Evaluating adaptive differential privacy model

O. DZIĘGIELEWSKA

olga.dziegielewska@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Kaliskiego St. 2, 00-908 Warsaw, Poland

Differential privacy is a statistical disclosure control that is gaining popularity in recent years due to easy application for the data collection mechanisms. Many variants of differential privacy are being developed for specific use cases and environments. One of them is adaptive differential privacy that modulates the generated noise in such a way, that the retrieved result is affected according to the risk profile of the asked query and the risk-accuracy tradeoff required for the queried database. This paper intends to evaluate the adaptive differential privacy using VIOLAS Framework and through assessing how the security characteristics satisfied by the adaptive differential privacy mitigate the risk of selected inference attacks.

Keywords: differential privacy, VIOLAS framework, information score, risk-accuracy, inference attacks, statistical databases security.

DOI: 10.5604/01.3001.0015.8603

1. Adaptive differential privacy

Differential privacy is a perturbative statistical disclosure control mechanism that involves adding appropriate random noise to the statistical queries to provide a countermeasure against attacks directed at statistical answers. The differential privacy model allows a data solicitor to collect data and infer meaningful information from the data without individual record attribution, i.e., the mechanism allows a solicitor to collect sensitive data, but the data cannot be attributed to any party [7].

The application of differential privacy involves several techniques, including the injection of mathematical noise in the collect data, data hashing, subsampling and randomized data injection. Finding an appropriate tradeoff between the data accuracy and anonymity while adding a noise to the data remains a challenge as datasets users need precise data to work on and the data owners need to comply with the data privacy regulations [7].

The original work on differential privacy [2] defines the model in a following definition: A randomized function K gives $\boldsymbol{\varepsilon}$-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(K)$:

$$\Pr\left[K(D_1) \in S\right] \leq \exp\left(\varepsilon\right) \times \Pr\left[K(D_2) \in S\right]$$

A mechanism K satisfying this definition assures that even if the participant removed his data from the data set, no outputs would release any statistics that could lead to identification of sensitive statistics related with the participant [2].

The adaptive differential privacy is a model proposed in [9] extends the base differential privacy model with information score and risk-accuracy metric. The metrics allow to modulate the generated noise in such a way, that the retrieved result is affected according to the risk profile of the asked query and the risk-accuracy tradeoff required for the queried database.

The adaptive differential privacy is a multi-stage method that requires configuration and preprocessing of the data associations. Data associations, along with their risk profile measure (called *S*) are stored in a database: *AR_DB*. Historical queries are stored in *HIST_DB*. The information score metric is calculated based on the sensitivity level of the asked query calculated based on the *AR_DB* and *HIST_DB* data [9].

The configuration, called a preset, required for the next stage of the computations, include:

- *a_preset* – required accuracy level of the database;
- *r_preset* – required risk level of the database.

Risk-accuracy metric is calculated based on the configuration setup and information score metric received in the previous step [9]. In the last stage, the noise is calibrated for the noise distribution function $D(d_q)$, according to the calculated risk-accuracy metric [9].

This modification of the base model changes the overall security and accuracy of the method as the noise added to the computed queries is calibrated and set for each individual retrieved query.

2. Adaptive model security features

In case of the differential privacy mechanisms there are several factors that must be considered and conditions that must be satisfied to be fully compliant with the model, which are: group privacy, composition, and closure under post-processing. Additionally, the adaptive model satisfies independent secrecy, which is not defined in the original differential privacy definition, but was defined through the adaptive model characteristics.

Group Privacy

Group Privacy is a feature that allows to control the privacy loss incurred by groups [5]. This feature renders to extending the base model of the ε-differential privacy, which is bounded by $exp(k\varepsilon)$, where k means a change of k items between two databases.

Let l_1 norm of a database x is denoted $||x||_1$. The l_1 distance between two databases x and y is $||x-y||_1$. $||x||_1$ is a measure of the size of a database x (i.e. the number of records it contains), and $||x-y||_1$ is a measure of how many records differ between x and y [5].

Any ε – differentially private mechanism M is $(k\varepsilon) -$ differentially private for groups of size k. That is, for all $||x-y||_1 \leq k$ and all $S \subseteq Range(M)$:

$$Pr[M(x) \in S] \leq exp(k\varepsilon)\, Pr[M(y) \in S]\,,$$

where the probability space is reflecting the characteristic of the mechanism M.

For all datasets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(M)$:

$$\Pr\,[M(D_1) \in S] \leq \exp\,(\varepsilon) \times \Pr[M(D_2) \in S].$$

The adaptive model still satisfies this feature, as the distribution function will not be affected beyond an acceptable margin.

Composition

A feature strictly connected with the quantification of privacy loss is composition. The quantification of loss also permits the analysis and control of cumulative privacy loss over multiple computations. Understanding the behavior of differentially private mechanisms under composition enables the design and analysis of complex differentially private algorithms from simpler differentially private building blocks [5].

As composition and group privacy are often mistaken and linked together, it must be stressed that composition and group privacy are not the same characteristic. The composition bounds improve upon the factor k, but do not yield the same gains for group privacy [5].

