Reply to the comments on manuscript "*A Systematic Review on Privacy-Preserving Distributed Data Mining*"

Submission tracking ID: 688-1668
Authors: Chang Sun, Lianne Ippel, Andre Dekker, Michel Dumontier, Johan van Soest

Dear reviewers and editors:

We thank the two reviewers for your generous and helpful comments on the manuscript and have revised the manuscript to address your concerns. The major revisions include adding more explanations about our review methodologies in Section 3.1 - Eligibility Criteria, adding figures to describe the review results in Section 4 - Results, and discussing the potential limitations of this review in Section 5.6. We updated the Acknowledgement section in the manuscript to express our gratitude to the reviewers' for their time and efforts to help us improve the manuscript.
Please find our response to the comments point by point in this letter. We believe that the manuscript is now suitable for publication in the Data Science Journal.

Kindest regards,
On behalf of all authors,
Chang Sun

Reply to the comments from reviewer 1:

Comment 1.1: The significance of the research is not clearly addressed and is not justified well other than data privacy issue in the paper.

- ➔ ***Reply 1.1****: Thank the reviewer for this comment. We addressed this comment by revising the last three paragraphs of the Introduction section [Page 2-3] to highlight the significance and contribution of our research. Briefly, from studying existing surveys, we found the majority of the published surveys have typically treated PPDDM as a specialised subtopic of either distributed data mining or privacy-preserving data mining. PPDDM methods which focus on both privacy issues and distributed data situations were under-reported and need a complete summary for the field. Therefore, our research is significant in that we conduct a comprehensive and complete analysis of existing PPDDM methods to present an overview of the field, identify major challenges, and provide some guidelines and recommendations for future research in this field. Our manuscript also encourages researchers to propose more innovative methods and apply their methods to practical/real-life use cases. In the revised manuscript, we highlighted our significance in the Introduction Section by describing the key contributions including proposing new 10-factor metrics for evaluating PPDDM methods, presenting the ambiguity of privacy and security, recognizing the reasons and the need to apply theoretical PPDDM methods to real-life use cases, providing a list of recommendations for future PPDDM studies.*

Comment 1.2: The discussion of 'global data miner' or meta-learning is not convincing enough. That plays an important role in distributed data mining. There are other meta-learning approaches other than SVM, DT, NN such as stacking, voting,

RandomCommitte. I would love to see some more discussions or an evaluation metric proposed for meta-learning.

- ➔ ***Reply 1.2:*** *We thank the reviewer for highlighting the importance of meta learning in the context of distributed data mining. Our overall objective is to provide researchers with a broad perspective across a variety of approaches in the field. Correspondingly, in section 2.3, we referred to two comprehensive survey papers on meta learning for interested readers to explore and obtain a more comprehensive view of those techniques. These survey papers give detailed explanations and development of stacking, voting, boosting and random committee. However, we very much appreciate the reviewer for this comment and agree that this category of PPDDM methods might be interesting to analyze and report for the researchers in this field. Therefore, we added this possibility for future research as a point in the Discussion Section. [section 5.6].*

Comment 1.3: The presentation of results via large, obscure tables sometimes makes it difficult to interpret. The use of a bar chart or pie chart can be useful here to interpret the comparisons.

- ➔ ***Reply 1.3:*** *We have added new figures (bars charts) to summarize the evaluation metrics. It's Figure 5 in the revised version.*

Comment 1.4: The case studies, although could be potentially the most interesting part of the paper, is barely covered (only a few examples such as authors explained data partitioning in figure 2)

- ➔ ***Reply 1.4:*** *If we understand the "case studies" correctly in the reviewer's comment, we expanded the section of secure multiparty computation protocols and added case studies based on this comment. Please see figure 1 in the revised paper. If the "case studies" the reviewer meant refers to the real-life/practical use cases, we discovered only 5% of reviewed papers involving real-life/practical uses cases in their study. We have discussed some of them in the result and discussion section in the review.*

Further comments:

Comment 1.5: Deep learning (e.g. CNN, LSTM) is now heavily used in distributed data ming area. The authors did pick some paper on neural network, however, completely ignored the discussion around deep learning and identify key metric for that. They used different performance measures other than what is described in the "Accuracy performance factor" such as mean average precision.

