How Researchers De-Identify Data in Practice

Wentao Guo, Paige Pepitone,[1] Adam J. Aviv,[2] and Michelle L. Mazurek
University of Maryland
[1]NORC at the University of Chicago
[2]The George Washington University

Abstract

Human-subjects researchers are increasingly expected to de-identify and publish data about research participants. However, de-identification is difficult, lacking objective solutions for how to balance privacy and utility, and requiring significant time and expertise. To understand researchers' approaches, we interviewed 18 practitioners who have de-identified data for publication and 6 curators who review data submissions for repositories and funding organizations. We find that researchers account for the kinds of risks described by *k*-anonymity, but they address them through manual and social processes and not through systematic assessments of risk across a dataset. This allows for nuance but may leave published data vulnerable to re-identification. We explore why researchers take this approach and highlight three main barriers to more rigorous de-identification: threats seem unrealistic, stronger standards are not incentivized or supported, and tools do not meet researchers' needs. We conclude with takeaways for repositories, funding agencies, and privacy experts.

1 Introduction

Social scientists, clinical trialists, and other researchers collectively generate extensive data about human subjects. Publishing this data bolsters reproducibility, empowers meta-analysis, and creates transparency [72]. As a result, data publication is not only seen as best practice but also increasingly required to publish papers [94] and receive research funding [7, 73]. The amount of published data is enormous and growing, with the Inter-university Consortium for Political and Social Research (ICPSR) archive—one among thousands of archives [80]—hosting over 20,000 datasets as of 2025 [42]. However, sharing data about humans risks social stigma, legal consequences, physical violence, and other harms [11, 15, 24]. Thus, researchers *de-identify* data: modifying data to make it harder to re-identify or learn information about individuals.

De-identification is hard. Researchers are expected to preserve data utility for various, often unspecified use cases. They must also avoid privacy pitfalls, as approaches from the ad hoc to *k*-anonymity [91] can fail in unexpected ways [17, 71, 77]. The privacy community has laser-focused on differential privacy (DP) as the solution. DP enables data release with measurable and guaranteed limits on inference about individuals [25]. Surveys [20, 40, 61, 65, 74, 83, 95] detail the myriad variants and improvements since DP's formalization in 2006, and the pace of research continues unabated [22, 23, 29, 39, 43, 45–47, 51, 52, 84, 104]. Some have argued that DP can support sharing of human-subjects research data by creating public statistics to accompany access-restricted datasets [79] or by creating entire synthetic or modified versions of datasets [6, 30, 54, 67].

However, real-world deployments of DP are largely limited to a few government agencies [37, 97] and tech companies [2, 4, 21, 62, 85], while a large portion of research data is de-identified using non-formal methods and published with minimal access restrictions in repositories such as ICPSR [42]. De-identification resources for researchers also generally do not recommend DP [35]. One reason is that researchers have concerns about suitability, as evidenced by widespread fears that the use of DP for the 2020 U.S. Census could undermine their work [69, 86, 102].

To empower human-subjects researchers to protect participants' privacy with state-of-the-art methods, we need to bridge the gap between them and the privacy community. To do this, we must ensure that tooling supports the use cases for publishing data in the first place, and that it is usable and useful enough to be adopted. It is possible this will ultimately require new methods or paradigms for data release.

As a crucial first step, we investigate researchers' current de-identification mindsets and practices, by interviewing 18 practitioners who have de-identified data for publication and 6 curators who review data submissions at repositories and funding organizations. We ask three main research questions:

1. How do researchers perceive re-identification threats?
2. How do researchers de-identify data in practice?
3. What challenges do researchers encounter?

We find that participants conceptualize risk in ways that align with *k*-anonymity [91], but they de-identify data following manual and social processes—evaluating indirect identifiers one or two at a time, estimating risk through discussion with collaborators—and not through systematic approaches such as *k*-anonymity. Thus, few ensure individuals cannot be singled out even by modest sets of identifiers. Some of their reasons may be familiar to security and privacy scholars: de-identification is a secondary concern that can clash with the primary task of publishing research papers [1], and despite great potential for harm, researchers doubt published data will be targeted by competent threat actors [88, 89]. Researchers also face challenges involving usability, capacity, and incentives, though not all feel these impediments are significant.

We also ask two targeted secondary research questions:

4. What is the role of external review?
5. What are researchers' perceptions of perturbative methods, such as differential privacy?

We find that review by curators can bolster de-identification but is undermined by communication issues; additionally, curators do not generally expect or recommend systematic de-identification approaches. Finally, researchers are skeptical that perturbative methods like differential privacy could meet their needs for data utility, but they are potentially open to new practices if funders expect them and provide support.

We conclude by highlighting three barriers to systematic de-identification, along with recommendations to privacy scholars and tool developers on meeting researchers' needs.

2 Background

We describe shifting expectations for publishing research data, de-identification techniques, and practitioners' perspectives.

Publishing research data. A growing open-science movement champions the benefits of publishing data, such as enabling replication and meta-analysis, as well as providing transparency to the public [72]. Starting in 2013, the U.S. Office of Science and Technology Policy has directed federal agencies to maximize public access to data from research they fund, while balancing confidentiality [38, 73]; it reported a huge increase in published data by 2021 [53]. Other countries [44, 76, 96], non-governmental research funders [7, 100, 103], and publication venues [94] also incentivize or require data publication.

Research data repositories can influence de-identification through guidelines and oversight by curation staff; example steps in the data submission workflow include review of disclosure risk, checks for format and quality, and creation of documentation [99]. Some repositories have built substantial infrastructure for reviewing de-identification, as with the Millennium Challenge Corporation's Disclosure Review Board [64]; others take less structured approaches. Repositories also facilitate restrictions on access and use, for example by reviewing access requests before users can download data, or by having users agree to not attempt re-identification [98].

De-identification. Human-subjects data is typically de-identified and/or access-restricted. De-identification involves removing or modifying attributes, which may be directly (e.g., email address, ID number) or indirectly (e.g., zip code, gender) identifying. However, de-identified data can still reveal information about individuals. We mainly consider the risk of *singling out*: locating a record[1] with a unique combination of indirect identifiers [90]. Singled-out records are *re-identified* once they are linked to individuals using external data (e.g., voter rolls) or an attacker's personal knowledge. Although other risks exist [49], we focus on singling out because it was the main risk discussed by participants.

Carvalho et al. review de-identification techniques [14]. Some techniques do not introduce inaccuracy: deleting data or *generalizing* values into broader categories (e.g., replacing cities with counties, using income brackets, and *top-coding* outlier values such as ages over 80 to *80+*). Other *perturbative techniques* produce data that is somewhat inaccurate: e.g., adding noise or swapping values among records.

De-identification can be structured around systematic approaches, such as *k*-anonymity, that measure the strength of a dataset's protection. A dataset is *k*-anonymous if each record shares the same values with at least *k-1* others across the *quasi-identifier*, a set of indirect identifiers chosen for their potential to link records with external data [91]; this is often achieved by deleting and generalizing data. The greater *k* is, the harder for an attacker to match a record to a particular individual. Unfortunately, *k*-anonymity and prominent variants [55, 56] are not designed to prevent re-identification using unconventional identifiers [70], and they can be reverse-engineered, depending on implementation [17].

In contrast, DP algorithms provide probabilistic limits on the impact of each individual's data on the final data release, capping the potential for re-identification. The limit is formally defined in relation to the ε parameter and cannot be exceeded, no matter how the data release is processed or what external data becomes available for linkage [27]. A common use case for DP is to protect aggregate statistics by adding random noise to the outputs, but methods have also been developed to release differentially private synthetic or modified versions of datasets [6, 30, 40, 54, 74].

Practitioners' considerations. De-identification can be difficult, especially for academic researchers who may need to publish datasets spanning hundreds of variables without the resources and expertise of government agencies or tech companies. Based on their own experience, researchers reflect

[1] A record is a single data unit, often corresponding to one person or household and/or one row in a dataset.

that dealing with k-anonymity [3, 59], DP [32, 33], and qualitative data [12] is rife with difficult decisions about which the research team may reasonably disagree [48]. Eastwick, for example, notes that it is "frustrating and disappointing that so much digital ink has been spilled over the importance of data sharing but very little of it has been devoted to helping researchers with (modestly) complex data sets" [48].

Increasing privacy tends to decrease data utility for future uses—but both are crucial, creating a tricky balancing act for practitioners. There is a general lack of objective answers, whether the question is how to generalize variables, what value of ε to use for DP [26], or which dataset characteristics are most important to preserve. Further complicating matters, the answers practitioners choose will inevitably support some use cases but not others, but they may want to support a range of use cases, not all of which are clear in advance.

Practitioners have resources to help with de-identification. Some tools de-identify data algorithmically [41, 78, 82, 92], and a limited body of work on usability shows that the effort to make existing tools user-friendly is only just beginning [75, 93]. Tools can also support decision-making: e.g., generating visualizations to help choose ε [68] or checking for potential problems [19]. Practitioners may refer to de-identification guides and frameworks [3, 13, 28, 57, 59], though existing online guides tend to have limited examples and sparse coverage of threats to de-identified data, and guides for researchers in particular rarely cover systematic approaches such as k-anonymity or DP [35].

There is unfortunately little work investigating researchers' mental models of disclosure risk; techniques and tools used in practice; and the impact of sociotechnical context. Peterson et al. investigated researchers' approaches to redacting qualitative data specifically [81]; our work covers both qualitative and quantitative data, focusing more on the latter.

Debating utility and compatibility with research norms. Some researchers and policymakers have concerns about how the inaccuracies introduced by DP affect data utility, spurred in large part by the use of DP in the 2020 U.S. Census [69, 102]. Some studies have sought to assuage specific concerns about utility, showing that despite introducing inaccuracies to the data, certain DP applications maintain distributions over variables [16], preserve replication of analyses [67], or support downstream applications [18]. On the other hand, some studies have argued that existing DP implementations are incapable of providing acceptable utility without sacrificing a meaningful privacy guarantee, at least for certain use cases [8, 31]; the inventor of DP herself writes that DP is the wrong tool for studying outliers or small datasets [26].

Beyond questions about utility for specific use cases, concerns regarding DP extend to deeper disagreements about what kinds of techniques and justifications are compatible with the norms of a research community. For example, altering data has not traditionally been a widely accepted scientific practice, even when done transparently to protect participant confidentiality [58]. In this vein, Sarathy and boyd argue that the Census's use of DP triggered controversy because it conspicuously exposed ways in which Census data is manipulated using statistical techniques to bolster accuracy and confidentiality—which clashed with many social scientists' expectations of the Census as a straightforward count of U.S. residents [10, 86]. As another example of tension with existing norms, Sarathy et al. found that data practitioners who interacted with a DP tool did not feel that the output statistics would enable them to conduct exploratory data analyses or replication, due in part to the lack of access to raw data [87].

It may be that DP's trade-offs make it a poor fit for a significant portion of human-subjects research datasets, at least for the foreseeable future. However, the same kinds of concerns over utility and compatibility apply to other de-identification approaches, and addressing them is key to increasing adoption of any de-identification method.

3 Method and participants

We conducted 18 semi-structured interviews with practitioners who have de-identified data for publication, as well as 6 interviews with curators who review data submissions for repositories and funding organizations. We held interviews on Zoom between September 2023 and May 2024.

Recruitment. We recruited practitioners from eight online repositories that require de-identification. We chose some repositories that host many types of data, such as ICPSR [42], and others specific to the sensitive topics of health, crime, and international development. We systematically reached out to recent submitters of public de-identified datasets, contacting 84 practitioners and yielding 14 interviews.

We recruited four more practitioners via non-systematic repository searches and professional connections, targeting diverse experiences such as working at NGOs, using k-anonymity, and publishing data in restricted access. As practitioners emphasized experiences both receiving and providing curation, we also recruited six curators for interviews by emailing repositories and funding organizations, in order to understand how they evaluate and ensure the quality of de-identification for submitted data. We continued recruiting until the research team agreed we had reached saturation, based on preliminary themes developed during analysis [34]. No two participants are from the same organization.

