

IHOP: Improved Statistical Query Recovery against Searchable Symmetric Encryption through Quadratic Optimization

Simon Oya
University of Waterloo

Florian Kerschbaum
University of Waterloo

Abstract

Effective query recovery attacks against Searchable Symmetric Encryption (SSE) schemes typically rely on auxiliary *ground-truth* information about the queries or dataset. Query recovery is also possible under the weaker *statistical* auxiliary information assumption, although statistical-based attacks achieve lower accuracy and are not considered a serious threat. In this work we present IHOP, a statistical-based query recovery attack that formulates query recovery as a quadratic optimization problem and reaches a solution by iterating over linear assignment problems. We perform an extensive evaluation with five real datasets, and show that IHOP outperforms all other statistical-based query recovery attacks under different parameter and leakage configurations, including the case where the client uses some access-pattern obfuscation defenses. In some cases, our attack achieves almost perfect query recovery accuracy. Finally, we use IHOP in a frequency-only leakage setting where the client’s queries are correlated, and show that our attack can exploit query dependencies even when PANCAKE, a recent frequency-hiding defense by Grubbs et al., is applied. Our findings indicate that statistical query recovery attacks pose a severe threat to privacy-preserving SSE schemes.

1 Introduction

Searchable Symmetric Encryption (SSE) schemes [37] allow a client to securely outsource a dataset to a cloud-storage provider, while still being able to perform secure queries over the dataset. Efficient SSE schemes [2, 4, 5, 8, 17, 19, 23, 24, 28, 30, 38] typically leak certain information during their initialization step or querying process, that an honest-but-curious service provider could exploit to recover the database or guess the client’s queries. This leakage typically consists of the access pattern, which refers to the identifiers of the

documents that match a query, and the search pattern, which refers to whether or not two queries are identical.

There are many attacks that exploit access and search pattern leakage, as well as auxiliary information, to either recover the underlying keywords of the client’s queries, or the dataset itself. In this work, we study *passive query recovery* attacks [1, 3, 9, 18, 26, 29, 31, 34] against SSE schemes that provide *keyword query* functionality. Namely, the dataset is a set of documents, each labeled with a list of keywords, and the client wants to be able to query for all the documents that match a particular keyword. The adversary is an honest-but-curious service provider that follows the SSE protocol, but might want to infer the keywords of the client’s queries from passively observing the system. We can broadly classify passive query recovery attacks depending on the nature of their auxiliary information into attacks that assume partial ground-truth knowledge about the dataset and/or the underlying keywords of the client’s queries (*ground-truth attacks*) [1, 3, 9, 18, 29], and those that *only* require statistical information about the keyword distribution in the dataset and the client’s querying behavior (*statistical-based attacks*) [18, 31, 34]. Previous work has shown that ground-truth attacks can achieve high query recovery rates [1, 9, 29], and thus pose a significant privacy threat to SSE schemes. Statistical-based attacks achieve lower query recovery accuracy, since they rely on a weaker auxiliary information assumption, although they are easier to adapt against access and search-pattern defense techniques [31].

In this work we focus on statistical-based query recovery attacks. The strongest attacks from this family are GraphM by Pouliot and Wright [34] and SAP by Oya and Kerschbaum [31]. GraphM performs query recovery by heuristically solving a quadratic optimization problem. This attack exploits the so-called volume co-occurrence information, which refers to how many documents two queries have in common, but uses the expensive PATH algorithm [42], and thus only works when the keyword universe size is small. On the other hand, SAP [31] efficiently solves a linear problem using individual keyword volume information as well as query fre-

To appear in: Proceedings of the 31st USENIX Security Symposium
August 10–12, 2022, Boston, MA, USA
<https://www.usenix.org/conference/usenixsecurity22>

quency information, but cannot take into account volume co-occurrence. In this paper, we propose a new statistical-based query recovery attack, that we call *IHOP* since it follows an Iterative Heuristic algorithm to solve a quadratic Optimization Problem. *IHOP* can be used with any quadratic objective function, and thus it can combine volume and frequency leakage information like *SAP* [31], exploit volume co-occurrence terms like *GraphM* [34], and is the first attack to exploit frequency co-occurrence terms. We use *IHOP* to optimize a maximum-likelihood-based objective function, and evaluate our attack in five real datasets under different settings. First, we consider the case where the adversary only uses volume leakage, where *GraphM* is state-of-the-art. We show that, both in SSE schemes that fully leak the access patterns during initialization [17, 24] and in those that (at least) leak the access pattern of those keywords that are queried [2, 4, 5, 8, 19, 23, 28, 30, 38], *IHOP* consistently outperforms *GraphM*, achieving higher accuracy while being orders of magnitude faster (e.g., *IHOP* achieves $\approx 99\%$ accuracy on Lucene dataset with $n = 1000$ keywords and finishes in 15 minutes, while *GraphM* achieves $\approx 56\%$ accuracy and takes almost three days to finish). Second, we consider the case where the adversary uses frequency leakage and the client sends queries independently, where *SAP* is state-of-the-art. We show that *IHOP* comfortably outperforms *SAP* and all other attacks, even in the case where the client uses two access-pattern-hiding defenses [6, 36].

Finally, we show that *IHOP* can also exploit query correlations in query recovery. For this, we consider a case where the client queries dataset entries individually and thus the attack must rely on query frequency information only. *PANCAKE*, a system proposed by Grubbs et al. [13], provably ensures that the frequency of access to each dataset entry is the same. Although *PANCAKE* does not hide query correlations, a preliminary analysis shows that it still provides sufficient protection in that case [14]. We adapt *IHOP* against *PANCAKE* when client’s queries follow a Markov process, and show that our attack can effectively exploit query correlations to recover keywords (in our example, *IHOP* achieves $\approx 48\%$ accuracy when the adversary has low-quality auxiliary information about the client’s querying behavior, and $\approx 69\%$ accuracy when the adversary has high-quality information).

In summary, we propose a statistical-based query recovery attack that outperforms all other known attacks when the adversary does not have ground-truth information on the client’s dataset or queries [18, 26, 31, 34]. Our results show that statistical-based query recovery attacks pose a serious threat to SSE schemes, since they can achieve high recovery rates without access to ground-truth information and can adapt against volume and frequency-hiding defenses.

2 Related work

Since their inception [8, 37], many authors have proposed SSE schemes with different privacy and utility trade-offs. Design-

ing SSE schemes without access pattern leakage implies leveraging expensive primitives such as ORAM [12] or PIR [7], which incurs expensive bandwidth, computational, and storage costs, while still being vulnerable to certain volume-based attacks [15, 20, 33]. In this work we target *efficient and deployable* SSE schemes that leak the access and search patterns [2, 4, 5, 8, 17, 19, 23, 24, 28, 30, 37, 38]. We consider SSE schemes where the client performs *keyword queries* (i.e., the client queries for a single keyword to retrieve all documents that match that keyword) and the adversary is an honest-but-curious service provider that performs a *query recovery* attack to guess the underlying keywords of the client’s queries.

Query recovery attacks can be broadly classified depending on the auxiliary data they require into ground-truth attacks [1, 3, 9, 18, 29], which partially know some of the client’s queries or database, and statistical-based attacks [18, 26, 31, 34], that have statistical information about the database and client’s querying behavior (e.g., a set of non-indexed documents).

Ground-truth attacks Islam et al. [18] propose one of the first query recovery attacks (*IKK*) that assumes full dataset knowledge and partial query knowledge. Their attack exploits access pattern leakage to compute volume co-occurrence matrices and solves a quadratic optimization problem using simulated annealing. Cash et al. [3] propose the count attack, which improves upon *IKK* by requiring only partial database knowledge. This attack follows an iterative algorithm that keeps reducing the set of candidate keywords for each unknown query until only one feasible candidate remains. Blackstone et al. [1] propose an attack based on subgraphs that follows different refinement heuristics to the count attack, outperforming it. Recently, Damie et al. [9] proposed a hybrid attack that uses ground-truth query knowledge, but only statistical dataset information, and Ning et al. [29] propose a ground-truth attack that achieves both keyword and document recovery.

Statistical-based attacks Even though *IKK* originally assumes ground-truth dataset and query knowledge [18], this attack can also be evaluated in the setting where only statistical information is available (see Sec. 5). The graph matching attack (*GraphM*) by Pouliot and Wright [34] uses volume information computed from access-pattern leakage, like *IKK*, but uses a more refined optimization function and looks for a solution using graph matching algorithms [39, 42]. The frequency-only attack (*Freq*) by Liu et al. [26] exploits search-pattern leakage and auxiliary information about the client’s querying behavior to perform query recovery. Recently, Oya and Kerschbaum [31] proposed *SAP*, an attack that exploits both volume and frequency information to efficiently solve a linear assignment problem using known optimal solvers [11].

Even though query recovery rates achieved by statistical-based attacks are typically below those achieved by ground-truth attacks, due to their weaker auxiliary knowledge assumptions, statistical-based attacks can be easily tuned [31] against

volume-padding defenses [6, 10, 32]. Ground-truth attacks, however, perform poorly against such defenses [36], since they are designed to use exact information about the data and the defenses typically randomize the leaked patterns.

3 Problem Statement

3.1 Overview

Our system model consists of two parties: a client and a server. The client owns a privacy-sensitive dataset that she wishes to store remotely on the server. The server offers storage services (e.g., it is a cloud storage provider), but is not trusted by the client. The client encrypts each document using symmetric encryption, and sends the encrypted documents to the server. Each document has a set of *keywords* attached to it, and the client wishes to be able to issue *keyword queries*, i.e., queries that retrieve all the documents that contain a particular a keyword.¹ In order to achieve this search functionality in the encrypted dataset, the client uses a Searchable Symmetric Encryption (SSE) scheme. This SSE scheme has an initialization step where the client generates an *encrypted search index*, which she sends to the server alongside the encrypted documents. Later, when the client wishes to query for a particular keyword, she generates a *query token* from that keyword and sends it to the server. The server evaluates the query token in the encrypted search index, which reveals which encrypted documents should be returned to the client in response.

The server is honest-but-curious, i.e., it follows the protocol specifications, but might be interested in learning sensitive information from the client from passive observation. We focus on an adversary that wants to guess the underlying keywords of the client’s issued query tokens, i.e., to perform a *query recovery attack*. This can in turn help the adversary guess the keywords attached to each encrypted document (database recovery attack).

Even though encrypting the documents and hiding which documents contain each keyword through the encrypted search index prevents the adversary from trivially matching query tokens to keywords, most efficient SSE schemes leak certain information that can be used to carry out a query recovery attack, with the help of certain auxiliary information. There are two typical sources of leakage in existing SSE schemes: access pattern leakage and search pattern leakage.

The access pattern of a query is the list of document identifiers that match the given query (i.e., the identifiers of the documents that the server returns to the client in response to the query). The search pattern refers to whether or not two

¹Note that some works consider datasets where each document is associated to a *single* keyword [10, 32] (e.g., the keyword can be the document’s publication date), and thus queries for different keywords always return disjoint sets of documents. The attacks we consider in this paper rely on volume co-occurrence leakage, which only occurs when at least some documents have more than one keyword.

N_d	Number of documents in the dataset.
n	Total number of keywords, $n \doteq \Delta_k $.
m	Total number of observed query tokens, $m \doteq \Delta_\tau $.
ρ	Number of queries issued by the client.
D	Dataset $D \doteq \{d_1, d_2, \dots, d_{N_d}\}$
Δ_k	Keyword universe $\Delta_k \doteq [k_1, k_2, \dots, k_n]$.
Δ_τ	Token universe $\Delta_\tau \doteq [\tau_1, \tau_2, \dots, \tau_m]$.
k_i	i th keyword, with $i \in [n]$.
τ_j	j th query token, with $j \in [m]$.
\mathbf{a}_j	Access pattern for token τ_j ($N_d \times 1$).
\mathbf{V}	Matrix of observed token volumes ($m \times m$).
$\tilde{\mathbf{V}}$	Matrix of auxiliary keyword volumes ($n \times n$).
\mathbf{f}	Vector of observed token frequencies ($m \times 1$).
$\tilde{\mathbf{f}}$	Vector of auxiliary keyword frequencies ($n \times 1$).
\mathbf{F}	Markov matrix of observed token freqs. ($m \times m$).
$\tilde{\mathbf{F}}$	Markov matrix of auxiliary keyword freqs. ($n \times n$).

Table 1: Summary of notation

queries are identical (i.e., whether they have the same underlying keyword). An adversary with *auxiliary information* about how often the client queries for particular keywords, or how many documents have certain keywords, can leverage this leakage to carry out a query recovery attack.

3.2 Formal Problem Description and Notation

We formalize the problem described above and introduce our notation, which we summarize in Table 1. We note that this general description is not tailored to a specific SSE scheme, but accommodates a broad family of SSE schemes that leak the access and search patterns [2, 4, 5, 8, 17, 19, 24, 28, 37, 38].

We use boldface lowercase characters for vectors (e.g., \mathbf{a}), and boldface uppercase characters for matrices (e.g., \mathbf{A}). The transposition of \mathbf{A} is \mathbf{A}^T , and $\mathbf{A}_{i,j}$ is the i, j th entry of \mathbf{A} . All products between matrices and vectors are dot products. We use $[n] \doteq \{1, 2, \dots, n\}$ for a positive integer n .

