International Journal of Population Data Science

Journal Website: www.ijpds.org

K-Anonymity Based Privacy Risk Budgeting System for Interactive Record Linkage

Kum, Hye-Chung[1*] and Jain, Prannay[1]

[1]Texas A&M University

Objective

Information privacy theory demonstrates mathematically that privacy is a budget constrained problem and that privacy preserving algorithms(e.g., differential privacy) must rely on a budgeting system. Thus, we design a privacy measure as a function of information disclosed to support incremental information disclosure required for safe interactive record linkage. The privacy measure will determine the increase in the privacy risk for any given information disclosed during record linkage.

Methods

Mathematically, the identity disclosure risk is inversely proportional to the number of entities in the population that share the information disclosed. If the information refers to one and only one person in the population, then the identity of the person has been fully disclosed by the information revealed. On the other hand, if the information disclosed is identical for multiple people(say n), then the information is less revealing as it could refer to any one of the n people. The larger the n, the lower the privacy risk. Thus, the anonymity-set size is defined as the number of people in the population that share the same identifying information. The privacy risk measure has one prespecified parameter k, which represents the minimum anonymity-set size to guarantee no privacy risk. That is, for any disclosed information, if the anonymity-set size is less than k, then a privacy risk is present and the risk score will be calculated. A commonly accepted threshold for k is 5 or 10. On the one hand, when all entities have anonymity-set size less than k, the privacy risk would be 100%. On the other hand, if all entities have anonymity-set size greater than or equal to k, the privacy risk would be 0%.

Results

The budgeting system contributes to the much-needed methods for protecting privacy while still supporting high quality interactive record linkage by allowing safer manual resolution of uncertain linkages. The budgeting system supports refining effective visual encoding techniques for incrementally revealing only the required information on an as-needed basis during manual resolution of uncertain linkages as well as refining the design for a visual interface to facilitate privacy preserving data standardization, cleaning, and conflict resolution for interactive record linkage. We evaluate the budgeting system with the NC voter registry data.

Conclusion

The k-anonymity based privacy risk budgeting system provides a mechanism where we can concretely reason about the tradeoff between the privacy risks due to information disclosed, accuracy gained, and biases reduced during interactive record linkage.

*Corresponding Author:
Email Address: kum@tamu.edu (H. Kum)

http://dx.doi.org/10.23889/ijpds.v1i1.241