5-2019

Responding to Some Challenges Posed by the Re-identification of Anonymized Personal Data

Herman T. Tavani
Rivier College

Frances Grodzinsky, ed.
Sacred Heart University, grodzinskyf@sacredheart.edu

Responding to Some Challenges Posed by the Re-identification of Anonymized Personal Data

Herman T. Tavani
Rivier University

Frances S. Grodzinsky
Sacred Heart University

Abstract

In this paper, we examine a cluster of ethical controversies generated by the *re-identification* of anonymized personal data in the context of big data analytics, with particular attention to the implications for personal privacy. Our paper is organized into two main parts. Part One examines some ethical problems involving re-identification of personally identifiable information (PII) in large data sets. Part Two begins with a brief description of Moor and Weckert's Dynamic Ethics (DE) and Nissenbaum's Contextual Integrity (CI) Frameworks. We then investigate whether these frameworks, used together, can provide us with a more robust scheme for analyzing privacy concerns that arise in the re-identification process (as well as within the larger context of big data analytics). This paper does not specifically address re-identification-related privacy concerns that arise in the context of the European Union's General Data Protection Regulation (GDPR). Instead, we examine those issues in a separate work.

Keywords: Re-identification, Privacy, Anonymized personal data, Big Data Analytics, Contextual Integrity, Dynamic Ethics

In this paper, we examine a cluster of ethical controversies generated by the *re-identification* of anonymized personal data in the context of big data analytics, with particular attention to the implications for personal privacy. By "re-identification," we mean the process of "using anonymized data to find individuals in public datasets" (Rijmenam 2016**)**. This process is often contrasted with "de-identification," which has been defined as a process for "treating sensitive personal data regarding individuals in such a way that the individuals cannot be identified" (https://www.cippguide.org/2010/09/21/de-identification-re-identification/). Typically, the de-identification of personal data involves the removal of certain specified (personal) identifiers — name, social security number, etc. One significant controversy involving re-identification (and big data analytics) is:

> *Privacy rules that apply to personally identifiable information (PII) do not apply to (de-identified) anonymized personal data, some of which can now be fairly easily re-identified.*

Since many people already are, or will soon be, affected by big data and its implications for the re-identification of their PII, we believe that a clear, explicit, and transparent privacy framework is needed. And because the field of big data (for a definition of big data, see Grodzinsky 2017) is also evolving in connection with other "emerging technologies," we believe that a robust privacy framework might also help us to address related issues that arise in those contexts as well. We argue that such a framework is possible if we use two frameworks, Moor and Weckert's Dynamic Ethics (DE) and Nissenbaum's Contextual Integrity (CI) in tandem, to better enable us to develop clearer and more robust policies that will protect users from the kinds of current privacy violations at risk via the re-identification of their previously anonymized personal data using big data analytics.

Our paper is organized into two main parts. Part One examines some ethical problems involving re-identification of personally identifiable information (PII) in large data sets. Part Two begins with a brief description of Moor and Weckert's DE and Nissenbaum's CI Frameworks. We then investigate whether these frameworks, used together, can provide us with a more robust scheme for analyzing privacy concerns that arise in the re-identification process (as well as within the larger context of big data analytics).

PART I:
Re-identification and De-identification in the Context of Big Data

Both "re-identification" and "de-identification" have challenged privacy in the context of big data. Big data is also evolving in connection with other "emerging technologies," such as ambient intelligence (AmI)/pervasive computing; therefore, a robust privacy framework might also help us to address related issues that arise in those contexts as well. Another challenge arises in the use of "scrubbed data" and netadata.

Re-identification of "Scrubbed" Data

Scrubbing data involves removing its identifying data. Lubarsky (2017) points out that in the case of PII, such as one's name, medical condition, social security number, etc., we have some legal protections regarding both its sale and disclosure. But he goes on to note that once data containing PII is "scrubbed," it can then be considered "anonymized data." Lubarsky also points out that scrubbed data is now commonly re-identified simply by "combining two or more sets of data to find the same user in both."

