CHAPTER 3

Information Privacy
Challenges and Opportunities for Technology and Measurement

Heng Xu and Nan Zhang

For educators and social scientists, one of the most exciting opportunities afforded by digital technologies is the ease of collecting and analyzing massive datasets that capture individual behavior and interactions in social, educational, and organizational processes. With the promise to study social and behavioral phenomena in unprecedented detail, many social scientists are attempting the methodological transition from taking measurements through surveys and experiments to extracting measurements from large amounts of data (Paxton & Griffiths, 2017). Reflecting this paradigm shift, Groves (2011) coined the term "organic data" for datasets that are organically generated (e.g., by ubiquitous sensors, on digital platforms such as social media) without an explicit research design. While some organic data are crucial for the proper functioning of the platform or device that generates them (e.g., step count for a pedometer), some other types of organic data may simply represent "data exhaust" (Harford, 2014), that is, digital trail left by human activities with no immediate use (e.g., the exact timestamp of each step sensed by the pedometer). In terms of scientific research, the use of "organic data" stands in contrast to that of "designed data," which originate from surveys or experiments that are specifically designed for research purposes.

The explosive growth of organic data collection sets the stage for innovative measurements, research designs, and practical applications. Yet, by its very nature of capturing the "digital footprints" of human activities and interactions (McFarland et al., 2016), the collection of organic data tends to raise privacy concerns for those whose personal information is being collected and analyzed. In educational settings, the collection of student data is usually governed by privacy laws and regulations such as the Family Educational Rights and Privacy Act in the USA. In organizational settings in general, privacy concerns are often pronounced when data are collected by organizations, or at least within the

https://doi.org/10.1017/9781009099813.005 Published online by Cambridge University Press

scope of an organization, on their employees (rather than by researchers on voluntary participants of lab experiments). For example, privacy concerns may arise when organizations collect and analyze the social media activities of their employees or prospective employees (Lee, 2018), as people's activities on social media tend to blur the boundary between professional and personal contexts (Abril et al., 2012). Similarly, employees may be concerned about their privacy when an organization uses sensory devices to track their whereabouts and calculate their "time off task" (Ravid et al., 2020), even though the collected data could be extremely useful for research studies on improving operational efficiency and workforce productivity.

The privacy concerns stemming from the collection of organic data can only be compounded when researchers or practitioners *link* the collected data with external data sources, such as administrative datasets, to reveal even more information about the individuals, for example when Twitter data are joined with voter-registration records to understand not only the expressed sentiments of an individual but also their socioeconomic status (Barberá et al., 2015). While the effective integration of (high-volume) organic data with (high-quality) administrative records is known to address many challenges in traditional research design (e.g., small sample size, nonresponse rate; Groves 2011; Ruggles 2014), the threat posed by data integration to individual privacy is also well documented in the literature. For example, in a landmark study, Sweeney (2002) demonstrated how cross-referencing an anonymized medical dataset with voter-registration records (publicly available for Cambridge, MA) revealed the medical records of William Weld, then Governor of Massachusetts.

Privacy issues arising from the collection and use of organic datasets are obviously complex, especially in educational and organizational settings, and have important legal implications (Abril et al., 2012; Peterson, 2016). Since the goal of this chapter is to provide a global perspective on the topic, we forgo a comprehensive legal treatment of the subject because the applicable laws vary considerably from country to country[1] – even from state to state in the USA (Russom et al., 2011). Instead, we refer readers to the recent law reviews (e.g., for the USA: Neace, 2019; for Australia: Koelmeyer & Josey, 2019) for the legal perspective on this topic, and focus this chapter on discussing what *researchers* need to be aware of, and

[1] For example, compared with US courts, European courts are known to be less preoccupied with protecting free speech when there is a trade-off to be made with privacy protection (Walker, 2012).

what precautions they need to take in terms of privacy protection, when handling the collection and analysis of organic datasets, *assuming* that the collection and analysis procedures already pass the legal compliance processes in the participating organizations. To this end, we will discuss the following three questions in order in the rest of this chapter.

From the perspective of data privacy, what types of information about an individual may be inferred from an organic dataset being collected?

From the perspective of individuals' privacy concerns, will knowledge of the organic data collection, the potential inference of personal information, and the intended use of the collected data make those individuals whose information is being collected concerned about their privacy?

Are there technical tools available to ameliorate privacy concerns while maintaining the utility of the collected organic data for research?

It is important to note that all three questions represent active research ideas in multiple disciplines, from computer science (Dwork & Roth, 2014) to psychology (Acquisti et al., 2020). In educational and organizational research, researchers are also starting to investigate issues pertinent to these questions (e.g., Alge et al., 2006; Bhave et al., 2020; Ravid et al., 2021). Nonetheless, there is still a dearth of research work that crosses disciplinary boundaries in answering these questions. Thus, we conclude the chapter with a discussion of future research topics that call for interdisciplinary collaborations in studying privacy-related phenomena in educational and employment settings.

