On the information leakage of differentially-private mechanisms

Mário Sérgio Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, Catuscia Palamidessi

On the information leakage of differentially-private mechanisms

Mário S. Alvim*
Universidade Federal de Minas Gerais

Miguel E. Andrés
LIX, École Polytechnique

Konstantinos Chatzikokolakis
CNRS, École Polytechnique

Pierpaolo Degano
Dipartimento di Informatica, Università di Pisa

Catuscia Palamidessi
INRIA and LIX, École Polytechnique

Abstract

Differential privacy aims at protecting the privacy of participants in statistical databases. Roughly, a mechanism satisfies differential privacy if the presence or value of a single individual in the database does not significantly change the likelihood of obtaining a certain answer to any statistical query posed by a data analyst. Differentially-private mechanisms are often oblivious: first the query is processed on the database to produce a true answer, and then this answer is adequately randomized before being reported to the data analyst. Ideally, a mechanism should minimize leakage—i.e., obfuscate as much as possible the link between reported answers and individuals' data—while maximizing utility—i.e., report answers as similar as possible to the true ones. These two goals, however, are in conflict with each other, thus imposing a trade-off between privacy and utility.

In this paper we use quantitative information flow principles to analyze leakage and utility in oblivious differentially-private mechanisms. We introduce a technique that exploits graph symmetries of the adjacency relation on databases to derive bounds on the min-entropy leakage of the mechanism. We consider a notion of utility based on identity gain functions, which is closely related to min-entropy leakage, and we derive bounds for it. Finally, given some graph symmetries, we provide a mechanism that maximizes utility while preserving the required level of differential privacy.

Keywords: Differential privacy, information flow, min-entropy leakage, gain functions, optimal mechanisms.

*Corresponding author: Mário S. Alvim - Universidade Federal de Minas Gerais, Computer Science Department, Av. Antônio Carlos, 6627, ICEx/Room 4010, CEP: 31270-010 Belo Horizonte, Minas Gerais, Brazil. Phone: +55 (31) 3409-5860. E-mail: msalvim@dcc.ufmg.br.

1 Introduction

Statistical databases store data of a large number of individuals, and data analysts are allowed to pose statistical queries about these data. Typical such queries include average values, total counting, or the percentage of the entries that satisfy a given property. Statistical databases are of crucial importance in many areas. For instance, medical databases can guide pharmacological research, and census databases can help authorities decide how to spend the budget of years to come. The field of *statistical disclosure control* concerns the problem of revealing accurate statistics about a set of individuals while preserving their privacy.

In principle we would like to consider aggregate information as public, and specific information about any individual as private. However, since the two kinds of information are intrinsically linked, it is not easy to make available the former without revealing the latter. Consider, for example, a database that stores the values of the salaries of a set of individuals, and consider the queries "what is the average salary of the people in the database?" and "how many people are in the database?". Both queries are about aggregate information, but by posing them immediately before and after the addition of a new individual to the database, the data analyst can infer exactly the salary of this individual.

Another important issue is the presence of *side information*, which is any information about individuals coming from sources external to the database itself (e.g., from prior beliefs, public sources, newspapers, or other databases). The combination of statistical queries and suitable side information can pose serious threats to the privacy of individuals [12].[1]

To tackle the problem of statistical disclosure control, Dwork has proposed the notion of *differential privacy* [12, 13, 14, 15], which has received great attention in the privacy community. Essentially, differential privacy ensures that the presence or absence of any individual in a database, or changing the data of any individual, does not significantly affect the probability of obtaining any specific answer for a certain query. Intuitively, this implies that it is "safe" for an individual to opt in (or out) a database, since their choice will not significantly affect the information the data analyst will obtain. An important feature of differential privacy is that it does not depend on side information.

There are several approaches in the literature to implement differentially-private mechanisms. For numeric queries, Dwork has proposed the simple method of adding *Laplacian* noise to the true answer to the query [12]. More sophisticated mechanisms correlate noise between queries, enabling more information to be extracted from the database while still preserving differential privacy [7, 25, 18]. For cases in which perturbing the answer is not an adequate option (e.g. for non-numeric queries), McSherry and Talwar have proposed the

[1] The combination of answers to statistical queries with side information can affect even the privacy of individuals not present in the database. For example, if one knows that a certain person's salary is exactly the same as the average salary of the people in the database, then the average-salary query will reveal the salary of this person, independently from whether they are in the database or not.

exponential mechanism [24].

Differential privacy relies on the randomization of the query answer, and thus imposes a trade-off in fulfilling two opposing goals. On one hand, *privacy* demands the minimization of the amount of information about the database (and, in particular, about individuals) revealed through the randomized answers. On the other hand, the mechanism should provide good *utility*, i.e., some adequate closeness between true and randomized answers.

In this paper we investigate the quantification of privacy and utility in differentially-private mechanisms, approaching the problem from the perspective of the well-established field of quantitative information flow.

Quantitative information flow. Completely secure systems are often impossible to obtain in practice, either by design or by technological constraints, and thus it is important to quantitatively analyze the leakage of such systems. The field of *quantitative information flow* is concerned with the amount of secret information one can infer from the observable behavior of an execution of a system.

Information theory is widely regarded as a natural framework to provide firm foundations to quantitative information flow. A system can be seen as an information-theoretic channel from secret inputs to observable outputs. The input has some *uncertainty* representing how easily an adversary can discover the secret. Uncertainty is evaluated through *entropy measures*, which vary according to the model of adversary and to the way one estimates the success of an attack [23]. One of the main notions of entropy in quantitative information flow is *min-entropy* [26, 9, 27], which is closely related to the risk of an adversary guessing the secret correctly in one try.

Independently of the adopted model of adversary and attack, a general principle of quantitative information flow is that leakage can be expressed as the difference between the *initial uncertainty* about the secret before the system is executed, and the *remaining uncertainty* after the execution is observed:

$$\textit{information leakage} = \textit{initial uncertainty} - \textit{remaining uncertainty}.$$

The observation of the execution is expected to increase the probabilistic knowledge about the secret, therefore decreasing the corresponding uncertainty and making the equation above non-negative.

Quantitative information flow vs. differential privacy. Both quantitative information flow and differential privacy measure of protection of sensitive information provided by a mechanism, and in both cases this measure is linked to the probabilistic knowledge that an adversary gains about the secret from the outcome of the mechanism. It is therefore natural to ask how they compare, and whether one is preferable to the other. In our opinion, there is no absolute criterion to prefer one to the other: it depends on the context in which we want to use a privacy mechanism. One main difference between the two notions is that quantitative information flow is an average measure, defined in terms of

the expected value of the probability of a successful attach with respect to all possible outcomes, while differential privacy is concerned with the worst-case, no matter how improbable it may be. Thus differential privacy reflects the perspective of an individual, for whom the consequences of a privacy breach may be dramatic, while quantitative information flow could be more suitable for companies, as they can usually amortize the costs of single incidents, and are more interested in optimizing the trade-off between average costs rather than avoiding all risks.

Another important difference between quantitative information flow and privacy is that the latter has a nice compositionality property: the combination of an ϵ_1-differentially private and an ϵ_2-differentially private mechanism gives an $(\epsilon_1 + \epsilon_2)$-differentially private mechanism. Hence the degradation of privacy under composition is controllable. In quantitative information flow there has been some work on the cascading of channels [16], which however is a rather specific form of composition, and rather different from the one of differential privacy. More general kinds of composition have been analyzed in [22] (including the one of differential privacy, under the name of "parallel composition with repeated input"), and in the context of a process calculus [8], but a neat compositionality result like the one of differential privacy has not been produced.

Contributions. This work investigates connections between differential privacy and quantitative information flow. We note that differentially-private mechanisms can be seen as information-theoretic channels from databases to reported answers, and we describe—and quantify—the level of privacy and utility of such mechanisms in terms of entropy min-entropy and leakage. We are motivated by the following fundamental questions.

1. Does ϵ-differential privacy induce a bound on the information leakage of a mechanism?
2. Does ϵ-differential privacy induce a bound on the utility of a mechanism?
3. Given a query and a value $\epsilon \geq 0$, can we construct a mechanism satisfying ϵ-differential privacy and also providing maximum utility?

To address those questions, we exploit the graph structure of the database domain to derive bounds on leakage, and, similarly, the graph structure that a query induces on the domain of true answers to derive bounds on utility. The main contributions of this paper are the following.

- We propose an information-theoretic framework to reason about information leakage and utility in differentially-private querying mechanisms.
- We explore graph-theoretic properties of the adjacency relation on databases, using two types of symmetries (distance-regularity and vertex-transitivity) that enable us to prove that differential privacy induces a bound on min-entropy leakage.

- Furthermore, we prove that this bound is tight, i.e., that there always exists a differentially-private mechanism that attains this bound.
- We prove that if the graph structure of the answers satisfies our symmetry conditions, then differential privacy induces a bound on utility, measured in terms of identity gain functions. We also prove that this bound is tight.
- As a side result, we prove that under the considered symmetry conditions the exponential mechanism is optimal, i.e., it provides maximum utility for a given degree of privacy.

Finally, we show that the opposite direction of the relation between quantitative information flow and differential privacy is not possible to establish in general. More precisely, a bound on the min-entropy leakage of a channel does not necessarily imply a bound on the level of differential privacy of the channel (i.e., on the parameter ϵ).

The remainder of this paper is organized as follows. In Section 2 we review basic concepts of information theory, quantitative information flow, differential privacy, and graph theory. In Section 3 we introduce a model for leakage and utility in oblivious differentially-private mechanisms. In Section 4 we exploit graph-symmetries (distance-regularity or vertex-transitivity) of the adjacency relation on the input to a channel to derive bounds on the a posterior min-entropy of that channel. We then apply this bound to differentially-private mechanisms to derive our results for leakage in Section 5 and for utility in Section 6. Finally, in Section 7 we discuss related work, and in Section 8 we conclude. Full proofs for all our results can be found in Appendix A.

A preliminary version of some of the results of this paper appeared in [1] and [2].

2 Preliminaries

2.1 The information-theoretic framework for quantitative information flow

Let A, B denote two discrete random variables with carriers $\mathcal{A} = \{a_0, \ldots, a_{n-1}\}$, $\mathcal{B} = \{b_0, \ldots, b_{m-1}\}$, and let π denote a probability distribution on A, called a *prior distribution*.

An *(information-theoretic) channel* is a triple $(\mathcal{A}, \mathcal{B}, M)$, where $\mathcal{A}$ is the *channel input*, $\mathcal{B}$ is the *channel output*, and M is a *channel matrix* of conditional probabilities. Each element $M_{a,b}$ represents the probability that B takes value b given that A has value a. Together, π and M induce a joint probability distribution p for A, B, defined as $p(a, b) = \pi_a M_{a,b}$. Note that p satisfies $p(a) = \pi_a$ and, for values of $\pi(a) > 0$, $p(b|a) = M_{a,b}$.

There are several measures of the information shared by A and B via the channel matrix. In this paper we will concentrate on the measures of vulnerability and min-entropy.

The *vulnerability* of A is defined as $V(A) = \max_{a \in \mathcal{A}} p(a)$, and it represents the probability that an adversary can correctly guess the value of A in a single try (a rational adversary chooses as a guess a value of a with maximum probability). Correspondingly, the *conditional vulnerability* of A given B is defined as $V(A|B) = \sum_{b \in \mathcal{B}} p(b) V(A|B = b) = \sum_{b \in \mathcal{B}} \max_{a \in \mathcal{A}} p(a)p(b|a)$, and it is the probability that the adversary can correctly guess the value of A in one try after having observed the value of B. It can be shown that ${}^{V(A|B)}/_{V(A)} \geq 1$, and intuitively this ratio represents by how much the adversary's probability of success is increased by the observation of the channel output.

For mathematical convenience, the vulnerability is usually converted into bits by taking its negative logarithm (in base 2). The *min-entropy* of A is then defined as $H_\infty(A) = -\log V(A)$, and the *conditional min-entropy* of A given B is defined as $H_\infty(A|B) = -\log V(A|B)$.

The *min-entropy leakage* (or simply *min-leakage*) of A to B is defined as the difference between the *a priori* (i.e., before observing the value of B) and the *a posteriori* (i.e., after observing the value of B) min-entropies of A: $I_\infty(A;B) = H_\infty(A) - H_\infty(A|B) = \log {}^{V(A|B)}/_{V(A)}$. It can be shown that $0 \leq I_\infty(A;B) \leq H_\infty(A)$, and that min-entropy leakage is not symmetric, i.e., that $I_\infty(A;B) \neq I_\infty(B;A)$ in general.

Min-capacity is the worst-case leakage over all input distributions: $C_\infty = \max_\pi I_\infty(A;B)$. It has been proven in [9] that C_∞ is realized at the uniform distribution, and that it equals the logarithm of the sum of the maxima of each column in the channel matrix, i.e., $C_\infty = \log \sum_{b \in \mathcal{B}} \max_{a \in \mathcal{A}} p(b|a)$.

2.2 Differential Privacy

Let $\mathcal{X}$ be the set of all possible databases. Two databases $x, x' \in \mathcal{X}$ are *adjacent* (or *neighbors*), written $x \sim x'$, if they differ for the presence or the value of exactly one individual. We call $\sim$ the *adjacency relation* on databases. Note that the structure $(\mathcal{X}, \sim)$ forms an undirected graph, where vertices are databases and edges connect every two adjacent databases.

A *(differentially-private) mechanism* is a probabilistic function $\mathcal{K}$ from $\mathcal{X}$ to some set of possible answers $\mathcal{Z}$, satisfying the property that the ratio between the probabilities of two adjacent databases to give a certain answer is bounded by e^ϵ, for some $\epsilon \geq 0$.

Definition 1 ([14])**.** *A mechanism $\mathcal{K}$ from $\mathcal{X}$ to $\mathcal{Z}$ satisfies* ϵ-differential privacy*, for some $\epsilon \geq 0$, if for all pairs $x, x' \in \mathcal{X}$, with $x \sim x'$, and all $S \subseteq \mathcal{Z}$:*

$$Pr[\mathcal{K}(x) \in S] \leq e^\epsilon \cdot Pr[\mathcal{K}(x') \in S]. \tag{1}$$

Since in this work we consider finite $\mathcal{X}, \mathcal{Z}$, all probability distributions are discrete, and in the above definition it is sufficient to consider probabilities of the form $Pr[\mathcal{K}(x) = z]$.

Intuitively, (1) means that an isolated individual has a negligible influence on a large database. It is usual to think of ϵ as a constant smaller than 1, so $e^\epsilon \approx (1 + \epsilon)$, and differential privacy ensures a small multiplicative difference in the distributions generated by neighbor databases.

2.3 Graph theory

Let $G = (\mathcal{V}, \sim)$ be a (undirected) graph with vertices in $\mathcal{V}$ and edges $\sim \subseteq \mathcal{V} \times \mathcal{V}$. We use the infix notation to denote that two elements v and w of $\mathcal{V}$ are connected by an edge, i.e., $v \sim w$ stands for $(v, w) \in \sim$. We also say, in this case, that v and w are *adjacent* or *neighbors*. The *distance* $d(v, w)$ between two connected vertices $v, w \in \mathcal{V}$ is the number of edges in a shortest path connecting them. We denote by $\mathcal{V}_{\langle d \rangle}(v)$ the subset of vertices in $\mathcal{V}$ that are at distance d from the vertex v. The *diameter* $d_{max}(G)$ of G is the maximum distance between any two connected vertices in $\mathcal{V}$, i.e., $d_{max}(G) = \max_{v,w \in \mathcal{V}} d(v, w)$. We denote by Δ_G the set $\{0, 1, \ldots, d_{max}\}$. When no confusion can arise, we use d_{max} and Δ instead of $d_{max}(G)$ and Δ_G, respectively. The *degree* of a vertex is the number of edges incident to it. G is called *regular* if every vertex has the same degree, and *k-regular* if all vertices have degree k. An *automorphism* of G is a permutation σ on the vertex set $\mathcal{V}$ such that for any pair of vertices v, w, if $v \sim w$, then $\sigma(v) \sim \sigma(w)$. The *automorphism group* of G is the set of all of its automorphisms. We denote by σ^k the composition of σ with itself k times, i.e., $\sigma^k(v) = \sigma(\sigma^{k-1}(v))$ for $k > 0$, and $\sigma^0(v) = v$.

For the purpose of this paper, we need to consider graphs with certain symmetry properties, namely the so-called *distance-regular graphs*, and the *vertex-transitive graphs*.

