TRANSACTIONS ON DATA PRIVACY 17 (2024) 123–146

When to use the k-rule? - Criteria for managing uniqueness and de-anonymization risk in social science survey data

Anja Perry*, **Wolfgang Zenk-Möltgen***

*GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6–8, 50667 Cologne, Germany.

E-mail: anja.perry@gesis.org, wolfgang.zenk-moeltgen@gesis.org

Received 17 July 2023; received in revised form 1 March 2024; accepted 28 July 2024

Abstract. Anonymization plays a large role for data sharing in the social sciences, where research subjects are often human. In this paper we are looking at k-anonymization, an anonymization strategy rarely used in the social sciences. This is due to high-dimensional socio-economic information necessary for social science research. Here, the k-rule is often too rigid and leads to information loss. We argue, however, that certain datasets need to be k-anonymized and suggest criteria to determine the need for this rule. We then apply our criteria to example datasets from a social science data archive. In doing so, we provide criteria for data curators to determine which level of anonymization to apply to data at hand and hands-on examples on how to apply them. We aim to improve workflows for data archives and support safe data sharing practices.

Keywords. Data confidentiality, k-anonymity, Quasi-identifier, Anonymization, Social science

1 Introduction

While open science and data sharing become increasingly important [9], ethical and legal considerations require anonymization to protect respondents' privacy [12]. This is especially the case for social science research where research subjects are often human. Anonymization is necessary to protect respondents from harm potentially caused by information revealed in the survey and to create trust between the respondents and the researchers and/or interviewers so that respondents are willing to respond truthfully to questions about sensitive information. A dataset can be viewed as not being anonymous when a unique sample case can be matched to a unique case in the underlying population [29].

When anonymizing data, it is often not sufficient to eliminate direct identifiers such as names, contact details, and IP addresses. Also indirect identifiers (sometimes called quasi-identifiers or key attributes) need to be taken into account. Indirect identifiers are "persistent demographic characteristics of people that might be used to discover their identities" (p. 2) [31]. One strategy, k-anonymity, may help to protect respondents of certain surveys [27]. A dataset is k-anonymous "for $k > 1$ if, for each combination of key attributes, at least k records exist in the data set sharing that combination" (p. 991) [6]. This also often leads to information loss [1] as information in the dataset needs to be coarsened into much broader

categories than when using alternative strategies. An alternative strategy often used is to coarsen single variables, e.g., country of origin, to make the underlying population larger but not necessarily to the stage of having k equal records in the sample [18]. k-anonymity goes beyond this and often demands that more variables, especially in combination with each other, need to be deleted or coarsened.

While research on anonymization strategies and on k-anonymity exists, recommendations on when to apply the k-rule in social science research are very rare. Social science data often contains highly detailed sociodemographic data and losing this information greatly reduces the analytical potential of the data [1, 7]. With ever more detailed data used in social science research, such as longitudinal data, network analyses, digital trace data, and big data linked to social science surveys [4], researchers and data curators should be prepared to apply stricter anonymization to protect respondents. Elliot, Mackey, and O'Hara [8] provide a general and very extensive guide on risk assessment and anonymization decision-making. To determine whether k-anonymity is necessary, the authors suggest a thorough risk assessment and, once deemed risky, further extensive steps that include establishing intruder scenarios, analytical approaches to estimate the risks, and penetration tests. In this paper we want to extend the Anonymization Decision-making Framework (ADF) by Elliot et al. [8] by looking at datasets considered at risk. In doing so, we simplify the process of strategy choice once risky data is identified and make this step more suitable for data curators in social science data archives.

Data curators at social science data archives receive multiple datasets each week to check and approve for publication. Besides making anonymization decisions, they perform several additional steps, such as checking for understandability, vetting the data and correcting mistakes, and adding further metadata. We suggest a simplified process for determining strict anonymization measures, such as the k-rule, by identifying criteria that help identify high risk data faster and more efficiently. We also apply these criteria to five datasets stored in the GESIS data archive. The GESIS data archive is only one example of a social science data hosting institution and our approach tested here can be applied to other data archive and data hosting institutions. While we cannot offer a strict rule, which is in line with Elliot et al.'s [8] probability/risk approach and the Five Safes approach [5], we highlight dataset criteria that are easy to observe and place them into the risk component framework by Müller, Blien, and Wirth [24], the risk assessment matrix by Elliot et al. [8] and the Five Safes approach by Desai et al. [5]. The paper aims to help curators identify risks that demand stricter anonymization measures more easily and improve anonymization processes.

The paper is structured as follows: We first present data risk assessment and typical anonymization strategies applied in social science data archives and k-anonymity as an alternative strategy. Then we present the concept of uniqueness and derive three criteria that determine when uniqueness in the data is problematic and when the k-rule needs to be applied. These criteria are applied to five datasets stored in the GESIS data archive. Finally, we discuss the findings, implications, and limitations of our work.

2 Risk assessment and anonymization in social science data archives

Data curators at social science data archives decide about data anonymization steps daily based on the nature of the data and the informed consent given by the respondents [16].[1] In doing so, they rely on two important frameworks: the ADF by Elliot et al. [8] and the Five Safes framework by Desai et al. [5]. The ADF is a guide through the evaluation of the data situation including the evaluation of the data environment, i.e., in which settings the data will be used, stakeholders' needs, e.g., those of researchers and data subjects, and the data themselves, i.e., their sensitivity [8]. Similarly, the Five Safes Framework regards five dimensions that interplay with each other [5]: safe people, safe projects, safe settings, safe data, and safe outputs. While assessing these aspects, one needs to consider the trade-off between data utility and maximum anonymization. Tied to this trade-off is the fact that "zero risk is not a realistic possibility" [8] (p.15) but rather a balance between protecting survey respondents and data utility (realistic risk principle) [8]. Consequently, the measures that data archives put in place need to be proportional to the data's risk (proportionality principle) [8] and to the interplay between data, data users, and the data environment [5]. Elliot et al. [8] therefore speak of "data that have been through an anonymization process" (p. 10) or, in shortened form, are "functionally anonymized".

Typically, data curators have evaluated and have solutions in place for most aspects of both frameworks. These are the stakeholders, e.g., maintaining the archive's reputation for data safety, having routines in place to evaluate informed consent, and the data environment and settings, e.g., having different access options in place ranging from open access to secure data enclaves.[2] Only the research data coming in is different every time and needs to be carefully evaluated during each ingest process. If the data pose a risk of re-identification, appropriate data protection measures need to be applied.

Data archives use different strategies to functionally anonymize data [18]. First, they eliminate variables from the data. This is often done with geographical information, which – if it remains in the data set – reduces the underlying population size and thereby increases the risk of re-identification combined with other information contained in the data set. They also top- or bottom-code variable values to hide outliers. These could be exceptionally high income, very high age, or a large number of children in countries where large families are rare. A third strategy is to aggregate categories of variable values such as age, profession, or country of origin [18, 22]. Each of the strategies mentioned enlarge the underlying population in matrix cells built by quasi-identifiers so that unique sample cases are not unique in the underlying population.