We denote the client’s dataset by $D \doteq \{d_1, d_2, \dots, d_{N_d}\}$, where N_d is the number of documents. We refer to the documents by their index, and assume that they are randomly shuffled during setup so that their index does not reveal any information about their content. Each document has a set of keywords attached to it, and keywords belong to the *keyword universe* $\Delta_k \doteq \{k_1, k_2, \dots, k_n\}$, of size n . The client encrypts each document $d \in D$, uses an SSE scheme to generate an encrypted search index, and sends the encrypted dataset and index to the server. In order to query for a particular keyword $k \in \Delta_k$, the client first generates a query token τ using k , and she sends it to the server. The server evaluates the query token τ on the encrypted search index, which reveals the access pattern, i.e., the indices of the documents that match the query. We represent the access pattern of a token τ_j as a column vector \mathbf{a}_j of length N_d whose ℓ th entry is 1 if d_ℓ matches the query, and 0 otherwise. We consider SSE schemes that reveal

the search pattern, i.e., they reveal whether or not two query tokens were generated with the same underlying keyword. Thus, in our notation we use τ_j and $\tau_{j'}$ ($j \neq j'$) for two tokens that have been generated with different keywords. We use $\Delta_\tau \doteq \{\tau_1, \tau_2, \dots, \tau_m\}$ to denote the set of all (m) distinct tokens observed by the adversary.

3.2.1 Adversary’s observation and auxiliary information

We use `obs` to denote the adversary’s observation, which corresponds to both the leakage of the SSE scheme during its initialization step and while the client performs queries. Typically, `obs` comprises the access and search patterns of the queries, i.e., $\text{obs} = [(\tau, \mathbf{a})_1, (\tau, \mathbf{a})_2, \dots, (\tau, \mathbf{a})_\rho]$, where $(\tau, \mathbf{a})_r$ is the query token and access pattern of the r th query. Most statistical-based query recovery attacks compute certain summary statistics from the observation before running the attack [3, 18, 31, 34]. These statistics are typically the *volume* and *frequency* information, which we define next. The matrix of observed volumes \mathbf{V} is an $m \times m$ matrix whose j, j' th entry contains the number of documents that match both query tokens τ_j and $\tau_{j'}$, i.e., $\mathbf{V}_{j,j'} = \mathbf{a}_j^T \mathbf{a}_{j'} / N_d$. The vector of observed query frequencies \mathbf{f} is a vector of length m whose j th entry contains the number of times the client queried for τ_j , normalized by the total number of queries ρ . The (Markov) matrix of observed query frequencies \mathbf{F} is an $m \times m$ matrix whose j, j' th entry contains the number of times the client sent token $\tau_{j'}$ followed by τ_j , normalized by the total number of queries with token $\tau_{j'}$ (which we denote by $\rho(\tau_{j'})$).

We use `aux` to denote the adversary auxiliary’s information. Since we consider statistical-based attacks, this information does not contain ground-truth knowledge about the queries and/or dataset. The adversary uses `aux` to compute the vectors and matrices $\tilde{\mathbf{V}}, \tilde{\mathbf{f}}, \tilde{\mathbf{F}}$, whose structure is the same as the variables that the adversary can compute for the observations ($\mathbf{V}, \mathbf{f}, \mathbf{F}$), defined above. In this paper, we generate the auxiliary information related to keyword volume by giving the adversary non-indexed documents [9, 31, 34] (i.e., documents not in the client’s dataset), and auxiliary frequency information by giving the adversary outdated query frequencies [31]. Note that the observed variables ($\mathbf{V}, \mathbf{f}, \mathbf{F}$) refer to volume and frequency statistics computed from the *query tokens*, but the auxiliary variables ($\tilde{\mathbf{V}}, \tilde{\mathbf{f}}, \tilde{\mathbf{F}}$) refer to the *keywords*. For example, $\tilde{\mathbf{V}}$ is an $n \times n$ matrix whose i, i' th entry contains an *approximation* of the percentage of documents that have both keywords k_i and $k_{i'}$ in the dataset. In most of the SSE schemes we consider in this paper, there is a one-to-one correspondence between keywords and tokens, which the adversary aims to guess.² We note that auxiliary information is *imprecise*: the actual percentage of documents with keywords k_i and $k_{i'}$ in the dataset will likely differ from $\tilde{\mathbf{V}}_{i,i'}$.

²We also consider two schemes where different tokens might correspond to the same keyword [13, 36]; we explain the details in Sections 5.2 and 6.

3.2.2 Leakage scenarios

We consider three different leakage scenarios in our paper. These scenarios abstract from the actual SSE scheme being used, and all SSE schemes that have *at least* this leakage are subject to the attacks we study in that setting.

- **S1: full access-pattern leakage, no queries.** In this scenario, the client uses a scheme whose initialization step leaks which documents are returned as a response to each query token. (This corresponds to the leakage type *L2* by Cash et al. [3].) A simple example of this is a scheme where the client uses deterministic encryption to generate a query token from a keyword, and simply sends the server the list of encrypted documents with the query tokens corresponding to the keywords they contain. In this setting, we run the attack before the client has performed any query. The adversary’s observation is therefore $\text{obs} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m]$, and thus the attack relies solely on volume information (\mathbf{V} and $\tilde{\mathbf{V}}$).

This setting includes academic proposals [17, 24] as well as commercial products (e.g., see Cash et al. [3]).

- **S2: access-pattern leakage for queries.** In this scenario, the initialization step does not leak any information. The client issues queries, and each token τ leaks its access pattern \mathbf{a} , i.e, the list of documents returned in response to that token. (This corresponds to the leakage type *L1* by Cash et al. [3].) Let $(\tau, \mathbf{a})_r$ be the query token and access pattern of the r th query and assume that the client performs ρ queries in total. The adversary’s observation is $\text{obs} = [(\tau, \mathbf{a})_1, (\tau, \mathbf{a})_2, \dots, (\tau, \mathbf{a})_\rho]$

In this scenario, some keywords might never be queried, so the adversary might not see all possible access patterns ($m \leq n$), contrary to *S1*. We study two cases: when the adversary does not have auxiliary frequency information and thus it relies on volume information only (\mathbf{V} and $\tilde{\mathbf{V}}$), and when the adversary can additionally exploit independent query frequencies (\mathbf{f} and $\tilde{\mathbf{f}}$) to aid the attack.

Most efficient SSE schemes exhibit at least this leakage [2, 4, 5, 8, 17, 19, 23, 24, 28, 30, 37, 38]

- **S3: frequency-only leakage.** Finally, we consider a setting identical to *S2* where each keyword matches exactly one document, and no two keywords match the same document. This means that $n = N_d$ and, without loss of generality, k_i matches only d_i . In this case, only frequency information is useful for query recovery: $\text{obs} = [(\tau)_1, (\tau)_2, \dots, (\tau)_\rho]$.

We consider this setting to evaluate the effectiveness of our attack in the presence of query dependencies without volume leakage (i.e., the adversary only uses \mathbf{F} and $\tilde{\mathbf{F}}$). This is important since our attack is the first query recovery attack that can exploit query correlations (besides the attack example in the appendix of [14]).

3.2.3 Adversary’s goal and success metrics

The goal of the adversary is to carry out a query recovery attack, i.e., to find the underlying keywords of each query token. The outcome of the attack is therefore an injective (one-to-one) mapping from the set of query tokens to the set of keywords, which we denote by $p(\cdot) : [m] \rightarrow [n]$. For example, $p(j) = i$ denotes that the adversary guesses that the underlying keyword of token τ_j is k_i . We sometimes represent this mapping as an $n \times m$ matrix \mathbf{P} where $\mathbf{P}_{i,j} = 1$ if $p(j) = i$, and 0 otherwise. We use \mathcal{P} to denote the set of all valid mappings \mathbf{P} , i.e., $\mathcal{P} \doteq \{\mathbf{P} | \mathbf{1}_n^T \mathbf{P} = \mathbf{1}_m^T, \mathbf{P} \mathbf{1}_m \leq \mathbf{1}_n, \mathbf{P} \in \{0, 1\}^{n \times m}\}$, where $\mathbf{1}_n$ is an all-ones column vector of length n .

A query recovery attack takes the observations `obs` and auxiliary information `aux` as input, and produces a mapping of query tokens to keywords \mathbf{P} as output. We measure the success of an attack as the *percentage of observed query tokens* ($\tau \in \Delta_\tau$) for which the attack *correctly* guesses their underlying keyword. This is the most popular success metric in previous works, and it is referred to as the *attack accuracy* [18, 26, 31, 34] or the *query recovery rate* [1, 3].

4 Quadratic Query Recovery with IHOP

In this section, we present our query recovery attack, which we call IHOP (Iteration Heuristic for quadratic Optimization Problems). We begin with an overview of statistical query recovery attacks, noticing that some solve linear optimization problems, while others are based on quadratic problems. We then propose IHOP, which uses a linear optimization solver to iteratively look for a (suboptimal) solution to a quadratic query recovery problem. IHOP is a general algorithm that can be used to minimize different quadratic objective functions. We particularize it to use the volume and frequency statistics we mentioned in Section 3.2.1, using a maximum likelihood-based objective function. We evaluate the performance of IHOP in Sections 5 and 6.

4.1 Linear and Quadratic Query Recovery

Most query recovery attacks can be framed as an optimization problem, where the adversary tries to find the assignment of keywords to query tokens \mathbf{P} that minimizes a certain objective function. These problems are typically linear or quadratic with respect to \mathbf{P} . Linear query recovery attacks [26, 31] can be formulated as

$$\mathbf{P} = \arg \min_{\mathbf{P} \in \mathcal{P}} \sum_{i \in [n]} \sum_{j \in [m]} c_{i,j} \cdot \mathbf{P}_{i,j}. \quad (1)$$

This problem follows the structure of a Linear Assignment Problem (LAP), which can be optimally solved with a computational cost of $O(n \cdot m + m^2 \cdot \log m)$ [11]. The constants $c_{i,j}$ represent the cost of assigning keyword k_i to token τ_j .

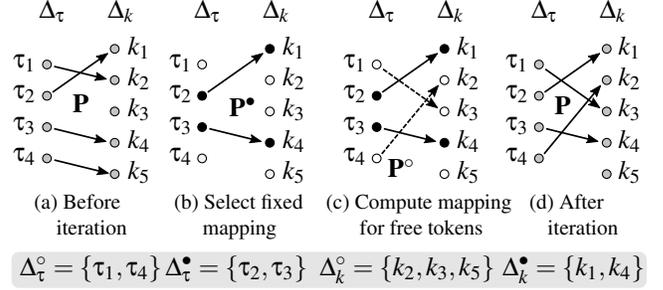


Figure 1: One iteration of IHOP

Quadratic query recovery attacks follow the formulation

$$\mathbf{P} = \arg \min_{\mathbf{P} \in \mathcal{P}} \sum_{i,i' \in [n]} \sum_{j,j' \in [m]} c_{i,i',j,j'} \cdot \mathbf{P}_{i,j} \cdot \mathbf{P}_{i',j'}, \quad (2)$$

where $c_{i,i',j,j'}$ is the cost of jointly assigning keyword k_i to token τ_j and $k_{i'}$ to token $\tau_{j'}$. This mathematical problem follows Lawler’s [25] formulation of the general Quadratic Assignment Problem (QAP), which is known to be NP-complete [35]. Existing attacks that follow this formulation thus rely on sub-optimal heuristic algorithms to find a solution [18, 34].

4.2 Quadratic Query Recovery via Linear Assignments

While there are efficient optimal solvers for the LAP (1), solvers for the QAP are suboptimal and heuristic. However, attacks based on a LAP cannot exploit the quadratic terms in a QAP that can contain valuable information for query recovery. We present IHOP, a query recovery attack that relies on efficient solvers for the LAP to iteratively solve a QAP.

We first explain a single iteration of our attack at a high level, and then we present its formal description. Figure 1 shows a toy example of an iteration of our attack. At the beginning of this iteration, the attack has an assignment of tokens to keywords \mathbf{P} (Fig. 1a). Then, it *fixes* some of those assignments at random (denoted \mathbf{P}^\bullet , see Fig. 1b) and *freed* the remaining keywords and query tokens. The sets of free query tokens Δ_τ° , fixed query tokens Δ_τ^\bullet , free keywords Δ_k° , and fixed keywords Δ_k^\bullet for this example are shown in the bottom of the figure. The attack re-computes an assignment of the free query tokens to the free keywords, leveraging the fixed assignment \mathbf{P}^\bullet to improve the re-estimation (Fig. 1c). This yields an update of the assignment \mathbf{P} and ends the iteration (Fig. 1d). The idea of this approach is that, if some assignments are fixed, we can use the quadratic terms involving those assignments and the assignments we wish to update while keeping the optimization problem linear.

