Bayesian Leakage Analysis: A Framework for Analyzing Leakage in Encrypted Search

Seny Kamara[*]
MongoDB & Brown University

Tarik Moataz[†]
MongoDB

Abstract

Sub-linear encrypted search algorithms (ESA) are highly efficient search algorithms that operate on end-to-end encrypted data. ESAs can be built using a variety of cryptographic primitives and can achieve different trade-offs between efficiency, expressiveness and leakage. Since the introduction of ESAs, cryptographers have focused on both minimizing and attacking their leakage but an important open problem in the field has been to provide a theoretical framework with which leakage can be analyzed and better understood.

In this work, we propose such a framework. We model leakage profiles as Bayesian networks and capture leakage attacks as statistical inference algorithms on these networks. We then formalize a notion we call *coherence* which, roughly speaking, captures the quality of the inference given some observed leakage and an auxiliary distribution. In this work, we focus on partial and full query recovery attacks, though our framework can be extended to capture data recovery attacks as well.

We then use our framework to study the coherence of two common leakage patterns—the query equality pattern and the volume pattern—against two well-known and powerful statistical inference techniques. In each case, we provide generic bounds on the coherence in the sense that they apply to arbitrary query and auxiliary distributions and concrete analyses for specific pairs of query and auxiliary distributions.

[*] `seny.kamara@mongodb.com`.

[†] `tarik.moataz@mongodb.com`.

Contents

1 Introduction **3**

2 Related Work **6**

3 Preliminaries **8**

3.1 Bayesian Networks . . . 8

4 Definitions **9**

5 Partial Recovery Against Query Equality **10**

5.1 Uniform Queries with Uniform Auxiliary . . . 11

5.2 Zipf Queries with Uniform Auxiliary . . . 11

5.3 Discussion . . . 12

6 Full Recovery Against Query Equality **13**

6.1 Uniform Queries and Uniform Auxiliary . . . 13

6.2 Uniform Queries with Zipf Auxiliaries . . . 14

6.3 Zipf Queries with Zipf Auxiliaries . . . 15

6.4 Uniform Queries with Distinct Uniform Auxiliary . . . 15

6.5 Zipf Queries with Distinct Zipf Auxiliaries . . . 16

6.6 Discussion . . . 17

7 Full Recovery Against Volume **18**

7.1 Uniform Data and Queries with Uniform Auxiliaries . . . 20

7.2 Zipf Data and Queries with Zipf Auxiliaries . . . 21

7.3 Zipf Data and Queries with Distinct Zipf Query Auxiliary and Uniform Data Auxiliary 21

7.4 Discussion . . . 22

8 Full Recovery Against Query Equality and Volume **23**

8.1 Uniform Data and Queries with Uniform Auxiliaries . . . 23

A Proofs for Section 5 (Partial Recovery Against Query Equality) **29**

B Proofs for Section 6 (Full Recovery Against Query Equality) **33**

C Proofs for Section 7 (Full Recovery Against Volume) **45**

D Proofs for Section 8 (Full Recovery Against Query Equality and Volume) **54**

E A Useful Lemma **56**

1 Introduction

Sub-linear encrypted search algorithms (ESA) are highly-efficient search algorithms that can be executed on end-to-end encrypted data. ESAs are the core building block in the design of a variety of end-to-end encrypted systems including encrypted cloud storage [12, 12, 47, 18] and encrypted databases [2, 31, 35, 54, 34, 44]. Sub-linear ESAs can be designed based on a variety of cryptographic primitives including structured encryption (STE), oblivious RAM (ORAM) and property-preserving encryption (PPE). Intuitively, an ESA should reveal no partial information about the data and/or queries but all sub-linear ESAs leak some information. This leakage is typically captured with a *leakage profile* that formally and precisely describes what the scheme reveals. While leakage profiles have proven to be an important conceptual and analytical tool, they are purely descriptive and do not (and are not meant to) provide any explanatory value.

The presence of leakage in ESAs has motivated several complimentary research agendas: *leakage cryptanalysis* which focuses on the design of attacks that try to exploit various leakage profiles in order to recover information about the data and/or queries [26, 45, 11, 8, 38, 42, 24, 25, 40, 41]; *leakage suppression* which focuses on techniques to design low- and even zero-leakage (sub-linear) ESAs [33, 32, 20, 48, 5]; and *leakage quantification* which focuses on quantifying the information revealed by a given leakage profile [28, 27, 39].

Auxiliary information. The challenge of formulating a useful "theory of leakage" has been open since [15, 14] first proposed the use of leakage profiles in the analysis of sub-linear ESAs. The main conceptual challenges in developing such a theory, however, has been that leakage profiles on their own do not capture everything that an adversary knows. In particular, they do not capture the adversary's auxiliary information about the query and/or data distribution which is a critical part of any inference attack. So while it is relatively straightforward to develop a framework to analyze leakage profiles as stand-alone objects (i.e., without properly integrating auxiliary information), such a framework would not provide any useful insights as to whether a profile is exploitable or not—which is ultimately the question we are interested in.

Formalizing attacks. In this work, we propose a new theoretical framework to analyze leakage profiles and their vulnerability to inference attacks. More precisely, our framework provides: (1) a new graphical/visual and intuitive representation of leakage profiles; (2) analytical techniques to bound the success probability of powerful classes of attacks; and (3) a natural way to incorporate the auxiliary information available to an adversary.

At a very high level, our framework makes use of a type of probabilistic graphical model called Bayesian networks to formalize and analyze leakage profiles. Adversaries are then modeled as Bayesian inference algorithms. More precisely, an inference attack can be viewed as a concrete instantiation of the following statistical inference problem. Let $\mathbf{X} = (\mathbf{D}, \mathbf{Q}, \mathbf{L})$ be a multi-variate random variable that consists of $\mathbf{D}$ which outputs a dataset, $\mathbf{Q}$ which outputs a query sequence and $\mathbf{L}$ which outputs leakage. In addition, let $\mathbf{A}$ be an adversary's prior over $\mathbf{Q}$ which is determined by its auxiliary information. An inference attack is an inference algorithm that, given an instantiation $\boldsymbol{\ell}$ of $\mathbf{L}$ and knowledge of $\mathbf{A}$, infers information about $\mathbf{Q}$ and/or $\mathbf{D}$. In most cases, the goal of the attack is to recover the instantiations $\boldsymbol{d}$ and/or $\boldsymbol{\ell}$ of $\mathbf{D}$ and $\mathbf{L}$, respectively. If the target of the attack is the query sequence then we refer to it as a *query recovery attack* whereas if the target is the data then we call it a *data recovery attack*. If the goal is to recover these instantiations in whole then the attack is a *full recovery attack* but if the goal is to recover a part of the instantiations then it is a *partial recovery attack*. If the goal is only to approximate the instantiations then we call it an *approximation attack*. While our framework can be used to analyze all of the previously-mentioned attack types, in this work we focus on full and partial query recovery attacks. [1]

Modeling leakage. In our framework, a leakage profile is represented as a Bayesian network which is a graphical representation of a multi-variate random variable and of the conditional dependencies among its variables. Our use of Bayesian networks has several advantages: (1) it provides us with a visual and intuitive representation of a leakage profile's "dependency structure"; and (2) it allows us to exploit this structure to derive bounds on an attack's success probability. More precisely, given a multi-variate random variable $\mathbf{X} = (\mathbf{D}, \mathbf{Q}, \mathbf{L})$ as above, a leakage profile is represented as a Bayesian network $\mathcal{N}_{\mathbf{X}}$ over $\mathbf{X}$ which we refer to as a *leakage network*. Intuitively, the usefulness of Bayesian networks is derived from the *Bayesian chain rule* which uses the dependency structure of the network to simplify the joint probability of $\mathbf{X}$ which in turn simplifies computations of the marginal distributions. In our setting, the dependency structure of the network is determined by the leakage profile so one can already see how some profiles may lead to harder or easier inference problems. Given a Bayesian network one can use analytical or computational tools to carry out different kinds of inference tasks on the marginals. As discussed above, the task we focus on this work is full and partial query recovery, i.e., inferring/estimating the instantiation $\boldsymbol{q}$ of $\mathbf{Q}$ given an instantiation $\boldsymbol{\ell}$ of $\mathbf{L}$.

Coherence. Notice that, so far, we have not incorporated auxiliary information. To do so, we consider an auxiliary distribution $\mathbf{A}$ that captures the information an adversary derives about $\boldsymbol{q}$ from its auxiliary information.[2] Given a leakage network, our goal is to bound the probability that an adversary recovers information about $\boldsymbol{q}$ given some leakage $\boldsymbol{\ell}$ derived from $\boldsymbol{q}$ and its auxiliary distribution $\mathbf{A}$. Here, recovering information about $\boldsymbol{q}$ is modeled as computing a *recovery function* g over $\boldsymbol{q}$. We formalize this intuition with a notion we call *coherence* [3] and study the coherence of various leakage networks against two unbounded adversaries we call the hypothesis adversary $\mathcal{A}_{\mathsf{hyp}}$ and the MAP adversary $\mathcal{A}_{\mathsf{map}}$.

$\mathcal{A}_{\mathsf{hyp}}$ is a partial query recovery adversary that, given $\boldsymbol{\ell}$, $\mathbf{A}$ and a recovery function g, outputs the element in the co-domain of g with the highest-probability pre-image with respect to the posterior distribution $\Pr\left[\,\mathbf{A} = \boldsymbol{q} \,\middle|\, \mathbf{L} = \boldsymbol{\ell}\,\right]$. In statistical terms, this corresponds to the MAP test in hypothesis testing where the hypotheses are of the form $H_s : g(\boldsymbol{q}) = s$, with $g : \mathbb{Q}^n \to \mathbb{S}$. $\mathcal{A}_{\mathsf{map}}$ is a full query recovery adversary that, given $\boldsymbol{\ell}$ and $\mathbf{A}$, returns the *maximum a-posteriori* (MAP) estimate which is the query sequence that maximizes the posterior distribution $\Pr\left[\,\mathbf{A} = \boldsymbol{q} \,\middle|\, \mathbf{L} = \boldsymbol{\ell}\,\right]$, viewed here as a function of $\boldsymbol{q}$. The MAP estimate is optimal in settings like full query recovery where the inferred sequence must be exactly the same as the instantiation $\boldsymbol{q}$ of $\mathbf{Q}$.[4] Notice, however, that $\mathcal{A}_{\mathsf{map}}$ uses its *auxiliary* distribution $\mathbf{A}$ to estimate $\boldsymbol{q}$ and *not* the query distribution $\mathbf{Q}$ so the optimality of the MAP estimate doesn't necessarily hold. The MAP adversary $\mathcal{A}_{\mathsf{map}}$ is still justified, however, whenever the posterior of the auxiliary distribution $\Pr\left[\,\mathbf{A} = \boldsymbol{q} \,\middle|\, \mathbf{L} = \boldsymbol{\ell}\,\right]$ converges in probability to

[1] As discussed in Section 2, a subset of data recovery attacks have already been studied formally whereas, as far as we know, no theoretical model has every been proposed that can capture query recovery attacks.

[2] In Bayesian terms, the auxiliary distribution is the adversary's prior but we call it the auxiliary to avoid confusion with the prior distribution in Baye's rule and Bayesian networks.

[3] The term coherence here is meant to capture the "quality" or "meaning" of the inference.

[4] More precisely, the MAP estimate is optimal in the sense that it minimizes the expected 0-1 loss.

the posterior of the query distribution $\Pr\left[\,\mathbf{Q} = \boldsymbol{q} \,\middle|\, \mathbf{L} = \boldsymbol{\ell}\,\right]$ as the size of the observed leakage $\boldsymbol{\ell}$ grows. In other words, under this condition, $\mathcal{A}_{\mathsf{map}}$ is asymptotically optimal.

Leakage patterns studied. We show coherence Theorems for three common leakage patterns/profiles: (1) the query equality which reveals if and when two queries are for the same label; (2) the volume pattern (sometimes referred to as the response length) which reveals the size of query responses; and (3) the combination of query equality and volume patterns. We chose to study these patterns because almost every searchable and structured encryption scheme leaks the query equality and the volume pattern [50, 21, 15, 14, 37, 36, 17, 4, 13, 12, 16, 49] (note that this is not meant to be an exhaustive list). Note that the volume pattern is also leaked by any encrypted search algorithm based on oblivious RAM [22, 51].

Analysis and concrete examples. For each of these patterns/profiles, we show general Theorems that bound the coherence against either $\mathcal{A}_{\mathsf{hyp}}$ or $\mathcal{A}_{\mathsf{map}}$ (as appropriate) for *any* pair of query and auxiliary distributions, and a variety of concrete Theorems for specific pairs of query and auxiliary distributions. For example, we consider the uniform distribution as well as a power-law distribution to capture real-world settings. For the power-law distribution, we chose the Zipf distribution since it is known to capture many publicly-available datasets and query logs, as shown in [29]. We also consider settings where the space of the query, data and auxiliary distributions are distinct. As far as we know, this has never been considered formally but it is an important case to study because, in practice, adversaries do not necessary know the client's exact query space or the exact dataset. We briefly summarize some of our findings:

- the coherence of the query equality pattern against full query recovery attacks is very small, even when the size of the query space m and the sequence length n are small. While we do not consider every possible pair of query and auxiliary distribution, this suggests that full query attacks against query equality leakage might be challenging to mount. The coherence bounds are tighter and significantly smaller when the query and auxiliary distributions have distinct support (or query space). Informally, we were able to show—under some assumptions on m and n—that when the query and auxiliary distributions are Zipf-distributed, the coherence ε is very small; more precisely,
$$\varepsilon \lesssim \frac{n}{2 \cdot e^{n-1} \cdot (\ln(2))^{n+1}}.$$
As a point of reference, for $m = 800$ and $n = 200$, the coherence is upper bounded by 2^{-576}.
- the coherence of the volume pattern against full query recovery attacks is also very small in all cases except when the auxiliary distribution is Zipf-distributed.
- the coherence of the query equality pattern against partial query recovery attacks that attempt to test whether the query is a known value is mixed. We were able to show that, when the query and auxiliary distributions are uniform, there is a limited space for which the coherence is small. However, when the query distribution is Zipf-distributed and the auxiliary distribution is uniform, our bounds show that the coherence is larger which suggests that query equality pattern can be damaging with respect to an adversary that just wants to know if the query matches a known value.

Limitations and future work. While our framework already provides a new and powerful way to model and analyze leakage, the analysis carried out in this work has limitations and should be viewed as the first step in a longer term research agenda. The most immediate limitation of our work is that it only considers i.i.d. leakage networks (i.e., where queries are sampled i.i.d.). Bounding the coherence of these networks when queries are dependent would be very interesting and non-trivial. A second limitation is that the leakage networks we analyze are only for static structures so studying the coherence of leakage networks for dynamic encrypted structures would also be interesting. A third limitation is that we do not consider settings where the query distributions can change as a function of the observed leakage. This can be captured using our framework and would likely be non-trivial to analyze. Of course, using our framework to study leakage networks that are more complex than the ones we consider would also be interesting. Finally, we note that while our framework is motivated by leakage in sub-linear ESAs, its applicability is not limited to encrypted search. In fact, in the full version of this work we will show how it can be used to study the leakage of other cryptographic primitives including, for example, secure multi-party computation (e.g., the information about the inputs that is revealed by the outputs).

2 Related Work

Leakage was first modeled in encrypted search in [15] so that it would be explicit and not "hidden under the rug" by making implicit assumptions. The motivation was so that, in the future, cryptanalysis could be performed. [14] further generalized the idea and proposed to parameterize security with a leakage profile. In [33], a nomenclature for leakage patterns and profiles was proposed. It has always been clear that leakage profiles only serve to *describe* leakage and not to understand it so developing a proper "theory of leakage" is one of the oldest open problems in the field. A modest step was taken Kamara, Moataz and Ohrimenko in [33] and Bost and Fouque in [9] where the authors use simulation to compare leakage profiles that are subset/supersets of each other. But the development of a real framework to analyze and study leakage has proven to be challenging. Here, we review related work that proposes such frameworks. We note that our focus is on comparing the *frameworks* that are described in these works and not the specific results proved using the frameworks.

Biased coin game. In [53], Wright and Pouliot propose a framework to study full data recovery attacks against the leakage of deterministic (DTE) and order-revealing encryption (ORE). At a high level, their approach consists of reducing the problem of recovering DTE- and ORE-encrypted data to winning two games the authors call the biased coin game (BCG) and the loaded dice game (LDG). In the (m, n)-BCG, a challenger holds m biased coins each of which lands heads with probability p_i. The challenger then samples a coin according to a prior distribution π and tosses that coin n times. It then provides its prior distribution over coins, the coin probabilities $(p_1, \ldots, p_m)$ and the results of the n coin tosses to an adversary whose goal is to guess which coin was chosen. The (m, n, d)-LDG is a generalization of the BCG to d-sided die. The authors then show how winning the BCG leads to a full data recovery attack on DTE and how winning the LDG leads to a full data recovery attack on ORE.

Quantitative information flow. In a pair of works, Jurado and Smith [28] and later Jurado, Palamidessi and Smith [27], present a comprehensive framework to analyze the leakage of deterministic and order-revealing encryption, respectively. Their approach is based on quantitative information flow (QIF) which is a theoretical framework originally proposed to study the information that a program reveals about a secret [1, 23]. For an introduction to QIF we refer the reader to [3]. At a very high level, the framework models a leakage pattern as a channel which, together with a prior distribution over the plaintexts, results in a distribution over posterior distributions which the authors call the *hyper-distribution.* This hyper-distribution is known to the adversary and, given some observed leakage, results in a specific posterior distribution. The framework also models different adversarial goals as *gain functions* g which can be thought of as loss functions from decision theory and machine learning. The prior g-vulnerability is then defined as the expected gain with respect to the prior distribution and the posterior g-vulnerability is defined as the expected gain over the hyper-distribution. The g-leakage is then defined as the difference or the quotient of the prior and posterior g-vulnerabilities. The authors study the g-leakage of DTE and ORE for various gain functions and prior distributions and use their Theorems to design and study mitigation techniques. Some results are quite surprising; e.g., the authors are able to show that ideal ORE is safe to use against an adversary that wishes to recover an entire column if the values in the column are sampled uniformly at random and the value space is larger than the number of rows.

Leakage inversion. Closer to our own work, Kornaropoulos, Moyer, Papamanthou and Psomas [39] propose a framework to study the leakage of searchable encryption schemes. Roughly speaking, their approach is to characterize the set of all databases (technically multi-maps) that lead to the same observed leakage as the target with respect to a certain leakage profile. This set is the target database's *reconstruction space* and the logarithm of its size is reported as the amount of information revealed about the target database. The framework of [39] quantifies leakage with respect to full data recovery attacks against (scheme specific) response identity leakage, which reveals the results of a query. [5] Furthermore, it handles auxiliary information that can be modeled as a predicate and that can be used to filter out items from the reconstruction space (e.g., "the data contains the word crypto").

PAC learning. Grubbs, Lacharite, Minaud and Paterson propose in [24] to use PAC learning [52] as a framework to study approximate data reconstruction attacks. More precisely, they show how, given $O\left(\frac{d}{\varepsilon}\log\frac{d}{\varepsilon\delta}\right)$ known queries sampled i.i.d, an adversary can recover an ε-approximation of a column with probability at least $1-\delta$. Here, d is the VC-dimension of a concept class needed for the reduction to PAC learning and an ε-approximation is, roughly speaking, a column whose entries will be incorrect with probability at most ε. Similarly to the leakage inversion framework, this approach focuses on data recovery attacks from response identity leakage but, unlike leakage inversion, it only applies to *known-query* attacks.

Summary. With respect to attacks, the BCG/LDG [53] and QIF [28, 27] frameworks model full data recovery attacks against frequency and order leakage. The leakage inversion [39] and PAC-based frameworks [24] model full data recovery attacks against response identity leakage. In this work, we focus on full and partial query recovery attacks against query equality, volume and joint query and volume leakage but our framework naturally and easily handles full and partial

[5] The response identity is sometimes referred to as the access pattern in the context of searchable symmetric encryption

data recovery attacks against any leakage profile that, as far as we know, has appeared in the literature. With respect to auxiliary information, the BCG/LDG framework handles auxiliary distributions that are within a certain statistical distance from the data distribution. The QIF framework assumes the adversary's auxiliary distribution is the same as the data distribution. Leakage inversion studies auxiliary distributions that can be modeled as predicates and the PAC-based framework assumes (non-distributional/perfect) auxiliary knowledge of client queries. Our framework makes no assumption about auxiliary information.

3 Preliminaries

Notation. The set of all binary strings of length n is denoted as $\{0,1\}^n$, and the set of all finite binary strings as $\{0,1\}^*$. We write $x \leftarrow \chi$ to represent an element x being sampled from a distribution χ, and $x \overset{\$}{\leftarrow} X$ to represent an element x being sampled uniformly at random from a set X. The output x of an algorithm $\mathcal{A}$ is denoted by $x \leftarrow \mathcal{A}$. Given a sequence $\mathbf{v}$ of n elements, we refer to its ith element as v_i or $\mathbf{v}[i]$. If S is a set then $\#S$ refers to its cardinality and 2^S to its powerset. We denote to the set of all functions from domain $\mathbb{X}$ to co-domain $\mathbb{Y}$ by $[\mathbb{X} \rightarrow \mathbb{Y}]$. Given a function $f : X \rightarrow Y$ and a sequence $\boldsymbol{x} \in X^n$, we sometimes write $f(\boldsymbol{x})$ to denote the sequence $(f(x_1), \ldots, f(x_n))$. We write $a \overset{\circ}{=} b$ to denote that a is defined as b. We denote Stirling numbers of the second kind by $\left\{ {n \atop k} \right\}$ and the falling factorial by $(m)_i$.

Probabilities. Given a discrete random variable X, we denote its distribution by $p_X(x)$ or $p(x)$ when X is clear. Given two discrete random variables X and Y, we denote the distribution of X conditioned on $Y = y$ for some y over the range of Y, by $p_X(x \mid y)$ or $p(x \mid y)$ when X is clear . This notation, which is common in Machine Learning, Statistics and the literature on Bayesian networks can lead to confusion so we note that when writing $p(x)$ or $p(x \mid y)$, p is a function, x is a variable of p and y is an instantiation of the random variable Y. In other words, $p(x) \overset{\circ}{=} f_X(\cdot)$, where f_X is the probability mass function of X and $p(x \mid y) \overset{\circ}{=} f_{X|Y=y}(\cdot)$, where $f_{X|Y=y}$ is the probability mass function of X conditioned on $Y = y$.

Leakage patterns and profiles. A leakage profile $\Lambda_\Sigma = (\mathcal{L}_{\mathsf{S}}, \mathcal{L}_{\mathsf{O}})$ is composed of a setup leakage $\mathcal{L}_{\mathsf{S}}$ and an operation leakage $\mathcal{L}_{\mathsf{O}}$. Each of these leakage functions can themselves be functions of various leakage patterns. In this work, all leakage functions and leakage patterns are *stateful*. We recall some leakage patterns that will appear throughout this work and refer the reader to [33] for a more comprehensive treatment:

- the *query equality* takes as input a data structure and a query and reveals if and when the query was repeated.
- the *response length* takes as input a data structure and an query and reveals the length of the query's response.

3.1 Bayesian Networks

Bayesian networks are a kind of probabilistic graphical model used to do probabilistic inference. More precisely, they can be used to represent a joint distribution and to infer the marginal distribution of some subset of random variables conditioned on the instantiation of another set of random variables. In our setting, we will use Bayesian networks to infer the instantiations of unobserved query variables conditioned on the instantiation of observed leakage variables.

Bayesian networks. A Bayesian network $\mathcal{N}_{\mathbf{X}}$ over a multi-variate random variable $\mathbf{X} = (X_1, \ldots, X_n)$ is a directed acyclic graph with the random variables X_i as vertices and directed edges between variables that are conditionally dependent. In addition, each node X_i with incoming edges is labeled with a *conditional probability table* defined as

$$\mathsf{cpt}(X_i) \overset{\circ}{=} \Big\{ p(x_i | z_1, \ldots, z_m) \Big\}_{(z_1,\ldots,z_m) \in \mathbb{Z}_1 \times \cdots \times \mathbb{Z}_m},$$

where $Z_1, \ldots, Z_m \in \mathbf{X}$ are the parents of X_i. We can partition the variables $\mathbf{X}$ into a subset of evidence variables $\mathbf{E} \subset \mathbf{X}$, a set of intermediate variables $\mathbf{I} \subset \mathbf{X}$ and a set of hidden variables $\mathbf{H} \in \mathbf{X}$ and use the Bayesian network to infer something about the hidden variables $\mathbf{H}$ given an instantiation $\boldsymbol{e}$ of the evidence variables $\mathbf{E}$. In this work, we will be interested in inferring the *maximum a-posteriori probability* (MAP) estimate which is defined as

$$\mathsf{map}_{\mathbf{H}|\boldsymbol{e}} \overset{\circ}{=} \arg\max_{\mathbf{h}} p(\mathbf{h} \,|\, \boldsymbol{e}) = \arg\max_{\mathbf{h}} \Pr\left[\, \mathbf{H} = \mathbf{h} \,|\, \mathbf{E} = \boldsymbol{e} \,\right]$$

and computing the MAP test which is defined as

$$\mathsf{hyp}_{f,\mathbf{H}|\boldsymbol{e}} \overset{\circ}{=} \arg\max_{s \in \mathbb{S}} \sum_{\mathbf{h} \in f^{-1}(s)} p(\mathbf{h} \,|\, \boldsymbol{e}) = \arg\max_{s \in \mathbb{S}} \sum_{\mathbf{h} \in f^{-1}(s)} \Pr\left[\, \mathbf{H} = \mathbf{h} \,|\, \mathbf{E} = \boldsymbol{e} \,\right]$$

where $f : \mathbb{H} \to \mathbb{S}$. The power of Bayesian networks comes from the *Bayesian network chain rule* which states that

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i \,|\, \mathsf{prnt}(X_i))$$

which is often more efficient to compute than the standard chain rule.