Lubarsky suggests that we can view personal data as residing on what he refers to as a "spectrum of identifiability," and he uses the metaphor of a staircase to illustrate this point. At the top of this hierarchy is personal data that can *directly* identify the individual (for example, one's name, telephone number, social security number, and so forth). The "second step" in this ladder is where we find the kind of indirect *identification* which includes data that is "unambiguously linked to an individual." (Lubarsky points out that 63% of the population can be uniquely identified by the combining three pieces of data: gender, date of birth, and zip code). Whereas the third step in Lubarsky's

“staircase” includes data can be “ambiguously” connected to multiple people—restaurant preferences, movies, etc.—the fourth step includes data that cannot be linked to any specific person (e.g., aggregated census data).

Lubarsky also describes a collection of approaches/tools/algorithms that can be used in the process of “scrubbing,” or removing, the “identifying information” from data. To scrub a data set of this kind of information, Lubarsky notes that four main categories or techniques can be used: “(1) removing data; (2) replacing data with pseudonyms; (3) adding statistical ‘noise’; and (4) aggregation.” Whereas (1) and (2) are used mainly for direct identifiers, (3) and (4) are used for indirect identifiers.” We will not elaborate on these techniques for “scrubbing data, since our main objective is in analyzing ethical aspects of anonymized data once it has become re-identified.

Some Recent Challenges Involving the Use of Metadata in Re-Identifying Individuals

Until recently, it had been assumed by many researchers working in the area of de-identifying personal data that “metadata” was not important, or at least was far less important than other kinds of data, with regard to concerns that critics had raised about re-identification issues. For example, Perez et al. (2018) note that metadata is still categorized as “non-sensitive,” even though it is the kind of data most produced by us in our “daily interactions and communications in the digital world.” In a case study involving Twitter users, Perez et al. set out to “quantify the uniqueness of the association between metadata and user identity” and also to understand better the effectiveness or ineffectiveness of potential “obfuscation strategies.” Perez et al. went on to note that via a “supervised learning algorithm,” they were able to “identify any user in a group of 10,000 people, with approximately 97.6% accuracy”.

Now, however, metadata is beginning to be seen as an increasingly important source as well. As Devlon (2018) points out, “geolocation metadata” acquired over time via apps in fitness wearables or smartphones “can produce unique, repeatable data patterns that can be directly associated with individuals.” And Solon (2018) points to some examples in which “metadata, rather than or in addition to actual content, has been at the heart of the re-identification process.” Solon also notes that metadata fits into a category that she describes as “context-setting information,” and she further points out that the amount of context-setting information currently being collected and stored will likely increase dramatically, as we continue our transition to autonomous vehicles and a cashless economy. While some might find this astonishing, Paul Ohm, who has raised concerns about what he calls “the surprising failure of anonymization,” has noted that “for almost every person on earth there is at least one fact about them stored in a computer database that an adversary could use” against that person (cited in Devlon 2018).

Some Controversial Cases Involving Re-identification

Two now classic cases—the Weld and Netflix cases—provided some earlier illustrations of how the re-identification of personal data could have serious implications for personal privacy. (We examine these and other cases in more detail in Part II, where we apply our framework that incorporates key elements of DE and CI.) The first case involves the re-identifying of (former Massachusetts governor) William Weld, based on the comparison of seemingly "random" pieces of data. Latayna Sweeney (supervising a graduate student at MIT) was able to re-identify Weld from three anonymized datasets. Sweeney (2006) described how she used a public dataset released by the Massachusetts Group Insurance Commission (aimed at improving healthcare and controlling costs in Massachusetts), a voter list, and some other pieces of data in successfully re-identifying Weld (https://epic.org/privacy/reidentification/). (See also https://www.cippguide.org/2010/09/21/de-identification-re-identification/).

