External.  
  
Hello,  
  
   Sorry that it took so long. I send enclosed the comments. According to  
the reviewers, we ask for a major review. Please, send us the revised  
version in two month. If you need more time, let me know, thank you for  
your submission to the journal.  
  
vicenç  
  
----------------------------------------------------------  
Transactions on Data Privacy           <https://nam10.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.tdp.cat%2F&data=05%7C01%7Ccdb327%40drexel.edu%7C67f8d7f60ec847bea5a208db1f9e1f36%7C3664e6fa47bd45a696708c4f080f8ca6%7C0%7C0%7C638138536241745177%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=VmTrHn7naIS%2B%2Fz5V7e%2BhRefwhF6QgNq%2F9rPwSaJD6xQ%3D&reserved=0>  
----------------------------------------------------------  
  
Authors: Cameron D. Bale, Jordan L. Fischer, Matthew J. Schneider, Steven  
Weber, Suzanne Chang  
Title:Legally Anonymizing Location Data Under the GDPR  
Paper-ID:  
  
ACCEPTANCE (scale 1-5):  
  
[ ] Accept as written  
[ ] Minor revision  
[x] Major revision (needs re-review)  
[ ] Reject (unsuitable for publication)  
[ ] Transfer to more suitable publication  
  
GENERAL EVALUATION:3  
(1: excellent; 5: bad)  
  
DETAILED EVALUATION: 3  
(1: completely agree/true; 5: disagree/false)  
  
Is the topic relevant to TDP? Yes  
Does the paper make a valuable contribution? Yes-ish  
Is the paper deep? No  
  
Is the presentation good (clarity and organization)? Ye  
Is the language acceptable? Yes  
Is the title appropriate? Yes, although they claim in the abstract ot be  
more general than just localization data  
Is the abstract appropriate? Not quite - the paper is an example of how  
you might approach it, but doesn't give the step-by-step guide it promises  
Are the references appropriate? yes  
  
TYPE OF PAPER:  
  
[] Research paper  
[x] Application paper  
[ ] Other (please specify):  
  
- Technical soundness:  
[ ] Technically correct (after careful check)  
[X] Apparently correct (after superficial check)  
[ ] Minor errors (please indicate them)  
[ ] Major errors (please indicate them)  
[ ] Unsupported claims  
  
- For applications and methods analysis:  
[ ] Conclusive analysis  
[x] Real world problems  
[ ] Toy problems  
[ ] Application description (no experiments described)  
  
Are the experiments explained with enough detail? Yes  
  
Can an interested reader reproduce the experiments? Yes if given access to  
the data  
  
  
REMARKS TO BE FORWARDED TO THE AUTHORS  
(please indicate here what needs to be revised).  
  
Overall perspective  
  
This is a clear summary of the legal and practical context of  
anonymization, some analysis of what GDPR regulations might mean  
statistically, and a case study applying those statistics to a specific  
case. It provides a useful example of the compromises necessary in  
practical anonymisation. However, the statistical analysis is both overly  
complicated and missing some important elements. Ultimately, this analysis  
all comes down to "have 2 or more observations on each unique set of  
identifiers, plus look at class disclosure" - I'm not sure that this is  
needed section 4 to come to this conclusion, although the actual applied  
work is of value. So I think this is a useful addition to the literature,  
but the middle section needs substantial revision.  
  
Specific comments.  
  
Abstract and sections 1, 2, 3  
The introduction and section 2 are very clearly written and set out the  
context very well. I like the continual focus on 'reasonableness'  
throughout the paper; the authors are continually considering the utility  
of their solutions.  
Is this really the 'first assessment determining whether existing  
anonymization solutions for location data are capable of legally  
anonymizing location data'? I thought quite a few poeple had studied it,  
and the literature review supports this. Is this perhaps teh first attempt  
to define general rules and apply to goegraphical data?  
  
Section 4 framework  
The referencing is a bit confusing here (and in the next section). N is  
used for columns rather than observastions, y[m,n] appears to refer to  
both the specific row varaible indexed by m and n, but also the subvector  
of matchable characteristics; later on d has two quite different meaings.  
Would be helpful to review these  
Singling out  
  - is eq 2 over complicated? Don't we just need inf[z] >=2?  
  - the criterion (of a correct match) assumes that the attacker \*knows\*  
that it is a correct match - this is a generally unsolved problem in SDC  
and its legit to assume this but I would note it  
  - Table 2 has an obvious attribute inference from the class (group)  
disclosure  int eh top two rows - you tackle this later but suggest you  
note here that this is a later problem (or remove the column)  
  - the authors make an important comment about k-anon vs their equation  
but surely there is no difference as you would assume that all the  
identifying/scanning vars are included in both k-anon or their set y[n,  
m]? If they are not indetifiyng variables, why include them in the  
uniqueness set?  
Linkability  
  - surely many-to-one does allow inference attacks via class disclosure  
(eg everyone visiting the covid testing centre was found to have covid, or  
everyone visiting the family planning centre was female)?  
  - it seems to me that the linkability criterion is jsut the Singling out  
criterion if the N terms used for Singling Out are those in the CI? If the  
CI has fewer vars than the Singling Out attributes then presumably no  
singling out => no linkability  
  - I think the linkability criterion is a red herring - it's very closely  
tied to singling out. The only thing it seems to add is the probability of  
sucesful attribution based on the numbers in the set  
  - "In practice, we suggest removing unique rows or non-essential columns  
in Y to prevent both types of linkages." - not really helpful; we can  
assume that the researcher would remove non-essential columns as a matter  
of good practice (if they're not doing that they are not going to get much  
out of the paper...)  
Inference:  
  - this surely isn't just limited to the dataset X eg we can see in table  
2 that there is a classs disclosure in the equivalized units  
  - p1012 para 1 - if you know the sex of an individual in X, isn't this  
just linkability again ie the one-to-one?  
  - I'm very confused about this metric for risk. Why are you using the  
average probabilities? If I were the data owner I would be much more  
concerned that some individuals can have their attributes identified with  
probability 1, and not that most people have no significant prob of  
attribute identification.  
  - as the authors note, k-anon doesn't prevent attribute disclosure  
(especially via class disclosure - it would perhaps be better to consider  
as a specific attack)  
Legaliity  
  [28]'s definition of uninformativeness is very closely tied to the  
differential privacy definition; so it is not helpful when considering  
attack context (which the authors do in the rest of the paper)  
  
