Appendix to "Serenade: An Approach for Differentially Private Greedy Redescription Mining"

Maiju Karjalainen, Esther Galbrun, and Pauli Miettinen[0000−0003−2271−316X]

University of Eastern Finland firstname.lastname@uef.fi

1 Reproducibility

The full source code, together with XML-formatted settings files and scripts to run every experiment, can be obtained from https://github.com/maijuka/Serenade.

The Mammals data set can be obtained upon request from the authors. The MIMIC data set can be obtained from https://physionet.org/content/mimiciii/1.4/. The VAA data set can be downloaded from http://urn.fi/urn:nbn:fi:fsd:T-FSD2701. The Psycho, Pref, SSC, SR, and RightW data sets can be downloaded from https://openpsychometrics.org/_rawdata.

The steepness parameter for computing the quality function (s in Equation 4) are presented in Table A1. The optimal steepness was calculated for each data set by searching for the value resulting in the smallest sensitivity. The corresponding sensitivities are also listed in Table A1.

data set	s	Δf
Mammals	0.02	0.0059
MIMIC	0.01	0.0025
VAA	0.03	0.0075
SSC	0.01	0.0025
Pref	0.01	0.0025
Psycho	0.01	0.0025
SR	0.02	0.0050
RightW	0.02	0.0050

Table A1. Steepness parameters s for the quality function and corresponding sensitivities Δf.

2 Proof of Lemma 4

IntialPairs (Algorithm 4) receives a budget of ε to choose I initial redescriptions. The budget is divided by the number of initial pairs I and by the sensitivity

Δf so that each redescription uses $\varepsilon/(I\Delta f)$. Then, we use the exponential mechanism to choose the redescriptions with a probability proportional to the budget (line 5), as shown in Equation 3. Since the exponential mechanism gives $2\varepsilon\Delta f$ -private results with a budget of ε, choosing one redescription is then $(\varepsilon/(I\Delta f)) \cdot (2\Delta f) = (2\varepsilon)/I$ -private. Repeating this for I redescriptions, the IntialPairs algorithm is $((2\varepsilon)/I) \cdot I = 2\varepsilon$ -private.

ExtendOne (Algorithm 3) receives a budget of ε to choose one extension to a given redescription. The budget is divided by the given sensitivity Δf, so that the extension is chosen using exponential mechanism with a probability proportional to $\varepsilon/\Delta f$, similarly as above. This makes extending a redescription once $(\varepsilon/\Delta f) \cdot (2\Delta f) = 2\varepsilon$ -private.

To compute the private quality for a redescription, we apply Laplace noise to the size of the intersection and the size of the union of the input queries' supports. With a budget ε, both operations receive a budget of $\varepsilon/2$, making the quality evaluation for one redescription ε -private.

Both algorithms SerenadeES and SerenadeCS use a total budget ε_{tot}, and the user decides how to divide it into $\varepsilon_{\text{init}}$, ε_{ext}, and $\varepsilon_{\text{qual}}$.

SerenadeES (Algorithm 2) first uses budget $\varepsilon_{\text{init}}/2$ by calling IntialPairs (line 2). Then, the loop starting on line 3 is executed K times, each time calling ExtendOne with a budget $\varepsilon_{\text{ext}}/(2K)$, making the total budget spent in this loop $\varepsilon_{\text{ext}}/2$. Finally, the quality is computed for all redescriptions in $\mathcal{Q}$ with a budget of $\varepsilon_{\text{qual}}/|\mathcal{Q}|$ to add Laplace noise to each redescription. As shown above for a budget ε both IntialPairs and ExtendOne were 2ε -private, and the quality computation is ε -private. With the budgets used in SerenadeES the algorithm is $(\varepsilon_{\text{init}}/2) \cdot 2 + (\varepsilon_{\text{ext}}/2) \cdot 2 + (\varepsilon_{\text{qual}}/|\mathcal{Q}|) \cdot |\mathcal{Q}| = \varepsilon_{\text{tot}}$ -private.

SerenadeCS (Algorithm 1) computes the initial pairs in the same way as SerenadeES. The loop on line 4 is executed $|\mathcal{Q}|$ times, and the loop on line 9 is executed at most L times, which means ExtendOne on line 9 is called at most $L|\mathcal{Q}|$ times with a budget of $\varepsilon_{\text{ext}}/(2L|\mathcal{Q}|)$ each time. Then the total budget spent for calling ExtendOne is $(\varepsilon_{\text{ext}}/(2L|\mathcal{Q}|)) \cdot L|\mathcal{Q}| = \varepsilon_{\text{ext}}/2$. The quality of the redescription is evaluated after each ExtendOne call (line 10) spending $\varepsilon_{\text{qual}}/(L|\mathcal{Q}|)$ budget each time, for at most $L|\mathcal{Q}|$ times, for total budget spent $\varepsilon_{\text{qual}}$. In total SerenadeCS is $(\varepsilon_{\text{init}}/2) \cdot 2 + (\varepsilon_{\text{ext}}/2) \cdot 2 + (\varepsilon_{\text{qual}}/L|\mathcal{Q}|) \cdot L|\mathcal{Q}| = \varepsilon_{\text{tot}}$ -private.

