January 17, 2023

By Electronic Mail

Ram D. Sriram, Director
National Institute of Standards and Technology
Software and Systems Division,
Information Technology Laboratory
100 Bureau Drive
Gaithersburg, MD 20899-8970

Re: Comments of ACM US Technology Policy Committee
De-Identifying Government Data Sets Report (NIST SP 800-188 3pd)

Dear Director Sriram:

ACM, the Association for Computing Machinery, is the world's largest and longest established association of computing professionals, representing approximately 50,000 individuals in the United States and more than 100,000 worldwide. ACM is a non-profit, non-lobbying and non-political organization whose U.S. Technology Policy Committee ("USTPC") is charged with providing policy and law makers throughout government with timely, substantive and apolitical input on computing technology, and the legal and social issues to which it gives rise.

Consistent with that charge, USTPC is pleased to submit the attached comments on the third draft of the Division's report on *De-Identifying Government Data Sets* ("Report") released in November of 2022 (NIST SP 800-188 3pd). We hope that our input is useful and, as requested in the Report, have compiled it utilizing the suggested spreadsheet template. USTPC wishes, however, to underscore two broader key points included in that document:

- The Report is an excellent synthesis of a huge amount of knowledge. USTPC recommends, however, that it be supplemented to explicitly provide greater clarity with respect to how the document might best be used by its multiple likely audiences. While the document provides a universally useful overview, we are concerned that absent such guidance two unintended problems could arise. Less technical readers could be intimidated by its detail, while more sophisticated readers might mistake it for comprehensive step-by-step guidance. Neither effect, of course, would be desirable and both may be avoided with further discussion in the Report; and

ACM U.S. Technology Policy Committee
1701 Pennsylvania Ave NW, Suite 200
Washington, DC 20006
+1 202.580.6555
acmpo@acm.org
www.acm.org/public-policy/ustpc

- The text of the Report would benefit from the precise and consistent use of risk-related terminology. Key distinctions in some cases are not drawn or maintained in the text, for example, between the terms “re-identification risk,” “privacy risk,” and “privacy loss.”

USTPC commends NIST for its excellent work in this sphere and stands ready to assist it in the future. Thank you for the opportunity to provide the attached recommendations. *

Respectfully submitted,

Alec Yasinsac
USTPC Vice Chair

Attached: Detailed Comments Spreadsheet

** USTPC’s recommendations were drafted for the Committee’s approval principally by members Harish Arunachalam, Arnon Rosenthal, and Stuart Shapiro. Simson Garfinkel, a lead author of the Report and USTPC Subcommittee Chair, was not engaged in that process.*

ACM U.S. Technology Policy Committee
1701 Pennsylvania Ave NW, Suite 200
Washington, DC 20006

+1 202.580.6555
acmpo@acm.org
www.acm.org/public-policy/ustpc

Comment #		Type (General / Editorial / Technical)	Starting Page # *	Starting Line #	Comment (include rationale)*	Suggested Change*
1	US Technology Policy Committee of the Association for Computing Machinery	Editorial	1	306	Typo	Replace “shoudl” with “should”
2	same	Technical	1	305	As the body of the document indicates, formal methods come with pros and cons; this statement thus is inaccurate.	Edit line to begin: "When they are available and all else being equal, formal..."
4	same	Technical	9	578	The references to privacy risk beg the question of what constitutes such risk, a concept that also has evolved over time.	Change text to read: "the amount of specifically-defined privacy risk that results..."
5	same	Technical	11	636	This statement is inconsistent with the adjacent statements as it appears to be referring to the broader privacy risk, i.e., downstream adverse impacts, rather than the re-identification risk per se.	Delete sentence, “Re-identifcation risk is typically a function of the adverse impacts that would arise if the re-identifcation were to occur and the likelihood of occurrence.”
6	same	Technical	14	737	Privacy loss and privacy risk are two distinct concepts conflated here. Depending on the context, low privacy loss might be enough to create the potential for significant adverse privacy consequences.	Edit the sentence to read: "releasing the data it produces will probably, but not necessarily, result in little..."
7	same	Technical	14	744	This equates privacy loss with privacy risk.	Replace "risk" with "loss."
8	same	Technical	14	745	This equates privacy loss with privacy risk.	Update the sentence to read: "considers the privacy loss of an individual from..."
9	same	Technical	15	772	This equates privacy loss with privacy risk.	Revise the sentence to read: "amount of privacy loss introduced..."
10	same	Technical	18	897	A risk assessment only evaluates risk. What is described here may more accurately be thought of as a risk-benefit analysis.	Change "risk assessments" to "risk-benefit assessments"
11	same	Technical	20	Note 12	This description of journalist risk here is incorrect in that the journalist is, in fact, trying to identify a specific person. A corrected footnote may be more applicable to RMP than UIRP.	Delete footnote text after the comma, “Some texts refer to UIRP as “journalist risk.” The scenario is that a journalist has obtained a de-identified file and is trying to identify one of the data subjects, but the journalist fundamentally does not care who is identified.”
12	same	Technical	21	978	The text here does not use standard risk modeling terms, including threats and vulnerabilities, although these terms are employed later in the document.	Employ relevant risk modeling and assessment terms consistently throughout the document.
13	same	Technical	32	1284	The charter also should frame DRB responsibilities in terms of how they relate to those of other relevant organizational components, especially those responsible for privacy, civil liberties, and enterprise risk.	Revise the text to read: "laws, as well as the responsibilities of other relevant organizational components, especially those with cognizance over privacy, civil liberties, and enterprise risk."
14	same	Technical	51	1936	The distinction between structured and unstructured data is more fundamental than is reflected here. De-identification of text narratives is almost a separate discipline.	Add text noting that de-identifying text narratives requires specialized expertise even within the broader domain of de-identification.
15	same	General			Risk-related terminology is used inconsistently (see comment # 12), and, in some cases, important distinctions are not drawn or maintained (e.g., between re-identification risk, privacy risk, and privacy loss).	Employ appropriate risk terminology consistently throughout the document while making relevant distinctions.
16	same	Technical	11	643/644	The ideas of redaction's contribution to dataset accuracy loss and non-ignorable bias are introduced abruptly here without contextualization. This counterproductively shifts the discussion of risk evaluation from a dataset standpoint to a specific activity-enabled risk (such as analytics/ machine learning activities).	Consider limiting the main text to discussion of general risk, moving discussion of specific activity-enabled risks to a citation/reference regarding, for example, 'non-ignorable biases’.
17	same	Editorial	15	802	Typo	Replace "Ulimited" with "unlimited"
18	same	Editorial	17	848	It is unclear whether "absent the data release" means "in the absence of data release"	Clarify