The controversial case involving Netflix had exposed some of the names of the 500,000 people who participated in an online contest sponsored by Netflix in 2006. In 2010 Arvind Narayanan (along with his team) was able to re-identify several individuals in the Netflix dataset of de-identified people (especially those who made more than nine entries). This incident resulted in Netflix being sued for privacy violation (under the Video Privacy Protection Act), and it also led to Netflix cancelling its second (planned) contest **(**Rijmenam 2016**).**

In a more recent case, Culnane et al. (2017) describe some ways in which patients in a "de-identified open health data set" in Australia were able to be successfully re-identified. The authors note that in August 2016, the Australian government, in compliance with its policy of open government data, had the federal Department of Health publish (online) the de-identified longitudinal medical billing records. The records in this data set included data pertaining to approximately 10% of Australian population, which amounted to roughly to 2.9 million people. In September 2016, Culnane et al. were able to decrypt the "IDs of suppliers" (doctors, midwives, and so forth) in that published report, and they informed the Department of Health about what they had been able to do. The Department then removed the dataset with the online billing records from online access. However, the authors also showed the government how easily patients could be re-identified without using decryption. For example, they noted that they were also able to link the "unencrypted parts of the record with known information about the individual." Culnane et al. were careful to point out that their primary aim had been to "inform" the data-sharing policy that had recently been put in place in Australia, by offering the government a "scientific demonstration of the ease of re-identification of the kind of data involved."

In order to encourage website developers to better their users' experience, third party companies use "session replay scripts" whose purpose is to record a user's actions while visiting a website or using an application on a mobile device in order to collect information used in data analytics. (Grodzinsky et al. 2018). One problem that has arisen is that the extent of gathered data is intrusive, going well beyond the stated objectives, and happening without user knowledge or consent. Unless manually redacted by the application developer (a clause in a third party company policy) to insure user privacy, the third party company could re-identify users who believed that their private information was protected. This has been particularly problematic in cases involving medical information, such as prescription data.

These cases illustrate the challenge, mentioned in the introduction to our paper, for protecting personal privacy in the context of recent and current de-identification schemes involving PII. So it would seem that much stronger, and more explicit, privacy protections (for data subjects) need to be built into the process of collecting and labeling, as well as mining and analyzing, the kinds of personal data that will eventually be anonymized through the de-identification process.

PART II:
Towards a Comprehensive Ethical Framework to Guide Future Research in Big Data Analytics

We next briefly describe and apply a model advanced by Moor (2005) and Moor and Weckert (2004), which the authors call the "dynamic ethics" approach and we refer to by the acronym *DE*. In proposing this model, Moor (2005, p. 118) argues that we can "improve our ethical approach" to emerging technologies by "recognizing that ethics is an "ongoing and dynamic enterprise," which "continually requires reassessment of the situation."

Moor notes that while there is a temptation to approach ethical aspects of emerging technologies by doing the ethical analysis first, we can foresee only so far into the future. But he also believes that "we need to do more to be more proactive and less reactive in doing ethics" (Moor, p. 118). However, this does not imply that we should adopt what Moor and Weckert describe as an "ethics-first" approach. The authors note that in the past, two kinds of (polar opposite) strategies have typically been used to address ethical issues affecting recent and emerging technologies: *ethics–last* and *ethics–first* frameworks. Historically, the ethics–last approach has been the "standard" model; typically, a new technology was introduced and then we thought about the ethical issues affecting it — so, from an ethical perspective, we had to play "catch up" with the new technology.

The ethics–first approach, on the contrary, is a more recent model; arguably, it was first used in the Human Genome Project (HGP) in the late 1980s and 1990s.The ELSI (**e**thical, legal, and **s**ocial **i**ssues) framework was perhaps the first and arguably best known example of this kind of approach. Many saw the ELSI framework as desirable because it was a "proactive" model.

But Moor and Weckert have argued that the ELSI framework is not ideal for use in *emerging* technologies, since, for example, it could halt future developments in an emerging technology, requiring that we place a moratorium on research in that technology. The authors also believe that having a moratorium would not necessarily result in an improvement regarding the ethical outlook for that technology.

Moor and Weckert also reject an ethics-last approach, which they believe would cause us revert back to "business as usual" and also potentially create a situation in which it is too late to make the necessary ethics adjustments. Instead, the authors suggest that their alternative "dynamic" strategy can be contrasted with a "static" view of ethics. Moor (2005, p. 118) argues that we also need to "establish better collaboration among ethicists, social scientists, social scientists and technologists" (i.e., a multi-disciplinary approach), in addition to developing "more sophisticated ethical analyses."

For example, Moor points out that “ethicists need to be informed about the nature of the technology and to press for an empirical basis for what is and what is not a likely consequence of its development and use.” He also worries that many conventional or traditional ethical theories are in themselves “often simplistic and do not give much guidance to particular situations.”