3 Proof of Lemma 2

Without memoization, the claim follows from Lemma 1. Indeed, the process is same, except that instead of selecting an extension of one redescription, all redescriptions in $\mathcal{Q}$ are tested.

With memoization, we can think of the process as sampling from a stream into which we add new elements (the extensions of the new redescriptions) in successive iterations. The proofs of correct sampling [2, 1] still hold, as the random values for each extension, are independent from each other, though constant over different runs for a memoized redescription.

Table A2. Example redescriptions obtained from the data sets.

data set	algorithm	q_L	q_R	noisy J	true J
`Psycho`	SerenadeES	Do you think tobacco has hurt you	[27.0 ≤ age]	0.317	0.231
		Do you find it difficult to pass urine in the presence of others	[age ≤ 33.0]	0.343	0.408
`SSC`	SerenadeES	[3.0 ≤ I have positive feelings about the way I approach my own sexual needs and desires] ∧ [23.0 ≤ age]	[3.0 ≤ I am very aware of the sexual aspects of myself eg habits thoughts beliefs]	0.428	0.375
`RightW`	SerenadeES	[5.0 ≤ You have to admire those who challenged the law and the majority's view by protesting for women's abortion rights, for animal rights, or to abolish school prayer]	[voted = No]	0.428	0.538
`Pref`	SerenadeES	[Q5 = Write a movie script]	([Q9 = Invent a story] ∧ [country = US]) ∨ [Q2 = Be a movie critic]	0.315	0.344
`SR`	SerenadeCS	[3.0 ≤ I am happiest when I am in my bed]	[3.0 ≤ I give people handmade gifts] ∧ [14.0 ≤ age ≤ 32.0]	0.651	0.725
`MIMIC`	SerenadeCS	Atrial fibrillation	[Uric Acid = 1] ∨ [Creatine Kinase (CK) = 0]	0.424	0.227
`VAA`	SerenadeES	[Gender = M]	[CountQ21:Fb ≤ 1.0] ∨ [W5:SeniorCare ≤ 0.0]	0.637	0.445

4 Further experimental results

4.1 Results from ReReMi

Figure A1 shows the results obtained with ReReMi on different data sets. These can be compared to results shown in Figure 1(c–e).

4.2 Results with the questionnaire datasets

We also evaluated the algorithms on three data sets consisting of answers to personality tests.[1] The Woodworth Psychoneurotic Inventory (`Psycho` data set) is a personality test originally designed to measure the emotional stability and

[1] Data sets available at `https://openpsychometrics.org/_rawdata/`.

data set	n, entities	$\mathbf{D}_L$ attributes	$\mathbf{D}_R$ attributes
`Mammals`	2 575	194 Boolean	48 numerical
`MIMIC`	46 065	107 Boolean	104 Boolean
`VAA`	1 656	9 categorical	107 categorical
`SSC` *	17 685	102 numerical	—
`Pref` *	13 502	13 categorical	—
`Psycho` *	6 019	119 Boolean	—
`SR` *	4 184	53 numerical	—
`RightW` *	2 863	61 numerical	—

* These data sets consist of a single table common to both sides.

Table A3. Data set properties.

the risk for shell shock of World War I soldiers. The Wagner Preference Inventory (`Pref` data set) is used to measure the hemispheric brain dominance with questions on preferred activities. The Multidimensional Sexual Self-Concept Questionnaire (`SSC` data set) consists of questions on sexual behaviours. These three data sets have the same variables on both sides; IntialPairs and ExtendOne are implemented to ensure the same attribute cannot be used in both queries of the same redescription. Figure A2 shows the results obtained with these datasets.

4.3 Scatter plots of true vs. noisy Jaccards

Figure A3 shows scatter plots of true vs. noisy Jaccards.

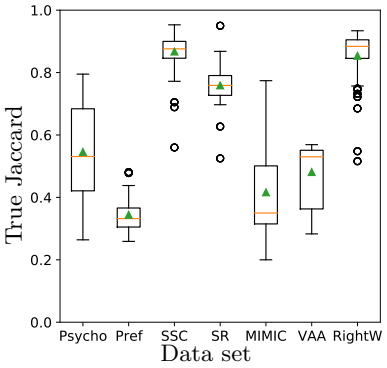

Fig. A1. Accuracy per data set with the ReReMi algorithm.