We formally describe the attack in Algorithm 1. The attack receives the observations `obs` and auxiliary information `aux`,

Algorithm 1 IHOP

```
1: procedure IHOP(aux, obs, niters, pfree)
2:   P ← SolveLinear(Δk, Δτ, 0, 0, 0, aux, obs)
3:   for i ← 1 . . . niters do
4:     Δτ◦ ←  $\lceil \frac{p_{free} \cdot m}{m} \rceil$  Δτ      ▷ Choose pfree · m tokens
5:     Δτ• = {τj | τj ∈ Δτ, τj ∉ Δτ◦}      ▷ Fixed tokens
6:     Δk• = {ki | i = p(j), ki ∈ Δk, τj ∈ Δτ•}
7:     Δk◦ = {ki | ki ∈ Δk, ki ∉ Δk•}
8:     P• = {τj → kp(j) | τj ∈ Δτ•}
9:     P◦ ← SolveLinear(Δk◦, Δτ◦, Δk•, Δτ•, P•, aux, obs)
10:    P ← combine(P◦, P•)
11:  return P
```

Algorithm 2 SolveLinear

```
1: procedure SolveLinear(Δk◦, Δτ◦, Δk•, Δτ•, P•, aux, obs)
2:   Get c ≡ {ci,i',j,j'} and d ≡ {di,j} using aux, obs.
3:   Solve the linear assignment problem:
```

$$\mathbf{P}^\circ = \arg \min_{\mathbf{P}^\circ \in \mathcal{P}^\circ} \sum_{k_i \in \Delta_k^\circ} \sum_{\tau_j \in \Delta_\tau^\circ} \left(\sum_{\tau_{j'} \in \Delta_\tau^\circ} \sum_{k_{i'} \in \Delta_k^\circ} c_{i,i',j,j'} \cdot \mathbf{P}_{i,j}^\circ \cdot \mathbf{P}_{i',j'}^\circ + d_{i,j} \cdot \mathbf{P}_{i,j}^\circ \right).$$

```
4:   return P◦
```

and two parameters: the number of iterations n_{iters} and the percentage of free tokens for each iteration p_{free} . The attack begins with an initialization step, where it computes \mathbf{P} by solving a linear problem *SolveLinear* that we specify below (Line 2). Then, it iterates n_{iters} times. At each iteration, the attack splits the set of query tokens Δ_τ at random into two groups: the group of *free* tokens Δ_τ° , which contains $\lceil p_{free} \cdot m \rceil$ tokens (Line 4), and the group of *fixed* tokens Δ_τ^\bullet , which contains all the other tokens (Line 5). Let Δ_k^\bullet be the set of keywords assigned to tokens in Δ_τ^\bullet by \mathbf{P} , and let Δ_k° be the set of all the keywords not in Δ_k^\bullet ; i.e., the *free* keywords (Lines 6–7). The fixed matching \mathbf{P}^\bullet is the bijective mapping of $\Delta_\tau^\bullet \rightarrow \Delta_k^\bullet$ extracted from \mathbf{P} (Line 8). Then, the attack looks for an assignment \mathbf{P}° of the free tokens Δ_τ° to the free keywords Δ_k° by solving a linear problem (Line 9). The attack finally updates \mathbf{P} using the newly computed \mathbf{P}° and the fixed assignment \mathbf{P}^\bullet , and this finishes the iteration.

A key component of this attack is the linear problem *SolveLinear*: this problem specifies how the adversary uses the auxiliary information *aux*, the observations *obs*, and the fixed mapping \mathbf{P}^\bullet to update the matching \mathbf{P} in each iteration. We show this problem in Algorithm 2. This problem is an instantiation of a QAP (2) that only considers quadratic terms that depend on \mathbf{P}° and \mathbf{P}^\bullet , but never between two terms in \mathbf{P}° . Thus, the problem is linear (in \mathbf{P}°) and can be written as a LAP (1). Since this problem is a LAP, it can be efficiently and optimally solved with the Hungarian algorithm [22, 25]. The constants c and d are computed from *obs* and *aux*, and

their actual expression depends on the SSE setting in question; below, we give expressions for these constants for the volume and frequency leakage scenarios we defined in Section 3.2.1. The constants $d_{i,j}$ are simply the terms $c_{i,i,j,j}$ in (2), that we write separately from c in Algorithm 2 to distinguish purely linear coefficients (d) from quadratic ones (c). Our attack therefore provides a *template* that researchers can use to instantiate quadratic query recovery attacks that optimize different objective functions, captured by c and d .

4.3 Coefficient Selection in IHOP

We follow an approach inspired by maximum likelihood estimation to set the values of the coefficients (c and d) in each iteration of *SolveLinear*. In short, *SolveLinear* finds the assignment between free keywords and tokens \mathbf{P}° that maximizes the *log-likelihood* of the observations *obs* given the auxiliary information *aux*.

First, we present the expressions when only one source of leakage is available to the adversary: either volume leakage, frequency leakage with independent queries, or frequency leakage with correlated queries. Then, we explain how we combine them when more than one source of leakage is available to the adversary. To derive these expressions, we leverage *binomial* and *Poisson* statistical models for the observations given the auxiliary information. We note that this is merely a tool to develop our expressions: we run the experiments in Sections 5 and 6 with real data.

4.3.1 Volume leakage.

Recall that \mathbf{V} is the matrix of observed token volumes, and $\tilde{\mathbf{V}}$ is the matrix of auxiliary keyword volumes. We use a binomial model to get the coefficients c and d . Namely, assume that, when keyword k_i corresponds to token τ_j (denoted $k_i \rightarrow \tau_j$), the number of documents with token τ_j (i.e., $N_d \cdot \mathbf{V}_{j,j}$) follows a binomial distribution with N_d trials and probability given by the auxiliary volume $\tilde{\mathbf{V}}_{i,i}$. Then, the log-likelihood cost of assigning token τ_j to keyword k_i is

$$d_{i,j} = -\log \Pr(\text{Bino}(N_d, \tilde{\mathbf{V}}_{i,i}) = N_d \cdot \mathbf{V}_{j,j}), \quad (3)$$

where *Bino* denotes a binomial distribution. We added a minus sign since we aim at maximizing the log-likelihood, but the linear problem in Alg. 2 is a minimization. Expanding this probability and ignoring the summands that do not affect the optimization problem, we get

$$d_{i,j} = -N_d [\mathbf{V}_{j,j} \log(\tilde{\mathbf{V}}_{i,i}) - (1 - \mathbf{V}_{j,j}) \log(1 - \tilde{\mathbf{V}}_{i,i})]. \quad (4)$$

We follow the same approach for the quadratic terms $c_{i,i',j,j'}$, i.e., we assume that when $k_i \rightarrow \tau_j$ and $k_{i'} \rightarrow \tau_{j'}$ the number of documents that have both tokens τ_j and $\tau_{j'}$ follows a Binomial distribution parametrized by $\tilde{\mathbf{V}}_{i,i'}$, and thus the quadratic term

is:

$$c_{i,i',j,j'} = -N_d [\mathbf{V}_{j,j'} \log(\tilde{\mathbf{V}}_{i,i'}) - (1 - \mathbf{V}_{j,j'}) \log(1 - \tilde{\mathbf{V}}_{i,i'})]. \quad (5)$$

4.3.2 Frequency leakage with independent queries.

Recall that \mathbf{f} is the vector of observed token frequencies, ρ is the total number of queries issued by the client, and $\tilde{\mathbf{f}}$ is the vector of auxiliary keyword frequencies. We use a Poisson model to derive the attack coefficients. This means that we assume that, when $k_i \rightarrow \tau_j$, the number of times the client sent token τ_j follows a Poisson distribution with rate $\rho \cdot \tilde{\mathbf{f}}_i$. Thus, the cost of assigning $k_i \rightarrow \tau_j$ is $d_{i,j} = -\log \Pr(\text{Pois}(\rho \tilde{\mathbf{f}}_i) = \rho \mathbf{f}_j)$. Expanding this expression and ignoring the summands that do not depend on both i and j , we get

$$d_{i,j} = -\rho \cdot \mathbf{f}_j \cdot \log \tilde{\mathbf{f}}_i. \quad (6)$$

In this case, the quadratic coefficients are $c_{i,i',j,j'} = 0$ since all queries are independent.

4.3.3 Frequency leakage with dependent queries.

We extend this approach to the case where we have dependent queries. Recall that $\rho(\tau_j)$ is the number of times the client queried using token τ_j , and that $\mathbf{F}_{j,j'}$ contains the number of times the client queried for token $\tau_{j'}$ followed by τ_j , normalized by $\rho(\tau_{j'})$. $\tilde{\mathbf{F}}_{i,i'}$ is the probability that the client queried for $k_{i'}$ followed by k_i , as computed from the auxiliary information. For each free token $\tau_j \in \Delta_\tau^\circ$, we first compute the overall probability that the client sent that token after sending *any* other free token $\tau_{j'} \in \Delta_\tau^\circ$ ($j \neq j'$):

$$\mathbf{F}_{j,\circ} \doteq \frac{\sum_{\tau_{j'} \in \Delta_\tau^\circ \setminus \tau_j} \rho(\tau_{j'}) \cdot \mathbf{F}_{j,j'}}{\sum_{\tau_{j''} \in \Delta_\tau^\circ} \sum_{\tau_j \in \Delta_\tau^\circ \setminus \tau_{j''}} \rho(\tau_{j'}) \cdot \mathbf{F}_{j'',j'}}.$$

The numerator counts the number of transitions from any other free token to τ_j and the denominator is a normalization term so that $\sum_{\tau_j \in \Delta_\tau^\circ} \mathbf{F}_{j,\circ} = 1$. We also define $\rho(\Delta_\tau^\circ)$ as the number of times the client queried for a free token other than τ_j , i.e., $\rho(\Delta_\tau^\circ) = \sum_{\tau_{j'} \in \Delta_\tau^\circ \setminus \tau_j} \rho(\tau_{j'})$

Once again we follow a Poisson model to derive the coefficients. When $k_i \rightarrow \tau_j$, the number of times the client queried for τ_j consecutively (i.e., $\rho(\tau_j) \cdot \mathbf{F}_{j,j}$) follows a Poisson distribution with rate $\rho(\tau_j) \cdot \tilde{\mathbf{F}}_{i,i}$. Likewise, the number of times the client queried for another free token followed by τ_j ($\rho(\Delta_\tau^\circ) \cdot \mathbf{F}_{j,\circ}$) follows a Poisson distribution with rate $\rho(\tau_\circ) \cdot \tilde{\mathbf{F}}_{i,\circ}$. Thus, similar to (6), we set

$$d_{i,j} = -[\rho(\tau_j) \cdot \mathbf{F}_{j,j} \cdot \log(\tilde{\mathbf{F}}_{i,i}) + \rho(\Delta_\tau^\circ) \cdot \mathbf{F}_{j,\circ} \cdot \log(\tilde{\mathbf{F}}_{i,\circ})]. \quad (7)$$

To get the quadratic terms, we assume that when $k_i \rightarrow \tau_j$ and $k_{i'} \rightarrow \tau_{j'}$, the number of times the client sent token τ_j after

sending $\tau_{j'}$ follows a Poisson distribution with rate $\rho(\tau_j) \cdot \tilde{\mathbf{F}}_{i,i'}$ (and vice-versa when the client queried for $\tau_{j'}$ after τ_j), which yields

$$c_{i,i',j,j'} = -[\rho(\tau_{j'}) \cdot \mathbf{F}_{j,j'} \cdot \log(\tilde{\mathbf{F}}_{i,i'}) + \rho(\tau_j) \cdot \mathbf{F}_{j',j} \cdot \log(\tilde{\mathbf{F}}_{i',i})]. \quad (8)$$

4.3.4 Leakage combinations

When the scheme allows both volume and frequency leakage, and the adversary has the corresponding auxiliary information, we combine the coefficients *additively*. For example, when the scheme leaks the volume and the (independent) query frequencies, we set $d_{i,j}$ as the sum of (4) and (6), and $c_{i,i',j,j'}$ is simply (5). The intuition behind this is that, since we are using log-likelihoods, adding coefficients is equivalent to multiplying probabilities, and thus this approach is the maximum likelihood estimator when the volume and frequency leakages are independent.

4.4 IHOP vs. other attacks

IHOP shares certain concepts with previous attacks; we clarify the differences between our attack and related work following. SAP [31] estimates the keywords of each query by solving a single LAP using a maximum likelihood-based approach. Therefore, SAP can be written as a single execution of *SolveLinear* with only linear (d) coefficients. IHOP can be seen as an extension of this idea that accommodates quadratic coefficients (c) and incorporates the linear program in an iteration loop (Alg. 1). This results in a more advanced and altogether different attack. The connection between IHOP and attacks like GraphM [34] and IKK [18] is that all these attacks aim at solving a QAP using different heuristics. IKK aims at minimizing a Frobenius norm $\|\tilde{\mathbf{V}} - \mathbf{PVP}^T\|_F$, while GraphM adds an additional linear term to this objective function. IHOP minimizes a totally different function, characterized by the coefficients c and d we explained above. IHOP's function is the first one that can incorporate query correlations into the optimization problem (by setting d as in (8) and (7)). Aside from the objective functions, the solvers for these three attacks are also completely different: GraphM uses a convex-concave relaxation method, IKK uses simulated annealing, and IHOP uses a novel iterative algorithm based on solving linear problems.

5 Evaluation: Volume-Leaking SSE Schemes

In this section, we compare the performance of our attack against other attacks on SSE schemes that leak the access pattern (S1 and S2 leakage scenarios in Section 3.2.1). We implement our experiments³ using Python3.8, and run our code in a machine running Ubuntu 16.04, with 1 TB RAM and an Intel(R) Xeon(R) CPU E7 (2.40GHz, 160 CPUs). We use

³<https://github.com/simon-oya/USENIX22-ihop-code>

Oya and Kerschbaum’s implementation of SAP [31] and the PATH algorithm [42] from GraphM package⁴ to implement Pouliot and Wright’s GraphM attack [34].