3.1 Data Privacy Issues in Organic-Data Collection

To understand what privacy issues may arise from the collection of an organic dataset, it is important to study three questions with regard to the potential of information disclosure from the collected data. The first question is about anonymity – this is, whether (parts of) the collected data could be linked back to a specific individual. The second question is on the feasibility of data inference – that is, among the parts of collected data that may be linked back to an individual, whether additional characteristics (e.g., demographics information) about the individual could be inferred from the collected data. Finally, the third question is about the interdependency between different individuals' private information – that is, whether the data linkable to one individual could be used to infer

https://doi.org/10.1017/9781009099813.005 Published online by Cambridge University Press

information about other individuals (e.g., colleagues, friends). We discuss these three questions respectively as follows.

3.1.1 *Anonymity*

A common misconception with regard to the anonymity afforded by a dataset is that no record can be linked back to an individual as long as all personal identifiable information (PII; e.g., name, national ID such as social security number) is removed from the data. Research in computer science has repeatedly shown that this is not the case. In a seminal work, Sweeney (2000) found that 87% of Americans can be uniquely identified based on a combination of ZIP code, gender, and date of birth, none of which is traditionally considered as PII. Since Sweeney's finding, considerable efforts were made to assess the risk of *reidentification* from a dataset stripped of PII. The ease of such reidentification was highlighted by a series of studies that demonstrated the feasibility of identifying an individual by linking the PII-free records with news stories (Yoo et al., 2018) or publicly available databases (Gymrek et al., 2013), exploiting the uniqueness of variable values (e.g., medical diagnosis code) in a record (Loukides et al., 2010; O'Neill et al., 2016), extracting patterns distinct to each individual from complex data types such as text (Jones et al., 2007), etc. Importantly, the risk of reidentification did not stay as a concern of academic interest but instead manifested as far-reaching incidents in practice. For example, soon after AOL released a dataset consisting of anonymized logs of search queries, a reporter was able to reidentify an AOL user by linking the released dataset with a public phonebook (Barbaro et al., 2006). Similarly, not long after Netflix released the anonymized movie ratings of its users, Narayanan and Shmatikov (2008) reidentified many Netflix users in the dataset by cross-referencing the movie ratings with those posted publicly on the Internet Movie Database website. Both incidents led to lawsuits on the ground of privacy violation and resulted in eventual settlements[2] from the companies releasing the datasets.

3.1.2 *Inference*

A key distinction between organic data and the traditional, designed data, is that the former is generated from a process beyond the control of a

[2] AOL: Case No. 1:11-cv-01014-CMH-TRJ (ED Va. Dec. 17, 2012). Netflix: In re Netflix Privacy Litigation, Case No. 5:11-cv-00379-EJD (ND Cal. Mar. 18, 2013).

researcher. In terms of privacy, what this means is that researchers may not be fully aware of the types of private information that could be inferred from an organic dataset, raising important privacy, ethical, and legal concerns for the collection and use of such a dataset (Oswald et al., 2020). This issue is particularly pronounced when the organic dataset contains variables that are rich in contextual information – for example, geolocations, text (e.g., emails), images, audios, and videos. Consider geolocation information as an example. While it ostensibly discloses only the whereabouts of an individual at a particular point in time, existing studies showed that geolocations shared over online social networks could be used to accurately infer a variety of demographic variables including age, gender, sexual orientation, education attainment, etc. (Zhong et al., 2015). Another example is the web-browsing history of an individual. Researchers often collect such information to gauge people's interests on certain topics (e.g., political view; Comarela et al., 2018). Yet web-browsing history has been shown to accurately reveal the gender and education attainment of an individual (Hu et al., 2007; Li et al., 2017). Perhaps unsurprisingly, similar inferences can be made based on a wide variety of context-rich variables, from an individual's search query logs (Weber & Castillo, 2010) to the user name chosen by an individual for an online social network (Wood-Doughty et al., 2018), from writings (even short ones like tweets; Yo & Sasahara, 2017) to audio records (Krauss et al., 2002), etc. It is important to note that this inference issue interacts with the aforementioned anonymity issue, as the additional demographic information being inferred also makes it more likely for a record to be linked back to an individual, for example, through cross-referencing with publicly available datasets like voter registration records. Once the linkage is established, the information in the public dataset then allows the inference of even more information about the individual, compounding the threat of private-information disclosure.

3.1.3 Interconnected Privacy

The information about one individual in an organic dataset could also be used to infer the information about others because the interdependency between different characteristics of the same individual, which leads to the inference issue discussed earlier, readily applies to the characteristics of different individuals. For example, researchers have long recognized the value of organic datasets in capturing the social relationships between different individuals (Knoke & Yang, 2019). Such relationships, in turn,

allow the inference of a wide variety of information for the individuals involved. For example, Wang et al. (2014) demonstrated that an individual's working relationships are the key for an accurate inference of work skills possessed by the individual, because such skills tend to be relatively homogeneous among people with close working relationships. Similarly, Dong et al. (2014) found that many demographic variables, including race, gender, and age groups, can be accurately inferred from an individual's social connections. Further, the way an individual's social connections change over time could also reveal the individual's demographic information (Dong et al., 2014). Note that the interdependency of different individuals' information extends beyond the *existence* of relationships between them, and may involve their interactions captured in the organic dataset. For example, Alsarkal et al. (2018) demonstrated how a conversation captured in a dataset, like a simple "Happy Birthday" message, could reveal the date of birth of an individual involved in the conversation.