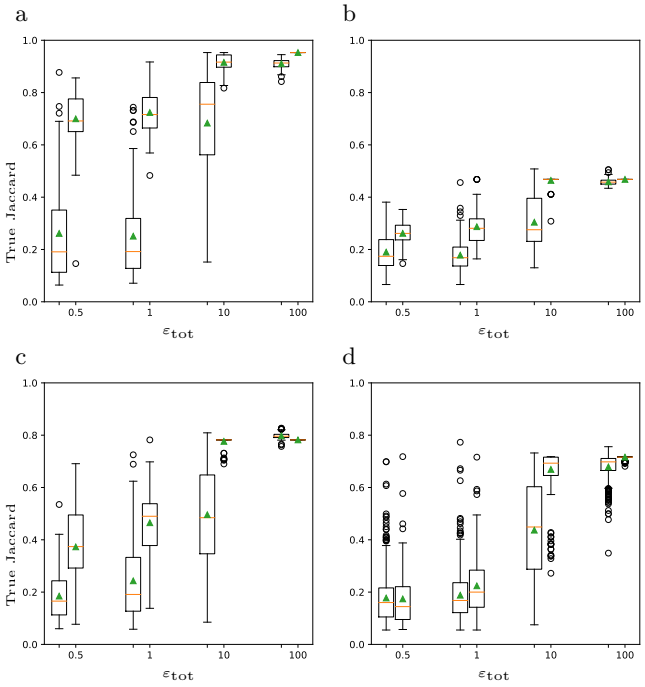

Fig. A2. Effects of the budget in terms of the true accuracy on different data sets. y-axis is always true Jaccard. x-axis: (a) Accuracy of the initial pairs, varying $\varepsilon_{\text{init}}$, with $\varepsilon_{\text{tot}} = 1$. (b)–(h) varying ε_{tot}, budget distribution d_3. This is the same as in Figure 1(a) for `Mammals`, but without NoSens. The data sets are (b) `SR`, (c) `RightW`, (d) `VAA`, (e) `SSC`, (f) `Pref`, (g) `Psycho`, and (h) `MIMIC`. Boxes in a group represent (left to right) SerenadeES and SerenadeCS. The horizontal bar and triangle within each box represent the median and the mean, respectively. Results are collected over five restarts.

4.4 Raw qualitative results

Table A2 shows some redescriptions mined from different datasets, as anecdotal examples of results one can expect by our algorithms with these kinds of data sets. A few of them are presented in Section 4.3.

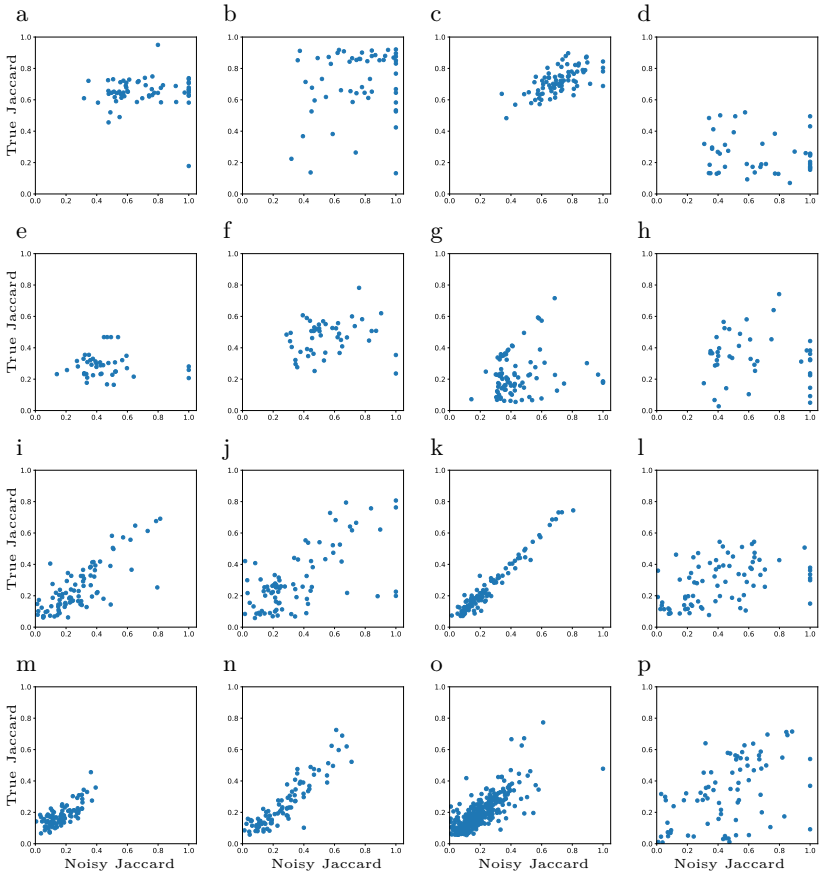

Fig. A3. Total budget 1, distribution d_3. The vertical axis represents the true Jaccard whereas the horizontal axis represents the noisy Jaccard. Top two rows: SerenadeCS. Bottom two rows: SerenadeES. (a) and (i) `SR`, (b) and (j) `RightW`, (c) and (k) `SSC`, (d) and (l) `VAA`, (e) and (m) `Pref`, (f) and (n) `Psycho`, (g) and (o) `MIMIC`, and (h) and (p) `Mammals`.