We use five public datasets: Enron, Lucene, Movies, News, and NYTimes. In Appendix C, we explain how we obtain and process these datasets, and show their statistical differences (Fig. 12). During this processing, we keep the 3000 most popular keywords of each dataset, and remove the rest. We download the query frequencies of each of those keywords for each week of 2020 from Google Trends⁵ using the gtab package [41]. We found that our results are qualitatively similar across datasets. Thus, in most of our experiments we show only results in a few datasets, and provide the complete results in Appendix E.

We compare IHOP with other *statistical-based* attacks, and disregard attacks that rely on ground-truth data [1, 3, 9, 29], since we do not consider this type of leakage in this paper.

- SAP [31] is an attack by Oya and Kerschbaum that uses search and access-pattern leakage and a maximum likelihood approach to solve a LAP (1). We use the original implementation by the authors. We set the hyperparameter $\alpha = 0.5$ as recommended by the authors.
- GraphM [34] is the graph matching attack by Pouliot and Wright, which aims at minimizing the function $(1 - \alpha)\|\tilde{\mathbf{V}} - \mathbf{PVP}^T\|_F^2 - \alpha\text{Tr}(\mathbf{P}^T\mathbf{C})$, where $\|\cdot\|_F$ denotes the Frobenius norm and \mathbf{C} is a maximum likelihood-based term. Following the original work [34], we use the PATH algorithm [42] to find a solution to this problem. We use $\alpha = 0.5$, since we empirically found it performs best.
- IKK [18] aims at minimizing $\|\tilde{\mathbf{V}} - \mathbf{PVP}^T\|_F$ using simulated annealing [21]. We set the initial temperature parameter of simulated annealing to $T = 100$, a cooling factor of $p_{cool} = 0.99995$, and end the algorithm when $T < 10^{-10}$. This yields running times close to GraphM’s.
- Freq is the frequency attack by Liu et al. [26], which uses exclusively frequency information, and simply maps each query token to the keyword whose frequency is closest in Euclidean distance.

We tested other generic algorithms for the QAP applied to $\|\tilde{\mathbf{V}} - \mathbf{PVP}^T\|_F$, namely the spectral algorithm by Umeyama [39] (Umeyama) and the fast projected fixed-point algorithm by Lu et al. [27] (FastPFP). We found these algorithms provide significantly lower accuracy than the other attacks in this list, so we omit them from the paper.

In our experiments, we vary the keyword universe size (n) and dataset size (N_d). Given a value of n , we simply build the keyword universe in each run of the experiment by selecting n keywords at random from the set of 3000. Given a value of N_d , we sample N_d documents from the dataset at random to

⁴<https://projects.cbio.mines-paristech.fr/graphm/>

⁵<https://trends.google.com/trends>

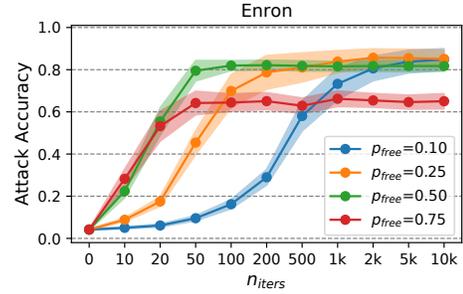


Figure 2: Evolution of IHOP’s accuracy with the number of iterations and different p_{free} values in Enron dataset. (S1)

build the client’s dataset. We also give the adversary N_d^{aux} non-indexed documents selected at random (i.e., they are not in the client’s dataset) as auxiliary information to compute $\tilde{\mathbf{V}}$. We repeat all of our experiments 30 times, measuring the query recovery accuracy and the running time of the attacks. We run each attack in a single thread, ensuring that our running time comparisons are fair. Shaded areas in our plots represent the 95% confidence intervals for the average accuracy.

5.1 Volume-only leakage attacks

We consider the setting where the adversary has auxiliary volume information only, and relies on access pattern leakage to estimate the keywords of each query. Since we do not consider frequency information, we set the coefficients of IHOP as in (4) and (5). We evaluate attacks that use volume (IHOP, SAP, IKK, GraphM) and exclude Freq from these experiments.

We first consider the case where the adversary observes the access pattern of each query token during initialization (S1, $m = n$). Later, we consider the case where only the access pattern of the queried tokens is leaked (S2, $m \leq n$).

IHOP parametrization (S1). We perform an initial experiment to understand how the number of iterations n_{iters} and the percentage of free tokens p_{free} affect the performance of IHOP. Figure 2 shows the attack accuracy (percentage of correctly guessed query tokens) with the number of iterations, for different values of p_{free} , in Enron dataset with $n = 500$ keywords, $N_d = 20000$ client documents, and $N_d^{aux} = 5000$ auxiliary information documents for the adversary. We see that the accuracy increases with the number of iterations and that it converges in all cases, reaching a point where increasing n_{iters} does not yield further improvement. With $p_{free} = 0.25$, the attack grows from an accuracy of 4.1% before iterating to an accuracy of 78.7% at 200 iterations, after running for only 38 seconds. Larger p_{free} values allow the algorithm to converge faster, since the algorithm re-computes more assignments per iteration in those cases. However, a small p_{free} increases the number of quadratic terms that are exploited in each iteration,

and thus yields higher asymptotic accuracy. We also note that smaller p_{free} implies faster running times, since each iteration of IHOP uses the Hungarian algorithm which has a cost that is $O(n \cdot m' + m'^2 \cdot \log m')$ [11], where $m' = \lceil m \cdot p_{free} \rceil$. The running time of IHOP for $p_{free} = 0.1, 0.25, 0.5$, and 0.75 in Enron was 0.14, 0.19, 0.34, and 0.46 seconds per iteration, respectively. We use $p_{free} = 0.25$ in the remainder of the paper, since it offers a good trade-off between convergence speed and asymptotic performance.

Attack comparison with N_d^{aux} (S1). We compare the performance of the different attacks in the S1 setting. We set the number of keywords to $n = 500$, since GraphM is computationally expensive. We build the client’s dataset with $N_d = 20000$ documents, and vary the amount of documents we give the adversary as auxiliary information $N_d^{aux} \in [500, 10000]$. Figure 3 shows the results (we used IHOP with 1000 iterations and $p_{free} = 0.25$). We see that the attacks’ accuracy increases as the quality of the auxiliary information grows. The accuracy depends on the dataset, but we see that IHOP achieves higher accuracy than other attacks except in the particular case of Enron with $N_d^{aux} = 500$ and 1000. The average running times of the attacks in Enron, which are similar across N_d^{aux} , are 33 seconds for SAP, 231 seconds for IHOP, 3.3 hours for GraphM, and 2.3 hours for IKK.

Attack comparison with n (S1). Next, we study how the keyword universe size affects the performance of the attacks. We set $N_d = 20000$ and $N_d^{aux} = 10000$, and progressively increase n from 500 to 1000. Figure 4 shows the results in Enron and Lucene datasets. The top plots show that the attack accuracy remains steady for IHOP, slightly decreases for GraphM, and significantly decreases for IKK. The bottom plots show the running times in logarithmic scale. We see that GraphM and IKK quickly become unfeasible as the keyword universe size increases: with $n = 1000$ keywords, GraphM takes around 3 days to finish, while IHOP ends in 15 minutes and achieves higher accuracy.

Attack comparison with n and fixed number of observed tokens m (S2). We repeat the previous experiment in the setting where only the access patterns of the queried keywords are leaked (S2). We fix the number of distinct keywords queries (i.e., distinct tokens observed) to $m = 500$, and increase n as above. This means that the adversary observes the access pattern of $m = 500$ query tokens and has to guess their corresponding keywords from a larger set $n \geq 500$. We show the results in Figure 5. As expected, the accuracy of all attacks decreases in this case compared to when all access patterns are observed. GraphM is particularly affected by this. This is due to a limitation of the objective function that GraphM minimizes ($\|\tilde{\mathbf{V}} - \mathbf{PVP}^T\|_F^2$ unfairly penalizes unassigned keywords when $n \geq m$). IHOP achieves the highest

accuracy among all attacks in all datasets, and a running time orders of magnitude below IKK and GraphM.

5.2 Volume and frequency leakage attacks

For the next set of experiments, we consider both volume and frequency leakage when the client sends queries independently. For each experiment, after selecting the keyword universe Δ_k as above, we take the frequency data from Google Trends for the keywords in Δ_k over the span of 2020. Then, we use the average frequencies of the first half of this year as the auxiliary information $\tilde{\mathbf{f}}$ and generate keywords independently at random following the average query frequencies of the second half of the year (this affects the observed frequencies of tokens \mathbf{f}). Since we consider independent query generation, the adversary uses \mathbf{f} and $\tilde{\mathbf{f}}$ in their attack. We set the coefficients of IHOP as the summation of (4) and (6), and use $n_{iters} = 1000$ and $p_{free} = 0.25$. Besides the attacks we considered above, we also evaluate Freq.

Attack comparison with ρ (S2). We first compare the attacks that rely on frequency information (IHOP, SAP, and Freq). We set $n = 3000$, since these attacks are efficient, and vary the number of queries ρ . Figure 6 shows the attack accuracy (recall that this is the percentage of *distinct* query tokens whose underlying keyword is correctly guessed by the attack). IHOP comfortably beats both SAP and Freq (+20% more accuracy than these attacks in all cases), which are the state-of-the-art in this setting. This is because IHOP is the only attack that can exploit both frequency and volume co-occurrence information at the same time.

Access-pattern obfuscation defenses (S2). We evaluate all attacks against two access-pattern obfuscation defenses: CLRZ [6] and OSSE [36]. CLRZ randomly adds and removes keywords from documents before outsourcing the database. The False Positive Rate (FPR) of CLRZ is the probability that it adds a keyword to a document that does not have it, and the True Positive Rate (TPR) is the probability that it keeps a keyword in a document that has it. CLRZ obfuscation occurs before outsourcing the database, and therefore querying for the same keyword twice produces the same (obfuscated) access pattern. OSSE achieves the same effect during query time: querying for a particular keyword can return documents that do not contain that keyword and miss some documents that do contain the keyword. Querying for the same keyword twice yields (with high probability) different access patterns in OSSE, providing a certain level of search pattern privacy.

We assume the adversary knows the defense parameters (TPR and FPR) and explain in Appendix A how we adapt all the attacks against these defenses. We set $n = 500$, since GraphM is too slow for larger keyword universes, and set the number of queries to $\rho = 500$. We set TPR=0.9999 and vary FPR for both defenses. Figure 7 shows the accuracy of the

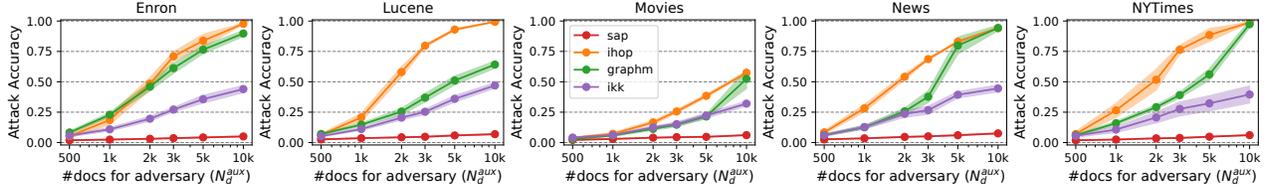


Figure 3: Attack comparison in all datasets vs. the number of non-indexed documents of the adversary (N_d^{aux}) (S1)

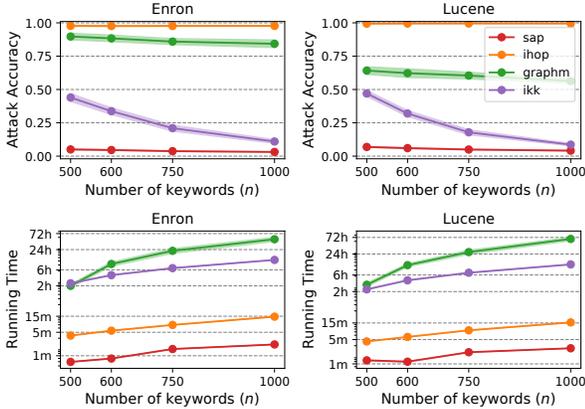


Figure 4: Attack comparison vs. number of keywords (n) (S1)

attacks against CLRZ and OSSE in Enron dataset. The figure shows that OSSE is a stronger defense than CLRZ, but we observe a smaller difference between these defenses than Shang et al. [36], since we have adapted the attacks against OSSE more efficiently. More importantly, we see a large accuracy gap between IHOP and the other attacks.⁶

Other defenses: discussion. We only considered the CLRZ and OSSE defenses because they affect the observed volumes \mathbf{V} without modifying the general leakage format we assume in this paper. We did not consider other defenses like the proposal by Patel et al. [32] or SEAL [10] because they pad the documents that are returned to each keyword independently, but do not consider how to pad volume co-occurrence. This works when each document has a single keyword (or attribute), but not in the *keyword query* scenario we consider in this paper unless the client replicates each document once per each of its keywords, which requires an unfeasible storage cost. We also disregarded PANCAKE [13] in this section since it hides frequencies but not volume (we consider this defense in the following section), and SWiSSSE [16] since it is an altogether more complex SSE scheme that does not fit our

⁶We observe an increase in the accuracy of IHOP when FPR=0.01. We only observed this effect in Enron and Movie datasets, and we conjecture it is an effect of these datasets’ distribution and how IHOP’s objective function interacts with it.