4 Definitions

A *leakage network* $\mathcal{N}_{\mathbf{X}}$ is a Bayesian network over a multi-variate random variable $\mathbf{X} = (X_1, \ldots, X_n)$ whose variables can be partitioned into a set of data variables $\mathbf{D} \subset \mathbf{X}$, a set of query variables $\mathbf{Q} \subset \mathbf{X}$, a set of intermediary variables $\mathbf{I} \subset \mathbf{X}$ and a set of leakage variables $\mathbf{L} \subset \mathbf{X}$. An auxiliary distribution is a random variable $\mathbf{A}$ over the same space as $\mathbf{Q}$. If $\mathbf{D} = \emptyset$, we say that $\mathcal{N}_{\mathbf{X}}$ is a query network and if $\mathbf{Q} = \emptyset$ we say that $\mathcal{N}_{\mathbf{X}}$ is a data network.

Coherence. Let $\mathcal{N}_{\mathbf{X}}$ be a leakage network, $\mathcal{A}$ be an adversary, $\mathbf{A}$ an auxiliary distribution and $g : \mathbb{D} \times \mathbb{Q}_1 \times \cdots \times \mathbb{Q}_n \to \mathbb{S}$ be a recovery function and consider the following probabilistic experiment:

- $\mathbf{CHR}_{\mathcal{A},\mathbf{A},g}(\mathcal{N}_{\mathbf{X}})$:
    1. the challenger samples the network $(\boldsymbol{q}, \boldsymbol{d}, \boldsymbol{i}, \boldsymbol{\ell}) \leftarrow \mathcal{N}_{\mathbf{X}}$;
    2. the adversary outputs $\mathbf{s} \leftarrow \mathcal{A}(\boldsymbol{\ell}, \mathbf{A})$;
    3. if $\mathbf{s} = g(\boldsymbol{d}, \boldsymbol{q})$ output 1 otherwise output 0.

Definition 4.1 (Coherence). *We say that a leakage network $\mathcal{N}_{\mathbf{X}}$ is $(\varepsilon, \mathcal{A}, \mathbf{A}, g)$-coherent if*

$$\left| \Pr\left[\, \mathbf{CHR}_{\mathcal{A},\mathbf{A},g}(\mathcal{N}_{\mathbf{X}}) = 1 \,\right] - \frac{1}{\#\mathbb{S}} \right| = \varepsilon.$$

Capturing different attacks. Definition 4.1 can be used to capture the coherence of a variety of attacks. For example, by setting g to the function $\varphi(\boldsymbol{q}, \boldsymbol{d})$ that returns $\boldsymbol{q}$, one can capture coherence against full query recovery. By setting g to the function $\psi(\boldsymbol{q}, \boldsymbol{d})$ that returns $\boldsymbol{d}$ one can capture coherence against full data recovery. One can also set g to a boolean function to capture coherence against attacks that try to recover a bit or, as we will study in Section 5, to functions $\delta_{\boldsymbol{q}^\star}(\boldsymbol{q})$ that outputs 1 if $\boldsymbol{q} = \boldsymbol{q}^\star$ and 0 otherwise in order to capture attacks that try to learn whether the query is equal to some known $\boldsymbol{q}^\star$.

Remark. Notice that the coherence is not an “absolute” security notion but a relative one in the sense that it depends on the auxiliary distribution. In other words, a particular leakage network could have low coherence when paired with a particular auxiliary distribution $\mathbf{A}$ but have high coherence when paired with another auxiliary distribution $\mathbf{A}'$.

The adversaries. We consider two adversaries $\mathcal{A}_{\mathsf{hyp}}$ and $\mathcal{A}_{\mathsf{map}}$. $\mathcal{A}_{\mathsf{hyp}}$ is a partial recovery adversary that, given leakage $\boldsymbol{\ell}$, an auxiliary distribution $\mathbf{A}$ and a recovery function g, computes the set $\mathsf{hyp}_{g,\mathbf{Q}|e}$ and outputs an element from it uniformly at random. $\mathcal{A}_{\mathsf{map}}$ is a full recovery adversary that, given leakage $\boldsymbol{\ell}$ and an auxiliary distribution $\mathbf{A}$, computes the set $\mathsf{map}_{\mathbf{A}|\boldsymbol{\ell}}$ and outputs a sequence from it uniformly at random.

5 Partial Recovery Against Query Equality

In this Section, we analyze the coherence of i.i.d. query equality networks against partial query recovery attacks. More precisely, we study leakage networks of the form $\mathcal{N}^{+}_{\mathbf{QEQ}}$ as described in Figure 1. We note that, technically, the Bayesian network we use to capture the query equality also reveals the size of the query space $\#\mathbb{Q}$ through the output length of the random function f. We add a $+$ in $\mathcal{N}^{+}_{\mathbf{QEQ}}$ to denote this and note that it is possible to construct a Bayesian network that captures only the query equality. Our first Theorem (Theorem 4 below) gives an upper bound on the coherence of such networks against point recovery functions. Due to space constraints, the proofs of all Theorems are in the appendix.

Theorem 1. *The i.i.d. query equality network $\mathcal{N}^{+}_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{hyp}}, \mathbf{A}, \delta)$-coherent with*

$$\varepsilon = \left| \frac{1}{m!} \sum_{\boldsymbol{q} \in \mathbb{Q}^n} \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \Bigg(\sum_{\boldsymbol{\ell} \in \mathbb{L}_{\delta(\boldsymbol{q})}} \bigg((m - \lambda_{\boldsymbol{\ell}})! \cdot \sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\, \mathbf{Q} = \boldsymbol{q}' \,\right] \bigg) \Bigg) - \frac{1}{\#\mathbb{S}} \right|$$

where $\delta : \mathbb{Q}^n \to \mathbb{S}$, $\mathbb{L}_{\delta(\boldsymbol{q})} \overset{\circ}{=} \{\boldsymbol{\ell} \in \mathbb{L} \mid \mathsf{hyp}_{\delta,\mathbf{Q}|\boldsymbol{\ell}} = \delta(\boldsymbol{q})\}$, $\mathbb{Q}^n_{\boldsymbol{\ell}} = \{\boldsymbol{q} \in \mathbb{Q}^n \mid q_i = q_j \text{ if } \ell_i = \ell_j,\ \forall i, j\}$, $\lambda_{\boldsymbol{\ell}}$ the number of unique leakage values in the sequence $\boldsymbol{\ell}$ and $m \overset{\circ}{=} \#\mathbb{Q} = \#\mathbb{L}$.

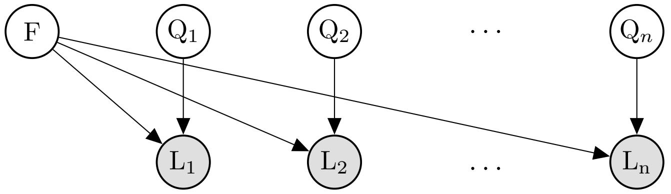


Figure 1: $\mathcal{N}^{+}_{\mathbf{QEQ}}$: the i.i.d. query equality network, where F outputs a function f chosen uniformly at random from $\mathbb{F} \overset{\circ}{=} [\mathbb{Q} \to \mathbb{L}]$, each $\mathrm{Q_i}$ outputs a query from $\mathbb{Q}$ and each $\mathrm{L_i}$ outputs leakage from $\mathbb{L}$. In addition, each $\mathrm{L_i}$ has a conditional probability table of the form $p(\ell_i \mid f, q_i) = 1$ if $\ell_i = f(q_i)$ and $p(\ell_i \mid f, q_i) = 0$ otherwise.

5.1 Uniform Queries with Uniform Auxiliary

We consider the case where both $\mathbf{Q}$ and $\mathbf{A}$ are multi-variate random variables composed of n independent uniform random variables and the (partial) recovery function is a point function that answers questions of the form: *is the query equal to $q^\star$?*

Theorem 2. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}^n_m$ and $\mathbf{A} \sim \mathcal{U}^n_m$, then $\mathcal{N}^{+}_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{hyp}}, \mathbf{A}, \delta)$-coherent with*

$$\varepsilon = \left|\Gamma - \frac{1}{2}\right|,$$

where,

$$\Gamma = \frac{m^n - (m-1)^n}{m^{2n}} \sum_{i=\lceil x_1 \rceil}^{m} (m)_i \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^n}{m^{2n}} \sum_{i=0}^{\lceil x_1 \rceil - 1} (m)_i \cdot \left\{ {n \atop i} \right\},$$

and $x_1 = (3m + 1 - \sqrt{5m^2 + 2m + 1})/2$.

5.2 Zipf Queries with Uniform Auxiliary

We consider the case where both $\mathbf{Q}$ and $\mathbf{A}$ are multi-variate random variables composed of n independent Zipf-distributed random variables. A random variable X is Zipf distributed with parameter s if for all $k \in \{1, \cdots, m\}$

$$\Pr\left[\, X = k \,\right] = \frac{k^{-s}}{H_{m,s}},$$

where $H_{m,s} = \sum_{i=1}^{m} 1/k^s$ is the general form of the harmonic number. We also assume the existence of a permutation $\pi : \mathbb{Q} \to [m]$ that maps every query in the query space $\mathbb{Q}$ to a particular rank in $[m]$. We denote by $\mathcal{Z}_{m,s}$ the Zipf distribution over a query space of size m and parameter s.

Theorem 3. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m,s}$ and $\mathbf{A} \sim \mathcal{U}^n_m$, then $\mathcal{N}^{+}_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{hyp}}, \mathbf{A}, \delta)$-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{2} - \Gamma_1, \Gamma_2 - \frac{1}{2} \right\},$$

where,

$$\Gamma_1 = \frac{m^n - (m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=\lceil x_1 \rceil}^{m} \frac{(m)_i}{i^{ns}} \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=0}^{\lceil x_1 \rceil - 1} \frac{(m)_i}{i^{ns}} \cdot \left\{ {n \atop i} \right\}.$$

and,

$$\Gamma_2 = \frac{m^n - (m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=\lceil x_1 \rceil}^{m} \frac{(m)_i}{(i!)^s} \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=0}^{\lceil x_1 \rceil - 1} \frac{(m)_i}{(i!)^s} \cdot \left\{ {n \atop i} \right\}$$

where $x_1 = (3m + 1 - \sqrt{5m^2 + 2m + 1})/2$.

5.3 Discussion

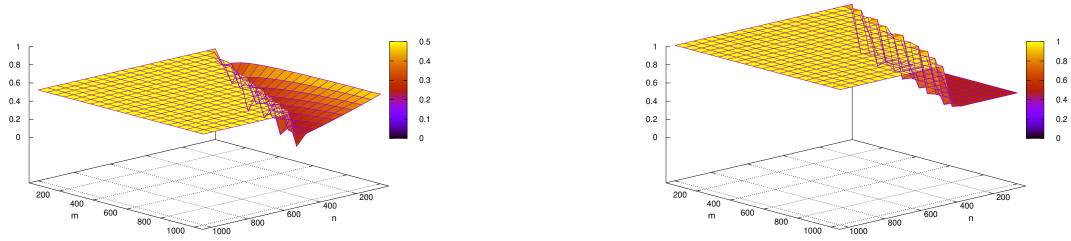
Theorems 2 and 3 analyze $\mathcal{N}^{+}_{\mathbf{QEQ}}$' s coherence against partial query recovery for various combinations of query and auxiliary distributions but they can be hard to understand intuitively. To address this, we plot them in Figure 2.

Uniform queries and auxiliaries. As illustrated by Figure 2a, $\mathcal{N}^{+}_{\mathbf{QEQ}}$'s coherence is close to 1/2 for a non-trivial number of combinations of m and n. This matches the intuition that, if n is fixed, the probability that an adversary can determine whether $q^\star$ is queried will be small except for a small range of m. Interestingly, this shows that there exists a sub-plane where m and n lead to smaller coherence. For example, for $m = 420$ and $n = 200$, the coherence is 0.02. Note that the graph was plotted with an increment of 20, so only 400 points from a possible 1 million points are plotted so it is very likely that there are points that reach even smaller coherence.

Zipf queries and uniform auxiliaries. Theorem 3 only provides an upper bound on the coherence but we can observe in Figure 2b that the values are larger than 0.5 for all 400 points plotted in the graph. Intuitively, this suggests that when the queries are sampled i.i.d. from a Zipf distribution, the query equality could reveal quite a bit of information about whether a query matches a known value (e.g., a known keyword) even if the adversary has no auxiliary information. Improving our bound, however, could also show that the coherence is smaller than our result suggests. Given the combinatorial complexity, obtaining a better bound seems challenging.

Importance of rebuilding. This analysis shows that the coherence significantly increases when the sequence length is larger (by some fraction) than the query space (see Figure 2a). However, when the size of the query space is larger than the sequence length, then the coherence is small. This implies the importance of *rebuild* protocols which, roughly speaking, reconstruct an encrypted structure in such a way that leakage (here, the query equality) is "reset". Rebuilding is a key component in a number of recent structured ESAs [33, 20], and is present in most oblivious RAM schemes [22, 51]. Our observation is that Theorems 2 and 3 can be used to schedule rebuilds. More precisely, given a query space $\mathbb{Q}$ of size m, one can safely use a rebuildable ESA that leaks the query equality for up to m/α queries before rebuilding, where $\alpha > 0$ is a constant. Note that if the bound of Theorem 3 can be further improved, it could also be used to better understand how large a query sequence should be before rebuilding.

A technicality on rebuilding. It is worth mentioning that, for a fixed n, rebuilding slightly changes the adversary's probability of winning the coherence experiment. In particular, there is a linear dependency between the winning probability and the number of rebuild operations. If the client rebuilds after n' queries, the probability that the adversary wins the coherence experiment



(a) Uniform queries and auxiliaries (Theorem 2).   (b) Zipf queries vs uniform auxiliariesm (Theorem 3).

Figure 2: Coherence of $\mathcal{N}^{+}_{\mathbf{QEQ}}$ against partial query recovery.

(when all random variables are independent) is

$$\Pr\left[\bigcup_{i=1}^{\lceil n/n'\rceil} \mathbf{CHR}_{\mathcal{A},\mathbf{A}_i,g}(\mathcal{N}_{\mathbf{X}_i}) = 1\right] \leq \lceil n/n' \rceil \cdot \Pr\left[\,\mathbf{CHR}_{\mathcal{A},\mathbf{A},g}(\mathcal{N}_{\mathbf{X}}) = 1\,\right]$$

where $\mathbf{Q} = (Q_1, \ldots, Q_{n'})$ and $\mathbf{A} = (A_1, \ldots, A_{n'})$, which follows from a union bound.

6 Full Recovery Against Query Equality

In this Section, we analyze the coherence of i.i.d. query equality networks against full query recovery attacks.

Theorem 4. *The i.i.d. query equality network $\mathcal{N}^{+}_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent, with*

$$\varepsilon = \left| \frac{1}{m!} \cdot \sum_{f\in\mathbb{F}} \left(\sum_{\boldsymbol{q}\in\mathbb{Q}_1^n} \frac{1}{\#S_{f(\boldsymbol{q})}} \cdot \Pr\left[\,\mathbf{Q} = \boldsymbol{q}\,\right] \right) - \frac{1}{m^n} \right|,$$

where $S_{f(\boldsymbol{q})} \overset{\circ}{=} \mathsf{map}_{\mathbf{A}|f(\boldsymbol{q})}$, $\mathbb{Q}_1^n \overset{\circ}{=} \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{f(\boldsymbol{q})}\}$ and $m \overset{\circ}{=} \#\mathbb{Q} = \#\mathbb{L}$.

Theorem 4 provides a closed form for the coherence against $\mathcal{A}_{\mathsf{map}}$ for arbitrary $\mathbf{Q}$ and $\mathbf{A}$. In Section 6.1 we study specific distributions and derive simpler bounds. Specifically, we consider: (1) uniform queries with uniform auxiliary distributions; and (2) uniform queries with Zipf auxiliary distributions, and (3) Zipf queries with Zipf auxiliary distributions.

6.1 Uniform Queries and Uniform Auxiliary

We consider the case where both $\mathbf{Q}$ and $\mathbf{A}$ are multi-variate random variables composed of n independent uniform random variables. We denote by $\mathcal{U}_m$ the uniform distribution with a support of size m.

Theorem 5. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}_m^n$ and $\mathbf{A} \sim \mathcal{U}_m^n$, then $\mathcal{N}_{\mathbf{QEQ}}^{+}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent, with*

$$\varepsilon = \left| \frac{1}{m^n} \cdot \sum_{i=1}^{m} \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right|,$$

where $m = \#\mathbb{Q}$.

While Theorem 5 applies to arbitrary m and n, it is not closed form. We show below, however, that when $m = n$ the coherence can be very small even for large values of n.

Corollary 6.1. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}^n$, $\mathbf{A} \sim \mathcal{U}^n$ and $m = n = \#\mathbb{Q}$, then $\mathcal{N}_{\mathbf{QEQ}}^{+}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \left(\frac{0.792}{\log(n+1)} \right)^n.$$

Proof. First notice that when $m = n$,

$$\sum_{i=1}^{m} \left\{ {n \atop i} \right\} = \sum_{i=0}^{n} \left\{ {n \atop i} \right\} = B_n,$$

where B_n is the Bell number. We also know from [7] that for all $n > 0$,

$$B_n \leq \left(\frac{0.792 \cdot n}{\log(n+1)} \right)^n$$

That is,

$$\varepsilon = \frac{1}{n^n} \cdot B_n - \frac{1}{n^n} \leq \left(\frac{0.792}{\log(n+1)} \right)^n - \frac{1}{n^n} \leq \left(\frac{0.792}{\log(n+1)} \right)^n.$$

■

6.2 Uniform Queries with Zipf Auxiliaries

We consider the case where $\mathbf{Q}$ is a multi-variate random variable composed of n independent uniform random variables and $\mathbf{A}$ is composed of n independent Zipf-distributed random variables.

Theorem 6. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}_m^n$ and $\mathbf{A} \sim \mathcal{Z}_{m,s}^n$, then $\mathcal{N}_{\mathbf{QEQ}}^{+}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max \left\{ \frac{1}{m^n}, \frac{1}{m^n} \cdot \sum_{i=1}^{m} i! \cdot \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right\}$$

where $m = \#\mathbb{Q}$.

While Theorem 6 applies to arbitrary values of m and n, it is not closed-form so we give an approximation when $n = m$ below.

Corollary 6.2. *For $n \geq 2$, if $\mathbf{Q} \sim \mathcal{U}_m^n$, $\mathbf{A} \sim \mathcal{Z}_{m,s}^n$ and $m = n$, then for large values of n, $\mathcal{N}_{\mathbf{QEQ}}^{+}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \lesssim \frac{n}{2 \cdot e^{n-1} \cdot (\ln(2))^{n+1}}.$$

Proof. First notice that when $m = n$,

$$\sum_{i=1}^{m} i! \cdot \left\{ {n \atop i} \right\} = \sum_{i=0}^{n} i! \cdot \left\{ {n \atop i} \right\} = F_n,$$

where F_n is the Fubini number (or the ordered Bell number). We also know from [6] that

$$F_n = \frac{n!}{2(\log(2))^{n+1}} + o((n-1)!) \approx \frac{n!}{2(\log(2))^{n+1}},$$

where the second equality holds for large numbers of n. In addition, for large values of n, we also have

$$F_n \cdot \frac{1}{n^n} - \frac{1}{n^n} \geq \frac{1}{n^n}$$

Finally, since $n! \leq n^{n+1}/e^{n-1}$ and putting it all together we obtain

$$\varepsilon \leq F_n \cdot \frac{1}{n^n} - \frac{1}{n^n} \leq F_n \cdot \frac{1}{n^n} \lesssim \frac{n}{2 \cdot e^{n-1} \cdot (\log(2))^{n+1}}.$$

■

6.3 Zipf Queries with Zipf Auxiliaries

We consider the case where both $\mathbf{Q}$ and $\mathbf{A}$ are multi-variate random variables composed of n independent Zipf-distributed random variables with the same underlying permutation π and parameter s.

Theorem 7. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m,s}$ and $\mathbf{A} \sim \mathcal{Z}^n_{m,s}$, then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{m^n} - \frac{1}{H^n_{m,s}} \cdot \sum_{i=1}^{m} (i!)^{1-s} \cdot \left\{ {n \atop i} \right\}, \frac{1}{H^n_{m,s}} \cdot \sum_{i=1}^{m} i^{-n \cdot s} \cdot \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right\},$$

where $m = \#\mathbb{Q}$.

6.4 Uniform Queries with Distinct Uniform Auxiliary

So far, we assumed that the query and auxiliary distributions share the same support which captures cases where the adversary knows the exact queries a client samples from. In the following subsections, we consider cases where the support of the query distribution $\mathbb{Q}$ and the auxiliary distributions $\mathbb{A}$ are distinct. Specifically, we are interested in the cases where $\mathbb{A} \subset \mathbb{Q}$ and where $\mathbb{Q} \subset \mathbb{A}$. The former captures cases where the adversary knows only a part of the client's support and the latter captures settings where the adversary has access to a distribution with support that includes the client's support. We consider uniform queries with uniform auxiliaries and Zipf queries with Zipf auxiliaries. We define $m_q \overset{\circ}{=} \#\mathbb{Q}$ and $m_a \overset{\circ}{=} \#\mathbb{A}$.

Theorem 8. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}^n_{m_q}$, $\mathbf{A} \sim \mathcal{U}^n_{m_a}$, and $\mathbb{A} \subset \mathbb{Q}$ then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon = \left| \frac{1}{m_q^n} \cdot \sum_{i=1}^{m_a} \left\{ {n \atop i} \right\} - \frac{1}{m_q^n} \right|.$$

In the following corollary we show that when $m_a = o(m_q)$, the coherence is significantly smaller than when the query and auxiliary distributions share the same support. This shows that full query recovery attacks are much harder when $\mathbb{A} \subset \mathbb{Q}$.

Corollary 6.3. *For all $n \in \mathbb{N}$, if $m_a = n$, $m_q = n \ln n$, $\mathbf{Q} \sim \mathcal{U}^n_{m_q}$ and $\mathbf{A} \sim \mathcal{U}^n_{m_a}$, then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \left(\frac{0.792}{\log^2 n}\right)^n.$$

The proof is omitted as it is similar to the proof of Corollary 6.1.

We now consider the case where $\mathbb{Q} \subset \mathbb{A}$.

Theorem 9. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}^n_{m_q}$, $\mathbf{A} \sim \mathcal{U}^n_{m_a}$, and $\mathbb{Q} \subset \mathbb{A}$ then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \frac{1}{m_q^n} \cdot \sum_{i=1}^{m_q} \left(\frac{m_q \cdot e}{m_a}\right)^i \cdot \left\{ {n \atop i} \right\}.$$

6.5 Zipf Queries with Distinct Zipf Auxiliaries

We consider the case where $\mathbf{Q}$ is a multi-variate random variable composed of n independent Zipf-distributed random variables over a support $\mathbb{Q}$ with parameter s_q and permutation π_q, and $\mathbf{A}$ is composed of n independent Zipf-distributed random variables over a support $\mathbb{A}$ with parameter s_a and permutation π_a. When $\mathbb{A} \subset \mathbb{Q}$, we use γ to refer to the maximum rank in $\mathbb{Q}$ of any query in $\mathbb{A}$. More formally, let $\gamma \in [m_q - m_a]$ such that for all $q \in \mathbb{A}$, $\pi_q[q] \geq \gamma$. When $\mathbb{Q} \subset \mathbb{A}$, we use θ to refer to the maximum rank of a query that belongs to $\mathbb{A}$ and not to $\mathbb{Q}$. More formally, let $\theta \in [m_q + 1]$ such that $\pi_a^{-1}[i] \in \mathbb{Q}$ for all $i \in [\theta]$ and $\pi_a^{-1}[\theta + 1] \notin \mathbb{Q}$.

Theorem 10. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m_q,s_q}$, $\mathbf{A} \sim \mathcal{Z}^n_{m_a,s_a}$, and $\mathbb{A} \subset \mathbb{Q}$ then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{m_q^n} - \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{m_a} \frac{\left\{ {n \atop i} \right\}}{(\gamma + i)^{s_q \cdot n}}, \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \cdot \sum_{i=1}^{m_a} i! \cdot \left\{ {n \atop i} \right\} - \frac{1}{m_q^n} \right\}$$

Though the upper bound obtained in Theorem 10 is not tight, we can show that there are cases where the coherence can be very small.

Corollary 6.4. *For all $n \geq 2$, if $m_a = n$, $m_q = n \ln n$, $\mathbf{Q} \sim \mathcal{Z}^n_{m_q,1}$, $\mathbf{A} \sim \mathcal{Z}^n_{m_a,1}$, $\gamma = n/\log(n)$, then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \lesssim \frac{n}{2 \cdot e^{n-1} \cdot (\ln(2))^{n+1}}.$$

Proof. Assuming $m_a = n$ and $s_q = s_a = 1$, then we have from the proof of Theorem 10 ,

$$\begin{aligned} \Pr[\mathbf{CHR} = 1] &\leq \frac{1}{H^n_{n\log(n),1} \cdot \gamma^n} \cdot \sum_{i=1}^{n} i! \cdot \left\{ {n \atop i} \right\} \\ &= \frac{(\log n)^n}{H^n_{n\log(n),1} \cdot n^n} \cdot F_n \end{aligned}$$

where F_n is the Fubini number. And we know that for all $n \geq 1$, $H_{n,1} \geq \log(n)$ which gives

$$\Pr\left[\, \mathbf{CHR} = 1 \,\right] \leq \frac{(\log n)^n}{(\log(n \log n))^n \cdot n^n} \cdot F_n \leq \frac{1}{n^n} \cdot F_n$$

And since $F_n \geq 2$, then

$$F_n \cdot \frac{1}{n^n} - \frac{1}{n^n} \geq \frac{1}{n^n}$$

which means that the second term of the maximum value in Theorem 10 is the upper bound of the coherence such that

$$\varepsilon \leq F_n \cdot \frac{1}{n^n} - \frac{1}{n^n} \leq F_n \cdot \frac{1}{n^n} \lesssim \frac{n}{2 \cdot e^{n-1} \cdot (\ln(2))^{n+1}},$$

where the second inequality follows from the same argument in Corollary 6.2.