The adaptive differential privacy provides stateless computations, i.e., each of the queries is calculated without the previous computations markup. However, the interim values of the information score metrics are dependent on the historical database results, therefore influencing the cumulative privacy loss for the queries. With every query call with the same parameters nonetheless, the privacy loss decreases as in such conditions added noise is altered in the distribution enhancing algorithm.

Closure under post-processing

Closure under post-processing is a feature meaning: a data analyst, without additional knowledge about the private database, cannot compute a function of the output of a differentially private algorithm M and make it less differentially private. That is, a data analyst cannot increase privacy loss, either under the formal definition or even in any intuitive sense, by analyzing outputs of the algorithm, no matter what auxiliary information is available [5].

Formally, the composition of a data-independent mapping f with an ε – differentially private algorithm M is also ε – differentially private:

Let $M : \mathbb{N}^{|X|}\, |X| \rightarrow R$ be a randomized algorithm that is (ε, δ) – differentially private. Let $f : R \rightarrow R'$ be an arbitrary randomized mapping. Then $f \circ M : \mathbb{N}^{|X|}\, |X| \rightarrow R'$ is $\varepsilon -$ differentially private [5].

For a deterministic function $f : R \rightarrow R'$ the result follows because any randomized mapping can be decomposed into a convex combination of deterministic functions, and a convex combination of differentially private mechanisms is differentially private [5]. It follows from the differential privacy definition, as ε-differential privacy composes in a straightforward way: the composition of two ε – differentially private mechanisms is also differentially private.

For any pair of neighboring databases $x,\ y$ with $\|x-y\|_1 \leq 1$, and any event $S \subseteq R'$, let $T = \{r \in R : f(r) \in S\}$.
Then:

$$Pr\big[f\big(M(x)\big) \in S\big] = Pr[M(x) \in T] \leq$$
$$\leq p(\varepsilon)Pr[M(y) \in T] =$$
$$= xp(\varepsilon)Pr[f(M(y)) \in S]$$

In case of the adaptive differential privacy algorithm, the condition still holds – since the results are modulated independently from each other, an analyst cannot infer any sensitive individual statistic without having access to the actual database. However, since the pre-processing of the metrics involve auxiliary databases, it must be assumed that an analyst does not have access to the raw dataset and auxiliary databases. The threat modelling included later in this paper covers auxiliary databases leakage and the risk related with the potential attacks at different assets is also analyzed and estimated in the risk assessment process.

Independent secrecy

Independent secrecy is a feature that is characteristic only to the adaptive model and can be defined as a feature that indicates robustness against post-processing attacks that take into consideration more than one source of information and calculations over multiple computations. In the adaptive model, even though the historical data are used to derive information score metric, the outputs generated per each query are independent from each other in terms of added noise. The algorithm is stateless, and the noise is generated specifically for each query, and collecting the data from a set of queries should not reveal any statistical characteristic that could be used to leverage a successful inference attack.

3. VIOLAS evaluation for the designed method

VIOLAS Framework [8] describes the necessary characteristic of the perfect SDC function; therefore, with use of the framework the protection mechanisms can be objectively assessed.

Every SDC method that is evaluated under the framework must be assessed separately under the same environmental conditions when being compared. The environmental security controls do not affect the score, as the framework is measuring the data quality and non-functional features rather than measuring the technical security of the implementation. However, the security of the environment itself cannot be omitted during risk assessment of the system, therefore implementation-layer security should also be reviewed in parallel while assessing the SDC method itself.

VIOLAS Framework uses the statistical liablity index (SLI) as a measure to indicate the effectiveness of the analyse SDC method. The higher the value, the better the method. The values are bounded by the system's criticality.

To calculate the SLI value four characteristics are taken into account: statistical confidentiality (sc), statistical integrity (si), statistical accuracy (sa) and statistical transparency (st). Each of those characteristics has a weighted value which is used to derive the end result. The weights ($\mathrm{w_{sc}}$, $\mathrm{w_{si}}$, $\mathrm{w_{sa}}$, $\mathrm{w_{st}}$) are being assigned by the owner of the system, e.g. in case where the most important factor is confidentiality of the statistical data, the weight for the statistical confidentiality can be increased and the other weight can be decreased.

Results of the experiment

To evaluate the adaptive differential privacy model under VIOLAS framework, a series of calculation had been made. The experiment evaluated three scenarios:

- *base differential privacy*, i.e., the originally defined model, with the typical security features characteristic for the model;
- *adaptive differential privacy*, i.e., differential privacy model extended with information score and risk-accuracy metrics, with the typical security features characteristic for the original model, but tweaked according to reflect additional features of the extended metrics;
- *arbitrary randomized noise generator function* applied over the query results, without differential privacy security features.

The outcome of the experiment shows estimated SLI values for all three scenarios.

