Statistical Journal of the IAOS 32 (2016) 271–274
DOI 10.3233/SJI-161001
IOS Press

Discussion of the synthetic data papers published in the previous issue[1]

Jörg Drechsler
Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg, Germany
E-mail: Joerg.Drechsler@iab.de

In our data driven society in which we expect that all major decisions are backed up by empirical evidence based on high quality data, broad access to these data is a must. However, the benefits of broad data access need to be balanced against potential risks of disclosure. Most data gathered by government agencies are collected under the pledge of confidentiality and the agencies have a legal and moral obligation to guarantee this pledge. Furthermore, if respondents get the impression that their data are not sufficiently protected they might refuse to participate or purposely provide wrong answers jeopardizing the quality of the collected data. Statistical agencies thus have to address this trade-off and much progress has been made in the last decades increasing the amount of data available for the general public while maintaining the confidentiality of the survey respondents.

Still, there are certain types of data for which addressing this trade-off is particularly difficult. Medical records containing sensitive information on health status are one example, another example are business data. These data are particularly difficult to protect since a few variables usually suffice to identify larger businesses in the data. At the same time the collected information is often sensitive since other establishments might gain an edge if they learn certain attributes about their competitors. For these reasons access to business data is very restricted. Most data collecting agencies do not offer access to their business data and if they do, the data can usually only be analyzed on the premises of the agency by sworn in researchers after a lengthy application process. Finding ways to simplify and broaden the access to business data for external researchers and the general public is thus a topic of intensive research.

The five papers published in the last issue of the SJIAOS under the title "synthetic establishment microdata around the world" are thus a welcome and timely contribution on this important topic and I really enjoyed reading all of them. I want to congratulate all authors for their very interesting and relevant contributions and the journal for putting together such a nice set of papers from some of the leading authors in the area of data confidentiality. Even though not all of the papers actually deal with establishment data and some papers use other approaches than generating synthetic data for data protection, they all deal with powerful and innovative methods for statistical disclosure control that could be used for protecting very sensitive data such as business data. Instead of discussing each paper in turn, which could be lengthy and boring, I tried to identify common themes of the different papers which I will comment on in this brief discussion.

The papers by Miranda and Vilhuber (MV), Wei and Reiter (WR), and McClure and Reiter (MR) deal with synthetic data according to the definition that is commonly used in the statistical disclosure control (SDC) community, i.e., synthetic data that are generated based on the ideas of multiple imputation for miss-

[1] Statistical Journal of the IAOS, vol. 32, no. 1, pp. 65–68, 2016, Statistical Journal of the IAOS, vol. 32, no. 1, pp. 69–80, 2016, Statistical Journal of the IAOS, vol. 32, no. 1, pp. 81–92, 2016, Statistical Journal of the IAOS, vol. 32, no. 1, pp. 93–108, 2016, Statistical Journal of the IAOS, vol. 32, no. 1, pp. 109–126, 2016, Statistical Journal of the IAOS, vol. 32, no. 1, pp. 127–135, 2016.

ing data. The papers address a wide range of topics, from new synthesis strategies to preserve fixed totals when disseminating magnitude microdata (WR), over disclosure risk assessments for synthetic data in general (MR) to a clever secondary use of synthetic microdata to avoid cell suppression when disseminating tabulations (MV). All papers illustrate that the synthetic data approach can be a viable solution if analytical validity should be maintained even if a high level of protection is required. However, as the assessments in WR illustrate, if the protection requirements are very strong, some loss in validity is inevitable.

Noting that Schmutte (S) also uses the term synthetic data to describe the output of a differentially private mechanism, I wondered if it is finally time to come up with a new label for artificial data that are generated based on Rubin's multiple imputation ideas. On the one hand any artificially generated data are synthetic data which often causes confusion. For example, I received several requests to write encyclopedia entries on synthetic data but I had to turn down the invitations since it was obvious that the editors had a different idea of synthetic data which was not related at all to statistical disclosure control or data protection. On the other hand potential users of the data shy away if they hear that they should be working with synthetic data. It sounds too much like made up data and especially less statistically trained users of the data will be skeptical if they can trust the results. Of course users of the data should be critical but it seems to me that most users are much less concerned about other perturbation methods such as swapping or noise addition that are often applied before the release even though their effects on analytical validity can be much more troublesome. My impression is that this is at least partially due to the label of the method. I must admit I don't have any good suggestions for a new label and it will be difficult changing a name that has been used for more than 20 years now. But I will be happy to adopt a new terminology if anyone comes up with a good suggestion.