3.2 Privacy Concerns Arising from Organic-Data Collection

Discussions in the last section clarified the many possible ways private information can be inferred from an organic dataset and linked back to an individual. While the disclosure of such information is clearly pertinent to the privacy of individual data subjects, it is important to recognize that people's need with regard to privacy is *not* about keeping all information about themselves secret, but about striking a proper balance between the need for disclosure and the need for secrecy (Acquisti et al., 2020). As such, a consensual view in the privacy research community is that whether the disclosure of certain private information is a "problem" – that is, whether such disclosure triggers people's *privacy concerns* – depends on the specific context, such as who the data subject is, why the information is disclosed, for what purpose, etc. Moreover, privacy concerns may arise for a wide variety of reasons: from the desire to keep certain information secret to the embarrassment of revealing certain activities outside the societal norm (Post, 2017), to concerns on the "Big Error" (Lazer et al., 2014; McFarland & McFarland, 2015) stemming from the hidden biases in the collected data. To understand how and why privacy concerns may arise from the collection and use of an organic dataset, it is important to examine two interrelated issues. The first is the conceptual issue of "*what is privacy*." The second is the operational issue of how to *measure* people's privacy concerns. We discuss these two issues respectively in the rest of this section.

https://doi.org/10.1017/9781009099813.005 Published online by Cambridge University Press

3.2.1 Conceptualization of Privacy

Although privacy has been extensively studied in social sciences (including philosophy, economics, psychology, law, and sociology), it is widely recognized that as a concept, privacy "is in disarray and nobody can articulate what it means" (Solove, 2006, p. 477). Scholars have proclaimed that "privacy is so muddled a concept that it is of little use" (Solove, 2007, p. 754). "Perhaps the most striking thing about the right to privacy," Thomson (1975) observes, "is that nobody seems to have any very clear idea what the right to privacy is" (p. 312). The wide scope of scholarly interests has resulted in a variety of conceptualizations of privacy, which leads Margulis (1977) to note that "theorists do not agree ... on what privacy is or on whether privacy is a behavior, attitude, process, goal, phenomenal state, or what" (p. 17).

Many efforts have been made by privacy scholars to develop a systematic understanding of privacy by integrating the different perspectives from different fields. The challenge, however, is that the conceptual picture that emerges is usually fragmented and discipline-specific (also see Smith et al., 2011 for a review). For instance, in law, perhaps the most famous conceptualization is to view "privacy as a right," first defined by Brandeis and Warren (1890) as "the right of the individual to be left alone" (p. 205). This stands in sharp contrast to the sociological view of privacy as a struggle for control between an individual and the society (Margulis 1977; Westin 1967), perhaps owing to the focus of sociology research on how the power and influence between individuals, groups, and institutions shape the collection and use of personal information in the society. Yet another conceptualization of privacy – in economics – is to define privacy as a value both in terms of its relevance to the information needed for efficient markets and its role as a piece of property (Acquisti et al., 2015). This is clearly distal to psychologists' view of privacy, which is often that of a perception or a feeling or an emotion. As Altman (1974) argues, there are many instances where no logical reason appears to exist for a person to feel that their privacy has been violated, yet that is precisely their perception. Almost all these conceptualizations have been developed (to different extents) by philosophers, who interpreted privacy as a state of "limited access or isolation" (Schoeman, 1984, p. 3), "being apart from others" (Weinstein, 1971, p. 626), etc.

The lack of conceptual clarity for privacy brings about considerable challenges not only for researchers but also for the government and organizations, making it extremely difficult for them to form coherent

policies regarding data practices. For research, the National Research Council (2011) found that having ad hoc definitions each capturing a small fraction of a complex social construct, without a common understanding of what the construct really means, leads to the balkanization of a field, sparse data, or even paucity of scholarly interest. For technological development, as Lederer et al. (2004, p. 440) pointed out, "one possible reason why designing privacy-sensitive systems is so difficult is that, by refusing to render its meaning plain and knowable, privacy simply lives up to its name." For news media, "privacy" is often used as a blanket term covering everything related to the unsettling consequences of applying the latest technologies (Hao, 2018). This, in turn, confuses the general public and leads to their difficulty in making decisions related to privacy, as observed by many existing studies (Debatin et al., 2009). In sum, while copious empirical evidence has shown privacy to be multidimensional, elastic, depending upon context, and dynamic in the sense that it varies with life experience, a lack of clear, concrete, measurable, and empirically testable conceptualization of privacy is still a major challenge facing today's privacy research and practices.

3.2.2 Measurement of Privacy Concerns

Given the clear lack of consensus on what "privacy" is, there has been a movement toward the measurement of people's *privacy concerns* as the central tenet in privacy research, as noted by Smith et al. (2011). A key reason for this focus is the prevailing belief that the recent wave of privacy laws and regulations, such as the European Union's General Data Protection Regulation, is mainly driven by policymakers' understanding, acknowledgment, and respect of citizens' privacy concerns (Solove, 2021). For social scientists collecting organic datasets, asking the data subjects about their privacy concerns also appears to be a straightforward solution because, intuitively, privacy only becomes a problem when those individuals whose private information is being collected and used are *concerned* about such collection and use.