The values describing statistical confidentiality, integrity, accuracy, and transparency remained constant throughout the experiment. For the base and adaptive differential privacy models, some additional conditions were considered as a result of the threat modelling process, therefore multiple, but constant values were used, and the results reflect actual SLI boundaries for those models under those conditions. The weights describing the importance of each of the criteria, were selected at random, but satisfying the definition of weighs for VIOLAS Framework [8].

The criticality of the system was set at a medium level, making the $SLI = 0.25$ being the upper bound for the end results.

Tab. 1. Static values of the experiment

v_{sc}	5
v_{sc_u}	10
v_{scm}	10
v_{scr}	2
v_{si}	10
v_{sim}	5
v_{sir}	10
v_{sa}	2
v_{sam}	5
v_{sam_u}	10
v_{sar}	2
v_{st}	10
v_{stm}	10
v_{str}	2
cr	4

Table 1 contains consolidated input for the constant values used to calculate SLIs, Table 2 shows variable values used for the experiment. Figure 1 and 2 contains consolidated SLI results for 50 random cases of w_{sc}, w_{si}, w_{sa}, w_{st}:

- dp is a lower bound of the SLI for base differential privacy;
- dp_u is an upper bound of the SLI for base differential privacy;
- dpm is a lower bound of the SLI for adaptive differential privacy;
- dpm_u is an upper bound of the SLI for adaptive differential privacy;
- $rand$ is an SLI value for an arbitrary randomized noise generator function.

The obtained results confirm that for almost all the tested weights, the lowest SLI values were scored by an arbitrary randomized noise generator function. The only exception was found in the 50th calculation, for which the lower bound of the adaptive differential privacy scored slightly worse than the arbitrary randomized noise generator function. In this case, the confidentiality, and the accuracy of the SDC method was marginalized, and the importance was put on the integrity of the method. Due to the characteristics of the adaptive differential privacy model, the statistical integrity of the results may be slightly impacted, in the contrast to the arbitrary randomized noise generator function, where the statistical integrity is rather consistent. However, even having that in mind, the final score difference is minimal.

When it comes to the base and adaptive differential privacy models, the base model scored better in cases where statistical integrity ruled over the statistical confidentiality and accuracy. In other cases, and in the equal distribution of weights, the adaptive model received higher SLI scores boundaries. It is however worth noting that in some cases, the scores space for the base and adaptive differential privacy models frequently overlap, i.e., SLI_{dp_u} is higher than SLI_{dpm}. This shows how sensitive the scoring framework is and it is crucial to properly assign the weights for a given system to properly validate the effectiveness of a given SDC method under the expected system's conditions and requirements.