A suggested above, DE is intended to work for *emerging technologies* (including, for our purposes, big data analytics in general, and the associated problem of re-identification of personal data in particular). So an analysis of ethical problems in this area will need to “be done continually” — i.e., as the field of big data analytics continues to develop, and as its potential social and ethical consequences become better understood. Using DE, we also need to:

A. differentiate between the factual/descriptive and normative components of issues affecting big data;
B. frame and (continually) revise policies, particularly our privacy and consent policies, affecting the collection, use, subsequent use, mining, and analysis of personal data in the context of big data analytics, as necessary — i.e., as:
    a. the factual data changes, or
    b. information about the potential social impacts becomes clearer.

Around the same time that Moor and Weckert were developing DE, Helen Nissenbaum was developing and refining her privacy theory of Contextual Integrity (CI). However, it was not until 2010, with the publication of her book *Privacy in Context* that she developed a “decision heuristic” for privacy which we believe also aligns nicely with DE.

Nissenbaum's CI privacy framework requires that the processes used in gathering and disseminating information (a) are “appropriate to a particular context” and (b) comply with norms that govern the flow of personal information in a given context (2004, p. 137). She refers to these two types of informational norms in the following way: (a) norms of appropriateness, and (b) norms of distribution. Whereas norms of appropriateness determine whether a given type of personal information is either appropriate or inappropriate to divulge within a particular context, norms of distribution restrict or limit the flow of information within and across contexts. When either norm has been “breached,” a violation of privacy occurs; conversely, the contextual integrity of the flow of personal information is maintained when both kinds of norms are “respected.”

Nissenbaum's theory demonstrates why we must always attend to the *context* in which information flows, and not to the nature of the information itself, in determining whether normative protection is needed. Nissenbaum’s framework—or “decision heuristic” –includes a series of specific guidelines that can be applied to her “contextual integrity” model in specific cases. One objective of her heuristic is to help us understand the “source or sources of trouble in new and emerging technologies,” while another is to help us to evaluate the “system or practice in question” (2010, p. 181). Nissenbaum’s decision heuristic includes nine steps:

1. Describe the new practice in terms of information flows.
2. Identify the prevailing context… and identify potential impacts from contexts nested in it…

3. Identify information subjects, senders, recipients.
4. Identify transmission principles
5. Locate applicable entrenched informational norms and identify significant points of departure.
6. Prima facie assessment…A breach of information norms yields a prima facie judgment that contextual integrity has been violated because presumption favors the entrenched practice.
7. Evaluation I: Consider moral and political factors affected by the practice in question…
8. Evaluation II: Ask how the system or practices directly impinge on values, goals and ends of the context...
9. On the basis of these findings, contextual integrity recommends in favor of or against systems or practices under study.... (2010, p. 182)

By comparing DE to this decision heuristic, we believe that we can make a strong case that the two should be used in tandem. It is interesting to note that both are "dynamic" in the sense that they demand the constant reassessment of 'entrenched norms' and 'significant points of departure'. Whereas DE focuses more on the macro-level, Nissenbaum (Steps 1-4) hones in on the specifics of systems under study that could, hypothetically, be those identified by DE (part A). Part B of DE, corresponds to Steps 5-9 of the Nissenbaum heuristic. DE might seem less effective in analyzing ethical aspects of emerging technologies where no clear and explicit policies yet exist. Also, that model does not specifically mention privacy per se (or at least not in a direct sense). But as we will see below, DE also aligns closely with key aspects of Moor's privacy theory which is compatible with CI.

Although both CI and DE use the metaphor of "context," CI does so more explicitly. However, it is important to note that DE is also compatible with, and in some ways expands upon, Moor's privacy theory, which, like Nissenbaum's, is *context based*. For example, Moor focuses on the notion of a "situation" (or zone or context), since an "individual or group" can enjoy privacy only in a "situation"—i.e., in a situation "with regard to others if and only if that individual or group is normatively protected from…information access by others" (Moor 1997). His theory, later described as the RALC (Restricted Access/Limited Control) theory of privacy (Tavani 2007) because it combines aspects of the traditional "control" and "restricted access" theories, also differentiates (a) the *concept* of privacy from both (b) the *justification* of privacy and (c) the *management* of privacy. As such, it has three important components (see Tavani and Moor [2001] for a detailed explanation of how these three component parts work together).