The papers by MV and WR also touch on an important area of research that did not get much attention in the literature so far: How can we quantify the risks of disclosure if more than one data product is released based on the same underlying microdata? In most cases the risks are only evaluated considering the information contained in the data product to be released. But the available information from the other data products needs to be taken into account when measuring the risks. There is no question that more information always results in an increase in the risk of disclosure. The important question is, whether these increased risks are still acceptable. But quantifying these risks is a difficult task, especially if the data products consist of different types of data, i.e. microdata and tabular data as is the case in both papers. WR mostly focus on the utility perspective by proposing strategies for generating synthetic data that match on totals that might have been published before. If the synthetic data exactly preserves the previously released information, traditional risk assessments based on the synthetic data alone will be sufficient since there is nothing that can be learned from the previous releases that is not available from the synthetic data. But this only holds if all previously published totals are preserved. However, it is exactly this scenario for which WR find unacceptably high risks of disclosure – at least under their very conservative risk assumptions. My impression is that measuring the risks of joint data releases is an underdeveloped area of research and holistic approaches for incorporating all published information when measuring disclosure risks would be an important area for future research.

A different approach for data protection based on noise infusion is discussed in the papers by MV and Abowd and McKinney (AM). Noise infusion is a common strategy for data protection and it is used, for example, by the U.S. Census Bureau for the Quarterly Workforce Indicators computed from the Longitudinal Employer Household Dynamics (LEHD). While MV use the noise approach as an alternative strategy to synthetic data for filling in cells in tables that would otherwise be suppressed, AM use the noise approach for protecting graph-based statistics. (As a side note I highly recommend the latter paper to anybody working with linked employer employee data (LEED) as it offers a gentle introduction to graph theoretic approaches for analyzing this type of data. It seems to me that graph theory can be a powerful tool for analyzing LEED and in my view it did not get the attention it deserves). Both papers illustrate that a high level of analytical validity is achievable using multiplicative noise. However, there is one considerable drawback of the approach. Compared to synthetic data it is very difficult to take the extra uncertainty that comes from the protection into account, especially since the noise infusion mechanism that is used is rather complex. This will generally lead to biased inferences obtained from the protected data. The bias might be small for point estimates at a high level of aggregation since the noise distribution is chosen to ensure that the expected bias is zero, but measures of association will always be affected. For example Chi-square tests for independence

of the variables contained in the released tables will always provide invalid p-values.

Of course this is an interesting question with room for heated discussions: Are perturbation techniques preferable that are known to distort all inferences but for which we expect that the impacts will be minor for many statistics of interest? Or should we always insist on perturbation methods for which it is possible in theory to obtain unbiased results but which often lead to substantial bias in practice? The statistician in me will always opt for the latter since methods that are known to generally lead to biased inferences should be avoided. The practitioner in me has a more balanced view on this issue. Shouldn't we aim at minimizing the negative impacts of the data protection methods being used? Maybe a small known bias is preferable to the risk of potentially substantial bias, especially if we cannot judge based on the released data how large this bias might be? Arguably, with the complex noise mechanism that is discussed in the two papers it will also be difficult to quantify the expected amount of bias but at least we can be confident that cell counts in tables with sufficiently large cells will only slightly be affected.

This also points to another important aspect regarding the discussions on data dissemination: I believe there is no such thing as the data user. Different users will be interested in completely different aspects of the data. Politicians or journalists will require different data products than academic researchers and it makes sense to use different instruments from the data protection tool kit depending on the user needs. For the former, noise infusion might be the best solution since they will often be satisfied with table counts. For the latter, the synthetic data approach potentially coupled with verification servers [1] that offer some information how close the inferences obtained from the protected data are to the inferences from the original data might be a better solution. Of course, once we are taking about access to the underlying business microdata, the synthetic data approach will be the only solution in my view since all other methods would either not sufficiently protect the data or would have to be applied so extensively that the resulting data would be useless.

