and colleagues (2019) stirred a lot of debate when recently showing how by the use of 15 demographic characteristics 99.98% of US-Americans can be correctly re-identified in any – also anonymised – data set. This number of demographic characteristics are often available when linking data sets. Rocher et al. conclude that it is unlikely that data sets will comply with the requirements of the GDPR and similar data protection standards for anonymization. This points to both legal and technical challenges for existing de-identification and release-and-forget models.

The risk of re-identification arises particularly when data are linked, i.e. when data sets of localisation data, temporal data and internet data (e.g. social media) are brought together. In this project, the methods for de- and re-identification of data sets will be investigated and the risk of de- and re-identification of data sets that have arisen from their linking will be evaluated. On the one hand, the risk of re-identification should be identified using statistical methods and, on the other both mathematic, computer-science and social-anthropological methods and procedures for de- and re-identification will be assessed and evaluated. The Secure Data Center at the GESIS Leibniz Institute for Social Sciences functions as a practical partner in this tandem, particularly the team of "Data Linking & Data Security". The Secure Data Center offers access to research data, which are subject to special access restrictions for data protection reasons, and advices on research with sensitive research data. Linking research data, especially with social media data, is an emphasis of the work.

The task of the *mathematical-computer-science* based part of the tandem is to investigate how the linking of data sets affects re-identification and de-identification just as the methods applied for data re-identification are studied. An area of interest is furthermore how the consequences of re-identification can be made available to users. We are particularly interested in cases in which the combining of data enables the detection of individuals' movements in space and time and thus the identification of the person. We investigate how existing methods for de-identification of such data can be expanded and thus how to protect data. For that purpose, it is of interest how the granularity and manipulation of spatial data and their attributes can be done in compliance with data protection law. We furthermore evaluate concepts for the provision of aggregated or synthetic data. The goal of de-identification is to make data available and accordingly we consider different methods for de-identification and their value for data users.

The focus of the *social anthropological part* of the project is how socio-material aspects contribute to protecting data against re-identification. Of interest are both scientific data practices and the configurations of the socio-material infrastructures at the data access point and their implication for data re-identification. The task will be to study and analyse how social science data practices – including their epistemological dimensions – transform through the risk assessment of both re-identification and of protection against re-identification. The effect of practices of re-identification protection will be investigated by way of participant observations, interviews and document analysis in and of the GESIS secure data centre. Of focal interest is the effects of the restriction of data sharing with controlled, scientific circles – i.e. so-called "trusted communities" – as well as the formal and informal norms and reward- and control systems of science: in what ways do the production of social trust relations, review processes and other control mechanisms of the sciences affect data practices and particularly the protection against data re-identification? On the other hand, it is also to be investigated how the change in data practices for increased data protection affect the epistemic practices of the sciences: whether changes can be observed in the scientific questions posed, the phenomena investigated and the categories and approaches applied in scientific research, as an effect of data re-identification risk assessments.

The two tandem projects will collaborate closely, as the object of study of both projects will be data sets of the GESIS Secure Data Centre as well as their data practices and methods for protection against re-identification. On the one hand, the mathematical-informatics project will develop procedures for protection against re-identification, which will be implemented in social science data sets. Social scientists will thus have to adjust their data practices to the altered procedures. These data practices will be the focus of the social anthropological research. On the other hand, the findings of the social anthropology project on how scientists use their structures and invent data practices that support the protection against data re-identification indicate what types of data combinations that are particularly in need of protection and should thus be attended to by the mathematical-informatics tandem part. The cooperation between the two tandem projects will increase both the breadth and the relevance of the results to protect against re-identification when linking data sets.

For questions to the socio-anthropological part, please do not hesitate to contact Estrid Sørensen (estrid.sorensen@rub.de). Questions concerning the math/computer-science tandem can be directed to Maike Buchin (maike.buchin@rub.de).

Please notice, the deadline for application will be extended until the right candidates are found. The project is planned to start in April 2021.