These strategies are insufficient for some datasets. These are often data with locally restricted geographical coverage or highly sensitive data, for example regarding respondents' health or sexuality. Also, the sample itself may be very visible, for example celebrities such as actors or politicians. If the sampling is based on sensitive characteristics, like identifying as LGBTQ+, respondents need to be especially protected. For high risk datasets, Elliot et al. [8] suggest stricter anonymization measures, such as k-anonymization or adding noise to the data by subsampling or swapping [18]. In addition, data can be put under restrictive access, such as strict contractual agreement for off-site data use or secure access in a safe room or a virtual data enclave [18][5]. Restrictive access, especially on-site data use, is

[1] This paper focuses only on the data themselves and not on legal restrictions imposed by, for example, incomplete or missing informed consent.

[2] These evaluations may need to be renewed in certain cases, e.g., for entirely new data types.

expensive for both the data infrastructure and the data user, and for some datasets this provision is disproportionate, for example when data re-use is expected to be low.

This paper will focus on k-anonymity – a method used to greatly reduce the risk of re-identification, but one rarely applied in social science research. There may, however, be datasets for which k-anonymization is useful in order to make data available to the public and avoid costly access restriction. We define k-anonymity in the next section and present its advantages and disadvantages.

3 k-anonymity

k-anonymity was first proposed by Samarati and Sweeney [27]. It comes from rule-based output statistical disclosure control and specifies that "a table may only be released if there are at least [k] observations for each cell" (p. 7) [26]. It is a concept that is often used for tables of counts such as censuses but can also be applied to microdata. When publishing medical trials in the medical field, patients' demographics are often aggregated in a way that no unique individual remains in the data [7]. Applying the k-rule to microdata means that strategies such as aggregating or top- and bottom-coding are applied in such a way that each record in the dataset is identical to at least $k-1$ other records in the values of a defined set of identifying variables [4][27]. Therefore, a specific individual cannot be uniquely identified. There are always $k-1$ individuals in the dataset that share the same combination of characteristics. In this paper we use the terms k-anonymization and k-rule synonymously.

k-anonymization leads to a maximum probability for respondents of being re-identified of $1/k$ (there may be other respondents with a combination of quasi-identifiers that have more than k individuals in their group). That is, in a 2-anonymous dataset, there will be at least two persons sharing the same quasi-identifiers. If all quasi-identifiers are known to an adversary, the probability of choosing the right person is 50:50. This maximum probability of re-identifying an individual is also defined as threshold risk [7]. The minimum threshold risk is suggested to be $k=3$; most typical is a k of 5 and it is rarely higher than that [31].

The k-rule represents a clear rule and is intuitive to understand. But k-anonymity also has some weaknesses: The k data twins in a k-anonymized dataset may share the same values in another substantive variable and thereby reveal sensitive information or preferences without identifying the respective individual. For example, every one of the k individuals indicate that they vote for the same political party, thereby revealing their shared political preferences. To avoid this, one can apply extensions to k-anonymity, such as l-diversity, p-sensitivity, or t-closeness [4, 6]. l-diversity means that "for each group of records sharing a combination of key attributes, there are at least l 'well-represented' values for each confidential attribute" (p. 991) [6]. p-sensitivity demands "at least p different values for each confidential attribute" (p. 991) [6] for the k individuals sharing the same key attributes. If both extensions are applied, no sensitive information will be revealed about the group of k individuals but they may offer an approximation, for example when most of the k individuals vote for one party compared to a much lower probability found in the population. Finally, t-closeness also takes the distribution of sensitive characteristics into account and defines that both distributions (among the k individuals and the population) should not differ more than a threshold t. The three extensions are not easy to implement, have their own shortcomings [6], and cause even further data restrictions [6, 8]. In this paper, we will not look further into those extensions.

Another major downside of the k-anonymity concept is that it is quite rigid. Aggarwal

[1] as well as El Emam and Dankar [7] point out that for high-dimensional data – data including a lot of sociodemographic information – the k-rule is not suitable. High-dimensional sociodemographics are usually found in social science survey data where they play an important role for data analyses. A high number of indirect identifiers and highly differentiated categories naturally lead to very small k, very often a k of 1. So, applying the k-rule means that this information needs to be coarsened tremendously, which greatly limits the analytical potential of social science data. At the same time, surveys usually have a sampling method that reduces the risk of re-identification. It is therefore not necessary to achieve k-anonymity in most cases. Thompson and Sullivan [31] and Elliot et al. [8] state that especially when drawing from a large population, sampling is a powerful tool for anonymizing data.

The k-rule is therefore only used sparsely for social science survey data. In this paper we argue that there are some cases that demand the k-rule and aim to derive dataset characteristics that help data curators decide when to apply this rule.

4 The uniqueness problem

Detailed sociodemographic information will lead to sample-unique individuals, i.e., respondents that do not share the same combination of characteristics with other respondents. One needs to distinguish between sample uniqueness and *population uniqueness* [2, 24]. Having a sample-unique person in the dataset is often easy to detect and not critical per se. It becomes critical when this person's characteristics are unique in the population as well. A person is identified when a sample-unique person is also population unique. The more quasi-identifiers that exist in a dataset and the more detailed they are, the more likely there will be population uniqueness, increasing the risk of a person being identified. For example, detailed sociodemographics such as geographical region, gender, profession, and marital status may identify a person when these characteristics are unique in this specific region. But also the underlying sub-population is decisive when determining population uniqueness. This could be a large number of children in Germany, where families with many children are rare, or an immigrant's country of origin. One person from Turkey in a German sample is not critical as this person represents the largest immigrant group in Germany with about 1.46 million people. But a sample-unique person originally from a country with very few immigrants in Germany is more likely to also be a population-unique person when few further characteristics are considered.

Identifying population uniqueness in survey data is a very complex and difficult task. Furthermore, a sample-unique person may be population unique but knowledge about it may be imperfect when one cannot rule out that another person with the same characteristics exists in the population. Müller et al. [24] therefore speak of *confidence of population uniqueness* rather than population uniqueness and define it as the probability that a person does not share the same combination of key variables with any other person in the population. To determine confidence of uniqueness, curators would need to look at each sample-unique case, which, depending on the dataset, can be many. To prove that these sample-unique cases are or are not population unique, they'd need to consider all possible sources for matching. This cannot be clearly defined as it depends on unknown factors such as intruder motivation, techniques a potential intruder may apply, and unknown data sources that a potential intruder has or will have in the future [8]. Hence, instead of determining population uniqueness, curators rather look at distributions in the underlying population as an indicator for potential population uniqueness, and reduce information

where subpopulations are small.

To determine whether the k-rule needs to be applied, we will now look at factors that increase confidence of population uniqueness for sample-unique cases. Building on previous work on identification risk by Marsh et al. [21], Müller et al. [24] provide a framework that incorporates the two mentioned components of identification risk – population uniqueness and confidence of population uniqueness – with two further factors: inclusion probability and compatibility.