Definition 2 (Distance-regular graph [10])**.** *A graph $G = (\mathcal{V}, \sim)$ is* distance-regular *if there exist integers b_d and c_d ($d \in \{0, \ldots, d_{max}\}$) (called* intersection numbers*) such that, for all vertices v, w at distance $d(v, w) = d$, there are exactly b_d neighbors of w in $\mathcal{V}_{\langle d+1 \rangle}(v)$, and c_d neighbors of w in $\mathcal{V}_{\langle d-1 \rangle}(v)$.*

The following proposition gives an alternative characterization of distance-regular graphs:

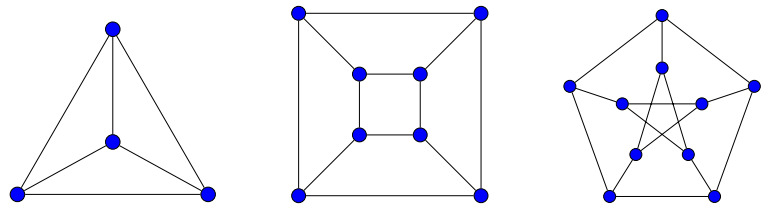
Proposition 3 ([10])**.** *A graph is distance-regular if, and only if, it is regular and, for any two vertices v and w, the cardinality of $\mathcal{V}_{\langle j \rangle}(v) \cap \mathcal{V}_{\langle k \rangle}(w)$ depends only on j, k, and $d(v, w)$.*

We now define vertex-transitive graphs.

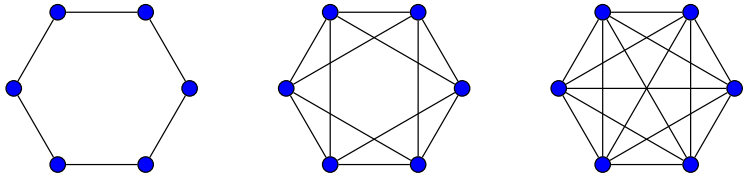
Definition 4 (Vertex-transitive graph)**.** *A graph $G = (\mathcal{V}, \sim)$ is called* vertex-transitive *if for any pair $v, w \in \mathcal{V}$ there exists an automorphism σ such that $\sigma(v) = w$.*

Many regular graphs are both distance-regular and vertex-transitive. This includes, in particular, the *Hamming graphs*, which play an important role in this paper since they are the natural structure of databases. A Hamming graph is characterized by two natural numbers, u and v, and it is defined as the graph whose set of nodes are tuples of u elements, where each element can take v different values, and with an adjacency relation defined by stipulating that two tuples are adjacent if they differ in the value of exactly one element.

Some examples of graphs that are both distance-regular and vertex-transitive are given in Figure 1. It is possible to show that Figures 1(a) and 1(b) are (isomorphic to) Hamming graphs, with (u, v) equal to $(1, 4)$ and $(3, 2)$, respectively.

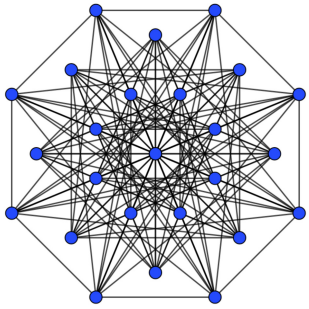


(a) Tetrahedral graph (b) Cubical graph (c) Petersen graph

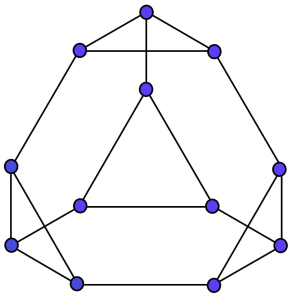


(d) A cycle (e) A cage (f) A clique

Figure 1: Some distance-regular and vertex-transitive graphs



(a) One of the Chang graphs, which are distance-regular and not vertex-transitive



(b) The truncated tetrahedron, which is vertex-transitive and not distance-regular

Figure 2: Examples of graphs distinguishing the properties of distance-regularity and vertex-trasitivity

Distance-regularity and vertex-transitivity, however, are incomparable properties, in the sense that neither of them implies the other. Figure 2(a) shows one of the Chang graphs, which is distance-regular and not vertex-transitive, whereas Figure 2(b) shows the truncated tetrahedron, which is an example of a graph that is vertex-transitive and not distance-regular.

Next, we show that distance-regular and vertex-transitive graphs have the property that, for any given distance d, the cardinality of $\mathcal{V}_{\langle d \rangle}(v)$ is the same for all vertices v, i.e., it depends only on d and not on v. This property will be relevant for some of the proofs in the paper.

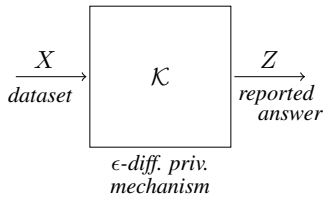


Figure 3: Mechanism $\mathcal{K}$

Proposition 5. *If a graph $G = (\mathcal{V}, \sim)$ is distance-regular or vertex-transitive, then, for every distance $d \in \Delta$, there exists a constant n_d which is equal to the cardinality of $\mathcal{V}_{\langle d \rangle}(v)$ for every $v \in \mathcal{V}$.*

3 A model of utility and privacy for statistical databases

Let $\mathcal{U} = \{0, 1, \ldots, u-1\}$ be a finite set of cardinality u representing the individuals participating in the database, and let $\mathcal{V} = \{v_0, v_1, \ldots, v_{v-1}\}$ represent the v different possible values for the sensitive attribute of each individual (e.g., disease name in a medical database). The case of multiple sensitive attributes can be modeled simply by considering $\mathcal{V}$ as a set of tuples. The absence of an individual from the database is modeled by a special value in $\mathcal{V}$. A database $x = \mathrm{x}_0 \ldots \mathrm{x}_{u-1}$ is a u-tuple where each $\mathrm{x}_i \in \mathcal{V}$ is the value of the individual i. The set of all databases is, therefore, $\mathcal{X} = \mathcal{V}^u$. As recalled in previous section, there is a natural graph structure associated to the set of databases, called Hamming graph. The vertices of this graph are the databases themselves, and the adjacency relation is defined by $x \sim x'$ if, and only if, x and x' differ in the value of exactly one individual.

A query on the database is a (deterministic) function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is the domain of the true answers $y = f(x)$ to the query. A (noisy) mechanism for f is a probabilistic mapping $\mathcal{K} : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ represents the domain of the answers reported by the mechanism, and does not necessarily coincide with $\mathcal{Y}$ (cf. Figure 3). We model such a mechanism as a channel $(\mathcal{X}, \mathcal{Z}, M)$, where $\mathcal{X}, \mathcal{Z}$ are the channel's inputs and outputs, respectively, and M is the channel matrix. The definition of differential privacy is directly expressed as a property of the channel: M satisfies ϵ-differential privacy if, and only if,

$$M_{x,z} \leq e^{\epsilon} M_{x',z} \qquad \text{for all } x, x' \in \mathcal{X} \text{ s.t. } x \sim x', \text{ and all } z \in \mathcal{Z}.$$

Let Y be a random variable ranging over $\mathcal{Y}$ and modeling the true answer to f. The mechanism $\mathcal{K}$ is often *oblivious*, meaning that the reported answer Z only depends on the true answer Y and not on the database X. The channel corresponding to an oblivious mechanism $\mathcal{K}$ can thus be decomposed into a channel from X to Y modeling the query f and a noisy channel $\mathcal{H}$ from Y to Z

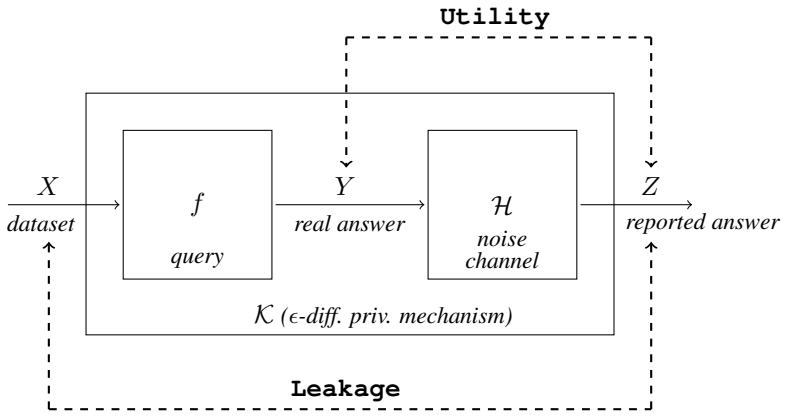


Figure 4: Leakage and utility for oblivious mechanisms

modeling the randomization of the true answer. These two channels are said to be *in cascade*, since the output of the first one is the input for the second one, as depicted in Figure 4.

The *leakage* of the channel associated to the mechanism is a measure of the information about the database that the adversary can obtain by observing the reported answer, hence it represents a relation between X and Z. On the other hand, the *utility* of the mechanism is a measure of how much one can learn about the true answer from the reported one, hence it represents a relation between Y and Z. Note that in an oblivious mechanism the utility depends only of the noise channel $\mathcal{H}$.

Following the standard approach in the differential privacy literature [17, 21], we represent the adversary's side information as a prior distribution on X.

4 Relating differential privacy and quantitative information flow

In this section we propose a general technique to derive bounds on the min-entropy leakage of ϵ-differentially-private channels. The bounds are derived for any channel whose input respects certain graph regularities. Note that in previous section we have formulated the notion of differential privacy in terms of a generic graph structure on the input, hence we do not need to assume that the inputs are databases. In Section 5 we will instantiate the technique to the channel $\mathcal{K}$, thus obtaining results about leakage, while in Section 6 we will instantiate it to the channel $\mathcal{H}$, thus obtaining results about utility.

More specifically, given a channel respecting ϵ-differential privacy, we show that if the graph structure of the channel input is distance-regular or vertex-transitive, it is possible to transform the channel matrix into a leakage-equivalent matrix with certain structural regularities that immediately allow to derive

bounds on the a posteriori min-entropy of the channel.

We emphasize now an important assumption of our analysis, and its implications.

Remark 6. *The bounds in this section are obtained under the assumption that the a priori distribution on the channel's input is uniform. This is not a restriction for when applying these bounds to leakage, since maximum min-entropy leakage is achieved on the uniform input distribution (as seen in Section 2) and, hence, any bound for the uniform distribution is also a bound for every other input distribution. In the case of utility the assumption of a uniform input distribution is more restrictive, but we will see that it still provides interesting results for several practical cases.*

4.1 Assumptions and notation

We start by setting some assumptions and notation.

Channels. We consider channels having input A and output B, with finite carriers $\mathcal{A} = \{a_0, \ldots, a_{n-1}\}$ and $\mathcal{B} = \{b_0, \ldots, b_{m-1}\}$, respectively. The probability distribution of A is assumed to be uniform. Furthermore, we assume that $|\mathcal{A}| = n \leq |\mathcal{B}| = m$. (where $|\cdot|$ represents the cardinality of a set). The latter assumption is without loss of generality, because if $n > m$ we can add to the matrix enough all-zero columns (i.e., columns containing only 0's), so as to match the number of rows. Note that adding all-zero columns does not change the min-entropy leakage of the channel.

We assume an adjacency relation $\sim$ on $\mathcal{A}$, so that the structure $(\mathcal{A}, \sim)$ is a (undirected) graph.

For simplicity, we will often (and especially in the proofs) represent the elements of $\mathcal{A}$ and $\mathcal{B}$ by their indexes, i.e., we will use the natural numbers i, j to denote a_i, b_j, respectively. We will also write $i \sim h$ to mean $a_i \sim a_h$, and $d(i, h)$ to denote $d(a_i, a_h)$, i.e., the distance between a_i and a_h in the graph structure.

We use M, M', M'' or N to range over channels. Since we represent the elements of $\mathcal{A}$ (resp. $\mathcal{B}$) by their indexes, typically the rows and the columns of the channels will be indexed by natural numbers from 0 to $n-1$ (resp. $m-1$). We typically use i, h, l to range over rows, and j, k to range over columns. Given a matrix M, we denote by $\max_j^M$ the maximum value of column j over all rows i, i.e., $\max_j^M = \max_i M_{i,j}$, and by $\max^M = \max_{i,j} M_{i,j}$ the maximum element of the matrix.

Recall that a channel matrix M satisfies ϵ-differential privacy if, for each column j and for each pair of rows i and h such that $i \sim h$, we have

$$\frac{1}{e^\epsilon} \leq \frac{M_{i,j}}{M_{h,j}} \leq e^\epsilon.$$

When necessary to avoid confusion, we annotate the a posteriori min-entropy and the min-entropy leakage with the channel M they refer to. I.e., we use the

notation $H_\infty^M(A|B)$ and $I_\infty^M(A;B)$, respectively.

4.2 The matrix transformation

Considering a channel matrix M having at least as many columns as rows, and assuming an uniform input distribution, the transformation on the channel matrix is divided into two steps. First, M is converted into a matrix M' in which each of the first n columns has a maximum in the diagonal, and the remaining columns are all 0's.

Then, we note that all-zero columns do not contribute to the a posteriori min-entropy leakage of the channel, nor to its min-entropy leakage, and, for simplicity, we erase them from M'. The result is a square matrix of dimension $n \times n$.

Second, under the assumption that the input domain is distance-regular or vertex-transitive, M' is converted into a matrix M'' which has the same maximum leakage, is still ϵ-differentially private, and has all the elements of the diagonal equal to its maximum element $\max^{M''}$.

A scheme of the transformation is shown in Figure 5, where Lemma 7 (*Step* 1) is the first step of our transformation, and the second is step either Lemma 8 (*Step* 2*a*) or Lemma 9 (*Step* 2*b*), depending on whether the graph structure is distance-regular or vertex-transitive, respectively.

Lemma 7. (Step 1) *Let M be a channel matrix of dimensions $n \times m$ s.t. $n \leq m$, and assume that M satisfies ϵ-differential privacy. Then it is possible to transform M into a matrix M' of the same dimensions satisfying the following conditions:*

(i) M' is a channel matrix: $\sum_{j=0}^{m-1} M'_{i,j} = 1$, for all $0 \leq i \leq n-1$;

(ii) each of the first n columns has a maximum in the diagonal: $M'_{i,i} = \max_i^{M'}$, for all $0 \leq i \leq n-1$;

(iii) the $m-n$ last columns only contain 0's: $M'_{i,j} = 0$, for all $0 \leq i \leq n-1$ and all $n \leq j \leq m-1$;

(iv) M' satisfies ϵ-differential privacy: $M'_{i,j} \leq e^\epsilon M'_{h,j}$, for all $0 \leq i, h \leq n-1$ s.t. $i \sim h$ and all $0 \leq j \leq m-1$; and

(v) $H_\infty^{M'}(A|B) = H_\infty^M(A|B)$, if A has the uniform distribution.

Lemma 8. (Step 2*a*) *Let M' be a square channel matrix of dimensions $n \times n$ that satisfies ϵ-differential privacy. Let $\sim$ be an adjacency relation on $\mathcal{A}$ such that the graph $(\mathcal{A}, \sim)$ is connected and distance-regular. Assume that the maximum value of each column is on the diagonal, i.e., $M_{i,i} = \max_i^M$ for all $0 \leq i \leq n-1$. Then it is possible to transform M' into a matrix M'' of the same dimension satisfying the following conditions:*

(i) M'' is a channel matrix: $\sum_{j=0}^{n-1} M''_{i,j} = 1$, for all $0 \leq i \leq n-1$;

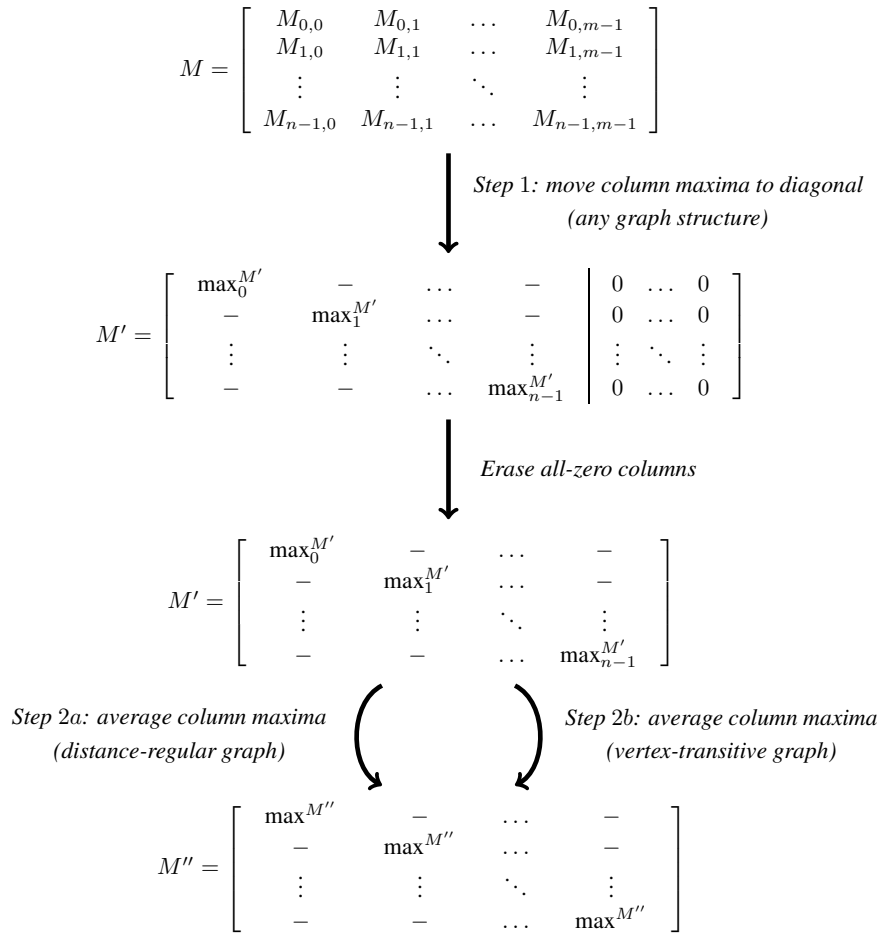


Figure 5: Matrix transformation for distance-regular and vertex-transitive graphs

(ii) the elements of the diagonal are all the same, and are equal to the maximum of the matrix: $M''_{i,i} = \max^{M''}$, *for all* $0 \leq i \leq n-1$;

(iii) M'' *satisfies* ϵ*-differential privacy:* $M'_{i,j} \leq e^{\epsilon} M'_{h,j}$, *for all* $0 \leq i, h, j \leq n-1$ *s.t.* $i \sim h$; *and*

(iv) $H_\infty^{M''}(A|B) = H_\infty^{M'}(A|B)$, *if* A *has the uniform distribution.*

Lemma 9. (Step 2*b*) *Consider a square channel matrix* M' *satisfying the assumptions of Lemma 8, except that we assume* $(\mathcal{A}, \sim)$ *to be vertex-transitive instead of distance-regular. Then it is possible to transform* M' *into a matrix* M'' *with the same properties as in Lemma 8.*

4.3 The bound on the a posteriori entropy of the channel

Once the transformation has been applied, and the channel matrix respects the properties of M'' as in Figure 5, we use the graph structure of $(\mathcal{A}, \sim)$ to determine a bound on the a posteriori entropy $H_\infty^{M''}(A|B)$. Since the matrix transformation preserves its a posteriori conditional min-entropy, the bound obtained is also valid for the original channel matrix M.