- ➔ ***Reply 1.5:*** *Thanks for the reviewer's valuable comment. As with other machine learning methods, evaluation metrics for deep learning were included in the performance measure factor. But we did not elaborate on the details of each metric. Owing to the high degree of heterogeneity in the reporting of performance measures across the surveyed papers, our goal was to determine whether any performance measure was reported rather than a survey of different performance measures. This is important given how different experimental conditions could heavily influence these measures and the interpretation of specific results. Correspondingly, we added more explanations in the performance measure factor paragraph [in section 3.3].*

Comment 1.6: I disagree with excluding topics around cloud computing, grid computing, edge computing, etc. They should be in the inclusion criteria as the platform used for training

and classification plays a big role in distributed data mining applications. I recommend including another evaluation metric for the platform.

- ➔ ***Reply 1.6:*** *We agree with the reviewer that cloud/grid/edge computing techniques have been used in many distributed machine learning/data mining tasks. A key metric in these techniques lies around computational performance rather than the analysis complexity of the problem (mining over heterogeneous and independent data sources that wish to collaborate on joint data analysis but are restricted by data privacy/data protection issues). For example, training a simple model to do an association study based on 1000 patients data collected by two hospitals would not necessarily benefit from HPC or cloud/grid/edge computing techniques. Therefore, we did not include cloud/grid/edge computing in our review. We highly appreciate the reviewer's comment and added the explanation in the paper to explain why we exclude topics in cloud/grid/edge computing techniques (in section 3.1 Eligibility Criteria).*

Comment 1.7: Important referencing is missed in many places. Authors need to justify some statements with important references. For example,on page 7, "There are plenty of algorithms across the data mining and statistics domain [ref of a survey paper?]" on page 8, "The accuracy performance includes accuracy, precision, recall, F1 score" .. please give examples with references.

- ➔ ***Reply 1.7:*** *We thank the reviewer for this comment. We have screened the full paper again and added new references to support the statements in the paper, including the two indicated cases.*

Reply to the comments from reviewer 1:

Comment 2.1: Despite presenting a systematic literature review, the paper is rather vague on some steps of the inclusion/exclusion of works. In particular, in the exclusion phase the authors refer to the exclusion criteria (Figure 2, steps 2-5) that was only implicitly presented in the text (i.e., paper covering topics on X, Y, Z), and has no explanation to how it was systematically conducted. For instance, were the topics searched for in the abstract/whole body? How were the topics searched for? By key-words, by reader's interpretation? How many people did the screening for these topics?

➔ ***Reply 2.1:*** *Thank the reviewer for this comment. We have rewritten the second paragraph in section 3.1 Eligibility Criteria to explain how we include and exclude the papers. Briefly, we searched topics in the title, keywords, and abstract to find relevant papers. The search strategy is described in 3.2. After getting all the papers from searching, reviewers first read the titles, keywords, and abstracts to exclude the papers that focus on the topics listed in exclusion criteria (section 3.2). This was conducted manually (based on the reviewer's understanding of the papers). If the titles, keywords, and abstracts are not sufficient to identify the paper, the reviewers read the full text of the paper to make a decision.*

Comment 2.2: Another missing step of the review is the backwards/forwards reference search, which is a well accepted technique for finding valid additional literature.

➔ ***Reply 2.2:*** *Thank the reviewer for this valuable comment about the backwards/forwards reference search method. While we considered a one-iteration forward reference search (review the papers from the references of the selected papers), we opted against this strategy as i) it introduces a bias in favour of what authors think is relevant to their narrative [Egger, Matthias. "Problems and limitations in conducting systematic reviews." Systematic reviews in health care meta-analysis in context (2001): 43-68.] and ii) We already had a considerable number of selected papers (231 papers) in which a backwards/forwards reference search would greatly increase the number of papers to be reviewed that may not be necessary, and iii) we mapped the referring/citing a network of reviewed papers in a graph (Fig. 6.) It shows three papers are frequently cited by other studies and have a major impact on the field. If we add more papers from the backwards/forwards reference search, our analysis of the referring/citing network would be more biased. However, we highly appreciate your comment and added this in the paper as one limitation of our review.*

Comment 2.3: And finally, I also mis details about how the 10 factors were generated. It looks like they emerged from the 231 papers, but they were also used to categorise the same 231 papers. I imagine this means several iterations to these papers were conducted in order to finish the review (and perhaps by multiple researchers), but that is not explained in the paper.