In RALC, *normative privacy* is differentiated from *descriptive* or *natural* privacy, depending on the kind of situation (or context) involved One has normative privacy in a situation where one is protected by explicit norms, policies, or laws that have been established to protect individuals (or groups) in that situation. One can "lose privacy" in a descriptive sense in a naturally privacy situation. But one's privacy is not only lost but also violated in a normatively private situation, by virtue of explicit norms that are breached in that situation or context. Nissenbaum's CI model, on the contrary, focuses exclusively on what would count as normatively private situation/context in RALC.

RALC also includes two important (privacy) *principles*, each of which dovetails nicely with DE: The Publicity Principle and The Adjustments Principle. According to the former principle, Moor (1997) states: “Rules and conditions governing private situations should be clear and known to the persons affected by them.” As Moor notes, our privacy is better protected “if we know where the zones of privacy are and under what conditions and to whom information will be given.” But Moor also notes that sometimes exceptional circumstances can arise and a privacy policy may need to be adjusted. So in his Adjustment Principle, Moor points out: “If special circumstances justify a change in the parameters of a private situation, then alterations should become an explicit and public part of the rules and conditions governing the private situation.” As noted above, we believe that both principles work well with DE and CI, especially in identifying and reinforcing the spirit of transparency needed for adequate privacy policies in particular, and (broader) ethical policies in general.

In the following section, we will show how both DE (in connection with some key aspects of RALC) and CI can be applied to controversies affecting re-identification in the larger context of big data analytics. It seems to us that current practices of re-identification are more of a “breach” of one or more norms, to use Nissenbaum’s metaphor, rather than an instance either of a “norm of appropriateness” or a “norm of distribution” in CI. But, when invoking Moor’s Adjustment Principle, we also believe that view could evolve over time, e.g., depending on some future research practices for big data and the kinds of policies that currently surrounding them.

Applying our Combined DE and CI Model to Classic and Recent Cases of Re-identification

Our comprehensive framework includes the following four steps:

1. Differentiate between the factual/descriptive and normative components of the new or emerging technology under consideration
    a. by identifying contexts, information flows, information subjects, senders and recipients
    b. Locating applicable entrenched informational norms and identifying significant points of departure.
2. Assess if norms of appropriateness and/or distribution have been violated.
3. Revise the policies affecting that technology as necessary, especially as the factual data or components change or as information about the potential social impacts becomes clearer.
4. Continue to evaluate the effectiveness of policies, practices and technical safeguards over time.

We next apply our model to four cases, beginning with the Weld and Netflix cases. Doing so, we believe, suggests the following with regard to the application of our four steps:

In the Weld and Netflix cases, we see that no adequate policies had been in place with regard to re-identifying individuals whose names were anonymized in the respective data sets In the Weld case, many people might have had an expectation that

their PII would not be able to be re-identified in situations involving subsequent uses of their data. But in the Netflix case, those participating in the (Netflix-sponsored) contest had been told that their names would remain anonymous. So, it would seem that people in this case had relied on an explicit agreement with Netflix that their PII would be protected from being re-identified in the future.

But when the new "factual data" revealed what had indeed happened in both cases, it became clear that better, and more, explicit policies were needed to protect individuals. Therefore, according to our model, we can infer:

- Big data analysts who design the code intended to protect personal privacy with respect to PII would need to "reassess" the de-identification schemes that had been used in those kinds of data sets. (Steps 1 and 2)
- It would also require government/health agencies/ organizations and corporations to re-assess their policies—framing new ones where needed, or revising existing policies where appropriate—in ways that can better protect an individual's PII from being so easily identified. (Step 3)
- The entrenched norms are evident in the contract with Netflix, but a bit murkier in the Weld case where there were no adequate policies, and only some possible expectations on the part of the participants. (Step 1)
- Nevertheless, the ability to re-identify individuals pointed out the potential to violate privacy in this context.(Step 2)
- Through the assessment undertaken in Step 2, we see a clear indication that goals of the context to protect privacy (step 3) were violated and, therefore, there would be a need to reconsider the policies of this system (step 3) rather than accepting it as is.