It is a known result in the literature (cfr. [9]) that when the distribution of A is uniform, the a posteriori min-entropy of a channel M'' is given by

$$H_\infty^{M''}(A|B) = -\log \frac{1}{n} \sum_{j \in \mathcal{B}} \max_j^{M''}. \tag{2}$$

Given that the diagonal elements of the matrix M'' are all equal to the maximum $\max^{M''}$, (2) becomes

$$H_\infty^{M''}(A|B) = -\log \max^{M''}, \tag{3}$$

and finding a bound on the a posteriori entropy of the channel M'' reduces to finding a bound on $\max^{M''}$.

We proceed by noting that ϵ-differential privacy induces a relation between the ratio of elements at any distance (rather than only for neighbor elements at distance 1).

Remark 10. *Let M be a matrix satisfying ϵ-differential privacy. Then, for any column j, and any pair of rows i and h we have that:*

$$\frac{1}{e^{\epsilon\, d(i,h)}} \leq \frac{M_{i,j}}{M_{h,j}} \leq e^{\epsilon\, d(i,h)}. \tag{4}$$

In particular, by taking $h = j$ in (4), and since the elements on the diagonal of M'' are equal to the maximum, we obtain that, for each element $M''_{i,j}$,

$$\max^{M''} \leq e^{\epsilon\, d(i,j)} M''_{i,j}. \tag{5}$$

Intuitively, (5) means that the value of $\max^{M''}$ cannot be increased by arbitrarily grabbing probability mass from other elements in the same line in M'', since ϵ-differential privacy imposes a maximum ratio between any two elements in the same column of the matrix. This observation motivates the next proposition. (Recall that $\mathcal{A}_{\langle d \rangle}(i)$ is the set of nodes $j \in \mathcal{A}$ at distance d from $i \in \mathcal{A}$, and we use $|\cdot|$ to denote the cardinality of a set. Moreover, Δ is the set of all possible distances between elements of $\mathcal{A}$.)

Proposition 11. *Let M be a channel matrix satisfying ϵ-differential privacy where for every $0 \leq i \leq n-1$ we have $M_{i,i} = max^M$. Then, for every node $i \in \mathcal{A}$,*

$$\max^M \leq \frac{1}{\sum_{d \in \Delta} \frac{|\mathcal{A}_{\langle d \rangle}(i)|}{e^{\epsilon d}}},$$

Proposition 11 can be made more precise when the graph structure of the channel input is vertex-transitive or distance-regular. By Proposition 5, in such graphs, for every distance $d \in \Delta$, the value of $|\mathcal{A}_{\langle d \rangle}(i)|$ is the same for every element $i \in \mathcal{A}$, it and depends only on d. We recall that such value is denoted by n_d.

Putting together the steps above, we obtain the main result in this section.

Theorem 12. *Consider a channel matrix M satisfying ϵ-differential privacy for some $\epsilon \geq 0$, assume that the probability distribution on A is uniform, and that $(\mathcal{A}, \sim)$ is either distance-regular or vertex-transitive. Then*

$$H_\infty^M(A|B) \geq -\log \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}}. \tag{6}$$

In case the matrix M can be transformed via Lemmata 7, 8, and 9 into a matrix M'' whose elements all satisfy (5) with equality, then the bound (6) holds with equality.

5 Application to leakage

We measure the leakage of a differentially-private mechanism as the the min-entropy leakage of the channel from X (databases) to Z (reported answers) of Figure 4. In this section we show that the graph structure $(\mathcal{X}, \sim)$ of the database domain presents the required symmetries (distance-regularity or vertex-transitivity) for the matrix transformation from the previous section, so we can instantiate our bound for the posterior min-entropy to the particular channel from X to Z.

As emphasized in Remark 6, min-entropy leakage is maximum when the input distribution is uniform, so the bounds derived in this section (Theorem 15, Proposition 20, and Proposition 22) are valid for all distributions on inputs. Moreover, since we model side information as input distributions, it follows that these bounds are valid for any side information the adversary may have.

5.1 Graph symmetries of the database domain

We recall that the graph structure of the database domain $(\mathcal{X}, \sim)$ is a Hamming graph. Since Hamming graphs present the symmetries of distance-regularity and vertex-transitivity (cfr. [20, 10]), and so does $(\mathcal{X}, \sim)$.

Proposition 13. *If $v \geq 2$, the graph $(\mathcal{V}^u, \sim)$ is a connected distance-regular graph with diameter $d_{max} = u$, and intersection numbers $b_d = (u - d)(v - 1)$ and $c_d = d$, for all $0 \leq d \leq d_{max}$.*

Proposition 14. *The graph $(\mathcal{V}^u, \sim)$ is a vertex-transitive graph.*

The relation between graph structures we consider in this paper is summarized in Figure 6. Figure 7 displays two examples of database structures $(\mathcal{V}^u, \sim)$. Note that when $|\mathcal{V}| = 2$, $(\mathcal{V}^u, \sim)$ is the u-dimensional hypercube.

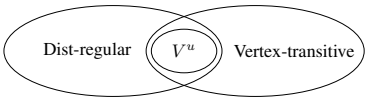
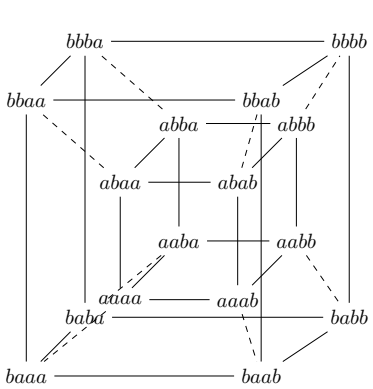
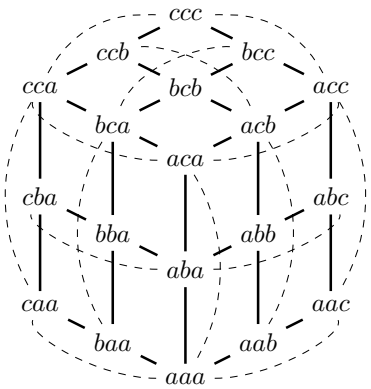


Figure 6: Venn diagram for the classes of graphs considered in Section 5.1.



(a) $u = 4, \mathcal{V} = \{a, b\}$ (4-dimensional hypercube)



(b) $u = 3, \mathcal{V} = \{a, b, c\}$ (for readability sake we only show part of the graph)

Figure 7: Two $(\mathcal{V}^u, \sim)$ graphs

5.2 The bound on leakage

Since the graph structure $(\mathcal{X}, \sim)$ of databases is both distance-regular and vertex-transitive, we can apply Theorem 12 to the channel from X to Z. Then, by (5), it follows that, for $j \in \mathcal{X}_{\langle d \rangle}(x)$ (i.e., every j in $\mathcal{X}$ at distance d from a given x), the transformed matrix M'' satisfies $M''_{x,j} \geq {}^{\max^{M''}}\!/_{e^{\epsilon d}}$.

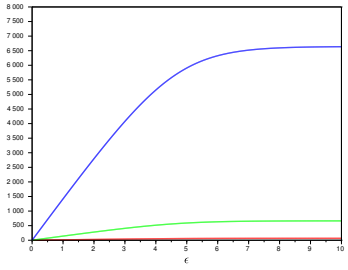
Each element j at distance d from x can be obtained by changing the value of d individuals in the u-tuple representing i. These d individuals can be chosen in $\binom{u}{d}$ possible ways, and for each of those we can change the value (with respect to the one in x) in $v - 1$ possible ways. Therefore $|\mathcal{X}_{\langle d \rangle}(x)| = \binom{u}{d}(v-1)^d$, and the number of databases at distance d from any x must be a constant value n_d given by

$$n_d = \binom{u}{d} (v-1)^d. \tag{7}$$

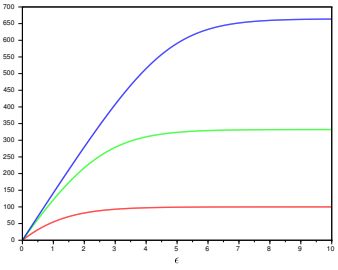
Using the value of n_d from (7) in Theorem 12 we obtain a function Bnd of u, v and ϵ defined as

$$Bnd(u, v, \epsilon) = u \log \frac{v\, e^{\epsilon}}{v - 1 + e^{\epsilon}},$$

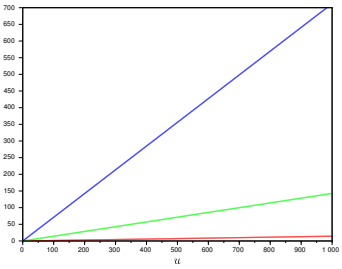
which is an upper bound for the leakage of the mechanism.



(a) $u = 10$ (lowest line), $u = 100$ (intermediate line), and $u = 1000$ (highest line) for fixed $v = 100$ and varying ϵ



(b) $v = 2$ (lowest line), $v = 10$ (intermediate line), and $v = 100$ (highest line) for fixed $u = 100$ and varying ϵ



(c) $\epsilon = 0.01$ (lowest line), $\epsilon = 0.1$ (intermediate line), and $\epsilon = 0.5$ (highest line) for fixed $v = 100$ and varying u

Figure 8: Graphs of $Bnd(u, v, \epsilon)$ for different configurations of u, v, ϵ

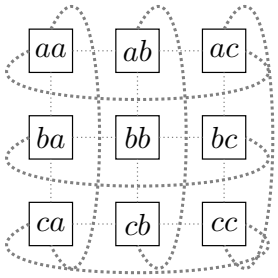
Theorem 15. *If $\mathcal{K}$ satisfies ϵ-differential privacy, then the information leakage is bounded from above as $I_\infty(X;Z) \leq Bnd(u, v, \epsilon)$.*

The bound $Bnd(u, v, \epsilon) = u \log {v\, e^\epsilon}/{(v-1+e^\epsilon)}$ is a continuous function in ϵ, has value 0 when $\epsilon = 0$, and converges to $u \log v$ as ϵ approaches infinity. Figure 8 shows the growth of $Bnd(u, v, \epsilon)$ for different configurations of u, v, ϵ.

Choosing an appropriate value of the parameter ϵ in differential privacy is not trivial, and in general one must consider several factors. The information contained in Figure 8 provides some insight into the implications, in terms of vulnerability, of the choice of ϵ in differential privacy. For instance, with $u = 100$, $v = 2$, and $\epsilon = 5$, the min-entropy capacity can be a high as 99.03 bits, giving a posterior vulnerability exceeding $1/2$, which intuitively means that this combination of parameters is not safe.

The next proposition shows that the bound obtained is tight.

Proposition 16. *For every u, v, and ϵ it is possible to define the mechanism $\mathcal{K}$ below, which provides ϵ-differential privacy and whose min-entropy leakage,*



(a) The datasets and their adjacency relation

	aa	*ab*	*ac*	*ba*	*ca*	*bb*	*bc*	*cb*	*cc*
aa	0	1	1	1	1	2	2	2	2
ab	1	0	1	2	2	1	2	1	2
ac	1	1	0	2	2	2	1	2	1
ba	1	2	2	0	1	1	2	1	2
ca	1	2	2	1	0	2	2	1	1
bb	2	1	2	1	2	0	1	1	2
bc	2	2	1	1	2	1	0	2	1
cb	2	1	2	2	1	1	2	0	1
cc	2	2	1	2	1	2	1	1	0

(b) The representation of the matrix, where each generic entry α stands for $\max^M / e^{\epsilon}\,\alpha$

Figure 9: All possible databases and highest min-entropy leakage matrix giving ϵ-differential privacy for Example 17.

for the uniform input distribution, is $I_\infty(X;Z) = Bnd(u,v,\epsilon)$.

$$\mathcal{K}_{x,z} = \frac{2^{Bnd(u,v,\epsilon)}}{v^u (e^\epsilon)^d}, \quad \textit{for every input } x \textit{ and output } z.$$

We now give an example of the use of $Bnd(u,v,\epsilon)$ as a bound for min-entropy leakage.

Example 17. *Assume we are interested in the eye color of a certain population* $\mathcal{U} = \{Alice, Bob\}$. *Let* $\mathcal{V} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, *where* `a` *stands for* absent *(i.e., the* null *value),* `b` *stands for* blue, *and* `c` *stands for* coal black. *Each dataset is a tuple* $\mathrm{x}_0\mathrm{x}_1 \in \mathcal{V}^2$, *where* x_0 *represents the eye color of Alice (cases* $\mathrm{x}_0 = b$ *and* $\mathrm{x}_0 = c$*), or that Alice is not in the dataset (case* $\mathrm{x}_0 = a$*), whereas* x_1 *provides the same kind of information for Bob. Note that* $v = 3$. *Fig 9(a) represents the graph structure* $(\mathcal{X}, \sim)$ *of the database domain, i.e., the set* $\mathcal{X}$ *of all possible datasets and its adjacency relation. Fig 9(b) represents the matrix with input* $\mathcal{X}$ *which provides* ϵ*-differential privacy and has the highest min-entropy leakage. In the representation of the matrix, the generic entry* α *stands for* $\max^M / e^{\epsilon}\,\alpha$, *where* $\max^M$ *is the highest value in the matrix, i.e.,* $\max^M = 2^{Bnd(u,v,\epsilon)} / v^u = \left(ve^\epsilon / v-1+e^\epsilon\right)^u \cdot 1/v^u = e^{2\epsilon}/(2+e^\epsilon)^2$.

Bounds on leakage do not imply differential privacy

It is important to note that the converse of Theorem 15 does not hold, i.e., a bound on the min-entropy leakage of a channel does not necessarily imply a bound on level of differential privacy of that channel (i.e., on the parameter ϵ). One reason is that the min-entropy is defined as an expected value, i.e., it is the result of averaging the contribution of all the columns to the leakage, while differential privacy represents the worst-case. Hence, there could be a column which breaks differential privacy entirely, for instance in the case in which the

$p_X(\cdot)$		y_1	y_2	$\cdots$	y_m
α	x_1	β	$\frac{1-\beta}{m-1}$	$\cdots$	$\frac{1-\beta}{m-1}$
$\frac{1-\alpha}{m-1}$	x_2	$\frac{1}{m}$	$\frac{1}{m}$	$\cdots$	$\frac{1}{m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\frac{1-\alpha}{m-1}$	x_n	$\frac{1}{m}$	$\frac{1}{m}$	$\cdots$	$\frac{1}{m}$

Figure 10: Input distribution and channel matrix for Example 18

column contains both zero and non-zero elements, and whose leakage, yet, does not contribute too much to the average (typically because the corresponding output has very low probability). In this case, the min-entropy leakage can be very small, and yet ϵ-differential privacy does not hold for any ϵ.

Another (related) reason is that min-entropy is sensitive to the values of the input distribution, whereas differential privacy is not. The following example illustrates this point.