Given the practical pertinence of measuring people's privacy concerns, the literature is replete with attempts to elicit self-reported privacy concerns with survey instruments, such as the Concern for Information Privacy (CFIP) instrument developed by Smith et al. (1996). Yet accurately gauging people's privacy concerns from self-reported data is known to be a challenge. In developing the 2016 US National Privacy Research Strategy, a subcommittee of the National Science and Technology Council

(2016) notes that people's self-reported privacy concerns are often diverse, dynamic, and situation-specific, not only challenging its reliable measurement at the individual level but making it "difficult to draw general conclusions about current privacy norms or predict how these norms may develop over time" (p. 10). Specifically, self-reported privacy concerns have been criticized for suffering from two main problems, *inflation* and *uncertainty* (Xu & Zhang, 2022).

The criticism of inflation mostly arises from the frequent observation that people could express heightened privacy concerns yet refuse to take even trivial actions to protect their own privacy (Acquisti & Grossklags, 2005; Beresford et al., 2012). For this reason, researchers and practitioners often cite people's self-reported privacy concerns as inflated, exaggerating their "true" level of concerns about privacy (Wittes & Liu, 2015). The literature has also noted multiple potential causes for the inflated responses. For example, Solove (2021) contended that inflated responses are natural because the survey instruments researchers usually use to elicit people's privacy concerns (e.g., the aforementioned CFIP) do not specify the context of data collection or use in sufficient detail. As such, it is perfectly reasonable for an individual to express a high level of *general* concerns yet not be concerned about privacy in a specific context. Complementary to this point is Hong and Thong's (2013) argument that, when privacy concerns are elicited in a hypothetical manner, inflated responses should be expected because a respondent could construe "privacy concerns" as their *expectation* of others' behavior in an ideal world (e.g., whether an ideal organization *should* collect its employee's private information). In other words, there is not really any cost (or trade-off to be considered) associated with an inflated response. Providing further evidence to the inflation of self-reported responses, Marreiros et al. (2017) demonstrated that self-reported privacy concerns became inflated immediately after exposure to information about privacy (e.g., after reading a newspaper article), no matter if such exposure is positive, neutral, or negative. In other words, even those individuals who were not very concerned about the collection of their private information could report a high level of concern after hearing *something* about privacy. This challenges the fundamental feasibility of eliciting people's "true" privacy concerns through a survey questionnaire.

Besides inflation, another frequent criticism is that people are often *uncertain* about their attitudes toward privacy concerns, leading to excessive variability in their self-reported privacy concerns. Uncertainty, or a respondent's lack of an attitude in a coherent form, is a common issue for self-

reported data (Bertrand & Mullainathan, 2001). What is unique for self-reported privacy concerns is that the uncertainty is not limited to respondents who are inattentive (Lelkes et al., 2012) but applies to a large part of the respondents. For example, many respondents do not seem to know the consequences of disclosing certain private information (Solove, 2021). Many others are unsure about how they feel toward privacy (Acquisti et al., 2015). Yet others are uncertain about whether the common tools for privacy protection (e.g., virtual private network, or VPN) are indeed effective (Gates, 2011). This excessive uncertainty has two pronounced consequences. One is the cue-seeking behavior it induces. That is, when survey respondents are reluctant to admit their uncertainties (e.g., for fear of being perceived as ignorant or naive; Acquisti et al., 2015), they tend to "cast around" for cues when answering survey questions about their privacy concerns (Adjerid et al., 2018). Unfortunately, such cues are rarely relevant to privacy, instead mostly concerning the design and appearance of a survey instrument (John et al., 2011) or even the physical environment surrounding the respondent when answering the question (Acquisti et al., 2015). The second consequence of excessive uncertainty is a phenomenon known as "privacy cynicism" (Hoffmann et al., 2016). That is, many people deliberately "discount risks or concerns" in order to cope with their uncertainty (Hoffmann et al., 2016). In either case, as different people respond to uncertainty in different ways (Powell & Baker, 2014), we tend to observe even more randomness in the self-reported privacy concerns.

To address the two criticisms, privacy scholars spent considerable efforts in recent years to refine the measurement of privacy concerns, in particular by carefully examining the role of *context* in the measurement process. Conceptually, Nissenbaum (2020) linked privacy concerns to a set of context-dependent situational norms. Operationally, researchers started adopting general-purpose survey instruments for measuring privacy concerns in specific contexts (e.g., Xu et al., 2012). In an educational or organizational setting, the proper contextualization of privacy-concern measurement is even more important given the empirical evidence that people's privacy concerns tend to differ considerably in workplace and personal settings (Auxier et al., 2019), and adults and adolescents tend to cope with their privacy concerns in different ways (Jia et al., 2015). To this end, researchers may benefit from a recently developed quantitative framework (Xu & Zhang, 2022) that estimates the degree of inflation and uncertainty from self-reported privacy concerns, making it possible for researchers to identify a proper contextualization and to correct both forms of bias when using self-reported privacy concerns.