4.1 Inclusion probability

Inclusion probability is the probability that an arbitrary individual to be identified is included in the dataset. By inclusion probability we refer to the term *representativeness*, used by Müller et al. [24]. Inclusion probability is high in full population data, such as a census. There, all individual units are surveyed and unique cases in the data are also population unique and directly identify a person [8]. Survey data are generally less risky compared to census or population data. Typically, survey data are collected from a sample of the population and there is usually no public knowledge about who was sampled and then who finally took part in the survey. Hence, sampling can already be a reliable protection measure.

In addition to pure coverage, biases also play a role in inclusion probability. The selection process of building a sample is usually a random selection process. If the sample is completely random, the probability that a person is in the sample is n/N (n = number of subjects in the sample, N = number of subjects in the population). When a sample is skewed, some parts of the population will have a higher probability of being sampled [24]. But there are also other techniques for selection, such as selecting persons from a specific region (e.g., when interviewers must travel to this region) or selecting by quota of some characteristic (e.g., to be able to get meaningful results for small subpopulations). If the criteria for selection or quota are known, the sample list can be re-created, thus involving a higher risk of identification for the respondents. Inclusion probability, and hence the de-identification risk, is increased when small populations are surveyed with a lot of information available about this population, for example when employees in a small firm are asked to take a survey. In this case, to have analytical value, the sample may cover a relatively large part of the population. Having a relatively large part of the population covered makes it less likely that a person with the same characteristics as the sample-unique person exists outside of the sample. The risk of identification can also be high when the sample is publicly known, for example a full sample of a known subpopulation.[3]

The aggregating and top coding of variables are anonymization measures that are applied to highly representative data [21]. By aggregating geographic information, one reduces the number of unique combinations but at the same time makes a sample-unique person less representative of the category they then stand for. For example, a person in the original dataset may have represented only a handful of people from a certain country currently living in Germany, but they would stand for a large number of people from that region or continent in the functionally anonymized version of the dataset. The goal for this strategy must be to aggregate groups in such a way that the underlying population is large enough to make population-unique cases unlikely. Another way to decrease identifiability is to draw a subsample of the realized sample before making data available to the public [21].

[3] This can also happen accidentally, when a school publishes information about which of their classes took part in mandatory competence testing. The information regarding which class was sampled is usually not public information.

There are two incidents that may make using sampling as a protection obsolete. These are participation knowledge and the presence of a snooper [28], a person purposly trying to re-identify respondents with access to the registry used for sampling, i.e., an employment or a membership database.

4.2 Compatibility

When trying to identify respondents, available information about a person must be *compatible* with the identifiers in the dataset. To be compatible, information in the data must be collected using the same definitions and during the same point in time [21]. Definitions may vary for certain classifications, such as professions or social status. Geographic information is usually highly compatible when standardized regions are used for reporting, such as NUTS regions [15]. They may not be precisely defined when respondents in a certain landscape (for example, the Bavarian Alps) are surveyed. Information can also change over time, and long field periods and delayed data publication can make the data less compatible to outside information [8, 21]. Finally, Marsh et al. [21] and Elliot et al. [8] also mention errors in data collection and misclassifications that work as an advantage towards "natural data protection" (p. 50) [8]. While these mistakes do still occur, newer survey technologies and automized data processing have reduced these problems over the years.

Re-identification is somewhat certain only if the available information exactly matches. One very effective anonymization strategy is therefore to aggregate or delete highly compatible information such as geographical information. But small and locally restricted surveys – for example surveys in specific city parts or in small cities as well as surveys of employees at all universities in a small state – have a higher risk as it is not feasible to delete the geographical information without drastically lowering the dataset's utility.

Very visible populations are also at risk of being identified as population unique due to compatibility. Skinner et al. [29] call them "figures in the public eye" and name examples such as a person from an unusual ethnic group aged 101, a chief of police with nine children and a PhD, and members of certain occupational categories such as politicians, actors, musicians. Information about figures in the public eye are locally or publicly available and often in a detailed way. The authors also argue that this issue is very complex as one cannot know which additional information might be taken into account and which information becomes available in the future [8].

Panel data are also critical regarding re-identification risk. Not only may further waves add more information about a specific individual [8], they may also render certain anonymization measures obsolete. This is the case when individuals move from one age group to another between waves, making the age variable more compatible to information about the individual's exact age [23]. In a data archive, it is often unclear whether further waves may follow due to a project extension. The subsequent waves may be published in the same archive or elsewhere with fewer anonymization measures.

The uniqueness problem as well as the two factors of inclusion probability and compatibility are hence relevant in social science surveys. This is, however, only one aspect in risk assessment [8]. In the next section we will examine data sensitivity as the second dimension of the risk assessment matrix.

5 Intersection with sensitivity

Elliot et al. (2020) discuss three aspects of *sensitivity*. For consent and expectation sensitivity we assume that data curators have routines in place to assess them. For our analysis here, we focus on data sensitivity which refers to the data's properties, such as sensitive topics or vulnerable, at-risk populations. The GDPR protects all personal data and in addition gives some indication in Art. 9 for special categories of data as "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, [...] genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation" [12]. This information can either identify a person directly, such as genetic or biometric data, or information according to Art. 9 of the GDPR is or has in the past been a reason for discrimination and persecution. Information according to Art. 9 of the GDPR almost always plays a role in social science research, for example when respondents are asked about political opinions or attitudes towards societal and political topics. But even information which does not fall under Art. 9 of the GDPR may be considered sensitive information by the person affected. This may be information regarding whether a test was passed or not, a personal opinion in an employee survey, or criminal actions committed or experienced by the respondent. Furthermore, data can also be sensitive because of the data subjects involved, for example a survey with immigrants or with members of the LGBTQ+ community. Hence, what is considered sensitive information differs from person to person.

Sensitivity plays an important role in data risk assessment [8] as it determines the harm a data breach can cause and the motivation for an intruder to identify a person in the data. It is only interesting to target a specific person or a dataset when additional, sensitive information about a person identified can be revealed. Marsh et al. [21] defines the probability of re-identification as $pr(identification) = pr(identification|attempt)pr(attempt)$ (p. 318) where sensitivity affects $pr(attempt)$, the probability that an intruder attempts to identify a person.[4] The more sensitive the data revealed by a data breach is, the greater the harm to the identified person.

Elliot et al. [8] combine the summary risk with the data's sensitivity in a risk assessment matrix. Summary risk takes into account the content of the data but also its environment, the people handling the data, etc. [5]. In this paper we define summary risk as the risk of re-identification, i.e., identifying a population-unique person in the data. We use this definition because most data in social science data archives are disseminated openly, and data users and their environment are often out of the archive's control. Access restrictions in form of access categories and secure data enclaves are, however, tools for handling the re-identification risk and we will refer to this trade-off when applying our suggested criteria to actual use cases.