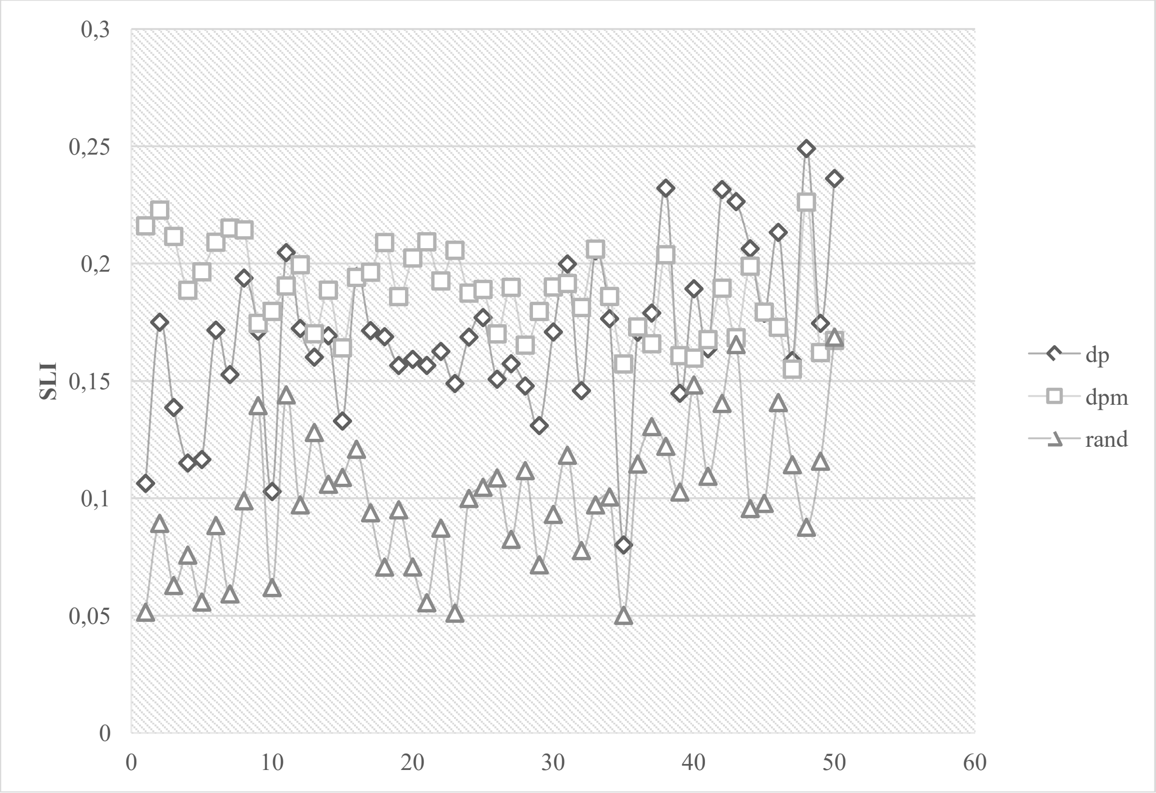

Fig. 1. Results of the experiment for lower bounds

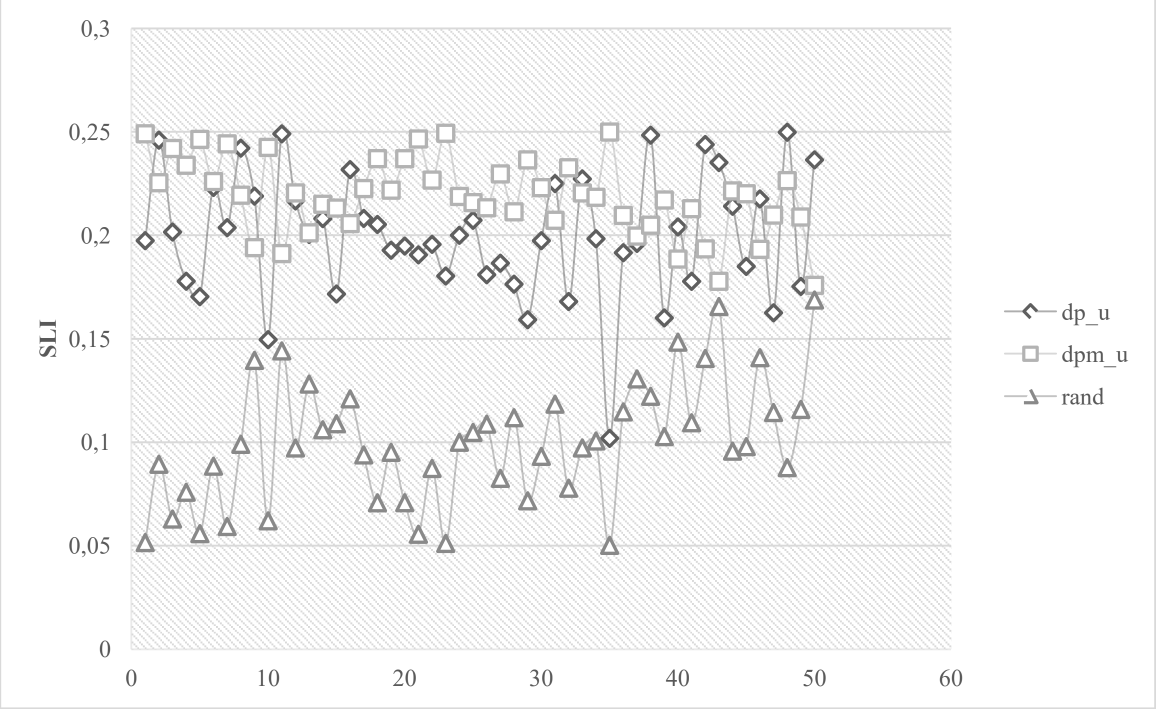

Fig. 2. Results of the experiment for upper bounds

Tab. 2. Variable values of the experiment sorted by the confidentiality weight

No.	w_{sc}	w_{si}	w_{sa}	w_{st}
1.	0,72768719	0,00772411	0,26304288	0,00154582
2.	0,56697839	0,19682362	0,02020022	0,21599777
3.	0,50248099	0,06459421	0,24222403	0,19070077
4.	0,50150999	0,12926696	0,36105916	0,00816389
5.	0,43019921	0,02926092	0,39828369	0,14225619
6.	0,41129868	0,19224656	0,13442088	0,26203387
7.	0,40718232	0,04655445	0,23161622	0,31464702
8.	0,38560219	0,24500421	0,03956339	0,32983022
9.	0,38086005	0,44775252	0,15543498	0,01595246
10.	0,37302844	0,0602368	0,50217933	0,06455543
11.	0,35412475	0,4707228	0,00499654	0,17015591
12.	0,35317876	0,23625866	0,1670999	0,24346268
13.	0,3212708	0,3905271	0,24813341	0,04006869
14.	0,31005211	0,2801789	0,20973858	0,20003041
15.	0,30912864	0,29482001	0,39189157	0,00415977
16.	0,29813374	0,35489023	0,09147491	0,25550113
17.	0,29312331	0,21942861	0,20927192	0,27817616
18.	0,29055414	0,1037781	0,2238659	0,38180185
19.	0,28894045	0,22568901	0,28617471	0,19919583
20.	0,28422166	0,10401156	0,27612164	0,33564514
21.	0,27091507	0,0278043	0,29730706	0,40397357
22.	0,26387117	0,18662955	0,27191715	0,27758213
23.	0,25080664	0,00567097	0,34837913	0,39514326
24.	0,25	0,25	0,25	0,25
25.	0,24133893	0,27373468	0,21381191	0,27111449
26.	0,2411239	0,293711	0,34529702	0,11986808
27.	0,23288824	0,16283798	0,31742887	0,28684491
28.	0,22766576	0,30928038	0,36815888	0,09489497
29.	0,22525431	0,10854561	0,45399614	0,21220394
30.	0,21158791	0,21599603	0,26310295	0,30931311
31.	0,20054099	0,34202138	0,1251635	0,33227413
32.	0,17655266	0,1389983	0,41010777	0,27434127
33.	0,1738486	0,23622124	0,11383035	0,4760998
34.	0,1737302	0,25312295	0,25839638	0,31475047
35.	0,17368734	0,0010207	0,74061508	0,08467688
36.	0,16584974	0,32329985	0,29215023	0,21870017
37.	0,13661006	0,40306139	0,26962761	0,19070094
38.	0,12926518	0,36105421	0,00816378	0,50151684
39.	0,12221164	0,26362997	0,44961922	0,16453918
40.	0,11758546	0,49180568	0,22987663	0,16073223
41.	0,11195586	0,29721153	0,36128138	0,22955124
42.	0,09785253	0,45246975	0,03077569	0,41890203
43.	0,06931337	0,57784849	0,07428275	0,27855539
44.	0,06092803	0,22847607	0,17987713	0,53071878
45.	0,04812502	0,23943005	0,32526199	0,38718295
46.	0,03337734	0,45431424	0,16226879	0,35003963
47.	0,03009932	0,32184733	0,4373183	0,21073505
48.	0,00554118	0,18870375	0,00110895	0,80464612
49.	0,00536092	0,32933173	0,37353906	0,29176829
50.	0,00081814	0,59363451	0,06787213	0,33767522

Base differential privacy values

As it was mentioned before, the retrieved SLI scores are a result of assigning static values for VIOLAS Framework measured characteristics, i.e., statistical confidentiality, integrity, accuracy, and transparency and calibrating its weights to validate the efficiency of the tested methods under different system's requirements.

To properly assign the values for the models in question, a threat modelling type of exercise has been performed to establish whether the condition is met. The results of this activity are presented in the Tables 3 and 4. It must be noted however, that the nature of the exercise was simplified and limited to identifying the most important factors affecting the VIOLAS Framework scores.

As a result of this activity, the values for the base differential privacy model characteristics were assigned as follows:

- $v_{sc} = 5$ as a lower limit for the statistical confidentiality;
- $v_{sc\_u} = 10$ as an upper limit for the statistical confidentiality;
- $v_{si} = 10$ as a value for the statistical integrity;
- $v_{sa} = 2$ as a value for the statistical accuracy;
- $v_{st} = 10$ as a value for the statistical transparency.

The statistical confidentiality in this case is ruled by the primary and the secondary data identification. As stated in the closure post-processing feature definition, the differential privacy mechanism should make it impossible to deduct any sensitive statistics based on the outcome of the asked query. What is only partially addressed by this feature are the inference attacks which use more than one source of information, e.g., database. In such scenario, two databases can be mutually exclusive, yet the metadata from both of them can serve as a potential threat in the secondary reidentification of the data.

However, since there's no negative theoretical bound, i.e., the differential privacy does not prevent from creating a mechanism under its conditions that would address this potential threat as well as even without specific considerations of this scenario in the design, there's no guarantee that the attack would succeed, two values for the statistical confidentiality were assigned to better reflect the actual SLI score.

Additional threat to the confidentiality is unauthorized access to the distribution function ($D(d_q)$) – in case $D(d_q)$ is leaked, then the confidentiality of all the retrieved statistics is affected. This threat should more likely be considered as a technical implementation-related one, therefore, when analyzing only the theoretical model, it could be potentially skipped. However, since there's a difference in the effects of such leakage for the base and adaptive differential privacy models, it was put in this analysis.

Tab. 3. Threat modelling results for base differential privacy

ID	**VIOLAS Characteristic**	**Abuse scenarios and comments**
v_{sc} v_{sc_u}	Statistical confidentiality	The characteristic is partially satisfied by the closure under post--processing feature. However, in the following cases, the statistical confidentiality, as defined in the VIOLAS Framework can be affected: • an attacker gains access to the noise distribution function $D(d_q)$; • an attacker uses more than one source of information to infer sensitive haracteristic.
v_{si}	Statistical integrity	The characteristic is fully satisfied as the integrity of the returned results is not calibrated in the base differential privacy model, i.e., for the same queries, the same results will be returned.
v_{sa}	Statistical accuracy	The characteristic is partially satisfied as the accuracy of the returned results is not calibrated in the base differential privacy model, i.e., the retrieved results will fully depend on the noise distribution function $D(d_q)$. In the following cases, the statistical accuracy, as defined in the VIOLAS Framework can be affected: • the noise distribution function $D(d_q)$ introduces too much noise over all the asked queries; • the noise distribution function $D(d_q)$ is never recalibrated overtime.
v_{st}	Statistical transparency	The characteristic is fully satisfied by differential privacy definition and the group privacy feature. In case the SDC did not satisfy this characteristic, the mechanism could not be defined as differential privacy mechanism.

