Going Beyond Obscurity: Organizational Approaches to Data Anonymization

VIKTOR HARGITAI, University of Copenhagen, Denmark
IRINA SHKLOVSKI, IT University of Copenhagen, Denmark
ANDRZEJ WĄSOWSKI, IT University of Copenhagen, Denmark

Anonymization is viewed as a solution to over-exposure of personal information in a data-driven society. Yet how organizations apply anonymization techniques to data remains under-explored. We investigate how such measures are applied in organizations through a small-scale interview study, asking whether anonymization practices are used, what approaches are considered practical and adequate, and how decisions are made to protect the privacy of data subjects while preserving analytical value. Our findings suggest that applying anonymization is a complex socio-technical process and it is less common than we had originally expected. Organizations that employ anonymization often view their practices as sensitive and resort to anonymity by obscurity alongside technical means. Rather than being a purely technical question of applying the right algorithms, anonymization in practice relies on multi-stakeholder collaborations across organizational boundaries. Here we observe engagement of what Suchman has called "multiple, located, partial perspectives" in ongoing discussions necessary to negotiate the technical as well as organizational complexities of implementing data anonymization.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**;

Keywords: Anonymization; decision making; GDPR; located accountabilities; organizational practices; privacy

ACM Reference Format:
Viktor Hargitai, Irina Shklovski, and Andrzej Wąsowski. 2018. Going Beyond Obscurity: Organizational Approaches to Data Anonymization. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2, CSCW, Article 66 (November 2018). ACM, New York, NY. 22 pages. https://doi.org/10.1145/3274335

1 INTRODUCTION

In 2016 Apple announced that they would use differential privacy, an advanced data anonymization approach, to protect the privacy of their users while analyzing their behavior. Reactions to this announcement were overwhelmingly positive [20]. However, Apple only granted a brief glimpse of their methods to an independent researcher, and no information was published about the details of their implementation. Barely a year later Tang et al. [49] published results from a painstaking study of the actual implementation of differential privacy in the MacOS Sierra operating system, arguing that Apple's approach may not be as effective as was originally suggested and urging Apple to make the details of their implementations across the various operating systems, devices and services transparent. Tang et al. [49] argued that Apple used noise parameters that were too large for the differential privacy implementation, rendering the process ineffective because the

Authors' addresses: Viktor Hargitai, University of Copenhagen, Denmark, viktorharg@gmail.com; Irina Shklovski, IT University of Copenhagen, Denmark, irsh@itu.dk; Andrzej Wąsowski, IT University of Copenhagen, Denmark, wasowski@itu.dk.

2573-0142/2018/11-ART66 $15.00
https://doi.org/10.1145/3274335

probability of de-anonymization becomes unacceptably high. The subsequent debate acknowledged the complexity of the approach and many argued that transparency about implementation is required for anonymization to be trustworthy and effective.

Data anonymization algorithms have been a staple of data management and statistical research for decades [29]. As data production, collection and use expanded in scale, data anonymization approaches and policies became increasingly utilized for addressing the emergent privacy risks [38]. Yet applying anonymization is far from trivial [17] and when organizations promise to anonymize the data they use, how can societies be sure that such promises are in fact fulfilled without insight into its usage in practice [21]? Implementation of anonymization algorithms requires making subtle assumptions on what constitutes sensitive information, along with enumerating and estimating the risks of possible information disclosure [29, 42, 43]. The usage of contemporary anonymization algorithms such as k-anonymity [47] and differential privacy [15] involves the calibration of stochastic methods to balance between reducing the risks and preserving the analytical value of the published data. Intricate heuristics have to be followed to map complex requirements of human data subjects to relevant parameters of the anonymization algorithms [4]. These factors all contribute to the considerable difficulty in making decisions when anonymizing datasets. As the Apple example above demonstrates, the definitions and policies are still being formulated and the public has limited knowledge about the practices and motivations involved in making decisions about anonymization. In this paper we investigate whether and how anonymization algorithms are implemented by organizations and companies in Europe. Our goal was to explore how anonymization approaches are used, whether they are considered practical and adequate, and how decisions are made to protect the privacy of data subjects while preserving analytical value.

We found that despite growing interest from a broad range of stakeholders and high expectations of the merits of data anonymization methods, few of the organizations we approached applied them in practice. This may be a symptom of anonymization being an afterthought, but the attitude is changing given the way the new European General Data Protection Regulation (EU GDPR) explicitly excludes data that are anonymized from its purview [41]. Our study and response rate were complicated by the fact that many of the organizations we contacted, consider the topic so sensitive that they are unwilling to discuss their handling of it. Some claim that such a policy of non-disclosure is in place to ensure that anonymization procedures are less vulnerable to de-anonymization attacks. This resembles a kind of anonymity through obscurity approach as an extra safeguard, although security experts recognize that relying on obscurity is ineffective. Swire [48] names more than a hundred thousand discussions of practitioners online, noting that most criticize security through obscurity. Among the six organizations that discussed their anonymization approaches with us, we found that decision-making varied considerably across practitioners working in different fields. Approaches to ensuring the viability of data anonymization for balancing privacy and analytical value often included the development of complex socio-technical systems of additional safeguards together with relatively nonchalant acknowledgements of the remote possibility of de-anonymization attacks. Achieving anonymization for these organizations was far more than implementing the right algorithm and our participants acknowledged that there was nothing definitive about their methods. Rather, these were contingent, dynamic processes that required managing different types of domain knowledge and disparate levels of technical expertise, transforming anonymization from a purely technical activity to a socio-material practice [45].

2 BACKGROUND

Data anonymization is not a new concern. Anonymization, de-identification, pseudonymization, and data masking all refer to methods of removing information that can be used to identify individuals. Conventional anonymization methods, such as k-anonymity, usually combine some

suppression, generalization, and perturbation based on hypotheses of what background knowledge a hypothetical adversary might possess, to reduce the risk of indirect identification. Although many researchers use the terms anonymization and de-identification interchangeably [7, 33], some have argued that the term de-identification more clearly captures what these approaches do and do not offer [42]. From a mathematical perspective perfect anonymization is not possible and available algorithms are only providing pseudo-anonymization. However, we use the term anonymization following the European General Data Protection Regulation (EU GDPR) that defines anonymization and distinguishes it from pseudonymization, which is relevant for data management approaches in Europe in particular [25, 41]. The regulation gives the term anonymization special meaning with significant legal implications.

2.1 Anonymization in the EU GDPR

In the European context, anonymization has recently gained new prominence due to the EU GDPR. According to the 2015 Eurobarometer report on data protection, 57% of EU citizens believe that regulation regarding the sharing of private data requires additional attention, and 70% are concerned that corporations and institutions use their data for other aims than the original purpose of data collection. Only about 15% feel that they have complete oversight over the usage of sensitive information provided through online channels [18]. Privacy concerns then become an issue for hopes of a lucrative data-driven economy. Yet research has shown that people may be willing to disclose more data if they believe their data will be anonymized [6]. GDPR addresses many of these concerns, and establishes anonymization as a possible solution to processing personal data. The regulation draws an important distinction between anonymizing such data and pseudonymizing it.

GDPR exempts anonymous data from its purview, defining anonymization as "data rendered anonymous in such a way that the data subject is not or no longer identifiable" [41] and requiring that de-anonymization is provably not possible even with additional information. The regulation also imposes further burdens on the data controller to "ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments" [41].

In comparison, Article 4 of the GDPR defines pseudonymization merely as: "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information" [41, 53]. This approach aims to reduce the risk of re-identification (which may be possible with additional information), but has no legal effect [41]. While the GDPR introduces some clarity in terms of identifiability and consequences, the requirements for anonymization and for what qualifies as "reasonably likely" are less clear [25, 38].