Both dimensions combined determine the severity of a data breach. Table 1 shows that a negligible risk of re-identification and low data sensitivity may be acceptable and control processes unnecessary. In this case, an intruder is less motivated to re-identify respondents as the high effort of identifying does not match the information that can potentially be revealed. On the other hand, when re-identification risk is high, data must always be protected, even when sensitivity is considered low. This is in line with sensitivity being subjective and re-identifying a person (i.e., confirming that a sample-unique person is also

[4] Another scenario, also described by Marsh et al. [21] is that an intruder wants to discredit a data holding institution, irrespective of which data is revealed about an individual. In this case, one can argue that, by considering data sensitivity during risk assessment, the data holder still sufficiently protects data subjects .

Table 1: Data risk assessment [8]

		Data Situation Sensitivity		
		low	medium	high
Summary Risk	high	essential	essential	essential
	medium	borderline	essential	essential
	negligible	unnecessary	borderline	borderline

population unique) must be avoided.

To find characteristics that demand the k-rule, we follow the same approach and look at components of re-identification (inclusion probability and compatibility) [24] and sensitivity [8] as three dimensions to be equally considered.

6 Criteria to determine the necessity of the k-rule

Following Elliot et al. [8], we consider sensitivity jointly with the data's re-identification risk, in this case inclusion probability and compatibility [24]. As laid out above, Elliot et al. [8] offer an extensive risk assessment to determine whether protection measures need to be put in place. The recommendation for which anonymization strategy to apply to data at high risk remains vague without pointing to specific measures for specific data situations. Rather, the authors suggest further risk analyses and penetrations tests, a lengthy procedure that data curators are unable to do on a day-to-day basis. Instead, as described in section "Risk assessment and anonymization in social science data archives", data curators often rely on well-established measures. These measures are the result of their long-term experience of checking and publishing data while protecting respondents in the data. There are, however, cases of datasets that demand anonymization strategies that go beyond those already mentioned. In these cases, the curators may know that the risk level demands stricter anonymization but they still find it hard to decide whether their usual strategies are sufficient or the stricter k-rule needs to be applied.

With k-anonymity being a very rigid rule leading to information loss, it should only be applied to social science survey data when other strategies cannot reduce the re-identification risk or the data's sensitivity, i.e., attractiveness for re-identification and harm caused. By extending the ADF framework by Elliot et al. [8] and going one step past the initial risk assessment laid out there, we want to give data curators a tool to help make the decision between applying less rigid anonymization strategies or k-anonymization. We cannot offer a clear rule for applying k-anonymity but can propose a rule of thumb: the k-rule should be applied in the riskiest scenario, i.e. when all three dimensions – inclusion probability, compatibility, and data sensitivity – are high (Table 2). However, the application of the k-rule should still be considered carefully and measures to reduce either of the three dimensions in order to avoid the k-rule must be favored. This principle-based approach, rather than a rule-based approach, is very much in line with the target audience of this paper: experienced data curators. As Desai et al. [5] point out, a principle-based rule fits best

Table 2: Risk assessment matrix based on Elliot et al. [8] to determine k-anonymity necessity

		Data Sensitivity: low	Data Sensitivity: high
Inclusion Probability *and* Compatibility	high		k-rule
	low		

when those applying the rule are experienced and when context needs to be considered in decision making rather than deciding purely according to a defined schema. We highly recommend using our approach as an indicator and not as a strict rule. This can leave room for other, less strict anonymization measures, if applicable, combined with a careful consideration along the spectrum between access and risk [5].

7 Criteria applied to five datasets

In this section we apply the above derived criteria to five datasets currently stored in the GESIS data archive. These datasets and the GESIS data archive only serve as examples, representative for other social science data archives and their data. We choose five examples to demonstrate how data content and sample characteristics impact the criteria we suggest and the respondents' re-identification risk and therefore the need for protective measures. We choose one dataset that covers the general population; this serves as a baseline comparison to the other examples as we expect that only few anonymization measures are needed for it. The remaining four datasets have higher risk factors because they can potentially include figures in the public eye, cover a restricted geographic area, cover a sample with publicly available compatible information for re-identification, or cover a vulnerable group. The five datasets have different access levels which we report in the data description section. According to the ADF and the Five Safes Framework, the access level also impacts the necessary protective measures. Both are in a trade-off between less anonymization but restrictive environment and more anonymization and less restrictive environment [5][8].

7.1 General population survey

7.1.1 Dataset description

We chose the Eurobarometer 92.3 dataset [11] as an example for a comparative social science survey with nationwide samples. The Eurobarometer survey series is conducted by the European Commission as a means of monitoring the public opinion of the population in European member and candidate countries. All Eurobarometer studies are publicly available. The survey contains repeated questions from previous waves of the Eurobarometer series as well as questions on special topics. As is often the case in social science surveys, the questionnaire contains many sociodemographic questions; postal code, sample point number, and interviewer number have been suppressed from the published dataset (cf. the dataset documentation).

Table 3: Risk assessment matrix based on Elliot et al. [8] applied to "Eurobarometer 92.3" inclusion probability

		Data Sensitivity	
		low	high
Inclusion Probability	high		
	low		

This Eurobarometer 92.3 dataset [11] contains data from 28 member states and five candidate countries and was collected between November and December of 2019. Overall, there are 32,543 respondents included in the dataset. The sampling was drawn in a multi-stage random procedure, first selecting primary sampling units and then selecting addresses by a random route procedure. The response rates for the EU member countries in the Eurobarometer 92.3 differ greatly from low rates at 19.9% (Germany) or 20.1% (Luxembourg) to high rates at 72.5% (Slovakia) or 78.0% (The Netherlands).

An intruder scenario, i.e., an attack to re-identify a respondent in this survey, is unlikely. The dataset covers the whole population with a comparatively low number of respondents, meaning that potential data twins outside the surveyed group are very likely, hence an attempt of re-identifying a person will probably not be successful.

7.1.2 Risk evaluation

Sensitivity: The Eurobarometer asks about country of origin and about political attitudes of the respondents. Both types of information can be used to discriminate against respondents and are therefore considered highly sensitive according to Art. 9 of the GDPR. The GDPR requires strict safeguarding measures for sensitive data, resulting in carefully chosen anonymization strategies when data is to be shared with the scientific community.

Inclusion probability: We consider inclusion probability of the data to be low (Table 3). Only a small proportion of the EU population was asked to participate so that one respondent in the final dataset stands for 15,800 EU citizens (an equivalent of 0.0063%). In the Netherlands, the country with the highest response rate, one respondent stands for 14,332 citizens (0.0070%) [10]. This makes it most difficult to assess whether a sample-unique person in the data is also population unique as it is impossible to rule out that a person with the same characteristics exists in the remaining population but was not surveyed. Hence, here sampling works very well as a protective measure against re-identification.