The statistical integrity for the base differential privacy model will be fully satisfied as the model itself does not specify any conditions or features that would affect the integrity of the same query asked multiple times by the same or other entities. Therefore, the value for this characteristic was assigned as its maximum.

When it comes to the accuracy of the retrieved results, the base differential privacy model does not always perform well. As it was proven in multiple research [3], [4] for some applications the base approach of the differential privacy is not acceptable in terms of efficiency and statistical accuracy. The two main issues that could be causing this are static character of the noise distribution function and too excessive noise for a given business use case of the statistical database. However, for some cases the achieved accuracy can be satisfactory, therefore the characteristic value was set at 2 ($v_{sa} = 2$).

The statistical transparency is an inherent feature of the differential privacy, as its definition and also group privacy feature emphasize that the elimination of the selected data sets from a database must not reveal any metadata allowing to identify the effect of the retrieved statistics before and after the elimination from the data sets. Therefore, the value for this characteristic was set at maximum.

Adaptive differential privacy values

The values for the adaptive differential privacy model characteristics were assigned as follows:

- $v_{scm} = 10$ as a value for the statistical confidentiality;
- $v_{sim} = 5$ as a value for the statistical integrity;
- $v_{sam} = 5$ as a lower limit for the statistical accuracy;
- $v_{sam\_u} = 10$ as an upper limit for the statistical accuracy;
- $v_{st} = 10$ as a value for the statistical transparency.

The statistical confidentiality in the adaptive differential privacy model tackles the grey area of the base model. To reiterate, the base model does not fully address a potential threat in the secondary reidentification of the data in case of using multiple mutually exclusive data sources, which in conjunction can be used to conduct an inference condition. Since the adaptive model, leverages on the associations table (AR_DB), which can be build using multiple data sources and models as long as the historical table ($HIST\_DB$), which covers the historical searched in the scope of the same database, the threat resulting primary and secondary data reidentification is addressed.

When it comes to the technical implementation-level threats that can affect the statistical confidentiality, in the base differential privacy model, $D(d_q)$ leakage was considered. In the adaptive model, leaking the distribution function only, will not give full advantage and lead to a full statistical disclosure. In the adaptive model, the distribution function is recalibrated dynamically per each request, therefore, leaking only an algorithm of the noise distribution, will not fully breach the confidentiality. However, in case more characteristics are leaked at once, i.e., $D(d_q)$, the AR_DB and $HIST\_DB$ databases, a_preset and r_preset, S values, then the issue will remain open.

Additional technical threat to the confidentiality in case of the adaptive model is insufficient quality of the AR_DB and $HIST\_DB$ databases. As it was stressed before in case the association table is not created before running the scheme, then the efficiency in terms of confidentiality and accuracy of the algorithm decreases.

Tab. 4. Threat modelling results for adaptive differential privacy

ID	VIOLAS Characteristic	Abuse scenarios and comments
v_{scm}	Statistical confidentiality	The characteristic is partially satisfied by the closure under post-processing feature. Additionally, the threats resulting from the base differential privacy model are addressed, by: • Calibrating the noise distribution function $D(d_q)$); • AR_DB and $HIST\_DB$ usage. The statistical confidentiality, as defined in the VIOLAS Framework can be affected by: • Insufficient quality of the AR_DB and $HIST\_DB$ • an attacker gaining access to all of the following: the noise distribution function $D(d_q)$, the AR_DB and $HIST\_DB$ databases, a_preset and r_preset, S values.
v_{sim}	Statistical integrity	The statistical integrity can be affected as repeated queries by design will return slightly different results.
v_{sam} v_{sam_u}	Statistical accuracy (5-10)	The characteristic can be fully satisfied as the accuracy of the returned results is calibrated. However, in the following cases, the statistical accuracy, as defined in the VIOLAS Framework can be affected: • the noise distribution function $D(d_q)$ introduces too much noise over high-risk queries; • an attacker overfloods the $HIST\_DB$ with artificially generated queries, affecting the quality of the $HIST\_DB$; • a_preset is wrongly set up.

v_{stm}	Statistical transparency (10)	The characteristic is fully satisfied by differential privacy definition and the group privacy feature. In case the SDC did not satisfy this characteristic, the mechanism could not be defined as differential privacy mechanism. Additionally, the transparency guarantee is increased by the adaptive noise distribution function.

The statistical integrity for the adaptive model will not be fully satisfied as the noise distribution function is dynamically calibrated per each database query, therefore, in case the same query is made to the database multiple times, every time, the retrieved result will be slightly different. Therefore, the value for this characteristic was lowered in contrast to the base model.