■

Theorem 11. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m_q,s_q}$, $\mathbf{A} \sim \mathcal{Z}^n_{m_a,s_a}$, and $\mathbb{Q} \subset \mathbb{A}$ then $\mathcal{N}^+_{\mathbf{VOL}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{m_q^n} - \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{\theta} \frac{\left\{ \begin{smallmatrix} n \\ i \end{smallmatrix} \right\}}{i^{s_q \cdot n}}, \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{\theta} (i!)^{1-s_q} \cdot \left\{ \begin{matrix} n \\ i \end{matrix} \right\} - \frac{1}{m_q^n} \right\}$$

6.6 Discussion

In this section, we plot and analyze the coherence of $\mathcal{N}^+_{\mathbf{QEQ}}$ against full query recovery attacks and studied in Theorems 5, 6, 8, 9, 10 and 11.[6] Note that these graphs (and all the remaining coherence plots) plot the log of the coherence.

Uniform queries and auxiliaries. Figure 3a plots the coherence for uniform queries and auxiliaries. Overall, one can see that the coherence is extremely small even for small and moderate values of m and n. One can also see that it shrinks significantly when m and n increase. For example, for $m = 800$ and $n = 200$, we found that $\varepsilon \leq 2^{-1011}$. It is worth noting that we limited the support size to $m = 1000$ due to computational limitations, but most leakage attacks tend to use datasets with a much larger query space. For example, the Enron dataset [10] has a query space of size $m > 10^5$ which would lead to much smaller coherence.

Uniform queries with Zipf auxiliaries. Figure 3b plots the coherence for uniform queries and Zipf auxiliaries. Compared to the uniform/uniform case above, the coherence is significantly larger even for large values of m and n. This is mainly due to our bound being loose and could potentially be improved. Note, however, that while the coherence is larger, it is still small and suggests that full query recovery over against i.i.d. query equality networks is challenging when the query distribution is uniform and the auxiliaries are Zipf. For example, for $m = 800$ and $n = 200$, we found that $\varepsilon \leq 2^{-576}$.

[6] In this section, we consider $m, n \geq 20$ which implies that for all the Theorems shown in Section 6, the coherence is upper bounded by the second term of the max function.

Uniform queries with distinct uniform auxiliaries. In this setting we consider two cases: (1) when $\mathbb{A} \subset \mathbb{Q}$; and (2) when $\mathbb{Q} \subset \mathbb{A}$. Note that the the second case, where the adversary and the client do not share the same query space, is more realistic. The coherence for the case $\mathbb{A} \subset \mathbb{Q}$ is plotted in Figures 3d. Here, we made an additional assumption that the size of the query space, m_q, is 10 times larger than the size of the auxiliary space, m_a.[7] In the distinct setting, the coherence is significantly smaller than when the query and auxiliary distributions share the same space space. One can also notice that it is extremely small even for small values of m_a and n. For example, for $m_a = 20$ and $n = 20$, we see that $\varepsilon \leq 2^{-106}$. And when we increase m_a and n, the coherence decreases significantly. For example, when $m_a = 800$ and $n = 200$, $\varepsilon \leq 2^{-1676}$. Figure 3c plots the coherence when $\mathbb{Q} \subset \mathbb{A}$ and $m_a = 10 \cdot m_q$. The shape of the coherence is similar to the case of uniform queries and auxiliaries but with smaller values—though the bound obtained in Theorem 9 is not tight. For example, when $m_q = 800$ and $n = 200$, $\varepsilon \leq 2^{-1094}$.

Zipf queries with distinct Zipf auxiliaries. We again consider the cases where: (1) $\mathbb{A} \subset \mathbb{Q}$; and (2) $\mathbb{Q} \subset \mathbb{A}$. Figure 3e plots the coherence when $\gamma = n/\log n$, where γ is the maximum rank in $\mathbb{Q}$ for any query in $\mathbb{A}$. For example, for $n = 200$ and $\gamma = 38$, which captures that the probability of querying from $\mathbb{A}$ is at most $\Pr[\,Q = 38\,]$ where $Q \sim \mathcal{Z}_{m_q,s_q}$. We observe that the coherence is small when m_a is small. This is intuitive because when the auxiliary space is small, the adversary can only guess a small number of sequences whereas the client can generate a large number of query sequences. For example when $m_a = 100$ and $n = 200$, we found that $\varepsilon \leq 2^{-704}$, whereas when $m_a = 800$ and $n = 200$, $\varepsilon \leq 2^{-345}$. We would also like to point out that the shape of the graph varies as a function of γ. As γ tends to 1, the coherence is much larger (see Theorem 10). Figure 3e plots the coherence when $\theta = \log m_q$. Recall that θ is the maximum rank of a query that belongs to $\mathbb{A}$ and not to $\mathbb{Q}$. For example, for $m_q = 200$, $\theta = 5$ means that there exist four queries in $\mathbb{Q}$ that have ranks 1 to 4 and no query in $\mathbb{Q}$ that has rank 5 in $\mathbb{A}$. We observe that the coherence is small when m_q is small. Given the bound shown in Theorem 11, this is intuitive since we made θ depend on the size of the query space. And the larger m_q is, the larger θ is and therefore the larger the coherence is. For example, when $m_q = 20$ and $n = 200$, $\varepsilon \leq 2^{-365}$, while when $m_q = 800$ and $n = 200$, $\varepsilon \leq 2^{-112}$.

A remark on rebuilding. Contrary to the partial recovery setting, the need for rebuilding is less clear in this setting. In fact, all the coherence values were small for our choice of parameters.

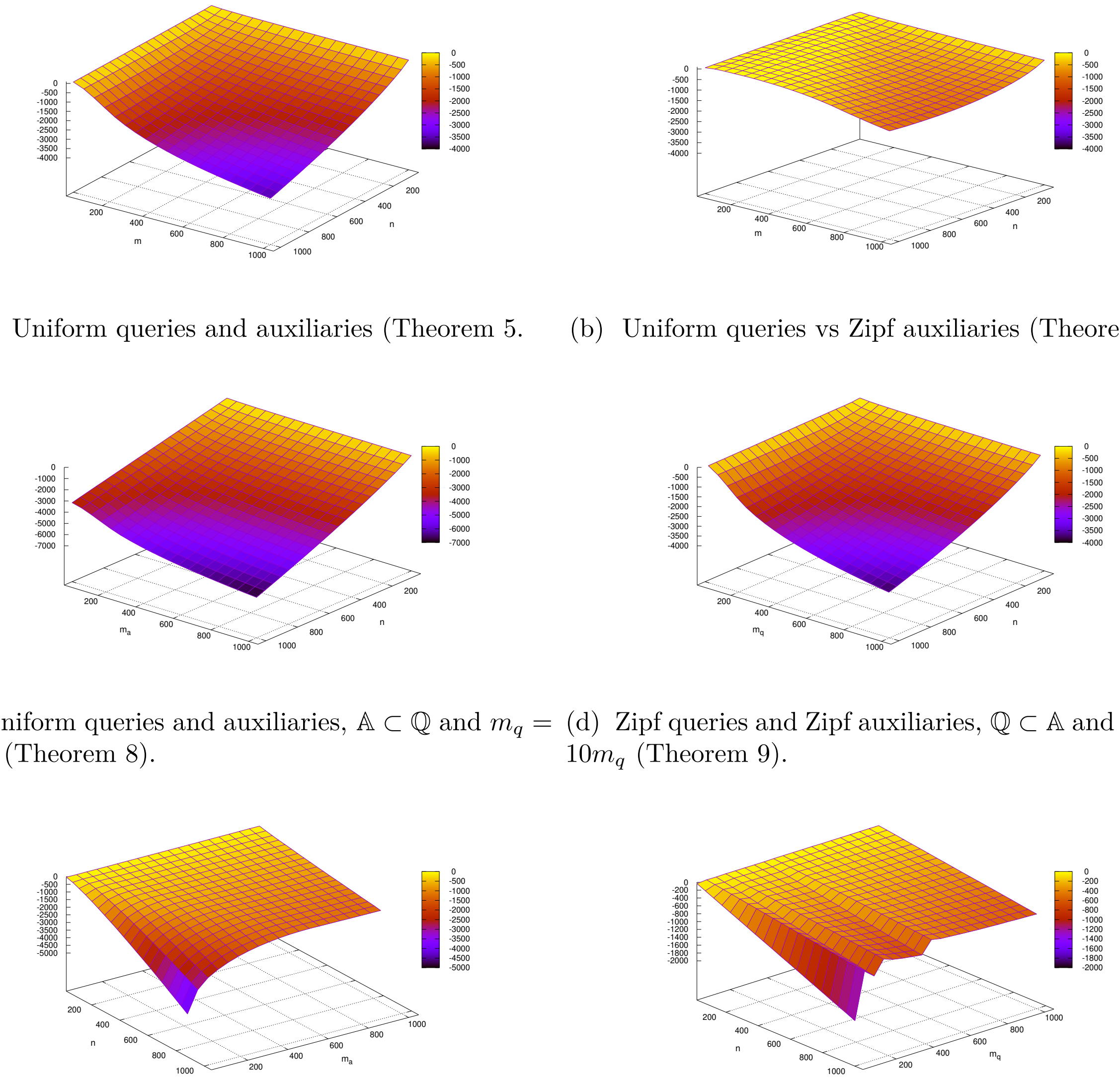
7 Full Recovery Against Volume

In this Section, we analyze the coherence of i.i.d. volume networks against full query recovery attacks. More precisely, we study leakage networks $\mathcal{N}_{\mathbf{VOL}}$ of the form described in Figure 4, where the random variable D outputs a function d from the space

$$\mathbb{D}_N = \left\{ d \in \left[\mathbb{Q} \to \{\star\}^{[N-m+1]}\right] \;\middle|\; \sum_{i=1}^{m} \#d(q_i) = N \right\}$$

and where $m = \#\mathbb{Q}$ and $N \in \mathbb{N}$ such that $N \geq m$. The functions $d \in \mathbb{D}$ are meant to model multi-maps which are data structures that map queries (usually called labels) to tuples. Here, N

[7] This factor was not picked arbitrarily. We observed that in the Enron email dataset [10], the keyword space of a single user' s inbox is at least 10 times smaller than the size of the keyword space of the entire dataset.



(a) Uniform queries and auxiliaries (Theorem 5.

(b) Uniform queries vs Zipf auxiliaries (Theorem 6).

(c) Uniform queries and auxiliaries, $\mathbb{A} \subset \mathbb{Q}$ and $m_q = 10m_a$ (Theorem 8).

(d) Zipf queries and Zipf auxiliaries, $\mathbb{Q} \subset \mathbb{A}$ and $m_a = 10m_q$ (Theorem 9).

(e) Zipf queries and Zipf auxiliaries, $\mathbb{A} \subset \mathbb{Q}$ and $\gamma = n/\log n$ (Theorem 10).

(f) Uniform queries and auxiliaries, $\mathbb{Q} \subset \mathbb{A}$ and $\theta = \log m_q$ (Theorem 11).

Figure 3: Log-coherence of $\mathcal{N}^{+}_{\mathbf{QEQ}}$ against full query recovery.

captures the size of the multi-map, i.e., the sum of its tuple lengths. Similar to the query equality case, we note that, technically, the Bayesian network we use to capture the volume also reveals the size of the query space $\#\mathbb{Q}$ as well as the size of the multi-map N through the output length of the function $d \in \mathbb{D}_N$. We add a + in $\mathcal{N}^{+}_{\mathbf{VOL}}$ to denote this.

Here, we often decompose the adversary' s auxiliary distribution $\mathbf{A}$ into a set of random variables $\mathbf{A_Q}$ which denote its auxiliary random variables over the queries and A_D which denotes its auxiliary distribution over the data. Our first Theorem (Theorem 12 below) gives an upper bound on the

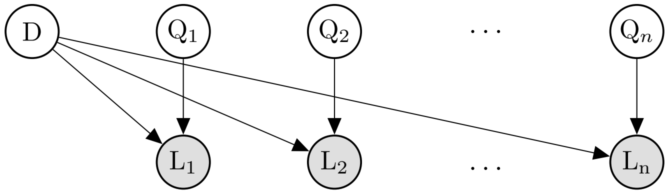


Figure 4: $\mathcal{N}^{+}_{\mathbf{VOL}}$: the i.i.d. volume network, where D outputs a multi-map d from $\mathbb{D}_N$, each $\mathrm{Q_i}$ outputs a query from $\mathbb{Q}$ and each $\mathrm{L_i}$ outputs leakage from $\mathbb{L}$. In addition, each $\mathrm{L_i}$ has a conditional probability table of the form $p(\ell_i \,|\, d, q_i) = 1$ if $\ell_i = \#d(q_i)$ and $p(\ell_i \,|\, d, q_i) = 0$ otherwise.

coherence of such networks.

Theorem 12. *The i.i.d. volume network $\mathcal{N}^{+}_{\mathbf{VOL}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \psi)$-coherent with*

$$\varepsilon = \left| \sum_{d \in \mathbb{D}_N} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\# S_{\#d(\boldsymbol{q})}} \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, D = d \,\right] \Big) - \frac{1}{m^n} \right|,$$

where $S_{\#d(\boldsymbol{q})} := \mathsf{map}_{\mathbf{A_Q}|\#d(\boldsymbol{q})}$ and $\mathbb{Q}_1^n := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{\#d(\boldsymbol{q})}\}$.

7.1 Uniform Data and Queries with Uniform Auxiliaries

We consider the case where $\mathbf{Q}$ and $\mathbf{A_Q}$ are multi-variate random variables composed of n independent uniform random variables and where D and A_D are uniform random variables over $\mathbb{D}_N$.

Claim 1. *For all $m, N \in \mathbb{N}$ such that $m \leq N$, if d is sampled uniformly at random from $\mathbb{D}_N$, then*

$$\Pr\left[\, D = d \,\right] = \frac{1}{\binom{N-1}{m-1}}.$$

Proof. First recall that $m \overset{\circ}{=} \#\mathbb{Q}$ and that d is sampled uniformly at random from $\mathbb{D}_N$ which is the set of functions that map $\mathbb{Q}$ to $\{\star\}^{N-m+1}$ such that

$$\#d(q_1) + \#d(q_2) + \cdots + \#d(q_m) = N.$$

where for all $i \in [m]$, $\#d(q_i) \geq 1$ (this follows from the fact that in a multi-map every label has tuple length at least 1). Our task then is to count the number of such functions. The number of solutions can be obtained by using a stars and bars argument. Consider m distinguishable bins and N balls. The number of ways the N balls can be allocated to the bins is equivalent to putting $m-1$ bars in $N-1$ positions between the stars (since the bins are not allowed to be empty in our case). And there are $\binom{N-1}{m-1}$ ways of setting the bars. This is exactly the same number of solutions the above equation can have.

■

Theorem 13. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}_m^n$, $\mathbf{A} \sim \mathcal{U}_m^n$, $D \sim \mathcal{U}_{\mathbb{D}_N}$ and $A_D \sim \mathcal{U}_{\mathbb{D}_N}$ then $\mathcal{N}^{+}_{\mathbf{VOL}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{m^n}, \frac{1}{m^n} \cdot \sum_{i=1}^{m'} \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right\}$$

where $m' = \min(m, \sqrt{2N})$

Similar to Corollary 6.1, we present below an asymptotic approximation when $\sqrt{2N} = n$ and $m < n$, that states that the coherence can be very small for large values of n.

Corollary 7.1. *If $\mathbf{Q} \sim \mathcal{U}_m^n$, $\mathbf{A} \sim \mathcal{U}_m^n$, $D \sim \mathcal{U}_{\mathbb{D}_N}$ and $A_D \sim \mathcal{U}_{\mathbb{D}_N}$ with $\sqrt{2N} = n$ and $m > \sqrt{2N}$ then $\mathcal{N}^+_{\mathbf{VOL}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \left(\frac{0.792 \cdot n}{m \cdot \log(n+1)}\right)^n.$$

We omit the details of the proof since it is similar to the one of Corollary 6.1.

7.2 Zipf Data and Queries with Zipf Auxiliaries

We consider the case where $\mathbf{Q}$ and $\mathbf{A_Q}$ are multi-variate random variables composed of n independent Zipf-distributed random variables and where D and A_D are uniform random variables over $\mathbb{D}_N$. In particular, we consider power-law shaped multi-maps where $\mathbb{D}_N$ now represents all permutations from $\mathbb{Q}$ to tuples that have sizes in $\mathbb{S}$ where

$$\mathbb{S} = \left\{\frac{N}{H_{m,s'}}, \frac{N}{2^{s'} \cdot H_{m,1}}, \cdots, \frac{N}{m^{s'} \cdot H_{m,s'}}\right\} \quad \text{and} \quad \mathbb{D}_N = \left\{d \in \left[\mathbb{Q} \to \{\star\}^{\mathbb{S}}\right]\right\}$$

Our choice of power-law-shaped multi-maps is not arbitrary. The evaluation of almost every leakage attack on exact keyword search [26, 11, 8, 30] uses datasets datasets that are power-law shaped, e.g., the Enron email dataset [10], the Wikipedia corpus [19], or the Arabidopsis Information Resource (TAIR) database [43].[8]

Theorem 14. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}_{m,s}^n$, $\mathbf{A} \sim \mathcal{Z}_{ms,}^n$, $D \sim \mathcal{U}_{\mathbb{D}_N}$ and $A_D \sim \mathcal{U}_{\mathbb{D}_N}$ then $\mathcal{N}^+_{\mathbf{QeVo}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max\left\{\frac{1}{m^n} - \frac{1}{H_{m,s}^n} \cdot \sum_{i=1}^m i^{-n \cdot s} \cdot \left\{\begin{matrix} n \\ i \end{matrix}\right\}, \frac{1}{H_{m,s}^n} \cdot \sum_{i=1}^m (i!)^{1-s} \cdot \left\{\begin{matrix} n \\ i \end{matrix}\right\} - \frac{1}{m^n}\right\}$$

7.3 Zipf Data and Queries with Distinct Zipf Query Auxiliary and Uniform Data Auxiliary

In this section, we consider the case where the adversary does not necessarily know the entire support and only knows a subset of the volumes in $\mathbb{S}$. In particular, we consider the adversary's query auxiliary distribution to be a Zipf with a support $\mathbb{A} \subset \mathbb{Q}$, a parameter s_a, and a permutation π_a; and its data auxiliary distribution to be uniform over a set of volumes $\mathbb{S}_{\mathbb{A}} \subset \mathbb{S}$. In particular, the possible multi-maps belong to the set $\bar{\mathbb{D}}_N$ where

$$\mathbb{D}_N = \left\{d \in \left[\mathbb{Q} \to \{\star\}^{\mathbb{S}_{\mathbb{A}}}\right]\right\}.$$

[8] Note however that in our analysis we pick $\mathbb{S}$ to have a Zipf-like distribution where every query maps to a different, unique volume in $\mathbb{S}$. One could modify this setup by allowing queries to map to the same volume but this would overcomplicates the analysis which is not necessary for this first work.

Theorem 15. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m_q,s_q}$, $\mathbf{A} \sim \mathcal{Z}^n_{m_a,s_a}$, $D \sim \mathcal{U}_{\mathbb{D}_N}$, $A_D \sim \mathcal{U}_{\bar{\mathbb{D}}_N}$ and $\mathbb{A} \subset \mathbb{Q}$ then $\mathcal{N}^+_{\mathbf{QeVo}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max\left\{\frac{1}{m^n} - \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{m_a} \frac{(m_a)_i \cdot \left\{{n \atop i}\right\}}{(m_q)_i \cdot (\gamma+i)^{s_q \cdot n}}, \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \cdot \sum_{i=1}^{m_a} \frac{(m_a)_i}{(m_q)_i} \cdot i! \cdot \left\{{n \atop i}\right\} - \frac{1}{m^n}\right\}$$

where γ the rank as defined in Section 6.5.

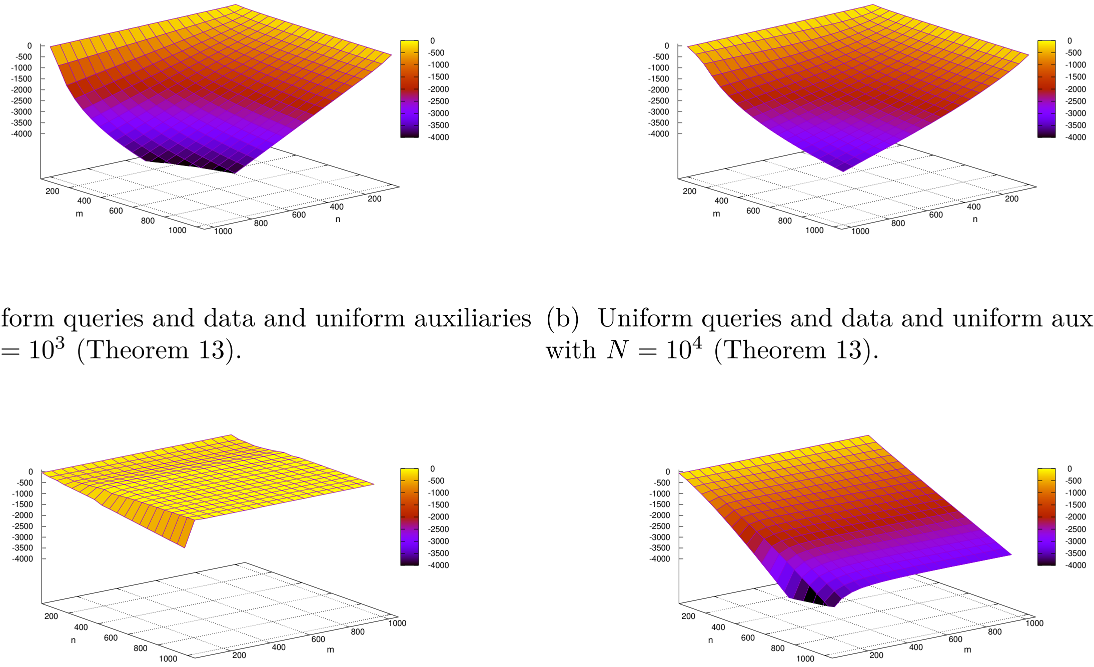
7.4 Discussion

As for the previous Theorems, it is challenging to interpret the coherence bounds in Theorems 13 and 15 so we plot them in Figure 5 for various values of m, n, N and γ.[9]

Uniform data and queries and uniform auxiliaries. We consider two cases with different multi-map sizes: (1) where $N = 10^3$; and (2) where $N = 10^4$. One can see that increasing the size of the multi-map N leads to slightly larger coherence for small values of m. This is expected since in Theorem 13, the summation goes to $m' = \min(\sqrt{2N}, m)$ and for large values of N and m, m' increases. Figure 5a plots the coherence when $N = 10^3$, and one can see that it is extremely small. For example, for $m = 100$ and $n = 200$, $\varepsilon \leq 2^{-415}$. The coherence decreases significantly as we increase both m and n. Interestingly, even though were only able to show an upper bound, $\mathcal{N}^+_{\mathbf{VOL}}$'s coherence is smaller than $\mathcal{N}^+_{\mathbf{QEQ}}$'s when the query and auxiliary distributions are uniform (see Section 6). In Figure 5b, we see that increasing N has a slight impact which is expected as highlighted above. For example, for $m = 100$ and $n = 200$, $\varepsilon \leq 2^{-411}$.

Zipf data and queries and Zipf auxiliaries. We consider two cases with different values of γ: (1) where $\gamma = \log n$; and (2) where $\gamma = n/\log n$. One can clearly see the impact of picking larger γ. Recall that γ is the maximum rank in $\mathbb{Q}$ for any query in $\mathbb{A}$. So the smaller it is, the higher the coherence is. The first case captures cases where the permutation of the adversary's Zipf distribution is *similar* to the client's, whereas the second case captures the opposite. Intuitively, γ can be thought of as a metric that captures some form of distance between the client and adversary' s Zipf distributions. In Figure 5c, the coherence is large compared to all the previous settings we analyzed. Part of this is because we were only able to obtain a lower bound—which means it could potentially improve in the future—but another part is simply because, for small values of γ, the MAP adversary does very well as it is be able to predict a larger number of sequences correctly. Recall that a crucial step in the proof of Theorem 15 is based on the observation that the number of query sequences that an adversary guesses is small compared to the uniform case (and sometimes can even be equal to 1), and this is due in part to the skewness of the distribution. While we mentioned above that the coherence is large compared to the previous cases, the concrete values are still small and do not suggest that $\mathcal{N}^+_v ol$ leakage is harmful even in this case. For example, for $m = 800$ and $n = 200$, $\varepsilon \leq 2^{-172}$. In Figure 5d, we observe that when we increase γ, the coherence decreases significantly. For example, for $m = 800$ and $n = 200$, $\varepsilon \leq 2^{-707}$. This is expected since γ is in the denominator of the bound.

[9] As for the previous section, since we pick $m, n \geq 20$, we can simply show that the coherence upper bounds in all the Theorems of this section are equal to the second term of the max function.



(a) Uniform queries and data and uniform auxiliaries with $N = 10^3$ (Theorem 13).

(b) Uniform queries and data and uniform auxiliaries with $N = 10^4$ (Theorem 13).

(c) Zipf queries and data and Zipf auxiliaries with $\gamma = \log n$ (Theorem 15).

(d) Zipf queries and data and Zipf auxiliaries with $\gamma = n/\log n$ (Theorem 15).

Figure 5: Log-coherence of $\mathcal{N}^{+}_{\mathbf{VOL}}$ against full query recovery.

