Chang Sun
Maastricht University, The Netherlands
chang.sun@maastrichtuniversity.nl

March 19, 2021

Submission of our manuscript "A Systematic Review on Privacy-Preserving Distributed Data Mining"

Dear editors,

Enclosed, please find our manuscript "A Systematic Review on Privacy-Preserving Distributed Data Mining" for consideration for publication in the Journal of Data Science.

In this systematic review, we reported the results and findings of a systematic review of privacy-preserving distributed data mining (PPDDM) techniques from reviewing 231 studies published in the past 20 years with summarizing the state of the art, the problems they address, and the challenges that remain outstanding. PPDDM techniques consider the issue of executing data mining algorithms on private, sensitive, and/or confidential data from multiple data parties while maintaining privacy. As an emerging field, PPDDM attracts increasing attention from both academia and industry. However, the majority of the published surveys have typically treated PPDDM as a specialised subtopic and did not provide a comprehensive review on the recent studies. Therefore, the aim of this systematic review is to provide a complete overview and detailed analyses of existing PPDDM techniques and offers new insights to the field. We hope this systematic review paper will serve as a helpful guide to past research and future opportunities in the area of PPDDM.

Due to the new factors involved in PPDDM techniques such as data partitioning problem, communication costs, and adversary behaviors, conventional data mining evaluation metrics are not adequate to evaluate new PPDDM techniques. To the best of our knowledge, there are no standard metrics for evaluating new PPDDM approaches. Therefore, in this review, we proposed a 10-factor metrics to assessed PPDDM studies including adversarial behavior of data party, data partitioning, experimented datasets, privacy/security analysis, privacy-preserving methods, data mining problems, analysis algorithms, complexity and cost, accuracy performance, and scalability. We applied these metrics to evaluate and compare the 231 selected studies in the review. We highlighted the characteristics of the 18 most cited studies and analyze their influence on other studies in the field. The results show an equal representation of horizontally and vertically partitioned data solutions and a wide range of privacy-preserving methods and data mining algorithms have been well-studied. We elaborately discussed the various definitions of privacy, differences between information privacy and information security in the PPDDM field. To minimize the ambiguity and confusion of definition of privacy in the future research, we also offer suggestions of how to make clear and applicable privacy descriptions to propose new PPDDM techniques. In the end, we provided a guideline based on the proposed evaluation metrics for researchers to conduct future research and publications in the PPDDM field.

This systematic review offers new insights into the important factors that should be considered to propose and evaluate new PPDDM techniques and how to bridge the gap between theoretical methods and practical applications in the field. The review results of all 231 studies (CSV file) will be published in a data repository (DOI: 10.6084/m9.figshare.14239937) for other researchers to retrieve or reuse our review results. The dataset is not privately saved in the repository at: https://figshare.com/s/cbb2317239ecfa48339f. We view our systematic review as an important contribution to the ongoing development of privacy-preserving distributed data mining, in which your journal has played an important role. The manuscript is not under review elsewhere and has not been published previously or accepted for publication by a peer-reviewed journal. The appropriate ethical guidelines were followed in the conduct of the research.

We are looking forward to your reply.

Sincerely yours,

Chang Sun, MSc