The accuracy of the retrieved results however is significantly increased compared with the base model. This is again thanks to the adaptive distribution function, which alternates between lowering and increasing the noise for the queries of the different risk profiles. It is technically possible to obtain perfect accuracy for the no-risk queries, what in some cases could be understood as an exception to the differential privacy model, since the model itself is a noise addition model. However, the mechanism as a whole solution remains in the boundaries of the differential privacy. It must be noted, however, that for the high-risk queries, the accuracy can still be affected, therefore, the values proposed for this characteristic were assigned as a lower and upper bound.

As for the technical-layer threats that can affect the accuracy is wrongly set a_preset, which would affect all the retrieved results. The second threat worth mentioning is feeding the $HIST\_DB$ with excessive number of artificially generated queries which would affect the retrieved metrics scores and influence the generated noise. The second scenario would only affect a local subset, i.e., only the queries which data was added to the $HIST\_DB$ would be affected. Both of those issues can be only addressed at a procedural and technical layer.

As in the base model, the statistical transparency is an inherent feature of the differential privacy, therefore, the value for this characteristic was set at maximum. Additionally, the transparency guarantee is the adaptive model is increased by the adaptive noise distribution function.

Arbitrary randomized noise generator values

The values for an arbitrary randomized noise generator function were assigned as follows:

- $v_{scr} = 2$ as a value for the statistical confidentiality;
- $v_{sir} = 10$ as a value for the statistical integrity;
- $v_{sar} = 2$ as a value for the statistical accuracy;
- $v_{str} = 2$ as a value for the statistical transparency.

An arbitrary randomized noise generator function must be understood as function which has a noise distribution function $D(d_q)$ however does not guarantee the differential privacy features. A randomized function stripped of the differential privacy premises can only fully satisfy the statistical integrity, if it is designed in such a way that the randomization function's input is the asked query, and no additional factors are treated as a seed for this function.

The confidentiality, accuracy, and the transparency, in some cases can be achieved, however, the observation of those characteristics would be based rather on anecdotal evidence rather than a systematic and repetitive occurrence. Depending on the database access, selected static and dynamic attacks [1] could be executed to infer sensitive statistics or the noise distribution function. Therefore, the values for $v_{scr}, v_{sir}, v_{sar}, v_{str}$ were set at 2.

4. Evaluating against inference attacks

Table 5 contains a summary of the best-case coverage that base differential privacy (in the Table 5 marked as DP), and adaptive differential privacy (marked as DPM) for the inference attacks. The inference attacks covered in the analysis were:

- $S1$: small and large query sets attacks [1];
- $S2$: linear equations attacks (including tracker) [1];
- $S3$: selection attacks [1];
- $D1$: complementation attacks [1];
- $D2$: insertion attacks [1].

Additionally, selected real case attacks were covered in the analysis:

- *GIC*: medical data reidentification [11], [12];
- *NF*: Netflix data reidentification [10].

The cells marked with the delta symbol (Δ) indicate that the SDC method (DP, DPM) remains effective under a specific attack (*S*1, *S*2, *S*3, *D*1, *D*2, *GIC*, *NF*). The cells marked with the inverted delta symbol (∇) indicate that the SDC method remains partially effective under the attack conditions.

As the results show, properly designed differential privacy mechanism could be effectively used to address the risks resulting from the base inference attacks models. However, as it was proven in the previous section of this chapter, there are several gaps in the base model assumptions and features, which could be leveraged on to conduct a successful attack. One of those characteristics is the secondary data reidentification through the multiple-source inference where the sources are mutually exclusive. This risk is addressed with the adaptive differential privacy model, what effectively improves the resistance of the method against the inference attacks.

Tab. 5. Effectiveness against selected attacks

	S1	S2	S3	D1	D2	GIC	NF
DP	Δ	Δ	Δ	Δ	Δ	∇	∇
DPM	Δ	Δ	Δ	Δ	Δ	Δ	Δ

Table 6 contains which of the differential privacy features can be attributed as a characteristic addressing the attack. The group privacy (denoted in the table as GP) addresses the attacks related to observing the changes of the results based on the database size (e.g., *S*1, *D*1, *D*2). The composition (denoted in the table as COM) deals with the attacks which involve a series of queries that can be put into a system of equations which result in sensitive data reidentification (e.g. S2, S3, GIC, NF).

The Group Insurance Commission (denoted in the tables as GIC [11], [12]) and the Netflix (denoted in the tables as NF [10]) attacks would be only partially addressed by the composition feature, as the attacks were multilayered. Both of them included tracker attacks phase (*S*2), however, the key difference between the generic tracker attacks class and the GIC and Netflix attacks was additional multi-source post-processing phase. This phase cannot be fully addressed by the composition feature of the differential privacy model.

However, the adaptive differential privacy model introduces independent secrecy feature (denoted in the table as IS), which satisfies the robustness against post-processing attacks. Thanks to that, the GIC and Netflix attacks multi-source post-processing phase of the attack could be mitigated.

The closure under post processing feature (denoted as the PP in the table) provides a guarantee that the sensitive statistic cannot be inferred based on the outputs of the database, therefore it will cover most of the single-source attacks. Yet again, the feature does not fully address multi-source type of attacks, therefore GIC and NF attacks can be only partially mitigated, as they leveraged on multiple databased to create patterns that allowed to infer sensitive data.

