Reviewer's report

Title: Privacy-preserving record linkage using Bloom filters

Version: 1 Date: 5 June 2009

Reviewer: Thomas Aden

Reviewer's report:

Major Compulsory Revisions
- None.

Minor Essential Revisions
1. It would be nice if the authors would stress that q-grams are applicable to
textual data only, i.e. strings are used to identify an individual (record) in a
database only. But attributes like names etc. might not be sufficient to identify
records or individuals unambiguously and additional data might need to be used.
Neither the method nor the protocol presented addresses this issue.

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
(or is identical to D). In such a case A would be able to link records from B to
individuals in its own database. This would presumably not be able otherwise.

3. The authors should discuss how the set of best matching pairs is determined,
i.e.
what is the minimal Dice similarity so that a pair is considered to be a match.
Is it possible to estimate a threshold value for the Dice coefficient
automatically, although different variables, such as number of records in
databases A and B, Bloom-filter length and number of hash functions influence
such a value?

4. A discussion on how to handle pairs (a,b) and (a,c) that both have the same
dice
similarity (or at least have a similarity that is above a threshold value) would
be valuable. It is obvious that if a and b apply to different individuals, at
least one of both pairs would represent a false positive. In numerous cases the
result of a record linkage process is a set of such pairs that are related to
each other.
Discretionary Revisions

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.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.