We next apply our model to the Australian Health Care (AHS) case, which involved the re-identification of individuals in the Australian government's open health data set. This case is perhaps even more controversial than the Weld and Netflix cases because the data set used in AHS also potentially reveals the PHI (Public Health Information) of individuals whose names have been re-identified. In fact, it seems that the only explicit policy that Australian government had implemented was one of "transparency" in the dissemination of the kind of health-related information involved —i.e., the (de-identified) longitudinal medical billing records— to be made publicly available (and the additional requirement that they be published online) so that all Australians could have easy access to that information. But in the wake of what was revealed in the study by Culnane et al., it was clear to the Australian government that it had to re-think its policy regarding transparency of medical billing records, until the problems affecting re-identification of individuals were resolved. The new factual data suggested that the government's "open data" policy was inadequate and needed to be "reassessed in light of the ethical implications involving the privacy problems resulting from the original policy. And this incident also suggests that any new or revised policies would also require continual reassessment, given the potential new ways in which people whose PII and PHI that resided in that kind of dataset might also be so easily be re-identified in the future.

The “new factual data” resulting in the Australian case also showed that PII, and possibly PHI as well, could be easily re-identified, regardless of whether that anonymized data had been encrypted. So, merely using strong encryption techniques in health data sets would not in itself be sufficient for revising or framing an adequate policy. Consider that Culnane et al. had also pointed out that they were easily able to link unencrypted parts of the data in that set in successfully re-identifying individuals.

Another worry involving the Australian case is that public health researchers depend on people volunteering to be research subjects in future health-related studies. But if potential research subjects cannot be better assured that their PHI can be protected, they may be less inclined to participate in the relevant research studies. This, in turn, could have negative consequences for advancements in health and medicine. So it is important to frame adequate policies.

We infer the following with regard to our four steps:

- No clear and explicit policy regarding de-identification process for anonymizing PII and PHI seems to have been in place.(Steps 1 and 2)
- After a study, it was clear that the government’s “open data” policy was inadequate and needed to be “reassessed in light of the ethical implications involving the privacy problems resulting from the original policy. (Step 3)
- If potential research subjects cannot be better assured that their PHI can be protected, they may be less inclined to participate in the relevant research studies.(Step 4)

We believe that the case involving session replay scripts is also interesting, because the departure from what could be presumed “informational norms” was most likely not anticipated by either the developers or the users of these scripts. A study by Englehardt (2017) demonstrated the need to revisit the privacy policies, especially because of the complexity of this emerging technology. In his study, Englehardt “… installed replay scripts from six of the most widely used services and found they all exposed visitors’ private moments to varying degrees.” (Englehardt 2017) “Scripts from FullStory, Hotjar, Yandex, and Smartlook were the most intrusive because, by default, they recorded all input typed into fields for names, e-mail addresses, phone numbers, addresses, Social Security numbers, and dates of birth” (Goodin, 2017). Rather than the information flowing only from a user to the application, it is also captured by third-party script vendors. (See Grodzinsky et al. for a detailed CI-related analysis of this issue). Although third- party vendors like FullStory (2016) had its own privacy policy, it placed the burden of protecting private information on the application developers in Full Story’s Acceptable Use policy. Thus, the burden to contain sensitive information was shifted to the website developer, who may or may not have been aware of this obligation and may or may not have been capable of implementing a technically-complicated, yet required redaction.

As new factual information about the social impacts of these scripts became clearer, companies began dropping Full Story and other third party vendors. Pharmacies, in particular were worried about possible HIPPA violations. In the case of Walgreens, sensitive information about medical conditions and prescriptions along with user names were sent to FullStory via Walgreens’ website, because third party analytics could have access to earlier verification questions and mouse tracking. Walgreens’ privacy policy had

assured customers that “Walgreen’s does not retain this data and cannot access or view your answers” (Englehardt 2017). Customers did not opt in to have their information leaked to Full Story, and they trusted the pharmacy to protect their privacy.

