Author's response to reviews

Title: Privacy-preserving record linkage using Bloom filters

Authors:

Rainer Schnell (Rainer.Schnell@uni-due.de)
Tobias Bachteler (Tobias.Bachteler@uni-due.de)
Jörg Reiher (Joerg.Reiher@uni-due.de)

Version: 2 Date: 28 June 2009

Author's response to reviews: see over
Dear editor,

Thank you for considering our manuscript for publication. We are grateful to the reviewers for their careful reading and suggestions, many of which we feel are helpful in clarifying aspects of our paper and which we have taken on board. Below we have indicated our changes and responses to the reviewers feedback.

Best regards,
Rainer Schnell

Reviewer 1 (Chaoyi Pang):

Minor Essential Revisions:

1) Line 4: “Existing protocols …. are inefficient in actual research practice.” => Many Existing protocols are inefficient in matching record pairs under encrypted identifiers where small errors may exist.

Response: We have revised the text in the abstract, so that it now reads: “Most existing protocols either drop any potentially matching record pairs with errors in identifiers or are inefficient in actual research practice.”

2) Page 2/line7: “…drop any potentially matching record pairs with errors in identifiers …” => … drop many potentially matched record pairs…

Response: The text on page 2 has been revised as above.

3) Figures are not properly labelled.

Response: Captions and labels have been changed according to the journal’s guideline.

4) References: Although this article including many good work on PP-linkage, some more recent papers may need to be cited. E.g., Xiaokui Xiao, Yufei Tao, Dynamic anonymization: accurate statistical analysis with privacy preservation, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, June 09-12, 2008, Vancouver, Canada.

Response: The paper by Xiao/Tao suggested by the reviewer does come from a related but different literature. We have added the following sentence to the first paragraph of the Introduction: “Methods for privacy preserving record linkage can be subsumed under the general field of privacy preserving data integration [4] which also comprises a vast literature on privacy preserving database joining and querying [references below are added here]. However, this work either uses exact identifier comparisons or does not address the problem of generating micro data sets which are usable for general statistical purposes.”


5) Quality of written English: Needs some language corrections before being published.
Response: The article has been carefully reviewed by two additional native speakers.

Reviewer 2 (Karen Olson):

Minor Essential Revisions:

1) Figures 3 and 4 are so much harder to read those those in Figure 2, probably because of the overlap. Could the range on the axes be restricted somewhat to make them more clear?
Response: We added a 5th figure which shows the interesting range mentioned in the text (recall levels above .75).

2) Comment on the problem of needing to share data that must contain elements not used for linkage, but can be potentially identifying.
Response: Generally, if two data files are linked the resulting data file is more easily deanonymized since there is more information available. However, this effect is independent of the specific record linkage procedure used. To clarify this point, we have added the following paragraph to the end of the discussion: “Of course, the recipient D has a greater chance of re-identifying individuals from the merged data file as compared to the individual files DBa and DBb. However, this pertains to any kind of data linkage and is not an intrinsic problem of the proposed method. Any real world application of record linkage protocols has to guarantee factual anonymity of the resulting micro data sets. This can be achieved in many different ways. A technical option would be the use of micro data disclosure control algorithms [the reference below is added here].”

Reviewer 3 (Catherine Quantin):

1) Background Section: “existing procedures either drop any potentially matching record pairs with errors in identifiers or are inefficient in actual research practice.” This assertion is not always true: if a probabilistic linkage method is used, this method can be applied whether the data was rendered anonymous by hashing or not, as probabilistic linkage is based on a binary comparison (a record pair is then matched or not matched).

Response: We have changed this sentence to reflect the point made by both this reviewer and reviewer 1 (see the responses to Reviewer 1).

2) Introduction section, last paragraph. The authors need to explain in more detail the advantages of working on nominative (not hidden by hashing or encryption) data, meaning the possibility of also comparing record bi-grams and not only the whole record. Of course, the comparison of record bi-grams is not possible after hash-coding records. The interest of Tim Church’s paper (reference 5) is to propose to create bi-grams before hashing. Then, it is possible to compare bi-grams, even after hashing. It would be very useful for the reader if this were mentioned in the introduction.

Response: We have changed the last paragraph of the Introduction so that it now reads: “Since these protocols require exact matching of identifiers, they do not tolerate any errors in identifiers: Due to the design specifications of cryptographic functions, the slightest input variation causes many changes to the output (ideally, a change of one input bit should cause a change in half of the output bits). By contrast, a protocol allowing one to evaluate a string similarity function will improve the linkage quality considerably. In addition, since records with variations of identifiers may have different characteristics than records with exact matching identifiers, restricting the linkage to exact matching records is not an option. Therefore, a method for approximate string matching in privacy-preserving record linkage is needed.”