8 Full Recovery Against Query Equality and Volume

In this Section, we analyze the coherence of i.i.d. query equality and volume networks against full query recovery attacks. More precisely, we study leakage networks $\mathcal{N}_{\mathbf{QeVo}}$ as described in Figure 6 with $\mathbb{D}_N$ as in Section 8. Similar to the query equality and volume cases, we add a $+$ in $\mathcal{N}^{+}_{\mathbf{QeVo}}$ to denote that our network reveals both m and N. Our first Theorem (Theorem 16 below) gives an upper bound on the coherence of such networks.

Theorem 16. *The i.i.d. query-volume network $\mathcal{N}^{+}_{\mathbf{QeVo}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon = \left| \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \sum_{d \in \mathbb{D}_N} \left(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\#S_{\boldsymbol{\ell}}} \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, D = d \,\right] \right) - \frac{1}{m^n} \right|,$$

where $S_{\boldsymbol{\ell}} \overset{\circ}{=} \mathsf{map}_{\mathbf{A_Q}|\boldsymbol{\ell}}$ and $\mathbb{Q}_1^n := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{\boldsymbol{\ell}}\}$ and $\ell_i = (f(q_i), \#d(q_i))$ for all $i \in [n]$.

The proof of this theorem is similar to Theorems 4 and 12.

8.1 Uniform Data and Queries with Uniform Auxiliaries

Let $\mathbf{Q}$ and $\mathbf{A_Q}$ be multi-variate random variables composed of n independent uniform random variables, and let D and A_D be uniform random variables.

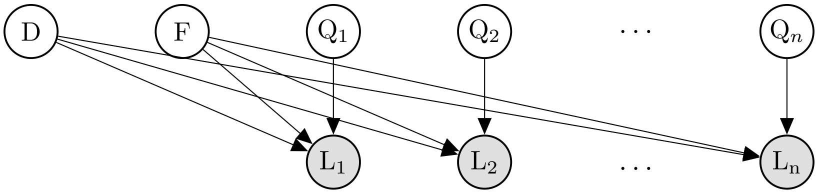


Figure 6: $\mathcal{N}^{+}_{\mathbf{QeVo}}$: the i.i.d. query equality and volume network, where D outputs a function d uniformly at random from $\mathbb{D}_N$, F outputs a function uniformly at random from $\mathbb{F} \overset{\circ}{=} [\mathbb{Q} \to \mathbb{L}]$ and each $\mathrm{Q_i}$ outputs a query from $\mathbb{Q}$ and each $\mathrm{L_i}$ outputs leakage from $\mathbb{L}$. In addition, each $\mathrm{L_i}$ has a conditional probability table of the form $p(\ell_i \,|\, d, f, q_i) = 1$ if $\ell_i = \big(f(q_i), \#d(q_i)\big)$ and $p(\ell_i \,|\, d, q_i) = 0$ otherwise.

Theorem 17. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}^n_m$, $\mathbf{A} \sim \mathcal{U}^n_m$, $D \sim \mathcal{U}_{\mathbb{D}_N}$ and $A_D \sim \mathcal{U}_{\mathbb{D}_N}$ then $\mathcal{N}^{+}_{\mathbf{QeVo}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon = \left| \frac{1}{m^n} \cdot \sum_{i=1}^{m} \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right|,$$

where $m = \#\mathbb{Q}$.

References

[1] *Cryptography and data security*, volume 112. Addison-Wesley, 1982.

[2] R. Ada Popa, C. Redfield, N. Zeldovich, and H. Balakrishnan. CryptDB: Protecting confidentiality with encrypted query processing. In *ACM Symposium on Operating Systems Principles (SOSP)*, pages 85–100, 2011.

[3] M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith. *The Science of Quantitative Information Flow.* Springer, 2020.

[4] G. Amjad, S. Kamara, and T. Moataz. Breach-resistant structured encryption. In *Proceedings on Privacy Enhancing Technologies (Po/PETS '19)*, 2019.

[5] G. Amjad, S. Patel, G. Persiano, K. Yeo, and M. Yung. Dynamic volume-hiding encrypted multi-maps with applications to searchable encryption. *Cryptology ePrint Archive*, 2021.

[6] J.-P. Barthélémy. An asymptotic equivalent for the number of total preorders on a finite set. *Discrete Mathematics*, 29(3):311–313, 1980.

[7] D. Berend and T. Tassa. Improved bounds on bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30(2):185–205, 2010.

[8] L. Blackstone, S. Kamara, and T. Moataz. Revisiting leakage abuse attacks. In *Network and Distributed System Security Symposium (NDSS '20)*, 2020.

[9] R. Bost and P. Fouque. Security-efficiency tradeoffs in searchable encryption. *Proc. Priv. Enhancing Technol.*, 2019(4):132–151, 2019.

[10] CALO Project. Enron Email Dataset. `https://www.cs.cmu.edu/~enron`, 2018.

[11] D. Cash, P. Grubbs, J. Perry, and T. Ristenpart. Leakage-abuse attacks against searchable encryption. In *ACM Conference on Communications and Computer Security (CCS '15)*, pages 668–679. ACM, 2015.

[12] D. Cash, J. Jaeger, S. Jarecki, C. Jutla, H. Krawczyk, M. Rosu, and M. Steiner. Dynamic searchable encryption in very-large databases: Data structures and implementation. In *Network and Distributed System Security Symposium (NDSS '14)*, 2014.

[13] D. Cash, S. Jarecki, C. Jutla, H. Krawczyk, M. Rosu, and M. Steiner. Highly-scalable searchable symmetric encryption with support for boolean queries. In *Advances in Cryptology - CRYPTO '13*. Springer, 2013.

[14] M. Chase and S. Kamara. Structured encryption and controlled disclosure. In *Advances in Cryptology - ASIACRYPT '10*, volume 6477 of *Lecture Notes in Computer Science*, pages 577–594. Springer, 2010.

[15] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky. Searchable symmetric encryption: Improved definitions and efficient constructions. In *ACM Conference on Computer and Communications Security (CCS '06)*, pages 79–88. ACM, 2006.

[16] I. Demertzis, D. Papadopoulos, and C. Papamanthou. Searchable encryption with optimal locality: Achieving sublogarithmic read efficiency. In *Advances in Cryptology - CRYPTO '18*, pages 371–406. Springer, 2018.

[17] M. Etemad, A. Küpçü, C. Papamanthou, and D. Evans. Efficient dynamic searchable encryption with forward privacy. *PoPETs*, 2018(1):5–20, 2018.

[18] B. A. Fisch, B. Vo, F. Krell, A. Kumarasubramanian, V. Kolesnikov, T. Malkin, and S. M. Bellovin. Malicious-client security in blind seer: a scalable private dbms. In *IEEE Symposium on Security and Privacy*, pages 395–410. IEEE, 2015.

[19] W. Foundation. Simple English Wikipedia. https://simple.wikipedia. org/, 2014.

[20] M. George, S. Kamra, and T. Moataz. Structured encryption and dynamic leakage suppression. In *Advances in Cryptology - EUROCRYPT 2021*, 2021.

[21] E.-J. Goh. Secure indexes. Technical Report 2003/216, IACR ePrint Cryptography Archive, 2003. See `http://eprint.iacr.org/2003/216`.

[22] O. Goldreich and R. Ostrovsky. Software protection and simulation on oblivious RAMs. *Journal of the ACM*, 43(3):431–473, 1996.

[23] J. W. Gray. Toward a mathematical foundation for information flow security. *J. Comput. Secur.*, 1(3–4):255–294, may 1992.

[24] P. Grubbs, M. Lacharité, B. Minaud, and K. G. Paterson. Pump up the volume: Practical database reconstruction from volume leakage on range queries. In D. Lie, M. Mannan, M. Backes, and X. Wang, editors, *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, pages 315–331. ACM, 2018.

[25] P. Grubbs, M. Lacharité, B. Minaud, and K. G. Paterson. Learning to reconstruct: Statistical learning theory and encrypted database attacks. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 1067–1083. IEEE, 2019.

[26] M. S. Islam, M. Kuzu, and M. Kantarcioglu. Access pattern disclosure on searchable encryption: Ramification, attack and mitigation. In *Network and Distributed System Security Symposium (NDSS '12)*, 2012.

[27] M. Jurado, C. Palamidessi, and G. Smith. A formal information-theoretic leakage analysis of order-revealing encryption. In *2021 IEEE 34th Computer Security Foundations Symposium (CSF)*, pages 1–16, 2021.

[28] M. Jurado and G. Smith. Quantifying information leakage of deterministic encryption. In *Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop*, CCSW'19, pages 129–139, New York, NY, USA, 2019. Association for Computing Machinery.

[29] S. Kamara, A. Kati, J. D. Maria, T. Moataz, A. Park, and A. Treiber. MAPLE: MArkov Process Leakage attacks on Encrypted Search. Technical report, IACR ePrint Cryptography Archive, 2023.

[30] S. Kamara, A. Kati, T. Moataz, T. Schneider, A. Treiber, and M. Yonli. Sok: Cryptanalysis of encrypted search with leaker–a framework for leakage attack evaluation on real-world data. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 90–108. IEEE, 2022.

[31] S. Kamara and T. Moataz. SQL on Structurally-Encrypted Data. In *Asiacrypt*, 2018.

[32] S. Kamara and T. Moataz. Computationally volume-hiding structured encryption. In *Advances in Cryptology - Eurocrypt' 19*, 2019.

[33] S. Kamara, T. Moataz, and O. Ohrimenko. Structured encryption and leakae suppression. In *Advances in Cryptology - CRYPTO '18*, 2018.

[34] S. Kamara, T. Moataz, A. Park, and L. Qin. A decentralized and encrypted national gun registry. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1520–1537. IEEE, 2021.

[35] S. Kamara, T. Moataz, S. Zdonik, and Z. Zhao. Opx: An optimal relational database encryption scheme. Technical report, IACR ePrint Cryptography Archive, 2020.

[36] S. Kamara and C. Papamanthou. Parallel and dynamic searchable symmetric encryption. In *Financial Cryptography and Data Security (FC '13)*, 2013.

[37] S. Kamara, C. Papamanthou, and T. Roeder. Dynamic searchable symmetric encryption. In *ACM Conference on Computer and Communications Security (CCS '12)*. ACM Press, 2012.

[38] G. Kellaris, G. Kollios, K. Nissim, and A. O. Neill. Generic attacks on secure outsourced databases. In *ACM Conference on Computer and Communications Security (CCS '16)*, 2016.

[39] E. M. Kornaropoulos, N. Moyer, C. Papamanthou, and A. Psomas. Leakage inversion. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. ACM, nov 2022.

[40] E. M. Kornaropoulos, C. Papamanthou, and R. Tamassia. The state of the uniform: Attacks on encrypted databases beyond the uniform query distribution. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1223–1240. IEEE, 2020.

[41] E. M. Kornaropoulos, C. Papamanthou, and R. Tamassia. Response-hiding encrypted ranges: Revisiting security via parametrized leakage-abuse attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1502–1519. IEEE, 2021.

[42] M. Lacharité, B. Minaud, and K. G. Paterson. Improved reconstruction attacks on encrypted data using range query leakage. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pages 297–314. IEEE Computer Society, 2018.

[43] P. Lamesch, K. Dreher, D. Swarbreck, R. Sasidharan, L. Reiser, and E. Huala. Using the arabidopsis information resource (tair) to find information about arabidopsis genes. *Current Protocols in Bioinformatics*, 30(1):1–11, 2010.

[44] MongoDB.

[45] M. Naveed, S. Kamara, and C. V. Wright. Inference attacks on property-preserving encrypted databases. In *ACM Conference on Computer and Communications Security (CCS)*, CCS '15, pages 644–655. ACM, 2015.

[46] S. Oya and F. Kerschbaum. Hiding the access pattern is not enough: Exploiting search pattern leakage in searchable encryption. In *USENIX Security Symposium*, pages 127–142, 2021.

[47] V. Pappas, F. Krell, B. Vo, V. Kolesnikov, T. Malkin, S.-G. Choi, W. George, A. Keromytis, and S. Bellovin. Blind seer: A scalable private dbms. In *Security and Privacy (SP), 2014 IEEE Symposium on*, pages 359–374. IEEE, 2014.

[48] S. Patel, G. Persiano, K. Yeo, and M. Yung. Mitigating leakage in secure cloud-hosted data structures: Volume-hiding for multi-maps via hashing. In L. Cavallaro, J. Kinder, X. Wang, and J. Katz, editors, *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 79–93. ACM, 2019.

[49] S. Patranabis and D. Mukhopadhyay. Forward and backward private conjunctive searchable symmetric encryption. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021.* The Internet Society, 2021.

[50] D. Song, D. Wagner, and A. Perrig. Practical techniques for searching on encrypted data. In *IEEE Symposium on Research in Security and Privacy*, pages 44–55. IEEE Computer Society, 2000.

[51] E. Stefanov, M. van Dijk, E. Shi, C. Fletcher, L. Ren, X. Yu, and S. Devadas. Path oram: An extremely simple oblivious ram protocol. In *ACM Conference on Computer and Communications Security (CCS '13)*, 2013.

[52] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[53] C. V. Wright and D. Pouliot. Early detection and analysis of leakage abuse vulnerabilities. *IACR Cryptol. ePrint Arch.*, page 1052, 2017.

[54] Z. Zhao, S. Kamara, T. Moataz, and S. Zdonik. Encrypted databases: From theory to systems. In *Conference on Innovative Data Systems Research (CIDR '21)*, 2021.

A Proofs for Section 5 (Partial Recovery Against Query Equality)

Theorem 1. *The i.i.d. query equality network* $\mathcal{N}^{+}_{\mathbf{QEQ}}$ *is* $(\varepsilon, \mathcal{A}_{\mathsf{hyp}}, \mathbf{A}, \delta)$*-coherent with*

$$\varepsilon = \left| \frac{1}{m!} \sum_{\boldsymbol{q} \in \mathbb{Q}^n} \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \Bigg(\sum_{\boldsymbol{\ell} \in \mathbb{L}_{\delta(\boldsymbol{q})}} \bigg((m - \lambda_{\boldsymbol{\ell}})! \cdot \sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\, \mathbf{Q} = \boldsymbol{q}' \,\right] \bigg) \Bigg) - \frac{1}{\#\mathbb{S}} \right|$$

where $\delta : \mathbb{Q}^n \to \mathbb{S}$*,* $\mathbb{L}_{\delta(\boldsymbol{q})} \overset{\circ}{=} \{\boldsymbol{\ell} \in \mathbb{L} \mid \mathsf{hyp}_{\delta, \mathbf{Q}|\boldsymbol{\ell}} = \delta(\boldsymbol{q})\}$*,* $\mathbb{Q}^n_{\boldsymbol{\ell}} = \{\boldsymbol{q} \in \mathbb{Q}^n \mid q_i = q_j \text{ if } \ell_i = \ell_j,\ \forall i, j\}$*,* $\lambda_{\boldsymbol{\ell}}$ *the number of unique leakage values in the sequence* $\boldsymbol{\ell}$ *and* $m \overset{\circ}{=} \#\mathbb{Q} = \#\mathbb{L}$*.*

Proof. For visual clarity, we will denote the random variable $\mathbf{CHR}_{\mathcal{A}, \mathbf{A}, g}(\mathcal{N}_{\mathbf{X}})$ simply as $\mathbf{CHR}$ throughout all the proofs in the paper. First, observe

$$\Pr\left[\, \mathbf{CHR} = 1 \,\right] = \sum_{\boldsymbol{\ell} \in \mathbb{L}^n} \Pr\left[\, \mathbf{CHR} = 1 | \mathbf{L} = \boldsymbol{\ell} \,\right] \cdot \Pr\left[\, \mathbf{L} = \boldsymbol{\ell} \,\right] \tag{1}$$

$$= \sum_{\boldsymbol{q} \in \mathbb{Q}^n} \Bigg(\sum_{\boldsymbol{\ell} \in \mathbb{L}^n} \Pr\left[\, \mathbf{CHR} = 1 | \mathbf{L} = \boldsymbol{\ell}, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, \mathbf{L} = \boldsymbol{\ell} \,\right] \Bigg) \tag{2}$$

We know that,

$$\Pr\left[\, \mathbf{L} = \boldsymbol{\ell} \,\right] = \frac{1}{m!} \sum_{f \in \mathbb{F}} \Bigg(\sum_{\mathbf{q} \in \mathbb{Q}^n} \Pr\left[\, \mathbf{L} = \boldsymbol{\ell} | F = f, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \Bigg)$$

$$= \frac{1}{m!} \sum_{\mathbf{q} \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Bigg(\sum_{f \in \mathbb{F}} \Pr\left[\, \mathbf{L} = \boldsymbol{\ell} | F = f, \mathbf{Q} = \boldsymbol{q} \,\right] \Bigg) \tag{3}$$

$$= \frac{(m - \lambda)!}{m!} \sum_{\mathbf{q} \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \tag{4}$$

Equation 3 simply follows from the observation that the only valid query sequences are the ones that coincide with the leakage. These query sequences are the ones that belong to the following set

$$\mathbb{Q}^n_{\boldsymbol{\ell}} = \Big\{ \boldsymbol{q} \in \mathbb{Q}^n | q_i = q_j \text{ if } l_i = l_j \ \forall i, j \Big\}$$

And Equation 4 follows from the fact that

$$\Pr\left[\, \mathbf{L} = \boldsymbol{\ell} | \mathbf{A} = \boldsymbol{q}, F = f \,\right] = \begin{cases} 1 & \text{if } \boldsymbol{q} \in \mathbb{Q}^n_{\boldsymbol{\ell}} \\ 0 & \text{otherwise.} \end{cases}$$

and therefore the number of functions f such that $\Pr\left[\, \mathbf{L} = \boldsymbol{\ell} | F = f, \mathbf{Q} = \boldsymbol{q} \,\right] = 1$ is equal to $(m - \lambda_{\boldsymbol{\ell}})!$ where $\lambda_{\boldsymbol{\ell}}$ is the number of unique leakage values. On the other hand, based on the definition of coherence experiment, we know that

$$\Pr\left[\, \mathbf{CHR} = 1 | \mathbf{L} = \boldsymbol{\ell}, \mathbf{Q} = \boldsymbol{q} \,\right] = \begin{cases} 1 & \text{if } \boldsymbol{\ell} \in \mathbb{L}_{\delta(\boldsymbol{q})} \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathbb{L}_{\delta(\boldsymbol{q})} = \{\boldsymbol{\ell} \in \mathbb{L}^n \mid \mathsf{hyp}_{\delta,\mathbf{A}|\boldsymbol{\ell}} = \delta(\boldsymbol{q})\}$, which is equivalent to

$$\Pr\left[\,\mathbf{CHR} = 1 | \mathbf{L} = \boldsymbol{\ell}, \mathbf{Q} = \boldsymbol{q}\,\right] = \begin{cases} 1 & \text{if } P_{\delta(\boldsymbol{\ell})} \geq P_i \forall\ i \in \mathbb{S} \setminus \{\delta(\boldsymbol{\ell})\} \\ 0 & \text{otherwise.} \end{cases}$$

where for all $i \in \mathbb{S}$,

$$P_i(\boldsymbol{\ell}) = \sum_{\boldsymbol{q} \in \delta^{-1}(i)} \Pr\left[\,\mathbf{A} = \boldsymbol{q} | \mathbf{L} = \boldsymbol{\ell}\,\right]$$

In the following, we compute the posterior so we can determine the quantities above. First, we have

$$\begin{aligned} \Pr\left[\,\mathbf{A} = \boldsymbol{q} | \mathbf{L} = \boldsymbol{\ell}\,\right] &= \Pr\left[\,\mathbf{L} = \boldsymbol{\ell} | \mathbf{A} = \boldsymbol{q}\,\right] \cdot \frac{\Pr\left[\,\mathbf{A} = \boldsymbol{q}\,\right]}{\Pr\left[\,\mathbf{L} = \boldsymbol{\ell}\,\right]} \\ &= \frac{\Pr\left[\,\mathbf{A} = \boldsymbol{q}\,\right]}{\Pr\left[\,\mathbf{L} = \boldsymbol{\ell}\,\right]} \cdot \frac{1}{m!} \sum_{f \in \mathbb{F}} \Pr\left[\,\mathbf{L} = \boldsymbol{\ell} | \mathbf{A} = \boldsymbol{q}, F = f\,\right] \\ &= \frac{\Pr\left[\,\mathbf{A} = \boldsymbol{q}\,\right]}{(m - \lambda)! \sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\,\mathbf{Q} = \boldsymbol{q}'\,\right]} \cdot \sum_{f \in \mathbb{F}} \Pr\left[\,\mathbf{L} = \boldsymbol{\ell} | \mathbf{A} = \boldsymbol{q}, F = f\,\right] \end{aligned}$$

And based on the value of $\Pr\left[\,\mathbf{L} = \boldsymbol{\ell} | \mathbf{A} = \boldsymbol{q}, F = f\,\right]$ that we have detailed above, we obtain

$$\Pr\left[\,\mathbf{A} = \boldsymbol{q} | \mathbf{L} = \boldsymbol{\ell}\,\right] = \begin{cases} \Pr\left[\,\mathbf{A} = \boldsymbol{q}\,\right] \cdot \left(\sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\,\mathbf{Q} = \boldsymbol{q}'\,\right] \right)^{-1} & \text{if } \boldsymbol{q} \in \mathbb{Q}^n_{\boldsymbol{\ell}} \\ 0 & \text{otherwise.} \end{cases}$$

Given the above, we obtain a more precise representation of the partition P_i, for $i \in \{0, 1\}$, such that

$$P_i(\boldsymbol{\ell}) = \sum_{\boldsymbol{q} \in \delta^{-1}(i) \bigcap \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\,\mathbf{A} = \boldsymbol{q}\,\right] \cdot \Bigg(\sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\,\mathbf{Q} = \boldsymbol{q}'\,\right] \Bigg)^{-1}.$$

Finally, plugging the results above in Equation 2, we obtain

$$\Pr\left[\,\mathbf{CHR} = 1\,\right] = \frac{1}{m!} \sum_{\boldsymbol{q} \in \mathbb{Q}^n} \Pr\left[\,\mathbf{Q} = \boldsymbol{q}\,\right] \Bigg(\sum_{\boldsymbol{\ell} \in \mathbb{L}_{\delta(\boldsymbol{q})}} \Bigg((m - \lambda_{\boldsymbol{\ell}})! \cdot \sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\,\mathbf{Q} = \boldsymbol{q}'\,\right] \Bigg) \Bigg)$$

where $\lambda_{\boldsymbol{\ell}}$ is the number of unique leakage value in the sequence $\boldsymbol{\ell}$. Finally, subtraction $1/\#\mathbb{S}$ and taking the absolute value concludes our proof.

■

Theorem 2. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}^n_m$ and $\mathbf{A} \sim \mathcal{U}^n_m$, then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{hyp}}, \mathbf{A}, \delta)$-coherent with*

$$\varepsilon = \left| \Gamma - \frac{1}{2} \right|,$$

where,

$$\Gamma = \frac{m^n - (m-1)^n}{m^{2n}} \sum_{i = \lceil x_1 \rceil}^{m} (m)_i \cdot \begin{Bmatrix} n \\ i \end{Bmatrix} + \frac{(m-1)^n}{m^{2n}} \sum_{i=0}^{\lceil x_1 \rceil - 1} (m)_i \cdot \begin{Bmatrix} n \\ i \end{Bmatrix},$$

and $x_1 = (3m + 1 - \sqrt{5m^2 + 2m + 1})/2$.