We required participants to be 18 years or older, speak English, and have experience de-identifying data or curating de-identified data. Participants described their organization and demographics in a two-minute pre-interview questionnaire (Appendix A). Afterward, they received a $50 Tango card, which can be redeemed for various monetary rewards.[2]

[2] https://www.tangocard.com/

Interview design. Semi-structured interviews with practitioners had six parts:

- **Background:** In what context participants de-identify data, and how they learned to do so.
- **Process:** How participants de-identify data, including techniques and implementation details.
- **Evaluation:** How participants determine when a dataset is sufficiently de-identified.
- **Challenges:** What challenges participants face in de-identification, and what tools or resources might help.
- **Threat modeling:** What threats to published data participants perceive, and how they understand the effectiveness of their de-identification measures.
- **Perturbation:** Under what circumstances participants think perturbative methods (e.g., DP) are appropriate, and what barriers prevent greater adoption.

We piloted our protocol with three experienced practitioners: a graduate student, the second author (a research scientist), and a consultant who helps businesses de-identify customer data. As a result, we focused our study on researchers and rephrased confusing questions. The two interview protocols for practitioners and curators are in Appendices B and C, respectively. Interviews lasted 62 minutes on average.

Thematic analysis. We aimed to both describe participants' de-identification practices and develop interpretive themes explaining why they follow these practices. Therefore, we conducted a template analysis, a form of thematic analysis that mixes both descriptive and interpretive coding [50]. The first two authors coded two transcripts[3] collaboratively to develop a qualitative codebook. As the codebook's high-level structure was stable, they coded five more interviews separately, meeting after each to resolve differences and develop preliminary themes. As the codebook structure stayed stable, remaining interviews were coded sequentially by the first and then second authors, and the research team worked together to further develop themes. We used the same codebook for both practitioner and curator interviews, because both covered similar thematic topics. Our codebook is linked in Appendix D.

Interviews covered both quantitative and qualitative data, but our analysis focused more on approaches to quantitative data—especially for de-identification processes (§4.2)—for two reasons. First, participants overall talked more about quantitative data, providing concrete detail that helped us better understand their processes without observing them at work. Second, existing work has taken a similar approach to study how practitioners de-identify qualitative data [81].

Limitations. As this study is qualitative, we avoid making generalizable claims. For context, we sometimes report the number of participants who expressed a particular idea, but this does not indicate its true prevalence. Thematic analysis is also an inherently subjective method where researchers develop nuanced interpretations of data relying on their own experience and knowledge. For a deeper and more rounded analysis, we involved two coders from diverse backgrounds, who both conducted interviews: the first author is a computer science graduate student who is familiar with academic literature and online guidance on de-identification, while the second author is a social scientist at a research company with hands-on experience de-identifying and publishing data. Both asked interview questions and interpreted responses in ways that were not intuitive to the other.

Our recruitment method contains potential bias. Some potential participants may have been restricted from taking part by organizational policy. As we recruited practitioners from repositories that require (and often review) de-identification, their practices may differ from those who submit data to repositories with less apparent oversight, such as OSF, Zenodo, and Figshare. We also recruited primarily from repositories that are used by researchers in the U.S., though participants' research topics span international cultural contexts.

Participants may have presented idealized accounts due to social desirability bias. We followed best practices to mitigate this, such as emphasizing that there are no right or wrong answers.

Participant information. Table 1 contains information about participants' backgrounds. Pseudonyms begin with P (practitioners) and C (curators). Of the 18 practitioners, 6 are professors, 5 are full-time or postdoctoral research scientists, 5 are in positions in which they oversee de-identification while coordinating research or managing data, and 2 are graduate students. Research topics mainly span health and healthcare (e.g., sexual behaviors, physicians' clinical practices), crime and criminal justice (e.g., intimate partner violence, recidivism), and international development (e.g., perceptions of local organized crime, standards of living). Participants cover a range of demographics and are mostly based in the U.S.; more details are in Appendix E.

When asked how they learned to de-identify data, 19 mentioned learning by doing, often adopting mentors' and collaborators' practices; 12 mentioned literature, including online guides, internal guidance, and research papers; and 10 mentioned coursework, job training, and other structured learning. Several emphasized the primacy of learning by doing: though P18 also learned through coursework and guides, they said, "The point at which I feel like I really fully understood options and requirements and best practices for de-identification came when I was required to make data publicly available ... working with data curation experts." Similarly, P15 had learned about de-identification in courses, but—perceiving a "gap" between computer science theory and what works in practice—they said, "You learn more by just actually having to do it and stress-test systems or datasets on your own."

[3] To generate transcripts, we used OpenAI's Whisper model locally to transcribe interview recordings, correcting the results manually.

Table 1: Participants' research areas, organization types, and counts of datasets de-identified and curated. Counts are based on participants' rough estimates and provide only order of magnitude. Blanks indicate we did not ask, not necessarily a value of zero.

	Research area	Organization	# de-id'ed	# curated
P1	Health(care)	Co./NGO	10–19	
P2	Health(care)	University	20+	100+
P3	Health(care)	University	20+	
P4	Health(care)	University	10–19	no answer
P5	Health(care)	University	5–9	
P6	Health(care)	University	5–9	
P7	Health(care)	University	1–4	
P8	Crime / justice	University	20+	
P9	Crime / justice	University	10–19	
P10	Crime / justice	University	5–9	
P11	International dev.	Co./NGO	20+	
P12	International dev.	Co./NGO	20+	
P13	International dev.	Co./NGO	10–19	
P14	International dev.	University	20+	
P15	International dev.	University	5–9	
P16	International dev.	University	5–9	
P17	Public policy	University	10–19	
P18	Psychology	University	10–19	
C1	Health(care)	Repo/funder[*]		500+
C2	Crime / justice	Repo/funder		40–99
C3	International dev.	Repo/funder		100+[†]
C4	International dev.	Repo/funder	20+	100+
C5	International dev.	Repo/funder		500+
C6	Social sciences	Repo/funder		no answer

[*] For confidentiality, we do not specify which curators work for repositories, funding agencies, or organizations that function as both.

[†] This interview was conducted with three curators from the same organization. As they shared similar experiences and perspectives, we treat them as if they were a single participant (C3), including for reported counts, except where otherwise noted. The number in this table is the combined number of datasets curated by all three individuals.

In terms of techniques, 21 described generalizing data into broader categories, 18 deleting data, 13 pseudonymizing or hashing IDs or names, and 7 adding noise. Other techniques were mentioned by 4 or fewer, including imputing values (e.g., replacing with the mean) and subsampling (withholding part of the dataset from release). One participant has used DP, and 3 have applied *k*-anonymity or helped others do so.

Some have awareness of whether and how data they publish is used: while 6 expressed uncertainty, 8 knew that it had been used for research, and 2 for policy development. Some mentioned that published data has been used in research training exercises, news articles, and machine learning models for health products. De-identified data—sometimes an exclusive, less private version—is also often used by funders to inform their own programs, such as public health initiatives. We did not ask why participants had published data, but several practitioners indicated it was a requirement for research funding or publication, while one said they try to publish data when appropriate out of support for open science.

4 Findings

To learn how privacy scholars can best support researchers in de-identifying data, we must first understand their current knowledge, practices, and challenges. We find that researchers think about risk in ways that align with *k*-anonymity, but they do not take systematic approaches that ensure against singling out. Our findings explore why.

We sometimes avoid specific attribution of participants' responses to reduce the risk of re-identification.

4.1 Perceptions of threats

While participants are concerned about serious potential consequences of re-identification, they find competent threat actors to be an unlikely hypothetical. At a high level, the way they think about risk aligns with *k*-anonymity.

Participants see serious consequences to re-identification. Six mentioned physical violence, ranging from domestic abusers retaliating against survivors sharing their experiences, to repressive governments, terrorist groups, and criminal organizations targeting people who view them negatively. Seven mentioned impacts on individuals' livelihoods or basic services, including retaliation from employers and denial of health insurance. Eleven mentioned social consequences, such as ostracization for culturally stigmatized kinds of sexual activity. Other concerns include incarceration and other legal consequences, as well as the need to respect data subjects' privacy and trust even absent a specific harm.

Potential harms can make participants reluctant to publish data at all. P1, who studies sexual practices and health in regions with stigma and criminalization, sometimes rules out journals that require data publication: "We're doing a research study based on sensitive data that we cannot share, but the journal requires us to share it, so we are between a rock and a hard place." P10 was required by a funder to publish transcripts describing traumatic experiences that could affect ongoing legal cases but felt uncomfortable doing so: "We were just so worried that our participants would be identifiable." They ended up publishing under access restrictions.

Participants doubt attacker capabilities and motivation. Despite an array of potential harms, the vast majority of participants find the threat of re-identification unlikely. Though C1's "nightmare" is a re-identified individual being denied health insurance based on family history, they added, "That's more like a sci-fi paranoia concern than it is a practical one."

Participants gave three main reasons for skepticism about the likelihood of real-world attacks. First, many believe de-identification creates enough of a barrier to deter attackers. P14, for instance, thinks no one would try because re-identifying their data is impossible. Adding some caveats, P4 believes that even though re-identification is probably possible, *motivated* attackers lack the capability to do so: "The level of sophistication and effort you would need to do that is beyond the real threat." As an example of a motivated but non-expert attacker, P4 described a data subject's spouse or partner who might try learn about a visit to the doctor for a sexual health issue; in this case, P4 argued that the more likely danger is the attacker acting harmfully based on incorrect conclusions. We note that while P14 and P4 both de-identify indirect identifiers using techniques such as generalization and noise addition, neither described following processes to ensure that individuals cannot be singled out.

Second, participants believe there would be little benefit to attackers. P2 said, "There's usually not much to gain monetarily—blackmail—or reputationally in these datasets." P13 and C4, who both do research in international development, both noted that their typical data subjects do not own bank accounts, arguably ruling out most profit-driven crime.

Third, participants perceive lower-hanging fruit for attackers to accomplish the same ends. This includes more accessible sources of data. P8 questions the point of taking de-identification seriously when criminal justice departments identify individuals in public records, shared datasets, and more: "Sometimes [department staff] don't really care, and there's probably all sorts of emails shooting around about those individuals while they're on supervision. But here we are, researchers, being terrified that we have that information." Other aspects of research infrastructure are also seen as lower-hanging fruit: to C3, the "biggest risk" is that systems might be hacked to expose data that has not been de-identified at all. P1 worries more about individuals being identified during data collection: "If police wanted to do a crackdown on illegal acts, the last place they would go is our dataset. … De-identification for us means that we need to set up a coffee shop with a special room where we do interviews." As C1 summed up, "It's just not the best access point for any bad actor—they're not going to come to a social science repository with intentions to identify 1 of 500 people that responded to a study in 1989. That's just kind of a bad premise."

Despite doubts, some are wary of underestimating threats. Uncertainty about threats can be a reason to take them seriously: P13 analogized, "Illegal armed groups employ lawyers, so I don't know why they wouldn't be able to employ data scientists, you know? So, I took that very seriously." Because P5 is unsure about the likelihood and real-world occurrence of re-identification attacks, they guess that the likeliest threat may come from other research participants, as well as participants' colleagues or friends; therefore, they intentionally avoid mentioning the existence of published data to their participants. Others simply stressed that de-identification should be strong, regardless of threats: C2 said, "Participants were guaranteed their safety, so we should ensure that it's protected."

Some actors may have reason to re-identify datasets, even if it brings them no benefit. Citing vaccine research targeted by conspiracy theorists, P1 expressed concern about attacks intended to publicly sabotage the project's reputation: "It would make a big headline against them, the funder, [and] against us, the organization that's doing this study, if a [re-identifiable] dataset comes out."

Participants largely perceive risk factors that align with *k*-anonymity's risk model. Related to the value *k*, nineteen mentioned the risk posed by uncommon or unique values of indirect identifiers as they explained how they de-identify data, with fifteen saying they look for uncommon *combinations* of values. Related to selecting a quasi-identifier, twelve explicitly noted that some types of data are more identifying, including (depending on context) location, age, occupation, and number of livestock, in contrast to other data types such as opinions of local officials. Four mentioned that richer data (e.g., with more attributes) is higher-risk, which can make it harder to achieve high values of *k*. Thirteen noted that external data can link records to identities, and five noted that personal knowledge of data subjects can do the same.