➔ ***Reply 2.3:*** *We added more explanations in section 3.3 - Metrics for Reviewing Papers "The authors initially generated and modified these evaluation metrics by reviewing 10% of the included articles. Then, the metrics have been discussed by the co-authors in several iterations of reviewing until an agreement has been made on these 10-factor evaluation metrics. Afterwards, all selected papers have been reviewed and assessed again using the metrics."*

Further comments:

Comment 2.4: Citations are glued to the text ("… efficiency[1,2]"), there should be a space between the text and the brackets.

➔ ***Reply 2.4:*** *Thank you for pointing out this problem. We have added space (used ~/cite{}) between text and reference brackets in the whole paper.*

Comment 2.5: In page 4, lines 29-32, 'Secure set union' has an unclear explanation.

➔ ***Reply 2.5:*** *We revised the Secure set union part with a more detailed explanation and a figure to describe how the protocol works by demonstrating an example.*

"Secure set union has been applied to the case where data parties want to jointly create unions of sets from rules and item sets shared by multiple parties but not leaking the owner of each set. To guarantee a secure computation, one approach is to apply a commutative encryption system in computing the set union [26,37]. A commutative encryption system can encrypt original data multiple times using different users' public keys. The final encrypted data can be decrypted without considering the order of the public keys in the encryption process [38]. In the secure set union protocol, one data party encrypts its own item sets using commutative encryption and transfers them to other parties. The receiver party encrypts both its own sets and the received encrypted sets and passes them to the next party. Once the data is encrypted by all parties, decryption can start at each party in any different order from the encryption order to preserve the ownership of itemsets. However, if one itemset is present at both data parties, then the number of this itemset will be exposed because of duplication."

Comment 2.6: In page 6, lines 9-10, links to databases are not relevant bibliography to the work, and should be presented as footnotes.

➔ ***Reply 2.6:*** *We agree with the review and we changed it to the footnotes.*

Comment 2.7: In page 6, line 34-35, why not fully-honest as well? Did that not happen in any paper? I would be curious to know that, as it seems this would be a limitation of the paper.

➔ ***Reply 2.7:*** *Thank the reviewer for this comment. The fully honest party is regarded as a fully-trusted and incorruptible party who will follow all the protocols and not be curious about any input/output from others. If data parties are trusted by each other, then it is not necessary to apply privacy-preserving methods to protect sensitive and private data. However, in the review, we did consider the third party to be fully honest, semi-honest, and malicious [in the paragraph of Adversarial behaviour of data parties in Section 3.3].*

Comment 2.8: In page 6, lines 40-41, If I had to guess, I would say 'untrusted', 'non-trusting', 'non-collaborative' refer to malicious behaviour. Does it mean you preferred not to classify them, or that they do not present a formal adversarial model?

➔ ***Reply 2.8:*** *Untrusted, non-trusting, non-collaborative can be either semi-honest or malicious because of the lack of further definition of these terms. We decided to not classify them into any category because 1) some studies did not provide privacy or security proof, we could not guess which adversarial behaviour they meant by (for example) "untrusted", 2) some papers did not study or mention "adversarial behaviour" or "corruption strategy" which might mean they did not the differences between these behaviours. For example, some papers used the terms such as 'untrusted', 'non-trusting', 'non-collaborative' only once without any explanation. It's not clear for us what adversarial behaviour they have considered in their method. Correspondingly, we added a reason stating why we did not classify them into any category - "Papers that use ambiguous expressions such as 'untrusted' or 'non-trusting' or*

'non-collaborative' are not classified into any category, because they did not clearly indicate the adversarial property of data parties, nor did they provide any privacy or security proof of their methods."

Comment 2.9: In page 9, Figure 2, steps are all numbered the same, and it is unclear why the excluding part begins in 2.