Proof. First, observe that the recovery function partitions the query space into two sets $\mathbb{S}_0 = \delta^{-1}(0)$ and $\mathbb{S}_1 = \delta^{-1}(1)$ such that

$$\mathbb{S}_0 = \Big\{ \boldsymbol{q} \in \mathbb{Q}^n \mid \exists\ i \in [n],\ q_i = q^\star \Big\} \quad \text{and} \quad \mathbb{S}_1 = \Big\{ \boldsymbol{q} \in \mathbb{Q}^n \mid \forall\ i \in [n],\ q_i \neq q^\star \Big\}.$$

From Theorem 1, we know that for $i \in \{0, 1\}$,

$$\begin{aligned} P_i(\boldsymbol{\ell}) &= \sum_{\boldsymbol{q} \in \mathbb{S}_i \bigcap \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\, \mathbf{A} = \boldsymbol{q} \,\right] \cdot \Big(\sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\, \mathbf{Q} = \boldsymbol{q}' \,\right] \Big)^{-1} \\ &= \frac{1}{m^n} \cdot \Big(\sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\, \mathbf{Q} = \boldsymbol{q}' \,\right] \Big)^{-1} \cdot \# \Big(\mathbb{S}_i \bigcap \mathbb{Q}^n_{\boldsymbol{\ell}} \Big) \end{aligned}$$

In order to quantify the partial coherence, the main challenge consists of determining the leakage sequences that compose the set $\mathbb{L}_\delta$. Based on the definition of $\mathbb{L}_\delta$, identifying these sequences boil down to solving the following inequality for all $\boldsymbol{\ell} \in \mathbf{L}$,

$$R(\boldsymbol{\ell}) = \frac{P_0(\boldsymbol{\ell})}{P_1(\boldsymbol{\ell})} = \frac{\# \Big(\mathbb{S}_0 \bigcap \mathbb{Q}^n_{\boldsymbol{\ell}} \Big)}{\# \Big(\mathbb{S}_1 \bigcap \mathbb{Q}^n_{\boldsymbol{\ell}} \Big)} \geq 1$$

It is easy to see that

$$\# \Big(\mathbb{S}_0 \bigcap \mathbb{Q}^n_{\boldsymbol{\ell}} \Big) = \lambda_{\boldsymbol{\ell}} \cdot m \cdot (m-1) \cdots (m - \lambda_{\boldsymbol{\ell}} + 2) = \frac{\lambda_{\boldsymbol{\ell}} \cdot m!}{(m - \lambda_{\boldsymbol{\ell}} + 1)!}.$$

The above is exactly the number of query sequences that coincide with $\boldsymbol{\ell}$ and that contain the query $q^\star$ as one of the queries. Similarly,

$$\# \Big(\mathbb{S}_1 \bigcap \mathbb{Q}^n_{\boldsymbol{\ell}} \Big) = (m-1) \cdot (m-2) \cdots (m - 1 - (\lambda_{\boldsymbol{\ell}} - 1)) = \frac{(m-1)!}{(m - \lambda_{\boldsymbol{\ell}} - 1)!},$$

where here instead we don't want any query to be equal to $q^\star$ and this explains why we start from $(m-1)$ in the quantity above. Given the above quantities, we can rewrite $R(\boldsymbol{\ell})$ as

$$R(\boldsymbol{\ell}) = \frac{\lambda_{\boldsymbol{\ell}} \cdot m}{(m - \lambda_{\boldsymbol{\ell}} + 1) \cdot (m - \lambda_{\boldsymbol{\ell}})}$$

Solving $R(\boldsymbol{\ell}) \geq 1$ is equivalent to solving $\lambda^2_{\boldsymbol{\ell}} - \lambda_{\boldsymbol{\ell}}(3m+1) + m^2 + m \leq 0$. This quadratic inequality has two roots such that

$$x_1 = \frac{3m + 1 - \sqrt{5m^2 + 2m + 1}}{2} \quad \text{and} \quad x_2 = \frac{3m + 1 + \sqrt{5m^2 + 2m + 1}}{2}$$

This implies that $\boldsymbol{\ell} \in \mathbb{L}_0$ iff $\lambda_{\boldsymbol{\ell}} \in \{\lceil x_1 \rceil, \cdots, \lfloor x_2 \rfloor\}$; but since $x_2 > m$, then $\lambda_{\boldsymbol{\ell}} \in \{\lceil x_1 \rceil, \cdots, m\}$. Given the result of Theorem 1, we have

$$\Pr\left[\, \mathbf{CHR} = 1 \,\right] = \frac{1}{m!} \sum_{\boldsymbol{q} \in \mathbb{Q}^n} \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \Big(\sum_{\boldsymbol{\ell} \in \mathbb{L}_{\delta(\boldsymbol{q})}} \Big((m - \lambda_{\boldsymbol{\ell}})! \cdot \sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Pr\left[\, \mathbf{Q} = \boldsymbol{q}' \,\right] \Big) \Big) \tag{5}$$

$$
\begin{aligned}
&= \frac{1}{m!\cdot m^{n}} \sum_{k\in\{0,1\}} \Bigg(\sum_{\boldsymbol{q}\in\mathbb{S}_k} \Bigg(\sum_{\boldsymbol{\ell}\in\mathbb{L}_{\delta(\boldsymbol{q})}} \Bigg((m-\lambda_{\boldsymbol{\ell}})! \cdot \sum_{\mathbf{q}'\in\mathbb{Q}^{n}_{\boldsymbol{\ell}}} \Pr\left[\,\mathbf{Q}=\boldsymbol{q}'\,\right] \Bigg)\Bigg)\Bigg) && (6)\\
&= \frac{1}{m!\cdot m^{n}} \sum_{\boldsymbol{q}\in\mathbb{S}_0} \Bigg(\sum_{i=\lceil x_1\rceil}^{m} \Bigg(\sum_{\boldsymbol{\ell}\in\mathbb{L}^{n}_{i}} \Bigg((m-i)! \cdot \sum_{\mathbf{q}'\in\mathbb{Q}^{n}_{\boldsymbol{\ell},i}} \Pr\left[\,\mathbf{Q}=\boldsymbol{q}'\,\right] \Bigg)\Bigg)\Bigg) \\
&+ \frac{1}{m!\cdot m^{n}} \sum_{\boldsymbol{q}\in\mathbb{S}_1} \Bigg(\sum_{i=0}^{\lceil x_1\rceil-1} \Bigg(\sum_{\boldsymbol{\ell}\in\mathbb{L}^{n}_{i}} \Bigg((m-i)! \cdot \sum_{\mathbf{q}'\in\mathbb{Q}^{n}_{\boldsymbol{\ell},i}} \Pr\left[\,\mathbf{Q}=\boldsymbol{q}'\,\right] \Bigg)\Bigg)\Bigg) && (7)\\
&= \frac{1}{m!\cdot m^{2n}} \sum_{\boldsymbol{q}\in\mathbb{S}_0} \Bigg(\sum_{i=\lceil x_1\rceil}^{m} \Bigg(\sum_{\boldsymbol{\ell}\in\mathbb{L}^{n}_{i}} \Big((m-i)! \cdot \#\mathbb{Q}^{n}_{\boldsymbol{\ell},i} \Big)\Bigg)\Bigg) \\
&+ \frac{1}{m!\cdot m^{2n}} \sum_{\boldsymbol{q}\in\mathbb{S}_1} \Bigg(\sum_{i=0}^{\lceil x_1\rceil-1} \Bigg(\sum_{\boldsymbol{\ell}\in\mathbb{L}^{n}_{i}} \Big((m-i)! \cdot \#\mathbb{Q}^{n}_{\boldsymbol{\ell},i} \Big)\Bigg)\Bigg) \\
&= \frac{1}{m!\cdot m^{2n}} \sum_{\boldsymbol{q}\in\mathbb{S}_0} \Bigg(\sum_{i=\lceil x_1\rceil}^{m} \Bigg(\sum_{\boldsymbol{\ell}\in\mathbb{L}^{n}_{i}} \Bigg((m-i)! \cdot \frac{m!}{(m-i)!} \Bigg)\Bigg)\Bigg) \\
&+ \frac{1}{m!\cdot m^{2n}} \sum_{\boldsymbol{q}\in\mathbb{S}_1} \Bigg(\sum_{i=0}^{\lceil x_1\rceil-1} \Bigg(\sum_{\boldsymbol{\ell}\in\mathbb{L}^{n}_{i}} \Bigg((m-i)! \cdot \frac{m!}{(m-i)!} \Bigg)\Bigg)\Bigg) && (8)\\
&= \frac{1}{m^{2n}} \sum_{\boldsymbol{q}\in\mathbb{S}_0} \Bigg(\sum_{i=\lceil x_1\rceil}^{m} \#\mathbb{L}^{n}_{i} \Bigg) + \frac{1}{m^{2n}} \sum_{\boldsymbol{q}\in\mathbb{S}_1} \Bigg(\sum_{i=0}^{\lceil x_1\rceil-1} \#\mathbb{L}^{n}_{i} \Bigg) \\
&= \frac{\#\mathbb{S}_0}{m^{2n}} \sum_{i=\lceil x_1\rceil}^{m} \binom{m}{i} \cdot i! \cdot \left\{ {n \atop i} \right\} + \frac{\#\mathbb{S}_1}{m^{2n}} \Bigg(\sum_{i=0}^{\lceil x_1\rceil-1} \binom{m}{i} \cdot i! \cdot \left\{ {n \atop i} \right\} && (9)\\
&= \frac{\#\mathbb{S}_0}{m^{2n}} \sum_{i=\lceil x_1\rceil}^{m} (m)_i \cdot \left\{ {n \atop i} \right\} + \frac{\#\mathbb{S}_1}{m^{2n}} \Bigg(\sum_{i=0}^{\lceil x_1\rceil-1} (m)_i \cdot \left\{ {n \atop i} \right\} \\
&= \frac{m^{n}-(m-1)^{n}}{m^{2n}} \sum_{i=\lceil x_1\rceil}^{m} (m)_i \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^{n}}{m^{2n}} \sum_{i=0}^{\lceil x_1\rceil-1} (m)_i \cdot \left\{ {n \atop i} \right\} && (10)
\end{aligned}
$$

Equality 5 is from the result of Theorem 4. In Equality 6, we simply partition the query space to $\mathbb{Q}^n = \mathbb{S}_0 \bigcup \mathbb{S}_1$. Equality 7 follows from our result above that $\boldsymbol{\ell} \in \mathbb{L}_\delta$ iff $m \geq \lambda_{\boldsymbol{\ell}} \geq \lceil x_1 \rceil$ and we denote by $\mathbb{L}^n_i$ the leakage sequences that have i unique values. Equation 8 holds since $\#\mathbb{Q}^n_{\boldsymbol{\ell},i} = m!/(m-i)!$ as shown in Theorem 4. Equation 9 follows from the fact that the set $\mathbb{L}^n_i$ consists of all sequences that have length n and that contain i unique values. This is a standard result from set partitioning and we will provide more details in Theorem 5, this quantity is equal to $\binom{m}{i} \cdot i! \cdot \left\{ {n \atop i} \right\}$. Equation 10 holds as $\#\mathbb{S}_1 = (m-1)^n$ and $\mathbb{S}_0 = \mathbb{Q}^n \setminus \mathbb{S}_1$.

$\blacksquare$

Theorem 3. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m,s}$ and $\mathbf{A} \sim \mathcal{U}^n_m$, then $\mathcal{N}^{+}_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{hyp}}, \mathbf{A}, \delta)$-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{2} - \Gamma_1, \Gamma_2 - \frac{1}{2} \right\},$$

where,

$$\Gamma_1 = \frac{m^n - (m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=\lceil x_1 \rceil}^{m} \frac{(m)_i}{i^{ns}} \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=0}^{\lceil x_1 \rceil - 1} \frac{(m)_i}{i^{ns}} \cdot \left\{ {n \atop i} \right\}.$$

and,

$$\Gamma_2 = \frac{m^n - (m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=\lceil x_1 \rceil}^{m} \frac{(m)_i}{(i!)^s} \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=0}^{\lceil x_1 \rceil - 1} \frac{(m)_i}{(i!)^s} \cdot \left\{ {n \atop i} \right\}$$

where $x_1 = (3m + 1 - \sqrt{5m^2 + 2m + 1})/2$.

Proof. Given that $\mathbf{Q} \sim \mathcal{Z}^n_{m,s}$, we have shown in Theorem 11 that any query sequence with i unique queries can be written as

$$\Pr[\,\mathbf{Q} = \boldsymbol{q}\,] = \frac{1}{H^n_{m,s}} \prod_{k=1}^{i} \frac{1}{k^{s \cdot a_k}} \le \frac{1}{H^n_{m,s} \cdot (i!)^s}.$$

And from Theorem 7,

$$\Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \ge \frac{1}{H^n_{m,s} \cdot i^{-ns}}.$$

From Theorem 2, we know that

$$\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &= \frac{1}{m! \cdot m^n} \sum_{\boldsymbol{q} \in \mathbb{S}_0} \Bigg(\sum_{i=\lceil x_1 \rceil}^{m} \Bigg(\sum_{\boldsymbol{\ell} \in \mathbb{L}^n_i} \Bigg((m-i)! \cdot \sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell},i}} \Pr[\,\mathbf{Q} = \boldsymbol{q}'\,] \Bigg) \Bigg) \Bigg) \\
&\quad + \frac{1}{m! \cdot m^n} \sum_{\boldsymbol{q} \in \mathbb{S}_1} \Bigg(\sum_{i=0}^{\lceil x_1 \rceil - 1} \Bigg(\sum_{\boldsymbol{\ell} \in \mathbb{L}^n_i} \Bigg((m-i)! \cdot \sum_{\mathbf{q}' \in \mathbb{Q}^n_{\boldsymbol{\ell},i}} \Pr[\,\mathbf{Q} = \boldsymbol{q}'\,] \Bigg) \Bigg) \Bigg) \\
&\le \frac{m^n - (m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=\lceil x_1 \rceil}^{m} \frac{(m)_i}{(i!)^s} \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=0}^{\lceil x_1 \rceil - 1} \frac{(m)_i}{(i!)^s} \cdot \left\{ {n \atop i} \right\} \qquad (11)
\end{aligned}$$

where $(m)_i$ is the falling factorial. And similarly, we can show that

$$\Pr[\,\mathbf{CHR} = 1\,] \ge \frac{m^n - (m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=\lceil x_1 \rceil}^{m} \frac{(m)_i}{i^{ns}} \cdot \left\{ {n \atop i} \right\} + \frac{(m-1)^n}{m^n \cdot H^n_{m,s}} \sum_{i=0}^{\lceil x_1 \rceil - 1} \frac{(m)_i}{i^{ns}} \cdot \left\{ {n \atop i} \right\}.$$

Leveraging the result from Lemma E.1 concludes the proof.

■

B Proofs for Section 6 (Full Recovery Against Query Equality)

Theorem 4. *The i.i.d. query equality network* $\mathcal{N}^+_{\mathbf{QEQ}}$ *is* $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$*-coherent, with*

$$\varepsilon = \left| \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_1} \frac{1}{\# S_{f(\boldsymbol{q})}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \Bigg) - \frac{1}{m^n} \right|,$$

where $S_{f(\boldsymbol{q})} \overset{\circ}{=} \mathsf{map}_{\mathbf{A}|f(\boldsymbol{q})}$, $\mathbb{Q}^n_1 \overset{\circ}{=} \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{f(\boldsymbol{q})}\}$ *and* $m \overset{\circ}{=} \#\mathbb{Q} = \#\mathbb{L}$.

Proof. Recall that F is the random variable that outputs a bijection uniformly at random in from $\mathbb{F} \overset{\circ}{=} [\mathbb{Q} \to \mathbb{L}]$, where $\#\mathbb{Q} = \#\mathbb{L} = m$. It follows then that for all $f \in \mathbb{F}$, $\Pr[\, F = f\,] = 1/m!$. Recall that $\mathbf{Q} = (\mathrm{Q}_1, \cdots, \mathrm{Q}_n)$ is a sequence of n random variables distributed according to the query distribution. We then have,

$$\begin{aligned}\Pr[\,\mathbf{CHR} = 1\,] &= \sum_{f \in \mathbb{F}} \Pr[\,\mathbf{CHR} = 1 | F = f\,] \cdot \Pr[\,F = f\,] \\ &= \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n} \Pr[\,\mathbf{CHR} = 1 | F = f, \mathbf{Q} = \boldsymbol{q}\,] \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \bigg).\end{aligned}$$

We decompose the query sequence space $\mathbb{Q}^n$ in two disjoint sets: (1) $\mathbb{Q}_1^n := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{\boldsymbol{\ell}}\}$; and (2) $\mathbb{Q}_1^n := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \notin S_{\boldsymbol{\ell}}\}$ where $S_{\boldsymbol{\ell}} := \mathsf{map}_{\mathbf{A}|\boldsymbol{\ell}}$, where $\boldsymbol{\ell} = f(\boldsymbol{q})$. It is clear that $\mathbb{Q}^n = \mathbb{Q}_1^n \bigcup \mathbb{Q}_2^n$ and therefore that,

$$\begin{aligned}\Pr[\,\mathbf{CHR} = 1\,] = \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \bigg(&\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \Pr[\,\mathbf{CHR} = 1 | F = f, \mathbf{Q} = \boldsymbol{q}\,] \\ &+ \sum_{\boldsymbol{q} \in \mathbb{Q}_2^n} \Pr[\,\mathbf{CHR} = 1 | F = f, \mathbf{Q} = \boldsymbol{q}\,] \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \bigg).\end{aligned}$$

Observe that the probability of the event $\{\mathbf{CHR} = 1 | F = f, \mathbf{Q} = \boldsymbol{q}\}$ is equal to $1/\#S_{\boldsymbol{\ell}}$ when $\boldsymbol{q} \in \mathbb{Q}_1^n$ since we uniformly at random pick a query in $S_{\boldsymbol{\ell}}$. In other words given the fact that we know that the sampled query sequence $\boldsymbol{q}$ is the set $S_{\boldsymbol{\ell}}$, then it is just a matter of guessing the right query sequence. On the other hand, the probability of the same event is equal to 0 when $\boldsymbol{q} \in \mathbb{Q}_2^n$ since by definition $\boldsymbol{q} \notin S_{\boldsymbol{\ell}}$. Then we have

$$\Pr[\,\mathbf{CHR} = 1\,] = \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\#S_{\boldsymbol{\ell}}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \bigg).$$

which concludes out proof.

■

Theorem 5. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}_m^n$ and $\mathbf{A} \sim \mathcal{U}_m^n$, then $\mathcal{N}_{\mathbf{QEQ}}^{+}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent, with*

$$\varepsilon = \left| \frac{1}{m^n} \cdot \sum_{i=1}^{m} \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right|,$$

where $m = \#\mathbb{Q}$.

Proof. We denote by F the random variable that is equal to a bijection in the set $\mathbb{F} := [\mathbb{Q} \to \mathbb{L}]$ such that $\#\mathbb{Q} = \#\mathbb{L} = m$. The variable F is uniformly distributed such that $\Pr[\,F = f\,] = 1/m!$. Let $\mathbf{Q} = (\mathrm{Q}_1, \cdots, \mathrm{Q}_n)$ and $\mathbf{A} = (A_1, \cdots, A_n)$ be two sequences of n random variables. We then have

$$\Pr[\,\mathbf{CHR} = 1\,] = \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\#S_{f(\boldsymbol{q})}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \bigg) \qquad (12)$$

$$= \frac{1}{m! \cdot m^n} \cdot \sum_{f \in \mathbb{F}} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\# S_{f(\boldsymbol{q})}} \Big) \tag{13}$$

$$= \frac{1}{m! \cdot m^n} \cdot \sum_{f \in \mathbb{F}} \Big(\sum_{i=1}^{m} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_{1,i}^n} \frac{1}{\# S_{f(\boldsymbol{q})}} \Big) \Big) \tag{14}$$

Equation 12 is the result of Theorem 4. Equation 13 holds since

$$\Pr\left[\,\mathbf{Q} = \boldsymbol{q}\,\right] = \prod_{i=1}^{n} \Pr\left[\,\mathrm{Q}_i = q_i\,\right] = 1/m^n,$$

and since the variables Q_i are independent. Equation 14 corresponds to the partitioning of the space $\mathbb{Q}_1^n$ into m disjoints sets such that the ith set is composed of query sequences with i unique queries. More formally, we have

$$\mathbb{Q}_{1,i}^n = \Big\{ \boldsymbol{q} = (q_1, \cdots, q_n) \in \ \mathbb{Q}_1^n \mid \#\mathsf{set}(\{q_j\}_{j \in [n]}) = i \Big\},$$

where $\mathsf{set}(\cdot)$ is a set and therefore does not account for redundant elements. It is also easy to see that

$$\mathbb{Q}_1^n = \bigcup_{i=1}^{m} \mathbb{Q}_{1,i}^n,$$

since any query sequence can have at most m unique queries and m is the size of the query space $\mathbb{Q}$. Now we need to perform two steps: (1) calculate the size of $\mathbb{Q}_1^n$; and (2) calculate the size of $S_{f(\boldsymbol{q})}$. First notice that we can rewrite $S_{f(\boldsymbol{q})}$ as follows:

$$\begin{aligned}
S_{f(\boldsymbol{q})} &= \mathsf{map}_{\mathbf{A}|\boldsymbol{\ell}} \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \Pr\left[\,\mathbf{A} = \boldsymbol{q}' \mid \mathbf{L} = f(\boldsymbol{q})\,\right] \Big\} \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \Pr\left[\,\mathbf{L} = f(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q}'\,\right] \cdot \frac{\Pr\left[\,\mathbf{A} = \boldsymbol{q}'\,\right]}{\Pr\left[\,\mathbf{L} = f(\boldsymbol{q})\,\right]} \Big\} \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \Pr\left[\,\mathbf{L} = f(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q}'\,\right] \cdot \prod_{j=1}^{n} \Pr\left[\,A_j = q'_j\,\right] \Big\} && (15) \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \sum_{f \in \mathbb{F}} \Pr\left[\,\mathbf{L} = f(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q}', F = f\,\right] \cdot \Pr\left[\,F = f\,\right] \cdot \prod_{j=1}^{n} \Pr\left[\,A_j = q'_j\,\right] \Big\} \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \sum_{f \in \mathbb{F}} \Pr\left[\,\mathbf{L} = f(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q}', F = f\,\right] \cdot \prod_{j=1}^{n} \Pr\left[\,A_j = q'_j\,\right] \Big\} && (16)
\end{aligned}$$

Equation 15 follows from the independence of Q_i's while Equation 16 holds since $\Pr\left[\,F = f\,\right]$ is a constant. Also note that

$$\Pr\left[\,\mathbf{L} = f(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q}', F = f\,\right] = \begin{cases} 1 & \text{if } f(q'_i) = f(q_i) \ \forall i \in [n] \\ 0 & \text{otherwise.} \end{cases}$$

We now need to compute the number of functions f for which

$$\Pr\left[\,\mathbf{L} = f(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q'}, F = f\,\right] = 1.$$

We show that the number of such functions is $(m - \lambda)!$, where λ is the number of unique elements in $f(\boldsymbol{q})$. First, notice that for fixed leakage $\boldsymbol{\ell} = f(\boldsymbol{q})$, the *possible* query sequences $\boldsymbol{q'} \in \mathbb{Q}^n$ that can lead to $\boldsymbol{\ell}$ are the sequences such that for all $1 \leq j, k \leq n$, $q'_j = q'_k$ if and only if $\ell_j = \ell_k$. In particular, we define

$$\mathbb{Q}^n_{f(\boldsymbol{q})} \overset{\circ}{=} \Big\{ \boldsymbol{q'} \in \mathbb{Q}^n \mid q'_j = q'_k \text{ iff } f(q_j) = f(q_k)\ \forall j, k \in [n] \Big\}.$$

Based on this observation, it is easy to see that the number of unique queries in any query sequence $\boldsymbol{q'} \in \mathbb{Q}^n_{f(\boldsymbol{q})}$ is equal to the number of unique leakage values in $f(\boldsymbol{q})$ such that $\#\mathsf{set}(\{q'_j\}_{j\in[n]}) = \#\mathsf{set}(\{f(q_j)\}_{j\in[n]})$. Let λ be the size of $\mathsf{set}(\{q'_j\}_{j\in[n]})$. That is, given a fixed leakage $f(\boldsymbol{q})$, and a *possible* query sequence $\boldsymbol{q'} \in \mathbb{Q}^n_{f(\boldsymbol{q})}$, the number of functions f that verify these constraints is equal to $(m - \lambda)!$. To see why, note that in both $\boldsymbol{q'}$ and $f(\boldsymbol{q})$ there are λ unique queries and leakage values, respectively. In particular, these λ values define a part of the bijection f but there are $m - \lambda$ points in the space that are still undefined. That is, there are $(m - \lambda)!$ possible functions, f, for the remaining values. So now we can plug this result in the equation above such that

$$\begin{aligned}
S_{f(\boldsymbol{q})} &= \arg\max_{\boldsymbol{q'} \in \mathbb{Q}^n_{f(\boldsymbol{q})}} \Big\{ (m - \lambda)! \cdot \prod_{j=1}^{n} \Pr\Big[\, A_j = q'_j \,\Big] \Big\} && (17)\\
&= \arg\max_{\boldsymbol{q'} \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Big\{ \prod_{j=1}^{n} \Pr\Big[\, A_j = q'_j \,\Big] \Big\} \\
&= \arg\max_{\boldsymbol{q'} \in \mathbb{Q}^n_{\boldsymbol{\ell}}} \Big\{ \frac{1}{m^n} \Big\} \\
&= \mathbb{Q}^n_{f(\boldsymbol{q})} && (18)
\end{aligned}$$

Equation 17 follows from Equation 16 by reducing the set of query sequences to the *possible* set of query sequences $\mathbb{Q}^n_{f(\boldsymbol{q})}$ and by counting the number of function f as discussed above. Equation 18 holds since all queries have the same probability equal to $1/m^n$.

Now that we have shown that $S_{f(\boldsymbol{q})} = \mathbb{Q}^n_{f(\boldsymbol{q})}$, we can reduce calculating the size of $S_{f(\boldsymbol{q})}$ to calculating the size of $\mathbb{Q}^n_{f(\boldsymbol{q})}$. In particular, without loss of generality, consider that $f(\boldsymbol{q})$ has λ unique values. Notice that $\mathbb{Q}^n_{f(\boldsymbol{q})}$ is the set of all query sequences composed of λ unique queries such that $q'_j = q'_k$ if and only if $f(q_j) = f(q_k)$, for all $j, k \in [n]$. The number of such sequences is equal to

$$m \cdot (m-1) \cdot (m-2) \cdots (m - \lambda + 1) = \frac{m!}{(m-\lambda)!}.$$

Moreover given that $S_{f(\boldsymbol{q})} = \mathbb{Q}^n_{f(\boldsymbol{q})}$, we can rewrite $\mathbb{Q}^n_{1,i}$ as

$$\mathbb{Q}^n_{1,i} = \Big\{ \boldsymbol{q} \in\ \mathbb{Q}^n_1 \mid \#\mathsf{set}(\{q_j\}_{j\in[n]}) = i \Big\}$$

$$
\begin{aligned}
&= \left\{ \boldsymbol{q} \in \ \mathbb{Q}^n | \ \#\mathsf{set}(\{q_j\}_{j\in[n]}) = i \text{ and } \boldsymbol{q} \in S_{f(\boldsymbol{q})} \right\} \\
&= \left\{ \boldsymbol{q} \in \ \mathbb{Q}^n | \ \#\mathsf{set}(\{q_j\}_{j\in[n]}) = i \text{ and } \boldsymbol{q} \in \mathbb{Q}^n_{f(\boldsymbol{q})} \right\} \\
&= \left\{ \boldsymbol{q} \in \ \mathbb{Q}^n | \ \#\mathsf{set}(\{q_j\}_{j\in[n]}) = i \right\} \\
&= \mathbb{Q}^n_i
\end{aligned}
$$

The last equality holds since the condition $\{\boldsymbol{q} \in \mathbb{Q}^n_{f(\boldsymbol{q})}\}$ is always true. In particular, given a permutation f, $q_j = q_k$ iff $f(q_j) = f(q_k)$, for all $j, k \in [n]$. So now plugging the above results in Equation 16, we obtain

$$
\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &= \frac{1}{m! \cdot m^n} \cdot \sum_{f\in\mathbb{F}} \Bigg(\sum_{i=1}^{m} \Bigg(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_i} \frac{(m-i)!}{m!} \Bigg) \Bigg) \\
&= \frac{1}{m! \cdot m^n} \cdot \sum_{f\in\mathbb{F}} \Bigg(\sum_{i=1}^{m} \sum_{j=1}^{\binom{m}{i}} \Bigg(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_{i,j}} \frac{(m-i)!}{m!} \Bigg) \Bigg) && (19) \\
&= \frac{1}{m! \cdot m^n} \cdot \sum_{f\in\mathbb{F}} \Bigg(\sum_{i=1}^{m} \binom{m}{i} \cdot \frac{(m-i)!}{m!} \cdot \#\mathbb{Q}^n_{i,j^*} \Bigg) \Bigg) && (20)
\end{aligned}
$$

Equation 19 follows from a decomposition of the set $\mathbb{Q}^n_i$ into $\binom{m}{i}$ subsets $\mathbb{Q}^n_{i,j}$. Every subset $\mathbb{Q}^n_{i,j}$ is composed of sequence of queries that have a unique set of queries, and notice that we can create $\binom{m}{i}$ possible sets of unique queries of size i. Equation 20 follows from the fact that these subsets have the same size for a fixed i. We prove this in the following claim.