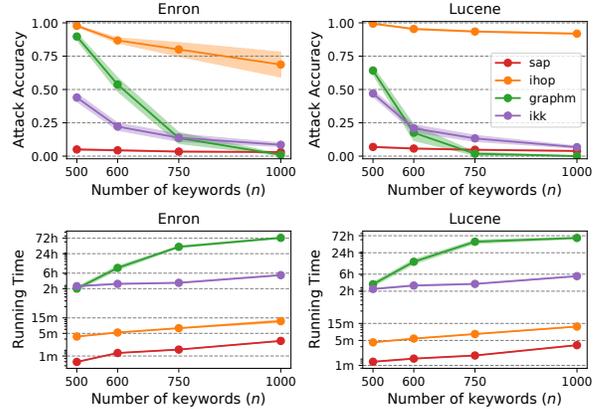


Figure 5: Attack comparison vs. number of keywords (n) when the adversary sees $m = 500$ distinct tokens (S2)

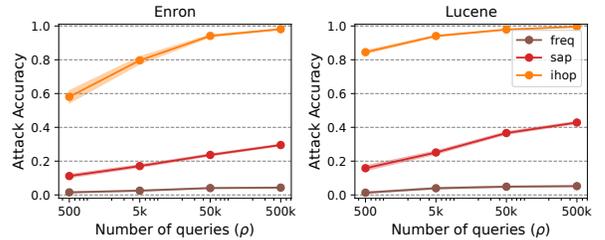


Figure 6: Comparison of attacks that use frequency leakage, with $n = 3000$ (S2)

description in this paper, and we believe deserves individual attention. We note that a purely ORAM-based defense can protect against IHOP, but these defenses come at a high overhead cost; our attack is effective against efficient defenses like CLRZ [6].

6 IHOP against PANCAKE with Query Dependencies

In this section, we evaluate the performance of IHOP in a setting where the client’s queries are correlated, i.e., querying for a particular keyword k_i affects the probability of the keyword chosen for the next query. This query model is interesting

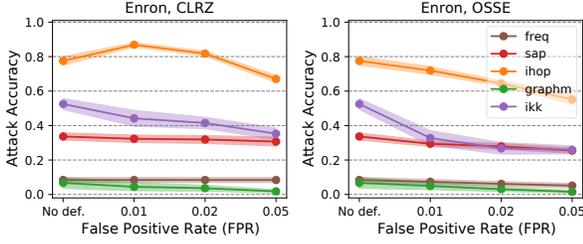


Figure 7: Attack comparison vs. CLRZ and OSSE in Enron dataset ($n = 500$, $\rho = 500$) (S2)

because humans rarely make independent decisions, and this type of query correlations have not been considered in attacks against SSE schemes before (except briefly by Grubbs et al. in their appendix [14]). As we explained in Sec. 4.3.3, IHOP can take these correlations into account. We consider a case where there is frequency-only leakage (S3) so that the attack’s success relies exclusively on exploiting these query dependencies. A particular case of frequency-only leakage is when the client simply queries for individual documents using their identifier or a “document title”.

Besides showing the effectiveness of IHOP with query correlations, the main contribution of this section is an evaluation of PANCAKE, the frequency hiding defense by Grubbs et al. [13], in the presence of query dependencies. We provide an overview of PANCAKE below, then explain how we adapt IHOP against PANCAKE, and finally introduce our experimental setup and show our results.

6.1 Overview of PANCAKE

PANCAKE [13] is a system that hides the frequency access of key-value datasets by using a technique called frequency smoothing. PANCAKE uses *document replication* and *dummy queries* to ensure that the access frequency to the key-values in the dataset is uniform. For compatibility with our notation, we consider the case where the keyword universe and dataset have the same size ($n = N_d$), and without loss of generality keyword k_i matches only document d_i ($i \in [n]$).

We use Figure 8 to summarize how PANCAKE works. In the figure, $n = N_d = 3$. Let \mathbf{f}^{real} be a vector of length $n + 1$, where its i th entry f_i^{real} contains the real query frequency of keyword k_i , and $f_{n+1}^{real} = 0$ represents the real query frequency of a dummy keyword-document pair $k_\delta - d_\delta$. We assume that the client knows \mathbf{f}^{real} for simplicity (this only benefits the client and not the attack). The client creates $R(i) = \lceil f_i^{real} / n \rceil \geq 1$ replicas for each document d_i ($i \in [n]$), and additionally creates $R(n + 1) = 2n - \sum_{i=1}^n R(i)$ dummy replicas d_δ . This makes a total of $2n$ document replicas, that are encrypted and sent to the server (in a random order). In Figure 8a, the client creates two replicas for d_1 , one for d_2 and d_3 , and two dummy replicas d_δ . We refer to the ℓ th replica of k_i by $r_{i,\ell}$, with $\ell \in$

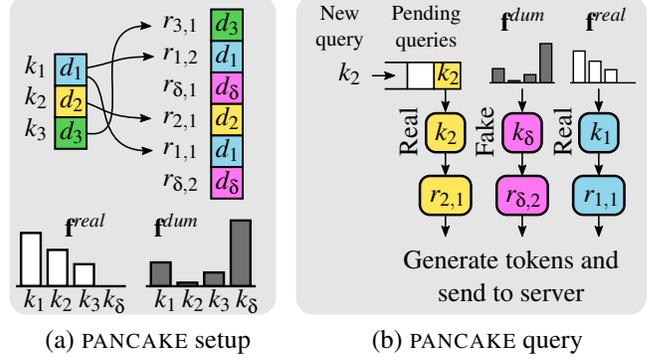


Figure 8: Overview of PANCAKE’s setup and queries.

$[R(i)]$. The client computes a vector of dummy frequencies $\mathbf{f}^{dum} = R(i) - f_i^{real} \geq 0$ for each $i \in [n + 1]$. For simplicity, we assume that the client knows the mapping of keywords to replicas and stores \mathbf{f}^{real} and \mathbf{f}^{dum} locally, and refer to the original paper for more advanced details [13].

When the client wishes to query for a keyword k_i , she places k_i inside a buffer of pending queries, and flips *three* unbiased coins (Fig. 8b). For each coin flip: if it shows heads, the client queries for the next keyword in her buffer of pending queries (e.g., k_2 in 8b). If the buffer is empty, she simply samples a keyword from \mathbf{f}^{real} and queries for it (e.g., k_1 in Fig. 8b). If the coin shows tails, the client samples a keyword from \mathbf{f}^{dum} and queries for it (e.g., k_δ in Fig. 8b). To query for a keyword, the client selects one of its replicas at random and generates a token that retrieves the document associated with that replica.

The probability that the client sends a query for keyword k_i is therefore $\frac{1}{2}f_i^{real} + \frac{1}{2}\mathbf{1}_i^{real} = R(i)/2n$. Since the client chooses one of the $R(i)$ replicas at random when querying for k_i , the access frequency of any replica is $1/2n$. Since each query token corresponds to one replica, the frequency of each token τ_j ($j \in [2n]$) is also $1/2n$ (uniform); i.e., query tokens are indistinguishable based on their frequencies.

6.2 Query dependencies against PANCAKE

PANCAKE ensures that each document replica is accessed with probability $1/2n$. However, when there are dependencies between the keywords that the client chooses, PANCAKE does not hide these dependencies. Grubbs et al. [13] provide strong evidence that PANCAKE obfuscates query correlations. Here, we perform another study of PANCAKE against dependent queries, and discuss our findings against the ones by Grubbs et al. [13] in Section 6.4. We consider the case where the client follows a Markov model to choose the keywords of her queries, and use \mathbf{F}^{real} to denote the $n \times n$ Markov matrix that characterizes the client’s querying behavior. We assume that the Markov process is irreducible and aperiodic so that it has a

unique stationary distribution, \mathbf{f}^{real} . This vector (\mathbf{f}^{real}) contains the probability that the client queries for each keyword (at any point in time) and can be computed analytically from \mathbf{F}^{real} . The client uses \mathbf{f}^{real} as input to PANCAKE to compute the replicas and dummy distribution. Even though the query frequency of each token is $1/2n$, the correlations between real queries cause correlations between query tokens which the adversary can leverage for query recovery (we show an example of this in Fig. 11, in the appendix).

Now we explain how we adapt our attack against PANCAKE. When the client queries ρ times, she sends a total of 3ρ tokens. The adversary observes the sequence of tokens $\text{obs} = [(\tau)_1, \dots, (\tau)_{3\rho}]$ and groups them, consecutively, into triples. Then, the adversary builds the $2n \times 2n$ matrix of observed frequencies \mathbf{F} by considering every two consecutive triples and counting each transition from a token in the first triple and a token in the second triple. This matrix is then normalized by columns, which ensures it is left-stochastic.

The adversary has an $n \times n$ auxiliary information matrix $\tilde{\mathbf{F}}$, which captures the client’s querying behavior (ideally, this matrix is close to \mathbf{F}^{real}). The adversary computes the stationary distribution of $\tilde{\mathbf{F}}$ (i.e., $\tilde{\mathbf{f}}$), and uses it to get the expected number of replicas for each keyword and the dummy profile that the client is using. Note that if $\tilde{\mathbf{F}}$ is far from \mathbf{F}^{real} , the adversary’s belief of the number of replicas per keyword and the dummy profile might be very far from the true ones. In Appendix B, we explain how the adversary can compute the matrix of *expected* frequencies between replicas given $\tilde{\mathbf{F}}$, which we denote by $\hat{\mathbf{F}}$. The adversary uses the matrix of observed token frequencies \mathbf{F} and expected replica frequencies $\hat{\mathbf{F}}$ to match tokens to replicas. This in turn yields a matching between tokens and keywords. This means that we adapt IHOP against PANCAKE by setting its c and d coefficients as in (8) and (7), using $\hat{\mathbf{F}}$ instead of $\tilde{\mathbf{F}}$ in these expressions.

6.3 Experimental Results

For our evaluation, we consider the case where the client’s database is an encyclopedia that she wishes to securely store on a server. Each document contains information about a particular topic, identified by its keyword. To build our datasets and compute the transition probabilities between the queries, we use data from the Wikipedia Clickstream dataset.⁷ We summarize our approach here, and provide full details in Appendix D. Using PetScan,⁸ we get all Wikipedia pages under the five categories ‘privacy’, ‘security’, ‘cryptology’, ‘politics’, and ‘activism’, including pages in subcategories. This yields a total of 5573 pages. Many of these pages have hyperlinks that allow a user to browse from one to another. We download the number of times Wikipedia users clicked on these hyperlinks to transition between these pages in 2020.

⁷<https://dumps.wikimedia.org/other/clickstream/>

⁸<https://petscan.wmflabs.org>

We build five keyword universes of size $n = 500$, each focused on one of the five categories above. For each of these keyword universes, we use the transition counts to build two 500×500 transition matrices: one with data from January 2020 to June 2020 (\mathbf{F}_{JAN}^{JUN}), and another one with data from July 2020 to December 2020 (\mathbf{F}_{JUL}^{DEC}). We use \mathbf{F}_{JUL}^{DEC} to generate the client’s real queries, i.e., $\mathbf{F}^{real} = \mathbf{F}_{JUL}^{DEC}$. We evaluate both the case where the adversary has low-quality auxiliary information ($\tilde{\mathbf{F}} = \mathbf{F}_{JAN}^{JUN}$, denoted $\text{aux} \downarrow$ in the plots) and high-quality information ($\tilde{\mathbf{F}} = \mathbf{F}_{JUL}^{DEC}$, denoted $\text{aux} \uparrow$).

We consider two values of the total number of queries: $\rho = 100\,000$ and $\rho = 500\,000$. These are large numbers, because the datasets contain $n = 500$ keywords, which generates 1 000 replicas with PANCAKE, and thus there are 1 000 possible query tokens. These tokens are all queried with the same probability (this is guaranteed by PANCAKE), and thus with $\rho = 100\,000$ (resp., $\rho = 500\,000$) the server observes ≈ 300 (resp., $\approx 1\,500$) transitions from (or to) each query token, which we believe is not a large number. We note that attacks that exploit query dependencies need such large number of queries to succeed (except in cases where query dependencies are unusually high).

We evaluate IHOP with $p_{free} = 0.25$ and use $n_{iters} = 10\,000$. Figure 9 shows the accuracy of IHOP vs. the number of iterations n_{iters} , for each of the five category-based keyword universes. As before, the accuracy in these plots is the percentage of query tokens (out of the $2n$ tokens) for which the adversary guesses the underlying keyword correctly. In each plot, the title shows the category, continuous and dashed lines denote $\rho = 100\,000$ and $\rho = 500\,000$, respectively, and each color denotes which defense is used and the quality of auxiliary information as explained above. We see that the attack accuracy and its evolution with n_{iters} varies significantly depending on the category (i.e., the underlying Markov model \mathbf{F}^{real} affects the algorithm’s convergence speed and accuracy). In most cases, PANCAKE significantly reduces the attack accuracy (blue vs. green lines, and orange vs. red lines). As expected, increasing the number of queries observed usually improves the attack. The exception to this is the green lines for some categories: this is because for these categories the imperfect auxiliary information ($\text{aux} \downarrow$) is misleading, so observing more queries misleads the attack even further. We see a slower convergence for IHOP compared to Fig. 2; this is because the attack has volume information in Fig. 2, which makes the linear coefficients d very helpful even when the fixed assignment in an iteration \mathbf{P}^\bullet contains many incorrect matchings. Our attack in this frequency-only setting (Fig. 9) relies mostly on the quadratic terms c , which are only truly helpful when the current assignment \mathbf{P} is already accurate. This makes the attack’s convergence slower.