A final common theme that I identified is that all papers which include some risk evaluations use very conservative assumptions about the available background knowledge when quantifying the risks. DR, WR, and S basically assume that the intruder possesses the information for every unit included in the database except for one record and the goal of the intruder is to get a good estimate for this last piece of information. The first two papers need this assumption for computational reasons, the last paper is based on the concept of differential privacy and a key element of the formal privacy guarantees offered by this concept is that it limits the amount of information an intruder can learn from the data even under these rather extreme assumptions. I noticed a trend in recent years that more and more papers that offer risk evaluations make similar restrictive assumptions. But I wonder if this is really helpful. Obviously, these assumptions are unrealistic but an argument often made is that if the estimated risks under these assumptions are low they are certainly even lower in more realistic scenarios and thus the data can be disseminated without any confidentiality concerns. While I agree that as long as the planned data release shows low risks under these assumptions everybody should be happy, I am not sure what the actual implications are if the risk evaluations indicate a high risk of disclosure. For example, WR find that one of their proposed methods lead to unacceptably high risks of disclosure under their conservative assumptions regarding the background knowledge of the intruder. But does that really mean we should not consider releasing the data? My worry is that we will become more and more overprotective if we only use these kinds of risk assessments and a lot of valuable information will never be disseminated. Of course it is desirable that we are able to quantify the risks of disclosure and the attractiveness of differential privacy is that it is the only existing strategy that guarantees a defined level of protection irrespective of the actual data, of any future data releases, or other sources of information that will be available to the intruder. But the important question is which price are we willing to pay for these guarantees?

In fact, this has been the major point of criticism regarding the concept of differential privacy. The implicit assumption that the intruder knows all records in the database except one requires strong protection mechanisms to sufficiently protect this one record. The mechanisms that have been proposed to achieve privacy under these conditions usually alter the data so much that no useful information can be learned from the released data. The paper by Schmutte is a valuable contribution to this debate as it is one of the few papers that actually evaluate the analytical validity of a differential private mechanism in practice. Interestingly, he finds that the validity is substantially higher than the theoretical lower bound for the privacy mechanism being used, although the validity is still unacceptably low in my view. But the paper also offers an

interesting economic perspective regarding privacy in which a privacy mechanism offers "production possibilities" for the statistical agencies. These possibilities need to be evaluated against the social preferences for privacy relative to accuracy to find an optimal solution. I think this concept addresses an important point: any release of information implies a reduction in data privacy. These are two sides of the same coin and statistical agencies should not aim at finding dissemination strategies with zero risk, since this is simply impossible. We have to face this fact and try to find the "socially optimal choice of privacy and accuracy" as the author puts it. However, finding this optimal choice is the major challenge even if we rely on differentially private mechanisms. We cannot simply include a question in the next Census asking which level of ε would be considered acceptable. And even if miraculously all citizens would fully understand the concept of differential privacy and would be able to quantify their preferred choice of accuracy versus privacy in terms of ε, it still would not be clear what the socially optimal choice would be. The smallest value of ε? The mean? The median? Would it be socially desirable that sufficiently accurate data would be released so that most citizens would benefit from reduced health care premiums because the health care providers would be able to identify high risk individuals and not offer them coverage? These are very difficult questions to answer and many advocates of differential privacy avoid any discussions about these questions by stating that their discipline is not the right one for providing answers here. But even though I agree that computer scientists or statisticians might not be the right persons to answer these questions, I still think these are the right questions to ask, since there will always be a trade-off between the benefits for the society from broad data access and the potential threats for individuals included in the data because the information learned about them might be used to their disadvantage. As long as we don't know how to identify the optimal choice between privacy and accuracy the most powerful tools to quantify and limit the risks will be useless.

Reference

[1] J.P. Reiter, A. Oganian and A.F. Karr, Verification servers: enabling analysts to assess the quality of inferences from public use data, *Computational Statistics and Data Analysis* **53** (2009), 1475–1482.