2.2 Understanding Anonymization in Practice

Anonymity has been an important area of research in many areas of computing, including CSCW, although the vast majority of this effort has been from the point of view of the end-users. Studies have shown that users may seek anonymity for a variety of reasons, but they are often unsuccessful. This is due to the lack of technical expertise [24], lack of awareness about the full extent of (meta)data collected [10, 32], underestimation of the data retention policies and inference capabilities of parties with access to data [19, 23], the perceived difficulty of using anonymization applications [3, 10, 23], the lack of support by some platforms and services for anonymous use [19], as well as governmental efforts to block access when using anonymization applications [44, 54]. When it comes to online security and privacy, non-experts tend to rely on oversimplified mental models to assess risks and make decisions, leaving them vulnerable to threats they do not account for [24, 40, 52]. Correa et al. have shown that posts on an anonymous social media site cover a rich spectrum of content

types and anonymity sensitivity, and noted that the level of anonymity required may differ [11]. Leavitt [27] demonstrated that the use of anonymous accounts to manage sensitive discussions and evade privacy issues was common, but the pervading reason for their popularity was their convenience. Fulfilling user anonymization needs is not trivial. For example, informed users can make decisions about their data that are mostly consistent with their values, but enabling their involvement using appropriate interfaces and services may be complicated [6]. It is no secret that anonymization applications have a usability problem. For example, the design of Tor and its mobile browser hardly facilitates feedback, user motivation, or learning of correct threat models, increasing risks [3]. Yet anonymization is not exclusively the responsibility of the user. In a data-driven economy, it is important to turn our attention to those that collect and use data rather than focusing on the practices of those that produce it.

Comparatively little research has investigated how anonymization is applied within organizations and what challenges they might face. Sedayao et al. reported unforeseen risks when directly implementing a limited anonymization framework in an enterprise, intended for protecting employee data [43]. However, they only note limitations of the particular approach they used, *k*-anonymity, and do not discuss broader organizational implementation issues. Adams & Blandford studied how security policies and procedures in a clinical setting are perceived by employees within an organization, showing that implementations of security and patient data management policies can become divisive when they are defined top-down by the IT department [2]. Elliot et al. [17] developed the notion of the data situation to discuss the context-dependent nature of anonymization as a process. They acknowledge that the technical application of algorithms is only a small part of the overall process of complex decision making, which also includes ethical considerations, public engagement and stakeholder analysis. On the basis of extensive research, Elliot et al. [17] propose a practical decision-making framework for applying anonymization given the requirements of the current and upcoming European legislation. However, aside from relatively general instructions for how organizations might implement anonymization, there is little discussion of what organizations already do and why. As CSCW research has repeatedly demonstrated, understanding current approaches and practices is important for further research and tool development.

2.3 Major Approaches to Anonymization

The area where anonymization was investigated first is statistical data analysis. Statistical Disclosure Limitation (SDL), also known as Statistical Disclosure Control, focuses on protecting the privacy of subjects in statistical data sets for research purposes. SDL is often applied to tabular data intended for public release, and is associated with a broad range of techniques including de-identification—exhibiting some overlap with anonymization, as well as aggregation, sampling, and query-based access methods, among others. The goal of SDL is often framed as finding an optimal balance in a given case with regard to trade-offs between information content and confidentiality. In their review of SDL methods, Matthews and Harel [29] note that multiple alternative conceptualizations of privacy and disclosure can impact risk assessment and selection of anonymization techniques. Moreover, despite the variety of SDL techniques it is difficult to ascertain what is "private enough" given the constant need to strike a balance between privacy and utility. Research on SDL is mature and some scholars have argued that SDL offers a framework to augment newer and more complex algorithms of today [22, 42]. However, SDL is highly structured and prescriptive. It clearly defines which data must be treated and how, making it more difficult to apply in the diverse context of big data. This has led the computing community to focus on probabilistic and combinatorial methods that are often advocated as generic and domain-independent.

Syntactic Anonymization Approaches. Concerns about data exposure have grown in lock step with advances in data collection and processing capabilities. Early on, Sweeney demonstrated that re-identification of subjects in U.S. Census data is not only possible, but relatively easy by linking these datasets with other publicly available data [46]. Sweeney challenged widespread assumptions that removing explicit identifiers from person-specific datasets would ensure the subjects' anonymity and developed one of the most popular formal models of anonymity: k-anonymity [47]. A tabular dataset is said to be k-anonymous if each valuation of a subset of columns (a quasi-identifier) occurs at least in k different rows in the table. A quasi-identifier [12] can be used to identify individuals using correlation. This anonymization method requires that data holders pinpoint quasi-identifiers with due diligence but acknowledges that there are limits to inferring what might constitute these. Thus even if relatively simple, it is already challenging to apply in practice.

The k-Anonymity protection model and related methods, all known as syntactic anonymity criteria, have received much criticism about their value for protecting privacy in the last two decades [14]. Those who challenge the model's suitability for contemporary applications have highlighted vulnerabilities stemming from unique characteristics in the data, questions about the viability of defining quasi-identifiers, as well as the outright failure of the model on high-dimensional datasets, among other factors [9]. However, k-anonymity has been and continues to be influential. Syntactic anonymization techniques, including k-anonymity, are often listed in guidebooks and legal documents on the topic, usually along with their weaknesses (e.g. [38]).

Differential Privacy. As criticisms of syntactic approaches mounted, alternative notions of privacy gained notice. In a series of theoretical results Dwork demonstrated that de-identified data do not really exist, either because subjects are still identifiable, or the data carry no information. Dwork introduced the notion of differential privacy, a data protection model that allows to control the trade-off between anonymity and usability of the data more explicitly [15]. A key principle of differential privacy is that anonymity loss should be quantified in a meaningful way. The privacy "price", or ϵ (epsilon), is used to calibrate the amount of randomness introduced into query results; achieved by adding random noise to the answers. Yet setting the value of ϵ can be problematic, as the Apple case has demonstrated. Computing a global sensitivity and selecting the ϵ value that both guarantees privacy and provides usable data is a difficult task that requires a complex analysis of the domain [9, 37].

Differential privacy also has other limitations: it does not offer protection to known cohesive groups or entire populations from conclusions based on aggregate results, and according to the Fundamental Law of Information Recovery, overly accurate answers to too many questions will eventually destroy anonymity (similar to other privacy-preserving analytical techniques) [15]. Moreover, differentially private algorithms or database mechanisms have to be developed for every use case separately, which has been a barrier to adoption, and its operating principles may be difficult to grasp without mathematical/statistical knowledge. Thus differential privacy is not an obvious and final answer to anonymization problems in industrial practice [9]. In light of these issues it is rather unclear how many data controllers would be willing to undertake the effort of adoption of differential privacy.

2.4 The Problem of Re-identification

In the early 2000s advances in anonymization inspired a lot of confidence and companies began releasing their datasets publicly. In 2006, Netflix released an anonymized sample consisting of the movie ratings of 500000 subscribers as part of the Netflix Prize; a public competition intended to improve their recommendation system. Narayanan and Shmatikov demonstrated severe limitations of the applied anonymization [35]. They introduced robust, general methods for identifying users

in supposedly anonymized datasets, and used this case to reflect on the challenges of anonymizing sparse datasets.

Rapid advances in re-identification techniques lead to laments about the "broken promises" [36] of anonymization as a technical solution to data privacy woes. The efficacy of anonymization as a solution to the problems of a data-driven society has been hotly debated since. The policy community has made efforts to rehabilitate the promise of anonymization as an approach to privacy-preserving data processing. Notably, Cavoukian et al. argued that anonymization should be encouraged, as they estimated re-identification risks based on proofs of concept in the literature to be low, and stricter regulatory measures would hinder potentially beneficial secondary uses of data [7]. In response, Narayanan et al. critiqued Cavoukian et al.'s approach as un-scientific, pointing to arbitrary assumptions about the knowledge of attackers, and disregard for significant privacy risks stemming from inference methods [33]. Narayanan et al. further argued that anonymization should not be seen as a default option by policymakers due to it providing no robustly quantifiable privacy [34]. Beyond the differing conceptualizations of data anonymization, there are ongoing debates about its merits for solving privacy issues associated with analyzing personal data. While researchers like Nathan Eagle see anonymization as key to enabling fair use of personal data [16], others have already proclaimed its failure in specific domains [13]. To reform policies, Rubinstein and Hartzog advocate more organizational responsibility and combined technical and regulatory measures tailored to minimize risks in a given context [22, 42]. However, despite recognition that purely technical approaches to anonymization have limited utility, it is curious that there has been little discussion of organizational processes and socio-technical approaches so far. To explore this question we leverage Suchman's concept of located accountabilities as a way to consider how application of anonymization processes is managed within organizations given their obvious socio-technical nature [45].