Even without mentioning it by name, participants often have a *k*-anonymity sort of mindset, with the goal that all records should have look-alikes across some set of indirect identifiers. P6 tries to "avoid putting clinicians into a group of less than five similar clinicians," in terms of identifiers such as age, race, occupation, and location. P7 said the only criterion they applied consistently was asking whether specific combinations of identifiers—such as medication, handedness, location, and age—matched with one record (bad), two or three ("okay, but it's still perilous"), or more (better). And P8 expressed confidence that even an attacker with a rich source of external data—e.g., a criminal record containing arrests, convictions, and location—would probably not be able to single out any record.

Participants' mindsets align less with DP's risk model. Risks other than singling out, such as attribute disclosure [49], did not arise in conversation. While ten mentioned the potential for reverse engineering de-identification (one of *k*-anonymity's potential weaknesses [17]), they largely mentioned attacks that could be mitigated by following best practices for *k*-anonymity and other non-formal techniques: e.g., assigning pseudonyms randomly rather than in alphabetical order, or ensuring sensitive information is removed from documentation. Additionally, as we detail in Section 4.2, many hesitate to treat all attributes as potential identifiers.

The sampling frame provides key peace of mind. Participants' datasets are mostly samples from a larger population, which means that an individual who is singled out in the dataset is not necessarily singled out in the population (assuming an attacker does not know the set of people who are in the dataset). As P13 explained, if only one man from Maryland was surveyed, "That doesn't necessarily mean, though, that combination is identifying, because there's so many people in the [broader] population with that combination." In total, six discussed the role of the larger sampling frame in reducing risk. Unlike other risk factors mentioned by participants, accounting for the sampling frame is somewhat at odds with k-anonymity: as P13's example shows, the sampling frame is a legitimate reason to tolerate records that are singled out. Similarly, recalling a recently reviewed dataset, C3 said a small group of look-alike records might be concerning in rural sites where the sample is ~50% of the population, but much less so in urban sites where the sample is ~2%.

4.2 How researchers de-identify data

Focusing on quantitative data,[4] we find that even though participants take steps to reduce the threat of singling out, they rarely evaluate risk systematically across identifiers or apply clear standards for when data is de-identified. Instead, they often approach data manually, inspecting one or two variables at a time, and they rely on informal and social processes, such as consulting collaborators, to make decisions.

Participants manage risk in limited subsets of identifiers. All participants incorporate traditional indirect identifiers—e.g., age, gender, and location—into their de-identification process. Many largely focus their attention on these. At P12's organization, practitioners start by creating a codebook to document information such as the risk each variable poses, how they will be de-identified (if at all), and the resulting impact on data utility; although each variable is documented, individual and household demographics are usually more scrutinized than opinions, working hours, and others "harder to be visible to an outside observer." However, several participants did describe incorporating less traditional identifiers, based on their knowledge of what information is distinctive among their research population; examples include drug use, insurance type, social media use, and records of prison incidents.

Though participants include some non-traditional identifiers, they are often reluctant to expand the scope, arguing that many variables pose minimal risk in context. P6 believes it is "a little silly" that date of healthcare service—important for many research questions, they noted—is an identifier to be removed under the U.S. Health Insurance Portability and Accountability Act (HIPAA). They explained, "[In my] department, we probably have [hundreds of] patients there that have a date of service of [date]. I can't tell you who any of them are." Similarly, P12 feels that practitioners in international development at their organization sometimes unnecessarily de-identify data types that a realistic attacker would not find useful, such as the type of toilet in the home.

To classify variables, researchers often talk through risk together: P16 said, "I get a little bit into the weeds sometimes about things, and I'm like, 'Ooh, they have two chickens, and nobody else has two chickens.' And my boss will be like, 'There's a very minute possibility that somebody would go to this village, and they probably have more chickens now.'"

Participants largely de-identify data by deleting and generalizing entire variables. Despite being a blunt tool, wholesale deletion is preferred by some because it is the "safe route," as described by P1, who prefers to automatically delete all variables unnecessary for replication. This is not only the best way to protect data subjects but also the easiest to justify: "It's much easier to answer the question afterwards of 'Why didn't you put it out?' Well, we thought it was too sensitive. But if we put it out and it so happens that those data somehow end up in the incorrect hands ... we're gonna have plenty of very hard questions." P6 similarly declined to provide any race and ethnicity data for a study, even though it was requested to help answer research questions, out of fear that individuals could be singled out in combination with other identifiers such as age and occupation details.

Deleting and generalizing entire variables are the most common way to handle indirect identifiers, with exceptions including top-coding outliers and adding noise to dates and locations. While this broad-strokes approach is more straightforward for humans to execute, prior work suggests that incorporating other methods such as local deletion [3, 59] may be beneficial for data utility.

Participants often prioritize certain variables to de-identify first. Location is a common first step in part because researchers tend to think about identifiability in terms of geographical areas: C2 starts with location because, for them, identifiers such as race and age would not be a major concern "unless it was a very small geographical area." Preferences are also based on data utility: given a record that could be singled out by age and gender, P12 would try to generalize age first, as "gender's typically an important variable."

Participants remove distinctive values, with varying notions of distinctiveness. Participants check for unique or uncommon values, often one variable at a time. Sometimes distinctive values are defined by a clear numerical threshold: citing guidelines from a funding agency, P9 generalizes identifiers such as ethnicity so that no value appears fewer than five times. More often, distinctiveness can be quite fluid: P11 said geographic locations that appear ten times would probably be too distinctive, but generalizing to at least "a couple hundred or a thousand" might be sufficient, adding, "We don't have a well-thought-out process, other than getting people in a room and thinking through that." P17 determines low counts based on "a sense kind of thing," saying they had deleted locations with counts of one or two in a dataset of 1,000, while the threshold might be twenty in a dataset of 20,000.

Some treat individual variables without distinctive values as non-identifying. P10 would treat zip code as an identifier in a small dataset where the same value might be shared by

[4] Due to prior work on how researchers redact qualitative data [81], this section mainly covers quantitative data except at the end.

only two records, but in a large dataset where "every zip code appears now hundreds of times, that's not an identifier to me." C4's organization has a tool that takes a dataset and suggests potential identifiers to focus on; because an earlier version produced "so many false positives that it wasn't very useful," they changed it to stop flagging variables that have a small number of evenly distributed values. While this heuristic importantly makes de-identification more manageable, it risks underestimating the power of even coarse variables to enable re-identification in combination with other variables [90].

Participants also take steps to reduce distinctive values without actually calculating distinctiveness within the dataset: for example, age is often generalized into standard buckets, and P12 has followed repository guidelines to top-code the top 5% of values for certain variables. Location is somewhat unique in that researchers may refer to real population counts outside the sample: C4 generalizes locations with fewer than 30–50,000 residents, while C2 considers populations below 10,000 small. However, both noted these thresholds are not hard cutoffs and can be raised for sensitive data or lowered for utility (leading to increased scrutiny of other variables). One practitioner echoed this concern with cutoffs: they took issue with a funding agency guideline on perturbing geocoordinates, arguing that resulting coordinates could probably still be narrowed down to 1,000 households in some rural areas.

Participants also remove distinctive *combinations* of values, though rarely systematically. Fifteen mentioned examining *crosstabs* (intersections of variables), though C3 noted many data submitters do not until asked to by curators. For example, P16 noted that owning a cow or a tin roof is distinctive when paired with certain locations, leading them to check those crosstabs. As a curator, P2 raised red flags about a dataset with unique combinations of location and race, as well as location and number of children; one of their recommendations was to top-code number of children, as higher values were particularly distinctive in certain locations.

Again, participants indicated that distinctiveness depends on context such as the sampling frame. Some, though, still refer to numerical thresholds: C2 and C3 both check for crosstabs with counts under ten, while C4 said they would start worrying if five or fewer records in a sensitive dataset shared a crosstab of location, occupation, and caste or religion. P2 does not typically calculate k-anonymity, but they keep in mind a goal of $k = 2$ or 3 when calculating crosstabs.

While participants often evaluate crosstabs, they rarely do so across a whole set of identifiers, as one would to achieve k-anonymity. Instead, they tend to choose pairs of identifiers that they deem risky. As P12 said, "Maybe it's somebody's position within the community, crosstabbed with their age or their gender. … It's not necessarily a scientific process. It's more knowing what to look for." P1 focuses on crosstabbing sensitive variables, such as those relating to sexual practices and health, adding, "I don't think we [crosstab] all the combinations. Quite frankly, I don't think we do three-way or four-way. … We look at some of the variables that we think could be sensitive."

One curator's repository reviews documentation provided by data submitters rather than the data itself, so they request summaries of specific crosstabs with the greatest expected impact, such as teachers' demographics and subject areas: "We know, based on prior submissions, that there are certain cross-tabulations—oftentimes recognizable demographic characteristics—where you might see identifiable things."

Measuring risk across a whole set of identifiers is seen as extreme. P4 argued that systematic crosstabbing would be too detrimental to utility: "We do look at, on some level, combinations of variables. But I would say right now, we don't do it in a particularly structured way. … It's very difficult when you have a very complex dataset to really, I think, achieve anything useful by doing [generalization] across multiple variables." Even without systematic crosstabbing, P4 often receives complaints from downstream users that important data, such as specific spoken languages, is missing due to de-identification.

A few have used k-anonymity in exceptional circumstances. One participant had a project with a vulnerable population and detailed questions about gender and sexual orientation. This rendered the data sensitive but also motivated the research team to represent data subjects "as they were." Their attempt to generalize data without erasing important identities left records that could be singled out, so they used sdcMicro [92] to achieve k-anonymity by deleting individual values in the data. Another participant has also applied k-anonymity once, to deal with the sensitive topic of sexual behaviors; they used a mix of generalization and noise.

Both emphasized the exceptional nature of these circumstances. The first participant's go-to solution for sensitive data is to restrict access, because k-anonymity "did take a lot out of the dataset." Similarly, the second participant stressed the sacrifice to utility: "The reason we think the protocol works from a privacy perspective is because we gave up so much. … There is a huge amount of analysis that we cannot do."

C4 helped a research team with k-anonymity once, even though they usually only calculate crosstabs of two or three variables at a time. They turned to k-anonymity due to the availability of external linking data, such as voter registration records. In contrast, C4's typical studies are in data-poor areas; i.e., their data may be the only data recorded about research participants. Therefore, preventing singling out is usually a less pressing threat: "The idea of me taking this and linking it to administrative records … we're not worried about that for a remote village in [a developing country]."

In contrast, one practitioner defaults to DP, citing a preference for "provable" methods. Despite some familiarity, they have not used k-anonymity, not only because they find it less protective but also because it undermines certain use cases: "If I want 4-anonymity, I need at least four rows that look the same in some sense. ... [When decision makers] are interested in learning about differences in their population, k-anonymity doesn't fit the bill." In contrast with participants who had used k-anonymity, they spoke positively about DP's impact on data utility, partly based on metrics that provide worst-case estimates of error caused by noise.

Participants often use tools to scaffold the de-identification process. P11, P12, and C4 use tools developed in-house by their organizations that operate on datasets and help plan de-identification, variable by variable. Capabilities supported by one or more of these tools include suggesting identifiers; providing an overview of each variable, including the most common value and five randomly chosen values; and automatically executing de-identification techniques. Several participants also scaffold de-identification using generic functions in statistical programs, such as Stata, to generate codebooks with summaries of variables. More niche cases of tools include one participant who has used public DP libraries and coded their own implementations, and a different participant who uses in-house scripts to search medical data for identifiers in free text fields and to scrape news websites for mentions of data subjects, which might raise their risk.

Almost no one uses tools that carry out de-identification algorithmically, such as sdcMicro [92], ARX [82], and μ-ARGUS [41]. Some may be unaware: P13 said, "If there are R packages [or] ... Stata commands [for de-identification], I'm not aware of them." The participant who had used sdcMicro to apply k-anonymity noted sdcMicro's usability hurdles: "If you're not familiar with k-anonymity and you're not familiar with R, you couldn't use it. It's not a point-and-click tool." We note that several de-identification tools do offer graphical user interfaces, including sdcMicro [5], but our participants did not describe using them.