Compatibility: Generally, the variables in the Eurobarometer 92.3 are highly aggregated and do not offer much compatibility to outside information – especially marital status and household composition variables do not allow conclusions about easily observable characteristics, like having young children. Regional information like community size is highly aggregated as well. Some countries, however, publish detailed regions at NUTS 2 and 3 levels [15]. This geographical information is highly compatible and – when combined with information from other variables, even as broad as gender and marital status – leads

Table 4: Risk assessment matrix based on Elliot et al. [8] applied to "Eurobarometer 92.3" compatibility

		Data Sensitivity	
		low	high
Compatibility	high		
	low		

to sample uniqueness that may become critical when matched with outside information to determine population uniqueness (Table 4).[5]

7.1.3 Conclusion

Due to its low inclusion probability, it is not necessary to apply the k-rule. Rather, the k-rule would greatly reduce the analytic potential of the variables already aggregated in the dataset. However, some countries publish very detailed regional information at NUTS 2 and 3 levels [15] which may make it possible to re-identify individuals when combined with other variables in the dataset. This level of detail could be reduced by aggregating regional information.

7.2 Dataset with political party members

7.2.1 Dataset description

The dataset "German Party Membership Study 2017" [20] observes political party members in Germany and is available only for researchers. When downloading the dataset, users must state their name and purpose of use. The data's purpose is the social reporting on German party members and a comparison over time.[6] The survey was conducted in 2017 among members of the parties which were represented in the German Bundestag.[7] Random samples were drawn via the party head offices based on membership lists. The total sample comprised 17,000 persons (p. 84) [19], the dataset contains 9,748 cases, which is a response rate of 57.3%. Before publishing the dataset, the data creators already applied anonymization measures: Answers to open questions were depersonalized and variables containing information about respondents' mandates on national, federal, and European level as well as membership of a national or federal party executive board were deleted.

One intruder scenario is that information about uniqueness may be revealed accidently if the dataset contains "figures in the public eye" [29]. Information about figures in the public eye is publicly available – for example through interviews – and can be very detailed; it is therefore compatible to information in the dataset. If the dataset includes such a person or a respondent enters the public sphere at a later time, an intruder may potentially be able

[5] For example, the Statistical Office of the Slovak Republic [30] publishes detailed data on gender, age, marital status, and region.

[6] The study is a continuation of similar cross-section studies from 1998 and 2009.

[7] In addition, the Free Democratic Party (FDP) was included as it was expected to re-enter the German Bundestag [19].

Table 5: Risk assessment matrix based on Elliot et al. [8] applied to "German Party Membership Study 2017" inclusion probability

		Data Sensitivity	
		low	high
Inclusion Probability	high		
	low		

to link information to the dataset to reveal their identity along with sensitive information provided during data collection. A public figure does not necessarily need to be famous at a national or subnational level, i.e., a federal or EU politician – a person may be locally well known, such as a mayor, and that increases the likelihood that local knowledge can be applied to identify persons in the dataset.

7.2.2 Risk evaluation

Sensitivity: In addition to demographics and form of involvement in party activities, the survey also asks about political viewpoints, voting decisions, and political engagement. This information is highly sensitive and protected under Art. 9 of the GDPR. In addition, the data may reveal political viewpoints that can potentially cause resentment within that person's party or local party chapter.

Inclusion probability: The group surveyed here is significantly smaller than the general population and defined by a public attribute, i.e., membership in one of Germany's popular parties. We consider inclusion probability of the data low (Table 5), given that a random sample of 3000 members per party was drawn.[8] The sample representation ranges between 0.6% of the members of Social Democratic Party of Germany (SPD) and 4.8% of the members of The Left (Die Linke); for numbers regarding the overall party members see Niedermayer [25]. This representation is even lowered by the fact that not all members in the sample participated in the survey.

Compatibility: The variables in the dataset were already reduced to lower the overall compatibility of the dataset. Information about mandates on national, federal, and European levels as well as membership of a national or federal party executive board was deleted, so that highly visible party members could not be identified. The sample in this survey is, however, problematic when it comes to potential re-identification. Skinner et al. [29] draw attention to "figures in the public eye" who need not necessarily be known nationwide, but can be known on a local level. Especially party members who are politically active on the community level are well known in their local area, with information such as their profession, family members, and engagement in organizations being local knowledge. While regional information is not available in the dataset, an unusual combination of characteristics in the data may lead to an individual that is known locally to have these characteristics. An

[8] Only 2000 members of the Christian Social Democrats were drawn, a party only active in Bavaria.

Table 6: Risk assessment matrix based on Elliot et al. [8] applied to "German Party Membership Study 2017" compatibility

		Data Sensitivity	
		low	high
Compatibility	high		
	low		

additional challenge is that sample-unique individuals may become nationwide "figures in the public eye" in the future. As an example, the sample contains a sample-unique female member of The Left coming from a Latin American country. The information on origin countries of party members is not generally available. But a politician of the same age rising in popularity may mention in a future interview that she is the only woman in her party from this specific country. This person could retrospectively be matched with the information in the dataset. The same may happen with further information about sample-unique public figures being revealed in past or future interviews which can be matched with information from the survey. This scenario becomes more likely if we think of local "figures in the public eye" and further information that is very likely unique (Table 6).

7.2.3 Conclusion

Due to its low inclusion probability, the k-rule should not be applied. However, the nature of the sample and the potential of containing current or future "figures in the public eye" [29] makes this dataset risky for re-identification. Currently, the dataset's availability is restricted, which is an appropriate measure to protect the dataset due to its potential compatibility to public knowledge. If this data's access restriction were to be waived in the future, the data should then be anonymized further, for example by aggregating highly compatible and sensitive information about the respondents' origin.

7.3 Data on a geographically restricted area

7.3.1 Dataset description

The dataset "Cologne Dwelling Panel" [3] samples dwellings (rather than individuals or households) in the two Cologne areas of Deutz and Mülheim to study gentrification in these neighborhoods. The dataset contains four waves collected between 2010 and 2016. A sample was randomly drawn from dwellings in the two neighborhoods. With every new wave, the same dwelling was surveyed again, irrespective of the individual(s) living there. 1009 dwellings were surveyed in wave one and 747 dwellings in wave four. Access to this data is restricted: Researchers who want to reuse the data must register with the GESIS data archive and state the purpose of their reuse.

A possible intruder may be a neighbor with very detailed local knowledge and knowledge about certain basic information about their neighbors that help to identify a respondent in the dataset. Re-identification can potentially reveal very sensitive information, such as same-sex relationships or opinions and attitudes towards the neighborhood and the city.

Table 7: Risk assessment matrix based on Elliot et al. [8] applied to "Cologne Dwelling Panel" inclusion probability

		Data Sensitivity	
		low	high
Inclusion Probability	high		
	low		

7.3.2 Risk evaluation

Sensitivity: The data may make sensitive information observable to neighbors – information that they previously did not have. For example, if one can identify their neighbor in the dataset and this person is living with someone of the same sex, the data can identify whether they are in a same-sex relationship rather than roommates. This makes the data highly sensitive (Art. 9, GDPR) as they may provide information for discrimination. Also, the data contain information provided by the interviewer rather than by the respondents themselves, such as a dirty or poor state of the dwelling.