3 METHOD

Our goal in this study was to learn how anonymization techniques are applied in practice, which techniques are typically selected and used as well as what kinds of policies guide decisions about anonymization. Where theoretically the legal context and internal organizational policies ought to work as guidelines for applying anonymization, in practice what such anonymization actually constitutes and how decisions about which data to anonymize likely vary dramatically. As we began our study we expected that different types of organizations may be more or less willing to discuss their anonymization practices. As such we identified and approached a wide range of organizations in Denmark, all of which act as data controllers because they collect some form of personal data for business purposes. We were willing to remain flexible and to adjust our engagement with each organization depending on their requirements. However, we were unprepared for the difficulty of finding entities that utilized anonymization at all. Even among the many (public and private) organizations whose primary activities involve processing and sharing sensitive personal data, very few admitted to the use of any anonymization methods.

As this became apparent, we adapted our data collection approach to contact entities further across Europe, focusing on organizations that publicly communicate their processing of personal data on a large scale and those that operate in sectors with especially strict regulation, whose practices would be most impacted by the GDPR. Thus, we systematically contacted census- and national statistical agencies, healthcare organizations, financial institutions providing personal finance services, telecommunication providers, major online advertising-, search-, and social network platform companies, data analysis- and data science companies working with consumer- or political research, IT companies that participated in large European public IT projects, highly digitized multi-national enterprises with a large number of employees in Europe, and other Internet

organizations and IT consultancy-, IT infrastructure-, and IT service providers (with particular focus on those emphasizing anonymization, privacy, or related terms in their public communication), as well as researchers at universities who have engaged with these sectors in their work.

3.1 Data collection

Data collection was conducted in the winter and spring of 2017 using snowball-sampling techniques. We collated a list of likely organizations to contact using the industry contacts available via our university and authors personal networks. We also searched for profiles on professional and academic portals (i.e. LinkedIn and Academia) for individuals and organizations that mentioned expertise in anonymization, collated job postings mentioning anonymization and collected media news items. The first author contacted directly close to 70 persons and organizations via email, briefly introducing the project and, when possible, attaching the set of interview questions we prepared in advance, in order to ensure the transparency of the study (see Appendix A). In our contact emails we explained that we are academics interested in understanding how data anonymization techniques are applied in practice and which techniques organizations find most useful. In order to best accommodate participants' individual preferences we offered to conduct the interview by email, in person, by phone or Skype. Even if they declined to participate for whatever reason, we asked everyone we contacted to recommend other relevant people or organizations for our study.

Two organizations (one private, one public) that are involved almost exclusively in personal data processing pointed us to their publicly available policies or guidelines. Our analysis of these documents showed that while quite different in form, they addressed matters of privacy relevant to the organizations' business practices. These documents typically mentioned confidentiality and identifiability, but they did not mention anonymization at all. These public-facing documents presented a set of policies each organization used for their data management but did not provide insight into how these policies were implemented. Formal documents may clearly prescribe action, but CSCW research has repeatedly demonstrated that employees engage in a range of coordination activities to make these prescribed actions a reality if these are in fact followed at all [5, 30]. Seven of those contacted declined to help without explanation.

In total, we received 12 responses with content. Six of the respondents only provided either incomplete answers to our initial questions (and were not willing to be interviewed more extensively), or provided answers about data handling practices that did not involve anonymization. For completness we detail these below. Two responses shared organizational policies or guidelines; one data scientist at a multi-national advertising and search company indicated that he could not answer our questions for security reasons, despite working on relevant research projects (implementing differential privacy), and having contributed to multiple academic papers on the subject; an IT manager at a multi-national pharmaceutical company provided general internal notes and slides about anonymization, pseudonymization, and other concepts relevant to decision making about the use of personal data in light of GDPR; an employee services lead at a multi-national IT company provided some answers about an internal HR feedback system designed to collect data anonymously; and an HR manager at a multi-national consumer research firm provided information, internal guidelines, and a contract template for collecting personal data confidentially. Another six respondents provided complete responses as well as answered follow up questions or agreed to direct interviews, each from a different field (see Table 1). Two interviewees requested that answers be handled confidentially and one other required that we sign a non-disclosure agreement prior to participation (no answers from this participant are presented unless these answers were also echoed by other respondents, however we used this data to inform our thinking). To respect all requests for confidentiality we provide little specific information on the interviewees beyond their position and the field in which they are engaged.

3.2 Sample description

Six complete interviews informed our analysis most (Table 1), although we relied on all data we gathered. These covered a broad range of sectors and anonymization approaches. For each organization we interviewed one representative. To aid in readability we use one-letter pseudonyms to refer to each participant. B from HEALTH and F from TELECOM were researchers that also engaged with data management and policy in various organizations alongside their academic pursuits and had the most theoretical expertise and practical experience with anonymization approaches. B (HEALTH) also served as the data protection officer (DPO) for a clinical center. The rest of the participants were an analyst (A from FINANCE), a software developer (G from SOFT) and two IT specialists (C from ENERGY and D from LEGAL) who were also involved to a different extent with organizational policy about anonymization and data handling.

Below are short descriptions of the relevant aspects of each organization:

A (FINANCE): Typically shared data with specific contracted third parties and the size of the datasets their organization anonymized varied depending on the situation. They employed goal-specific anonymization techniques in order to comply with legal and regulatory requirements.

B (HEALTH): Shared data with specific contracted third parties and dataset size varied depending on the situation. They employed different anonymization techniques based on the characteristics of the dataset, to comply with legal and regulatory requirements, and expressed additional ethical concerns as reasons for anonymizing.

C (ENERGY): Typically worked with very large datasets that required distributed infrastructures, and shared data with specific contracted third parties. They were looking to anonymization techniques to comply with legal and regulatory requirements, and also saw anonymization as an additional competitive advantage, although this level of implementation was still in the planning stages.

D (LEGAL): Worked with medium and even small-scale datasets. They were the only organization that explicitly prepared data for public release. They employed an anonymization system combining named entity recognition and human review in order to comply with legal and regulatory requirements.

Table 1. Interviewee description

In text	Sector	Interviewee position	Anonymization approach	Additional security efforts	Data sharing	Country	Interview via
A FINANCE	Finance	Analyst	Goal-specific	Legal, IT security	Limited	Denmark	Email
B HEALTH	Healthcare	DPO	Dataset-specific	Limited access	Limited	Hungary	Email
C ENERGY	Energy	Startup CEO	Planned	Legal	Limited	France	Skype
D LEGAL	Legal/public data	IT Developer	NER+human-in-the-loop	IT security	Public	Denmark	In person + email
G SOFT	Software development	Software developer	Masking	Internal access only	Internal	Denmark	Email
F TELECOM	Telecom	Researcher	Privacy-through-security	Auditing queries	None, query results only	UK	Skype

G (SOFT): Worked with medium and even small-scale datasets. They had to anonymize externally received data for internal use. They employed masking techniques to comply with legal and regulatory requirements.

F (TELECOM): Were typically working with very large datasets that required distributed infrastructures. They developed an audited query-based system to provide answers instead of data to contracted third parties. They employed this approach to comply with legal and regulatory requirements, but expressed additional ethical concerns as reasons for anonymizing.

3.3 Data analysis

All interviews were collated and those that were conducted as conversations were recorded and transcribed. Data analysis followed an iterative approach with each round of coding focusing the investigation. The first author conducted an in-depth coding of all interview data following techniques derived from a grounded theory approach [8]. After the first round of coding, initial themes and insights were discussed with the second author and another round of thematic coding was conducted. The themes that coalesced after the second round of coding were discussed by all authors, finalizing analysis and relevant insights.

4 FINDINGS

Data anonymization is an actively discussed subject: it is granted a prominent role in current and upcoming legal policies about personal data, researchers from multiple disciplines contribute new methods or challenge assumptions on a regular basis, and companies like Apple even advertise their approach as a competitive advantage. However, despite an intensive, extended period of searching for potential interview subjects who engage with anonymization, very few of the organizations we had contacted indicated that they applied anonymization techniques of any kind to the data they used. Although many noted that they vaguely planned to implement anonymization of some sort at some point in the future, the most common reason for declining participation in our study was that the companies we contacted did not perform anonymization on the personally identifiable information they might be collecting and using. The practitioners that we did interview can be loosely describe according to their approach to anonymization although some used more than one of these approaches: (a) some use pseudonymization or anonymization/de-identification in a limited manner, (b) some use or have used a syntactic anonymization method similar to k-anonymity, and (c) some are using or developing advanced methods such as for example those based on differential privacy.