Informal and social processes dominate for deciding when data is de-identified. P12 reflected,

> [During this interview], I'm like, wow, a lot of this is very subjective. And I never realized it, because we have a very clear process. ... You document, you understand why the decision was made. You talk to multiple people along the process about making that decision. But in the end, it is subjective. The IRB and the project team can disagree. The clients and the project team can disagree. And then it goes down to who has the authority in that position or if there's any compromise that can be made.

Instead of quantifiable metrics and standards, participants often apply their own expertise via manual inspection of data, reflective exercises, and discussions with colleagues. P11 described generalizing location: "We worked our way up as a team until we got to a place where we thought, well, at this level, no one's gonna be able to identify somebody based off of all the combination of other indirect identifiers, like gender, age. ... It was a conversation with the project team, understanding the context of where they're working, not like a flowchart where you're following some hard, set rule." P14 determined that data (with location generalized but other variables left untouched, including age, gender, and number of children) could not be re-identified in a "thought experiment": "We sit around, we look at this and think about what would make it easier or harder based on our own experiences." Similarly, P4 makes many decisions via discussion, rather than an "objective assessment of uniqueness."

As a curator, P2 is discontent that they decide when data is de-identified through informal conversations with submitters: "It's hard because I feel like there's no right answer." But some see this as a natural solution, given inherent subjectivity in de-identifying data. As researchers "reasonably disagree" over which variables to include, how to generalize values, and how to run crosstabs, P1's organization approximates "an internal peer-reviewed process" to hash out compromises.

Some consult advisory bodies such as IRBs for a second opinion. P12's IRB reviews all data for publication: "At least somebody outside the project team is taking a look at the data and being able to raise different questions." And when curators at C1's repository are unsure, for example about appropriate thresholds for low counts, they can consult a committee to ask, "Is this as bad as we think this is?" P10 wants IRBs to play a greater role in overseeing de-identification: "[IRBs] make sure you're thinking about de-identification, but they don't ask ... how do you know when you've pulled enough variables out and de-identified adequately?"

Participants may rely on informal processes in part because they lack awareness of alternatives. P13 said they are not aware of "any hard and fast sort of checks" to determine when a dataset has been adequately de-identified. Explaining how they pick categories for generalization, P1 said, "It's an art. It's a judgment call. ... I don't think there's any pre-established statistical or algorithm to do any of that."

With vague notions of how data is re-identified, most participants guess there is remaining risk. Several, including C1, said they lack a detailed understanding of how attackers might re-identify data: "I don't know what that mindset's like. I'm sure that the right tools in bad hands are probably far more capable than I could ever imagine."

In this light, many believe their data could technically be re-identified, but some view it as outside of the relevant scope of threats. P11 guessed that, on a scale of threats from 1 (a random person) to 10 (a well-funded government actor), their published data is protected up to level 8, making them feel "pretty confident" that they have adequately protected their research participants. C1 suggested their repository is not responsible for preventing attacks that leverage external linking data: "What we want to be able to say is we took the precautions that were necessary to make sure that you were not identified using the data solely that we put out. Now that's put with another piece of information, that's on someone that had the intention to do that, and we can't prevent against that."

Without being prompted, 13 of 18 practitioners brought up and stressed the importance of access restrictions and data use agreements. Access restrictions help avoid uncomfortable compromises between privacy and utility: despite believing attackers with personal knowledge of data subjects might be able to re-identify them, P10 feels "pretty safe" because their data cannot be accessed without IRB approval and a research proposal. Curators share this outlook. When identifiers like county and race are risky but important, C2 retains them but restricts access, believing that malicious actors would not make it through the application process. And C3 sometimes publishes restricted versions of public data with, for example, fewer variables generalized. Still, making data public is common, and we note that we recruited all but one practitioner via publicly available data they had published.

De-identifying qualitative data relies on informal and social processes to an even greater extent. Participants search manually for identifiers in free text, usually redacting them or replacing them with a label indicating the data type (e.g., "city"). In contrast with quantitative data, generalization is an uncommon technique. Participants aim to remove distinctive values and combinations of values, though they refrain from counting to assess distinctiveness, instead relying on critical thinking and conversations with other researchers. No one mentioned using tools, instead emphasizing getting other researchers to read through and provide a second opinion.

4.3 Challenges

Participants' de-identification processes are shaped in part by various challenges concerning usability, capacity, and incentives. The perceived magnitude of these challenges varies, though, with some characterizing them as significant impediments and others feeling that de-identification is easy.

Participants are unsure how best to de-identify data. Some feel forced to make unwanted sacrifices to utility, such as P9, who finds it often "unfortunate" but necessary to combine underrepresented races and ethnicities into an *Other* category. Noting that "time is so important," P4 and P8 both emphasized the "huge struggle" of de-identifying times and dates without "screwing up people's analysis" of seasonal trends, impacts of specific events and policy changes, and more.

Participants called out the need for techniques to better balance privacy with utility. P2 explained that k-anonymity is less useful for sample datasets than for censuses, as it does not account for additional privacy from the sampling frame. Another called for improved DP techniques to enable better utility at lower values of ε, noting a yawning gap between theory and practice: "If I were to pick a value of ε that was so small that [it] might make the privacy community very happy but wouldn't let any policymaker figure out how to get supplies to folks after an earthquake or after a military coup or something, then that's just off the table."

Some trace difficulties back to training and learning resources. Due to gaps in their de-identification training during their PhD, P10 only gradually began de-identifying indirect identifiers later in their career, giving a recent example where a research assistant made them realize that wedding date is identifying. One participant completed the only de-identification certification they know of, but they found it not so personally applicable due to its focus on HIPAA. Some are dissatisfied with online resources. P9 found some that give "general pointers about things to be mindful of" but no comprehensive guide "that talks really thoroughly about the process"; similarly, P7 read many webpages with guidance but wished they had found a "formal guidance document."

Some also blame difficulties on vague expectations and standards. P4 wants a more detailed technical specification than HIPAA's standard, as they struggled to interpret it in cases where, for example, dates were perturbed, or patients in longitudinal datasets crossed the threshold of age 89. P1 said submitting data to funders can result in unnecessarily "spinning our wheels" over subjective judgment calls, especially when a project has multiple funders or when staff at funding organizations change. Curators backed up this finding, with several stressing lessons learned about setting de-identification expectations as early in the process as possible.

Participants have limited capacity. De-identification can take extraordinary concentration and time: C4 described the "tedious process" of manually poring over thousands of variables in a single dataset as a practitioner. This is particularly hard for curators, who sometimes try to review submissions with a fine-toothed comb despite the volume: C3 said, "We've had codebooks come to us with six datasets—six codebooks that were over a thousand pages long each. ... You have to look at it really carefully because you don't wanna skim over that one question where they ask about household negotiation that ends in violence." Thus, as C5 said, publishing data can take years: "We have a never-ending backlog of these things."

Still, some see this effort-heavy approach as imperative. C4 said, as a practitioner, "I don't care how good my automated process is. I'm going to at least want to look at those variables once at the end, just because this is such a big part of what we talk about when we talk about research ethics." Similarly, C3 said, "To do this thoroughly—and it has to be done thoroughly to be done right—there isn't really a way to shorten it."

Disincentives and misaligned expectations are pervasive. De-identification is also often low priority, in terms of both researchers' workflows and funding—with three participants separately describing it as an "afterthought" or "last thought."

As a result, participants admit, data is de-identified with more haste and by less qualified researchers than ideal. P13 said, "A lot of times, we don't even remember we have to do it until the very end. It sometimes even happens that we're doing it after the contract has ended—the resources that are left to dedicate to it are not as much as they should be." Sometimes, P1 noted, funding runs dry, or the researcher most familiar with the data leaves before de-identifying it. Participants also put off adoption of perceived best practices: P4 said, "I would like us to develop some approaches for reporting on uniqueness, *k*-anonymity in the data. It's just low priority."

Participants gave various explanations for why they or others do not prioritize de-identification. Suggesting a collective mindset issue, P11 said practitioners they have worked with "generally" find it unimportant. P16 faulted systemic disincentives in academia, where de-identification is not "necessarily for your own research benefit." C3 said data submitters have yet to adjust to evolving research norms, noting that their organization had only recently prioritized data submission as a deliverable when funding projects. In the same vein, P1 suggested that some funders are also working through growing pains, as grants typically have no funds for de-identification, and even when they do, the funding is rarely enough to do it well. They added that the requirement to submit de-identified data has occasionally come as a last-minute surprise, when the agreed-upon deliverable was a report: "Once we are in that final stage of closing the project, the data are requested. ... It was never clear from the beginning, it was never budgeted."

However, de-identification is often perceived as easy. Several noted that they find de-identification straightforward in comparison to other tasks such as securely storing identifiable data. P18 (who focuses on generalizing location and sometimes deleting demographic attributes) explained that following their procedure makes decisions easy: "I don't think it's that difficult. ... You just need to make sure you build the steps in." We note this sentiment was common but far from universal: others emphasized the difficulty of de-identification, including P10, who said de-identifying interview transcripts was "an awful lot of work—I have never spent so many hours on something in my life."

4.4 The role of external review

Curators can and do raise the quality of de-identification through review. In practice, though, practitioners report receiving minimal feedback, and the two groups sometimes struggle with poor communication and differing priorities.

Models for curation differ greatly. Some curators only review data; others help de-identify it, with one even offering this as a paid service with fees dependent on the amount of work. One repository actually prefers submitters to leave indirect identifiers for curators to de-identify, in the interest of preserving utility. On the other hand, one curator reviews data at some point *after* publication to avoid bottlenecks, while another reviews documentation only, not the data itself, to avoid holding identifiable data that could be exposed by hackers or Freedom of Information Act requests.

External review can strengthen de-identification. All six curators (as well as both practitioners with curation experience) described catching mistakes and raising standards, including spotting direct identifiers like names and ID numbers, flagging indirect identifiers like employment history and movement patterns, calculating distinctive crosstabs like women with tertiary education, and checking for other vulnerabilities like pseudonyms assigned in alphabetical order.

Some practitioners described beneficial experiences with external oversight. One said curators at the Millennium Challenge Corporation (MCC) occasionally request further de-identification to fix distinctive crosstabs, though MCC is one of few research funders to their knowledge that "actually has a rigorous process." Another has had ICPSR consultants evaluate risk by searching for distinctive values and crosstabs. Though ICPSR had "never come up with any re-identification risk," they felt it was valuable because ICPSR's "data-driven" analysis complemented their own "theory-driven" approach: "They don't do it through a set of theories about what the system is like; they do it through computation. Whereas we do it more with like, this is what we understand to be the variables that could pose risk." Another practitioner had presented an early de-identification plan to an external ethics group that demanded improvements, such as adding noise to dates.

Most submitters report minimal feedback from curators. Though we asked each practitioner whether they had received feedback on de-identification, P12 was the only one who described curators requesting changes to strengthen protections. Most said curators had never asked for any changes to de-identification, typically focusing on data quality instead. Having never received substantial feedback on de-identification from one repository, one practitioner questioned the rigor of their review: "I don't really know what capacity they have on the other end to really be making sure that things have been done appropriately, before [data] actually goes up on the Web. I suspect that they're mostly just relying on us to get it right."

One possible explanation for the lack of feedback is that our participants may have simply done a particularly good job: P9 believed this to be the case, saying, "[The curators] clearly thought [the data] was adequately de-identified," and adding that a thorough review included many questions, but mainly about data quality due to a lack of de-identification issues. Another is that curators may not always have more rigorous standards than practitioners. Based on their time as a practitioner, one curator believes most curators have low standards: "I worked with other government agencies that have this same on-paper commitment to transparency. And [in reality] the expectation was you did the de-identification yourself, and you put the data up yourself, and that was that—no one ever checked." One other possible explanation is that curators take de-identification into their own hands without communicating it: rather than ask submitters to fix issues with indirect identifiers, C2's repository typically handles that internally. Likewise, if C5's repository makes de-identification changes, submitters are typically not informed until publication.