Inclusion probability: Wave one of the Cologne Dwelling Panel [3] sampled 1009 dwellings in two Cologne neighborhoods. Both neighborhoods combined have 57,794 inhabitants, resulting in an inclusion probability of 1.75% on the individual level. Wave four with 747 interviews yields an inclusion probability of 1.29%. While both percentages do not sound like much, they are much higher than in the general population survey described in the section above. The two neighborhoods combined can be compared to a medium-sized German city (20,000 – 100,000 inhabitants). Cities of this size are usually anonymized in general population surveys, i.e., the name of a city of that size cannot be revealed. This is because unique combinations of characteristics paired with local knowledge may identify a population-unique person. Here in this case, the metadata reveals which neighborhoods are sampled as it is important information for replicating the study and for considering important local aspects that influence gentrification. Hence, we consider inclusion probability to be high (Table 7).

Compatibility: The geographic area covered by this panel survey is very small and the information of where individuals live (neighborhood boundaries) is highly compatible with available outside knowledge. Usually, geographic information is aggregated on a much higher level to protect respondents, which is not possible in this case. Due to the local restriction, we must assume that local knowledge is available which can be matched with the data. This may be information gained through local activities, such as volunteering, engagement in clubs or local events, or through interactions in social media groups connected to the neighborhood. Some things asked in the survey may also be observable by neighbors, for example that a partner and/or children live in the same household. In addition, information available on social media should be considered in this case. The survey focuses on gentrification, a topic intensively discussed in large cities such as Cologne. This is especially true in the surveyed neighborhoods as they are highly affected by gentrifi-

Table 8: Risk assessment matrix based on Elliot et al. [8] applied to "Cologne Dwelling Panel" compatibility

		Data Sensitivity	
		low	high
Compatibility	high		
	low		

cation. This survey and participation in it may have therefore been of wide interest in these neighborhoods and may have been discussed online. Hence, other things need to be considered – response knowledge and the possibility of additional information shared by the respondent in the same social media outlet, such as family pictures, pictures of their house/apartment, etc. The panel design leads to further increased compatibility for certain information available in the data. It can indicate when a person moves in or out of a dwelling, when a new partner moves in or out, or when a baby is born. Children's age is aggregated in the data. However, when a child jumps from one category to the next over the course of the panel, the exact age may become apparent or at least becomes narrowed down to smaller age brackets. Considering these factors, compatibility with information available or observable is very high (Table 8).

7.3.3 Conclusion

This dataset is of high risk due to the inclusion probability for a population of a small, locally restricted area and the compatibility of the information included, and also due to a multi-wave design. While inclusion probability cannot be reduced, reducing compatibility may also be difficult due to detailed information coming from the panel design. In addition, information contained in the data can be highly sensitive. Hence, applying the k-rule is advised for this dataset.

At the time of writing this paper, the fourth wave of the study was under review at the GESIS data archive, and a fifth wave was being prepared by the research team. Further waves add further information to the initial waves, making the data more compatible with outside information and narrowing the time gap between collection and publication of the data. After discussing the dataset with the data curator responsible, the research team was advised to apply k-anonymity to previously published waves and all following waves.

Panels with with temporary funding are most difficult to assess for data curators. Often they can not foresee whether further panel waves will be added to the existing data later. Hence, curators must work with precaution and therefore not only check for existing reidentification risks in the data but also think of potential critical combinations that may enter the data with further waves. This means that curators do not necessarily need to identify actual population uniqueness and sensitive information connected with that individual in the data but rather think in terms of potential re-identification scenarios with further data being added.

7.4 Dataset with known sample and compatible information publicly available

7.4.1 Dataset description

The dataset "Factors influencing the data sharing behavior of researchers in sociology and political science" [33] is based on a survey that took place among researchers who published in a defined group of sociology and political science journals. The survey was conducted by one of the paper's co-authors. Data access is restricted to reuse for scientific purposes and can be requested by registering with GESIS and stating the purpose of the data usage. The data are hierarchical: the journal level and the article level nested within the journals. The study authors selected ten journals and all their issues in a defined time span and derived a list of authors. The authors' e-mail addresses of 1011 articles were collected to conduct the survey. Authors of 446 articles participated in the survey, covering 44.1% of all of the articles.

A possible intruder scenario may be that someone directly looks for a person in this known sample or tries to identify as many researchers as possible to reveal information that may damage that person's reputation as a researcher. This may be information that this researcher is not sharing data on purpose or is not citing data properly and thereby ignoring practices of good scientific conduct.

7.4.2 Risk evaluation

Sensitivity: The data include information about religion, which is considered highly sensitive information and needs to be protected (Art. 9, GDPR). Furthermore, it contains opinions about scientific conduct, such as attitude to and actual data sharing and whether respondents cite the data they use. While some of this actual behavior is also publicly available, revealing the researchers' attitude towards aspects of scientific conduct has the potential to damage their reputation.

Inclusion probability: Although the respondents' names and email addresses are removed from the original dataset, the selected journal names and publication years are published in an article [32]. As a result, the list of authors in the gross sample is publicly available and can be retrieved from the journal websites, including additional information such as the authors' university or institute affiliations. Hence, the data are a rare case for which the exact sampling procedure is publicly available and can be exactly replicated. This contrasts with other common sampling procedures where random samples are drawn from a given population and the actual sample is unknown. Not all contacted lead authors participated in the survey, reducing the re-identification risk to a certain degree. An intruder does not know with certainty who participated in the survey. But we still consider a response rate of 44.1% as being high enough to pose a risk to the respondents. Hence, we consider inclusion probability of this data as high (Table 9).

Compatibility: Information about respondents in the dataset, such as level of career, employer type, gender, and country of residence, is publicly available. Researchers often publish their CVs with detailed information on their website. Also, their employers (universities and research institutes) often publish similar information and at least the most important career stages on their websites. Both public sources also typically contain a list of publications, including those in the journals used for the sampling. Hence, compatibility

Table 9: Risk assessment matrix based on Elliot et al. [8] applied to "Factors influencing the data sharing behavior of researchers in sociology and political science" inclusion probability

		Data Sensitivity	
		low	high
Inclusion Probability	high		
	low		

Table 10: Risk assessment matrix based on Elliot et al. [8] applied to "Factors influencing the data sharing behavior of researchers in sociology and political science" compatibility

		Data sensitivity	
		low	high
Compatibility	high		
	low		

of information in the dataset and information available in public source is very high (Table 10).

7.4.3 Conclusion

When assessing the risk of this dataset, we had access to the original dataset and could analyze sample and population uniqueness. The focus here is not only on certain variables that may reveal population-unique individuals but rather the whole sample that is at risk of being re-identified. This is due to the rather high inclusion probability (the sample can be fully replicated) and the widely available, nearly fully compatible information about respondents of this special population. Hence, the high risk of re-identification cannot be resolved by altering certain groups of very detailed variables; the k-rule is advised.