We summarize the results of our experiment in Figure 10. Here, each box represents the accuracy of the attack against *all* categories (30 accuracy values for each category, 5 categories) in the last iteration of the attack ($n_{iters} = 10\,000$). This figure

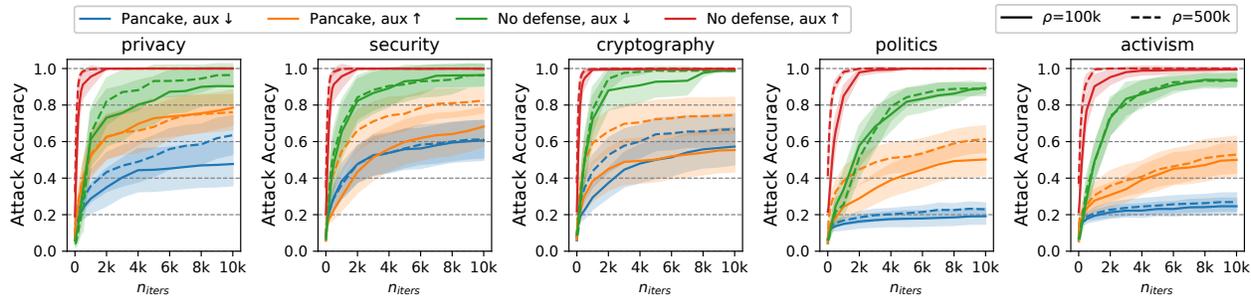


Figure 9: Accuracy of IHOP against PANCAKE with the number of iterations n_{iters} , for different category-based datasets and number of queries ρ . $\text{aux} \uparrow$ and $\text{aux} \downarrow$ denote high-quality and low-quality auxiliary information, respectively. (S3)

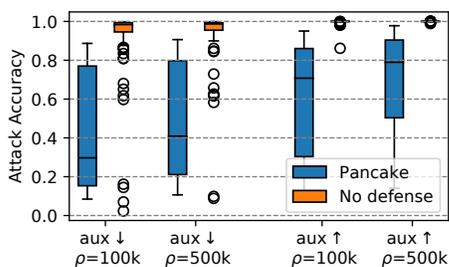


Figure 10: Summary of results of IHOP against PANCAKE with query dependencies. (S3)

shows more clearly the advantage of using PANCAKE over no defense when there are query dependencies. For $\rho = 100\,000$ queries, using PANCAKE decreases the average attack accuracy from 93.6% (no defense) to 41.9% (PANCAKE), when the auxiliary information is low-quality. With high-quality auxiliary information, the accuracy decreases from 99.7% to 60.5% thanks to PANCAKE. Although the decrease is somewhat significant, our results confirm that PANCAKE is vulnerable to query correlations, which differs from the findings in the preliminary analysis by Grubbs et al. [14]. The accuracy gap between the attack without defense and with PANCAKE is even smaller when the number of queries increases.

6.4 Discussion

Our experiments show that PANCAKE might not provide enough protection when there are query dependencies, which differs from the findings by Grubbs et al. [13]. The reason for this is that Grubbs et al. consider a case where the client’s dataset contains mappings of keywords to document identifiers, and of document identifiers to documents. The client queries first for a keyword, followed by a query for a document that contains that keyword. This creates dependencies between pairs of queries. Their dataset contains $\approx 532\,000$ entries, which results in over one million replicas. They see

that, even when the client issues 10 million queries, the joint distribution of consecutive accesses is almost flat. This is reasonable, since with one million replicas, 10 million queries might not be enough information for the adversary. Even with enough queries, carrying out IHOP in this case would be computationally prohibitive due to the size of the problem.

To summarize: we have provided the first piece of evidence that there are cases where PANCAKE does not provide sufficient protection against query correlations. PANCAKE’s protection against correlated queries thus depends on the keyword universe and dataset sizes, as well as the strength of the correlations. Developing statistical-based query recovery attacks able to deal with large keyword universe sizes and studying PANCAKE’s protection for different correlation levels is an interesting future research line.

7 Conclusion

We proposed IHOP, a new attack on SSE schemes that uses statistical (non-ground-truth) auxiliary information to recover the client’s queries. Our attack formulates query recovery as a quadratic optimization problem and uses a novel iteration heuristic that relies on efficient optimal linear solvers to find a suitable solution. IHOP is the first query recovery attack that can leverage both quadratic volume and frequency terms.

We evaluated our attack on real datasets against SSE schemes that exhibit typical leakage patterns, showing that it outperforms all other statistical-based query recovery attacks. In four out of the five datasets we consider, IHOP achieves almost perfect query recovery accuracy ($\approx 100\%$) in the full access-pattern leakage setting when the adversary receives a number of non-indexed documents equal to half the dataset size, while being more than one order of magnitude faster than previous attacks. When the adversary sees the access patterns of a subset of all the keywords only, IHOP widely outperforms all other attacks (92% accuracy in Lucene; previous best achieves 7%), while being significantly faster (12 minutes vs. over 5 hours). We verified IHOP consistently outperforms other attacks when frequency information is available, as well

as when efficient access-pattern obfuscation techniques are applied. Finally, we demonstrate that IHOP can exploit query dependencies by adapting our attack against the PANCAKE frequency-smoothing defense. Our results in a small dataset confirm that PANCAKE might not provide sufficient protection against query recoveries and urge for a more thorough analysis of PANCAKE under query dependencies.

Acknowledgments

We gratefully acknowledge the support of NSERC for grants RGPIN-05849, IRC-537591, and the Royal Bank of Canada for funding this research. This work benefited from the use of the CrySP RIPPLE Facility at the University of Waterloo.

References

- [1] Laura Blackstone, Seny Kamara, and Tarik Moataz. Revisiting leakage abuse attacks. In *Network and Distributed System Security Symposium (NDSS)*, page TBD, 2020.
- [2] Raphael Bost. Σοφος: Forward secure searchable encryption. In *ACM Conference on Computer and Communications Security (CCS)*, pages 1143–1154, 2016.
- [3] David Cash, Paul Grubbs, Jason Perry, and Thomas Ristenpart. Leakage-abuse attacks against searchable encryption. In *ACM Conference on Computer and Communications Security (CCS)*, pages 668–679, 2015.
- [4] David Cash, Stanislaw Jarecki, Charanjit Jutla, Hugo Krawczyk, Marcel-Cătălin Roşu, and Michael Steiner. Highly-scalable searchable symmetric encryption with support for boolean queries. In *International Cryptology Conference (CRYPTO)*, pages 353–373, 2013.
- [5] Yan-Cheng Chang and Michael Mitzenmacher. Privacy preserving keyword searches on remote encrypted data. In *International Conference on Applied Cryptography and Network Security (ACNS)*, pages 442–455, 2005.
- [6] Guoxing Chen, Ten-Hwang Lai, Michael K Reiter, and Yinqian Zhang. Differentially private access patterns for searchable symmetric encryption. In *IEEE International Conference on Computer Communications (INFOCOM)*, pages 810–818, 2018.
- [7] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 41–50, 1995.
- [8] Reza Curtmola, Juan Garay, Seny Kamara, and Rafail Ostrovsky. Searchable symmetric encryption: improved definitions and efficient constructions. *Journal of Computer Security*, 19(5):895–934, 2011.
- [9] Marc Damie, Florian Hahn, and Andreas Peter. A highly accurate query-recovery attack against searchable encryption using non-indexed documents. In *USENIX Security Symposium*, pages 143–160, 2021.
- [10] Ioannis Demertzis, Dimitrios Papadopoulos, Charalampos Papamanthou, and Saurabh Shintre. SEAL: Attack mitigation for encrypted databases via adjustable leakage. In *USENIX Security Symposium*, pages 2433–2450, 2020.
- [11] Michael L Fredman and Robert Endre Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM (JACM)*, 34(3):596–615, 1987.
- [12] Oded Goldreich and Rafail Ostrovsky. Software protection and simulation on oblivious RAMs. *Journal of the ACM (JACM)*, 43(3):431–473, 1996.
- [13] Paul Grubbs, Anurag Khandelwal, Marie-Sarah Lacharité, Lloyd Brown, Lucy Li, Rachit Agarwal, and Thomas Ristenpart. PANCAKE: Frequency smoothing for encrypted data stores. In *USENIX Security Symposium*, pages 2451–2468, 2020.
- [14] Paul Grubbs, Anurag Khandelwal, Marie-Sarah Lacharité, Lloyd Brown, Lucy Li, Rachit Agarwal, and Thomas Ristenpart. PANCAKE: Frequency smoothing for encrypted data stores. Cryptology ePrint Archive, Report 2020/1501, 2020. <https://eprint.iacr.org/2020/1501>.
- [15] Paul Grubbs, Marie-Sarah Lacharité, Brice Minaud, and Kenneth G Paterson. Pump up the volume: Practical database reconstruction from volume leakage on range queries. In *ACM Conference on Computer and Communications Security (CCS)*, pages 315–331, 2018.
- [16] Zichen Gui, Kenneth G Paterson, Sikhar Patranabis, and Bogdan Warinschi. SWiSSSE: System-wide security for searchable symmetric encryption. Cryptology ePrint Archive, Report 2020/1328, 2020. <https://eprint.iacr.org/2020/1328>.
- [17] Warren He, Devdatta Akhawe, Sumeet Jain, Elaine Shi, and Dawn Song. Shadowcrypt: Encrypted web applications for everyone. In *ACM Conference on Computer and Communications Security (CCS)*, pages 1028–1039, 2014.
- [18] Mohammad Saiful Islam, Mehmet Kuzu, and Murat Kantarcioglu. Access pattern disclosure on searchable encryption: Ramification, attack and mitigation. In *Network and Distributed System Security Symposium (NDSS)*, volume 20, page 12, 2012.

- [19] Seny Kamara, Charalampos Papamanthou, and Tom Roeder. Dynamic searchable symmetric encryption. In *ACM Conference on Computer and Communications Security (CCS)*, pages 965–976, 2012.
- [20] Georgios Kellaris, George Kollios, Kobbi Nissim, and Adam O’neill. Generic attacks on secure outsourced databases. In *ACM Conference on Computer and Communications Security (CCS)*, pages 1329–1340, 2016.
- [21] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.
- [23] Kaoru Kurosawa. Garbled searchable symmetric encryption. In *International Conference on Financial Cryptography and Data Security (FC)*, pages 234–251, 2014.
- [24] Billy Lau, Simon Chung, Chengyu Song, Yeongjin Jang, Wenke Lee, and Alexandra Boldyreva. Mimesis aegis: A mimicry privacy shield—a system’s approach to data privacy on public cloud. In *USENIX Security Symposium*, pages 33–48, 2014.
- [25] Eugene L Lawler. The quadratic assignment problem. *Management Science*, 9(4):586–599, 1963.
- [26] Chang Liu, Liehuang Zhu, Mingzhong Wang, and Yu-An Tan. Search pattern leakage in searchable encryption: Attacks and new construction. *Information Sciences*, 265:176–188, 2014.
- [27] Yao Lu, Kaizhu Huang, and Cheng-Lin Liu. A fast projected fixed-point algorithm for large graph matching. *Pattern Recognition*, 60:971–982, 2016.
- [28] Muhammad Naveed, Manoj Prabhakaran, and Carl A Gunter. Dynamic searchable encryption via blind storage. In *IEEE Symposium on Security and Privacy (SP)*, pages 639–654, 2014.
- [29] Jianting Ning, Xinyi Huang, Geong Sen Poh, Jiaming Yuan, Yingjiu Li, Jian Weng, and Robert H Deng. Leap: Leakage-abuse attack on efficiently deployable, efficiently searchable encryption with partially known dataset. In *ACM Conference on Computer and Communications Security (CCS)*, pages 2307–2320, 2021.
- [30] Wakaha Ogata, Keita Koiwa, Akira Kanaoka, and Shin’ichiro Matsuo. Toward practical searchable symmetric encryption. In *International Workshop on Security (IWSEC)*, pages 151–167, 2013.
- [31] Simon Oya and Florian Kerschbaum. Hiding the access pattern is not enough: Exploiting search pattern leakage in searchable encryption. In *USENIX Security Symposium*, page TBD, 2021.
- [32] Sarvar Patel, Giuseppe Persiano, Kevin Yeo, and Moti Yung. Mitigating leakage in secure cloud-hosted data structures: Volume-hiding for multi-maps via hashing. In *ACM Conference on Computer and Communications Security (CCS)*, pages 79–93, 2019.
- [33] Rishabh Poddar, Stephanie Wang, Jianan Lu, and Raluca Ada Popa. Practical volume-based attacks on encrypted databases. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 354–369, 2020.
- [34] David Pouliot and Charles V Wright. The shadow nemesis: Inference attacks on efficiently deployable, efficiently searchable encryption. In *ACM Conference on Computer and Communications Security (CCS)*, pages 1341–1352, 2016.
- [35] Sartaj Sahni and Teofilo Gonzalez. P-complete approximation problems. *Journal of the ACM (JACM)*, 23(3):555–565, 1976.
- [36] Zhiwei Shang, Simon Oya, Andreas Peter, and Florian Kerschbaum. Obfuscated access and search patterns in searchable encryption. In *Network and Distributed System Security Symposium (NDSS)*, page TBD, 2021.
- [37] Dawn Xiaoding Song, David Wagner, and Adrian Perrig. Practical techniques for searches on encrypted data. In *IEEE Symposium on Security and Privacy (SP)*, pages 44–55, 2000.
- [38] Emil Stefanov, Charalampos Papamanthou, and Elaine Shi. Practical dynamic searchable encryption with small leakage. In *Network and Distributed System Security Symposium (NDSS)*, pages 72–75, 2014.
- [39] Shinji Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):695–703, 1988.
- [40] Cornelis J Van Rijsbergen, Stephen Edward Robertson, and Martin F Porter. *New models in probabilistic information retrieval*, volume 5587. British Library Research and Development Department London, 1980.
- [41] Robert West. Calibration of google trends time series. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2257–2260, 2020.
- [42] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. A path following algorithm for the graph matching

problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2227–2242, 2008.