3.3 Technological Solutions for Privacy Protection

In addition to the quest of understanding people's privacy concerns, privacy researchers also pursued a technical solution to the problem of privacy protection, with the goal being to maintain the utility of the organic dataset for research while eliminating the collection or inference of private information. The key rationale here is an observation that, even though an organic dataset usually consists of data about individuals, social scientists rarely use such datasets to study one individual specifically. Instead, the research goal is almost always to identify patterns or relationships that hold for a large part of the population. As research in statistics has long shown the feasibility of recovering accurate patterns from datasets that have undergone drastic changes (e.g., noise insertion) at the individual-record level (Osborne, 1991), the idea of the technical solution for privacy protection is to develop a process for *anonymizing* a dataset so as to block all potential links to the individual data subjects yet maintain the patterns and relationships that are of research interest. The existing techniques for data anonymization can be largely partitioned into two categories, *data removal* and *noise insertion*, respectively, which we discuss as follows.

3.3.1 Data Removal

Since the goal of data anonymization is to prevent any individual from being identified from the anonymized dataset, a natural idea for anonymizing a dataset is to remove the part of data that could be used to identify an individual. The data removal mechanism is rooted in this idea. Its initial implementations focused on removing variables that are obvious identifiers, such as name, address, social security number, etc. These implementations were challenged by the aforementioned discovery (Sweeney, 2000) that ZIP code, gender, and date of birth in combination can uniquely identify 87% of Americans. Following this discovery, a plethora of data removal techniques were developed to detect and rectify the issues caused by such "quasi-identifiers" (e.g., Machanavajjhala et al., 2007; Sweeney, 2002), forming the bulk of the existing literature for the data removal techniques (for a review, see Fung et al., 2010). A common idea followed by these techniques is to first identify the individuals at risk of being identified before removing the minimum necessary information to block such identifications. The removal of information could be operationalized by marking a value as missing (i.e., replacing its real value with N/A) or by

https://doi.org/10.1017/9781009099813.005 Published online by Cambridge University Press

reducing the precision of a value through granularity change (e.g., replacing Los Angeles with California).

A key advantage of data removal techniques is their transparency to the subsequent data analysis processes. Since most data removal techniques alter only the values in a dataset but not the structure of the dataset, a researcher who uses the dataset can readily apply any traditional statistical analysis tool over the privacy-preserved data with little change. Unfortunately, this advantage turns into a problem when researchers start integrating an anonymized dataset with other data sources. Specifically, the computer science literature has noted that, without making substantial assumptions about what the other data sources are, it is simply impossible for a data removal mechanism to block all possible links to an individual without making the data useless for research (Ganta et al., 2008). In other words, while data removal is highly attractive due to its simplicity of operation, it does not offer any rigorous anonymity guarantee that can withstand potential integration with other datasets. For this reason, there is a sharp divide between researchers and practitioners on the use of data removal mechanisms today. On the one hand, data removal remains popular in practice, frequently used or recommended by firms and government agencies for compliance with privacy laws and regulations (Article 29 Data Protection Working Party, 2014; Rocher et al., 2019). Yet, on the other hand, technical research on data removal all but ceased in the last decade, with technical researchers shifting their focus to noise insertion techniques, which we discuss next.

3.3.2 Noise Insertion

As discussed earlier, research in statistics, specifically calibration (Osborne, 1991), has long demonstrated the feasibility of recovering summary statistics from noise-ridden data. Leveraging this finding, earlier work on noise insertion simply anonymized a dataset by adding independent and identically distributed Gaussian noise to variable values before developing statistical techniques to recover the statistics of interest (Agrawal & Srikant, 2000). Unfortunately, this method was later found to be ineffective for privacy protection because the inherent correlation between different variables in the original dataset allows spectral methods (Bernardi & Maday, 1997) to separate the independent noise from the correlated data, thereby nullifying the anonymization achieved through noise insertion (Huang et al., 2005). This problem was solved by the development of differential privacy (Dwork et al., 2006), which guarantees that, for any

individual in the dataset and any statistics of the dataset, the statistics would be indistinguishable from the same statistics of a dataset that has the individual's record removed. In other words, an individual's privacy derives from the fact that no one can learn anything new from the dataset that it cannot already learn from a dataset without the individual's data.

A key advantage of differential privacy, compared with the data removal mechanism, is that it provides a rigorous guarantee that holds *no matter* what external data sources may be available. This makes differential privacy extremely attractive in research and practice. As a result, it became the de-facto standard for the noise insertion mechanism today, and has been adopted by high-tech firms such as Apple (Tang et al., 2017) and Google (Erlingsson et al., 2014) as well as government agencies such as the US Census Bureau in its 2020 decennial Census (Abowd, 2018). To researchers who use the (anonymized) dataset, however, differential privacy presents a challenge, as the majority of implementations for differential privacy do not allow researchers access to the raw dataset, instead requiring them to interactively query the data to obtain (noisy) estimates of statistics required for research. This means that the traditional tools for statistical analysis cannot be directly applied. Instead, new tools need to be developed that take into account the way differential privacy performs noise insertion. While differentially private tools for tasks like linear regression are available (Wang, 2018), there are many other statistical analysis tools, like structural equation modeling, for which no differentially private version has been developed.