Curators generate distrust by pushing for weaker access restrictions and de-identification. Although curators often push for stronger access restrictions, pushing for weaker restrictions is a notable source of disagreement. C3 said submitters sometimes propose more restrictive access levels than necessary: "Our goal is to publish as much data as possible, as safely and securely as possible. If, in our review, we find actually this data is not as risky . . . we push the evaluator to consider decreasing the access type." Sometimes, submitters and curators then come to a consensus: C3 described one case where looking at crosstabs in a meeting with the submitters persuaded C3 to up the access restrictions.

However, some curators make unilateral decisions. As submitters often lack experience due to the "revolving door of personnel working at these places," C5's repository often disregards their recommendations on access restrictions: "It doesn't really matter what they propose, because [the repository] will have the final say. . . . Occasionally the implementing partner will propose restricted because they think there's red flags of some sort. And then [the curation team] comes along and says, 'There's nothing wrong with this; we could publish this as public.'" For practitioners, lacking control over access level can skew their de-identification approach. P9 said one of their submitted datasets was "rendered useless by the amounts of de-identification we had to do" for this reason: "I could say I want you to put it in the highest level of security, but they don't necessarily have to do what I say."

When access is restricted, curators also sometimes push for weaker de-identification than submitters find appropriate. P10 negotiated extensively with curators who felt indirect identifiers such as wedding date and number of children had been unnecessarily removed: "They felt like if you've removed all the really obvious things—like the person's [name, state, town of residence, and date of birth]—then that's probably enough." P10 chalked this disagreement up to different threat models: curators felt that "the chances of somebody working that hard" on re-identification were slim enough to disregard non-obvious identifiers, while P10 felt journalists might try to re-identify the data in order to contact data subjects.

Relatedly, funders sometimes request data for their *own* use that is more disclosive than practitioners are comfortable providing. One participant recounted a funding agency's request for data with only direct identifiers removed (but not GPS coordinates, which the agency argued are not personally identifiable), particularly objecting to the lack of justification: "We have no idea what the hell they wanted it for." After negotiation involving lawyers, they compromised, sharing a dataset with GPS coordinates coarsened and an attribute removed that could be used in active government programs that target illegal activity. Another participant brought up objections about the same funding agency, explaining that their organization has had to push back a couple times on requests for photos of research sites without face blurring: "We're not there to generate a photo for you to put on your PowerPoint presentation—we're there to monitor the site."

Curators are stymied by disengaged submitters. While many submitters are communicative, three curators described a pattern of submitters who stop responding, causing delays and more work. To address de-identification issues, C1 needs input from submitters on how to proceed, but submitters often disappear: "It's at the end of their grant, they're onto their next thing. . . . [The issue] sits in somebody's email for three to six months, and then we have this ticket for this project that's sitting there—our supervisors are clamoring for answers." C2 added that some submitters reject further responsibility for the data outright, asserting that they have satisfied their grant obligations by submitting data. In some cases, disengaging is a crude but effective privacy-preserving tactic: when C2 requests more disclosive data, "Most people will just ignore our emails if they don't want to give us the data."

4.5 Perturbative techniques

Due to controversy over inaccuracies introduced by DP methods [10, 69, 86, 102], we asked about perceptions of perturbing data. We find participants are largely unfamiliar with such methods and cautious about the impact on utility, but they are open to adopting new approaches if it were an expectation.

Participants have little familiarity with perturbation. Aside from one who uses DP, practitioners' experience with perturbative techniques is mostly limited to occasionally adding noise to locations and dates. Like many, P16 expressed a general sense of unfamiliarity: "I just haven't encountered them. I don't think I've heard of anybody that uses them on a regular basis." Even curators generally do not recall encountering perturbed data.

Participants are cautious about the impact on utility. Many are concerned that perturbative techniques might not preserve important relationships between variables. For example, C3 worries about losing the ability to compare results between highly educated men and highly educated women. Caution sometimes stems from lack of familiarity: P14 said they needed a "better understanding of the implications" including the impact on cross-tabular analyses, which their funding agency values. However, despite some knowledge of perturbative techniques, P2 remains wary; they employ multivariate analyses that involve "not just a single main effect or even a single interaction, but many variables," which they believe are not well supported by existing perturbative techniques.

On DP, C4 feels that too little is known about its impact on small datasets, given few real-world instances. For example, they said a researcher might design a randomized controlled trial that is just barely powered due to limited funding, then later find that DP would make their study unpowered.

Many practitioners are open to new approaches, pending guidance and support. Though standards like *k*-anonymity and DP are viewed as extreme, many would change their practices if directed to. For example, P13 said the only reason they have not done more perturbation is that it is not "traditionally requested" by funders, and P1 explained that a funder asking for synthetic data would be "all the motivation we need: someone telling us we want it, and we wanna pay for it."

Participants noted the need for support to make these changes. For example, P1 has wanted to publish synthetic data, but they lack the know-how to do so without losing univariate and bivariate distributions or other statistical properties: "That's a whole different level of expertise that's needed, and I wouldn't say we have the skills to do that properly."

5 Discussion

Norms around de-identifying research data are in transition. Both practitioners and curators are adjusting to new workflows, with misunderstandings often creating friction between the two groups. After decades of de-identification research, practices are being updated at large scale, but not necessarily in the way scholars have advocated for. Even though our participants all take steps to reduce the risk of re-identification via indirect identifiers, in all but a handful of cases the question of whether singling out—the fundamental threat addressed by *k*-anonymity—has been prevented is left unanswered. In many cases, some individuals in published data can likely be singled out by combinations of as few as two or three identifiers. If re-identified, they may be at risk of stigma and reprisals from domestic abusers, governments, employers, and more.

We call out three interlinked reasons for the existing state of practice. Addressing them requires providing researchers with better support, including de-identification tooling tailored to their needs. Therefore, we provide recommendations for repositories, funding agencies, and privacy experts to help researchers de-identify data.

5.1 Hurdles to more rigorous de-identification

Threats seem unrealistic. Most researchers doubt competent adversaries would go to the effort of re-identifying data, which calls into question the need for changes. Adversaries are not entirely hypothetical, though—the journalists who used Grindr data to out a Catholic priest demonstrate that some will re-identify online data to further an agenda [9]. Threats also change; though C1 described a re-identified data subject being denied health insurance as a "sci-fi paranoia concern," technologies even recently considered futuristic, like generative AI [63] and facial recognition databases [36], are now commonplace. Despite the cost of scraping mountains of personal data online—often in potential violation of regulations and terms [36, 63]—these technologies have gathered enough investors and customers to thrive. Without strong protection, research data could be the next target.

Research data is downplayed as a target in part because personal data handled carelessly by government officials or sold by brokers is seen as an easier and more fruitful target. Leading by example across all levels of government and instituting stronger protections for commercial user data would better protect everyone's privacy and also create greater motivation for researchers to protect published data.

Participants often feel they have made re-identification sufficiently difficult to dissuade likely threat actors, but they also expressed lacking detailed understanding of re-identification attacks. Educational resources could help researchers make more informed decisions by explaining re-identification mechanisms; currently, many online de-identification guides have gaps in their coverage of threats, and very few walk readers through real case studies of re-identification [35].

Stronger standards are not incentivized or supported. Though practitioners care about protecting data subjects, de-identification is ultimately a secondary concern that not only takes time from the primary task of publishing papers but also lowers the quality of data in the eyes of their peers. Participants cited concerns about loss of data utility when discussing systematic and perturbative approaches, and P4 has even fielded complaints from downstream data users about information removed during de-identification. Thus, practitioners face similar incentive dilemmas as other professionals who are expected to execute secondary security and privacy tasks [1].

Practitioners feel that funders and other stakeholders generally do not have clear expectations for de-identification, and they rarely receive feedback on de-identification during review by curation staff at repositories. As a result, systematic approaches to de-identification are often de-prioritized. Instead, methods are largely passed down from colleague to colleague. This may partly explain why several said de-identification is easy: they follow familiar processes and are not expected to make more difficult decisions, such as where to sacrifice utility for a higher standard of protection.

Tools do not meet researchers' needs. Tools that de-identify data algorithmically [41, 78, 82, 92] could improve both privacy (e.g., by ensuring more systematically that no one is singled out) and utility (e.g., by applying less destructive techniques than deleting or generalizing entire variables). However, they are rarely used. The simplest reason is a lack of awareness. Researchers need to be informed and persuaded that these tools are relevant to them.

However, even when they are aware, researchers feel that existing algorithmic tools and the approaches they take, such as *k*-anonymity, are not tailored to their needs. Perceived limitations include unacceptable impacts on data utility, poor usability, and an inability to account for risk within the sampling frame (§4.1). Some limitations identified by participants inherently lack neat fixes: trading utility for privacy is somewhat inevitable, and assessing risk within a sampling frame requires acquiring or guessing information about the larger population whose data was not collected. Still, these barriers could be lowered by designing tools to researchers' needs.

5.2 Recommendations

For repositories, funding agencies, and curators. Participants indicated that funders are critical drivers of change, but they need to provide clear expectations and sufficient funding for de-identification from the get-go. De-identifying real-world research datasets—which often span hundreds of variables—is a huge undertaking, and researchers do not feel this is reflected in how it is funded and supported. More funding would reduce the current pressure to de-identify data as quickly as possible, and funders and repositories could provide more support through helplines and consultations. Exemplary de-identification efforts could also be recognized as part of awards for data publications and open science [101].

Repositories and funders can consider several measures to reduce strain between submitters and curators. Submitters may appreciate greater transparency and agency in the de-identification review process. Without feedback, they sometimes assume that curators are not conducting a serious review. And when they lack agency, they may act adversarially for the sake of privacy—throwing away data utility in case a repository disregards their recommended access restrictions, or employing lawyers to get out of sharing certain data. Correspondingly, curators may appreciate stronger requirements (e.g., in research grants) for practitioners to see data publication through to the end, to avoid leaving curators in the lurch with unfamiliar, inadequately de-identified data.

For privacy scholars and technologists. Privacy experts can empower human-subjects researchers by helping them consider risk more systematically when de-identifying data; by providing them with more nuanced methods for preserving utility; and by creating tools that make the whole process more accessible. To do so, though, they must account for researchers' concerns when designing metrics, algorithms, and tools. As an example effort to bridge this gap, MinBlur is a *k*-anonymity algorithm for social scientists expressly designed to account for some of the same priorities mentioned by our participants: users specify which identifiers are more important to preserve for analysis, and they provide input to estimate risk within the sampling frame [66]. Such efforts to center researchers' perspectives in design could produce tools that are not only more useful for researchers but also make a more compelling case for uptake.

Our work has uncovered some of researchers' constraints for de-identification methods, but future work should lay out these constraints more systematically—considering factors such as how utility should be measured for various use cases, what parameters would help researchers customize the process, and how different approaches might be at least initially (in)compatible with existing research norms. Ideally, de-identification methods should be designed from square one to fit these constraints, as much as possible while still providing privacy protection. Human-subjects researchers should also be involved in the design and testing of new methods, ideally in user studies that resemble real-world conditions—for example, by having participants bring their own data.

Future work could also involve human-subjects researchers in the design of DP methods specifically. Advocates should not only demonstrate the trade-offs of DP repeatedly and rigorously across many kinds of data (including small sample datasets) and future research use cases (including complex multivariate analyses), but also guide researchers through the implications for their work. Some have taken early steps, evaluating DP for replication of randomized controlled trials [67] and for redistricting using Census data [18].

User-centered design can also help privacy experts create de-identification tooling that is not only acceptable to practitioners, but also *usable* and *useful*—crucial qualities for complex tools to be adopted [60]. Tools should incorporate functions that researchers already use tools for, such as generating in-depth summaries of variables to guide de-identification. These functions can be built out—e.g., summarizing crosstabs or providing metrics of risk and utility—but not at the cost of eliminating their initial usefulness.

6 Conclusion

Research data publication is becoming an expectation and requirement. In 24 interviews, we studied how researchers approach the complex task of de-identifying human-subjects data. We find that while lessons of de-identification research have percolated into various fields of research practice, leading to widespread awareness of concepts related to *k*-anonymity (indirect identifiers, singling out, etc.), researchers are largely not following systematic processes that ensure protection against basic re-identification threats such as singling out. Collaborative work bridging this gap has the potential to improve privacy for research data subjects, usability for researchers, and quality for research broadly.