The general agreement among respondents was that the number of individuals and organizations using anonymization approaches declines sharply when moving from limited to more advanced methods not only because of technical complexity but also because of the difficulty with balancing privacy and data utility. Among the interviewees the majority were only using one form of anonymization—although all were aware of limitations in their different approaches and some took additional steps to protect data. In what follows we discuss why and how decisions about anonymization were made and what drove application of different kinds of approaches. We then describe our participants' concerns about the limits of anonymization and the additional measures they at times utilized to mitigate these limitations. Finally, we discuss the apparent problem of organizational secrecy around anonymization practices and the implications of the reticence to disclose these practices for the protective promise of anonymization as a privacy safeguard.

4.1 Why is Anonymization applied?

Despite the limited number of responses in our sample, the reasons for anonymization, decisions for which anonymization mechanisms to employ and how to apply these varied covering a full spectrum

of practices. For example, A from FINANCE and B from HEALTH operated in highly regulated environments, thus compliance was one of their primary drivers. Personal data regulations in many countries in Europe are quite heavy duty and the GDPR will generate even more regulatory pressure for implementing anonymization in the commercial context.

Yet regulation was not the sole driver. B, the data protection officer (DPO) from HEALTH, was very much concerned about ethics: "I publish my work in scientific papers, give lectures in medical informatics conferences and inform colleagues at my workplace. I want to represent the interests of the research subjects (patients) and protect their privacy rights." Here practical organizational concerns directly intersected with academic research and ethical discussions that, when coming from an institutionalized position of the DPO, could create an organizational space for ethical debate. B (HEALTH) pointed out that they also relied on an explicit "code of conduct on how to obtain research data," which guided how decisions regarding anonymization needed to be made.

C, whose ENERGY startup was just in the process of planning for anonymization implementations, expressed their concerns primarily with an eye to business development: "I see anonymization as an enabler, in the sense that the good anonymization techniques would help us develop our business and more innovative services." In this case, the use of anonymization was seen as an ethical and potentially significantly advantageous business practice: "I think like everybody's talking about these big data services market, which is supposed to be huge, but we can assume that big part of this big data market depends on good anonymization, which means that billions of dollars, huge opportunity." The implications of the GDPR at the European level clear drive anticipation of business opportunities alongside technical concerns for the efficacy of anonymization techniques.

D explained that LEGAL were focused on public court order data and saw the need for anonymization as a process that would automate an existing manual practice and as a way to demonstrate that anonymization processes can be done more efficiently thus motivating broader application to generate public data. As D (LEGAL) explained: "while this [was the] first Danish system … anonymization process … it was risky business. But we could prove that if it would be successful, it would be a big gain." The researcher F from TELECOM noted that telecom companies are very much aware of the sensitive nature of the location data they possess despite this area being currently not very strictly regulated. The sensitivity of such data has been previously debated with some researchers arguing that despite telecom data being potentially very useful for public good purposes, such data are so difficult to anonymize that making them too freely available is risky [50].

Alongside these high level concerns of ethics, business or regulation, our participants also shared much more practical reasons for using anonymization in organizational contexts. For example, A from FINANCE and B from HEALTH anonymized data in order to enable data sharing with external parties such as supporting research projects or engaging with business partners. B (HEALTH) also made their decisions directly in response to research data requests: "In the case of small data sets, simple de-identification method is used. When larger data set is processed, then I investigate the anonymisation issue in detail, taking into account scientific literature, the project needs and the IT capabilities." Thus anonymization was often applied in response to particular data needs rather than as a matter of course.

4.2 How is Anonymization applied?

When exploring decisions for how to anonymize data, we observed a diversity of approaches. Some, such as G (SOFT) or D (LEGAL), described previously established frameworks that were consistently applied to all data they were anonymizing. In contrast, A (FINANCE), B (HEALTH) and C (ENERGY) explained that they developed, or expected to develop, anonymization procedures tailored to each specific dataset. The approach of F (TELECOM) was the most extensive, proposing alternative methods involving radical limitations and oversight on data use.

Working With Established Frameworks. D from LEGAL and G from SOFT explained that their organizations had developed established frameworks over time. In both cases the process involved a significant engagement with the relevant stakeholders identifying what types of information were considered sensitive, developing proposals and sketching out solutions collaboratively. Eventually, these processes were codified in frameworks that combine IT systems and structured organizational activities of human involvement in the process.

D (LEGAL) was part of the team that developed a system where requirements were established through a lengthy, multilateral discussion involving two collaborating IT companies, the Courts of Denmark, and language technology researchers: "there was a careful discussion between the courts and the specific company about the set of rules that they should follow, the anonymizations, what could be accepted, what shouldn't be accepted." The courts made requirements about what should be considered sensitive information and how it should be handled, as well as came up with multiple possible solution scenarios. The IT companies, in collaboration with the researchers, focused on providing input about what they considered technically feasible. The software they eventually developed takes unstructured court documents as input (with no markup, and sometimes also formatting and spelling issues), performs named entity recognition (NER) [31] trained for this legal text domain to select anonymization candidates, and outputs structured XML files that can be further processed automatically in the document pipeline. The obtained system implements a mixed technical and manual solution in the style of SDL.

In light of the strict requirements about sensitive information and the limitations of language technology at the time of development (the project started in 2006), it was clear that human oversight in the anonymization process was crucial. This meant that the NER was optimized for recall, selecting every possible anonymization candidate to save reading time for the human editors: "given that those people … had to go through every document, then made sense to give them a little more control, instead of making the automatic anonymization focus on very high precision. … I decided on focusing on recall, because that would help the manual process a lot." The candidates are ranked by confidence of recognition with pseudonymous labels based on semantic categories (as specified by the Courts of Denmark, e.g. persons, companies, etc., with entities numbered according to order of occurrence in the document). The trained human editors then evaluate and finalize the anonymized output. A drawback of this approach is that different names are occasionally not recognized as referring to the same entity and mislabeled, which the editors have to recognize and correct. The system is currently in use and applied to court documentation. Substantial data homogeneity made it feasible to develop such a structured single-approach system because it is applied to a very predictably organized dataset. The manual approach was also necessary because certain data were simply too difficult to manage automatically: "there were things like ethnicity—should that be anonymized? … we might have been able to find trigger words, but the problem, the markup of candidates for anonymization would have been very fuzzy." The company instead built a complex socio-technical process that resulted in anonymized datasets where the human editors continuously engaged in coordination and repair work.

G from SOFT described a different highly collaborative process that demonstrated reliance on networks of working relations across organizational and disciplinary boundaries. This software development and consulting firm uses data from municipalities for production and testing of public IT systems. As they G described their process: "[…] the initial approach was based on customer requirements, and a desire to reach as high a masking level as possible. We now have a company-wide framework for data masking, used to ensure the ability to do cross-solution/platform testing." Although we were not able to get a specific technical description of their anonymization method, the technique appears to be derived from relatively lightweight syntactic approaches. Selection of fields that should be masked is conducted collaboratively with the business unit, using their

knowledge of the data to evaluate what can be used to identify individuals (directly or by inference), and decide what constitutes sensitive information. The security unit, with assistance from legal and the project team are also involved in decision-making regarding the masking setup, which is reflected in the company's policies. Despite potentially quite heterogeneous data, G explained that SOFT were able to develop a company-wide framework because their use of the data is very specific and pre-defined. In this case, converting the data to the right format is relatively straightforward and the question that needs to be considered is which fields to designate as sensitive: "Within the organisation we have multiple stakeholders: security, legal, business units, who all have to be on board for the project to succeed. Security and legal need to know what we do, so policies reflect our setup. Business units have knowledge of the data we are tasked with masking, so we need them for analysis prior to masking their data to ensure we don't miss anything." There is then a codified highly collaborative process for identifying sensitive data in any dataset, done with acknowledgement of what Suchman called located accountabilities [45] where responsibilities are discussed, shared, assigned and worked through.