Comment #		Type (General / Editorial / Technical)	Starting Page # *	Starting Line #	Comment (include rationale)*	Suggested Change*
19	same	Editorial	37	1443	The paper rightly asks for DRB powers to be explicitly defined, e.g., whether its instructions are mandatory or advisory.	Separately specify the power for positive vs. negative decisions by DRB. E.g., in some organizations, DRB might be allowed to say “We agree that this is low risk – go ahead”, but not to veto. In others, vice versa.
20	same	Technical	34	1360	The treatments of “Prescriptive” and “Performanc-Based” should be more specific.	· Require a high level of specificity in prescriptive instructions, e.g., “if you do this, do it in this form”. · Require that instructions be identified as mandatory or optional. State whether each condition is meant to be necessary or sufficient, e.g., “HIPAA specifies two conditions, either of which is sufficient.”
21	same	Technical	47	1803	Another option of including "secure computation": These are techniques for securely implementing the abstract computation the user wants, i.e., for carrying out a computation without leaking info. But the result may still be sensitive. One still needs to approve releasing the result of the computa-tion. The consumer thus still needs to be aware of release policies.	Include “Secure Computation” as one of the specified items.
22	same	General		1320	Much of the information required for DRB also is needed for routine operations, e.g., to describe what is collected.	Include a recommendation that DRBs be sufficiently flexible in format to permit them to accept information to be delivered within a data engineering tool.
23	same	General	46	1759	Two key points are not addressed in this paragraph: 1) encryption is not recommended because it's reversible; 2). No justification for not using Hashing is provided.	Add: Although information may be protected by encryption, an attacker who obtains the encryption key can access the information. Accordingly, hashing – which is not reversible – should be used to the maximum extent feasible.
24	same	Editorial	47	1797	"Attenuation bias" is not defined. Rather than introduce a new designated term for a single use, consider simply expounding in text.	Define this term and provide an example of its use in a sentence. Alternatively, revise the text to read: "leads to especially severe errors on differences, correlations, and regressions".
25	same	Editorial	7	487-498	The document is an excellent synthesis of a huge amount of knowledge. But will it be usable by the intended audience? The Intended Audience and Purpose sections suggest that agency staff are the primary audience. However, the sheer volume of material and its technical sophistication may make its use challenging for them and, in fact, more suited to researchers and technologists creating de-identification tools. At the same time, the document does not constitute step-by-step guidance; this should be explicitly noted.	Rewrite the Purpose and Intended Audience sections to provide greater clarity on how the document would best be used by its different audiences. Be clear that the document provides an overview useful to all, but less technical readers may wish to skip some sections. Conversely, also note that more sophisticated readers should not consider it to be comprehensive sufficient step-by-step guidance.
26	same	General	14	748	The word "Mechanisms" is used in different contexts throughout the document with different meanings.	Use a word or phrase term that denotes mathematical algorithms.
27	same	Editorial	45	1715	Sentence does not scan	Revise
28	same	Editorial	50	1901	Determination of DNA sequence/whole genome construction is possible within a certain accepted error threshold.	Rephrase the statement to include threshold information.
29	same	Technical	50	1934	It is also possible to identify certain diseases and whether an individual is at risk for certain hereditary conditions. For example, presence of the HER2 gene genetic information is a marker for breast cancer risk	Rephrase the statement to convey that genetic information can identify numerous factors about an individual ranging from their ancestry to medical conditions.
30	same	Technical	65		There are many approaches captured in the document that are published as open-source packages, repositories, and tools (e.g., FHE). Use of those tools and techniques comes with its own risks and challenges. It would be good to briefly talk about their use, especially the importance of using only those code repositories that are actively maintained.	Consider including a section on using open-source software and its attendant risks.