3.4 Future Research

It is clear from the earlier discussions that there is still much to be learned about the information privacy issues surrounding organic data collection and use, especially in an educational or workplace setting. While numerous inference channels have been identified, we do not yet have the full picture of what private information may be inferred from an organic dataset. The scientific community has not yet converged on a consensual definition of privacy, nor the effective means to capture people's privacy concerns. The development of technical solutions for privacy protection is also a work in progress. While all these unknowns may lead to pessimism for educational or organizational researchers wanting to take advantage of the rich knowledge afforded by organic datasets, we note that they also represent opportunities for them to contribute to the literature of information privacy, in particular to the understanding of privacy issues in educational and

workplace settings. For this reason, we conclude the chapter with a discussion of future research that calls for the participation of educational and organizational researchers in the interdisciplinary efforts required to address the existing and emerging privacy challenges.

As discussed earlier, the scientific community has not formed a consensus on the definition of privacy in decades, and will unlikely converge on a consensus anytime soon. Interestingly, there are two distinct paths through which privacy researchers are attempting to address this lack of consensus. On the one hand, there are behavioral scientists who are making extensive efforts to *refine* the concepts of privacy by revealing more and more factors that affect people's perceptions or concerns of privacy (Dinev et al., 2015; Smith et al., 2011), including attention, cognition, emotion, motivation, environment, etc. On the other hand, computer scientists who develop technical solutions for privacy protection tend to treat the conceptualization of privacy as an afterthought, articulated not according to what it means to everyday people but based on what is mathematically feasible to achieve (e.g., the aforementioned different privacy guarantee). We argue that both paths could be detrimental to addressing the privacy issues *in practice*. While further refinement of privacy conceptualization could reveal factors that affect people's attitudes, beliefs, and perceptions about privacy invasions, it also risks overcomplicating the problem and producing a (somewhat defeatist) belief that everything related to privacy is fluid and must be examined on a case-by-case basis. On the other hand, oversimplifying a complex concept like privacy by rolling all of its dimensions into a singular technical definition is problematic too, as doing so may lead to a technical solution that is designed to meet everyone's need but indeed satisfies no one, causes confusions among everyday people, and make it difficult for them to make decisions with regard to the use of technical solutions (Debatin et al., 2009).

We believe that one way to address the problem of overcomplication and oversimplification in privacy research is to develop middle-range (Merton, 1968), context-contingent theories that, instead of attempting to identify the overarching features of privacy that operate in all social processes, simply aim to consolidate the empirical regularities related to privacy concerns and behavior in a specific set of similar contexts, like educational or workplace settings. A key reason behind our argument for developing context-contingent theories is the recognition that there is already ample empirical evidence suggesting the wide variation of privacy-related phenomena across contexts. For example, privacy concern and privacy-seeking behavior were found to be strongly correlated within

one context (Dienlin & Trepte, 2015), virtually uncorrelated in another context (Reynolds et al., 2011), and negatively correlated in a third context (Sheehan & Hoy, 1999). As there is little evidence that an overarching theory actually exists to explain the privacy phenomena across all contexts, from a practical perspective, it may be more productive to pursue a context-contingent theory that can *approximate* what privacy means to most people in a set of specific contexts, so as to enable the development of practical solutions that can address most people's privacy concerns in these contexts. Clearly, developing context-contingency theories requires domain expertise for the corresponding contexts. Thus, we believe that the participation of educational and organizational researchers is essential for advancing our understanding of the privacy phenomena in educational and workplace settings, especially given the rapidly increasing popularity of Big Data technologies that use organic data collection and analysis to improve operational efficiency.

REFERENCES

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2867–2867).

Abril, P. S., Levin, A., & Del Riego, A. (2012). Blurred boundaries: Social media privacy and the twenty-first-century employee. *American Business Law Journal*, *49*(1), 63–124.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514.

(2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, *30*(4), 736–758.

Acquisti, A., & Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE Security & Privacy*, *3*(1), 26–33.

Adjerid, I., Peer, E., & Acquisti, A. (2018). Beyond the privacy paradox: Objective versus relative risk in privacy decision making. *MIS Quarterly*, *42*(2), 465–488.

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (pp. 439–450).

Alge, B. J., Ballinger, G. A., Tangirala, S., & Oakley, J. L. (2006). Information privacy in organizations: Empowering creative and extrarole performance. *Journal of Applied Psychology*, *91*(1), 221–232.

Alsarkal, Y., Zhang, N., & Xu, H. (2018). Your privacy is your friend's privacy: Examining interdependent information disclosure on online social networks. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 892–901).

Altman, I. (1974). Privacy: A conceptual analysis. In D. H. Carson (Ed.), *Man-environment interactions: Evaluations and applications: Part 2* (pp. 13–28). Environmental Design Research Association.

Article 29 Data Protection Working Party. (2014). *Opinion 05/2014 on anonymisation techniques*. https://ec.europa.eu/justice/article-29/documentation/

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019). *Americans and privacy: Concerned, confused and feeling lack of control over their personal information*. Pew Research Center. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Barbaro, M., Zeller, T., & Hansell, S. (2006). A face is exposed for AOL searcher no. 4417749. *New York Times*. https://www.nytimes.com/2006/08/09/technology/09aol.html

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological Science*, *26*(10), 1531–1542.

Beresford, A. R., Kübler, D., & Preibusch, S. (2012). Unwillingness to pay for privacy: A field experiment. *Economics Letters*, *117*(1), 25–27.