Due to these risks, an earlier version of the data was withdrawn from the GESIS data archive[9] and a reduced and k-anonymized version of the data was published for secondary use. The goal was to anonymize the data so that $k >= 5$ and the following anonymization measures were applied:

- Three variables (age, level of career, and employer type) were completely dropped
- Gender (IF01) was dichotomized (male, female/other)
- Religion (IF04) was dichotomized (Christian, Muslim, Jewish, other vs. agnostic, atheist)

[9]The primary researcher still possesses the original data to comply with good scientific practices and to keep the data for 10 years [17]. This original data is not distributed through the GESIS data archive.

- Country (IF05) was dichotomized (US, non-US)

The newly published dataset now has a k of 21, meaning that for any individual in the dataset with a certain combination of characteristics, there are at least 20 "data twins" with the same combination of characteristics in the data.[10] Thus, the data could be made available for other researchers in an anonymous form but it now has much lower potential for analysis.

7.5 Dataset with a sensitive sample

7.5.1 Dataset description

The EU LGBTI II survey by the European Union Agency for Fundamental Rights [14] is a large-scale survey from 2019, surveying LGBTI persons (lesbian, gay, bisexual, trans and inter). It focuses on respondents' views and experiences, especially regarding discrimination, violence, and harassment in different areas of life, including employment, education, healthcare, housing, and other services.

The survey was conducted in EU member states, Northern Macedonia, and Serbia. The sample is a non-random availability sample to achieve a representative sample of LGBTI persons across the EU. Respondents were targeted via social media, dating apps, LGBTI civil society organizations, and more. 141,621 individuals participated in the survey, representing 0.8% of the estimated target population and 0.04% of the EU population. Data access is restricted and only accessible after signing a data use agreement. Special regulations apply for publishing results based on the data.

LGBTI persons are still discriminated against and even experience violence in some countries or in parts of society in the more progressive countries. An intruder may try to use this dataset to identify LGBTI persons to reveal their sexual orientation, gender identity, or intersex traits which may otherwise not be visible to them.

7.5.2 Risk evaluation

Sensitivity: The data are highly sensitive as the sample itself is defined by individuals' sexuality and gender identity, both being sensitive information according to Art. 9 of the GDPR. LGBTI persons are often still discriminated against; furthermore, the data contains information such as experiences of sexual assault, attacks, and harassment. Inter and trans persons may also reveal personal medical information in the survey, such as variation of sex characteristics and body interventions. Respondents also answered questions about their general health and mental health.

Inclusion probability: While each respondent in the survey represents 0.8% of the target population (LGBTI persons in the EU), they represent only 0.04% of the EU population. It is further often difficult to assess who belongs to the target population. The sampling procedure for this survey itself is based on estimates. This is because statistics regarding LGTBI persons do not exist and the group is not accessible using conventional sampling techniques [13]. We therefore consider inclusion probability to be low and comparable to other large-scale general population surveys (Table 11).

[10] Although the data creators aimed for a lower k ($k = 5$), the measures yielded a k of 21. Not implementing any one of the measures would have yielded a k lower than 5.

Table 11: Risk assessment matrix based on Elliot et al. [8] applied to "The EU LGBTI II Survey, 2019" inclusion probability

		Data Sensitivity	
		low	high
Inclusion Probability	high		
	low		

Table 12: Risk assessment matrix based on Elliot et al. [8] applied to "The EU LGBTI II Survey, 2019" compatibility

		Data Sensitivity	
		low	high
Compatibility	high		
	low		

Compatibility: Compatibility is very low. This is due to relatively little information requested about the respondents' demographics. Geographical information that is typically highly compatible with outside information is reduced to country of residence and broad categories of self-assessed city size. This restriction is a very helpful tool for anonymizing data. Further quasi-identifiers that can be observed by outsiders include household size and composition as well as religion. With the restricted geographical information, we consider it impossible to conclude that a sample-unique respondent is population unique on a broad national level (Table 12).

7.5.3 Conclusion

Despite the highly sensitive nature of the sample, we do not advise using the k-rule. Inclusion probability and compatibility are fairly low so that the current anonymization level paired with the restricted access offers sufficient protection.

8 Discussion

In the previous section we looked at five different datasets and assessed their risk to determine whether the k-rule needs to be applied. In two cases we advised using the k-rule. These are cases in which the sample is either highly geographically restricted, i.e., containing highly compatible geographic information, or the sample is known or can be exactly replicated, meaning that inclusion probability for respondents is high in the data. In both cases also the two remaining criteria are high but both particular data characteristics make it difficult to apply other, less rigid anonymization strategies. In the three cases for which

we did not recommend k-anonymity, either one or more of the three factors in focus were low or the data already was under appropriate access restrictions when one or more factors seemed to demand caution.

k-anonymity is a well-known anonymization rule that is rarely used for social science survey data. We use the risk component framework by Müller et al. [24] and the risk assessment matrix by Elliot et al. [8] to derive criteria of when to use this rigid k-rule in the social science field. These criteria are high sensitivity paired with high inclusion probability of the population surveyed and high compatibility of the information with potential outside sources for matching. When all three criteria are met, the k-rule should be applied.

This paper positions the k-rule in the social science survey research. The reason for not applying the k-rule usually lies in the highly dimensional sociodemographic data typically surveyed in social science research. This aspect often makes the k-rule unsuitable as it leads to large information loss [1, 7]. At the same time, sampling often protects survey respondents. By applying the risk component framework [24] and the risk assessment matrix [8] to examples from the GESIS data archive, we can demonstrate their practical applicability to actual use cases when determining the anonymization level for the respective datasets.

We find the criteria most useful when applied to the example datasets from the GESIS data archive. In particular, we advised using the k-rule based on our criteria for two of the five datasets. One dataset contains highly geographically restricted information, i.e., highly compatible geographic information, which cannot be deleted without losing the data's analytic potential. The other dataset used a publicly available sampling procedure, making the data highly representative for its population. Both characteristics were paired with high levels for the remaining two criteria but especially the two characteristics pointed out here make it impossible to apply other, less rigid, anonymization strategies.

Our aim is to provide data curators with criteria to determine when to use conventional anonymization strategies [18] and when these are not sufficient so that the k-rule needs to be applied. This will improve data risk assessment procedures and decision making in data archives. The k-rule is to be applied when all three suggested criteria are considered high. In the two use cases for which we recommend the k-rule, at least one factor cannot be deleted without losing the data's analytic potential. This is the case when highly representative, compatible, or sensitive aspects of the data are profound to the analytical potential.

In both cases critical information is revealed in the metadata – the city neighborhoods where data was collected or the sampling procedure. This raises the need to not only focus on the data themselves when anonymizing but also on their metadata and what they may reveal. When faced with critical data that need k-anonymization, researchers and data curators still have the option to choose between strict anonymization (and likely restricting the information remaining in the data) or restricted data access, for example through a data use agreement or in a data enclave [18].