These examples demonstrate that anonymization as an approach is not a single algorithm but a socio-technical framework that includes algorithmic processing together with collaborative human oversight working across multiple boundaries, thus acknowledging how partial and located knowledge about what constitutes sensitive data can be. However, these two approaches are very different in how the social and the technical are combined into a coherent process. Where LEGAL has implemented a kind of "human-in-the-loop" system with human coders filling in for the inadequacies of technology, SOFT's framework was there to explicitly support discussion of options and possibilities as part of decision-making. In an interesting iteration on this idea, F from TELECOM argued for a much more constrained and technically bounded solution where to keep the promise of anonymity while extracting value from such data, an alternative method is necessary involving "privacy-through-security." The dataset would be on a secure server with an audited question-and-answer system. Queries would have to be submitted for (human or algorithmic) review first, and when approved, the user would only receive the result, without direct access to the data. All user activity would be recorded and monitored, with the possibility of automatically restricting access: "basically, how can we ensure that the data will be used anonymously, not because the data is anonymous, but because everything has been put in place to ensure that the data will be used anonymously." In this way, the organization of F (TELECOM) can be in control as they remain legally responsible. This vision also included an acknowledgement for the need to have a social component to compliment the technical systems: "we want to make sure that we have people locally that can help us decide what are questions that are ok to be looking at, not ok to be looking at, so we're putting in place what we call a local governance board." In this conception local advisors are expected to be involved in the decision making about the design and pre-set limits of such a system [28, 39].

This is reminiscent of Dwork's work where the effort is really on controlling what kinds of queries can be allowed on any dataset [15], but also clearly recognizes the material and located nature of anonymization and goes beyond what Kunda has termed "corporate attempts at normative control" [26]. Discussing how to gracefully implement the social and the technical components however, lead to an admission that both aspects of the system are visions in search of implementation. How these might need to be implemented to ensure that the local engagement and the technical components might be gracefully integrated together remains an open question, leading us back to old questions of how networks of relations might be managed to build located accountabilities [45]. The fact that complex socio-technical solutions are at least considered, and often actually used, in these technologically advanced data-centric organizations challenges the very idea of convenience and effectiveness of generic anonymization solutions.

Tailored Anonymization Procedures. The practice of tailoring anonymization procedures to each data request is a different approach and is likely more common in organizations that engage with many different third parties. For A (FINANCE) and B (HEALTH) the data were typically generated as part of the services their organizations offered to customers. These data then were utilized for purposes other than the direct provision of service, requiring special handling. C explained that ENERGY currently store data and do processing for data holders and have conducted pilot projects with researchers based on data from their customers: "I think the first objective here is to make sure that you give an adequate protection to people, so you need to identify what does it mean—adequate protection—and on the second hand you need to make sure that you provide enough useful information to the service." As this ENERGY startup explores anonymization, C emphasizes that there can be significant differences between services in terms of what is considered useful information, and thus each case requires discussion of what (if any) anonymization approach should be taken.

The conditions that defined whether and how anonymization was applied differed depending on the context of data sharing for C (ENERGY) and A (FINANCE) and, especially in the case of B (HEALTH), on the size of the dataset. For B (HEALTH) where smaller datasets allowed for simple de-identification methods primarily because they did not include enough information for de-anonymization in cross-referencing scenarios, larger and more diverse datasets required significant oversight. As the DPO explained: "When a larger dataset is processed, then I investigate the anonymization issue in detail, taking into account scientific literature, the project needs and the IT capabilities." The issue of size for B (HEALTH) brought up a lot of interesting concerns. Some of these larger, more complex projects required development of a selection and anonymization algorithm, which asks crucial questions about funding and responsibilities, demanding involvement of different parties for decision-making. For example this could include people from the local operation team, or the IT company contracted for data processing.

The way that specific features of the dataset should be handled, in terms of suppression or generalization, were evaluated on a case-by-case basis (for A (FINANCE) and B (HEALTH)) with actions taken based on the value of the data and how challenging it is to anonymize them. As B (HEALTH) explained: "Generally the demographic data (birth date, birth place, resident address, ZIP code), numeric identifiers (social security, patient identifier in the HIS system) are removed or de-identified." Given the specificity of health data, in these cases physiological parameters are rarely perturbed, but in the case of longitudinal data, dates are distorted to prevent known-date attacks. In some cases, B (HEALTH) chose to perform de-anonymization attacks and to compute probabilities to express risks, in order to review the results before finalizing the approach.

In all cases the ideal approach was typically negotiated by multiple stakeholders, but in some cases (A (FINANCE), C (ENERGY)) experience with relevant anonymization methods or lack thereof could prevent the necessary agreement on the route of action. For B (HEALTH) the medical research ethics committee and the information security leader also needed to be consulted about key aspects of the anonymization approach, and the final decision was made by the president of the clinical center on the basis of collaborative input. Yet even when discussions were not quite so formalized (A (FINANCE), C (ENERGY)), they spanned many departments (i.e. legal, customer service, IT support) and were crucial to understanding and managing the risks and responsibilities involved in releasing any dataset. For example, C (ENERGY) explained: "The choice of techniques will be done between product marketing people, and the development people. And then the other important player would be the client, the customer, then there is the public body in charge of privacy protection—like, in France we have CNIL."[1] It is clear from our data that the organizational

[1]CNIL is Commission Nationale de l'Informatique et des Libertés, the French national data protection authority.

complexity of decision making about the choice of technique or framework for anonymization can vary significantly depending on types of data, organizational configurations as well as how responsibilities are assigned and managed. What at the first glance appears to be a technical decision turns out to be a complex balancing of socio-technical processes where partial knowledges [45] must be managed as anonymization relies on imagining future attackers or future problems. After all, predictions of futures are always partial.

4.3 Beyond anonymization

Anonymization is not an easy solution and our data provide no indication that anonymized datasets were ever treated as unproblematic. The problematic nature of anonymization was discussed in two ways. On the one hand, anonymization procedures could limit the utility of the data and required changes in how data might need to be treated. On the other hand, the threat of re-identification was always present and acknowledged. F (TELECOM) argued that while data anonymization/de-identification methods like k-anonymity might have historically been adequate, they do not scale to address the challenges associated with contemporary data capture and analytical practices, and fail outright when it comes to location and mobility data: "The thing is, historically, it worked ok … one of the first, *k*-anonymity algorithm … worked well when we were sending data in paper form, on floppy disks, and fairly small datasets, like Excel file type data, CSV data. The big issue is, in today's world the data that people are actually interested in are fairly big, over time, very big dimensional data of human behavior … the technique of anonymizing such data just does not scale … it worked well in Excel with five columns, it does not scale to data in which you have hundreds of thousands of points about a single person over a period of a year." The argument here is that such an approach does not prevent re-identification in the face of current threats but the promise of anonymization could lead to legal decisions being based on the fact that the data have been anonymized. Such concerns illustrated the problem for many in our project—the privacy threats addressed by anonymity had to be serious enough to warrant potential reductions in the utility of the resulting dataset, yet it was clear that understanding and estimating real or imagined risk was very difficult: "It helps, and yes of course, doing something like it makes it more difficult, but … the current state of technology is such that anonymization does not prevent re-identification."

For C (ENERGY) and A (FINANCE), the fact that anonymized data cannot be linked with other data sources such as social media and public registers could become a major issue. The suppression and generalization techniques applied on their datasets could further limit data utility in some cases. B (HEALTH) and A (FINANCE) noted that anonymization methods sometimes had to be changed in cases where algorithms resulted in data that did not produce valuable results. As B (HEALTH) explained: "Since large data sets are requested by participants of bigger scientific projects, therefore sometimes the implementation needs to be changed. When the expected results are not obtained, or the algorithms do not produce any valuable results, then the anonymization or the record selection process has to be changed." C (ENERGY) noted that developments regarding anonymization in the smart energy field are primarily enabled and driven by agencies and other governmental organizations. They are tasked with finding a balance between protecting the privacy of citizens and sustaining innovation for economic growth—a trade-off where the right position is rarely evident.