Bernardi, C., & Maday, Y. (1997). Spectral methods. In P. G. Ciarlet & J. L. Lions (Eds.), *Handbook of numerical analysis* (Vol. 5; pp. 209–485). Elsevier.

Bertrand, M., & Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *American Economic Review*, *91*(2), 67–72.

Bhave, D. P., Teo, L. H., & Dalal, R. S. (2020). Privacy at work: A review and a research agenda for a contested terrain. *Journal of Management*, *46*(1), 127–164.

Brandeis, L., & Warren, S. (1890). The right to privacy. *Harvard Law Review*, *4*(5), 193–220.

Comarela, G., Durairajan, R., Barford, P., Christenson, D., & Crovella, M. (2018). Assessing candidate preference through web browsing history. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 158–167).

Debatin, B., Lovejoy, J. P., Horn, A. K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, *15*(1), 83–108.

Dienlin, T., & Trepte, S. (2015). Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviors. *European Journal of Social Psychology*, *45*(3), 285–297.

Dinev, T., McConnell, R. A., & Smith, H. J. (2015). Informing privacy research through information systems, psychology, and behavioral economics: Thinking outside the "APCO" box. *Information Systems Research*, *26*(4), 639–655.

Dong, Y., Yang, Y., Tang, J., Yang, Y., & Chawla, N. V. (2014, August). Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 15–24).

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265–284). Springer.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407.

Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1054–1067).

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, *42*(4), 1–53.

Ganta, S. R., Kasiviswanathan, S. P., & Smith, A. (2008, August). Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 265–273).

Gates, G. W. (2011). How uncertainty about privacy and confidentiality is hampering efforts to more effectively use administrative records in producing U.S. national statistics. *Journal of Privacy and Confidentiality*, *3*(2), 3–40.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75* (5), 861–871.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, *339*(6117), 321–324.

Hao, K. (2018, October 21). Establishing an AI code of ethics will be harder than people think. *MIT Technology Review*.

Harford, T. (2014). Big data: A big mistake?. *Significance*, *11*(5), 14–19.

Hoffmann, C. P., Lutz, C., & Ranzini, G. (2016). Privacy cynicism: A new approach to the privacy paradox. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *10*(4).

Hong, W., & Thong, J. Y. (2013). Internet privacy concerns: An integrated conceptualization and four empirical studies. *MIS Quarterly*, *37*(1), 275–298.

Hu, J., Zeng, H. J., Li, H., Niu, C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 151–160).

Huang, Z., Du, W., & Chen, B. (2005). Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (pp. 37–48).

Jia, H., Wisniewski, P. J., Xu, H., Rosson, M. B., & Carroll, J. M. (2015). Risk-taking as a learning process for shaping teen's online information privacy behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 583–599).

John, L. K., Acquisti, A., & Loewenstein, G. (2011). Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of Consumer Research*, *37*(5), 858–873.

Jones, R., Kumar, R., Pang, B., & Tomkins, A. (2007). "I know what you did last summer" query logs and user privacy. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (pp. 909–914).

Knoke, D., & Yang, S. (2019). *Social network analysis*. Sage Publications.

Koelmeyer, A., & Josey, N. (2019). Employment and privacy: Consent, the 'privacy act' and biometric scanners in the workplace. *LSJ: Law Society of NSW Journal*, *57*, 76–77.

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*(6), 618–625.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Lederer, S., Hong, J. I., Dey, A. K., & Landay, J. A. (2004). Personal privacy through understanding and action: Five pitfalls for designers. *Personal and Ubiquitous Computing*, *8*(6), 440–454.

Lee, D. (2018, November 27). Predictim babysitter app: Facebook and Twitter take action. *BBC News*. https://bbc.com

Lelkes, Y., Krosnick, J. A., Marx, D. M., Judd, C. M., & Park, B. (2012). Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology*, *48*(6), 1291–1299.

Li, H., Zhu, H., & Ma, D. (2017). Demographic information inference through meta-data analysis of Wi-Fi traffic. *IEEE Transactions on Mobile Computing*, *17*(5), 1033–1047.

Loukides, G., Denny, J. C., & Malin, B. (2010). The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, *17*(3), 322–327.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). *l*-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3–es.

Margulis, S. T. (1977). Conceptions of privacy: Current status and next steps. *Journal of Social Issues*, *33*(3), 5–21.

Marreiros, H., Tonin, M., Vlassopoulos, M., & Schraefel, M. C. (2017). "Now that you mention it": A survey experiment on information, inattention and online privacy. *Journal of Economic Behavior & Organization*, *140*, 1–17.

McFarland, A. D., Lewis, K., & Goldberg, A. (2016). Sociology in the era of big data: The ascent of forensic social science. *American Sociologist*, *47*, 12–35.

McFarland, D. A., & McFarland, H. R. (2015). Big data and the danger of being precisely inaccurate. *Big Data & Society*, July–December, 1–4.

Merton, R. K. (1968). *Social theory and social structure*. Simon & Schuster.

National Research Council. (2011). *The importance of common metrics for advancing social science theory and research: A workshop summary*. National Academies Press.