3) Pages 3-4-5, Sub section: Three-party protocols (of Related work): First paragraph: the reference dated 1979, is out of date; the new machines have made the proposed method obsolete, and systems used to attack networks have considerably evolved over the last 30 years. This paragraph could be removed.

Response: The paragraph describes a procedure for privacy preserving record linkage which, we accept was suggested early in the literature in this field, is still proposed and discussed by recent papers (cf. reference [13]). Therefore, we felt it important to mention in the Related work section. To our knowledge, Boruch and Cecil (1979) were the first to describe such an architecture thoroughly, so we feel that in accordance with good research practice, it is important to cite them.

4) Second paragraph: “identifiers are transformed according to phonetic rules and subsequently encrypted with a one-way hash function”. This protocol is quite old and has been revised to avoid false positive links. In a more recent paper, authors have proposed using successive probabilistic linkage on data whether or not it has previously been phonetically transformed (Methods of Information in Medicine, 2005).

Response: We have added a reference to the more recent publication of the reviewer.

5) Concerning the protocol regarding reference 24
5.1) The authors should add comments on its performance.

Response: We have added a sentence to the paragraph, “The performance of the protocol depends crucially on the set of reference strings. Unless this is a super set of the original strings (which is difficult to find in advance), the performance is rather disencouraging.”

5.2) The description of the protocol is too detailed and very difficult to follow. It needs to be synthesised.
Response: The description is explained in just 6 lines and focuses on the essential details only. The only way to make it easier to follow would be to expand it. To adhere to the journal’s guideline to be concise we elected to make no change beyond the requested comment on its performance.

5.3) Moreover, it would be preferable to present the protocol regarding reference 26 just after the protocol regarding reference 24.

Response: We have changed the text as proposed.

5.4) In fact, the two protocols compute the Euclidean distance between the records, and it would be very useful to focus on their differences, and discuss their approaches, advantages and disadvantages, so as to highlight the proposed solution.

Response: The main intent of our article is to introduce a new method of privacy preserving record linkage. However, good research practice advocates mention of alternative approaches. To compare these approaches and to discuss the relative merits of them would mean to spend too much valuable space to what is perhaps more suited to a review article of this area.

5.5) The authors should also indicate if this distance corresponds to the standard Euclidean distance or the normal one.

Response: In [24] no Euclidean distance is computed at all. In [26], the standard Euclidean distance is computed, not the normal(ized) Euclidean distance. We have revised the text accordingly.

6) The protocol regarding reference 6, is not sufficiently detailed, which makes it difficult to understand what it can really do. It may be removed.

Response: We mention reference 6 because it is an important alternative approach to ours. In our opinion, a related work section should mention and briefly describe alternative approaches but is not the place to explain them in detail.

7) Concerning the first paragraph of page 5, (last protocol), the authors said that it is possible to use “data sets without identifiers”. The authors should then indicate if they evaluated the concordance or discordance based on hashed data.

Response: We accept and are grateful to the reviewer for pointing out the error in this passage. We have changed it to “They suggest using anonymized versions of the data sets for a first linkage step that is capable of classifying a large portion of record pairs correctly as matches or mismatches.”

8) Page 5, Sub section: Two party protocols: This paragraph is not clear and does not give precise information regarding the proposed protocols. More details should be added particularly regarding references 26 and 32.

Response: See response to item 6)

9) The expression “impractical for linking large databases” is repeated several times in the text.

Response: Our entire manuscript contains the expression “impractical for linking large databases” only once.

10) Page 6, second paragraph: In the cited example, the DICE value seems to be very close, when the number of bi-grams based on nominative (not hidden) data and when the number of bits is set to 1. This single example is not sufficient to demonstrate that the results are always similar.
Response: We have given the example to demonstrate how our method works. It is not intended as a proof or even a demonstration that the results are always similar. On the contrary, it is important to understand that the similarity depends on the number of hash functions used given the length of the bit array. We have made this possibility clear in the section “Results for simulated databases” on page 8.

11) Page 7, a bracket should be added in the first formula.
Response: We have added an appropriate pair of brackets.