Example 18. *Let $(\mathcal{X}, \mathcal{Y}, M)$ be a channel such that $|\mathcal{X}| = n$ and $\mathcal{Y} = m$. Assume that the channel matrix is such that $p(x_1) = \alpha$ and $p(x_i) = \frac{1-\alpha}{n-1}$ for $2 \leq i \leq n$, and let $p(y_1 \mid x_1) = \beta$, $p(y_j \mid x_1) = \frac{1-\beta}{m-1}$ for $2 \leq j \leq m$, and $p(y_j \mid x_i) = \frac{1}{m}$ otherwise. This channel is represented in Figure 10. Simple calculations show that the min-entropy leakage of the channel approaches 0 as α approaches 0, independently of the value of β.*

Differential privacy, however, depends only on the value of β, more precisely, the parameter of differential privacy is $\max\{\ln \frac{1}{m\beta}, \ln m\beta, \ln \frac{m-1}{m(1-\beta)}, \ln \frac{m(1-\beta)}{m-1}\}$, and such parameter is unbound and goes to infinity as β approaches 0.

Leakage of "skinny" matrices ($m < n$)

The construction of the matrix for Proposition 16 gives a square matrix of dimension $\mathcal{V}^u \times \mathcal{V}^u$. Often, however, the range of $\mathcal{K}$ is fixed, as it is usually related to the possible answers to the query f. Hence it is natural to consider the scenario in which we are given a number $r < \mathcal{V}^u$, and want to consider only those $\mathcal{K}$'s whose range has cardinality at most r. Proposition 20 shows that in this restricted setting we can find a better bound than that from Theorem 15.

Lemma 19. *Let $\mathcal{K}$ be a mechanism with input X, where $\mathcal{X} = \mathcal{V}^u$, providing ϵ-differential privacy. Assume that $r = |Range(\mathcal{K})| = v^\ell$, for some $\ell < u$. Let M be the matrix associated with $\mathcal{K}$. Then it is possible to build a square matrix M' of size $v^\ell \times v^\ell$, with row and column indices in $\mathcal{A} \subseteq \mathcal{X}$, and a binary relation $\sim' \subseteq \mathcal{A} \times \mathcal{A}$ such that $(\mathcal{A}, \sim')$ is isomorphic to $(\mathcal{V}^\ell, \sim_\ell)$, and such that:*

(i) M' is a channel matrix: $\sum_{j=0}^{m-1} M'_{i,j} = 1$ for all $0 \leq i \leq n-1$;

(ii) $M'_{i,j} \leq (e^\epsilon)^{u-l+d} M'_{h,j}$ for all $i, h \in \mathcal{X}$ and $j \in \mathcal{Z}$, where d is the $\sim'$-distance between i and h;

(iii) the elements of the diagonal are all equal to the maximum element of the matrix: $M'_{i,i} = \max^{M'}$ *for all* $i \in \mathcal{X}$*; and*

(iv) $H_\infty^{M'}(X|Z) = H_\infty^{M}(X|Z)$*, if* X *has the uniform distribution.*

Now we are ready to prove the proposition.

Proposition 20. *Let* $\mathcal{K}$ *be a mechanism with associated channel matrix* M*, and let* $r = |Range(\mathcal{K})|$*. If* $\mathcal{K}$ *provides* ϵ-differential privacy *then the min-entropy leakage associated with* $\mathcal{K}$ *is bounded from above as follows:*

$$I_\infty^M(X;Z) \leq \log \frac{r\,(e^\epsilon)^u}{(v-1+e^\epsilon)^\ell - (e^\epsilon)^\ell + (e^\epsilon)^u},$$

where $\ell = \lfloor \log_v r \rfloor$*.*

The bound above can be significantly smaller than that from Theorem 15. For instance, when $r = v$ it becomes

$$\log \frac{v\,(e^\epsilon)^u}{v-1+(e^\epsilon)^u},$$

which for large values of u is much smaller than $Bnd(u,v,\epsilon)$. This does not contradict the fact that the bound $Bnd(u,v,\epsilon)$ is strict—it is strict when we are free to choose the range, which here is fixed.

Leakage about an individual

Protecting an entire database is not the primary goal of differential privacy—some information is expected to be revealed, otherwise the query would not be useful. Instead, differential privacy aims at protecting the value of any single individual, and its definition induces a straightforward bound on the corresponding min-entropy leakage. To derive this straightforward bound, we will use the following result.

Proposition 21. *Let* $(\mathcal{X}, \mathcal{Z}, M)$ *be a channel, with associated input random variable* X *and output random variable* Z*, such that for all* $x, x' \in \mathcal{X}$ *and all* $z \in \mathcal{Z}$*, there is an* $\epsilon \geq 0$ *such that the conditional probabilities of* M *satisfy* $p(z|x)/p(z|x') \leq e^\epsilon$*. Then the min-entropy leakage from this channel is bounded by*

$$I_\infty(X;Z) \leq \log e^\epsilon.$$

Now we derive the first bound on the leakage about an individual. Let x be a database and $x^- = x \backslash \{\mathrm{x}_i\}$ be the database obtained from x after removing the data x_i relative to individual i. The tuple $x^- \in \mathcal{V}^{u-1}$ contains the given (and known) values of all other $u-1$ individuals. If the mechanism outputs answers from a set $\mathcal{Z}$, then the probability of the mechanism producing any answer $z \in \mathcal{Z}$ given any two neighbor databases x and x^- can be related as follows.

$$\begin{aligned}
\frac{p(z|x)}{p(z|x^-)} &= \frac{p(z)p(x|z)}{p(x)} \cdot \frac{p(x^-)}{p(z)p(x^-|z)} && \text{(by the Bayes law)}\\
&= \frac{p(x|z)}{p(x)} \cdot \frac{p(x^-)}{p(x^-|z)}\\
&= \frac{p(x^-|z)p(\mathrm{x}_i|x^-,z)}{p(x^-)p(\mathrm{x}_i|x^-)} \cdot \frac{p(x^-)}{p(x^-|z)} && \text{(by the chain rule of probabilities)}\\
&= \frac{p(\mathrm{x}_i|x^-,z)}{p(\mathrm{x}_i|x^-)} && (8)
\end{aligned}$$

Intuitively, (8) means that the probability of an output z from the mechanism depends on the choice of the value x_i for the fixed individual, once the rest of the database x^- has been fixed.

Hence, we can create a channel whose input X_i ranges over the values $\mathcal{V}$ and represents the value of our individual of interest, and whose output Z ranges over a set of answers $\mathcal{Z}$. Note that in this way we take into consideration all possible input databases where the values of the other individuals are exactly those of x^- and only the value of the selected individual varies.

The min-entropy leakage about an individual in this scenario is given, as usual, in terms of the corresponding a priori and a posteriori min-entropies:

$$\begin{aligned}
I_\infty^{x^-}(X_i;Z) &= H_\infty^{x^-}(X_i) - H_\infty^{x^-}(X_i|Z)\\
&= \log \frac{\sum_z p(z) \max_{\mathrm{x}_i} p(\mathrm{x}_i|x^-,z)}{\max_{\mathrm{x}_i} p(\mathrm{x}_i|x^-)}. && (9)
\end{aligned}$$

If the mechanism respects ϵ-differential privacy for some value of ϵ, then for all $\mathrm{x}, \mathrm{x}' \in \mathcal{V}$ and all $z \in \mathcal{Z}$, ${}^{p(z|\mathrm{x})}/_{p(z|\mathrm{x}')} \leq e^\epsilon$, and hence, by (8) also ${}^{p(\mathrm{x}_i|x^-,z)}/_{p(\mathrm{x}_i|x^-)} \leq e^\epsilon$, and we can apply Proposition 21 to the channel from X_i to Z under a fixed x^- to derive a bound on the leakage:

$$I_\infty^{x^-}(X_i;Z) \leq \log e^\epsilon \qquad (10)$$

With our model, however, we can find a better bound, as shown in the next proposition.

Proposition 22. *Assume that $\mathcal{K}$ satisfies ϵ-differential privacy. Then the information leakage for an individual is bounded from above by*

$$I_\infty^{x^-}(X_i;Z) \leq \log \frac{v\, e^\epsilon}{v - 1 + e^\epsilon}.$$

Note that this bound on leakage for an individual neither depends on the size u of $\mathcal{U}$, nor on the database x^- that we fix, which is in accordance with the fact that the guarantees provided by differential-privacy are independent of any side information. (Indeed, note that the bound on leakage is precisely ${}^{Bnd(u,v,\epsilon)}/_u$.)

Finally, note that the bound given by (10) differs from that of Proposition 22 up to an additive factor of $\log v/(v-1+e^{\epsilon})$. This factor is always non-positive, and it becomes non-negligible only when $e^{\epsilon}-1$ is comparable to v. Since ϵ is usually taken to be a small constant, our bound gives best improvements with respect to the traditional one when v is small. For instance, in a database that collects information about the incidence of a certain disease in a population, we could have $v = 3$ (any individual is either *affected*, *not affected*, or *absent* from the database). If we take $\epsilon = 1.35$, the usual bound on leakage is $\log e^{\epsilon} \approx 1.95$, whereas our bound corrects it by an additive factor of $\log v/(v-1+e^{\epsilon}) \approx -0.97$, meaning that our bound is approximately 50% tighter.

6 Application to utility

In this section we consider the relation between differential privacy and the utility of oblivious mechanisms. In such mechanisms, utility is a property of the noise channel $\mathcal{H}$ of Figure 4, which maps true answer $y \in \mathcal{Y}$ into reported answer $z \in \mathcal{Z}$ according to conditional probability given by $\mathcal{H}_{y,z}$. We use the results of Section 4 to derive bounds on the utility of $\mathcal{H}$, and to construct the optimal mechanism.

We start by noting that the data analyst does not necessarily take the output z of the randomization mechanism $\mathcal{H}$ as a guess for the true answer y, since some Bayesian post-processing can be applied to maximize the probability of a correct guess. For each reported answer z the data analyst can remap their guess to a value $y' \in \mathcal{Y}$ according to some strategy.

The standard way to define utility is via *gain functions* (cfr. [6]). These are functions of the form $gain : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $gain(y, y')$ represents the reward for an adversary who guesses the answer y' when the correct answer is y. It is natural to define the utility of the noise channel $\mathcal{H}$ as the expected gain. In the following definition, which formalizes this concept, $p(y)$ is the prior probability of the true answer y, and $p(y'|y)$ is the probability of guessing y' when the true answer is y. Note that the definition depends implicitly on the reported answer z, since the data analyst guesses a y' on the basis of the observed z.

Definition 23. *The utility of the noise matrix $\mathcal{H}$ is:*

$$\mathcal{U}(Y, Z) = \sum_{y} p(y) \sum_{y'} p(y'|y) gain(y, y') \tag{11}$$

The following characterization of the utility makes explicit the role of the observed z, and of the remapping function used:

Proposition 24. *Assuming that the data analyst uses a remapping function $guess : \mathcal{Z} \to \mathcal{Y}$, we have*

$$\mathcal{U}(Y, Z) = \sum_{y,z} p(y, z) gain(y, guess(z)). \tag{12}$$

We focus here on the so-called *identity gain function*, which is defined as

$$gain_{id}(y, y') = \begin{cases} 1 & \text{if } y = y', \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

The use of identity gain functions in the context of differential privacy was also investigated by Ghosh et al. [17].[2] Let δ_x represent the "point" probability distribution which has value 1 on x and 0 elsewhere. Hence

$$gain_{id}(y, guess(z)) = \delta_y(guess(z)).$$

Intuitively, the function $gain_{id}$ fits situations in which there is little reason to prefer one answer over another, except if it is the correct answer. [3]

By substituting *gain* with $gain_{id}$ in (12) we obtain

$$\mathcal{U}(Y, Z) = \sum_{y,z} p(y, z)\delta_y(guess(z)), \tag{14}$$

which tells us that the utility is highest when $guess(z) = y$ is chosen to maximize $p(y, z)$. Hence this is the data analyst's best strategy. Under such a maximizing remapping, we have

$$\begin{aligned} \mathcal{U}(Y, Z) &= \sum_z \max_y p(y, z) \\ &= \sum_z \max_y (p(y)\, p(z|y)) \qquad \text{(by the Bayes law)} \end{aligned} \tag{15}$$

The formula thus obtained trepresents the Bayes risk, which is the converse of the min-vulnerability. Thus there is a correspondence between $\mathcal{U}$ and the a posteriori min-entropy, expressed by the following proposition.

Proposition 25. *Assume that function gain is the identity and the function guess is optimal. Then:*

$$\mathcal{U}(Y, Z) = \sum_z \max_y (p(y)\, p(z|y)) = V(Y|Z) = 2^{-H_\infty(Y|Z)}.$$

Whereas leakage is related to the difference between the a priori and a posteriori min-entropy of the channel, utility concerns only the a posteriori min-entropy. Note that leakage is a comparative measure, i.e., it tells how much the adversary's probability of success has improved with respect to an initial situation, whereas utility is an absolute measure, as it is only concerned about how much the mechanism's output tells about the true answer to the query.

[2] Instead of gain functions, [17] uses the dual notion of *loss functions*, but the final result is equivalent.

[3] In more general cases, not studied here, the answer domain could be endowed with a notion of distance, and the gain function could take into account the "proximity" of the reported answer to the true one. Intuitively, in this case a "close" answer, even if wrong, should be considered better than a "distant" one.

6.1 The bound on utility and construction of the optimal mechanism

In this section we show that in some special cases, the fact that $\mathcal{K}$ provides ϵ-differential privacy determines a bound on the utility under the identity gain function. We start by observing that the adjacency relation on $\mathcal{X}$ induces, via the query f, an adjacency relation on $\mathcal{Y}$:

Definition 26. *Given $y, y' \in \mathcal{Y}$, with $y \neq y'$, we say that y and y' are adjacent (notation $y \sim y'$), if, and only if, there exist $x, x' \in \mathcal{V}^u$ with $x \sim x'$ such that $y = f(x)$ and $y' = f(x')$.*

Since $\sim$ is symmetric on databases, it is also symmetric on $\mathcal{Y}$, therefore also $(\mathcal{Y}, \sim)$ forms an undirected graph.

Using the above concept of neighborhood for the inputs of the noise channel $\mathcal{H}$, we can show that in an oblivious mechanism of the form represented in Figure 4, $\mathcal{K}$ satisfies ϵ-differential privacy with respect to neighbor databases if, and only if, $\mathcal{H}$ satisfies ϵ-differential privacy with respect to neighbor answers.

Proposition 27. *In an oblivious setting (cf. Figure 4), if the query function f is deterministic, then the mechanism $\mathcal{K}$ satisfies ϵ-differential privacy with respect to $\mathcal{X}$ if, and only if, the noise channel $\mathcal{H}$ satisfies ϵ-differential privacy with respect to $\mathcal{Y}$.*

The link established by the above proposition will help us determine a bound on the utility of $\mathcal{H}$. Note that, as was also the case in the previous Section, the bounds derived in this Section depend on the graph structure of the channel input satisfying either distance-regularity or vertex-transitivity. However, whereas the graph structure on databases is guaranteed to always satisfy both types of symmetries, the graph structure on true answers may present only one of them, or neither (cfr. Figure 2). Hence, the two alternative ways of performing the second step of the matrix transformation of Figure 5 will be particularly useful here.

In the following, we use n_d to represent the number of nodes in $\mathcal{Y}$ at distance d from another node in $y \in \mathcal{Y}$. We recall that in any distance-regular or vertex-transitive graph, such number depends only on d and not on y (cf. Proposition 5).

Theorem 28. *Consider a noise channel $\mathcal{H}$ satisfying ϵ-differential privacy for some $\epsilon > 0$. Assume that the distribution of Y is uniform and that $(\mathcal{Y}, \sim)$ is either distance-regular or vertex-transitive. Then we have:*

$$\mathcal{U}(Y, Z) \leq \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}} \tag{16}$$

Provided $(\mathcal{Y}, \sim)$ is distance-regular or vertex-transitive, the above bound is tight, in the sense that there is a channel matrix whose utility coincides with

the bound. Indeed, for $0 \leq i, j \leq n-1$, define $\mathcal{H}$ as follows.

$$\mathcal{H}_{i,j} = \frac{\gamma}{e^{\epsilon\, d(i,j)}}, \tag{17}$$

where

$$\gamma = \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}}. \tag{18}$$

The following proposition shows that this definition gives a channel with maximal utility.

Theorem 29. *Assume $(\mathcal{Y}, \sim)$ is distance-regular or vertex-transitive and that the distribution of Y is uniform. Then the matrix $\mathcal{H}$ defined by (17) is a channel matrix that satisfies ϵ-differential privacy and has maximal utility:*

$$\mathcal{U}(Y, Z) = \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}}$$

Note that the definition of $\mathcal{H}$ given by (17) needs the condition of distance-regularity or vertex-transitivity in order for n_d to be defined. Furthermore, if the distribution on $\mathcal{Y}$ is not uniform, then the utility of such $\mathcal{H}$ is not necessarily optimal. These are strong limitations for the results in this section, because the structure of $(\mathcal{Y}, \sim)$ and the distribution on $\mathcal{Y}$ depend on the query f, and in general the above conditions are not granted by f.