National Science and Technology Council. (2016). *National privacy research strategy*. https://www.nitrd.gov/pubs/NationalPrivacyResearchStrategy.pdf

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy* (pp. 111–125). IEEE.

Neace, G. (2019). Biometric privacy: Blending employment law with the growth of technology. *UIC Law Review*, *53*, 73–112.

Nissenbaum, H. (2020). *Privacy in context*. Stanford University Press.

O'Neill, L., Dexter, F., & Zhang, N. (2016). The risks to patient privacy from publishing data from clinical anesthesia studies. *Anesthesia & Analgesia*, *122* (6), 2017–2027.

Osborne, C. (1991). Statistical calibration: A review. *International Statistical Review*, *59*(3), 309–336.

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 505–533.

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638.

Peterson, D. (2016). Edtech and student privacy: California law as a model. *Berkeley Technology Law Journal*, *31*(2), 961–995.

Post, R. C. (2017). Data privacy and dignitary privacy: Google Spain, the right to be forgotten, and the construction of the public sphere. *Duke Law Journal*, *67*, 981–1072.

Powell, E. E., & Baker, T. (2014). It's what you make of it: Founder identity and enacting strategic responses to adversity. *Academy of Management Journal*, *57*(5), 1406–1433.

Ravid, D. M., Tomczak, D. L., White, J. C., & Behrend, T. S. (2020). EPM 20/20: A review, framework, and research agenda for electronic performance monitoring. *Journal of Management*, *46*(1), 100–126.

Ravid, D. M., White, J. C., & Behrend, T. S. (2021). Implications of COVID-19 for privacy at work. *Industrial and Organizational Psychology*, 14(1–2), 194–198.

Reynolds, B., Venkatanathan, J., Gonçalves, J., & Kostakos, V. (2011, September). Sharing ephemeral information in online social networks: Privacy perceptions and behaviours. In *IFIP Conference on Human-Computer Interaction* (pp. 204–215). Springer.

Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*(1), 1–9.

Ruggles, S. (2014). Big microdata for population research. *Demography*, *51*(1), 287–297.

Russom, M. B., Sloan, R. H., & Warner, R. (2011). Legal concepts meet technology: A 50-state survey of privacy laws. In *Proceedings of the*

2011 Workshop on Governance of Technology, Information, and Policies (pp. 29–37).

Schoeman, F. D. (Ed.). (1984). *Philosophical dimensions of privacy: An anthology*. Cambridge University Press.

Sheehan, K. B., & Hoy, M. G. (1999). Flaming, complaining, abstaining: How online users respond to privacy concerns. *Journal of Advertising*, *28*(3), 37–51.

Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: An interdisciplinary review. *MIS Quarterly*, *35*(4), 989–1015.

Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, *20*(2), 167–196.

Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review*, *154*(3), 477–560.

(2007). 'I've got nothing to hide' and other misunderstandings of privacy. *San Diego Law Review*, *44*, 745–772.

(2021). The myth of the privacy paradox. *George Washington Law Review*, *89*(1), 1–51.

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

(2002). *k*-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570.

Tang, J., Korolova, A., Bai, X., Wang, X., & Wang, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. *arXiv preprint arXiv:1709.02753*.

Thomson, J. J. (1975). The right to privacy. *Philosophy & Public Affairs*, *4*(4), 295–314.

Walker, R. K. (2012). The right to be forgotten. *Hastings Law Journal*, *64*, 257–286.

Wang, Y. X. (2018). Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence* (pp. 93–103).

Wang, Z., Li, S., Shi, H., & Zhou, G. (2014). Skill inference with personal and skill connections. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical papers* (pp. 520–529).

Weber, I., & Castillo, C. (2010). The demographics of web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 523–530).

Weinstein, M. A. (1971). The uses of privacy in the good life. In J. R. Pennock & J. W. Chapman (Eds.), *Nomos XIII: Privacy* (pp. 624–692). Atherton Press.

Westin, A. F. (1967). *Privacy and freedom*. Atheneum.

Wittes, B., & Liu, J. C. (2015, May 21). *The privacy paradox: The privacy benefits of privacy threats*. Center for Technology Innovation at Brookings.

Wood-Doughty, Z., Andrews, N., Marvin, R., & Dredze, M. (2018). Predicting Twitter user demographics from names alone. In *Proceedings of the 2nd*

Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (pp. 105–111).

Xu, H., Teo, H. H., Tan, B. C., & Agarwal, R. (2012). Effects of individual self-protection, industry self-regulation, and government regulation on privacy concerns: A study of location-based services. *Information Systems Research*, *23*(4), 1342–1363.

Xu, H., & Zhang, N. (2022). From contextualizing to context-theorizing: Assessing context effects in privacy research. *Management Science*, *68*(10), 7065–7791.

Yo, T., & Sasahara, K. (2017). Inference of personal attributes from tweets using machine learning. In *2017 IEEE International Conference on Big Data* (pp. 3168–3174). IEEE.

Yoo, J. S., Thaler, A., Sweeney, L., & Zang, J. (2018). Risks to patient privacy: A re-identification of patients in Maine and Vermont statewide hospital data. *Journal of Technology and Science Education*, *2018*, 2018100901.

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., & Xie, X. (2015). You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining* (pp. 295–304).