➔ ***Reply 2.9:*** *Thank you for your comments on the criteria and the workflow. We have made changes in the workflow based on your comments and described more details in the answer of your other comments related to the inclusion/exclusion criteria [in Reply 2.1, 2.3, 2.10].*

Comment 2.10: In page 10, lines 6-9, here is some excluding criteria, but how was that systematically executed? Does this refer to step 2? How was step 2 executed: manually, automatically searching those key words? In the body/abstract/keywords of the paper?

➔ ***Reply 2.10:*** *Yes, it refers to step 2. We have revised some sentences in this section. For step 2, we have answered it in your previous comments. We did it manually by reviewing first the titles, keywords, and abstracts of the papers. If necessary, we also read the full text of the paper to exclude the irrelevant papers. The reason to do it manually because 1) searching keywords such as "cloud computing" or "grid computing" to exclude some papers that are actually relevant to our review, for example, the paper just discusses/mentioning the impact of cloud computing on PPDDM. 2) we also exclude papers that focus on privacy-preserving data publishing/data collection/data management, these studies do not have specific keywords we can use to exclude the papers unless we read the content.*

Comment 2.11: In page 10, Figure 3's caption, unclear what "relations" mean here. Only clear when you read the text below.

➔ ***Reply 2.11:*** *Thank you for pointing this out. We revised the caption of the figure to "Papers are presented as nodes and clustered by the search terms. The number of papers in each cluster is labelled in the figure. The edges show which search terms were used to find the papers. For example, the 23 nodes in the purple cluster were found from using search terms PPDDM, PPHDM, and PPVDM."*

Comment 2.12: In page 10, lines 42-44, this does not look like a conclusion I can take from the graph. Were the 10 metrics used to generate the graph somehow? This comment looks misplaced.

➔ ***Reply 2.12:*** *The 10 metrics were not used to generate the graph. The conclusion "a large number of papers (71 papers from PPDDM, 22 papers from PPDML) did not indicate what data partitioning problems their method can solve in their titles, abstracts, and keywords." was made because the 71 and 22 papers from PPDDM and PPDML do not have keywords "horizontal" or "vertical" in their titles, keywords, and abstracts. Therefore, readers will miss the papers by searching "privacy-preserving" and "horizontally/vertically partitioned data". Readers have to read the whole paper to know which data partitioning problem the study is solving.*

Comment 2.13: In page 10, lines 10-11, does that mean they assume the worse?

➔ ***Reply 2.13:*** *Do you mean page 11 line 10-11? (line 10-11 on page 10 is a figure), "However, it is worth noting that more than 30% of selected papers did not state a clear assumption that which adversarial behaviour their approach can deal with." If we understand correctly, "they assume the worse" means "they assume the malicious adversarial"? From the paper we*

reviewed, if they did not explicitly indicate the adversarial behaviour in the paper, then 1) they did not provide or discuss privacy proof, 2) it is not clear what possible attacks or adversarial behaviour will happen at the data parties and can be prevented by the proposed method. As reviewers/readers, we cannot make an assumption about this based on limited evidence from the paper.

Comment 2.14: In page 11, footnote, link does not work.

➔ ***Reply 2.14:*** *The data repository was not public when we submitted the manuscript. Now it is publicly available and the link is on the footnote of Page 12 in the revised paper. And the link to the data repository is working now at: https://figshare.com/s/cbb2317239ecfa48339f*

Comment 2.15: In page 12, line 37, what does "manner" mean here? Change of terminology?

➔ ***Reply 2.15:*** *We apologize for the confusion. We changed the text to "privacy preservation" - The majority of studies describe "privacy preservation" very briefly in their own understanding.*

Comment 2.16: In page 13, line 25, "In the rest [of the] papers", "of the" missing.

➔ ***Reply 2.16:*** *We changed the text to "in the rest of the papers"*

Comment 2.17: In page 14, line 40, "SMC.[69, 70] combined", remove the dot.

➔ ***Reply 2.17:*** *We apologize for the mistake. The dot is removed in the revised version.*

Comment 2.18: In page 17, the Discussion section could use some more structure (sub-sections).