Claim 2. *For all $j \in \mathbb{N}$, we have*

$$\#\mathbb{Q}^n_{i,j} = i! \cdot \left\{ {n \atop i} \right\}.$$

Proof. Recall that $\mathbb{Q}^n_{i,j}$ is the set of sequences composed of i unique queries. Given a fixed j, the set of unique queries is also fixed and is equal to $\{q^j_1, \cdots, q^j_i\}$, where $q^j_i \in \mathbb{Q}$ for all $i \in [m]$. Counting the number of sequences in $\mathbb{Q}^n_{i,j}$ is equivalent to the following partitioning problem: given a set of n elements, in how many ways can we partition it in i blocks, such that the blocks are distinguishable? To see why, note that the elements in this question are the indexes of the query sequence. Furthermore, a block can be thought of as the assignment of subset of indexes to a query q^j_k for $k \in [i]$.

This answer to the above question in the case of *indistinguishable* block is a standard counting problem where the number of ways is equal to Stirling number of second kind

$$\left\{ {n \atop i} \right\} = \frac{1}{i!} \cdot \sum_{j=1}^{i} (-1)^i \binom{i}{j} (i-j)^n$$

However, in our cases, the i blocks are distinguishable and therefore any permutation should be accounted for which then gives that the total number of ways of partitioning is equal to $i! \cdot \left\{ {n \atop i} \right\}$.

■

Given Claim 20, we now have

$$\begin{aligned}\Pr[\,\mathbf{CHR}=1\,] &= \frac{1}{m!\cdot m^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m}\binom{m}{i}\cdot\frac{(m-i)!}{m!}\cdot i!\cdot\genfrac\{\}{0pt}{}{n}{i}\Big)\Big)\\ &= \frac{1}{m!\cdot m^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m}\genfrac\{\}{0pt}{}{n}{i}\Big)\Big)\\ &= \frac{1}{m^n}\cdot\sum_{i=1}^{m}\genfrac\{\}{0pt}{}{n}{i}\end{aligned}$$

Subtracting by $1/m^n$ and taking the absolute value conclude our proof.

■

Theorem 6. *For all $n\in\mathbb{N}$, if $\mathbf{Q}\sim\mathcal{U}_m^n$ and $\mathbf{A}\sim\mathcal{Z}_{m,s}^n$, then $\mathcal{N}^{+}_{\mathbf{QEQ}}$ is $(\varepsilon,\mathcal{A}_{\mathsf{map}},\mathbf{A},\varphi)$-coherent with*

$$\varepsilon\leq\max\left\{\frac{1}{m^n},\frac{1}{m^n}\cdot\sum_{i=1}^{m}i!\cdot\genfrac\{\}{0pt}{}{n}{i}-\frac{1}{m^n}\right\}$$

where $m=\#\mathbb{Q}$.

Proof. We showed in Equation 14 of the proof of Theorem 5 that

$$\Pr[\,\mathbf{CHR}=1\,]=\frac{1}{m!\cdot m^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m}\Big(\sum_{\boldsymbol{q}\in\mathbb{Q}_{1,i}^n}\frac{1}{\#S_{f(\boldsymbol{q})}}\Big)\Big)$$

where $\mathbb{Q}_{1,i}^n$ is the set of query sequences with i unique queries and such that the queries $\boldsymbol{q}\in S_{f(\boldsymbol{q})}$, for some fixed $f\in\mathbb{F}$. More formally, we have

$$\mathbb{Q}_{1,i}^n=\Big\{\boldsymbol{q}=(q_1,\cdots,q_n)\in\ \mathbb{Q}^n\mid\#\mathsf{set}(\{q_j\}_{j\in[n]})=i\text{ and }\boldsymbol{q}\in S_{f(\boldsymbol{q})}\Big\},$$

We can partition the set $\mathbb{Q}_{1,i}^n$ more in such a way that it contains a fixed set of unique queries. Notice that we have $\binom{m}{i}$ possible combinations. That is, we can write

$$\mathbb{Q}_{1,i}^n=\bigcup_{j=1}^{\binom{m}{i}}\mathbb{Q}_{1,i,j}^n$$

where $\mathbb{Q}_{1,i,j}^n$ is only composed of query sequences that have the same set of unique queries. As a result, we can rewrite the Equation above as

$$\Pr[\,\mathbf{CHR}=1\,]=\frac{1}{m!\cdot m^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m}\Big(\sum_{j=1}^{\binom{m}{i}}\Big(\sum_{\boldsymbol{q}\in\mathbb{Q}_{1,i,j}^n}\frac{1}{\#S_{\boldsymbol{\ell}}}\Big)\Big)\Big).$$

However, we can show that among all of the possible combinations only a single one is valid. This holds given that the probability mass function of the Zipf distribution is strictly non-increasing.

Claim 3. *There exists a* $j^* \in \{1, \cdots, \binom{m}{i}\}$ *such that*

$$\mathbb{Q}^n_{1,i} = \mathbb{Q}^n_{1,i,j^*}.$$

Proof. Consider a query sequence $\boldsymbol{q} \in \mathbb{Q}^n_{1,i}$, and recall that there are different combinations of i queries (there are $\binom{m}{i}$). Given that the A_i's are Zipf-distributed, among all of the possible combinations, there is only a single one, $j^\star$, such that $\mathbb{Q}^n_{1,i,j^*} \neq \emptyset$. In other words, the only sequences $\boldsymbol{q}$ that verify $\boldsymbol{q} \in S_{f(\boldsymbol{q})}$ are the ones in $\mathbb{Q}^n_{1,i,j^*}$. To see why, notice that the probability mass function of a Zipf distribution is a strictly non-increasing function: if $k > r$, then $\frac{k^{-s}}{H_{m,s}} < \frac{r^{-s}}{H_{m,s}}$. So the best combination of queries is the one that has ranks $(1, \cdots, i)$, and therefore these queries are $(\pi^{-1}(1), \cdots, \pi^{-1}(i))$. Without loss of generality, we refer to this combination as the $j^\star$th combination. More formally, recall that we have shown as part of the proof in Theorem 5 that

$$S_{f(\boldsymbol{q})} = \arg \max_{\boldsymbol{q}' \in \mathbb{Q}^n_{f(\boldsymbol{q})}} \left\{ \prod_{j=1}^{n} \Pr\left[\, A_j = q'_j \,\right] \right\},$$

where

$$\mathbb{Q}^n_{f(\boldsymbol{q})} = \left\{ \boldsymbol{q}' \in \mathbb{Q}^n \mid q'_j = q'_k \text{ iff } f(q_j) = f(q_k)\ \forall j,k \in [n] \right\}.$$

We can rewrite the equation above as

$$S_{f(\boldsymbol{q})} = \arg \max_{\boldsymbol{q}' \in \mathbb{Q}^n_{f(\boldsymbol{q})}} \left\{ \frac{1}{H^n_{m,s}} \prod_{i=1}^{\lambda} (a_i^{-s})^{k_i} \right\},$$

where λ is the number of unique values in $f(\boldsymbol{q})$, $\mathbf{a} = (a_1, \cdots, a_\lambda)$ are the ranks of the query sequence $\boldsymbol{q}'$ and $\mathbf{k} = (k_1, \cdots, k_\lambda)$ are the occurrences of the unique values in $f(\boldsymbol{q})$. Observe that the query sequence that maximizes the above quantity is the query sequence where $\mathbf{a} = (1, \cdots, \lambda)$, for a fixed $\mathbf{k}$, s, and λ.

To summarize, for $j \in \{1, \cdots, m\} \setminus j^\star$, and for all $\boldsymbol{q} \in \mathbb{Q}^n_{1,i}$ with the jth combination of unique queries, we have $\boldsymbol{q} \notin S_{f(\boldsymbol{q})}$, which then implies that $\mathbb{Q}^n_{1,i,j} = \emptyset$.

■

Given the result of the above claim, we then have

$$\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &= \frac{1}{m! \cdot m^n} \cdot \sum_{f \in \mathbb{F}} \left(\sum_{i=1}^{m} \left(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j^\star}} \frac{1}{\# S_{f(\boldsymbol{q})}} \right) \right) \\
&\leq \frac{1}{m! \cdot m^n} \cdot \sum_{f \in \mathbb{F}} \left(\sum_{i=1}^{m} \# \mathbb{Q}^n_{1,i,j^\star} \right) \\
&= \frac{1}{m! \cdot m^n} \cdot \sum_{f \in \mathbb{F}} \left(\sum_{i=1}^{m} i! \cdot \left\{ {n \atop i} \right\} \right) \\
&= \frac{1}{m^n} \cdot \sum_{i=1}^{m} i! \cdot \left\{ {n \atop i} \right\}
\end{aligned} \tag{21}$$

where Equation 21 holds since counting the number of query sequences in $\mathbb{Q}^n_{1,i,j^\star}$ is the same as counting the number of ways we can partition a set of n elements into i blocks where the blocks are distinguishable – refer to the proof of Theorem 5 for more details. That is, $\mathbb{Q}^n_{1,i,j^\star} = i! \cdot \left\{ {n \atop i} \right\}$. [10]

Leveraging the result of Lemma E.1 where the lower bound is 0 concludes our proof.

■

Theorem 7. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m,s}$ and $\mathbf{A} \sim \mathcal{Z}^n_{m,s}$, then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{m^n} - \frac{1}{H^n_{m,s}} \cdot \sum_{i=1}^{m} (i!)^{1-s} \cdot \left\{ {n \atop i} \right\}, \frac{1}{H^n_{m,s}} \cdot \sum_{i=1}^{m} i^{-n \cdot s} \cdot \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right\},$$

where $m = \#\mathbb{Q}$.

Proof. Given the result from Theorem 4, we have

$$\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &= \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_1} \frac{1}{\# S_{f(\boldsymbol{q})}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \Bigg) \\
&= \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{i=1}^{m} \Bigg(\sum_{j=1}^{\binom{m}{i}} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j}} \frac{1}{\# S_{f(\boldsymbol{q})}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \Bigg) \Bigg) \Bigg) \\
&= \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{i=1}^{m} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j^\star}} \frac{1}{\# S_{f(\boldsymbol{q})}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \Bigg) \Bigg) \qquad (22)
\end{aligned}$$

where the three equalities above follow from the same arguments in Theorem 5 except that the probability mass function of the multi-variate random variable $\mathbf{Q}$ is Zipf-distributed and therefore is a function of the query sequence contrary to the case of the uniform distribution where all sequences are equally likely. In particular, given that $\mathbf{Q} \sim \mathcal{Z}^n_{m,s}$, then for any $\mathbf{q} \in \mathbb{Q}^n_{1,i,j^\star}$ we have

$$\Pr[\,\mathbf{Q} = \boldsymbol{q}\,] = \prod_{k=1}^{n} \Pr[\,Q_k = q_k\,] = \frac{1}{H^n_{m,s}} \cdot \prod_{k=1}^{i} \frac{1}{(k^s)^{a_k}},$$

where a_k is the multiplicity of the kth query in the $j^\star$th combination and $\sum_{k=1}^{i} a_k = n$. Recall that $j^\star$ represents the index of the combination that corresponds to the i queries that maximize the posterior, and as shown in Theorem 6, $j^\star$ corresponds to the i first queries with the highest ranks in $\mathbf{Q} \sim \mathcal{Z}^n_{m,s}$ given a permutation π. In the following, we are interested in obtaining a lower-bound of the above probability mass function independently of the multiplicities a_k which would allow us to later derive a lower-bound for the coherence probability. In particular, given that $k \leq i$ we have

$$\sum_{k=1}^{i} s \cdot a_k \cdot \log(k) \leq \sum_{k=1}^{i} s \cdot a_k \cdot \log(i)$$

[10] Note that this bound is very loose because we consider $\# S_{f(\boldsymbol{q})}$ to be the maximum (i.e., 1) for all query sequences in $\mathbb{Q}^n_{1,i,j^\star}$. There are many query sequences, however, that do not belong to $S_{f(\boldsymbol{q})}$ but that we are still accounting for in our worst-case bound. In other words, there are query sequences in $\boldsymbol{q} \in \mathbb{Q}^n_{1,i}$ where the combination is $j^\star$, but $\boldsymbol{q} \notin S_{f(\boldsymbol{q})}$. Obtaining a tighter bound is an interesting open problem that would require a more complex counting argument.

$$\sum_{k=1}^{i} \log(k^{s\cdot a_k}) \leq n \cdot \log(i)$$

$$\frac{1}{H^n_{m,s}} \cdot i^{-n\cdot s} \leq \frac{1}{H^n_{m,s}} \cdot \prod_{k=1}^{i} \frac{1}{(k^s)^{a_k}}$$

Given the above, we can now obtain a lower-bound for Equation 22 such that

$$\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &\geq \frac{1}{m! \cdot H^n_{m,s}} \cdot \sum_{f\in\mathbb{F}} \Bigg(\sum_{i=1}^{m} i^{-n\cdot s} \Big(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_{1,i,j^\star}} \frac{1}{\#S_{f(\boldsymbol{q})}} \Big)\Bigg) \\
&\geq \frac{1}{m! \cdot H^n_{m,s}} \cdot \sum_{f\in\mathbb{F}} \Bigg(\sum_{i=1}^{m} i^{-n\cdot s} \Big(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_{1,i,j^\star}} \frac{1}{i!} \Big)\Bigg) && (23) \\
&= \frac{1}{m! \cdot H^n_{m,s}} \cdot \sum_{f\in\mathbb{F}} \Bigg(\sum_{i=1}^{m} i^{-n\cdot s} i! \cdot \left\{ {n \atop i} \right\} \cdot \frac{1}{i!} \Big)\Bigg) && (24) \\
&= \frac{1}{H^n_{m,s}} \cdot \sum_{i=1}^{m} i^{-n\cdot s} \cdot \left\{ {n \atop i} \right\}.
\end{aligned}$$

Inequality 23 follows from the observation that for any query sequence $\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j^\star}$, we have $\#S_{f(\boldsymbol{q})} \leq i!$. The upper-bound is reached when all i unique queries appear exactly the same number of times in $\boldsymbol{q}$. Equality 24 holds as the size of the set $\mathbb{Q}^n_{1,i,j^\star}$ is equal to $i! \cdot \left\{ {n \atop i} \right\}$, refer to Theorem 6 for more details.

We now show an upper bound for Equation 22 by first observing that

$$\Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \frac{1}{H^n_{m,s}} \cdot \prod_{k=1}^{i} \frac{1}{(k^s)^{a_k}} \leq \frac{1}{H^n_{m,s}} \frac{1}{(i!)^s},$$

and this holds since multiplicities a_k for $k \in [i]$ are non-increasing with the higher associated to smallest rank – recall that this is the way how one could maximize the Zipf probability mass function as shown in Theorem 6. Given the above inequality and the fact that $\#S_{f(\boldsymbol{q})} \geq 1$ for all $\boldsymbol{q} \in \mathbb{Q}^n$, it is easy to show that

$$\Pr[\,\mathbf{CHR} = 1\,] \leq \frac{1}{H^n_{m,s}} \cdot \sum_{i=1}^{m} (i!)^{1-s} \cdot \left\{ {n \atop i} \right\}.$$

Leveraging the result of Lemma E.1, we conclude our proof.

■

Theorem 8. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}^n_{m_q}$, $\mathbf{A} \sim \mathcal{U}^n_{m_a}$, and $\mathbb{A} \subset \mathbb{Q}$ then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon = \left| \frac{1}{m^n_q} \cdot \sum_{i=1}^{m_a} \left\{ {n \atop i} \right\} - \frac{1}{m^n_q} \right|.$$

Proof. From Theorem 4, we know

$$\begin{aligned}\Pr[\,\mathbf{CHR}=1\,] &= \frac{1}{m_q!}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{\boldsymbol{q}\in\mathbb{Q}_1^n}\frac{1}{\#S_{f(\boldsymbol{q})}}\cdot\Pr[\,\mathbf{Q}=\boldsymbol{q}\,]\Big)\\ &= \frac{1}{m_q!\cdot m_q^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m_q}\Big(\sum_{\boldsymbol{q}\in\mathbb{Q}_{1,i}^n}\frac{1}{\#S_{f(\boldsymbol{q})}}\Big)\Big)\end{aligned}$$

Now, observe that an adversary can only output query sequences that it knows which implies that any query sequence that is composed of queries that are not in $\mathbb{A}$ will never be part of $S_{f(\boldsymbol{q})}$. More formally, if $\boldsymbol{q}\in\mathbb{Q}_{1,i}^n\setminus\mathbb{A}_{1,i}^n$ then $\boldsymbol{q}\notin S_{f(\boldsymbol{q})}$. Also observe that the number of possible unique queries in $\mathbb{A}_{1,i}^n$ is m_a which is smaller than $m_q=\#\mathbb{Q}$. Then, we obtain

$$\Pr[\,\mathbf{CHR}=1\,] = \frac{1}{m_q!\cdot m_q^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m_a}\Big(\sum_{\boldsymbol{q}\in\mathbb{A}_{1,i}^n}\frac{1}{\#S_{f(\boldsymbol{q})}}\Big)\Big)$$

Following the same proof in Theorem 5, we can show that for all $\boldsymbol{q}\in\mathbb{A}_{1,i}^n$ for $i\in[m_a]$,

$$\#S_{f(\boldsymbol{q})}=\frac{m_a!}{(m_a-i)!}.$$

Putting everything together we obtain,

$$\begin{aligned}\Pr[\,\mathbf{CHR}=1\,] &= \frac{1}{m_q!\cdot m_q^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m_a}\Big(\sum_{j=1}^{\binom{m_a}{i}}\frac{(m_a-i)!}{m_a!}\cdot\#\mathbb{A}_{1,i,j}^n\Big)\Big) && (25)\\ &= \frac{1}{m_q!\cdot m_q^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m_a}\binom{m_a}{i}\cdot\frac{(m_a-i)!}{m_a!}\cdot i!\cdot\left\{{n\atop i}\right\}\Big) && (26)\\ &= \frac{1}{m_q^n}\cdot\sum_{i=1}^{m_a}\left\{{n\atop i}\right\}.\end{aligned}$$

Note that in Equation 25, we further partition the set $\mathbb{A}_{1,i}=\bigcup_{j=1}^{\binom{m_a}{i}}\mathbb{A}_{1,i,j}$ where $\mathbb{A}_{1,i,j}$ represents the set of all query sequences that have a fixed set of i distinct queries. Equation 26 holds as the size of the set $\mathbb{A}_{1,i,j}$ is equal to $i!\cdot\left\{{n\atop i}\right\}$ following the same argument made in Theorem 5. Finally, subtracting $1/m_q^n$ and taking the absolute value concludes our proof.

■

Theorem 9. *For all $n\in\mathbb{N}$, if $\mathbf{Q}\sim\mathcal{U}_{m_q}^n$, $\mathbf{A}\sim\mathcal{U}_{m_a}^n$, and $\mathbb{Q}\subset\mathbb{A}$ then $\mathcal{N}_{\mathbf{QEQ}}^{+}$ is $(\varepsilon,\mathcal{A}_{\mathsf{map}},\mathbf{A},\varphi)$-coherent with*

$$\varepsilon\leq\frac{1}{m_q^n}\cdot\sum_{i=1}^{m_q}\Big(\frac{m_q\cdot e}{m_a}\Big)^i\cdot\left\{{n\atop i}\right\}.$$

Proof. Similar to Theorem 8, we have

$$\Pr[\,\mathbf{CHR}=1\,] = \frac{1}{m_q!\cdot m_q^n}\cdot\sum_{f\in\mathbb{F}}\Big(\sum_{i=1}^{m_q}\Big(\sum_{\boldsymbol{q}\in\mathbb{Q}_{1,i}^n}\frac{1}{\#S_{f(\boldsymbol{q})}}\Big)\Big)$$

$$
\begin{aligned}
&= \frac{1}{m_q! \cdot m_q^n} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{i=1}^{m_q} \Bigg(\sum_{j=1}^{\binom{m_q}{i}} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j}} \frac{1}{\# S_{f(\boldsymbol{q})}} \Bigg) \Bigg) \Bigg) \\
&= \frac{1}{m_q! \cdot m_q^n} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{i=1}^{m_q} \Bigg(\sum_{j=1}^{\binom{m_q}{i}} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j}} \frac{(m_a - i)!}{m_a!} \Bigg) \Bigg) \qquad (27) \\
&= \frac{1}{m_q! \cdot m_q^n} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{i=1}^{m_q} \binom{m_q}{i} \cdot i! \cdot \begin{Bmatrix} n \\ i \end{Bmatrix} \frac{(m_a - i)!}{m_a!} \Bigg) \qquad (28) \\
&= \frac{1}{m_q^n} \cdot \sum_{i=1}^{m_q} \binom{m_q}{i} \binom{m_a}{i}^{-1} \cdot \begin{Bmatrix} n \\ i \end{Bmatrix}
\end{aligned}
$$

Equality 27 follows from the fact that the most likely sequences the adversary outputs given a leakage sequence will include query sequences that in $\mathbb{A}^n$. This is why the size of the set $S_{f(\boldsymbol{q})}$ is function of m_a and not m_q. Equality 28 holds since the size of the set $\mathbb{Q}^n_{1,i,j}$ is equal to $i! \cdot \left\{ {n \atop i} \right\}$ as shown in Theorem 5. We further simplify the above equation by observing that for all $1 \leq k \leq n$,

$$
\left(\frac{n}{k} \right)^k \leq \binom{n}{k} \leq \left(\frac{n \cdot e}{k} \right)^k ,
$$

We then obtain,

$$
\Pr \left[\, \mathbf{CHR} = 1 \, \right] \leq \frac{1}{m_q^n} \cdot \sum_{i=1}^{m_q} \left(\frac{m_q \cdot e}{m_a} \right)^i \cdot \begin{Bmatrix} n \\ i \end{Bmatrix} .
$$

Subtracting $1/m^n$ and taking the absolute value completes the proof.

■

Theorem 10. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{Z}^n_{m_q,s_q}$, $\mathbf{A} \sim \mathcal{Z}^n_{m_a,s_a}$, and $\mathbb{A} \subset \mathbb{Q}$ then $\mathcal{N}^+_{\mathbf{QEQ}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$
\varepsilon \leq \max \left\{ \frac{1}{m_q^n} - \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{m_a} \frac{\left\{ {n \atop i} \right\}}{(\gamma + i)^{s_q \cdot n}}, \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \cdot \sum_{i=1}^{m_a} i! \cdot \begin{Bmatrix} n \\ i \end{Bmatrix} - \frac{1}{m_q^n} \right\}
$$

Proof. Similar to Theorem 9, we have

$$
\begin{aligned}
\Pr \left[\, \mathbf{CHR} = 1 \, \right] &= \frac{1}{m_q!} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{i=1}^{m_q} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j^\star}} \frac{1}{\# S_{f(\boldsymbol{q})}} \cdot \Pr \left[\, \mathbf{Q} = \boldsymbol{q} \, \right] \Bigg) \Bigg) \\
&= \frac{1}{m_q!} \cdot \sum_{f \in \mathbb{F}} \Bigg(\sum_{i=1}^{m_a} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{A}^n_{1,i,j^\star}} \frac{1}{\# S_{f(\boldsymbol{q})}} \cdot \Pr \left[\, \mathbf{Q} = \boldsymbol{q} \, \right] \Bigg) \Bigg) \qquad (29)
\end{aligned}
$$

Equation 29 follows from the fact that the adversary can only output sequences in $\mathbb{A}^n$ and that the possible number of unique queries is at most m_a. Recall that $\mathbb{A}_{1,i,j^\star}$ represents the set of all query sequences that belong to $S_{f(\boldsymbol{q})}$ and that are composed of a fixed set of i unique queries. Now, let's denote by γ the maximum rank in $\mathbb{Q}$ of any query in $\mathbb{A}$. More formally, let $\gamma \in [m_q - m_a]$ such that for all $q \in \mathbb{A}$, $\pi_q[q] \geq \gamma$, then we can write for all $\boldsymbol{q} \in \mathbb{A}_{1,i,j^\star}$,

$$\Pr[\,\mathbf{Q} = \boldsymbol{q}\,] = \prod_{k=1}^{n} \Pr[\,Q_k = q_k\,] = \frac{1}{H^n_{m_q,s_q}} \cdot \prod_{k=\gamma}^{i+\gamma} \frac{1}{(k^{s_q})^{a_k}}$$

Following the same algebraic computations as in Theorem 7, we can show that for all $i \in [m_a]$,

$$\frac{1}{H^n_{m_q,s_q} \cdot (\gamma+i)^{s_q \cdot n}} \leq \frac{1}{H^n_{m_q,s_q}} \cdot \prod_{k=\gamma}^{i+\gamma} \frac{1}{(k^{s_q})^{a_k}} \leq \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}}$$

Given the above, we can rewrite Equation 29 such that

$$\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &\geq \frac{1}{m_q! \cdot H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \cdot \sum_{f \in \mathbb{F}} \left(\sum_{i=1}^{m_a} \left(\sum_{\boldsymbol{q} \in \mathbb{A}^n_{1,i,j^\star}} \frac{1}{\#S_{f(\boldsymbol{q})}} \right) \right) \\
&\geq \frac{1}{m_q! \cdot H^n_{m_q,s}} \cdot \sum_{f \in \mathbb{F}} \left(\sum_{i=1}^{m_a} \left(\sum_{\boldsymbol{q} \in \mathbb{A}^n_{1,i,j^\star}} \frac{1}{i! \cdot (\gamma+i)^{s_q \cdot n}} \right) \right) \qquad (30) \\
&\geq \frac{1}{H^n_{m_q,s}} \cdot \sum_{i=1}^{m_a} \frac{{n \brace i}}{(\gamma+i)^{s_q \cdot n}}
\end{aligned}$$

Equation 30 holds since as shown in Theorem 6, for all $\boldsymbol{q} \in \mathbb{A}^n_{1,i,j^\star}$, $\#S_{f(\boldsymbol{q})} \leq i!$. Similarly, we show that

$$\Pr[\,\mathbf{CHR} = 1\,] \leq \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \cdot \sum_{i=1}^{m_a} i! \cdot \left\{ \begin{matrix} n \\ i \end{matrix} \right\}$$

Using Lemma E.1 we conclude our proof.