12) Page 8, Test on simulated databases: Simulations are limited to very specific situations (a fixed filter length of 1000 bits and tri-grams). Why not try other situations?
Response: We intended to introduce a new method and to show that this method works in principle. Presenting the results of a detailed investigation of its properties is the focus of a forthcoming paper and Phd thesis.

13) Page 8, Procedures and data sets:
13.1) Orthographic pre-treatments are carried out. We need to know if the performance of the method would be the same without these orthographic pretreatments.
Response: Our method is suggested to be used as an alternative when customary similarity functions can not be evaluated due to privacy concerns. However, this objection pertains to any similarity function, whether it be evaluated privacy preserving or not. That is, if we compare our method to Bigrams calculated from plain text names, the effect of the orthographic treatments is held constant.

13.2) The situations envisaged, for these orthographic errors (changing one character per name) are very limited. In usual situations it is also possible to observe the inversion of two letters (two modified characters), double letters, and errors on many consecutive letters (ch instead k,…).
Response: As we have indicated in response to item 12), our intenation was to introduce a new method and to show that this method works in principle. Moreover, we report the results of a test using real data comprising real orthographic errors. Of course, much additional testing under various conditions is needed. In the near future we will perform such tests thoroughly and report the results in a forthcoming paper, citing this current manuscript as the first description of the method.

14) Figure 2, the number of sub figures of this figure 2 should be limited to 3. It would be interesting to add other figures corresponding to other simulation results, (with bi-grams, changing the filter length…).
Response: Using these plots, we want to demonstrate monotone decline of the Bloom-curve as the number of hash functions increases. In our opinion all 6 subfigures are necessary to make this point.

15) The authors need to indicate if the performance (which is good) would be the same when the parameters have been changed. If possible, it would be very interesting to show the evolution of the results depending on different parameters. In a real situation, due to the absence of a gold standard, it would be very difficult to make the right choice regarding the parameters. So, some advice about the choices would be very useful and could be included in particular in the protocol (pages 9–10). This would help in the choice of the parameters depending on the string length; for example, a special strategy could be proposed for “short strings”.
Response: As mentioned in the response to item 12), we will present the results of a detailed investigation of the method’s properties in a forthcoming paper.
16) The method used for linkage is not mentioned. After calculating the similarity between two strings, which linkage method will be applied? The authors should mention if they recommend the use of deterministic or probabilistic methods.

Response: The text (p. 9, Procedures and data sets, sentence 1) has been revised. “In the context of an evaluation of different probabilistic record linkage procedures for research purposes, we conducted a test of the proposed procedure on two German private administration databases.”

Reviewer 4 (Thomas Aden):

Minor Essential Revisions:

1) It would be nice if the authors would stress that q-grams are applicable to textual data only.

Response: We have added the following paragraph to the end of the section “A Protocol for privacy-preserving record linkage”: “Our protocol is intended to match string attributes only. However, as outlined above there are various protocols for matching attributes exactly in a privacy preserving manner. These protocols can be used for the comparison of numerical attributes in combination.”

2) The authors might discuss that the protocol for privacy preserving record linkage is secure only as long as a data holder A does not collaborate with a recipient D.

Response: Again, this problem is independent of the specific record linkage procedure used. To clarify this point, we have added the following paragraph to the end of the discussion section: “Another problem might be if the recipient D transfers the merged data file to one of the data holders. The latter would then be able to re-identify individuals in the other holder’s data base. However, this threat is common to any linked data file. If such a transfer cannot be prevented by legal means, the only remedy for this problem is a second trusted party E. E supersedes the recipient D in the protocol, merges the files DBa and DBb in his place, and perturbs the merged data file using disclosure control algorithms. Afterwards, E transfers the perturbated file to the recipient D.”

3) and 4) The authors should discuss how the set of best matching pairs is determined/discussion on how to handle pairs (a,b) and (a,c) that both have the same dice similarity.

Response: Our method is not proposed as a classification method, but is intended to be used instead if customary similarity functions can not be evaluated due to privacy concerns. However, these objections pertain to any similarity function, whether be evaluated privacy preserving or not.

5) It would be interesting to see results for scenarios with a very large number of records, i.e. 100.000 - 1.000.000 records in each database.

Response: This would mean at least 100,000*100,000 comparisons, far more than normally performed in record linkage applications. Usually, blocking strategies are adopted to reduce the number of comparisons. String similarities are calculated just within the blocks which seldomly contain up to or more than 10,000 records. That is, a scenario with 100,000 – 1,000,000 records would be quite unrealistic.