6.2 Examples

Although our method for the construction of the optimal noise channel requires strong conditions (distance-regularity or vertex-transitivity), there are some interesting scenarios in which they are satisfied. Furthermore:

Remark 30. *Our method can be applied also when the conditions of Theorem 29 are not met—we can always add "artificial" adjacencies to the graph structure so as to meet those conditions. Instead of $(\mathcal{Y}, \sim)$, for computing the distance in (17) we use a structure $(\mathcal{Y}, \sim')$ that satisfies the conditions of Theorem 29, and such that $\sim \subseteq \sim'$. The matrix constructed in this way provides ϵ-differential privacy, but may be non-optimal. In general, the smaller $\sim'$, the higher the utility.*

The following are two simple scenarios in which the conditions are satisfied:

- $(\mathcal{Y}, \sim)$ is a *clique*: every element has exactly $|\mathcal{Y}| - 1$ adjacent elements.
- $(\mathcal{Y}, \sim)$ is a *ring*: every element has exactly two adjacent elements. This is similar to the case of the counting queries considered in [17], with the difference that our "counting" is in arithmetic modulo $|\mathcal{Y}|$.

The next two examples illustrate queries that give rise to the clique structure and to a line structure—which is then transformed into a ring by adding artificial adjacencies to the graph—, and show the corresponding matrices. Note that the matrices generated by our method can be rather different, depending on the structure of $(\mathcal{Y}, \sim)$.

Example 31. *Consider a database with electoral information in which each entry corresponds to a ballot cast, which is described by three fields:*

- elector*: a unique (anonymized) identifier assigned to an elector;*
- candidate*: the name of the candidate the elector voted for; and*
- city*: the name of the city where the ballot was cast.*

Consider the query "What is the city with the greatest number of votes for a given candidate c?", and assume that the data analyst models the utility with the identity gain function.[4] *Note that every two answers are neighbors, so the graph structure of the domain of answers is a clique.*

Consider the scenario where the set of cities is $\mathrm{City} = \{A, B, C, D, E, F\}$ *and assume for simplicity that there is a unique answer for the query, i.e., there are no two cities with exactly the same number of individuals voting for candidate c. Table 1 shows two alternative mechanisms providing ϵ-differential privacy (with $\epsilon = \ln 2$).*

The first one, M_1, is based on the truncated geometric mechanism method used in [17] for counting queries, here adapted to the case where every two distinct answers are neighbors. More precisely, the mechanism is defined by mapping $A, \ldots, F$ into the interval of natural numbers $[0, 6]$, then constructing the truncated geometric mechanism as in [17], where the parameter α (representing the minimum ratio between the elements of the matrix) is calculated imposing both the ϵ-differential privacy constraint (with $\epsilon = \ln 2$) and that all natural numbers in $[0, 6]$ are adjacent to each other. We obtain that $\alpha^5 = 2$, i.e. $\alpha \approx 1.5$.

The second mechanism, M_2, is constructed according to (17). Theorem 29 ensures that, for the uniform input distribution, M_2 gives optimal utility. With the uniform input distribution, indeed, we have $\mathcal{U}(M_1) = 0.2243 < 0.2857 = \mathcal{U}(M_2)$.

Even for non-uniform distributions, our mechanism still provides better utility. For instance, for $p(A) = p(F) = 1/10$ and $p(B) = p(C) = p(D) = P(E) = 1/5$, we have $\mathcal{U}(M_1) = 0.2415 < 0.2857 = \mathcal{U}(M_2)$. This should not come as a surprise: the geometric mechanism, as well as the Laplacian mechanism, performs well when they are defined on the natural metrics of the answer domain. Here, the structure $[0, 6]$ is sort of imposed artificially.

[4] Clearly there are more interesting gain function, for instance with a suitable gain function we could distinguish the case in which the answer is totally wrong from an answer that is "almost correct" in the sense that the guessed town was the second best for the candidate. However, this is out of the scope of this paper as here we consider only binary gain functions.

(a) M_1: adapted truncated geometric mechanism

In/Out	A	B	C	D	E	F
A	0.534	0.060	0.053	0.046	0.040	0.267
B	0.465	0.069	0.060	0.053	0.046	0.307
C	0.405	0.060	0.069	0.060	0.053	0.353
D	0.353	0.053	0.060	0.069	0.060	0.405
E	0.307	0.046	0.053	0.060	0.069	0.465
F	0.267	0.040	0.046	0.053	0.060	0.534

(b) M_2: our mechanism

In/Out	A	B	C	D	E	F
A	2/7	1/7	1/7	1/7	1/7	1/7
B	1/7	2/7	1/7	1/7	1/7	1/7
C	1/7	1/7	2/7	1/7	1/7	1/7
D	1/7	1/7	1/7	2/7	1/7	1/7
E	1/7	1/7	1/7	1/7	2/7	1/7
F	1/7	1/7	1/7	1/7	1/7	2/7

Table 1: Mechanisms for the city with higher number of votes for candidate c

Example 32. *Consider the same database of the previous example, and assume now the query of interest "What is the number of votes for candidate c?". Each answer has at most two neighbors, and hence the graph structure on the domain of answers is a line. Assume, for simplicity, that* 5 *individuals participated in the election.*

Table 2 shows again the two mechanisms providing ϵ-differential privacy ($\epsilon = \ln 2$) that we considered in previous example: the truncated geometric mechanism M_1 (in which, now, $\alpha = 2$, because we only have the constraint of $(\ln 2)$-differential privacy), and ours, M_2. Note that in order to apply our method we have first to apply Remark 30 to transform the graph structure from a line into a ring.

Considering the uniform prior distribution, the utility of M_1 is now greater than that of M_2: 4/9 versus 8/21, respectively. This does not contradict our theorem, because our matrix is guaranteed to be optimal only in the case of a ring structure, not of a line as we have in this example. (In fact, if the structure were a ring, i.e., if the last row were adjacent to the first one, then M_1 would not provide ϵ-differential privacy.) In case of a line, as here, the truncated geometric mechanism has been proved optimal [17].

7 Related work

To the best of our knowledge, the first work to investigate the relation between differential privacy and information-theoretic leakage for an individual was [3]. Their definition of channel was for a given database, and the channel inputs were all possible databases adjacent to it. Two bounds on leakage were presented, one

(a) M_1: truncated $\frac{1}{2}$-geom. mechanism

In/Out	0	1	2	3	4	5
0	2/3	1/6	1/12	1/24	1/48	1/48
1	1/3	1/3	1/6	1/12	1/24	1/24
2	1/6	1/6	1/3	1/6	1/12	1/12
3	1/12	1/12	1/6	1/3	1/6	1/6
4	1/24	1/24	1/12	1/6	1/3	1/3
5	1/48	1/48	1/24	1/12	1/6	2/3

(b) M_2: our mechanism

In/Out	0	1	2	3	4	5
0	8/21	4/21	2/21	1/21	2/21	4/21
1	4/21	8/21	4/21	2/21	1/21	2/21
2	2/21	4/21	8/21	4/21	2/21	1/21
3	1/21	2/21	4/21	8/21	4/21	2/21
8	2/21	1/21	2/21	4/21	8/21	4/21
5	4/21	2/21	1/21	2/21	4/21	8/21

Table 2: Mechanisms for the counting query (5 voters)

for the min-entropy, and one for Shannon entropy. Our bound in Proposition 22 improves the min-entropy bound of [3].

Barthe and Köpf [5] were the first to investigate the more challenging connection between differential privacy and the min-entropy leakage for the set of all possible databases. They considered "end-to-end" differentially-private mechanisms, which correspond to what in this paper we call the mechanism $\mathcal{K}$, and proposed, like we do, to interpret these mechanisms as information-theoretic channels. Barthe and Köpf provided a bound for leakage, but pointed out that it was not tight in general. They also observed that for any number of individuals u and level of privacy ϵ one can construct a channel whose maximal leakage is $u \log {(2e^\epsilon)}/{(e^\epsilon+1)}$, and concluded therefore that the bound must be at least as high as such expression. Another difference between their work and ours is that [5] captures the case in which the focus of differential privacy is on hiding participation of individuals in a database, whereas we consider both the participation and the values of the individuals.

Clarkson and Schneider also considered differential privacy as a case study of their proposal for quantification of integrity [11]. They analyzed database privacy conditions from the literature (such as differential privacy, k-anonymity, and l-diversity) for utility quantification. In particular, they studied the relationship between differential privacy and a notion of leakage (being different from ours as their definition is based on Shannon entropy) and they provided a tight bound on leakage.

Heusser and Malacaria [19] were among the first to explore the application of information-theoretic concepts to databases queries. They proposed to model database queries as programs, which allows for statistical analysis of the infor-

mation leaked by the query. However, [19] did not relate information leakage to differential privacy.

Ghosh et al. [17] aimed at obtaining optimal-utility mechanisms while preserving differential privacy. They proposed adding noise to the output of the query according to the geometric mechanism. Their framework provides a general definition of utility for a mechanism M that captures any possible side information and preferences the users of M may have. They proved that the geometric mechanism is optimal in the particular case of counting queries. Our results in Section 6 do not restrict to counting queries, yet we only consider the case of the identity gain (loss) function.

Finally, our definition of the channel matrix in (17) corresponds to the exponential mechanism of McSherry and Talwar [24] when the quality function relating two answers i and j is taken to be $-d(i,j)$, that is, the negative of the distance between them. Therefore, under the hypothesis of Theorem 28, it follows that the exponential mechanism is an optimal way to maximize utility as measured by the identity gain-function, while preserving differential privacy.

8 Conclusion

We have investigated the relations of ϵ-differential privacy with leakage, and utility, extending our previous work [1, 2]. Our main contribution has been the development of a general technique for determining these relations depending on the graph structure of the input domain, induced by the adjacency relation and by the query. We have considered two particular structures, the distance-regular graphs, and the vertex-transitive graphs, which allowed us to obtain tight bounds on leakage and on utility. We also constructed an optimal noise channel satisfying ϵ-differential privacy for some special cases.

As future work, we plan to extend our result to other kinds of utility functions, among which we feel promising the g-leakage framework [4].

Acknowledgements We are very grateful to the anonymous referees: Their comments and suggestions helped us to improve the paper in a substantial way.

Mário S. Alvim was a postdoctoral research associate at the Mathematics Department of the University of Pennsylvania when part of this research was performed, and he was partially supported by the AFOSR MURI "Science of Cyber Security: Modeling, Composition, and Measurement" as AFOSR grant FA9550-11-1-0137. The work of Konstatinos Chatzikokolakis and Catuscia Palamidessi was partially supported by the European 7th FP grant MEALS, the project ANR-12-IS02-001 PACE, the INRIA Large Scale Initiative CAPPRIS, and the INRIA Associated team PRINCESS. Part of this work was carried on during a visit of Pierpaolo Degano at LIX, partially supported by a Digiteo/INRIA grant, and part within the MIUR-PRIN project Security Horizons.

References

[1] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential Privacy: on the trade-off between Utility and Information Leakage. In Gilles Barthe, Anupam Datta, and Sandro Etalle, editors, *Postproceedings of the 8th International Worshop on Formal Aspects in Security and Trust (FAST)*, volume 7140 of *Lecture Notes in Computer Science*, pages 39–54, Leuven, Belgium, March 2011. Springer.

[2] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. On the relation between Differential Privacy and Quantitative Information Flow. In Jiri Sgall Luca Aceto, Monika Henzinger, editor, *38th International Colloquium on Automata, Languages and Programming (ICALP)*, volume 6756 of *Lecture Notes in Computer Science*, pages 60–76, Zurich, Switzerland, 2011. Springer.

[3] Mário S. Alvim, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy versus quantitative information flow. Technical report, INRIA and LIX, Ecole Polytechnique, 2010. http://hal.archives-ouvertes.fr/hal-00548214/en/.

[4] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *Proceedings of the 25th IEEE Computer Security Foundations Symposium (CSF)*, pages 265–279, 2012.

[5] Gilles Barthe and Boris Köpf. Information-theoretic bounds for differentially private mechanisms. In *Proceedings of the 24th IEEE Computer Security Foundations Symposium (CSF)*, pages 191–204. IEEE Computer Society, 2011.

[6] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc., 1994.

[7] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 609–618. ACM, 2008.

[8] Christelle Braun, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Compositional methods for information-hiding. In Roberto Amadio, editor, *Proceedings of the 11th International Conference on the Foundations of Software Science and Computation Structures (FOSSACS'08)*, volume 4962 of *Lecture Notes in Computer Science*, pages 443–457. Springer, 2008.

[9] Christelle Braun, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Quantitative notions of leakage for one-try attacks. In *Proceedings of the*

25th Conf. on Mathematical Foundations of Programming Semantics, volume 249 of *Electronic Notes in Theoretical Computer Science*, pages 75–91. Elsevier B.V., 2009.

[10] A.E. Brouwer, A.M. Cohen, and A. Neumaier. *Distance Regular Graphs*. Ergebnisse der Mathematik 3.18. Springer-Verlag, 1989.

[11] M. R. Clarkson and F. B. Schneider. Quantification of integrity. *Mathematical Structures in Computer Science*, 2011. To appear.

[12] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *33rd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[13] Cynthia Dwork. Differential privacy in new settings. In Moses Charikar, editor, *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 174–183. SIAM, 2010.

[14] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–96, 2011.

[15] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In Michael Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 371–380, Bethesda, MD, USA, May 31 - June 2 2009. ACM.

[16] Barbara Espinoza and Geoffrey Smith. Min-entropy leakage of channels in cascade. In Gilles Barthe, Anupam Datta, and Sandro Etalle, editors, *Proceedings of the International Worshop on Formal Aspects in Security and Trust (FAST)*, volume 7140 of *Lecture Notes in Computer Science*, pages 70–84. Springer, 2011.

[17] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the 41st annual ACM Symposium on Theory of Computing (STOC)*, pages 351–360, New York, NY, USA, 2009. ACM.

[18] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 61–70, Washington, DC, USA, 2010. IEEE Computer Society.

[19] Jonathan Heusser and Pasquale Malacaria. Applied quantitative information flow and statistical databases. In Pierpaolo Degano and Joshua D. Guttman, editors, *Proceedings of the International Workshop on Formal Aspects in Security and Trust (FAST 2009)*, volume 5983 of *Lecture Notes in Computer Science*, pages 96–110. Springer, 2009.

[20] Wilfried Imrich and Sandi Klavžar. *Product graphs, structure and recognition*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2000.

[21] Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. Cryptology ePrint Archive, Report 2008/144, 2008. http://eprint.iacr.org/.

[22] Yusuke Kawamoto, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Compositionality Results for Quantitative Information Flow. In Gethin Norman and William H. Sanders, editors, *Proceedings of the 11th International Conference on Quantitative Evaluation of SysTems (QEST 2014)*, volume 8657 of *Lecture Notes in Computer Science*, pages 368–383. Springer, 2014. IDEX Digital Society project.

[23] Boris Köpf and David A. Basin. An information-theoretic model for adaptive side-channel attacks. In Peng Ning, Sabrina De Capitani di Vimercati, and Paul F. Syverson, editors, *Proceedings of the 2007 ACM Conference on Computer and Communications Security (CCS 2007)*, pages 286–296. ACM, 2007.

[24] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007)*, pages 94–103. IEEE Computer Society, 2007.

[25] Aaron Roth and Tim Roughgarden. Interactive privacy via the median mechanism. In *Proc. of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 765–774, 2010.

[26] Geoffrey Smith. On the foundations of quantitative information flow. In Luca de Alfaro, editor, *Proceedings of the 12th International Conference on Foundations of Software Science and Computation Structures (FOSSACS 2009)*, volume 5504 of *LNCS*, pages 288–302, York, UK, 2009. Springer.

[27] Geoffrey Smith. Quantifying information flow using min-entropy. In *Eighth International Conference on Quantitative Evaluation of Systems, QEST 2011, Aachen, Germany, 5-8 September, 2011*, pages 159–167. IEEE Computer Society, 2011.

A Proofs

Proofs are given in the order the corresponding results appear in the main text.

Before proceeding, we providing additional notation, definitions, and auxiliary results needed for the proofs in the reminding of this section.

A.1 Auxiliary notation, definitions, and results

Adjacency relation on channel outputs. At the end of the trasformation operated by Lemma 7, and after removing the all-zero columns, we have a square matrix in which the maximum of each column is in the diagonal, as in Figure 11. On the columns $\mathcal{B}$ of this matrix we define a a graph structure derived from that of $\mathcal{A}$. This will be useful to prove the Step $2a$ and Step $2b$ of the matrix transformation in Section 4. The definition is simply the following: given two elements $j_1, j_2 \in \mathcal{B}$, we stipulate that $j_1 \sim j_2$ if, and only if, j_1, j_2 are also in $\mathcal{A}$ [5], and $j_1 \sim j_2$ in $\mathcal{A}$.

In this way $(\mathcal{B}, \sim)$ is also a graph structure, and the notions of distance, diameter, set of vertices at distance d from a given vertex, etc., are defined as usual. Also $(\mathcal{A}, \sim)$ and $(\mathcal{B}, \sim)$ are isomorphic hence all the properties that hold for $\mathcal{A}$ hold also for $\mathcal{B}$.