➔ ***Reply 2.18:*** *We agree with the reviewer and structured the discussion section into six subsections:*

5.1. Inadequate definition and measurement of privacy
5.2. Ambiguity between privacy and security
5.3. Inadequate experiments and practical use cases
5.4. Challenge of linking data in vertically partitioned data scenario
5.5. A recommendation list of key parameters for PPDDM studies
5.6. Potential limitations

Comment 2.19: In page 18, line 1, "privacy violence", do you mean violation?

➔ ***Reply 2.19:*** *We apologize for the mistake and we changed the word to "privacy violation" in the revised version.*

Comment 2.20: In page 18, lines 23-24, the comment on decision trees needs more explanation.

➔ ***Reply 2.20:*** *Thank the reviewer for this comment. We added more explanations to the decision tree example in Section 5.2 on page 20.*

"A typical example is building a decision tree on vertically partitioned data in a privacy-preserving way. The decision tree model can be securely and correctly built up. However, to some extent, the decision tree, as an output, leaks information about the input data [109]. Decision tree algorithm splits nodes based on attributes or features, while the splitting decision is dictated by the data. When the final decision tree is completed, the leaf nodes in the tree might reveal some information about the input data such as class counts.

Therefore, releasing the final decision tree to all participating parties could potentially breach privacy."

Comment 2.21: In page 18, lines 27-29 and line 34, this might arise from the fact that any information can potentially be used to infer sensitive information from people or organisations participating in the distributed DM/ML schema. For instance, you cannot foresee the impact of linkage attacks without knowing the data present in other datasets. And getting this knowledge is nearly impossible. I would guess this is why most papers refrain from distinguishing which data is more/less sensitive, instead treating every single piece of data equally protection-worthy.

➔ ***Reply 2.21:*** *Thank the reviewer for this helpful comment. We agree that any information can potentially be used to infer sensitive information. However, data protection laws such as GDPR, do categorize personal data into different sensitive levels. For example, genetic and biometric information, data about people's racial or ethnic origin, political opinions, religious and philosophical beliefs are all into special category data which requires additional conditions. Moreover, data parties/organizations also classify data into different confidential/sensitivity levels for the use and management of the data. Different requirements and standards are designed for data in different sensitivity levels. Therefore, we think it is necessary to know "Which data are deemed sensitive or require protection, and why?"*

Comment 2.22: In page 19, lines 11-12, reference missing for the national citizen identifier in the NL.

➔ ***Reply 2.22:*** *We added "wet algemene bepalingen burgerservicenummer)" as reference in the paper.*

Comment 2.23: In page 21, lines 27 and 32, strange extra hyphen in the first column of the table.

➔ ***Reply 2.23:*** *Thank the reviewer for pointing out this. We removed all hyphens that were in the wrong positions.*

Comment 2.24: In page 22, lines 14-15, I don't understand how the authors came to this conclusion, was it something not discussed in the reviewed papers?

➔ ***Reply 2.24:*** *If I understand the comment correctly, I assume the reviewer meant this conclusion "Future research will preferably balance this trade-off depending on their specific use cases and the purpose of the data analysis.". Some of the reviewed papers did recognize the trade-off of privacy preservation and efficiency in general. However, we want to highlight that in the special time or use cases, for example, in the COVID-19 pandemic, this trade-off has a different balance between individual privacy and model performance (such as accuracy and efficiency) for reasons of substantial public interest (in the area of public health in this case).*

Therefore, We elaborated on this conclusion by adding - "In the outbreak of the COVID-19 pandemic, the processing of sensitive data including personal health data is allowed by the General Data Protection Regulation (GDPR) and other European data protection laws in order to protect against serious cross-border threats to health [121]" in the revised paper.

Comment 2.25: In page 18, lines 38-39, this could also be a reflection of the lack of forward reference search. What if there are follow up works with experiments?

➔ ***Reply 2.25:*** *We agree with the reviewer that some papers in our review could have follow-up work/papers with experiments. Some of these follow-up papers are included in this review because they matched our search criteria. But we indeed have not searched explicitly for this scenario. We highly appreciate this comment from the reviewer and added it as one of the potential limitations in the Discussion section [Section 5.6].*

Comment 2.26: In page 22, the work lacks discussion of their limitations.

➔ ***Reply 2.26:*** *Thanks for this valuable point. We agree with the reviewer and we added a subsection in the Discussion Section to describe the limitations we have for this review.*