Tab. 6. Base and *adaptive* differential privacy features effectiveness against selected attacks

	S1	S2	S3	D1	D2	GIC	NF
GP	Δ			Δ	Δ		
COM		Δ	Δ			∇	∇
IS	Δ	Δ	Δ	Δ	Δ	Δ	Δ
PP	Δ	Δ	Δ	Δ	Δ	∇	∇

5. Summary

Adaptive differential privacy changes the approach of understanding security and accuracy tradeoff for the dynamic statistical disclosure control methods as the generated noise is calibrated and custom set for each query.

Mitigation of inference attacks is the key feature of the proposed method, but thanks to its adaptive character it provides more accuracy than the base differential privacy model. However, in cases when the desirable feature of the statistical disclosure control method is statistical integrity, as defined in VIOLAS framework, the base differential privacy may be a better option.

It must be noted that SDC methods typically focus on the data layer of the system, as the statistical inference attacks are classified as business logic abuse rather than environment or implementation related attacks. The attacks leverage on the vulnerable data model design as in the inference attack scenarios it is assumed that the access to the dataset is granted by default to a certain group of system users, and the users abuse the legitimate data-level access.

However, since the security of the working environment also plays a major part in the overall security of the system, other factors, procedural and technical, must also be considered while assessing the risk of the system, but they were not covered in the scope of this paper.

6. Bibliography

[1] Denning D., *Cryptography and Data Security*, Addison-Wesley Publishing Company, Inc., USA, 1982.

[2] Dwork C., “Differential Privacy”, in: *Automata, Languages and Programming*: *33rd International Colloquium*, *ICALP 2006*, *Venice*, *Italy*, *July 10–14*, *2006*, *Proceedings*, Part II, LNCS 4052, 1–12, Springer, 2006.

[3] Dwork C., Kenthapadi K., McSherry F., Mironov I., Naor M., “Our data, ourselves: Privacy via distributed noise generation”, in: *Advances in Cryptology – EUROCRYPT 2006*, LNCS 4004, 486–503, Springer, 2006.

[4] Dwork C., McShelly F., Nissim K., Smith A., “Calibrating Noise to Sensitivity in Private Data Analysis”, in: *Theory of Cryptography*, TCC 2006, LNCS 3876, 265–284, Springer, 2006.

[5] Dwork C., Roth A., “The Algorithmic Foundations of Differential Privacy”, *Foundations and Trends in Theoretical Computer Science*, Vol. 9, Issue 3–4, 211–407 (2014).

[6] Dzięgielewska O., Szafrański B., “A brief overview of basic inference attacks and protection controls for statistical databases”, *Computer Science and Mathematical Modelling*, No. 4, 19–24 (2016).

[7] Dzięgielewska O., “Anonymization, tokenization, encryption. How to recover unrecoverable data”, *Computer Science and Mathematical Modelling*, No. 6, 9–13 (2017).

[8] Dzięgielewska O., “Evaluating Quality of Statistical Disclosure Control Methods – VIOLAS Framework”, in: *Privacy in Statistical Databases*: UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings, LNCS 12276, pp. 299–308, Springer, Cham 2020.

[9] Dzięgielewska O., “Defeating Inference Threat with Scoring Metrics”, *Proceedings of the 36th IBIMA Conference*, 4–5 November 2020, Granada, Spain, 10413–10419.

[10] Narayanan A., Shmatikov V., *How to Break Anonymity of the Netflix Prize Dataset.* arXiv:cs/0610105v2 [cs.CR], 2006.

[11] Ohm P., “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”, *UCLA Law Review*, Vol. 57, 1701–1777 (2010).

[12] Sweeney L., *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon University, Pittsburgh 2000.

Ewaluacja adaptacyjnego modelu prywatności różnicowej

O. DZIĘGIELEWSKA

Prywatność różnicowa to metoda ochrony statystycznych baz danych, która w ostatnich latach zyskuje popularność ze względu na łatwość jej zastosowania dla mechanizmów gromadzenia danych. Istnieje wiele wariantów prywatności różnicowej dla konkretnych przypadków i środowisk użycia. Jednym z wariantów jest adaptacyjna prywatność różnicowa, która moduluje generowany szum w zależności od profilu ryzyka zadanego zapytania oraz wybranego poziomu kompromisu między ryzykiem a dokładnością wyniku dla przeszukiwanej bazy danych. Artykuł ma na celu ocenę adaptacyjnej prywatności różnicowej, wykorzystując VIOLAS Framework i analizę tego, w jaki sposób charakterystyki bezpieczeństwa zapewniane przez adaptacyjną prywatność różnicową zmniejszają ryzyko wybranych ataków wnioskowaniem.

Słowa kluczowe: prywatność różnicowa, framework VIOLAS, metryka wyniku informacyjnego, metryka dokładności, ataki wnioskowaniem, bezpieczeństwo statystycznych baz danych.