Automorphisms. Let $(\mathcal{V}, \sim)$ be a graph, $v, w \in \mathcal{V}$ be two of its vertices, and let Γ be an automorphism group for $(\mathcal{V}, \sim)$. We define the set of automorphisms that map v to w as

$$\Gamma_{v \mapsto w} = \{\sigma \in \Gamma \mid \sigma(v) = w\}.$$

[5] Recall that for simplicity we represent the elements of $\mathcal{A}$ and $\mathcal{B}$ by their indexes.

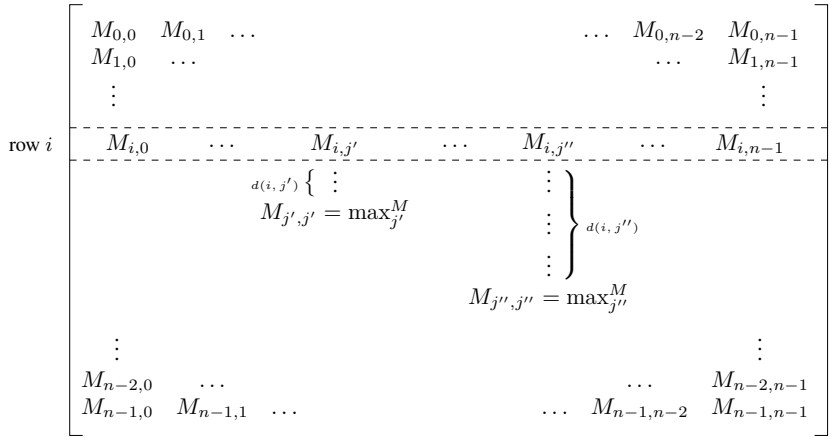


Figure 11: The relation between elements of a row i and the elements in the diagonal

Note that $\Gamma_{v\mapsto w} \cap \Gamma_{v\mapsto w'} = \emptyset$ for all $w \neq w'$, and that $\Gamma = \bigcup_w \Gamma_{v\mapsto w}$ for any $v \in \mathcal{V}$.

Lemma 33 states that in a vertex-transitive graph, starting from any vertex v, there exist as many automorphisms mapping v to w as to any other vertex w'. This auxiliary result is used in the proof of Lemma 9.

Lemma 33. *Let $(\mathcal{V}, \sim)$ be a finite vertex-transitive graph, $n = |\mathcal{V}|$ and Γ its full automorphism group. Then, for all $v, w, w' \in \mathcal{V}$:*

$$|\Gamma_{v\mapsto w}| = |\Gamma_{v\mapsto w'}| = \frac{|\Gamma|}{n}$$

Proof. Given two automorphisms σ and ρ, let $\rho \circ \sigma$ denote the composition of ρ with σ, i.e., the automorphism one would obtain by first applying σ to every vertex of the graph, and then applying ρ to the resulting mapping. By extension, given an automorphism group Γ and an automorphism ρ, we write $\rho \circ \Gamma$ for the automorphism group obtained by composing ρ with every $\sigma \in \Gamma$.

Assume that $|\Gamma_{v\mapsto w}| < |\Gamma_{v\mapsto w'}|$ for some $v, w, w' \in \mathcal{V}$. Since the graph is vertex-transitive, there exists an automorphism ρ such that $\rho(w') = w$.

Consider the set of automorphisms $\rho \circ \Gamma_{v\mapsto w'}$. This set contains $|\Gamma_{v\mapsto w'}|$ distinct automorphisms (since $\rho \circ \sigma = \rho \circ \sigma'$ implies $\sigma = \sigma'$). Moreover these automorphisms map v to w, and therefore we have $\rho \circ \Gamma_{v\mapsto w'} \subseteq \Gamma_{v\mapsto w}$, which is a contradiction since by hypothesis we have $|\Gamma_{v\mapsto w}| < |\Gamma_{v\mapsto w'}| = |\rho \circ \Gamma_{v\mapsto w'}|$.

Thus $|\Gamma_{v\mapsto w}| \geq |\Gamma_{v\mapsto w'}|$ and by exchanging w, w' we get $|\Gamma_{v\mapsto w}| = |\Gamma_{v\mapsto w'}|$. Finally from $\Gamma = \bigcup_w \Gamma_{v\mapsto w}$ we get $|\Gamma_{v\mapsto w}| = \frac{|\Gamma|}{n}$. □

A.2 Preliminaries (Section 2)

Proposition 5. *If a graph $G = (\mathcal{V}, \sim)$ is distance-regular or vertex-transitive, then, for every distance $d \in \Delta$, there exists a constant n_d which is equal to the cardinality of $\mathcal{V}_{\langle d\rangle}(v)$ for every $v \in \mathcal{V}$.*

Proof. Let us first examine the case in which G is distance-regular. Consider the alternative definition of distance-regularity given by proposition 3, and assume $w = v$ and $j = k = d$. We have $\mathcal{V}_{\langle d\rangle}(v) = \mathcal{V}_{\langle j\rangle}(v) \cap \mathcal{V}_{\langle k\rangle}(w)$ whose cardinality, by Proposition 3, depends only on j, k (both equal d), and on $d(v, w)$. Finally, note that $d(v, w) = d(v, v) = 0$.

Consider now the case in which G is vertex-transitive. Given two vertices v and w, we have that there exists an automorphism σ such that $\sigma(v) = w$. By simple induction it can be shown that for every $d \in \Delta$, if $\mathcal{V}_{\langle d\rangle}(v) = \{v_1, v_2, \ldots, v_n\}$, then $\mathcal{V}_{\langle d\rangle}(w) = \{\sigma(v_1), \sigma(v_2), \ldots \sigma(v_n)\}$, which means that the cardinality of $\mathcal{V}_{\langle d\rangle}(v)$ depends only on d. □

A.3 Relating differential privacy and quantitative information flow (Section 4)

Lemma 7. (Step 1) *Let M be a channel matrix of dimensions $n \times m$ s.t. $n \leq m$, and assume that M satisfies ϵ-differential privacy. Then it is possible to*

transform M into a matrix M' of the same dimensions satisfying the following conditions:

(i) *M' is a channel matrix: $\sum_{j=0}^{m-1} M'_{i,j} = 1$, for all $0 \leq i \leq n-1$;*

(ii) *each of the first n columns has a maximum in the diagonal: $M'_{i,i} = \max_i^{M'}$, for all $0 \leq i \leq n-1$;*

(iii) *the $m-n$ last columns only contain 0's: $M'_{i,j} = 0$, for all $0 \leq i \leq n-1$ and all $n \leq j \leq m-1$;*

(iv) *M' satisfies ϵ-differential privacy: $M'_{i,j} \leq e^{\epsilon} M'_{h,j}$, for all $0 \leq i, h \leq n-1$ s.t. $i \sim h$ and all $0 \leq j \leq m-1$; and*

(v) *$H_{\infty}^{M'}(A|B) = H_{\infty}^{M}(A|B)$, if A has the uniform distribution.*

Proof. For $0 \leq k \leq m-1$, define Col_k as the set of columns indexes j of M which have maximum in row k, and in no other row with index smaller than k, i.e.,

$$Col_k = \{j \mid M_{k,j} = \max_j^M \text{ and } \forall i < k.\ M_{i,j} < \max_j^M\}$$

Note that, by construction, each j belongs to exactly one set Col_k. Consequently we have

$$\bigcup_k Col_k = \{0, 1, \ldots, m\} \tag{19}$$

and

$$Col_h \cap Col_k = \emptyset \text{ for } h \neq k \tag{20}$$

Note that some Col_k may be empty. In particular, $Col_k = \emptyset$ for all $k > n$.

Now, define the matrix M' as:

$$M'_{ik} = \sum_{j \in Col_k} M_{ij} \text{ for } 0 \leq i \leq n-1\ ,\ 0 \leq k \leq m-1$$

We prove now that M' satisfies the required properties.

It is easy to see that M' is a valid channel matrix thanks to properties (19) and (20).

We prove now that the maximum of each column $k \leq n$ is in the diagonal (condition (ii)):

$$M'_{kk} = \sum_{j \in Col_k} M_{kj} \geq \sum_{j \in Col_k} M_{ij} = M'_{ik}$$

Condition (iii) is satisfied because $Col_k = \emptyset$ for all $k > n$.

Furthermore, M' satisfies ϵ-differential privacy (condition (iv)):

$$M'_{ik} = \sum_{j \in Col_k} M_{ij} \leq \sum_{j \in Col_k} e^{\epsilon} M_{hj} = e^{\epsilon} \sum_{j \in Col_k} M_{hj} = e^{\epsilon} M'_{hk}$$

Finally, observe that:

$$\sum_k \max_k^{M'} = \sum_k M'_{kk} = \sum_k \sum_{j \in Col_k} M_{kj} = \sum_k \sum_{j \in Col_k} \max_j^M = \sum_j \max_j^M$$

from which it immediately follows that $H_\infty^{M'}(A|B) = H_\infty^M(A|B)$ (recall that A has the uniform distribution and therefore the a posteriori entropy is a function of the sum of the maximum of each column), so condition (v) is satisfied. □

Lemma 8. (Step 2*a*) *Let M' be a square channel matrix of dimensions $n \times n$ that satisfies ϵ-differential privacy. Let $\sim$ be an adjacency relation on $\mathcal{A}$ such that the graph $(\mathcal{A}, \sim)$ is connected and distance-regular. Assume that the maximum value of each column is on the diagonal, i.e., $M_{i,i} = \max_i^M$ for all $0 \leq i \leq n-1$. Then it is possible to transform M' into a matrix M'' of the same dimension satisfying the following conditions:*

(i) M'' is a channel matrix: $\sum_{j=0}^{n-1} M''_{i,j} = 1$, for all $0 \leq i \leq n-1$;

(ii) the elements of the diagonal are all the same, and are equal to the maximum of the matrix: $M''_{i,i} = \max^{M''}$, for all $0 \leq i \leq n-1$;

(iii) M'' satisfies ϵ-differential privacy: $M'_{i,j} \leq e^\epsilon M'_{h,j}$, for all $0 \leq i, h, j \leq n-1$ s.t. $i \sim h$; and

(iv) $H_\infty^{M''}(A|B) = H_\infty^{M'}(A|B)$, if A has the uniform distribution.

Proof. We define each element $M''_{i,j}$ of the new matrix M'' as an averaging, or "smoothing", of the elements of M' that are at same distance from i as j is:

$$M''_{i,j} = \frac{1}{n|\mathcal{A}_{\langle d(i,j)\rangle}(i)|} \sum_{k \in \mathcal{B}} \sum_{h \in \mathcal{A}_{\langle d(i,j)\rangle}(k)} M'_{h,k}$$

We prove that this definitions satisfies Conditions (i) – (iv). We start with Condition (i).

$$\begin{aligned}\sum_{j \in \mathcal{B}} M''_{i,j} &= \sum_{j \in \mathcal{B}} \frac{1}{n|\mathcal{A}_{\langle d(i,j)\rangle}(i)|} \sum_{k \in \mathcal{B}} \sum_{h \in \mathcal{A}_{\langle d(i,j)\rangle}(k)} M'_{h,k} \\ &= \frac{1}{n} \sum_{k \in \mathcal{B}} \sum_{j \in \mathcal{B}} \frac{1}{|\mathcal{A}_{\langle d(i,j)\rangle}(i)|} \sum_{h \in \mathcal{A}_{\langle d(i,j)\rangle}(k)} M'_{h,k}\end{aligned}$$

Note that for every i, $\mathcal{B} = \bigcup_{d \in \Delta} \mathcal{B}_{\langle d\rangle}(i)$, and for different values of d the sets $\mathcal{B}_{\langle d\rangle}(i)$ are disjoint. Therefore by splitting the summation over $j \in \mathcal{B}$ we obtain

$$\begin{aligned}&= \frac{1}{n} \sum_{k \in \mathcal{B}} \sum_{d \in \Delta} \sum_{j \in \mathcal{B}_{\langle d\rangle}(i)} \frac{1}{|\mathcal{A}_{\langle d\rangle}(i)|} \sum_{h \in \mathcal{A}_{\langle d\rangle}(k)} M'_{h,k} \\ &= \frac{1}{n} \sum_{k \in \mathcal{B}} \sum_{d \in \Delta} \sum_{h \in \mathcal{A}_{\langle d\rangle}(k)} M'_{h,k} \sum_{j \in \mathcal{B}_{\langle d\rangle}(i)} \frac{1}{|\mathcal{A}_{\langle d\rangle}(i)|}\end{aligned}$$

since $\sum_{j \in \mathcal{B}_{\langle d \rangle}(i)} \frac{1}{|\mathcal{A}_{\langle d \rangle}(i)|} = \sum_{j \in \mathcal{A}_{\langle d \rangle}(i)} \frac{1}{|\mathcal{A}_{\langle d \rangle}(i)|} = 1$ (recall that the definition of distance in $\mathcal{B}$ is given as that in $\mathcal{A}$), we obtain

$$= \frac{1}{n} \sum_{k \in \mathcal{B}} \sum_{d \in \Delta} \sum_{h \in \mathcal{A}_{\langle d \rangle}(k)} M'_{h,k}$$

and now the summations over h and d can be joined together

$$= \frac{1}{n} \sum_{k \in \mathcal{B}} \sum_{h \in \mathcal{A}} M'_{h,k}$$

now, reorganizing the summations, and considering that M' is a channel matrix, and that $|\mathcal{B}| = n$ we have

$$\begin{aligned} &= \frac{1}{n} \sum_{h \in \mathcal{A}} \sum_{k \in \mathcal{B}} M'_{h,k} \\ &= \frac{1}{n} \sum_{h \in \mathcal{A}} 1 \\ &= 1 \end{aligned}$$

which implies that condition (i) is satisfied.

We now turn our attention to the elements of the diagonal, which are all identical because $M''_{i,i} = \frac{1}{n} \sum_{k \in \mathcal{B}} M'_{k,k}$. To fulfill condition (ii) we still need to show that $M''_{i,i} = \max_i^{M''}$ for all $i \in \mathcal{A}$.

$$\begin{aligned} M''_{i,j} &= \frac{1}{n|\mathcal{A}_{\langle d(i,j) \rangle}(i)|} \sum_{k \in \mathcal{B}} \sum_{h \in \mathcal{A}_{\langle d(i,j) \rangle}(k)} M'_{h,k} \\ &\leq \frac{1}{n|\mathcal{A}_{\langle d(i,j) \rangle}(i)|} \sum_{k \in \mathcal{B}} \sum_{h \in \mathcal{A}_{\langle d(i,j) \rangle}(k)} M'_{k,k} \qquad \begin{array}{r}\text{(since the biggest element} \\ \text{is in the diagonal)}\end{array} \\ &= \frac{1}{n} \sum_{k \in \mathcal{B}} M'_{k,k} \frac{1}{|\mathcal{A}_{\langle d(i,j) \rangle}(i)|} \sum_{h \in \mathcal{A}_{\langle d(i,j) \rangle}(k)} 1 \\ &= \frac{1}{n} \sum_{k \in \mathcal{B}} M'_{k,k} \frac{|\mathcal{A}_{\langle d(i,j) \rangle}(k)|}{|\mathcal{A}_{\langle d(i,j) \rangle}(i)|} \end{aligned}$$

and because the graph is distance-regular, the number of vertices at any arbitrary distance (and, in particular, at distance $d(i,j)$) from any vertice is always the same. Hence $|\mathcal{A}_{\langle d(i,j) \rangle}(k)| = |\mathcal{A}_{\langle d(i,j) \rangle}(i)|$ and we can conclude

$$\begin{aligned} &= \frac{1}{n} \sum_{k \in \mathcal{B}} M'_{k,k} \cdot 1 \\ &= M''_{i,i} \end{aligned}$$

Since A has the uniform distribution, the conditional min-entropy $H_\infty^{M''}(A|B)$ is given by the sum of the maximum elements of each column of M''. Since this sum is preserved in the transformation, we have $H_\infty^{M''}(A|B) = H_\infty^{M'}(A|B)$, hence Condition (iv) is satisfied.

It remains to show that M'' satisfies ϵ-differential privacy (condition (iii)).

Since $d(i, i') = 1$, from the triangular inequality we have:

$$d(i', j) - 1 \leq d(i, j) \leq d(i', j) + 1$$

Thus, there are 3 possible cases:

1. $d(i, j) = d(i', j)$

   The result is immediate since $M''_{i,j} = M''_{i',j}$.