A Adapting IHOP and SAP against CLRZ and OSSE

We adapt IHOP and SAP against CLRZ following the approach by Shang et al. [36, Appendix D]. Namely, the adversary knows the TPR and FPR of CLRZ and computes the $n \times n$ matrix of *expected* keyword volumes $\hat{\mathbf{V}}$ after the defense is applied. Recall that $\tilde{\mathbf{V}}_{i,i'}$ is an estimation (from the auxiliary information) of the probability that a document has both keywords k_i and $k_{i'}$. Let $\tilde{\mathbf{V}}_{i,i'}^{\text{NOT}}$ be an estimation of the probability that a document has *neither* keywords k_i nor $k_{i'}$ (in our experiments, the adversary computes this from the auxiliary data set). Then, the i, i' ’th entry of $\hat{\mathbf{V}}$ is [36, Appendix D]

$$\hat{\mathbf{V}}_{i,i'} = \begin{cases} i \neq i' : & \text{TPR}^2 \cdot \tilde{\mathbf{V}}_{i,i'} + \text{FPR}^2 \cdot \tilde{\mathbf{V}}_{i,i'}^{\text{NOT}} \\ & + \text{TPR} \cdot \text{FPR} \cdot (1 - \tilde{\mathbf{V}}_{i,i'} - \tilde{\mathbf{V}}_{i,i'}^{\text{NOT}}), \\ i = i' : & \text{TPR} \cdot \tilde{\mathbf{V}}_{i,i'} + \text{FPR} \cdot \tilde{\mathbf{V}}_{i,i'}^{\text{NOT}}. \end{cases} \quad (9)$$

To adapt the different attacks against CLRZ, the adversary simply uses $\hat{\mathbf{V}}$ instead of $\tilde{\mathbf{V}}$ in the attack’s coefficients.

We explain how we adapt the attacks against OSSE. Recall that, in OSSE, the adversary observes access patterns (with each entry obfuscated with TPR and FPR), but does not know whether or not two access patterns correspond to the same keyword, since each access pattern has been generated with fresh randomness. Let m be the number of distinct queried keywords in OSSE (the number of distinct observed access patterns could be up to ρ ; i.e., the number of queries). Following Shang et al. [36], the adversary first clusters the observed access patterns into m groups (we assume the adversary knows m for simplicity [36]). Let C_j be the j ’th cluster, for $j \in [m]$. To build the off-diagonal j, j' ’th entries ($j \neq j'$) of the matrix of observed volumes \mathbf{V} , the adversary computes the *average* number of documents in common between one access pattern from C_j and one from $C_{j'}$, i.e.,

$$\mathbf{V}_{j,j'} = \frac{1}{|C_j| \cdot |C_{j'}|} \sum_{\mathbf{a} \in C_j} \sum_{\mathbf{a}' \in C_{j'}} \frac{\mathbf{a}^T \mathbf{a}'}{N_d}.$$

The diagonal entries are $\mathbf{V}_{j,j} = \frac{1}{|C_j|} \sum_{\mathbf{a} \in C_j} \frac{\mathbf{a}^T \mathbf{a}}{N_d}$. The adversary uses this observation matrix \mathbf{V} and the auxiliary matrix $\hat{\mathbf{V}}$ above (9) to run the attacks against OSSE.

B Adapting IHOP against PANCAKE

We provide details into how we adapt IHOP against PANCAKE. As we mention in the main text, the client chooses keywords for her queries following a Markov process characterized by the $n \times n$ matrix \mathbf{F}^{real} , with stationary profile \mathbf{f}^{real} . Every time

Algorithm 3 Compute \mathbf{F} in PANCAKE

```

1: procedure COMPUTEFREQOBS( $n, T_1, T_2, \dots, T_\rho$ )
2:    $\mathbf{F} \leftarrow \mathbf{0}_{2n \times 2n}$ 
3:   for  $i \leftarrow 1 \dots \rho - 1$  do
4:     for  $\tau_j \in T_i, \tau_{j'} \in T_{i+1}$  do
5:        $\mathbf{F}_{j',j} \leftarrow \mathbf{F}_{j',j} + 1$ 
6:   Normalize columns of  $\mathbf{F}$ .
7:   return  $\mathbf{F}$ 

```

the client queries the server, PANCAKE creates three query slots, selects three keywords, and sends the corresponding query tokens to the server (see Section 6.1). Each of these keywords can either be *dummy* ($k \sim \mathbf{f}^{\text{dum}}$), *real fake* ($k \sim \mathbf{f}^{\text{real}}$), or an actual real query (sampled from \mathbf{F}^{real} , according to the previous real query). Let T_1, T_2, \dots, T_ρ be the set of query token triplets observed by the adversary. The adversary builds the matrix of observed token frequencies \mathbf{F} by counting all token transitions in consecutive triplets (Algorithm 3).

Then, the adversary builds the $2n \times 2n$ expected matrix of *replica* frequencies $\hat{\mathbf{F}}$ given the auxiliary information $\tilde{\mathbf{F}}$. First, the adversary computes $\tilde{\mathbf{f}}$ from $\tilde{\mathbf{F}}$, and gets the number of replicas of each keyword and the dummy keyword distribution $\tilde{\mathbf{f}}^{\text{dum}}$ following PANCAKE’s specifications. Let $\tilde{R}(i)$ be the number of replicas for keyword k_i and $\mathbf{v} : [2n] \rightarrow [n]$ be a mapping from replicas to keywords, both computed from $\tilde{\mathbf{f}}$.

If we generate keyword queries following PANCAKE’s specifications and using $\tilde{\mathbf{F}}, \tilde{\mathbf{f}}$ and $\tilde{\mathbf{f}}^{\text{dum}}$, for every two keywords k and k' in consecutive triplets, one of the following events happens:

- A. k' was sampled from $\tilde{\mathbf{f}}^{\text{dum}}$.
- B. k was sampled from $\tilde{\mathbf{f}}^{\text{dum}}$ but k' was not.
- C. Neither k nor k' were sampled from $\tilde{\mathbf{f}}^{\text{dum}}$, but at least one is a “real fake” query sampled from $\tilde{\mathbf{f}}$.
- D. Both k and k' are real queries.

Following PANCAKE specifications, one can see that $\Pr(A) = 0.5$ and $\Pr(B) = 0.25$. We computed the probabilities of events C and D empirically, since they are constants that just depend on PANCAKE’s specifications: $\Pr(C) \approx 0.145$ and $\Pr(D) \approx 0.105$. Note that k and k' are independent except in event D . If both k and k' are keywords for real queries (event D), then k' was generated right after k with probability 0.81, two queries after k with probability 0.17, and three queries after k with probability 0.02 (we also determined these probabilities empirically; they are constants that only depend on PANCAKE’s protocol). Summarizing, with probability 0.5 we have that $k' \sim \tilde{\mathbf{f}}^{\text{dum}}$, with probability $0.25 + 0.145 = 0.395$ we have $k' \sim \tilde{\mathbf{f}}$, and otherwise k' depends on k . Putting this together, the transition probabilities between two keywords in consecutive triplets are:

$$\mathbf{G} = 0.105 \cdot (0.81\tilde{\mathbf{F}} + 0.17\tilde{\mathbf{F}}^2 + 0.02\tilde{\mathbf{F}}^3) + 0.395 \cdot \tilde{\mathbf{f}}\mathbf{1}_n^T + 0.5 \cdot \tilde{\mathbf{f}}^{\text{dum}}\mathbf{1}_n^T.$$

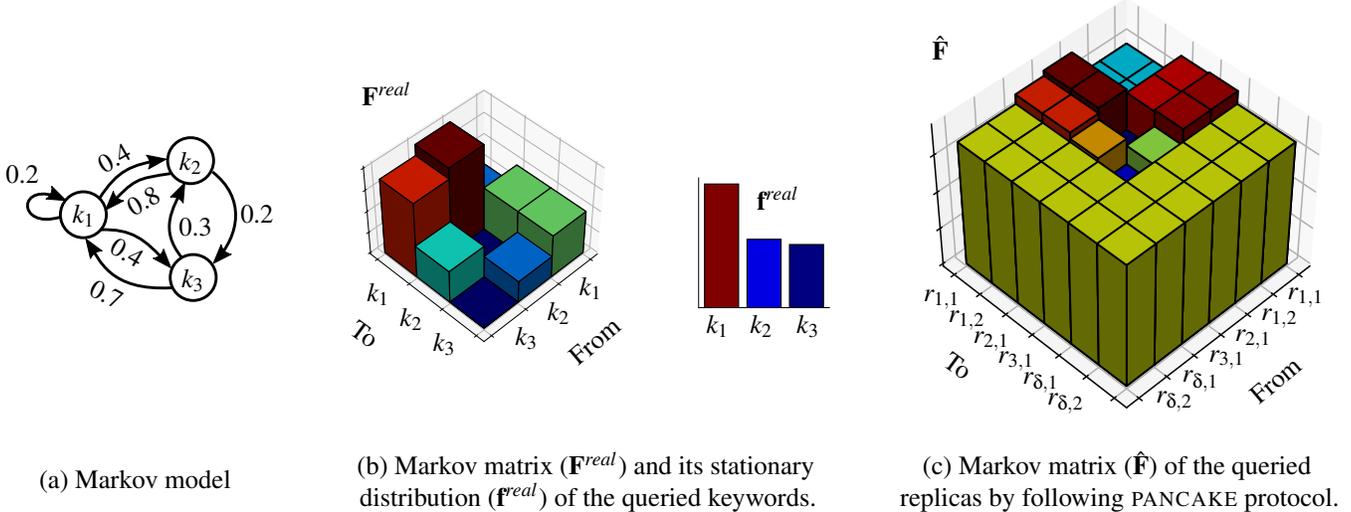


Figure 11: Example of PANCAKE under correlated queries.

Matrix $\hat{\mathbf{F}}$ follows this formula, expanded to the space of replicas:

$$\hat{\mathbf{F}}_{j,j'} = \mathbf{G}_{v(j),v(j')} / \tilde{R}(v(j)). \quad (10)$$

Figure 11 shows an example of a Markov model with three keywords (Fig. 11a), the Markov matrix and stationary profile (Fig. 11b), and the expected transition between replicas $\hat{\mathbf{F}}$ (10) (Fig. 11c). In the plot, we used $\tilde{\mathbf{F}} = \mathbf{F}^{real}$ to compute $\hat{\mathbf{F}}$, for illustration purposes. The stationary profile of $\hat{\mathbf{F}}$ is uniform, but we can see that $\hat{\mathbf{F}}$ is not. The matrix of observed frequencies \mathbf{F} is a (noisy) version of $\hat{\mathbf{F}}$, with randomly permuted rows and columns. We use IHOP to estimate this permutation by setting the coefficients of IHOP as in (8) and (7), but using $\hat{\mathbf{F}}$ instead of $\tilde{\mathbf{F}}$. The result of this is a mapping \mathbf{P} from tokens to replicas, which we map to keywords using $v(\cdot)$.

C Dataset Generation

Table 2 shows the dataset names, the type and number of documents they contains, and the source URL that we used to download them. We download the datasets from the source we show in the table and parse them so that each document is a string (an email, news article, or movie plot summary). Following related work [9], for Enron dataset we only take documents in the `_sent_mail` folder; for Lucene, we remove the signature at the end of each email that begins with "To unsubscribe", since this message appears in all emails. Then, we follow a series of steps to extract the keywords of each document:

1. We extract each word in the document using the regular expression `(regex) \w+`.
2. We convert all words to lower-case and ignore words that contain non-alpha characters [33].

3. We ignore English stopwords, and words whose length is not between 3 and 20 characters [33].
4. We apply Porter stemmer [40] to the remaining words, and keep the 3000 most popular stems of each dataset. This stemming process is the most popular approach to extract keywords in related work [1, 3, 9, 18, 29, 33].

These stems are the keywords in our evaluation. We save, for each dataset, the different words that yielded each of the 3000 stems. Then, for each of those words, we download their query frequencies for each week of 2020 from Google Trends.⁹ We use the `gtab` package [41] to fix normalization issues of Google Trends. The query frequency of a particular stem is the sum of frequencies for each of its keywords. For example, "time" is a popular stem in Enron dataset. The words that resulted in this stem are "time", "timed", "timely", etc. We downloaded the query frequencies for each of those words, and assumed that queries for each of those words trigger a match in documents that contain the stem "time".

