

A Privacy Attack on Multiple Dynamic Match-key based Privacy-Preserving Record Linkage

Submitted for double-blind review.

Appendix

Database name	Extracted time	Number of records
NCVR	October 2019	7,688,308
	August 2019	7,607,211
	February 2019	7,449,896
	October 2018	8,114,702
	October 2017	7,846,174
	October 2011	6,233,661
MVR	September 2016	7,325,991
	January 2016	7,217,440
	September 2014	7,357,341
	June 2013	7,341,914

Table 1: Number of records in each snapshot of NCVR and MVR that we used for the experimental evaluation of the attack.

Attribute	October 2019 missing (%)	August 2019 2 months missing/changed (%)	February 2019 8 months missing/changed (%)	October 2018 1 year missing/changed (%)	October 2017 2 years missing/changed (%)	October 2011 8 years missing/changed (%)
<i>FirstName</i> (F)	0.001	0.001 / 0.023	0.001 / 0.063	0.001 / 0.160	0.001 / 0.246	0.001 / 1.027
<i>MiddleName</i> (M)	6.821	6.804 / 0.164	6.812 / 0.496	6.860 / 1.030	6.839 / 1.826	6.467 / 8.104
<i>LastName</i> (L)	0.0	0.0 / 0.106	0.0 / 0.358	0.0 / 0.730	0.0 / 1.317	0.0 / 5.583
<i>BirthYear</i> (B)	0.0	0.0 / 0.003	0.0 / 0.009	0.0 / 0.025	100.0 / 24.99	100.0 / 24.93
<i>StreetAddress</i> (S)	0.0	0.0 / 1.434	0.0 / 4.383	0.0 / 10.102	0.0 / 16.18	0.0 / 47.72
<i>ZipCode</i> (Z)	10.52	10.47 / 0.807	10.10 / 2.329	11.91 / 4.944	11.89 / 8.093	0.007 / 23.92

Table 2: Percentages of missing and changed (without missing) values for different attributes in the NCVR databases. Missing percentages are calculated for individual databases whereas changed percentages are calculated for database pairs with different time intervals between them. Every earlier database is compared with the October 2019 snapshot of the NCVR. Note that for the October 2017 and October 2011 snapshots there were no attribute called Birth year, resulting in 100% missing percentage. However, for the attack, we manually calculated an estimate for the birth year of each record using the attribute Age and those manually calculated values are used here to measure the change percentages.

Attribute	September 2016 missing (%)	January 2016 8 months missing/changed (%)	September 2014 2 years missing/changed (%)	June 2013 3 years missing/changed (%)
<i>FirstName</i> (F)	0.0	0.0 / 0.038	0.0 / 0.070	0.0 / 0.101
<i>MiddleName</i> (M)	7.474	7.551 / 0.106	7.806 / 0.201	7.943 / 0.348
<i>LastName</i> (L)	0.0	0.0 / 0.347	0.0 / 0.854	0.0 / 1.115
<i>BirthYear</i> (B)	0.0	0.0 / 0.011	0.0 / 0.020	0.0 / 0.027
<i>StreetAddress</i> (S _d)	0.004	0.004 / 0.557	0.004 / 1.447	0.004 / 2.007
<i>StreetName</i> (S _n)	0.004	0.004 / 0.397	0.004 / 1.052	0.004 / 1.474
<i>ZipCode</i> (Z)	0.0	0.0 / 0.033	0.0 / 0.086	0.0 / 0.122

Table 3: Percentages of missing and changed (without missing) values for different attributes in the MVR databases. Missing percentages are calculated for individual databases whereas changed percentages are calculated for database pairs with different time intervals between them. Every earlier database is compared with the September 2016 snapshot of the MVR.

	Attack Scenario	Step 1	Step 2	Step 3	Step 4	Step 5
NCVR	Comb	231	137	0.0004	0.0010	195
	Attr	1498	773	0.0010	0.0024	230
	Dom	6566	3293	0.0067	0.0125	227
MVR	Comb	252	147	0.0005	0.0016	281
	Attr	3403	1942	0.0027	0.0060	307
	Dom	7352	3849	0.0059	0.0162	302

Table 4: Time taken in seconds for each step of the attack when using different databases NCVR and MVR. Results are shown in seconds and averaged over all database pairs with different time intervals. Results are further categorised to attack scenarios **Comb**, **Attr**, and **Dom** as discussed in the Results section.

Database pairs	Original Prec/Reca (%)	After applying recommendation 2 with different x values						Maximum difference percentage Prec/Reca
		x = 50 Prec/Reca (%)	x = 20 Prec/Reca (%)	x = 10 Prec/Reca (%)	x = 5 Prec/Reca (%)	x = 2 Prec/Reca (%)	x = 1 Prec/Reca (%)	
NCVR-2m	98.9/98.1	98.9/98.1	98.9/98.1	98.9/98.1	98.9/98.1	99.0/98.1	99.9/97.8	+1.04/-0.21
NCVR-8m	96.4/98.5	96.4/98.5	96.4/98.5	96.4/98.5	96.4/98.5	96.7/98.4	99.9/97.5	+3.59/-1.00
NCVR-1y	95.7/97.9	95.7/97.9	95.7/97.9	95.7/97.9	95.8/97.9	96.1/97.8	99.8/96.7	+4.23/-1.18
NCVR-2y	94.4/94.0	94.5/94.0	94.5/94.0	94.5/94.0	94.5/94.0	95.0/94.0	99.6/93.2	+5.36/-0.94
NCVR-8y	87.7/82.5	87.7/82.5	87.7/82.5	87.7/82.5	87.7/82.5	88.1/82.4	97.3/81.4	+10.4/-1.29
MVR-8m	99.2/99.5	99.2/99.5	99.2/99.5	99.2/99.5	99.2/99.5	99.3/99.5	99.9/99.5	+0.69/-0.04
MVR-2y	98.6/98.8	98.6/98.8	98.6/98.8	98.6/98.8	98.6/98.8	98.7/98.8	99.4/98.7	+0.79/-0.06
MVR-3y	98.1/98.3	98.1/98.3	98.1/98.3	98.1/98.3	98.1/98.3	98.1/98.3	98.9/98.2	+0.86/-0.07

Table 7: Comparison of precision (Prec) and recall (Reca) of the linkage when improvement method proposed in recommendation 2 is applied. Maximum difference percentage in both precision and recall for each row is also presented. Different database pairs from both NCVR and MVR are used for the evaluation. Database pairs are labelled with their time difference where 'm' refers to months and 'y' refers to years. Results are shown for different x values where match-key values are set to missing of those values have a frequency $> x$.