2. $d(i, j) = d(i', j) - 1$

   We define the set of neighbors of h "one step further away" from k:

   $$\mathcal{F}_{h,k} = \{h' \sim h \mid h' \in \mathcal{A}_{\langle d(h,k)+1 \rangle}(k)\}$$

   Equivalently, we can see $\mathcal{F}_{h,k}$ as the set of neighbors of h that are not at distance $d(h, k) - 1$ or $d(h, k)$ from k. Since the graph is distance-regular, the number of elements in this set is given by the intersection number $b_{d(h,k)}$, i.e., $|\mathcal{F}_{h,k}| = b_{d(h,k)}$. The following inequalities hold for any $h, h' \in \mathcal{A}$:

   $$\begin{aligned} M'_{h,k} &\leq e^\epsilon M'_{h',k} && \forall h' \in \mathcal{F}_{h,k} && \text{(diff. privacy)} \Rightarrow \\ b_{d(h,k)} M'_{h,k} &\leq e^\epsilon \sum_{h' \in \mathcal{F}_{h,k}} M'_{h',k} && && \text{(sum of the above)} \end{aligned}$$

   Fix now a distance d and sum the above inequalities for all vertices at distance d from k:

   $$\sum_{h \in \mathcal{A}_{\langle d \rangle}(k)} b_d M'_{h,k} \leq e^\epsilon \sum_{h \in \mathcal{A}_{\langle d \rangle}(k)} \sum_{h' \in \mathcal{F}_{h,k}} M'_{h',k}$$

   Note that, being the graph distance-regular, and by the definition of the intersection number c_{d+1}, each $h' \in \mathcal{A}_{\langle d+1 \rangle}(k)$ is contained in $\mathcal{F}_{h,k}$ for exactly c_{d+1} different $h \in \mathcal{A}_{\langle d \rangle}(k)$. So the right-hand side above sums all vertices of $\mathcal{A}_{\langle d+1 \rangle}(k)$ exactly c_{d+1} times each. Thus we get that for all $k \in \mathcal{B}, d \in \Delta$:

   $$b_d \sum_{h \in \mathcal{A}_{\langle d \rangle}(k)} M'_{h,k} \leq e^\epsilon \; c_{d+1} \sum_{h \in \mathcal{A}_{\langle d+1 \rangle}(k)} M'_{h,k} \tag{21}$$

Finally, note that $c_{d+1}|\mathcal{A}_{\langle d+1\rangle}(i)| = b_d|\mathcal{A}_{\langle d\rangle}(i)|$ (both sides count the number of edges between a vertex at distance d and a vertex at distance $d+1$) and therefore we have

$$\frac{c_{d+1}}{b_d} = \frac{|\mathcal{A}_{\langle d\rangle}(i)|}{|\mathcal{A}_{\langle d+1\rangle}(i)|}$$

Now we pick $d = d(i,j)$ to conclude that

$$\begin{aligned}
M''_{i,j} &= \frac{1}{n|\mathcal{A}_{\langle d\rangle}(i)|}\sum_{k\in\mathcal{B}}\sum_{h\in\mathcal{A}_{\langle d\rangle}(k)} M'_{h,k} \\
&\le e^{\epsilon}\,\frac{1}{n|\mathcal{A}_{\langle d\rangle}(i)|}\frac{c_{d+1}}{b_d}\sum_{k\in\mathcal{B}}\sum_{h\in\mathcal{A}_{\langle d+1\rangle}(k)} M'_{h,k} \qquad \text{(from (21))} \\
&= e^{\epsilon}\,\frac{1}{n|\mathcal{A}_{\langle d+1\rangle}(i)|}\sum_{k\in\mathcal{B}}\sum_{h\in\mathcal{A}_{\langle d+1\rangle}(k)} M'_{h,k} \\
&= e^{\epsilon} M''_{i',j}.
\end{aligned}$$

3. $d(i,j) = d(i',j) + 1$

   This case is analogous to the case where $d(i,j) = d(i',j) - 1$.

□

Lemma 9. (Step 2*b*) *Consider a square channel matrix M' satisfying the assumptions of Lemma 8, except that we assume $(\mathcal{A},\sim)$ to be vertex-transitive instead of distance-regular. Then it is possible to transform M' into a matrix M'' with the same properties as in Lemma 8.*

Proof. Let Γ be the automorphism group of $(\mathcal{A},\sim)$. Since $(\mathcal{A},\sim)$ and $(\mathcal{B},\sim)$ are isomorphic, it follows that Γ is also the automorphism group of $(\mathcal{B},\sim)$. For all $i,j\in\mathcal{A}$ we define the elements of M'' as:

$$M''_{i,j} = \frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma} M'_{\sigma(i),\sigma(j)}$$

We now show that M'' satisfies Conditions (i) – (iv). We start with Condition (i):

$$\begin{aligned}
\sum_{j=0}^{n-1} M''_{i,j} &= \sum_{j=0}^{n-1}\frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma} M'_{\sigma(i),\sigma(j)} \\
&= \frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma}\sum_{j=0}^{n-1} M'_{\sigma(i),\sigma(j)} \\
&= \frac{1}{|\Gamma|}\sum_{\sigma\in\Gamma} 1 \qquad \text{(since } M_{\sigma(i),\cdot} \text{ is a prob. distribution)} \\
&= 1
\end{aligned}$$

Then we show that $M''_{i,i} = M''_{j,j}$ for all $i, j \in \mathcal{A}$.

$$\begin{aligned}
M''_{i,i} &= \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} M'_{\sigma(i),\sigma(i)} \\
&= \frac{1}{|\Gamma|} \sum_{k=0}^{n-1} \sum_{\sigma \in \Gamma_{i \mapsto k}} M'_{\sigma(i),\sigma(i)} && \text{(since } \Gamma = \bigcup_k \Gamma_{i \mapsto k}\text{)} \\
&= \frac{1}{|\Gamma|} \sum_{k=0}^{n-1} \sum_{\sigma \in \Gamma_{i \mapsto k}} M'_{k,k} && \text{(} \sigma(i) = k \text{ since } \sigma \in \Gamma_{i \mapsto k}\text{)} \\
&= \frac{1}{|\Gamma|} \sum_{k=0}^{n-1} M'_{k,k} |\Gamma_{i \mapsto k}| \\
&= \frac{1}{|\Gamma|} \sum_{k=0}^{n-1} M'_{k,k} \frac{|\Gamma|}{n} && \text{(by Lemma 33)} \\
&= \frac{1}{n} \sum_{k=0}^{n-1} M'_{k,k}
\end{aligned}$$

And we can conclude that every element in the diagonal of M'' is the same, as they are the average of the diagonal elements of M'. To fulfil condition (ii) we still need to show that $M''_{j,j} = \max_j^{M''}$ for all $j \in \mathcal{A}$.

$$\begin{aligned}
M''_{j,j} &= \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} M'_{\sigma(j),\sigma(j)} \\
&\geq \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} M'_{\sigma(h),\sigma(j)} && \text{(for all } h \in \mathcal{A}\text{, since all maxima} \\
&&& \qquad \text{are in the diagonal of } M'\text{)} \\
&= M''_{h,j}
\end{aligned}$$

Then we show that M'' provides ϵ-differential privacy (condition (iii)). Let $i, h \in \mathcal{A}$ such that $i \sim h$. Note that $\sigma(i) \sim \sigma(h)$ for all $\sigma \in \Gamma$ since σ is an automorphism. Thus for all $j \in \mathcal{A}$ we have:

$$\begin{aligned}
M''_{i,j} &= \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} M'_{\sigma(i),\sigma(j)} \\
&\leq \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} e^{\epsilon} M'_{\sigma(h),\sigma(j)} && \text{(}\epsilon\text{-diff. priv. of } M, \sigma(i) \sim \sigma(h)\text{)} \\
&= e^{\epsilon} M''_{h,j}
\end{aligned}$$

Finally, we show that $H_\infty^{M''}(A|B) = H_\infty^{M'}(A|B)$ (condition (iv)).

$$H_\infty^{M''}(A|B) = -\log \frac{1}{n} \sum_{i=0}^{n-1} M''_{i,i} \qquad \text{(the maxima of } M'' \text{ are in the diagonal and are all equal)}$$

$$
\begin{aligned}
&= -\log \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} M'_{\sigma(i),\sigma(i)} \\
&= -\log \frac{1}{n} \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} \sum_{i=0}^{n-1} M'_{\sigma(i),\sigma(i)} \\
&= -\log \frac{1}{n} \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} \sum_{i=0}^{n-1} M'_{i,i} && \text{(because } \sigma \text{ is an automorphism)} \\
&= -\log \frac{1}{n} \sum_{i=0}^{n-1} M'_{i,i} \\
&= H_\infty^{M'}(A|B) && \text{(the maxima of } M \text{ are also in the diagonal)}
\end{aligned}
$$

□

Proposition 11. *Let M be a channel matrix satisfying ϵ-differential privacy where for every $0 \leq i \leq n-1$ we have $M_{i,i} = max^M$. Then, for every node $i \in \mathcal{A}$,*

$$\max^M \leq \frac{1}{\sum_{d \in \Delta} \frac{|\mathcal{A}_{\langle d \rangle}(i)|}{e^{\epsilon d}}},$$

Proof. The elements of any given row i of M represent a probability distribution, so they sum up to 1: $\sum_j M_{i,j} = 1$. Then we can derive the following inequalities.

$$
\begin{aligned}
\sum_j \left(\frac{\max^M}{e^{\epsilon d(i,j)}} \right) &\leq 1 && \text{(by (5))} \\
\sum_{d \in \Delta} \left(|\mathcal{A}_{\langle d \rangle}(i)| \frac{\max^M}{e^{\epsilon d}} \right) &\leq 1 && \text{(grouping elements by distance)}
\end{aligned}
$$

And from the above we can conclude.

$$\max^M \leq \frac{1}{\sum_{d \in \Delta} \frac{|\mathcal{A}_{\langle d \rangle}(i)|}{e^{\epsilon d}}}$$

□

Theorem 12. *Consider a channel matrix M satisfying ϵ-differential privacy for some $\epsilon \geq 0$, assume that the probability distribution on A is uniform, and that $(\mathcal{A}, \sim)$ is either distance-regular or vertex-transitive. Then*

$$H_\infty^M(A|B) \geq -\log \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}}. \tag{6}$$

In case the matrix M can be transformed via Lemmata 7, 8, and 9 into a matrix M'' whose elements all satisfy (5) with equality, then the bound (6) holds with equality.

Proof. The inequality follows directly from (3) and Proposition 11.

To prove the second part of the result, note that if a (transformed) matrix M'' is such that all of its elements satisfy (5) with equality, then trivially condition bound (6) also holds with equality. Recalling that the transformations given by Lemmata 7, 8, and 9 do not change the a posterior min-entropy of the channel, if M'' satisfies (6) with equality, so does M. □

A.4 Application to leakage (Section 5)

Proposition 13. *If $v \geq 2$, the graph $(\mathcal{V}^u, \sim)$ is a connected distance-regular graph with diameter $d_{max} = u$, and intersection numbers $b_d = (u-d)(v-1)$ and $c_d = d$, for all $0 \leq d \leq d_{max}$.*

Proof. The vertices of $(\mathcal{V}^u, \sim)$ are u-tuples $(v_1, \ldots, v_u)$, $v_i \in \mathcal{V}$ and two vertices are adjacent if, and only if, they differ in exactly one element v_i. Then the distance between two vertices is the number of elements in which they differ. Let $x_1, x_2 \in \mathcal{V}^u$ with $d(x_1, x_2) = d$, so they differ in exactly d elements. To go at distance $d+1$ from x_1 we can select any of the remaining $u-d$ elements and change it in $v-1$ possible ways, so the total number is $(u-d)(v-1)$, which only depends on d, not on x_1, x_2. Similarly, by changing one of the differing elements of x_2 to match the value of x_1 we get a vertex at distance $d-1$, and there are d such elements. □

Proposition 14. *The graph $(\mathcal{V}^u, \sim)$ is a vertex-transitive graph.*

Proof. The Hamming graph is the Cartesian product of u complete graphs of size v. Complete graphs are trivially vertex-transitive and it is known [20] that a Cartesian product is vertex-transitive if, and only if, each of its factors is so. □

Theorem 15. *If $\mathcal{K}$ satisfies ϵ-differential privacy, then the information leakage is bounded from above as $I_\infty(X;Z) \leq Bnd(u, v, \epsilon)$.*

Proof. We start by deriving the following.

$$\begin{aligned}
\sum_{d\in\Delta} \frac{n_d}{e^{\epsilon d}} &= \sum_{d\in\Delta} \frac{\binom{u}{d}(v-1)^d}{e^{\epsilon d}} && \text{(by (7))} \\
&= \frac{1}{e^{\epsilon u}} \sum_{d\in\Delta} \binom{u}{d} (v-1)^d e^{\epsilon(u-d)} && \text{(multiplying each term by } e^{\epsilon u}/e^{\epsilon u}\text{)} \\
&= \frac{(v-1+e^\epsilon)^u}{e^{\epsilon u}} && \text{(by binomial expansion)} \qquad (22)
\end{aligned}$$

And we can use the above in Theorem 12 as follows.

$$\begin{aligned} H_\infty^M(A|B) &\geq -\log \frac{1}{\sum_{d\in\Delta} \frac{n_d}{e^{\epsilon\, d}}} && \text{(by Theorem 12)} \\ &= -\log \frac{1}{(v-1+e^\epsilon)^u / e^{\epsilon u}} && \text{(by (22))} \\ &= -\log \left(\frac{e^\epsilon}{v-1+e^\epsilon}\right)^u && (23) \end{aligned}$$

Since by hypothesis the input distribution is uniform, the min-entropy of X is maximum, i.e., $H_\infty(X) = \log|\mathcal{V}^u|$. We then use this fact to derive that

$$\begin{aligned} I_\infty^M(X;Y) &= H_\infty(X) - H_\infty^M(X|Y) && \text{(by definition)} \\ &\leq \log v^u + \log \left(\frac{e^\epsilon}{v-1+e^\epsilon}\right)^u && \text{(by (23))} \\ &= u \log \frac{v\, e^\epsilon}{v-1+e^\epsilon} \end{aligned}$$

□

Proposition 16. *For every u, v, and ϵ it is possible to define the mechanism $\mathcal{K}$ below, which provides ϵ-differential privacy and whose min-entropy leakage, for the uniform input distribution, is $I_\infty(X;Z) = Bnd(u,v,\epsilon)$.*

$$\mathcal{K}_{x,z} = \frac{2^{Bnd(u,v,\epsilon)}}{v^u (e^\epsilon)^d}, \quad \textit{for every input } x \textit{ and output } z.$$

Proof. The adjacency relation in $\mathcal{X}$ determines a graph structure $(\mathcal{X}, \sim)$. Let $\mathcal{Z} = \mathcal{X}$ and define the matrix of $\mathcal{K}$ as follows, where $d = d(x,z)$:

$$p(z|x) = \frac{2^{Bnd(u,v,\epsilon)}}{v^u (e^\epsilon)^d} \tag{24}$$

We need to show that $p(\cdot|x)$ is a probability distribution for every x:

$$\begin{aligned} \sum_{z\in\mathcal{Z}} \frac{2^{Bnd(u,v,\epsilon)}}{v^u (e^\epsilon)^d} &= \frac{2^{Bnd(u,v,\epsilon)}}{v^u} \sum_{z\in\mathcal{Z}} \frac{1}{(e^\epsilon)^d} \\ &= \frac{2^{Bnd(u,v,\epsilon)}}{v^u} \sum_{d\in\Delta} \frac{n_d}{(e^\epsilon)^d} \\ &= \frac{2^{Bnd(u,v,\epsilon)}}{v^u} \cdot \frac{1}{\max^M} && \text{(by Proposition 11)} \\ &= \frac{2^{Bnd(u,v,\epsilon)}}{v^u} \cdot \frac{v^u e^0}{2^{Bnd(u,v,\epsilon)}} && \text{(max occurs when } d=0 \text{ in (24))} \\ &= 1 \end{aligned}$$

We now show that $\mathcal{K}$ provides ϵ-differential privacy. For every $x, x' \in \mathcal{X}$ such that $x \sim x'$, and for every $z \in \mathcal{Z}$ we have:

$$\begin{aligned}
\frac{p(z|x)}{p(z|x')} &= \frac{2^{Bnd(u,v,\epsilon)}}{v^u (e^\epsilon)^{d(x,z)}} \cdot \frac{v^u (e^\epsilon)^{d(x',z)}}{2^{Bnd(u,v,\epsilon)}} && \text{(by (24))} \\
&= e^{\epsilon(d(x',z)-d(x,z))} \\
&\leq e^{\epsilon d(x,x')} && \text{(by the triangle inequality)} \\
&= e^\epsilon && (d(x,x') = 1 \text{ since } x \sim x')
\end{aligned}$$

Finally, let us calculate $I_\infty(X;Z) = H_\infty(X) - H_\infty(X|Z)$. Since input distribution is uniform, we have:

$$H_\infty(X) = -\log \frac{1}{v^u}$$

Moreover, we know from (3) that $H_\infty(X|Z) = -\log \max^{M''}$, and to obtain the maximum value we take $d = 0$ in (24):

$$\begin{aligned}
H_\infty(X|Z) &= -\log \frac{2^{Bnd(u,v,\epsilon)}}{v^u e^0} \\
&= -\log \frac{2^{Bnd(u,v,\epsilon)}}{v^u}
\end{aligned}$$

Now, by subtracting the value of $H_\infty(X|Z)$ from the value of $H_\infty(X)$ we obtain $I_\infty(X;Z) = Bnd(u,v,\epsilon)$.