One limiting factor is the lack of clear thresholds between low and high inclusion probability, compatibility, and sensitivity. This also only results in indications of when to consider the k-rule rather than a clear rule of when to apply it. Data creators and data curators should carefully assess the data at hand based on these three indications and rule out other anonymization measures before applying the k-rule. This approach, instead of providing clear thresholds and a clear anonymization rule, is in line with Elliot et al. [8] and their risk/probability-based approach as well as with the Five Safes approach [5]. Desai et al. [5] point to the advantages of a principle-based, rather than rule-based, approach – when decision makers are experienced and when it is advised to take context, such as access

restrictions, into account. Elliot et al. [8] state that one can only rely on probabilities of attempts and success of potential re-identification. Due to too many unknown factors this risk can never be zero. This premise is in contrast to strict thresholds and, consequently, to a strict rule that determines the application of k-anonymity. Further application of the suggested criteria to additional datasets, especially to various risky subpopulations, may make our recommendations more precise and easier to apply in data curation.

Author contributions. A.P.: Conceptualization, Writing, Methodology; W.Z-M.: Methodology, Formal Analysis

Acknowledgements

We thank Oliver Watteler for many discussions about this topic as well as Markus Cziesla and Christian Prinz for weekly discussion about data ingest at the GESIS data archive. The insights into your work have greatly helped this paper to evolve. We also thank Holger Döring for providing critical feedback and very helpful comments. Also, we are thankful for very fruitful discussions about this paper with colleagues from the department of Data Services for the Social Sciences at GESIS, at the IASSIST 2023 conference, and for very valuable feedback from two anonymous reviewers.

References

[1] C. C. Aggarwal. On k-Anonymity and the Curse of Dimensionality. In *Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005*, Trondheim, Norway, 2005.

[2] J. G. Bethlehem, W. J. Keller, and J. Pannekoek. Disclosure Control of Microdata. *Journal of the American Statistical Association*, 85(409):38–45, Mar. 1990.

[3] J. Blasius. Cologne Dwelling Panel. https://doi.org/10.7802/2035, 2020. Data retrieved from the GESIS Data Archive.

[4] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati. k-Anonymity: From Theory to Applications. *Transactions on Data Privacy*, 16:25–49, 2023.

[5] T. Desai, F. Ritchie, and R. Welpton. Five Safes: designing data access for research, 2016.

[6] J. Domingo-Ferrer and V. Torra. A Critique of k-Anonymity and Some of Its Enhancements. In *2008 Third International Conference on Availability, Reliability and Security*, pages 990–993. IEEE, Mar. 2008.

[7] K. El Emam and F. K. Dankar. Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637, Sept. 2008.

[8] M. Elliot, E. Mackey, and K. O'Hara. *The anonymisation decision-making framework 2nd Edition: European practitioners' guide*. UKAN, Manchester, 2020.

[9] European Commission. Open access & Data management - H2020 Online Manual. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm, 2017.

[10] European Commission. Standard Eurobarometer 92 – Autumn 2019 - First results. https://europa.eu/eurobarometer/api/deliverable/download/file?deliverableId=71659, 2019.

[11] European Commission. Eurobarometer 92.3 (2019): Standard Eurobarometer 92. https://doi.org/10.4232/1.13564, 2020. Data retrieved from the GESIS Data Archive.

[12] European Parliament and Council of the European Union. General Data Protection Regulation 2016/678. https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2018.

[13] European Union Agency for Fundamental Rights. *A long way to go for LGBTI equality: technical report*. Publications Office of the European Union, Luxembourg, 2020. OCLC: 1242909141.

[14] European Union Agency for Fundamental Rights. The EU LGBTI II Survey, 2019. https://doi.org/10.4232/1.13733, 2021. Data retrieved from the GESIS Data Archive.

[15] Eurostat. *Statistical regions in the European Union and partner countries NUTS and statistical regions 2021*. Publications Office of the European Union, Luxembourg, 2022 edition edition, 2022. OCLC: 1356305766.

[16] German Data Forum (RatSWD). Data Protection Guide, 2nd edition. https://doi.org/10.17620/02671.57, 2020. Publisher: German Data Forum (RatSWD) Version Number: 1.

[17] German Research Foundation (DFG). Guidelines for Safeguarding Good Research Practice. Code of Conduct. https://doi.org/10.5281/zenodo.6472827, Apr. 2022.

[18] ICPSR. Guide to social science data preparation: Best practice throughout the data life cycle. https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

[19] M. Klein, P. Becker, L. Czeczinski, Y. Lüdecke, B. Schmidt, and F. Springer. Die Sozialstruktur der deutschen Parteimitgliedschaften. Empirische Befunde der Deutschen Parteimitgliederstudien 1998, 2009 und 2017. *Zeitschrift für Parlamentsfragen*, 50(1):81–98, 2019.

[20] M. Klein, T. Spier, and C. Strünck. German Party Membership Study 2017: Party member survey. https://doi.org/10.7802/2326, 2021. Data retrieved from the GESIS Data Archive.

[21] C. Marsh, C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D. Lievesley, and N. Walford. The Case for Samples of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(2):305, 1991.

[22] G. J. Matthews and O. Harel. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5(none), Jan. 2011.

[23] R. Mitra, S. Blanchard, I. Dove, C. Tudor, and K. Spicer. Confidentiality challenges in releasing longitudinally linked data. *Transactions on Data Privacy*, 13(2):151 – 170, 2020.

[24] W. Müller, U. Blien, and H. Wirth. Identification Risks of Microdata. *Sociological Methods and Research*, 24(2), 1995.

[25] O. Niedermayer. Parteimitgliedschaften im Jahre 2017. *Zeitschrift für Parlamentsfragen*, 49(2):346–371, 2018.

[26] F. Ritchie and M. Elliot. Principles- Versus Rules-Based Output Statistical Disclosure Control In Remote Access Environments. *IASSIST Quarterly*, 39(2):5, Dec. 2015.

[27] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, Computer Science Laboratory, SRI International, 1998.

[28] R. Sarathy and K. Muralidhar. The Security of Confidential Numerical Data in Databases. *Information Systems Research*, 13(4):389–403, Dec. 2002.

[29] C. Skinner, C. Marsh, S. Openshaw, and C. Wymer. Disclosure control for census microdata. *Journal of Official Statistics*, 10(1), 1994.

[30] Statistical Office of the Slovak Republic. Datacube - Demographics and social statistics. https://datacube.statistics.sk/, 2022.

[31] K. A. Thompson and C. Sullivan. Mathematics, risk, and messy survey data. *IASSIST Quarterly*, 44(4), Dec. 2020.

[32] W. Zenk-Möltgen, E. Akdeniz, A. Katsanidou, V. Naßhoven, and E. Balaban. Factors influencing the data sharing behavior of researchers in sociology and political science. *Journal of Documentation*, 74(5):1053–1073, Aug. 2018.

[33] W. Zenk-Möltgen, A. Katsanidou, E. Akdeniz, V. Naßhoven, and E. Balaban. Replication data for: Factors influencing the data sharing behavior of researchers in sociology and political sciences (Version 2.0). https://doi.org/10.7802/2284, 2021. Data retrieved from the GESIS Data Archive.