Using our framework to evaluate session replay scripts prior to their implementation might have identified these privacy problems at the onset. Going forward, because this technology is here to stay, DE would be helpful to assess shared responsibility for private data between application developers and third party vendors, and user opt-in policies. It could also be used to

Finally, if we apply our model to the Session Replay Scripts, we infer that Steps 1-3 are relevant:

- Rather than the information flowing only from a user to the application, it is also captured by third-party script vendors without user consent. (Steps 1 and 2)
- The departure from what could be presumed “informational norms” was most likely not anticipated by either the developers or the users of these scripts. (Step 3)
- Review and possibly revise existing policies as new implementations are implemented. (Steps 3 and 4)
- Reassess privacy policies in systems using session replay scripts (See Grodzinsky et al 2018 for a detailed analysis). (Step 4)

Conclusion

As we have attempted to demonstrate in our analysis of the above cases, combining/integrating key aspects of the DE and CI frameworks to create a more comprehensive one and applying it to the problematic and often unanticipated problem of privacy in the re-identification of anonymized personal data can lead to a more comprehensive assessment. Going forward, we also believe that applying these frameworks in tandem with system development might alleviate the necessity to continuously have to ‘put out fires’ in the future, on both a case-by-case and an ad-hoc basis. Our workable, comprehensive framework moves us in the direction of sounder policy development and the protection of PII and PHI. In doing research for the session replay case, we were pleased to note that the Future of Privacy Forum (Gray, 2018), has developed a checklist for “best practices” of privacy professionals around this emerging technology. In evaluating third-party vendors, it suggests:

1. Evaluating and inquiring into script developers’ terms and privacy policies.
2. Carefully selecting which pages on a site are appropriate for session replay scripts.
3. Placing technical safeguards on the side of collection (client side) rather than use side (server).
4. Continuing to evaluate the effectiveness of policies, practices and technical safeguards over time, by reviewing site implementation and report analytics.

Although the Forum focuses on the specific case of session replay scripts, we can see how it also reinforces the position we have taken in this paper in combining DE and CI.

We believe that using our framework brings us to a very similar conclusion with respect to privacy protection in the context of emerging technologies.

Acknowledgments

An earlier version of this paper, titled “Ethical Aspects of Big Data (Analytics),” was delivered as a talk by Herman Tavani at the Thought Leadership Network (TLN) Seminar on Research with Big Data, Bentley University, April 25, 2017. Tavani is grateful for the comments and suggestions received from those attending that seminar, and some of those suggestions have been incorporated into our paper. We also draw from some material in a previously presented talk, titled “Looking for the Full Story: Ethical issues associated with session replay scripts,” by Frances Grodzinsky, Keith Miller, and Marty J. Wolf, delivered at the Ethicomp 2018 Conference. In developing the present paper, the authors have also drawn from some concepts and distinctions introduced in a few of their previously published works, including Grodzinsky and Tavani (2010, 2011) and Tavani (2016). Finally, we wish to acknowledge our gratitude for comments received from attendees at a paper session at the CEPE 2019 Conference (Old Dominion University, Norfolk, VA), where an earlier version of this work was presented on May 29, 2019. We have also incorporated some of those comments into the published version of this paper.

References

CIPP (Certified Information Privacy Professional Guide). “De-Identification and Re-Identification.” Available at https://www.cippguide.org/2010/09/21/de-identification-re-identification/. Accessed 11/5/2018.

Culnane, C., Rubenstein, B., and Teague, V. (2017). “Health Data in an Open world.” Report: University of Melbourne. Available at: https://about.unimelb.edu.au/newsroom/news/2017/december/research-reveals-de-identified-patient-data-can-be-re-identified. Accessed 4/24/19.

Devlon B. (2018). “The Anonymization Myth.” *TWDI*. Available at https://tdwi.org/articles/2018/09/04/dwt-all-anonymization-myth.aspx. Accessed 11/5/2108.

Englehardt, S. (2017). No boundaries: Exfiltration of Personal Data by Session-Replay Scripts. https://freedom-to-tinker.com/2017/11/15/no-boundaries-exfiltration-of-personal-data-by-session-replay-scripts/. Accessed 9 December 2017.

EPIC (Electronic Privacy Information Center). “Re-Identification.” Available at (https://epic.org/privacy/reidentification/. Accessed 11/5/2108.