■

Theorem 11. *For all* $n \in \mathbb{N}$, *if* $\mathbf{Q} \sim \mathcal{Z}^n_{m_q,s_q}$, $\mathbf{A} \sim \mathcal{Z}^n_{m_a,s_a}$, *and* $\mathbb{Q} \subset \mathbb{A}$ *then* $\mathcal{N}^+_{\mathbf{VOL}}$ *is* $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$*-coherent with*

$$\varepsilon \leq \max\left\{ \frac{1}{m_q^n} - \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{\theta} \frac{{n \brace i}}{i^{s_q \cdot n}}, \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{\theta} (i!)^{1-s_q} \cdot \left\{ \begin{matrix} n \\ i \end{matrix} \right\} - \frac{1}{m_q^n} \right\}$$

Proof. Similar to Theorem 10, we have

$$\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &= \frac{1}{m_q!} \cdot \sum_{f \in \mathbb{F}} \left(\sum_{i=1}^{m_q} \left(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j^\star}} \frac{1}{\#S_{f(\boldsymbol{q})}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \right) \right) \\
&= \frac{1}{m_q!} \cdot \sum_{f \in \mathbb{F}} \left(\sum_{i=1}^{\theta} \left(\sum_{\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j^\star}} \frac{1}{\#S_{f(\boldsymbol{q})}} \cdot \Pr[\,\mathbf{Q} = \boldsymbol{q}\,] \right) \right) \qquad (31)
\end{aligned}$$

In Equation 31, we know that if the number of unique queries exceeds θ, then there is no possible query sequence that can belong to $S_{f(\boldsymbol{q})}$. To see why, notice that when the number of unique queries is strictly higher than θ them any sequence the adversary is going to output has to contain a query that belongs to $\mathbb{A}$ but not to $\mathbb{Q}$. Now, we need to bound the probability of observing a query sequence $\boldsymbol{q}$. Following the same steps in Theorem 7, we can show that for all $\boldsymbol{q} \in \mathbb{Q}^n_{1,i,j^\star}$,

$$\frac{1}{H^n_{m_q,s_q} \cdot i^{s_q \cdot n}} \leq \prod_{k=1}^{n} \Pr\left[\, Q_k = q_k \,\right]$$

and,

$$\Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] = \prod_{k=1}^{n} \Pr\left[\, Q_k = q_k \,\right] = \frac{1}{H^n_{m_q,s_q}} \cdot \prod_{k=1}^{i} \frac{1}{(k^{s_q})^{a_k}} \leq \frac{1}{H^n_{m_q,s_q} \cdot (i!)^{s_q}}$$

The last inequality holds as the multiplicities, $(a_k)_{k\in[i]}$, have to verify $a_1 \geq a_2 \geq \cdots \geq a_i$ given that the highest rank will always be assigned to the most frequent query in a given query sequence, refer to Theorem 6 for more details. Given the above, we then obtain

$$\frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{\theta} \frac{\left\{ {n \atop i} \right\}}{i^{s_q \cdot n}} \leq \Pr\left[\, \mathbf{CHR} = 1 \,\right] \leq \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{\theta} (i!)^{1-s_q} \cdot \left\{ {n \atop i} \right\}.$$

Applying Lemma E.1 concludes our proof.

■

C Proofs for Section 7 (Full Recovery Against Volume)

Theorem 12. *The i.i.d. volume network* $\mathcal{N}^+_{\mathbf{VOL}}$ *is* $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \psi)$*-coherent with*

$$\varepsilon = \left| \sum_{d\in\mathbb{D}_N} \Big(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_1} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, D = d \,\right] \Big) - \frac{1}{m^n} \right|,$$

where $S_{\#d(\boldsymbol{q})} := \mathsf{map}_{\mathbf{A_Q}|\#d(\boldsymbol{q})}$ *and* $\mathbb{Q}^n_1 := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{\#d(\boldsymbol{q})}\}$.

Proof. We have by definition,

$$\begin{aligned}
\Pr\left[\, \mathbf{CHR} = 1 \,\right] &= \sum_{d\in\mathbb{D}_N} \Pr\left[\, \mathbf{CHR} = 1 | D = d \,\right] \cdot \Pr\left[\, D = d \,\right] \\
&= \sum_{d\in\mathbb{D}_N} \Big(\sum_{\boldsymbol{q}\in\mathbb{Q}^n} \Pr\left[\, \mathbf{CHR} = 1 | D = d, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, D = d \,\right] \Big)
\end{aligned}$$

We then divide the query sequence space $\mathbb{Q}^n$ in two disjoint sets: (1) $\mathbb{Q}^n_1 := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{\#d(\boldsymbol{q})}\}$, and (2) $\mathbb{Q}^n_1 := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \notin S_{\#d(\boldsymbol{q})}\}$ where $S_{\#d(\boldsymbol{q})} := \mathsf{map}_{\mathbf{A_Q}|\#d(\boldsymbol{q})}$. Similarly to Theorem 4, we can easily show that

$$\begin{aligned}
\Pr\left[\, \mathbf{CHR} = 1 \,\right] = \sum_{d\in\mathbb{D}_N} \Big(&\sum_{\boldsymbol{q}\in\mathbb{Q}^n_1} \Pr\left[\, \mathbf{CHR} = 1 | D = d, \mathbf{Q} = \boldsymbol{q} \,\right] \\
&+ \sum_{\boldsymbol{q}\in\mathbb{Q}^n_2} \Pr\left[\, \mathbf{CHR} = 1 | D = d, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, \mathbf{Q} = \boldsymbol{q} \,\right] \cdot \Pr\left[\, D = d \,\right] \Big)
\end{aligned}$$

$$= \sum_{d \in \mathbb{D}_N} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\# S_{\# d(\boldsymbol{q})}} \cdot \Pr[\, \mathbf{Q} = \boldsymbol{q} \,] \cdot \Pr[\, D = d \,] \Big).$$

Finally, subtracting $1/m^n$ and taking the absolute value concludes the proof.

■

Theorem 13. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}_m^n$, $\mathbf{A} \sim \mathcal{U}_m^n$, $D \sim \mathcal{U}_{\mathbb{D}_N}$ and $A_D \sim \mathcal{U}_{\mathbb{D}_N}$ then $\mathcal{N}_{\mathbf{VOL}}^{+}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon \leq \max \left\{ \frac{1}{m^n}, \frac{1}{m^n} \cdot \sum_{i=1}^{m'} \left\{ {n \atop i} \right\} - \frac{1}{m^n} \right\}$$

where $m' = \min(m, \sqrt{2N})$

Proof. First, we rewrite the result of Theorem 12 as

$$\begin{aligned} \Pr[\, \mathbf{CHR} = 1 \,] &= \sum_{d \in \mathbb{D}_N} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\# S_{\# d(\boldsymbol{q})}} \cdot \Pr[\, \mathbf{Q} = \boldsymbol{q} \,] \cdot \Pr[\, D = d \,] \Big) \\ &= \frac{1}{m^n \cdot \binom{N-1}{m-1}} \cdot \sum_{d \in \mathbb{D}_N} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\# S_{\# d(\boldsymbol{q})}} \Big) \end{aligned}$$

where the second Equation follows from plugging in the probability mass functions of D and $\mathbf{Q}$ and from $\mathbb{Q}_1^n := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{\# d(\boldsymbol{q})}\}$. Now we are interested in finding a more concrete representation of the set $S_{\# d(\boldsymbol{q})}$ so we can characterize the possible query sequences. Specifically, we have by definition

$$\begin{aligned} S_{\# d(\boldsymbol{q})} &= \mathsf{map}_{\mathbf{A_Q} | \# d(\boldsymbol{q})} \\ &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \Pr[\, \mathbf{A_Q} = \boldsymbol{q}' \,|\, \mathbf{L} = \# d(\boldsymbol{q}) \,] \Big\} \\ &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \Pr[\, \mathbf{L} = \# d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}' \,] \cdot \frac{\Pr[\, \mathbf{A_Q} = \boldsymbol{q}' \,]}{\Pr[\, \mathbf{L} = \# d(\boldsymbol{q}) \,]} \Big\} \\ &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \sum_{d' \in \mathbb{D}_N} \Pr[\, \mathbf{L} = \# d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', A_D = d' \,] \cdot \Pr[\, A_D = d' \,] \cdot \Pr[\, \mathbf{A_Q} = \boldsymbol{q}' \,] \Big\} \\ &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \sum_{d' \in \mathbb{D}_N} \Pr[\, \mathbf{L} = \# d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', A_D = d' \,] \cdot \frac{1}{m^n \cdot \binom{N-1}{m-1}} \Big\} && (32) \\ &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \sum_{d' \in \mathbb{D}_N} \Pr[\, \mathbf{L} = \# d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', A_D = d' \,] \Big\} && (33) \end{aligned}$$

Equation 32 holds since both $\mathbf{A_Q}$ and A_D are uniform while Equation 33 holds because removing constants does not change the argmax. On the other hand, note that

$$\Pr\left[\,\mathbf{L} = \#d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', A_D = d'\,\right] = \begin{cases} 1 & \text{if } \#d'(q'_i) = \#d(q_i) \ \forall i \in [n] \\ 0 & \text{otherwise.} \end{cases}$$

In particular, the constraint $\#d'(q'_i) = \#d(q_i)$ implies that the query sequences $\mathbf{q'}$ have to have at least λ unique queries where λ is the number of unique volumes such that $\lambda := \#\mathsf{set}(\{\#d(q_i)\}_{i\in[n]})$. More formally, the set of such query sequences is

$$\mathbb{Q}^n_{\#d(\boldsymbol{q})} = \Big\{ \boldsymbol{q}' \in \mathbb{Q}^n \mid q'_i \neq q'_j \text{ if } \#d(q_i) \neq \#d(q_j) \ \forall i, j \in [n] \Big\}.$$

It is important to note that the number of unique queries can be larger than λ and still verify the condition above. For instance, consider queries that have the same volume. We denote the number of unique queries in $\boldsymbol{q}'$ to be equal to γ where $\gamma := \#\mathsf{set}(\{q'_i\}_{i\in[n]})$. Furthermore, the condition also implies that, given $\boldsymbol{q}'$, only a subset of functions in $\mathbb{D}_N$ are possible. We denote this by

$$\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} = \Big\{ d' \in \mathbb{D}_N \mid \#d'(q'_i) = \#d(q_i) \ \forall i \in [n] \Big\}.$$

So we can rewrite Equation 33 as

$$S_{\#d(\boldsymbol{q})} = \arg \max_{\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q})}} \Big\{ \#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} \Big\}.$$

Now we need to identify which query sequence(s) maximize $\#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'}$. Notice that this set is largest when the number of unique queries γ is the smallest. The reason is that the more queries we know, the more of the function we know and, therefore, the remaining queries will have to "share" a much smaller set of volumes. We prove this observation more formally in the claim below.

Claim 4. *Given a volume leakage $\#d(\boldsymbol{q})$,*

$$\max_{\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q})}} \#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} = \max_{\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q}),\lambda}} \#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} = \binom{N - \sum_{l \in \mathsf{Uniq}} l - 1}{m - \lambda - 1}$$

where $\mathbb{Q}^n_{\#d(\boldsymbol{q}),\gamma} = \Big\{ \boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q})} \mid \#\mathsf{set}(\{q'_i\}_{i\in[n]} = \gamma) \Big\}$, $\mathsf{Uniq} = \mathsf{set}(\{\#d(q_i)\}_{i\in[n]})$ *and* $\lambda = \#\mathsf{Uniq}$.

Proof. Consider a query sequence $\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q})}$ with γ unique queries, i.e., $\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q}),\gamma}$. Without loss of generality, assume that these queries are $\{q_{m-\gamma+1}, \cdots, q_m\}$. In addition we know that these γ queries have volumes in $\mathsf{set}(\{\#d(q_i)\}_{i\in[n]})$ and that at least λ have distinct volumes—this simply follows from the definition of $\mathbb{Q}^n_{\#d(\boldsymbol{q})}$. For now, assume that the $\gamma - \lambda$ queries got assigned to arbitrary volumes. We denote the volumes of the queries $\{l^\star_1, \cdots, l^\star_\gamma\}$ such that the first λ elements in this set are unique. Now counting the elements in $\#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'}$ is equivalent to solving the following equation

$$\#d'(q_1) + \cdots + \#d'(q_{m-\gamma}) = N - \sum_{i=1}^{\gamma} l^\star_i$$

Based on a stars and bars argument (refer to Claim 1 for more details), we have the number of possible solutions and therefore the number of possible multi-maps is equal to

$$\binom{N-\sum_{i=1}^{\gamma} l_i^\star - 1}{m-\gamma-1}.$$

In the case of $\gamma = \lambda$, the above quantity is equal to

$$\binom{N-\sum_{i=1}^{\lambda} l_i^\star - 1}{m-\lambda-1}.$$

The only remaining step is to show that the above value reaches its maximum when $\gamma = \lambda$. For this we make use of Pascal's rule that states that for al $k \in [n-1]$,

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

where we start by $n = N - \sum_{i=1}^{\lambda} l_i^\star - 1$ and $k = m - \lambda - 1$ and we apply several recursive steps before reaching $n' = N - \sum_{i=1}^{\gamma} l_i^\star - 1$ and $k' = m - \gamma - 1$ since $n' \le n$ and $k' \le k$. Note that all of these recursive steps are additive and therefore we have for all $\gamma \ge \lambda$

$$\binom{N-\sum_{i=1}^{\lambda} l_i^\star - 1}{m-\lambda-1} \ge \binom{N-\sum_{i=1}^{\gamma} l_i^\star - 1}{m-\gamma-1}$$

which concludes our proof.

■

Given the claim above we can rewrite $S_{\#d(\boldsymbol{q})}$ as

$$\begin{aligned} S_{\#d(\boldsymbol{q})} &= \arg \max_{\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q}),\lambda}} \left\{ \#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} \right\} \\ &= \arg \max_{\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q}),\lambda}} \left\{ \binom{N - \sum_{l \in \mathsf{Uniq}} l - 1}{m - \lambda - 1} \right\} \\ &= \mathbb{Q}^n_{\#d(\boldsymbol{q}),\lambda} = \left\{ \boldsymbol{q}' \in \mathbb{Q}^n \mid q'_i = q'_j \text{ iff } \#d(q_i) = \#d(q_j)\ \forall i,j \in [n] \right\} \end{aligned} \tag{34}$$

Equation 34 holds because of the following: recall that $\mathbb{Q}^n_{\#d(\boldsymbol{q}),\lambda}$ is the set of possible query sequences $\boldsymbol{q}'$ such that $q'_i \neq q'_j$ if $\#d(q_i) \neq \#d(q_j)$ for all $i, j \in [n]$. That is, it leaves open the possibility that a different query can occur in the positions in $d(\boldsymbol{q})$ that have the same volume. If that occurs then the number of unique queries will strictly exceed λ which is never going to happen since $\boldsymbol{q}' \in \mathbb{Q}^n_{\#d(\boldsymbol{q}),\lambda}$ which has the restriction that $\boldsymbol{q}'$ has exactly λ unique queries. In other words, every volume is mapped to a unique query. More formally, we have $q'_i = q'_j$ iff $\#d(q_i) = \#d(q_j)$ for all $i, j \in [n]$. Now it is easy to see that the size of $S_{\#d(\boldsymbol{q})}$ is

$$S_{\#d(\boldsymbol{q})} = m \cdot (m-1) \cdots (m - \lambda + 1) = \frac{m!}{(m-\lambda)!}.$$

Given the above result and setting $\alpha = (m^n \cdot \binom{N-1}{m-1})^{-1}$, we obtain

$$\begin{aligned}
\Pr[\,\mathbf{CHR} = 1\,] &= \alpha \cdot \sum_{d \in \mathbb{D}_N} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \Big) \\
&= \alpha \cdot \sum_{j=1}^{m'} \Big(\sum_{d \in \mathbb{D}_j} \Big(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \Big)\Big) && (35) \\
&= \alpha \cdot \sum_{j=1}^{m'} \Big(\sum_{d \in \mathbb{D}_j} \Big(\sum_{i=1}^{j} \sum_{\boldsymbol{q} \in \mathbb{Q}_{1,i}^n} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \Big)\Big)\Big) && (36)
\end{aligned}$$

Equation 35 is a decomposition of the space $\mathbb{D}_N$ into m' subsets $\mathbb{D}_j$ such that each subset $\mathbb{D}_j$ is the set of functions d with j unique volumes, i.e., $j = \#\mathsf{set}(\{\#d(q_i)\}_{i\in[m]})$. It is easy to see that $\mathbb{D}_N = \bigcup_{j=1}^{m'} \mathbb{D}_j$ where m' is the maximum on the number of possible unique volumes a multi-map can have. We will show below that $m' = \min(m, \sqrt{2N})$. Equation 36 represents a decomposition of the query space into subsets where each subset $\mathbb{Q}_{1,i}^n$ corresponds to the query sequences $\boldsymbol{q} \in \mathbb{Q}_1^n$ with i unique queries, i.e.,

$$\mathbb{Q}_{1,i}^n = \Big\{ \boldsymbol{q} \in \ \mathbb{Q}_1^n \mid \#\mathsf{set}(\{q_j\}_{j\in[n]}) = i \Big\}.$$

Claim 5. *We show that there does not exist $d \in \mathbb{D}$ such that*

$$\#\mathsf{set}\Big(\Big\{ \#d(q_i) \Big\}_{i\in[m]} \Big) > \min\Big(m, \sqrt{2N} \Big)$$

Proof. Given that the sum of all the m volumes should not exceed N, we want to maximize the non-zero values of the following equation:

$$x_1 + 2x_3 + \cdots + (N - m + 1)x_{N-m+1} = N,$$

where $\{1, \cdots, N - m + 1\}$ are the possible volumes. The maximum is reached when we select the smallest volumes that can lead to N, since otherwise we will reach N faster and therefore use less unique volumes. In particular finding the maximum number of unique volumes θ is equivalent to solving the equation:

$$\sum_{i=1}^{\theta} i = N \equiv \theta^2 + \theta - 2N = 0,$$

where the solution is

$$\theta = \frac{\sqrt{8N+1} - 1}{2} \leq \sqrt{2N}$$

which follows from the fact that $\sqrt{8N+1} \leq \sqrt{8N} + \sqrt{1} = 2\sqrt{2N} + 1$. However we know that the number of unique volumes cannot exceed m, the size of the query space, so the number of unique volumes is at most $\min(m, \sqrt{2N})$.

■

Given the above results, we now have

$$
\begin{aligned}
\Pr[\,\mathbf{CHR}=1\,] &= \alpha\cdot\sum_{j=1}^{m'}\Big(\sum_{d\in\mathbb{D}_j}\Big(\sum_{i=1}^{j}\Big(\sum_{k=1}^{\binom{m}{i}}\sum_{\boldsymbol{q}\in\mathbb{Q}^n_{1,i,k}}\frac{1}{\#S_{\#d(\boldsymbol{q})}}\Big)\Big)\Big) && (37)\\
&= \alpha\cdot\sum_{j=1}^{m'}\Big(\sum_{d\in\mathbb{D}_j}\Big(\sum_{i=1}^{j}\Big(\sum_{k=1}^{\binom{m}{i}}\sum_{\boldsymbol{q}\in\mathbb{Q}^n_{1,i,k}}\frac{(m-i)!}{m!}\Big)\Big)\Big) && (38)\\
&\le \alpha\cdot\sum_{j=1}^{m'}\Big(\sum_{d\in\mathbb{D}_j}\Big(\sum_{i=1}^{j}\Big(\sum_{k=1}^{\binom{m}{i}} i!\cdot\left\{{n \atop i}\right\}\frac{(m-i)!}{m!}\Big)\Big)\Big) && (39)\\
&= \alpha\cdot\sum_{j=1}^{m'}\Big(\sum_{d\in\mathbb{D}_j}\Big(\sum_{i=1}^{j}\binom{m}{i} i!\cdot\left\{{n \atop i}\right\}\frac{(m-i)!}{m!}\Big)\Big)\\
&= \alpha\cdot\sum_{j=1}^{m'}\Big(\sum_{d\in\mathbb{D}_j}\Big(\sum_{i=1}^{j}\left\{{n \atop i}\right\}\Big)\Big)\\
&\le \alpha\cdot\Big(\sum_{i=1}^{m'}\left\{{n \atop i}\right\}\Big)\cdot\sum_{j=1}^{m'}\Big(\sum_{d\in\mathbb{D}_j}1\Big) && (40)\\
&\le \alpha\cdot\Big(\sum_{i=1}^{m'}\left\{{n \atop i}\right\}\Big)\cdot\#\mathbb{D}_N = \frac{1}{m^n}\cdot\sum_{i=1}^{m'}\left\{{n \atop i}\right\}
\end{aligned}
$$

Equation 37 is a further decomposition of the query sequence space into subsets $\mathbb{Q}^n_{1,i,k}$ where we fix a particular combination of the i unique queries. Note that there are $\binom{m}{i}$ ways to pick i unique queries for a query space of size m. Equation 38 simply replaces $\#S_{\#d(\boldsymbol{q})}$ with its value which we previously computed. In Equation 39, the size of $\mathbb{Q}^n_{1,i,k}$ is $i!\cdot\left\{{n \atop i}\right\}$ (we refer the reader to Theorem 5 for the detailed argument). Note also that we have accounted for all possible query sequences in $\mathbb{Q}^n_{1,i,k}$ which is an upper-bound. To see why, note that there might be query sequences $\boldsymbol{q'}\in\mathbb{Q}^n_{1,i,k}$ with a fixed i queries such that i is larger than the number of unique volumes in $S_{\#d(\boldsymbol{q})}$, for a particular multi-map d.[11] Equation 40 is a simple upper bound where we set $j=m'$ so that the sum $\sum_{i=1}^{m'}\left\{{n \atop i}\right\}$ is the largest and is a constant that no longer depends on which function we pick. Finally, leveraging the result of Lemma E.1 concludes our proof.

■

Theorem 14. *For all $n\in\mathbb{N}$, if $\mathbf{Q}\sim\mathcal{Z}^n_{m,s}$, $\mathbf{A}\sim\mathcal{Z}^n_{ms,}$, $D\sim\mathcal{U}_{\mathbb{D}_N}$ and $A_D\sim\mathcal{U}_{\mathbb{D}_N}$ then $\mathcal{N}^+_{\mathbf{QeVo}}$ is $(\varepsilon,\mathcal{A}_{\mathsf{map}},\mathbf{A},\varphi)$-coherent with*

$$
\varepsilon\le\max\left\{\frac{1}{m^n}-\frac{1}{H^n_{m,s}}\cdot\sum_{i=1}^{m}i^{-n\cdot s}\cdot\left\{{n \atop i}\right\},\ \frac{1}{H^n_{m,s}}\cdot\sum_{i=1}^{m}(i!)^{1-s}\cdot\left\{{n \atop i}\right\}-\frac{1}{m^n}\right\}
$$

[11] This is a crucial step in the proof which makes our bound quite loose. In order to tighten the bound, one would need to find which combination of i unique queries are *valid* in the sense that i is equal to the number of unique volumes in $\#d(\boldsymbol{q})$ for a specific function d. We leave this question as an interesting open problem.