7 Ethics considerations

The University of Maryland Institutional Review Board approved this study. Participants gave informed consent, including for automated transcription. We limit personal information in this paper to lower the risk of re-identification. We did not attempt to re-identify any published data.

8 Open science

Our research protocols are in Appendices B and C. Our qualitative codebook, recruitment message templates, and consent form are in our supplementary materials at https://osf.io/4tgpv/. We have not published interview transcripts, due to the risk they pose if participants or their organizations are re-identified. In particular, participants sometimes spoke critically of their peers and organizations, which could have negative repercussions on their employment and career.

Acknowledgments

This work was supported in part by NSF grant OAC-2232863. NORC at the University of Chicago also provided funding to support the time dedicated to this project. We are grateful to Jason Walter for helping to clean interview transcripts. We also thank Nathan Reitinger, Wellington Barbosa, and Simson Garfinkel for providing valuable insights on this project.

References

[1] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *IEEE Cybersecurity Development Conference*, November 2016. https://doi.org/10.1109/SecDev.2016.013.

[2] Ahmet Aktay, Shailesh Bavadekar, Gwen Cossoul, John Davis, Damien Desfontaines, Alex Fabrikant, Evgeniy Gabrilovich, Krishna Gadepalli, Bryant Gipson, Miguel Guevara, Chaitanya Kamath, Mansi Kansal, Ali Lange, Chinmoy Mandayam, Andrew Oplinger, Christopher Pluntke, Thomas Roessler, Arran Schlosberg, Tomer Shekel, Swapnil Vispute, Mia Vu, Gregory Wellenius, Brian Williams, and Royce J. Wilson. Google COVID-19 Community Mobility Reports: Anonymization process description (version 1.1). https://doi.org/10.48550/arXiv.2004.04145, November 2020.

[3] Olivia Angiuli, Joe Blitzstein, and Jim Waldo. How to de-identify your data: Balancing statistical accuracy and subject privacy in large social-science data sets. *Queue*, 13(8), 2015. https://doi.org/10.1145/2838344.2838930.

[4] Apple Differential Privacy Team. Learning with privacy at scale. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

[5] Thijs Benschop and Matthew Welch. sdcApp manual. https://sdcappdocs.readthedocs.io/en/latest/, May 2023.

[6] Raffael Bild, Klaus A. Kuhn, and Fabian Prasser. SafePub: A truthful data anonymization algorithm with strong privacy guarantees. *PoPETs*, 2018(1), 2018. https://doi.org/10.1515/popets-2018-0004.

[7] Bill & Melinda Gates Foundation. Open Access policy. https://openaccess.gatesfoundation.org/, 2024.

[8] Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Computing Surveys*, 55(8), December 2022. https://doi.org/10.1145/3547139.

[9] Michelle Boorstein, Marisa Iati, and Annys Shin. Top U.S. Catholic Church official resigns after cellphone data used to track him on Grindr and to gay bars. *The Washington Post*, July 2021. https://www.washingtonpost.com/religion/2021/07/20/bishop-misconduct-resign-burrill/.

[10] danah boyd and Jayshree Sarathy. Differential perspectives: Epistemic disconnects surrounding the U.S. Census Bureau's use of differential privacy. *Harvard Data Science Review*, Special Issue 2, June 2022. https://hdsr.mitpress.mit.edu/pub/3vj5j6i0/release/3.

[11] Rebecca Campbell, Rachael Goodman-Williams, and McKenzie Javorka. A trauma-informed approach to sexual violence research ethics and open science. *Journal of Interpersonal Violence*, 34(23–24), 2019. https://doi.org/10.1177/0886260519871530.

[12] Rebecca Campbell, McKenzie Javorka, Jasmine Engleton, Kathryn Fishwick, Katie Gregory, and Rachael Goodman-Williams. Open-science guidance for qualitative research: An empirically validated approach for de-identifying sensitive narrative data. *Advances in Methods and Practices in Psychological Science*, 6(4), 2023. https://doi.org/10.1177/25152459231205832.

[13] Loredana Caruccio, Domenico Desiato, Giuseppe Polese, Genoveffa Tortora, and Nicola Zannone. A decision-support framework for data anonymization with application to machine learning processes. *Information Sciences*, 613, October 2022. https://doi.org/10.1016/j.ins.2022.09.004.

[14] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luís Antunes. Survey on privacy-preserving techniques for microdata publication. *ACM Computing Surveys*, 55(14s), July 2023. https://doi.org/10.1145/3588765.

[15] Phaik Yeong Cheah, Decha Tangseefa, Aimatcha Somsaman, Tri Chunsuttiwat, François Nosten, Nicholas P. J. Day, Susan Bull, and Michael Parker. Perceived benefits, harms, and views about how to share data responsibly: A qualitative study of experiences with and attitudes toward data sharing among research staff and community representatives in Thailand. *Journal of Empirical Research on Human Research Ethics*, 10(3), 2015. https://doi.org/10.1177/1556264615592388.

[16] Miranda Christ, Sarah Radway, and Steven M. Bellovin. Differential privacy and swapping: Examining de-identification's impact on minority representation and privacy preservation in the U.S. Census. In *IEEE S&P '22*, May 2022. https://doi.org/10.1109/SP46214.2022.9833668.

[17] Aloni Cohen. Attacks on deidentification's defenses. In *USENIX Security '22*, August 2022. https://www.usenix.org/conference/usenixsecurity22/presentation/cohen.

[18] Aloni Cohen, Moon Duchin, JN Matthews, and Bhushan Suwal. Private numbers in public policy: Census, differential privacy, and redistricting. *Harvard Data Science Review*, Special Issue 2, 2022. https://doi.org/10.1162/99608f92.22fd8a0e.

[19] Louise Corti, Vernon Gayle, Jon Johnson, Myles Offord, Cristina Magder, and Anca Vlad. QAMyData. UK Data Service, 2020. https://ukdataservice.ac.uk/about/research-and-development/past-projects/qamydata/.

[20] Damien Desfontaines and Balázs Pejó. SoK: Differential privacies. *PoPETs*, 2020(2), 2020. https://doi.org/10.2478/popets-2020-0028.

[21] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Neural Information Processing Systems*, 2017. https://www.microsoft.com/en-us/research/publication/collecting-telemetry-data-privately/.

[22] Kai Dong, Zheng Zhang, Chuang Jia, Zhen Ling, Ming Yang, Junzhou Luo, and Xinwen Fu. Relation mining under local differential privacy. In *USENIX Security '24*, August 2024. https://www.usenix.org/

conference/usenixsecurity24/presentation/dong-kai.

[23] Wei Dong, Qiyao Luo, Giulia Fanti, Elaine Shi, and Ke Yi. Almost instance-optimal clipping for summation problems in the shuffle model of differential privacy. In *ACM CCS '24*, October 2024. https://doi.org/10.1145/3658644.3690225.

[24] Gabrielle Drake. The ethical and methodological challenges of social work research with participants who fear retribution: To 'do no harm'. *Qualitative Social Work*, 13(2), 2014. https://doi.org/10.1177/1473325012473499.

[25] Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, Lecture Notes in Computer Science, 2006. https://doi.org/10.1007/11787006_1.

[26] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2), October 2019. https://doi.org/10.29012/jpc.689.

[27] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4), 2014. https://doi.org/10.1561/0400000042.

[28] Khaled El Emam. *Guide to the De-Identification of Personal Health Information*. Auerbach Publications, April 2013. https://doi.org/10.1201/b14764.

[29] Shuya Feng, Meisam Mohammady, Han Wang, Xiaochen Li, Zhan Qin, and Yuan Hong. DPI: Ensuring strict differential privacy for infinite data streaming. In *IEEE S&P '24*, May 2024. https://doi.org/10.1109/SP54263.2024.00124.

[30] Mohamed R. Fouad, Khaled Elbassioni, and Elisa Bertino. A supermodularity-based differential privacy preserving algorithm for data anonymization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7), July 2014. https://doi.org/10.1109/TKDE.2013.107.

[31] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security '14*, August 2014. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew.

[32] Simson L. Garfinkel, John M. Abowd, and Sarah Powazek. Issues encountered deploying differential privacy. In *WPES '18*, October 2018. https://doi.org/10.1145/3267323.3268949.

[33] Simson L. Garfinkel and Philip Leclerc. Randomness concerns when deploying differential privacy. In *WPES '20*, September 2020. https://doi.org/10.1145/3411497.3420211.

[34] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough?: An experiment with data saturation and variability. *Field Methods*, 18(1), February 2006. https://doi.org/10.1177/1525822X05279903.

[35] Wentao Guo, Aditya Kishore, Adam J Aviv, and Michelle L Mazurek. A qualitative analysis of practical de-identification guides. In *ACM CCS '24*, October 2024. https://doi.org/10.1145/3658644.3690270.

[36] Kashmir Hill. What happens when our faces are tracked everywhere we go? *The New York Times*, March 2021. https://www.nytimes.com/interactive/2021/03/18/magazine/facial-recognition-clearview-ai.html.

[37] Shlomi Hod and Ran Canetti. Differentially private release of Israel's National Registry of Live Births. In *IEEE S&P '25*, May 2025. https://doi.org/10.1109/SP61157.2025.00101.

[38] John P. Holdren. Increasing access to the results of federally funded scientific research. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf, February 2013.

[39] Naoise Holohan, Stefano Braghin, and Mohamed Suliman. Securing floating-point arithmetic for noise addition. In *ACM CCS '24*, October 2024. https://doi.org/10.1145/3658644.3690347.

[40] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. SoK: Privacy-preserving data synthesis. In *IEEE S&P '24*, May 2024. https://doi.org/10.1109/SP54263.2024.00002.

[41] Anco Hundepool, Ramya Ramaswamy, Peter-Paul de Wolf, Luisa Franconi, Ruth Brand, and Josep Domingo. μ-ARGUS. https://research.cbs.nl/casc/mu.htm, July 2021.

[42] ICPSR. https://www.icpsr.umich.edu/web/pages/ICPSR/index.html, 2025.

[43] Jacob Imola, Amrita Roy Chowdhury, and Kamalika Chaudhuri. Metric differential privacy at the user-level via the earth-mover's distance. In *ACM CCS '24*, October 2024. https://doi.org/10.1145/3658644.3690363.

[44] Japan Science and Technology Agency. JST policy on open access to research publications and research data management. https://www.jst.go.jp/EN/about/openscience/policy_openscience_en_r4.pdf, April 2022.

[45] Tianxi Ji and Pan Li. Less is more: Revisiting the Gaussian mechanism for differential privacy. In *USENIX Security '24*, August 2024. https://www.usenix.org/conference/usenixsecurity24/presentation/ji.

[46] Bo Jiang, Jian Du, Sagar Sharma, and Qiang Yan. Budget recycling differential privacy. In *IEEE S&P '24*, May 2024. https://doi.org/10.1109/SP54263.2024.00212.

[47] Jiankai Jin, Chitchanok Chuengsatiansup, Toby Murray, Benjamin I. P. Rubinstein, Yuval Yarom, and Olga Ohrimenko. Elephants do not forget: Differential privacy with state continuity for privacy budget. In *ACM CCS '24*, October 2024. https://doi.org/10.1145/3658644.3670281.

[48] Samantha Joel, Paul W. Eastwick, and Eli J. Finkel. Open sharing of data on close relationships and other sensitive social psychological topics: Challenges, tools, and future directions. *Advances in Methods and Practices in Psychological Science*, 1(1), 2018. https://doi.org/10.1177/2515245917744281.

[49] Daniel Kifer, John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census. http://arxiv.org/abs/2209.03310, September 2022.

[50] Nigel King and Joanna M. Brooks. *Template Analysis for Business and Management Students*. SAGE, 2017.

[51] Nicolas Küchler, Emanuel Opel, Hidde Lycklama, Alexander Viand, and Anwar Hithnawi. Cohere: Managing differential privacy in large scale systems. In *IEEE S&P '24*, May 2024. https://doi.org/10.1109/SP54263.2024.00122.