Goodin, D. (2017). No, you're not being paranoid. Sites really are watching your every move, https://arstechnica.com/tech-policy/2017/11/an-alarming-number-of-sites-employ-privacy-invading-session-replay-scripts/. Accessed 1 December 2017.

Gray, S. (2018). Understanding Session Replay Scripts—a Guide for Privacy Professionals. Future of Privacy Forum (March 5, 2018), https://fpf.org/2018/03/05/understanding-session-replay-scripts-a-guide-for-privacy-professionals/. Accessed 12 May 2018.

Grodzinsky, F. S. (2017). "Why Big Data Needs the Virtues." In *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic and Ethics*. T.M. Powers, Ed. Springer, Philosophical Studies Series, Vol .128. pp. 221-234.

Grodzinsky, F. S., Miller, K. W. and Wolf, M. J. (Forthcoming). Looking for the Full Story: Ethical issues associated with session replay scripts. ETHICOMP 2018 (September, 2018).

Grodzinsky, F.S., and Tavani, H. T. (2010). "Applying the 'Contextual Integrity' Model of Privacy to Personal Blogs in the Blogosphere" (with Frances Grodzinsky). *International Journal of Internet Research Ethics*, Vol. 3, No. 1, 2010, pp. 38-47.

Grodzinsky, F. S., and Tavani, F.S. (2011). "Privacy in the Cloud: Applying Nissenbaum's Theory of Contextual Integrity" (with Frances Grodzinsky). *Computers and Society*, Vol. 42, No. 1, 2011, pp. 38-47.

Lubarsky, B. (2017). "Re-identification of 'Anonymized Data.'" *Georgetown Technology Law Review*, 202. Available at https://perma.cc/86RR-JUFT. Accessed 11/5/2108.

Moor, J. H. (1997). "Towards a Theory of Privacy for the Information Age." *Computers and Society*, Vol. 27, no. 3, pp. 27-32.

Moor, J. H. (2005). "Why We Need Better Ethics for Emerging Technologies." *Ethics and Information Technology*, Vol. 7, No. 3, pp. 111-119.

Moor, J. H., and Weckert, J. (2004). "Nanoethics: Assessing the Nanoscale from an Ethical Point of View." In D. Baird, A. Nordmann, and J. Schummer, eds. *Discovering the Nanoscale*. Amsterdam, The Netherlands: IOS Press, pp. 301–10.

Nissenbaum, H. (2004). "Privacy as Contextual Integrity." *Washington Law Review*, Vol. 79, No. 1, pp. 119–57.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Palo Alto, CA: Stanford University Press.

Perez, B., Musolesi, M., and Gianluca, S. (2018). “You are your Metadata: Identification and Obfuscation of Social Media Users using Metadata Information.” *Association for the Advancement of Artificial Intelligence*. Available at: https://www.ucl.ac.uk/~ucfamus/papers/icwsm18.pdf. Accessed 11/5/2108.

Rijmenam, M. van. (2016). “The Re-Identification of Anonymous People with Big Data.” Available at: https://datafloq.com/read/re-identifying-anonymous-people-with-big-data/228. Feb. 10. Accessed 11/5/2018.

Solon, O. (2018). “Data is a fingerprint: why you aren't as anonymous as you think online.” *The Guardian*. Available at: https://www.theguardian.com/world/2018/jul/13/anonymous-browsing-data-medical-records-identity-privacy. Accessed 4/24/2019.

Sweeney, L., and Malin, B. (2006). Trail re-identification and unlinkability in distributed databases. Doctoral dissertation. https://dl.acm.org/citation.cfm?id=1168443. Accessed 4/15/2019.

Tavani, H. T. (2007). “Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy.” *Metaphilosophy*, Vol. 38, No. 1, pp. 1-22.

Tavani, H. T. (2016). *Ethics and Technology: Controversies, Questions, and Strategies in Ethical Computing*. 5th ed. Hoboken, NJ: John Wiley and Sons.

Tavani, H. T., and Moor, J. H. (2001). “Privacy Protection, Control of Information, and Privacy-Enhancing Technologies.” *Computers and Society*, Vol. 31, No. 1, pp. 6-11.