□

Lemma 19. *Let $\mathcal{K}$ be a mechanism with input X, where $\mathcal{X} = \mathcal{V}^u$, providing ϵ-differential privacy. Assume that $r = |Range(\mathcal{K})| = v^\ell$, for some $\ell < u$. Let M be the matrix associated with $\mathcal{K}$. Then it is possible to build a square matrix M' of size $v^\ell \times v^\ell$, with row and column indices in $\mathcal{A} \subseteq \mathcal{X}$, and a binary relation $\sim' \subseteq \mathcal{A} \times \mathcal{A}$ such that $(\mathcal{A}, \sim')$ is isomorphic to $(\mathcal{V}^\ell, \sim_\ell)$, and such that:*

(i) M' is a channel matrix: $\sum_{j=0}^{m-1} M'_{i,j} = 1$ for all $0 \leq i \leq n-1$;

(ii) $M'_{i,j} \leq (e^\epsilon)^{u-l+d} M'_{h,j}$ for all $i, h \in \mathcal{X}$ and $j \in \mathcal{Z}$, where d is the $\sim'$-distance between i and h;

(iii) the elements of the diagonal are all equal to the maximum element of the matrix: $M'_{i,i} = \max^{M'}$ for all $i \in \mathcal{X}$; and

(iv) $H_\infty^{M'}(X|Z) = H_\infty^{M}(X|Z)$, if X has the uniform distribution.

Proof. (Sketch) We first apply a procedure similar to that of Lemma 7 to construct a square matrix of size $v^\ell \times v^\ell$ which has the maximum values of each column in the diagonal. (In this case we construct an injection from the columns to rows containing their maximum value, and we eliminate the rows that at the end are not associated with any column.) Then define $\sim'$ as the projection of $\sim_u$ on $\mathcal{V}^\ell$, which satisfies condition (ii). Finally, apply the procedure in Lemma 8, or equivalently the procedure in Lemma 9, on the structure $(\mathcal{X}, \sim')$ to make all elements in the diagonal equal to the maximum element of the matrix (condition (iii)). Note that this procedure preserves the property of condition (ii), and conditional min-entropy (condition (iv)). Also the matrix obtained is a valid channel matrix (condition (i)). □

Proposition 20. *Let $\mathcal{K}$ be a mechanism with associated channel matrix M, and let $r = |Range(\mathcal{K})|$. If $\mathcal{K}$ provides* ϵ-differential privacy *then the min-entropy leakage associated with $\mathcal{K}$ is bounded from above as follows:*

$$I_\infty^M(X;Z) \;\leq\; \log \frac{r\,(e^\epsilon)^u}{(v-1+e^\epsilon)^\ell - (e^\epsilon)^\ell + (e^\epsilon)^u},$$

where $\ell = \lfloor \log_v r \rfloor$.

Proof. We first consider the case where $r = v^\ell$ for some ℓ. We transform the matrix M associated with $\mathcal{K}$ by applying Lemma 19, and let M' be the resulting matrix. Let $\max^{M'}$ be the value of every element in the diagonal of M', i.e., $\max^{M'} = M'_{i,i}$ for every row i, and let $\mathcal{A}'_{\langle d \rangle}(i)$ be the set of elements whose $\sim'$-distance from i is d. Note that for every $j \in \mathcal{A}'_{\langle d \rangle}(i)$ we have that $M'_{j,j} \leq M'_{i,j}(e^\epsilon)^{u-\ell+d}$, hence

$$M'_{i,j} \geq \frac{\max^{M'}}{(e^\epsilon)^{u-\ell+d}}$$

Furthermore each element j at $\sim'$-distance d from i can be obtained by changing the value of d individuals in the ℓ-tuple representing i (remember that $(\mathcal{A}, \sim')$ is isomorphic to $(\mathcal{V}^\ell, \sim_\ell)$). We can choose those d individuals in $\binom{\ell}{d}$ possible ways, and for each of these individuals we can change the value (with respect to the one in i) in $v-1$ possible ways. Therefore

$$|\mathcal{A}'_{\langle d \rangle}(i)| = \binom{\ell}{d}(v-1)^d$$

Taking into account that for $M'_{i,i}$ we need not to divide by $(e^\epsilon)^{u-\ell+d}$, we obtain:

$$\max^{M'} + \textstyle\sum_{d=1}^{\ell} \binom{\ell}{d}(v-1)^d \frac{\max^{M'}}{(e^\epsilon)^{u-\ell+d}} \;\leq\; \sum_j M'_{i,j}$$

Since each row represents a probability distribution, the elements of row i must sum up to 1. Hence:

$$\max^{M'} + \textstyle\sum_{d=1}^{\ell} \binom{\ell}{d}(v-1)^d \frac{\max^{M'}}{(e^\epsilon)^{u-\ell+d}} \;\leq\; 1 \tag{25}$$

Simple calculations, similar to those of the proof of Theorem 15, give:

$$\max^{M'} \leq \frac{(e^\epsilon)^u}{(v-1+e^\epsilon)^\ell-(e^\epsilon)^\ell+(e^\epsilon)^u}$$

Therefore:

$$I_\infty^{M'}(X;Z) = H_\infty(X) - H_\infty^{M'}(X|Z) \qquad \text{(by definition)} \tag{26}$$

$$= \log v^u + \log \sum_{j=1}^{v^\ell} \max^{M'} \tfrac{1}{v^u} \tag{27}$$

$$= \log v^u + \log \frac{1}{v^u} + \log(v^\ell \max^{M'}) \tag{28}$$

$$\leq \log \frac{v^\ell\,(e^\epsilon)^u}{(v-1+e^\epsilon)^\ell - (e^\epsilon)^\ell + (e^\epsilon)^u} \qquad \text{(by (25))} \tag{29}$$

Consider now the case in which r is not of the form v^ℓ. Let ℓ be the maximum integer such that $v^\ell < r$, and let $m = r - v^\ell$. Transform the matrix M associated with $\mathcal{K}$ by collapsing the m columns with the smallest maxima into the m columns with highest maxima. I.e., let $j_1, j_2, \ldots, j_m$ the indices of the columns which have smallest maxima values, i.e., $\max_{j_t}^M \leq \max_j^M$ for every column $j \neq j_1, j_2, \ldots, j_m$. Similarly, let $k_1, k_2, \ldots, k_m$ be the indices of the columns with maximum values. Recalling the definition of the "collapsing" operator introduced in Section 4.1, we define

$$N = M[j_1 \to k_1][j_2 \to k_2] \ldots [j_m \to k_m]$$

Finally, eliminate the m all-zero columns to obtain a matrix with exactly v^ℓ columns. It is easy to show that

$$I_\infty^M(X;Z) \leq I_\infty^N(X;Z)\frac{r}{v^\ell}$$

After transforming N into a matrix M' with the same min-entropy leakage as described in the first part of this proof, from (26) we conclude

$$I_\infty^M(X;Z) \leq I_\infty^{M'}(X;Z)\frac{r}{v^\ell} \leq \log \frac{r\,(e^\epsilon)^u}{(v-1+e^\epsilon)^\ell - (e^\epsilon)^\ell + (e^\epsilon)^u}$$

□

Proposition 22. *Assume that $\mathcal{K}$ satisfies ϵ-differential privacy. Then the information leakage for an individual is bounded from above by*

$$I_\infty^{x^-}(X_i;Z) \leq \log \frac{v\,e^\epsilon}{v-1+e^\epsilon}.$$

Proof. Fix a database x, and a particular individual i in $\mathcal{U}$. The possible ways in which we can change the value of i in x are $v-1$. All the new databases obtained in this way are adjacent to each other, i.e., the graph structure associated with

the input is a clique of v nodes. Recall that n_d is the number of elements of the input at distance d from a given element x. In this case we have

$$n_d = \begin{cases} 1 & \text{for } d = 0, \\ v - 1 & \text{for } d = 1, \\ 0 & \text{otherwise.} \end{cases}$$

By substituting this value of n_d in Theorem 12, we get

$$\begin{aligned} H_\infty^{x^-}(X_i|Z) &\geq -\log \frac{1}{1 + \dfrac{v-1}{e^\epsilon}} \\ &= -\log \frac{e^\epsilon}{v - 1 + e^\epsilon} \end{aligned} \tag{30}$$

The particular individual can present v different values, and thus in the case the input distribution is uniform its min-entropy is $H_\infty^{x^-}(X_i) = \log v$.

$$\begin{aligned} I_\infty^{x^-}(X_i; Z) &= H_\infty^{x^-}(X_i) - H_\infty^{x^-}(X_i|Y) && \text{(by definition)} \\ &\leq \log v + \log \frac{e^\epsilon}{v - 1 + e^\epsilon} && \text{(by (30))} \\ &= \log \frac{v\, e^\epsilon}{v - 1 + e^\epsilon} \end{aligned}$$

Since the min-entropy leakage is maximum in the case of the uniform input distribution, the result follows.

□

Proposition 21. *Let $(\mathcal{X}, \mathcal{Z}, M)$ be a channel, with associated input random variable X and output random variable Z, such that for all $x, x' \in \mathcal{X}$ and all $z \in \mathcal{Z}$, there is an $\epsilon \geq 0$ such that the conditional probabilities of M satisfy $p(z|x)/p(z|x') \leq e^\epsilon$. Then the min-entropy leakage from this channel is bounded by*

$$I_\infty(X; Z) \leq \log e^\epsilon.$$

Proof. First let us calculate the a posteriori min-entropy of the channel.

$$
\begin{aligned}
-H_\infty(X|Z) &= \log \sum_z p(z) \max_x p(x|z) && \text{(by definition)} \\
&= \log \sum_z \max_x p(z)p(x|z) \\
&= \log \sum_z \max_x p(x)p(z|x) \\
&\leq \log \sum_z \max_x p(x)e^\epsilon p(z|\hat{x}) && \text{(for any fixed } \hat{x} \in \mathcal{X}\text{, by } \epsilon\text{-d.p.)} \\
&= \log \sum_z e^\epsilon p(z|\hat{x}) \max_x p(x) \\
&= \log \left(e^\epsilon \max_x p(x) \sum_z p(z|\hat{x}) \right) \\
&= \log \left(e^\epsilon \max_x p(x) \right) \\
&= \log e^\epsilon + \log \max_x p(x) \\
&= \log e^\epsilon - H_\infty(X) && \text{(by definition of min-entropy)}
\end{aligned} \tag{31}
$$

By substituting (31) in the definition of min-entropy leakage, we conclude the proof.

$$
\begin{aligned}
I_\infty(X;Z) &= H_\infty(X) - H_\infty(X|Z) \\
&\leq H_\infty(X) + \log e^\epsilon - H_\infty(X) \\
&= \log e^\epsilon
\end{aligned}
$$

□

A.5 Application to utility (Section 6)

Proposition 24. *Assuming that the data analyst uses a remapping function* $guess : \mathcal{Z} \to \mathcal{Y}$*, we have*

$$
\mathcal{U}(Y,Z) = \sum_{y,z} p(y,z) gain(y, guess(z)). \tag{12}
$$

Proof. Let δ_x represent the probability distribution which has value 1 on x and

0 elsewhere.

$$\begin{aligned}
\mathcal{U}(Y,Z) &= \sum_y p(y) \sum_{y'} p(y'|y) gain(y,y') && \text{(by (11))} \\
&= \sum_y p(y) \sum_{y'} \left(\sum_z p(z|y) p(y'|z) \right) gain(y,y') \\
&= \sum_y p(y) \sum_{y'} \left(\sum_z p(z|y) \delta_{y'}(guess(z)) \right) gain(y,y') && \text{(as } y' = guess(z)\text{)} \\
&= \sum_y p(y) \sum_z p(z|y) \sum_{y'} \delta_{y'}(guess(z)) gain(y,y') \\
&= \sum_y p(y) \sum_z p(z|y) gain(y, guess(z)) \\
&= \sum_{y,z} p(y,z) gain(y, guess(z)) && (32)
\end{aligned}$$

□

Proposition 25. *Assume that function gain is the identity and the function guess is optimal. Then:*

$$\mathcal{U}(Y,Z) = \sum_z \max_y (p(y)\, p(z|y)) = V(Y|Z) = 2^{-H_\infty(Y|Z)}.$$

Proof. Just substitute (15) in the definition of conditional min-entropy: $H_\infty(Z \mid Y) = -\log \sum_z \max_y ((p(y)\, p(z|y))$. □

Proposition 27. *In an oblivious setting (cf. Figure 4), if the query function f is deterministic, then the mechanism $\mathcal{K}$ satisfies ϵ-differential privacy with respect to $\mathcal{X}$ if, and only if, the noise channel $\mathcal{H}$ satisfies ϵ-differential privacy with respect to $\mathcal{Y}$.*

Proof. The probability of $\mathcal{K}$ producing an arbitrary reported answer z given an arbitrary database x can be calculated as

$$\begin{aligned}
\mathcal{K}_{x,z} &= \sum_y p(y|x) \cdot \mathcal{H}_{y,z} && \text{(since the mechanism is oblivious)} \\
&= p(f(x)|x) \cdot \mathcal{H}_{f(x),z} && \text{(since } y = f(x) \text{ is deterministic)} \\
&= \mathcal{H}_{f(x),z} && \text{(since } p(f(x)|x) = 1\text{)}
\end{aligned}$$

Hence, it follows immediately that $\frac{\mathcal{K}_{x,z}}{\mathcal{K}_{x',z}} \leq e^\epsilon$ if, and only if, $\frac{\mathcal{H}_{f(x),z}}{\mathcal{H}_{f(x'),z}} \leq e^\epsilon$.

□

Theorem 28. *Consider a noise channel $\mathcal{H}$ satisfying ϵ-differential privacy for some $\epsilon > 0$. Assume that the distribution of Y is uniform and that $(\mathcal{Y}, \sim)$ is either distance-regular or vertex-transitive. Then we have:*

$$\mathcal{U}(Y, Z) \leq \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}} \tag{16}$$

Proof. Since $(\mathcal{Y}, \sim)$ is distance-regular or vertex-transitive, and the distribution on $\mathcal{Y}$ is uniform, we can apply Theorem 12 to derive that $H_\infty^M(Z|Y) \geq -\log \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}}$. Then we just substitute this result in Proposition 25. □

Theorem 29. *Assume $(\mathcal{Y}, \sim)$ is distance-regular or vertex-transitive and that the distribution of Y is uniform. Then the matrix $\mathcal{H}$ defined by (17) is a channel matrix that satisfies ϵ-differential privacy and has maximal utility:*

$$\mathcal{U}(Y, Z) = \frac{1}{\sum_{d \in \Delta} \frac{n_d}{e^{\epsilon\, d}}}$$

Proof. First we prove that the matrix as defined in (17) is a channel matrix, i.e., that each row is a probability distribution.

$$\begin{aligned}
\sum_{j \in \mathcal{Z}} \mathcal{H}_{i,j} &= \sum_{j \in \mathcal{Z}} \frac{\gamma}{e^{\epsilon d(i,j)}} \\
&= \gamma \sum_{j \in \mathcal{Z}} \frac{1}{e^{\epsilon d(i,j)}} \\
&= \gamma \sum_{d \in \Delta} \frac{n_d}{e^{\epsilon d}} && \text{(by (18))} \\
&= \gamma \frac{1}{\gamma} \\
&= 1
\end{aligned}$$

Now we show that the mechanism respects ϵ-differential privacy. For every i, i' such that $i \sim i'$ and every j:

$$\begin{aligned}
\frac{\mathcal{H}_{i,j}}{\mathcal{H}_{i',j}} &= \frac{\gamma}{e^{\epsilon d(i,j)}} \cdot \frac{e^{\epsilon d(i',j)}}{\gamma} && \text{(by definition of } \mathcal{H}) \\
&= \frac{e^{\epsilon d(i',j)}}{e^{\epsilon d(i,j)}} \\
&= e^{\epsilon (d(i',j) - d(i,j))} \\
&\leq e^{\epsilon d(i',i)} && \text{(by the triangular inequality)} \\
&= e^{\epsilon} && \text{(since } i \sim i',\ d(i,i') = 1)
\end{aligned}$$

Finally, we show that the utility is maximum.

$$
\begin{aligned}
\mathcal{U}(Y,Z) &= \sum_{z\in\mathcal{Z}} \max_y (p(y)\,\mathcal{H}(z|y)) && \text{(by (15))}\\
&= \sum_{z\in\mathcal{Z}} \max_y \frac{1}{|\mathcal{Y}|}\mathcal{H}(z|y) && \text{(since } Y \text{ is uniform)}\\
&= \frac{1}{|\mathcal{Y}|}\sum_{z\in\mathcal{Z}} \max_y \frac{\gamma}{\min_d e^{\epsilon d(z,y)}} && \text{(by (17))}\\
&= \frac{1}{|\mathcal{Y}|}\sum_{z\in\mathcal{Z}} \gamma && \text{(maximum is } d=0)\\
&= \frac{1}{|\mathcal{Y}|}\cdot|\mathcal{Z}|\gamma\\
&= \gamma && \text{(since } |\mathcal{Y}|=|\mathcal{Z}|=n)
\end{aligned}
$$

□