[52] Nada Lahjouji, Sameera Ghayyur, Xi He, and Sharad Mehrotra. ProBE: Proportioning privacy budget for complex exploratory decision support. In *ACM CCS '24*, October 2024. https://doi.org/10.1145/3658644.3670394.

[53] Eric S. Lander. Public access congressional report. https://www.whitehouse.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf, November 2021.

[54] Hyukki Lee and Yon Dohn Chung. Differentially private release of medical microdata: An efficient and practical approach for preserving informative attribute values. *BMC Medical Informatics and Decision Making*, 20, 2020. https://doi.org/10.1186/s12911-020-01171-5.

[55] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *IEEE International Conference on Data Engineering*, April 2007. https://doi.org/10.1109/ICDE.2007.367856.

[56] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), March 2007. https://doi.org/10.1145/1217299.1217302.

[57] Abdul Majeed and Seong Oun Hwang. A practical anonymization approach for imbalanced datasets. *IT Professional*, 24(1), 2022. https://doi.org/10.1109/MITP.2021.3132330.

[58] Annette N. Markham. Fabrication as ethical practice: Qualitative inquiry in ambiguous internet contexts. *Information, Communication, & Society*, 15(3), 2012. https://doi.org/10.1080/1369118X.2011.641993.

[59] Alishah Mawji, Holly Longstaff, Jessica Trawin, Dustin Dunsmuir, Clare Komugisha, Stefanie K. Novakowski, Matthew O. Wiens, Samuel Akech, Abner Tagoola, Niranjan Kissoon, and J. Mark Ansermino. A proposed de-identification framework for a cohort of children presenting at a health facility in Uganda. *PLOS Digital Health*, 1(8), August 2022. https://doi.org/10.1371/journal.pdig.0000027.

[60] Susan E. McGregor, Elizabeth Anne Watkins, Mahdi Nasrullah Al-Ameen, Kelly Caine, and Franziska Roesner. When the weakest link is strong: Secure collaboration in the case of the Panama Papers. In *USENIX Security '17*, August 2017. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/mcgregor.

[61] Fredrik Meisingseth and Christian Rechberger. SoK: Computational and distributed differential privacy for MPC. *PoPETs*, 2025(1), 2025. https://doi.org/10.56553/popets-2025-0023.

[62] Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins. Facebook privacy-protected full URLs data set. https://doi.org/10.7910/DVN/TDOAPG/DGSAMS, June 2020.

[63] Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson, and Nico Grant. How tech giants cut corners to harvest data for A.I. *The New York Times*, April 2024. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

[64] Disclosure Review Board. https://www.mcc.gov/about/org-unit/disclosure-review-board/.

[65] Àlex Miranda-Pascual, Patricia Guerra-Balboa, Javier Parra-Arnau, Jordi Forné, and Thorsten Strufe. SoK: Differentially private publication of trajectory data. *PoPETs*, 2023(2), 2023. https://doi.org/10.56553/popets-2023-0065.

[66] Kirsten N. Morehouse, Benedek Kurdi, and Brian A. Nosek. Responsible data sharing: Identifying and remedying possible re-identification of human participants. *American Psychologist*, 2024. https://doi.org/10.1037/amp0001346.

[67] Soumya Mukherjee, Aratrika Mustafi, Aleksandra Slavković, and Lars Vilhuber. Assessing utility of differential privacy for RCTs. http://arxiv.org/abs/2309.14581, September 2023.

[68] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. Visualizing privacy-utility trade-offs in differentially private data releases. *PoPETs*, 2022(2), April 2022. https://doi.org/10.2478/popets-2022-0058.

[69] Priyanka Nanayakkara and Jessica Hullman. What's driving conflicts around differential privacy for the U.S. Census. *IEEE Security & Privacy*, 21(5), 2023. https://doi.org/10.1109/MSEC.2022.3202793.

[70] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE S&P '08*, May 2008. https://doi.org/10.1109/SP.2008.33.

[71] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets: A decade later. https://www.cs.princeton.edu/~arvindn/publications/de-anonymization-retrospective.pdf, May 2019.

[72] National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Research Data and Information, and Committee on Toward an Open Science Enterprise. *Open Science by Design: Realizing a Vision for 21st Century Research*. National Academies Press (US), July 2018. https://www.ncbi.nlm.nih.gov/books/NBK525421/.

[73] Alondra Nelson. Ensuring free, immediate, and equitable access to federally funded research. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf, August 2022.

[74] Boel Nelson and Jenni Reuben. SoK: Chasing accuracy and privacy, and catching both in differentially private histogram publication. *Transactions on Data Privacy*, 13(3), December 2020. https://www.tdp.cat/issues16/abs.a387a20.php.

[75] Ivoline C. Ngong, Brad Stenger, Joseph P. Near, and Yuanyuan Feng. Evaluating the usability of differential privacy tools with data practitioners. In *SOUPS '24*, August 2024. https://www.usenix.org/conference/soups2024/presentation/ngong.

[76] Office of the Chief Science Advisor of Canada. Roadmap for open science. https://science.gc.ca/site/science/en/office-chief-science-advisor/open-science/roadmap-open-science, February 2020.

[77] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57, 2010. https://papers.ssrn.com/abstract=1450006.

[78] OpenAIRE. Amnesia. https://amnesia.openaire.eu/, 2023.

[79] OpenDP Team. The OpenDP white paper. https://projects.iq.harvard.edu/files/opendifferentialprivacy/files/opendp_white_paper_11may2020.pdf, May 2020.

[80] Heinz Pampel, Nina Leonie Weisweiler, Dorothea Strecker, Michael Witt, Paul Vierkant, Kirsten Elger, Roland Bertelmann, Matthew Buys, Lea Maria Ferguson, Maxi Kindling, Rachael Kotarski, and Vivien Petras. re3data – indexing the global research data repository landscape since 2012. *Scientific Data*, 10, 2023. https://doi.org/10.1038/s41597-023-02462-y.

[81] Emma I. C. Peterson, Valerie Zhao, Dan Byrne, and Blase Ur. MARI: Semi-automated, human-in-the-loop redaction of text corpora. In *SOUPS '23*, August 2023. https://www.usenix.org/conference/soups2023/presentation/peterson-poster.

[82] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. Flexible data anonymization using ARX—current status and challenges ahead. *Software: Practice and Experience*, 50(7), July 2020. https://doi.org/10.1002/spe.2812.

[83] René Raab, Pascal Berrang, Paul Gerhart, and Dominique Schröder. SoK: Descriptive statistics under local differential privacy. *PoPETs*, 2025(1), 2025. https://doi.org/10.56553/popets-2025-0008.

[84] Zachary Ratliff and Salil Vadhan. A framework for differential privacy against timing attacks. In *ACM CCS '24*, October 2024. https://doi.org/10.1145/3658644.3690206.

[85] Ryan Rogers, Subbu Subramaniam, Sean Peng, David Durfee, Seunghyun Lee, Santosh Kumar Kancha, Shraddha Sahay, and Parvez Ahammad. LinkedIn's Audience Engagements API: A privacy preserving data analytics system at scale. https://doi.org/10.48550/arXiv.2002.05839, November 2020.

[86] Jayshree Sarathy and danah boyd. Statistical imaginaries, state legitimacy: Grappling with the arrangements underpinning quantification in the U.S. Census. *Critical Sociology*, 2024. https://doi.org/10.1177/08969205241270898.

[87] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. Don't look at the data! How differential privacy reconfigures the practices of data science. In *CHI '23*, April 2023. https://doi.org/10.1145/3544548.3580791.

[88] M. Angela Sasse and Ivan Flechais. Usable security: Why do we need it? How do we get it? In Lorrie Faith Cranor and Simson Garfinkel, editors, *Security and Usability: Designing Secure Systems That People Can Use*. O'Reilly, 2005. https://discovery.ucl.ac.uk/id/eprint/20345/.

[89] Bruce Schneier. The psychology of security. In *AFRICACRYPT '08*, 2008. https://doi.org/10.1007/978-3-540-68164-9_5.

[90] Latanya Sweeney. Simple demographics often identify people uniquely. https://dataprivacylab.org/projects/identifiability/, 2000.

[91] Latanya Sweeney. *k*-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05), October 2002. https://doi.org/10.1142/S0218488502001648.

[92] Matthias Templ, Bernhard Meindl, Alexander Kowarik, Johannes Gussenbauer, Organisation For Economic Co-Operation And Development, Statistics Netherlands, and Pascal Heus. sdcMicro. https://cloud.r-project.org/web/packages/sdcMicro/index.html, March 2024.

[93] Joana Tomás, Deolinda Rasteiro, and Jorge Bernardino. Data anonymization: An experimental evaluation using open-source tools. *Future Internet*, 14(6), 2022. https://doi.org/10.3390/fi14060167.

[94] TOP Factor. All journals. https://topfactor.org/journals?factor=Data+Transparency.

[95] Michael Carl Tschantz, Shayak Sen, and Anupam Datta. SoK: Differential privacy as a causal property. In *IEEE S&P '20*, May 2020. https://doi.org/10.1109/SP40000.2020.00012.

[96] UK Research and Innovation. UKRI open access policy. https://www.ukri.org/publications/ukri-open-access-policy/, November 2023.

[97] U.S. Census Bureau. Disclosure avoidance for the 2020 Census: An introduction. https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html, November 2021.

[98] U.S. National Science and Technology Council. Desirable characteristics of data repositories for federally funded research. https://doi.org/10.5479/10088/113528, May 2022.

[99] Veerle Van den Eynden and Louise Corti. Advancing research data publishing practices for the social sciences: From archive activity to empowering researchers. *International Journal on Digital Libraries*, 18, 2017. https://doi.org/10.1007/s00799-016-0177-3.

[100] Wellcome. Open access policy. https://wellcome.org/grant-funding/guidance/open-access-guidance/open-access-policy, January 2024.

[101] White House Office of Science and Technology Policy. The White House Office of Science & Technology Policy Open Science Recognition Challenge. https://www.challenge.gov/?challenge=ostp-year-of-open-science-recognition-challenge.

[102] Michael Wines. The 2020 Census suggests that people live underwater. There's a reason. *The New York Times*, April 2022. https://www.nytimes.com/2022/04/21/us/census-data-privacy-concerns.html.

[103] World Health Organization. Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance. https://www.who.int/publications/i/item/9789240044968, April 2022.

[104] Kai Zhang, Yanjun Zhang, Ruoxi Sun, Pei-Wei Tsai, Muneeb Ul Hassan, Xin Yuan, Minhui Xue, and Jinjun Chen. Bounded and unbiased composite differential privacy. In *IEEE S&P '24*, May 2024. https://doi.org/10.1109/SP54263.2024.00108.

A Pre-interview questionnaire

Participant demographics

1. What is your age?
   - 18–24 | 25–34 | 35–44 | 45–54 | 55–64 | 65+ | Prefer to self-describe ________ | Prefer not to state
2. What is your gender?
   - Male | Female | Non-binary | Prefer to self-describe ________ | Prefer not to state
3. What is your race? (You can select more than one option.)
   - American Indian or Alaska Native | Asian | Black or African American | Hispanic or Latino | Native Hawaiian or Pacific Islander | White | Prefer to self-describe ________ | Prefer not to state
4. What is the highest level of formal education that you have completed or are currently pursuing?
   - No high school degree | High school graduate, diploma, or equivalent (for example, GED) | Trade, technical, or vocational training | Associate's degree | Bachelor's degree | Graduate or professional degree | Prefer to self-describe ________ | Prefer not to state
5. What country (or countries) do you live in?
   - United States | Other ________ | Prefer not to state

Organization details (practitioners only)

1. Which of the following best describes the organization where you de-identify data?
   - Government agency | Academic institution | Healthcare organization | Research institute | Private company | Nonprofit | Prefer to self-describe ________
2. What is your role (i.e., job title) at your organization? ________
3. Approximately how many people at your organization are involved in de-identifying data at least occasionally? ________
4. Approximately how many employees work at your organization in total? ________
5. As part of this project, we are also analyzing practical guides that explain how to de-identify data. If you think of any that are important for us to include, please consider sharing names, links, or files here.[5] ________
6. Is there anything else we should know before we begin this interview? ________

Organization details (curators only)

1. Does your organization maintain a repository where datasets are published?
   - Yes | No | Prefer to self-describe ________
2. Does your organization review datasets that it contracted or funded (i.e., you are the client)?
   - Yes, all datasets we review are contracted or funded by us | Yes, some of the datasets we review are contracted or funded by us | No, never | Prefer to self-describe ________
3. What is your role (i.e., job title) at your organization? ________
4. Approximately how many people at your organization are involved in reviewing de-identified data at least occasionally? ________
5. Approximately how many employees work at your organization in total? ________
6. As part of this project, we are also analyzing practical guides that explain how to de-identify data. If you think of any that are important for us to include, please consider sharing names, links, or files here. ________
7. Is there anything else we should know before we begin this interview? ________

B Practitioner interview protocol

Background
I'll start off with some basic definitions, since sometimes these terms mean different things to different people. We are interested in how you de-identify datasets containing personal information. By de-identify, we mean modifying data (or the interface for viewing it) to make it harder to identify who the data refers to. There are many ways to do this, but it usually includes removing or masking direct identifiers like name and phone number. It also often includes modifying indirect identifiers such as zip code and age. When the goal is to minimize the re-identification risk, this is sometimes called anonymization. Is there anything I can clarify about the scope of this interview?