Given the public nature of their data output, D explained that LEGAL paid a lot of attention to ensuring the security of the system and the overall document pipeline, such as during file transfers, and in terms of access. D (LEGAL) pointed out that "[…] even if all references to people, companies etc. are anonymized perfectly, then there is a risk that, I expect, will grow in importance: By automatic cross-referencing various data sources it is likely that anonymized documents can be enriched with the previously removed references to people, etc. Currently, it is difficult to construct

countermeasures for that sort of reverse anonymization. I'm convinced that the world is undergoing a learning process now, which might bring solutions." Despite the involvement of human editors, predicting what could be done with proliferating datasets of different kinds becomes a guessing game where, echoing F from TELECOM, anonymization "does something" but in and of itself it is not enough.

G explained that SOFT were very much aware that their masking approach does not really make data anonymous, and is susceptible to re-identification by an attacker with the right knowledge. However, G noted: "our take is that any data that is truly anonymous is not usable for testing (functional testing and above)" as it eliminates essential structural information, so no one could rightfully claim to use it. When it comes to the drawbacks of masking, the testing process has to be adjusted, and testers (i.e. customers) need to get used to working on masked data. G noted that SOFT managed the risk of re-identification by limiting the extent of masked data use to internal operations only, as well as limiting discussions of their approaches outside the small teams that engaged in data masking. Anonymization was only part of a range of restrictive practices that all organizations tried to employ. For SOFT, however, this sort of secrecy became integral to achieving the balance between utility and privacy protection. Considering legislative changes like the GDPR, G (SOFT) argued that data masking is the one possible way of compliance. However, SOFT continuously reevaluated their approach to optimize their setup and to address new threats from more open datasets and new re-identification schemes. They have also considered working with synthetic data, but concluded that it is infeasible due to the complexity of the degrees of freedom involved in generating it. Instead they try to make re-identification as hard as possible while retaining the necessary utility, only use the data in-house, on specific test setups, and rely on obscurity as an additional security measure.

F (TELECOM) noted that his research team proposed a radical solution, which was essential because adding noise to location data yields decreasing returns in terms of privacy with respect to mobility traces. The team demonstrated that, due to individuals' highly unique spatio-temporal traces in "[...] credit card and mobile phone and other location data types, four points, four places and times where someone was, is sufficient to uniquely re-identify him 95% of the time." This led the team to develop highly restrictive approaches, by limiting access to data and creating a system of monitoring and oversight for data queries from external parties.

In all cases we heard different discussions of how access to anonymized data could be and often was managed and restricted. For C (ENERGY) and A (FINANCE), any external parties receiving access to anonymized data must be subject to some form of contractual agreement. In some cases secure storage devices are used when transferring the data, as an additional security measure to prevent unauthorized access. B (HEALTH) noted that out of consideration for any unanticipated risks, only the members of the research project that requested the data may have access to it. Such extensive efforts speak in part to the limits of anonymization approaches, but in part also to the difficulty with trusting any anonymization approach in the face of continuous technological development. As F (TELECOM) explained: "Anonymization is such a convenient idea, it is so easy, it's beautiful, like oh you take the data and then someone smart comes and invents an algorithm that's … the data and you can do anything you want with it … but then it turns out the dataset is subject to you know short term or medium term re-identification." Despite the efforts, there was a sense of risk and impermanence to anonymization techniques and so the representatives of the organizations and companies we spoke to repeatedly came back to the socio-technical puzzle they were solving and the impossibility of a purely technical solution.

5 DISCUSSION & CONCLUSION

Although anonymization has been featured in public discourse, emphasized in private data regulations, and emerged as an active object of academic inquiry in multiple disciplines, of the nearly 70 organizations we approached far fewer indicated that they use these approaches than we had expected. At the same time, many organizations that did engage with anonymization treated it as a sensitive topic and thus refused to answer our questions or limited the information they provided due to security concerns. Approaches to anonymization among our respondents and their reasons for granting access were diverse. They expressed significantly different opinions, echoing some of the ongoing debates about the viability of data anonymization. Perspectives ranged from general optimism about the possibilities of emergent anonymization methods to complete dismissal of conventional anonymization practices and the policies that regulate them.

At the same time, even among participants that did agree to engage with our study, their willingness to disclose details of organizational anonymization approaches varied. Among those who offered complete answers, some asked for confidentiality, with their answers handled anonymously. In these cases, their disclosure of information went through a legal and security approval process. Curiously, the stated reasons for reticence in this kind of disclosure were explained less by concerns about loss of competitive advantage and more by additional security measures to limit the potential for de-identification attacks. This sort of anonymization-by-obscurity is potentially problematic and warrants deeper consideration. While some responses cited security reasons for not providing more information, reasons for treating the subject as sensitive are not easy to pinpoint. Such attitude can also be motivated by the data anonymization methods being seen as trade secrets, which could be the case for example where industrial research is being done, or alternatively by concerns about gaps between actual practices and legal or internal policies, potentially resulting in a lack of control over the data, which could negatively affect their reputation if published. To us this suggests a significant immaturity of practice and the variability of policies around application of anonymization mechanisms and disclosure of these policies troubles the expectation of trust increasingly required for the data economy to operate.

5.1 The Problem of Anonymity-by-obscurity

Concerns about de-identification threats in many cases lead to efforts to control threats through additional measures such as contractual agreements about data use or limited data releases (only to project members, etc.). One of the major ways of limiting exposure, however, was through efforts of obscurity. In many cases companies claimed that they were unable to explain their approaches to us precisely because this was their way of limiting the possibility of external attack on their anonymized datasets. We see this as a significant problem, and a limitation to its uptake. Obscurity, understood as secrecy regarding privacy protection practices, is likely the first reflex for many organizations, and may be interpreted as a sign of immaturity of industrial practice in this respect. Security practice has shown extensively that obscurity protects secrets only as long as an exploit has not been found. Returning to the Apple example, while the company has lauded its own efforts at applying differential privacy, highlighting the complexity of this technique as particularly important, the refusal to disclose how exactly the algorithms were being applied motivated researchers to investigate and led to a public relations problem.

From the point of view of privacy protection, the practice of anonymity-through obscurity has an additional drawback: it could be seen as an unwillingness to engage with the data subjects, and the broader society in a dialog about what should be protected and why. After all, being secretive about your practices makes it impossible to openly discuss them. Anonymity is often touted as a solution when data usage is discussed with respect to public concerns. The idea that when datasets

are anonymized privacy risks are sufficiently minimized to assuage most worries underlies new laws and policy implementations. Surprisingly, little research has examined how users react to their data being anonymized by others, although Brush et al. [6] suggest that users may be more willing to share location data publicly if they believe it to be anonymized. This suggests that if consumers come to believe their data are anonymized in a way that does not expose them, they may be willing to share more data - which can be problematic if what "anonymization" means is unclear. A deeper understanding of the socio-technical processes that produce anonymized data in practice are crucial for ensuring that when an organization promises anonymization, it is a trustworthy promise.

5.2 Located accountabilities of anonymization

The heterogeneity and complexity of data anonymization approaches is rarely made explicit in research, policies, or public discourse. Our data begin to surface contrasts in decision-making between various fields of application but also the very similar struggles they encounter. While the interviewees working with court orders and software testing used a predetermined and well defined anonymization approach or framework, the other responses highlighted more complex decision making. For G (SOFT) and D (LEGAL), the technical approaches were invariable and the necessary human decision-making needed to be done either before (G (SOFT)) or after (D (LEGAL)) the application of their particular anonymization method. In contrast, for the rest of the participants, the selection of the anonymization method itself had to be adapted or developed every time for each particular case or dataset, requiring complex negotiation throughout the process. Yet even with pre-set frameworks, their development in the first place required a complex socio-technical process involving networks of relations within organizations working across domain boundaries [45].

In Suchman's terms, while these organizations appear to be struggling to present a decontextualized front of technological superiority that solves the apparent problems of privacy through sophisticated highly complex technical solutions, the reality is kept tenuously together through significant efforts of working across boundaries and acknowledging that knowledges are always partial [45]. Each interviewee had their own perspective on the benefits and risks of the data anonymization practices in their own cases. Through their responses, they have also outlined—directly or implicitly—their views about the viability of data anonymization as a solution to protect privacy while preserving analytical value. These opinions show interesting contrasts, and may help understand the debates surrounding anonymization.