Section 5 case study  
I found this section clearly structured and written, and would be suitable  
and useful to circulate to a wdier audience  
Table 8 - would be really helpful to see the the quantiles split between  
SK vs Seoul so that we can get a better sense of how coarsening affectsw  
utility - I'm guessing  that within Seoul it has much less effect even at  
97.5% than in SK as a whole  
The coarsening analysis is interesting and instructive; however, I'm not  
sure about the aggregation analysis. It seems to me that the key element  
is dropping of the longitudinal component, not the aggregation. Wouldn't  
you get the same results for both coarsening and aggregation if you  
dropped the longitudinal links from both, as one should be the invesrse of  
the other?  
p1022 "one-to-one linkage or a many-to-one linkage, which prevents  
Linkability" - only the latter prevents linkability but phrasing implies  
both?  
  
Discussion  
As the authors note, the Working Group also looked at randomisation as  
well as generalization - this paper only consideres the latter. Woudl be  
useful just to add a sentence explaing why (not enough space, too many  
options, not feeeling it's a good idea?)  
  
  
Typos  
p1001: "(Id.) "? Is this a missing reference?  
p1021: second "d=1" shoudl be "d=2"?  
  
  
  
----------------------------------------------------------------  
  
Authors: Bale, Fischer, Scneider, Weber, Chang  
Title: Legally anonymizing location data under the GDPR  
Paper-ID:  
  
ACCEPTANCE:  
[X] Minor revision  
  
GENERAL EVALUATION:3  
  
DETAILED EVALUATION: 3  
  
Is the topic relevant to TDP? Yes  
Does the paper make a valuable contribution? Ok  
Is the paper deep? No  
  
Is the presentation good (clarity and organization)? 3  
Is the language acceptable? 3  
Is the title appropriate? 3  
Is the abstract appropriate? 4  
Are the references appropriate? 3  
  
  
TYPE OF PAPER:  
  
[X] Research paper  
[X] Application paper  
  
- Technical soundness:  
[X] Apparently correct (after superficial check)  
  
- For applications and methods analysis:  
[X] Real world problems  
  
Are the experiments explained with enough detail?  
  
Can an interested reader reproduce the experiments?  
  
REMARKS TO BE FORWARDED TO THE AUTHORS  
  
The introduction is very long about the discussion personal data vs  
non-personal data, and this is further extended in Section 2. Then,  
Section 3 and 4 is also about topics and concepts well known. References  
are missing.  
  
For location privacy, see also the recent paper in the journal (and  
references therein):  
A Survey on Privacy in Human Mobility  
Anna Monreale, Roberto Pellungrini  
Transactions on Data Privacy 16:1 (2023) 51 - 82  
  
I think that the authors need to further justify why the consider this  
location data and not just a standard database to propose/illustrate their  
approach.  
  
In Section 3 the authors mention singling out and reference 12. The  
authors include some critizism to this definition, and I think that it is  
relevant. That "differential privacy and likely goes above what would be  
determined as reasonable", I agree (even this is a matter of the epsilon  
parameter and the big epsilon that some companies use with very simple  
data!). So, even for particular definitions, what is the appropriate level  
of privacy is a matter of disagreement.  
  
The authors use "inf" in eq. 1. "min" works as well (the set is finite).  
Single out is just about unique records. So, maybe it is worth to refer to  
definitions of uniqueness by Skinner and  Elliot -- see (J.R.Statist.Soc.  
B (2002) 64 Part 4 855-867 A measure of disclosure risk for microdata) and  
references.  
  
For linkability and record linkage there is all the literature by Winkler,  
and other authors that have extensively used record linkage to evaluate  
disclosure risk (e.g. Torra). They discuss worst-case scenarios.  
Masking/protection does not avoid linkage.  
  
Section 4.3 is mainly about attribute disclosure. There are different ways  
to evaluate disclosure. Authors one is one of them. Interval disclosure is  
another way (as described by several authors and implemented in e.g.  
sdcMicro by Templ). See also some of the surveys or books on data privacy.  
  
There are quite a few alternatives for location data protection. The  
approach described in Section 5 is one of them. It is difficult to see in  
what extent the method is "good" in comparison with others, and about the  
compliance with the regulation, the authors already discuss at the end of  
the paper the difficulty.  
  
In overall, the paper seems more an incomplete survey-like paper with a  
case study than really proposing something really novel. The discussion is  
interesting, but incomplete. There is extensive literature on risk  
assessment that seems to me missing in the text. Also on location/mobility  
data.