Let's start with some brief questions about what you do.

- At a high level, could you please describe the kind of datasets that you've de-identified?
- To the extent that you are aware, what are these datasets used for after de-identification? (By whom?)
- Are you required to comply with any regulations or guidelines relating to de-identification?
- Approximately how many datasets have you de-identified?
  - ○ Note: a dataset should be counted separately if it requires reassessing how de-identification should be done (e.g., which variables to remove, what buckets to use for generalization).
- How did you learn how to de-identify data?

[5] This information was incorporated into a separate study [35].

Process

Now, let's get into the process of de-identification.

- Think back to one of the more typical datasets you have de-identified. Could you please walk me through the process of de-identifying this specific dataset?
  - Who was involved, and how did they fit into the timeline?
  - What techniques did you use?
  - You mentioned [using *x* de-identification technique]. How do you decide [see examples below]?
    - *k*-anonymity: what value of *k* to use, or what counts as a quasi-identifier
    - Differential privacy: what value of epsilon to use
    - Generalization: how to pick categories (including whether there's a threshold for when it's generalized enough), or which variables to generalize
    - Suppression: which values/records/variables to delete
    - Swapping: which records to swap values between
    - Adding noise: what data should be changed and how
- To what extent is this process different for other datasets?
- If you and someone else at your organization or in your research group were to de-identify the same dataset independently, would they end up doing it the same way that you do?

Evaluation

- Is there anything you do to determine when a dataset has been adequately de-identified? (Are there any specific tests or metrics?)
- Have you ever received feedback on a dataset you de-identified? (What and from whom? Did you need to make any changes?)
- Have you ever de-identified a dataset too much or too little? For example, so much so that it was no longer useful, or so little that you had to change or redo something to protect people's privacy?
  - More generally, how do you think about balancing privacy and utility?
- Has your de-identification process changed over time? (How, why?)

Challenges

- Are there aspects of de-identifying data that you find frustrating or challenging?
- Have you used tools or scripts that help you de-identify data?
  - Are these helpful and easy to use?
  - Do you have any suggestions for how to improve these, or suggestions for new tools or resources to make the de-identification process easier?

Threat modeling

Now, I'm going to shift our conversation towards your perceptions of threats to de-identified data.

- Do you have a process for assessing the risk posed by the release or use of de-identified data? This is sometimes called threat modeling.
- For the datasets you have de-identified, do you think there would be significant risks to data subjects if they were re-identified, and if so, what are they?
- In your opinion, what is the likelihood that someone would actually try to re-identify anyone in the datasets you've de-identified? (Who?)
- Let's imagine a variety of potential threat actors who range in technical sophistication, resources, and motivation. An example of a weaker threat is someone casually looking at a dataset who happens to recognize an individual, while a stronger threat might be an intelligence agency investigating a person of interest. How well do you think datasets you de-identify are protected against different kinds of threat actors?
  - *[if not protected from weaker threats]* If de-identification doesn't protect against weaker threats, what do you see as the point of it?
  - *[if differences between threats]* What do you think stronger threats could do that weaker threats wouldn't?
  - *[if not protected from stronger threats]* Do you think it's feasible for de-identified data to be protected against stronger threats and also still be useful? (Do you know how you would attempt to do so?)
  - *[if protected from stronger threats]* Could you tell me more about what gives you confidence that the data you de-identify is protected against these threats?
  - *[if uncertain]* Do people at your organization and in your professional circle talk about specific threat actors and how de-identification may or may not protect against them? (In these conversations, who are the threats, and how well does de-identification protect against them?)
- If there was a significantly [higher/lower] risk that someone would try to re-identify any of the datasets you've de-identified, would it change the way in which you do de-identification? (What's the first thing you would change?)

Perturbation

Before we wrap up, I want to spend some time talking about a specific family of de-identification techniques that involve introducing errors to the data. Instead of deleting data or generalizing it into broader categories, some techniques make it so that data about individuals is incorrect, for example by swapping values between rows, or by adding random changes to values. Terms used to describe these techniques include data perturbation, generating synthetic data, and differential privacy.

- *[if they've mentioned using perturbative techniques]* You mentioned that you use error-introducing techniques at least sometimes. How important are these techniques as part of your de-identification toolkit? When are they appropriate or inappropriate to use?
- *[else]* Do you ever use any techniques that introduce error?
  - *[if yes]* How important are these techniques as part of your de-identification toolkit? When are they appropriate or inappropriate to use?
  - *[if no]* Is there a reason you've avoided using these techniques? When might they be appropriate or inappropriate to use?
- What do you see as the barriers, if any, that keep you from using techniques that introduce error to a greater extent?

Conclusion

- [if no clear answer in the pre-questionnaire] In the questionnaire, you didn't mention any how-to guides that you think are important for us to analyze. That's totally fine—just wanted to check if you thought of any since then.
- Before we wrap up, is there anything else you think we should know about your work de-identifying data?

C Curator interview protocol

Background

I'll start off with some basic definitions, since sometimes these terms mean different things to different people. We are interested in how you de-identify datasets containing personal information. By de-identify, we mean modifying data (or the interface for viewing it) to make it harder to identify who the data refers to. There are many ways to do this, but it usually includes removing or masking direct identifiers like name and phone number. It also often includes modifying indirect identifiers such as zip code and age. When the goal is to minimize the re-identification risk, this is sometimes called anonymization. Is there anything I can clarify about the scope of this interview?

Let's start with some brief questions about what you do.

- At a high level, could you please describe the kind of datasets that you review?
- To the extent that you are aware, what are these datasets used for after de-identification? (By whom?)
- Are reviewed datasets required to comply with any regulations or guidelines relating to de-identification?
- Approximately how many datasets have you reviewed de-identification for?
  - Note: a dataset should be counted separately if it requires reassessing how de-identification should be done (e.g., which variables to remove, what buckets to use for generalization).

- Do you have experience de-identifying data yourself?
- How did you learn about how data should be de-identified?

Process
Now, let's get into how you review de-identification.

- Think back to one of the more typical datasets you have reviewed. Could you please walk me through the process of reviewing this specific dataset?
    - Who was involved, and how did they fit into the timeline?
    - What information do you receive from submitters that helps you review their de-identification?
- Do you provide feedback to submitters? (What feedback?)
    - Have you asked submitters to change their de-identification? (Have they agreed, and what happens if they disagree?)

Expectations

- Do you have specific expectations for how data should be de-identified?
- Are there any techniques you expect or recommend against?
- You mentioned [*x* de-identification technique]. How do you decide [see examples below]?
    - *k*-anonymity: what value of *k* is appropriate, or what counts as a quasi-identifier
    - Differential privacy: what value of epsilon should be used
    - Generalization: how categories should be picked (including whether there should be a threshold for when it's generalized enough), or which variables to generalize
    - Suppression: which values/records/variables should be deleted
    - Swapping: which records should have values swapped
    - Adding noise: what data should be changed and how

Evaluation

- How do you determine when a dataset has been adequately de-identified? (Are there any specific tests or metrics?)
- How do you decide what access level a dataset should be published at—if there are options for different access levels?
    - Have you ever decided that a dataset should not be published at all for reasons relating to de-identification?
- Have you ever reviewed a dataset that was de-identified too much or too little? For example, so much so that it was no longer useful, or so little that something needed to be changed or redone to protect people's privacy?
    - More generally, how do you think about balancing privacy and utility?
- What are the most common or egregious de-identification mistakes that appear in data you review?
    - If you could give just one piece of advice to submitters before they start de-identifying data, what would you recommend?
- After a dataset is published, do you ever do any follow-up review of its de-identification?

Consistency and change

- To what extent is your review process different depending on the dataset?
- If you and someone else at your organization were to review the same dataset independently, would they end up doing it the same way that you do?
- Has your process for reviewing de-identification changed over time? (How, why?)

Challenges and tools/resources

- Are there aspects of reviewing de-identification that you find frustrating or challenging?
- Have you used tools or scripts that help you review de-identification?
    - Are these helpful and easy to use?
    - Do you have any suggestions for how to improve these, or suggestions for new tools or resources to make the review process easier?

Threat modeling
Now, I'm going to shift our conversation towards your perceptions of threats to de-identified data.

- Do you have a process for assessing the risk posed by the release or use of de-identified data? This is sometimes called threat modeling.
- For the datasets you have reviewed, do you think there would be significant risks to data subjects if they were re-identified, and if so, what are they?
- In your opinion, what is the likelihood that someone would actually try to re-identify anyone in the datasets you've reviewed? (Who?)
- Let's imagine a variety of potential threat actors who range in technical sophistication, resources, and motivation. An example of a weaker threat is someone casually looking at a dataset who happens to recognize an individual, while a stronger threat might be an intelligence agency investigating a person of interest. How well do you think datasets you approve are protected against different kinds of threat actors?
    - *[if not protected from weaker threats]* If de-identification doesn't protect against weaker threats, what do you see as the point of it?
    - *[if differences between threats]* What do you think stronger threats could do that weaker threats wouldn't?
    - *[if not protected from stronger threats]* Do you think it's feasible for de-identified data to be protected against stronger threats and also still be useful? (Do you know how you would attempt to do so?)
    - *[if protected from stronger threats]* Could you tell me more about what gives you confidence that the data you review is protected against these threats?
    - *[if uncertain]* Do people at your organization and in your professional circle talk about specific threat actors and how de-identification may or may not protect against them? (In these conversations, who are the threats, and how well does de-identification protect against them?)

Perturbation
Before we wrap up, I want to spend some time talking about a specific family of de-identification techniques that involve introducing errors to the data. Instead of deleting data or generalizing it into broader categories, some techniques make it so that data about individuals is incorrect, for example by swapping values between rows, or by adding random changes to values. Terms used to describe these techniques include data perturbation, generating synthetic data, and differential privacy.

- In your opinion, when are these techniques appropriate or inappropriate to use?
- What do you see as the barriers, if any, that keep these techniques from being used to a greater extent?

Conclusion

- [if no clear answer in the pre-questionnaire] In the questionnaire, you didn't mention any how-to guides that you think are important for us to analyze. That's totally fine—just wanted to check if you thought of any since then.
- Before we wrap up, is there anything else you think we should know about your work reviewing de-identification?

D Interview codebook

Our codebook is in the supplementary material: https://osf.io/4tgpv/

E Participant demographics

In this appendix only, we disaggregate the three participants collectively referred to as C3 (see footnote to Table 1). Thirteen participants are men, and thirteen are women. Nineteen are White, two are Asian, one is Black or African American, one is Hispanic or Latino, one is Middle Eastern or North African, one identified as mixed, and one did not disclose. Two are ages 18–24, six are 25–34, twelve are 35–44, five are 45–54, one is 55–64, and one did not disclose. The highest level of formal education achieved or in progress is graduate or professional degree for twenty-four, and bachelor's degree for two. Twenty-four live in the U.S., and two live in the UK.