Where D (LEGAL) was optimistic about future developments in anonymization solving privacy issues, D was critical about the protection their project provides against newly emergent re-identification threats. For C (ENERGY), anonymization is seen as an enabler of more innovative and competitive products, although finding the right approach for particular applications is expected to be challenging. G explained that SOFT struggle to balance data viability with anonymization, claiming that truly anonymous data would be of no value at all. While B (HEALTH), C (ENERGY) and F (TELECOM) brought up protecting privacy as a key objective, others seemed to focus more on compliance with legal or internal policies. The concerns expressed in our data suggest that despite the care and effort involved, even the most complex anonymization approaches seem to offer too limited protections and may lead to a false sense of security, not addressing acknowledged and emergent threats. Anonymization then seems to be a little bit of an organizational "hot potato" where even discussing it can be seen as a dangerous form of disclosure that could lead to technical or public relations attacks. At times, it seemed that anonymization was seen as a necessary evil to be navigated in order to be able to cope with problems stemming from privacy regulations but in all cases it required immense internal organizational effort.

5.3 Socio-technical Anonymization as Method

How are decisions about anonymization made? In the organizations that we engaged there seems to be at least a general agreement and some guidelines that define whether and how anonymization is applied. In some cases, all datasets of a certain type and used for a particular purpose are anonymized using an established and painstakingly maintained framework. In other cases, there are particular conditions that may lead to different levels of effort in anonymization. However, across all of our respondents we observed an emphasis on a collaborative process comprised of diverse stakeholders that was necessary in order to identify how anonymization might be applied. The multi-stakeholder decision process addressed issues that ranged from setting goals and identifying risks, to decisions about which specific data ought to be masked. In most cases, what kinds of queries were allowed on even anonymized datasets is also tightly controlled unless the datasets (as in the legal case) are released publicly.

Part of the reason for this extended system where the application of algorithms is only one aspect, is because in all of the organizations having a clear idea of who is responsible for data handling and for responding "should something go wrong" is crucial in order to accomplish the use of data at all. Yet these responsibilities play a role in anonymization decisions to a different extent depending on the data and the purpose. After all, different types of data demand different levels of concern and effort. This depends on legislation and policy (financial and medical data are extremely regulated) or technical challenges of anonymization (location data) or both (energy data are increasingly regulated and potentially difficult to anonymize). Elliot et al. point to the idea of a "data situation", advocating the importance of considering the relationship between data and environment in anonymization decisions [17]. However, their report is oriented explicitly towards the organizational unit implementing the technical components and, although they emphasize the importance of considering legal and ethical obligations, they note that "consulting with your stakeholders is a useful exercise" [17, p. 101]. Our data suggest that in fact consulting with stakeholders is central to doing anonymization and not merely an exercise.

The socio-technical entanglement that enables anonymization in practice goes against much of the discussion and rhetoric around anonymization. Some discussions such as the public debate between Cavoukian and Narayanan suggest that the technical, social, and legal approaches to data disclosure are often in opposition. The legal approaches may over-optimistically rely on algorithmic guarantees, while the technical approaches are far more pessimistic, as they develop increasingly complex techniques that are potentially difficult to implement in practice [1, 9, 15, 34]. More social approaches, many of which are thoroughly researched and discussed in CSCW, seek to enhance awareness of problems, to produce analysis of potential impacts and, most importantly, to nudge and influence "folk theories" [40, 51] or "naïve" [24] behavior towards particular visions of generic "good security practices." The formulation around anonymization in GDPR suggests that the legal community has come to believe that technical implementations can offer the benefits of data without the attendant privacy issues and has moved to develop policies and guidelines based on this [4]. Such policies are intended to motivate the "right" kinds of behavior on the part of data controllers and end users alike. Yet in practice we found that organizations develop processes that merge all of these aspects pragmatically, demonstrating that instead of opposition, the different facets must work together to produce usable approaches to data and anonymization. The solution then is to create diverse teams despite the difficulty of cross-domain collaboration, enabling the engineers, legal and customer experts to develop anonymization approaches together. It appears that anonymization is not about getting the IT department or data science wizards to throw some algorithms and make the problems go away and never will be. Instead, we see a need to support negotiation, "mutual learning and partial translations" [45] as a matter of course. The more complex the anonymization problem

the more involved the process and it requires moving beyond claims to universality, demanding instead practical wisdom in the production of anonymization as a material practice.

6 LIMITATIONS AND FUTURE WORK

Although our research has produced insights into how anonymization is done in some organizations, our findings have been limited due to the problems of access. We believe that research on how anonymization is done in practice across different sectors is necessary. In order to understand what are the potential security issues and implications, we need a way to document particulars of decision-making processes in more detail. Observing anonymization as it is carried out would provide an opportunity for an inquiry into the differences between policies and practice as well as an understanding of how accountabilities are located. While addressing the perspectives of stakeholders who engage with anonymization in different ways, such as policy makers, managers, and data subjects, we also may want to ask whether there is a rift between practitioners' and the wider public's understanding.

The lack of published information about organizational data anonymization practices raises several questions, one of the most pressing being whether the promise of anonymization is indeed broken [36]. The GDPR will likely drive further efforts towards implementing anonymization by organizations that are as yet not using such tools. Thus it is imperative to analyze best practices, addressing the diversity of needs and complications across the different aspects of the data rich world. Anonymity through obscurity is not going to help this process. Where legislation such as GDPR emphasizes technical solutions, demanding that anonmyzation techniques withstand the test of time, our study demonstrates that the technical problems are only one part of the process. Arguably, the interdependent socio-technical process of anonymization required that organizations are able to support productive internal crossing of boundaries. Clearly we need transparency and dialogue about what anonymity actually means in practice. Yet our study suggests that transparency of the mechanics of technical approaches will not be enough. Rather, we need to seek transparency for how decisions are made in the course of developing and applying anonymization alogrithms.

A INTERVIEW QUESTIONS

(1) What kind of data do you capture/data sources do you work with?
(2) At which stages of your work process does anonymization play a role?
(3) How do you decide what anonymization approach to use?
(4) What decisions are made when implementing a specific method?
(5) Why and when may the approach or its implementation change?
(6) How do you determine which features of the data should be anonymized?
(7) How do you decide and control who can access the dataset/queries?
(8) Who is involved in making these choices?
(9) Do you measure/monitor the effects of anonymization on the data itself or your work?
(10) Do you share the (anonymized) data, your analytical capabilities, or the results of your analysis with anyone outside your organization?
(11) Do you experience any drawbacks of not being able to identify individuals in the data?
(12) Do you discuss your approach to anonymization with anyone within the organization? Outside the organization? Why? For what purpose?
(13) How confident are you that this approach to anonymization will stand up to potential de-anonymization attacks?

ACKNOWLEDGMENTS

We thank our participants for their generous participation and anonymous reviewers for constructive feedback. The work is supported by the EU Horizon 2020 Research and Innovation program under grant agreement No.: 732027.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. ACM, New York, NY, USA, 308–318. https://doi.org/10.1145/2976749.2978318

[2] Anne Adams and Ann Blandford. 2005. Bridging the gap between organizational and user perspectives of security in the clinical domain. *International Journal of Human-Computer Studies* 63, 1 (July 2005), 175–202. https://doi.org/10.1016/j.ijhcs.2005.04.022

[3] Hala Assal, Stephanie Hurtado, Ahsan Imran, and Sonia Chiasson. 2015. What's the Deal with Privacy Apps?: A Comprehensive Exploration of User Perception and Usability. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia (MUM '15)*. ACM, New York, NY, USA, 25–36. https://doi.org/10.1145/2836041.2836044

[4] Jane Bambauer, Krishnamurty Muralidhar, and Rathindra Sarathy. 2013. Fool's Gold: An Illustrated Critique of Differential Privacy. *Vanderbilt Journal of Entertainment and Technology Law* 16 (2013), 701. http://heinonline.org/HOL/Page?handle=hein.journals/vanep16&id=733&div=&collection=

[5] Pernille Bjørn and Kjetil Rødje. 2008. Triage Drift: A Workplace Study in a Pediatric Emergency Department. *Computer Supported Cooperative Work (CSCW)* 17, 4 (Aug. 2008), 395–419. https://doi.org/10.1007/s10606-008-9079-2

[6] A.J. Bernheim Brush, John Krumm, and James Scott. 2010. Exploring End User Preferences for Location Obfuscation, Location-based Services, and the Value of Location. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. ACM, New York, NY, USA, 95–104. https://doi.org/10.1145/1864349.1864381

[7] Ann Cavoukian and Khaled El Emam. 2011. *Dispelling the myths surrounding de-identification: Anonymization remains a strong tool for protecting privacy.* Information and Privacy Commissioner of Ontario, Canada.