Figure 12 shows the volume of each keyword in the datasets. Recall that the volume of a keyword is the percentage of documents that contain such keyword. We sort the keywords of each dataset in decreasing volume order in the plot. In email datasets (Enron and Lucene), we see that keywords have lower volumes than in other datasets, which indicates that each email uses specific terminology that is not very common among other emails. Datasets that consist of news articles (News and NYTimes) lie in the opposite extreme: their documents use a similar vocabulary, which yields keywords with overall high volumes. The volume distribution of the dataset that consists of movie summaries (Movies) is somewhere in between these two extremes. This plot shows that the datasets we consider are statistically varied. We note that the

⁹<https://trends.google.com/trends>

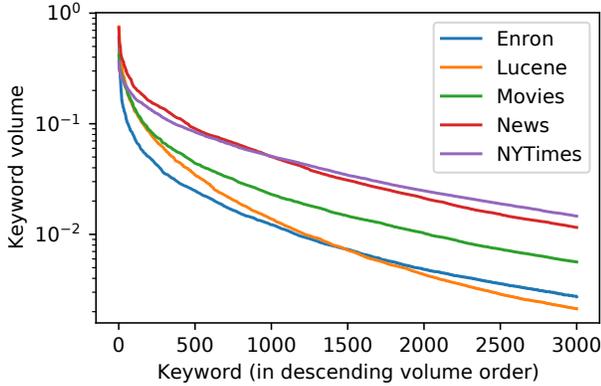


Figure 12: Keyword volume distribution for each of the datasets we consider.

volume distribution is not necessarily correlated with attack accuracy. There are many variables that affect attack accuracy (e.g., keyword volume uniqueness, keyword co-occurrence uniqueness, higher-order dependencies between keywords), and we cannot capture all of them in a single plot.

D Wikipedia Dataset Generation

We explain how we generate the keyword universes and transition matrices \mathbf{F}^{real} for our experiments with query correlations (Section 6.3). As we mention in the main text, we consider a scenario where each keyword-document pair represents a particular topic or webpage, and we use the Wikipedia clickstream dataset to build a realistic transition matrix \mathbf{F}^{real} between keywords. Thus, we refer to keywords as *pages* in this appendix.

We use PetScan¹⁰ to retrieve all Wikipedia pages under the categories ‘*privacy*’, ‘*security*’, ‘*cryptography*’, ‘*politics*’, and ‘*activism*’, including pages in subcategories up to a depth of two. This yields 5573 pages. We download the number of times Wikipedia users transitioned between those pages in 2020 by querying the Wikipedia Clickstream dataset.¹¹ Let G_{big} be a graph where the nodes are these pages, and the edges between two nodes are the number of times users transitioned between the pages represented by such nodes.

We explain how we build a keyword universe Δ_k of size $n = 500$ centered around a category. We start with a subgraph $G_{cat} \subset G_{big}$ with nodes (pages) from that category only. We remove all nodes with a degree ≤ 1 and we keep removing nodes with the smallest degree in G_{cat} until the subgraph has a size smaller than n . If the subgraph already had less nodes than n , we add nodes from G_{big} (the ones that share

more edges with G_{cat} are added first) until G_{cat} has n nodes. Then, let k_{big} be the node in $G_{big} \setminus G_{cat}$ that has more edges connecting to G_{cat} (and let $s(k_{big})$ be this number of edges), and let k_{cat} be the node in G_{cat} with smallest degree (and let $s(k_{cat})$ be this degree). If $s(k_{cat}) \leq 1$ or if $s(k_{cat}) < s(k_{big}) - 2$, we remove k_{cat} from G_{cat} and replace it with k_{big} . Otherwise, we finish the building process, and the remaining n nodes in G_{cat} are the keyword universe Δ_k for the category.

Finally, we build \mathbf{F}_{JUL}^{DEC} from Δ_k as follows (the process for \mathbf{F}_{JAN}^{JUN} is the same, with data from different months). We get all transitions between pages in Δ_k from July to December (2020), and also get the number of times users accessed pages in Δ_k from *other* sources (namely, transitions from pages named ‘*other-empty*’, ‘*other-external*’, ‘*other-internal*’, ‘*other-search*’, ‘*other-other*’ in the Clickstream dataset). We use these transitions from other sources to compute the query probability of the pages in Δ_k when the user starts a browsing session (we denote these probabilities by \mathbf{p}_{other}).

The columns of \mathbf{F}_{JUL}^{DEC} are transition profiles, i.e., the i th column is a vector that represents the probability of querying for a particular page after querying for k_i . To compute this profile for a page $k_i \in \Delta_k$, we count the transitions from k_i to all other $k_{i'} \in \Delta_k$ and normalize this so that it adds up to one. We also consider that there is a probability 0.05 of *restarting* the browsing session, i.e., with probability 0.05 the user simply chooses a new page by following \mathbf{p}_{other} . This ensures the Markov process is irreducible and aperiodic. If k_i is a *sink page*, i.e., a page the user can transition to but that does not have outgoing transitions to other pages in Δ_k , we simply consider that the user restarts the browsing session after that page, so the transition profile is just \mathbf{p}_{other} .

E Evaluation Results in All Datasets

For completeness, below we show the results of all our experiments in the five datasets. Figures 13, 14, 15, 16, and 17 below correspond to the experiments shown in Figures 2, 4, 5, 6, and 7, respectively. In all these figures, we can see that the results are qualitatively similar: the trends of IHOP with n_{iters} and p_{free} are the same for all datasets in Figure 13, and IHOP outperforms all other attacks in the remaining experiments.

¹⁰<https://petscan.wmflabs.org>

¹¹<https://dumps.wikimedia.org/other/clickstream/>

Dataset	Size	Content	Source
Enron	30033	Emails	https://www.cs.cmu.edu/~./enron/
Lucene	63597	Emails	https://mail-archives.apache.org/mod_mbox/lucene-java-user/
Movies	42304	Movie plots	http://www.cs.cmu.edu/~ark/personas/
News	49931	Articles	https://www.kaggle.com/snapcrack/all-the-news/version/4
NYTimes	299607	Articles	https://archive.ics.uci.edu/ml/datasets/bag+of+words

Table 2: Datasets that we consider.

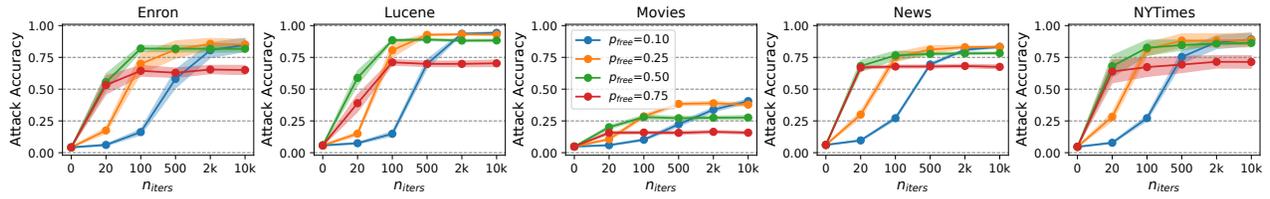


Figure 13: Evolution of IHOP's accuracy with the number of iterations and different p_{free} values. (S1)

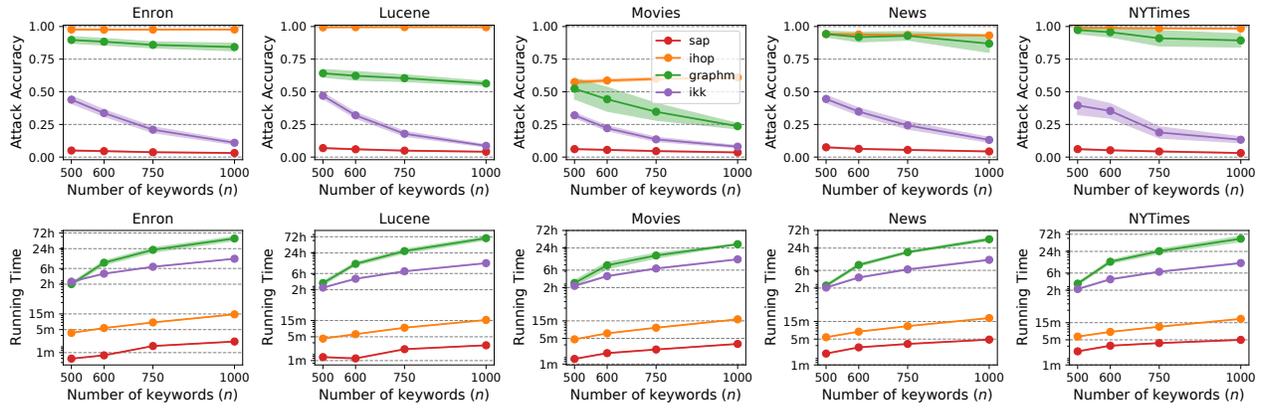


Figure 14: Attack comparison vs. number of keywords (n) (S1)

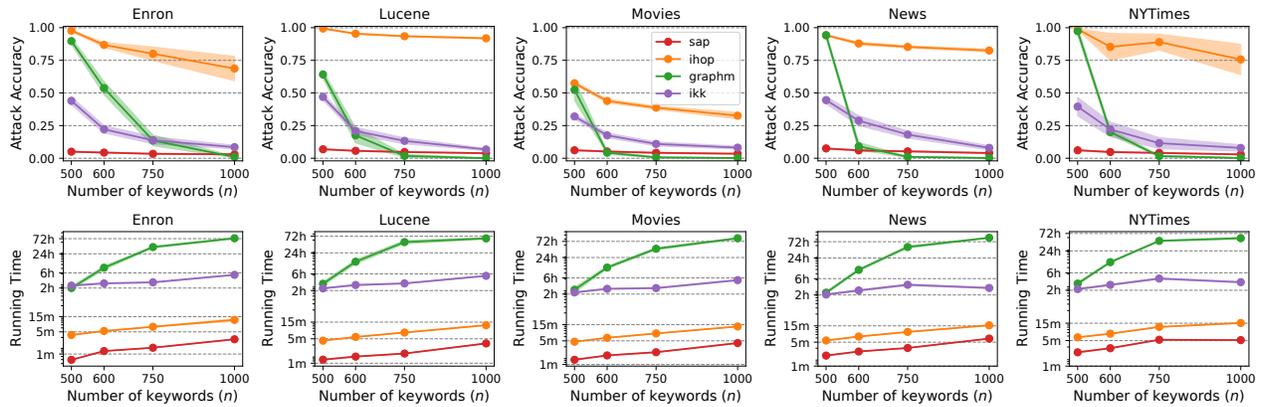


Figure 15: Attack comparison vs. number of keywords (n) when the adversary sees $m = 500$ distinct tokens (S2)

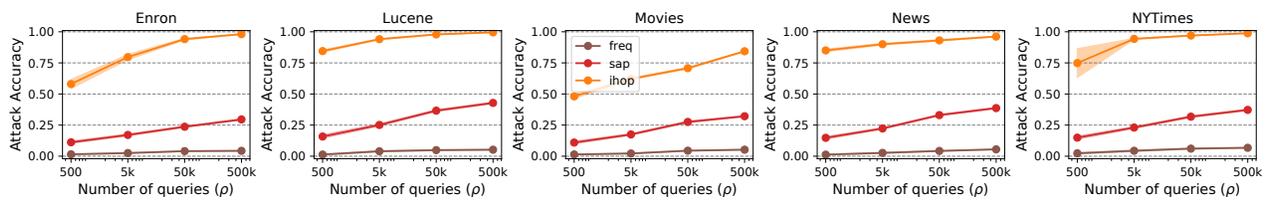


Figure 16: Comparison of attacks that use frequency leakage, with $n = 3000$ (S2)

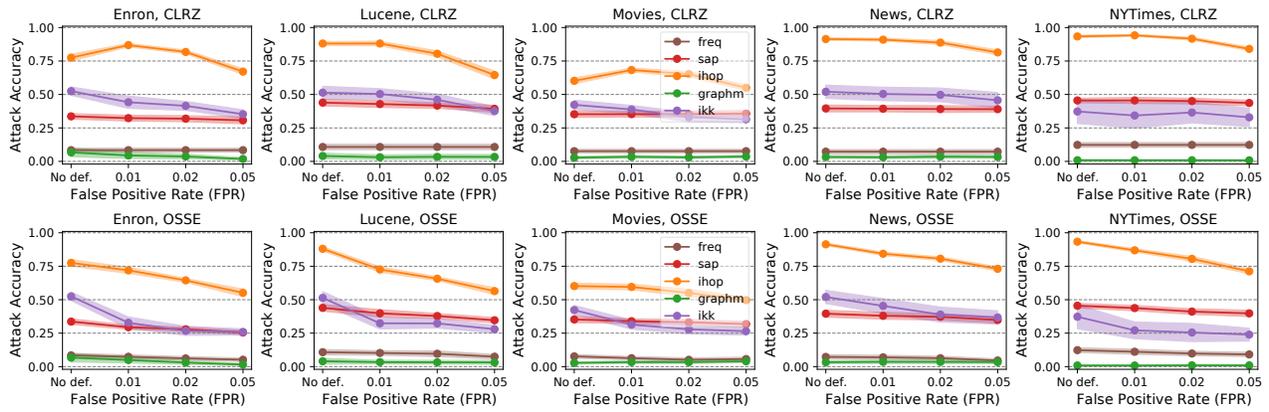


Figure 17: Attack comparison vs. CLRZ and OSSE ($n = 500$, $\rho = 500$) (S2)