Proof. From Theorem 12, we know

$$\begin{aligned}
\Pr\left[\,\mathbf{CHR}=1\,\right] &= \sum_{d\in\mathbb{D}_N}\left(\sum_{\boldsymbol{q}\in\mathbb{Q}_1^n}\frac{1}{\#S_{\#d(q)}}\cdot\Pr\left[\,\mathbf{Q}=\boldsymbol{q}\,\right]\cdot\Pr\left[\,D=d\,\right]\right)\\
&= \frac{1}{m!}\cdot\sum_{d\in\mathbb{D}_N}\left(\sum_{\boldsymbol{q}\in\mathbb{Q}_1^n}\frac{1}{\#S_{\#d(\boldsymbol{q})}}\cdot\Pr\left[\,\mathbf{Q}=\boldsymbol{q}\,\right]\right) \qquad (41)
\end{aligned}$$

Equation 41 holds as the number of multi-maps with volumes in $\mathbb{S}$ is equal to $m!$. Given that we fix $\mathbb{S}$, the only aspect that can vary is how we assign the m keywords to the m unique volumes which can be done in $m!$ ways. In the following, we focus on calculating $S_{\#d(\boldsymbol{q})}$.

$$\begin{aligned}
S_{\#d(\boldsymbol{q})} &= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n}\left\{\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,|\,\mathbf{L}=\#d(\boldsymbol{q})\,\right]\right\}\\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n}\left\{\Pr\left[\,\mathbf{L}=\#d(\boldsymbol{q})\,|\,\mathbf{A}=\boldsymbol{q}'\,\right]\cdot\frac{\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,\right]}{\Pr\left[\,\mathbf{L}=\#d(\boldsymbol{q})\,\right]}\right\}\\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n}\left\{\Pr\left[\,\mathbf{L}=\#d(\boldsymbol{q})\,|\,\mathbf{A}=\boldsymbol{q}'\,\right]\cdot\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,\right]\right\}\\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n}\left\{\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,\right]\cdot\sum_{d'\in\mathbb{D}_N}\Pr\left[\,\mathbf{L}=\#d(\boldsymbol{q})\,|\,\mathbf{A}=\boldsymbol{q}',D'=d'\,\right]\cdot\Pr\left[\,D'=d'\,\right]\right\}\\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n}\left\{\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,\right]\cdot\sum_{d'\in\mathbb{D}_N}\Pr\left[\,\mathbf{L}=\#d(\boldsymbol{q})\,|\,\mathbf{A}=\boldsymbol{q}',D'=d'\,\right]\right\}\\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n_{d(\boldsymbol{q})}}\left\{\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,\right]\cdot\sum_{d'\in\mathbb{D}_N}\Pr\left[\,\mathbf{L}=\#d(\boldsymbol{q})\,|\,\mathbf{A}=\boldsymbol{q}',D'=d'\,\right]\right\} \qquad (42)\\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n_{d(\boldsymbol{q})}}\left\{\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,\right]\cdot(m-\lambda_{d(\boldsymbol{q})})\right\} = \arg\max_{\boldsymbol{q}'\in\mathbb{Q}^n_{d(\boldsymbol{q})}}\left\{\Pr\left[\,\mathbf{A}=\boldsymbol{q}'\,\right]\right\} \qquad (43)
\end{aligned}$$

Equation 42 follows from the fact that the only query sequences for which the conditional probability is not null are the one that belong to $\mathbb{Q}^n_{d(\boldsymbol{q})}$ where

$$\mathbb{Q}^n_{d(\boldsymbol{q})} = \left\{\boldsymbol{q}\in\mathbb{Q}^n\ |q_i=q_j \text{ if } \#d(q_i)=\#d(q_j),\ \ \forall\ i,j\in[n]\right\}.$$

And given that $\mathbf{A}$ is Zipf-distributed, then as shown in Theorem 6, the query sequences that maximize $S_{\#d(\boldsymbol{q})}$ are the ones composed of the $\lambda_{d(\boldsymbol{q})}$ queries with the highest ranks, where $\lambda_{d(\boldsymbol{q})}$ represents the number of unique volumes in $d(\boldsymbol{q})$. In addition, these queries also verify the fact that the volume appearing the most in a query sequence will be assigned to the query with the highest rank, the second appearing most with the second highest rank and so on and so forth. And similar to Theorem 7, we can show that

$$1\leq S_{\#d(\boldsymbol{q})}\leq\lambda_{d(\boldsymbol{q})}!$$

We are also going to leverage a previous result shown in Theorem 7 where given that $\mathbf{Q}\sim\mathcal{Z}^n_{m,s}$, then for all $\boldsymbol{q}\in\mathbb{Q}^n$,

$$\Pr\left[\,\mathbf{Q}=\boldsymbol{q}\,\right]\geq\frac{1}{H^n_{m,s}}\cdot i^{-n\cdot s}$$

Plugging the above results in Equation 41, we obtain

$$\Pr[\,\mathbf{CHR}=1\,] = \frac{1}{m!}\cdot\sum_{d\in\mathbb{D}_N}\left(\sum_{i=1}^{m}\left(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_{1,i}}\frac{1}{\#S_{\#d(\boldsymbol{q})}}\cdot\Pr[\,\mathbf{Q}=\boldsymbol{q}\,]\right)\right) \tag{44}$$

$$\geq \frac{1}{m!}\cdot\sum_{d\in\mathbb{D}_N}\left(\sum_{i=1}^{m}\left(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_{1,i}}\frac{1}{i!}\cdot\frac{1}{H^n_{m,s}}\cdot i^{-n\cdot s}\right)\right)$$

$$= \frac{1}{m!}\cdot\sum_{d\in\mathbb{D}_N}\left(\sum_{i=1}^{m}\frac{1}{i!}\cdot\frac{1}{H^n_{m,s}}\cdot i^{-n\cdot s}\sum_{j=1}^{\binom{m}{i}}\#\mathbb{Q}^n_{1,i,j}\right)$$

$$= \frac{1}{m!}\cdot\sum_{d\in\mathbb{D}_N}\left(\sum_{i=1}^{m}\frac{1}{i!}\cdot\frac{1}{H^n_{m,s}}\cdot i^{-n\cdot s}\cdot\#\mathbb{Q}^n_{1,i,j^\star}\right) \tag{45}$$

$$= \frac{1}{m!}\cdot\sum_{d\in\mathbb{D}_N}\left(\sum_{i=1}^{m}\frac{1}{H^n_{m,s}}\cdot i^{-n\cdot s}\cdot\left\{{n \atop i}\right\}\right) \tag{46}$$

$$= \frac{1}{H^n_{m,s}}\cdot\sum_{i=1}^{m} i^{-n\cdot s}\cdot\left\{{n \atop i}\right\}$$

Equation 44 results from the fact that every multi-map $d\in\mathbb{D}_N$ has m distinct volumes. Equation 45 follows from the argument above that there is only a unique set of i unique queries that maximizes $\#S_{d(\boldsymbol{q})}$. Finally, Equation 46 follows from the fact that the size of $\mathbb{Q}^n_{1,i,j^\star}$ is equal to $i!\cdot\left\{{n \atop i}\right\}$, refer to Theorem 5 for more details.

Similarly, we can compute a upper bound as follows. First, as shown in Theorem 11, we have

$$\Pr[\,\mathbf{Q}=\boldsymbol{q}\,] \leq \frac{1}{H^n_{m,s}}\cdot(i!)^{-s}$$

and following the same steps as above and with the observation that for all $\boldsymbol{q}\in\mathbb{Q}^n_{1,i}$, $\#S_{\#d(\boldsymbol{q})}\geq 1$, we can show that

$$\Pr[\,\mathbf{CHR}=1\,] \leq \frac{1}{H^n_{m,s}}\cdot\sum_{i=1}^{m}(i!)^{1-s}\cdot\left\{{n \atop i}\right\}.$$

Finally, using the result of Lemma E.1 concludes our proof.

■

Theorem 15. *For all $n\in\mathbb{N}$, if $\mathbf{Q}\sim\mathcal{Z}^n_{m_q,s_q}$, $\mathbf{A}\sim\mathcal{Z}^n_{m_a,s_a}$, $D\sim\mathcal{U}_{\mathbb{D}_N}$, $A_D\sim\mathcal{U}_{\bar{\mathbb{D}}_N}$ and $\mathbb{A}\subset\mathbb{Q}$ then $\mathcal{N}^+_{\mathsf{QeVo}}$ is $(\varepsilon,\mathcal{A}_{\mathsf{map}},\mathbf{A},\varphi)$-coherent with*

$$\varepsilon \leq \max\left\{\frac{1}{m^n}-\frac{1}{H^n_{m_q,s_q}}\cdot\sum_{i=1}^{m_a}\frac{(m_a)_i\cdot\left\{{n \atop i}\right\}}{(m_q)_i\cdot(\gamma+i)^{s_q\cdot n}},\ \frac{1}{H^n_{m_q,s_q}\cdot\gamma^{s_q\cdot n}}\cdot\sum_{i=1}^{m_a}\frac{(m_a)_i}{(m_q)_i}\cdot i!\cdot\left\{{n \atop i}\right\}-\frac{1}{m^n}\right\}$$

where γ the rank as defined in Section 6.5.

Proof. From Theorem 12, we know that

$$\Pr[\,\mathbf{CHR}=1\,] = \sum_{d\in\mathbb{D}_N}\left(\sum_{\boldsymbol{q}\in\mathbb{Q}^n_1}\frac{1}{\#S_{\#d(q)}}\cdot\Pr[\,\mathbf{Q}=\boldsymbol{q}\,]\cdot\Pr[\,D=d\,]\right)$$

$$= \frac{1}{m_q!} \cdot \sum_{d\in\mathbb{D}_N} \Bigg(\sum_{\boldsymbol{q}\in\mathbb{Q}_1^n} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \cdot \Pr\left[\,\mathbf{Q} = \boldsymbol{q}\,\right]\Bigg) \tag{47}$$

$$= \frac{1}{m_q!} \cdot \sum_{d\in\mathbb{D}_N} \Bigg(\sum_{i=1}^{m_a} \Bigg(\sum_{j=1}^{\binom{m_q}{i}} \Bigg(\sum_{\boldsymbol{q}\in\mathbb{A}_{1,i,j}^n} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \cdot \Pr\left[\,\mathbf{Q} = \boldsymbol{q}\,\right]\Bigg)\Bigg)\Bigg) \tag{48}$$

where $\mathbb{A}_{1,i,j}$ introduced in Equation 48 represents the set of all query sequences that belong to $S_{\#d(\boldsymbol{q})}$ and that are composed of a fixed set of i unique queries. On the other hand, and as shown in Theorem 14, we know that

$$\begin{aligned}
S_{\#d(\boldsymbol{q})} &= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}_{d(\boldsymbol{q})}^n} \Big\{ \Pr\left[\,\mathbf{A} = \boldsymbol{q}'\,\right] \cdot \sum_{d'\in\mathbb{D}_N} \Pr\left[\,\mathbf{L} = \#d(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q}', D' = d'\,\right] \Big\} \\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}_{d(\boldsymbol{q})}^n} \Big\{ \Pr\left[\,\mathbf{A} = \boldsymbol{q}'\,\right] \cdot \sum_{d'\in\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'}} \Pr\left[\,\mathbf{L} = \#d(\boldsymbol{q}) \mid \mathbf{A} = \boldsymbol{q}', D' = d'\,\right] \Big\} && (49) \\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}_{d(\boldsymbol{q})}^n} \Big\{ \Pr\left[\,\mathbf{A} = \boldsymbol{q}'\,\right] \cdot \#\mathbb{D}_{d(\boldsymbol{q}),\boldsymbol{q}'} \Big\} \\
&= \arg\max_{\boldsymbol{q}'\in\mathbb{Q}_{d(\boldsymbol{q})}^n} \Big\{ \Pr\left[\,\mathbf{A} = \boldsymbol{q}'\,\right] \cdot (m_a - \lambda_{d(\boldsymbol{q})})! \Big\} = \arg\max_{\boldsymbol{q}'\in\mathbb{Q}_{d(\boldsymbol{q})}^n} \Big\{ \Pr\left[\,\mathbf{A} = \boldsymbol{q}'\,\right] \Big\} && (50)
\end{aligned}$$

Equation 49 simply specifies the set of possible databases $\mathbb{D}_{N,d(\boldsymbol{q})}$ for which the set $S_{\#d(\boldsymbol{q})}$ is not empty such that for all $\boldsymbol{q}' \in \mathbb{Q}_{d(\boldsymbol{q})}^n$,

$$\mathbb{D}_{d(\boldsymbol{q}),\boldsymbol{q}'} = \Big\{ d' \in \bar{\mathbb{D}}_N \mid \#d(q_i) = \#d'(q_i') \ \forall i \in [n] \Big\}.$$

Also recall from Theorem 6 that since $\mathbf{A}$ are Zipf-distributed, the query sequences that maximizes $\#S_{\#d(\boldsymbol{q})}$ are the ones composed of the $\lambda_{\#d(\boldsymbol{q})}$ queries with the highest ranks based on the permutation π_a. Leveraging this observation, we can rewrite Equation 48 such that

$$\begin{aligned}
\Pr\left[\,\mathbf{CHR} = 1\,\right] &= \frac{1}{m_q!} \cdot \sum_{d\in\mathbb{D}_N} \Bigg(\sum_{i=1}^{m_a} \Bigg(\sum_{\boldsymbol{q}\in\mathbb{A}_{1,i,j^\star}^n} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \cdot \Pr\left[\,\mathbf{Q} = \boldsymbol{q}\,\right]\Bigg)\Bigg) \\
&= \frac{1}{m_q!} \cdot \sum_{i=1}^{m_a} \Bigg(\sum_{d\in\mathbb{D}_N^i} \Bigg(\sum_{\boldsymbol{q}\in\mathbb{A}_{1,i,j^\star}^n} \frac{1}{\#S_{\#d(\boldsymbol{q})}} \cdot \Pr\left[\,\mathbf{Q} = \boldsymbol{q}\,\right]\Bigg)\Bigg) && (51)
\end{aligned}$$

Equation 51 swaps two summations and replaces $\mathbb{D}_N$ by $\mathbb{D}_N^i$ as the latter is the only subset of possible multi-maps for which the set $S_{\#d(\boldsymbol{q})}$ is non-empty, for all $\boldsymbol{q} \in \mathbb{A}_{1,i,j}^n$. More formally we define $\mathbb{D}_N^i$ such that

$$\mathbb{D}_N^i = \Big\{ d \in \mathbb{D}_N \mid d(q) \in \mathbb{S}_{\mathbb{A}} \ \forall q \in \{\pi_a^{-1}(1), \cdots, \pi_a^{-1}(i)\} \Big\}.$$

In the following, we are interested in computing the size of the set $\mathbb{D}_N^i$, for all $i \in [m_a]$.

Claim 6. *For all $i \in [m_a]$, and $N \in \mathbb{N}$, we have*

$$\#\mathbb{D}_N^i = \binom{m_a}{i} \cdot i! \cdot (m_q - i)!$$

Proof. By definition $\mathbb{D}_N^i$ is composed of all multi-maps such that i a-priori fixed queries have volumes in $\mathbb{S}_{\mathbb{A}}$. The number of ways we can assign i volumes from m_a volumes is equal to $i! \cdot \binom{m_a}{i}$ since the order of the assignment matters. Assigning the remaining queries to the remaining volumes in $\mathbb{S}$ can be done in $(m_q - 1)!$.

■

In the following, we will compute a lower and an upper bound for the $\{\mathbf{CHR} = 1\}$ event. But before, recall that since $\mathbb{A} \subset \mathbb{Q}$ and $\mathbf{Q} \sim \mathcal{Z}^n_{m_q,s_q}$, we can write for all $\boldsymbol{q} \in \mathbb{A}_{1,i,j}$,

$$\frac{1}{H^n_{m_q,s_q} \cdot (\gamma + i)^{s_q \cdot n}} \leq \Pr[\, \mathbf{Q} = \boldsymbol{q} \,] \leq \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}}$$

where γ is the rank value defined in Section 6. And that for all $\boldsymbol{q} \in \mathbb{A}^n_{1,i,j^\star}$, we have

$$1 \leq \# S_{\# d(\boldsymbol{q})} \leq i!$$

First, the upper bound can be calculated as follows,

$$\begin{aligned}
\Pr[\, \mathbf{CHR} = 1 \,] &\leq \frac{1}{m_q!} \cdot \sum_{i=1}^{m_a} \Bigg(\sum_{d \in \mathbb{D}_N^i} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{A}^n_{1,i,j^\star}} \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \Bigg) \Bigg) && (52) \\
&= \frac{1}{m_q!} \cdot \sum_{i=1}^{m_a} \left(\frac{\# \mathbb{D}_N^i \cdot \# \mathbb{A}^n_{1,i,j^\star}}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \right) && (53) \\
&= \frac{1}{m_q!} \cdot \sum_{i=1}^{m_a} \left(\frac{\binom{m_a}{i} \cdot i! \cdot (m_q - i)! \cdot i! \cdot \left\{ {n \atop i} \right\}}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \right) && (54) \\
&= \frac{1}{H^n_{m_q,s_q} \cdot \gamma^{s_q \cdot n}} \cdot \sum_{i=1}^{m_a} \frac{(m_a)_i}{(m_q)_i} \cdot i! \cdot \left\{ {n \atop i} \right\} && (55)
\end{aligned}$$

Similarly, we can show that the lower bound is equal to

$$\Pr[\, \mathbf{CHR} = 1 \,] \geq \frac{1}{H^n_{m_q,s_q}} \cdot \sum_{i=1}^{m_a} \frac{(m_a)_i}{(m_q)_i \cdot (\gamma + i)^{s_q \cdot n}} \cdot \left\{ {n \atop i} \right\}$$

Finally, using Lemma E.1 concludes the proof.

■

D Proofs for Section 8 (Full Recovery Against Query Equality and Volume)

Theorem 16. *The i.i.d. query-volume network $\mathcal{N}^+_{\mathbf{QeVo}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon = \left| \frac{1}{m!} \cdot \sum_{f \in \mathbb{F}} \sum_{d \in \mathbb{D}_N} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\# S_{\boldsymbol{\ell}}} \cdot \Pr[\, \mathbf{Q} = \boldsymbol{q} \,] \cdot \Pr[\, D = d \,] \Bigg) - \frac{1}{m^n} \right|,$$

where $S_{\boldsymbol{\ell}} \overset{\circ}{=} \mathsf{map}_{\mathbf{A_Q}|\boldsymbol{\ell}}$ and $\mathbb{Q}_1^n := \{\boldsymbol{q} \in \mathbb{Q}^n | \boldsymbol{q} \in S_{\boldsymbol{\ell}}\}$ and $\ell_i = (f(q_i), \# d(q_i))$ for all $i \in [n]$.

Theorem 17. *For all $n \in \mathbb{N}$, if $\mathbf{Q} \sim \mathcal{U}_m^n$, $\mathbf{A} \sim \mathcal{U}_m^n$, $D \sim \mathcal{U}_{\mathbb{D}_N}$ and $A_D \sim \mathcal{U}_{\mathbb{D}_N}$ then $\mathcal{N}^{+}_{\mathbf{QeVo}}$ is $(\varepsilon, \mathcal{A}_{\mathsf{map}}, \mathbf{A}, \varphi)$-coherent with*

$$\varepsilon = \left| \frac{1}{m^n} \cdot \sum_{i=1}^{m} \genfrac\{\}{0pt}{}{n}{i} - \frac{1}{m^n} \right|,$$

where $m = \#\mathbb{Q}$.

Proof. We use the same notation as the proofs of both Theorem 5 and Theorem 13. Given our assumptions, we can rewrite the general form of the coherence from Theorem 16 as

$$\Pr\left[\,\mathbf{CHR} = 1\,\right] = \frac{1}{m! \cdot m^n \cdot \binom{N-1}{m-1}} \sum_{f \in \mathbb{F}} \Bigg(\sum_{d \in \mathbb{D}_N} \Bigg(\sum_{\boldsymbol{q} \in \mathbb{Q}_1^n} \frac{1}{\# S_{\boldsymbol{\ell}}} \Bigg) \Bigg)$$

where $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_n)$ and $\ell_i = (f(q_i), \#d(q_i))$ for all $i \in [n]$. Now we further study the set $S_{\boldsymbol{\ell}}$ to better understand its structure. We have

$$\begin{aligned}
S_{\boldsymbol{\ell}} &= \mathsf{map}_{\mathbf{A_Q}|\boldsymbol{\ell}} \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \Pr\left[\,\mathbf{A_Q} = \boldsymbol{q}' \,|\, (\mathbf{L}_1, \mathbf{L}_2) = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)\,\right] \Big\} && (56) \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \left\{ \Pr\left[\,\mathbf{L}_1 = \boldsymbol{\ell}_1 \,|\, \mathbf{A_Q} = \boldsymbol{q}', \mathbf{L}_2 = \boldsymbol{\ell}_2\,\right] \cdot \frac{\Pr\left[\,\mathbf{A_Q} = \boldsymbol{q}' \,|\, \mathbf{L}_2 = \boldsymbol{\ell}_2\,\right]}{\Pr\left[\,\mathbf{L}_1 = \boldsymbol{\ell}_1 \,|\, \mathbf{L}_2 = \boldsymbol{\ell}_2\,\right]} \right\} && (57) \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \left\{ \Pr\left[\,\mathbf{L}_1 = \boldsymbol{\ell}_1 \,|\, \mathbf{A_Q} = \boldsymbol{q}', \mathbf{L}_2 = \boldsymbol{\ell}_2\,\right] \cdot \frac{\Pr\left[\,\mathbf{L}_2 = \boldsymbol{\ell}_2 \,|\, \mathbf{A_Q} = \boldsymbol{q}'\,\right] \cdot \Pr\left[\,\mathbf{A_Q} = \boldsymbol{q}'\,\right]}{\Pr\left[\,\mathbf{L}_2 = \boldsymbol{\ell}_2\,\right]} \right\} && (58) \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Big\{ \Pr\left[\,\mathbf{L}_1 = f(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', \mathbf{L}_2 = \#d(\boldsymbol{q})\,\right] \cdot \Pr\left[\,\mathbf{L}_2 = \#d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}'\,\right] \Big\} \\
&= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n} \Bigg\{ \Bigg(\sum_{f \in \mathbb{F}} \Pr\left[\,\mathbf{L}_1 = f(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', \mathbf{L}_2 = \#d(\boldsymbol{q}), F = f\,\right] \Bigg) \\
&\cdot \Bigg(\sum_{d' \in \mathbb{D}} \Pr\left[\,\mathbf{L}_2 = \#d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', A_D = d'\,\right] \Bigg) \Bigg\}
\end{aligned}$$

In Equation 57 we decompose $\mathbf{L}$ into two random variables $(\mathbf{L}_1, \mathbf{L}_2)$. Equation 58 holds using Bayes' rule for three events while Equation 58 follows from the standard Bayes' rule. First, note that

$$\Pr\left[\,\mathbf{L}_1 = f(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', \mathbf{L}_2 = \#d(\boldsymbol{q}), F = f\,\right] = 1,$$

if $\mathbf{q}' \in \mathbb{Q}^n_{f(\boldsymbol{q})}$, where

$$\mathbb{Q}^n_{f(\boldsymbol{q})} \overset{\circ}{=} \Big\{ \boldsymbol{q}' \in \mathbb{Q}^n \mid q'_j = q'_k \text{ iff } f(q_j) = f(q_k)\ \forall j, k \in [n] \Big\}.$$

Note that the second condition $\{\mathbf{L}_2 = \#d(\boldsymbol{q})\}$ does not change which query sequence is possible. And from Theorem 5, we know that the number of possible functions f is equal to $(m - \lambda)!$ where λ is the number of unique queries in $f(\boldsymbol{q})$. On the other hand, we know from Theorem 13 that $\Pr\left[\,\mathbf{L}_2 = \#d(\boldsymbol{q}) \,|\, \mathbf{A_Q} = \boldsymbol{q}', D' = d'\,\right]$ is equal to 1 if $\boldsymbol{q} \in \mathbb{Q}^n_{\#d(\boldsymbol{q})}$ where

$$\mathbb{Q}^n_{\#d(\boldsymbol{q})} = \Big\{ \boldsymbol{q}' \in \mathbb{Q}^n \mid q'_i \neq q'_j \text{ if } \#d(q_i) \neq \#d(q_j)\ \forall i, j \in [n] \Big\}$$

and the number of possible functions d is equal to $\#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'}$ where

$$\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} = \Big\{ d' \in \mathbb{D}_N \mid \#d'(q'_i) = \#d(q_i) \ \forall i \in [n] \Big\}$$

The most important observation here is that

$$\mathbb{Q}^n_{f(\boldsymbol{q})} \subseteq \mathbb{Q}^n_{\#d(\boldsymbol{q})}$$

This holds since several combinations of queries remain possible in the positions in $\#d(\boldsymbol{q})$ where the volume is the same which, in turn, leads to more query sequences. This is however not true for $\mathbb{Q}^n_{f(\boldsymbol{q})}$ since we know the exact position of the unique queries. Based on this observation, we rewrite $S_{\boldsymbol{\ell}}$ as

$$\begin{aligned} S_{\boldsymbol{\ell}} &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n_{f(\boldsymbol{q})}} \Big\{ (m-\lambda)! \cdot \#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} \Big\} \\ &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n_{f(\boldsymbol{q})}} \Big\{ \#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'} \Big\} \\ &= \arg\max_{\boldsymbol{q}' \in \mathbb{Q}^n_{f(\boldsymbol{q})}} \left\{ \binom{N - \sum_{i\in P} \#d(q) - 1}{m-\lambda-1} \right\} \qquad (59) \\ &= \mathbb{Q}^n_{f(\boldsymbol{q})} \end{aligned}$$

where P is a set of λ indices such that for all $i, j \in P$, $f(q_i) \neq f(q_j)$. Equation 59 follows from Theorem 13 where we computed the size of $\#\mathbb{D}_{\#d(\boldsymbol{q}),\boldsymbol{q}'}$. Note that the size of this set is constant and does not vary since: (1) we know that there are λ unique queries; and (2) given $\boldsymbol{\ell}$, we know *exactly* all the volumes of all the λ unique queries. Also recall that $\#\mathbb{Q}^n_{f(\boldsymbol{q})} = m!/(m-\lambda)!$, so using the same query space decomposition techniques from the proof of Theorem 5 we can show that

$$\Pr[\,\mathbf{CHR} = 1\,] = \frac{1}{m^n} \cdot \sum_{i=1}^{m} \left\{ {n \atop i} \right\}. \qquad (60)$$

Finally, subtracting $1/m^n$ and taking the absolute value ends our proof.

■

E A Useful Lemma

When analyzing coherence, it is often the case that one cannot compute the exact probability that an adversary wins the coherence experiment but instead can only obtain a lower and/or an upper bound. In this case, the following Lemma can be useful to derive an upper bound on the coherence.

Lemma E.1. *If a, b and ε are reals in $[0,1]$ such that $|a-b| = \varepsilon$, and if there exists $u, l \in \mathbb{R}^+$ such that $l \leq a \leq u$, then*

$$\varepsilon \leq \max\Big\{ b - l, u - b \Big\}$$

Proof. First, we have by assumption that $l - b \leq a - b \leq u - b$. We need to consider two cases. If $a - b \geq 0$, then $a - b = \varepsilon$ and therefore

$$\varepsilon \leq u - b.$$

Otherwise if $a - b < 0$, then $a - b = -\varepsilon$ and therefore

$$\varepsilon \leq b - l.$$

This concludes the proof.

■