[8] Kathy Charmaz. 2014. *Constructing Grounded Theory.* SAGE.

[9] Chris Clifton and Tamir Tassa. 2013. On syntactic anonymity and differential privacy. In *Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on*. IEEE, 88–93.

[10] Gregory Conti and Edward Sobiesk. 2007. An Honest Man Has Nothing to Fear: User Perceptions on Web-based Information Disclosure. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS '07)*. ACM, New York, NY, USA, 112–121. https://doi.org/10.1145/1280680.1280695

[11] Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto, and Krishna P Gummadi. 2015. The Many Shades of Anonymity: Characterizing Anonymous Social Media Content.. In *ICWSM*. 71–80.

[12] Tore Dalenius. 1986. Finding a needle in a haystack or identifying anonymous census records. *Journal of official statistics* 2, 3 (1986), 329.
[13] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.
[14] Josep Domingo-Ferrer and Vicenç Torra. 2008. A critique of k-anonymity and some of its enhancements. In *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*. IEEE, 990–993.
[15] Cynthia Dwork, Aaron Roth, and others. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
[16] Nathan Eagle. 2009. Engineering a common good: fair use of aggregated, anonymized behavioral data. In *First international forum on the application and management of personal electronic information*.
[17] Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. 2016. *The Anonymisation Decision - Making Framework*. UKAN, University of Manchester.
[18] Special Eurobarometer. 2015. *431 DATA PROTECTION REPORT Fieldwork: March 2015, Publication: June 2015*.
[19] Andrea Forte, Nazanin Andalibi, and Rachel Greenstadt. 2017. Privacy, Anonymity, and Perceived Risk in Open Collaboration: A Study of Tor Users and Wikipedians. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1800–1811. https://doi.org/10.1145/2998181.2998273
[20] Andy Greenberg. 2016. Apple's 'differential privacy' is about collecting your data – but not your data. *Wired (June 13, 2016)* (June 2016). Retrieved February 13, 2018 from https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/
[21] Andy Greenberg. 2017. How One of Apple's Key Privacy Safeguards Falls Short. *Wired* (Sept. 2017). Retrieved February 13, 2018 from https://www.wired.com/story/apple-differential-privacy-shortcomings/
[22] Woodrow Hartzog and Ira Rubinstein. 2017. The Anonymization Debate Should Be About Risk, Not Perfection. *Commun. ACM* 60, 5 (April 2017), 22–24. https://doi.org/10.1145/3068787
[23] Ruogu Kang, Stephanie Brown, and Sara Kiesler. 2013. Why Do People Seek Anonymity on the Internet?: Informing Policy and Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2657–2666. https://doi.org/10.1145/2470654.2481368
[24] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. 2015. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, Ottawa, 39–52. https://www.usenix.org/conference/soups2015/proceedings/presentation/kang
[25] Waltraut Kotschy. 2016. *The new General Data Protection Regulation-Is there sufficient pay-off for taking the trouble to anonymize or pseudonymize data*.
[26] Gideon Kunda. 1992. *Engineering Culture: Control and Commitment in a High-tech Corporation*. Temple University Press.
[27] Alex Leavitt. 2015. "This is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 317–327. https://doi.org/10.1145/2675133.2675175
[28] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2017. Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology* (2017).
[29] Gregory J. Matthews and Ofer Harel. 2011. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys* 5 (2011), 1–29. https://doi.org/10.1214/11-SS074
[30] Naja Holten Møller and Pernille Bjørn. 2011. Layers in Sorting Practices: Sorting out Patients with Potential Cancer. *Computer Supported Cooperative Work (CSCW)* 20, 3 (June 2011), 123–153. https://doi.org/10.1007/s10606-011-9133-3
[31] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.
[32] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. 2017. Privacy Expectations and Preferences in an IoT World. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, Santa Clara, CA, 399–412. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/naeini
[33] Arvind Narayanan and Edward W Felten. 2014. No silver bullet: De-identification still doesn't work. *White Paper* (2014), 1–8.
[34] Arvind Narayanan, Joanna Huey, and Edward W. Felten. 2016. A Precautionary Approach to Big Data Privacy. In *Data Protection on the Move*. Springer, Dordrecht, 357–385. https://doi.org/10.1007/978-94-017-7376-8_13
[35] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 111–125.

[36] Paul Ohm. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review* 57 (2010), 1701. http://heinonline.org/HOL/Page?handle=hein.journals/uclalr57&id=1713&div=&collection=

[37] Catuscia Palamidessi. 2015. Quantitative approaches to the protection of private information: State of the art and some open challenges. In *International Conference on Principles of Security and Trust*. Springer, 3–7.

[38] Article 29 Data Protection Working Party. 2014. Opinion 05/2014 on Anonymisation Techniques. http://www.pdpjournals.com/docs/88197.pdf

[39] OPAL Project. 2018. About OPAL - OPAL Project. Retrieved February 13, 2018 from http://www.opalproject.org/about-us/

[40] Emilee Rader, Rick Wash, and Brandon Brooks. 2012. Stories As Informal Lessons About Security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*. ACM, New York, NY, USA, 6:1–6:17. https://doi.org/10.1145/2335356.2335364

[41] General Data Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)* 59 (2016), 1–88.

[42] Ira S. Rubinstein and Woodrow Hartzog. 2016. Anonymization and Risk. *Washington Law Review* 91 (2016), 703. http://heinonline.org/HOL/Page?handle=hein.journals/washlr91&id=719&div=&collection=

[43] Jeff Sedayao, Rahul Bhardwaj, and Nakul Gorade. 2014. Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues. In *2014 IEEE International Congress on Big Data*. 601–607. https://doi.org/10.1109/BigData.Congress.2014.92

[44] Irina Shklovski and Nalini Kotamraju. 2011. Online Contribution Practices in Countries That Engage in Internet Blocking and Censorship. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1109–1118. https://doi.org/10.1145/1978942.1979108

[45] Lucy Suchman. 2002. Located accountabilities in technology production. *Scandinavian Journal of Information Systems* 14, 2 (Jan. 2002). http://aisel.aisnet.org/sjis/vol14/iss2/7

[46] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671 (2000), 1–34.

[47] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[48] Peter P Swire. 2004. A model for when disclosure helps security: What is different about computer and network security. *J. on Telecomm. & High Tech. L.* 3 (2004), 163.

[49] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. 2017. Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12. *arXiv:1709.02753 [cs]* (Sept. 2017). http://arxiv.org/abs/1709.02753 arXiv: 1709.02753.

[50] Linnet Taylor. 2016. The ethics of big data as a public good: which public? Whose good? *Phil. Trans. R. Soc. A* 374, 2083 (Dec. 2016), 20160126. https://doi.org/10.1098/rsta.2016.0126

[51] Rick Wash. 2010. Folk models of home computer security. ACM Press, 1. https://doi.org/10.1145/1837110.1837125

[52] Rick Wash. 2012. Folk Security. *IEEE Security Privacy* 10, 6 (Nov. 2012), 88–90. https://doi.org/10.1109/MSP.2012.144

[53] Matt Wes. 2017. Looking to comply with GDPR? Here's a primer on anonymization and pseudonymization. *International Association of Privacy Professionals* (2017). Retrieved February 13, 2018 from https://iapp.org/news/a/looking-to-comply-with-gdpr-heres-a-primer-on-anonymization-and-pseudonymization/

[54] Philipp Winter and Stefan Lindskog. 2012. How the Great Firewall of China is Blocking Tor. In *Free and Open Communications on the Internet*. USENIX.

Received April 2018; revised July 2018; accepted September 2018