頻度分析をk-匿名性で緩和する検索可能共通鍵暗号

山岡 裕司† 牛田 芽生恵† 伊藤 孝一†

†株式会社富士通研究所
〒211-8588 川崎市中原区上小田中 4-1-1 (研 S403)
{yamaoka.yuji, ushida.mebae, ito.kouichi}@jp.fujitsu.com

あらまし 検索可能暗号は，クラウドサービスなどの外部サーバに，機密情報を検索できる形で暗号化して格納し，検索内容をサーバ管理者にも知られないようにする技術である．確定的な共通鍵暗号による暗号文をクエリーとする検索可能共通鍵暗号は，検索処理量が少ないという特長がある．しかし，十分に検索がおこなわれた状態での頻度分析への対策がなく，検索内容や検索対象文書内容がサーバ管理者に知られる恐れが大きい．本稿では，索引と検索キーワードに共通のk-匿名性を持たせることにより，頻度分析耐性と検索処理量のトレードオフを選択できる新しい方式を提案する．$k \geq 2$ とすることで，従来より頻度分析耐性を向上させることができる．

Searchable Symmetric Encryption with Frequency Analysis Resistance based on k-Anonymity

Yuji Yamaoka† Mebae Ushida† Kouichi Itoh†

†FUJITSU LABORATORIES LTD.
1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki 211-8588, Japan (LS403)
{yamaoka.yuji, ushida.mebae, ito.kouichi}@jp.fujitsu.com

Abstract Searchable encryption is technology that lets users encrypt and store their confidential information in outside servers in a searchable manner, and that prevents the server administrators from knowing search contents. Searchable symmetric encryption has the advantage of efficiency in the search. However, it does not take measures against the frequency analysis in the state that the searches has been performed enough. This paper proposes a new method that the users make the search keywords satisfy k-anonymity to be able to select the trade-off between the frequency analysis resistance and the search workload. The method can improve the frequency analysis resistance with $k \geq 2$.

1 はじめに

昨今，文書管理のアウトソース化が有望な選択肢になっている．クラウドコンピューティング技術の発展などにより，ストレージサービスである外部サーバに文書を置くことで，文書管理の運用コストを下げられる場合がある．

しかし，機密情報を含む文書をサーバに置く場合は，暗号化が必要になる．サーバの管理者が不正をしたり，ハッカーが管理者権限を奪取したりすると，平文の文書からは機密情報が漏洩してしまうためである．

一方，文書管理における利便性の観点からは，キーワードによる文書の検索ができることが必要である．

そのため，文書と検索キーワードをサーバに

対し暗号化したままサーバで検索できるようにする技術である，検索可能暗号が多数研究されている [3].

検索可能暗号の主な研究課題として，頻度分析耐性と検索処理量の両立がある．頻度分析耐性とは，暗号化されている検索クエリーの頻度や，検索結果の頻度から，平文を推定する攻撃に対する耐性のことである．検索処理量とは，検索時に必要なサーバの計算量である．検索処理量は，検索キーワードを含む文書数 m に対し $O(m)$ であるのが理想であり，全文書数 n に対し $\Omega(n)$ であるのは非実用的である．

SSE（Searchable Symmetric Encryption，検索可能共通鍵暗号）は，共通鍵暗号による確定的な暗号化による暗号文をクエリーとする方式 [4, 7, 6, 10] で，検索処理量を理想である $O(m)$ にできるのが特長である．確定的な暗号化とは，同じ平文をいつも同じ暗号文にする方式である．基本的な SSE では，検索キーワードの暗号文をそのままクエリーとする．そのため，暗号文の頻度がキーワードの頻度と一致し，頻度分析耐性が低いという課題がある．また，これまでの研究では，索引への頻度分析に対する耐性を高める提案 [4, 7, 6, 10] がされてきたが，十分に検索がおこなわれた状態での頻度分析は考慮されていなかった [9]．たとえば，文書数が膨大な場合は特に，検索結果が同じになったクエリー同士は同じ検索キーワードであるとの推定や，検索結果の頻度分析によりその平文が推定できてしまう恐れがある．

1.1 貢献と構成

そこで本稿では，パラメーター $k \geq 2$ に対し，検索処理量が $O(km)$ 程度に悪化する代わりに，k を大きくするほど頻度分析耐性が向上する方式を提案する．本稿の主な貢献は次の通りである．

- 頻度分析耐性と検索処理量のトレードオフを選択できる，新しい SSE 方式を提案する．頻度分析耐性は，十分に検索がおこなわれた状態での耐性を含む．

提案方式を簡単に説明すると，転置索引と検索キーワードに共通の k-匿名性を持たせる方式である．全キーワードを，k 個以上のキーワードで構成されるグループのいずれかに分類し，転置索引と検索ではグループ ID を使うようにする．

本稿の以降の構成は次の通りである．第 2 節では，本稿での主な用語と表記を説明する．第 3 節では，検索可能暗号への主な攻撃である頻度分析と選択平文攻撃について説明する．第 4 節では，関連研究について説明する．第 5 節では，提案方式について説明する．第 6 節でまとめる．

2 準備

本稿での主な用語と表記を説明する．

2.1 文書，キーワード

利用者にあたるクライアントは，サーバに文書 $M_1, M_2, ..., M_n$ の n 文書を置く（置いている）とする．現実的には文書の追加や削除がおこなわれるが，本稿の提案方式はそれらを妨げるものではないため，簡単のためすでに n 文書が置かれた状態を考える．

キーワードの全体集合 $W = \{w_1, w_2, ..., w_v\}$ は有限で変化しないものとする．簡単のため，$\forall i, M_i \subseteq W$ とする．現実的には，たとえばキーワードは単語とし，文書は形態素解析などにより単語集合とみなすことで，本モデルを適用することができる．

文書 ID を $I(M)$ と表す．すなわち，次が成り立つ．

$$\forall i, I(M_i) = i$$

2.2 確定的／確率的暗号

平文 p について，確定的な暗号化による暗号文を $E(p)$ と表す．同一の平文 p, p' に対し，暗号文 $E(p), E(p')$ も同一になる．なお，本稿での確定的な暗号化は復号できる必要がない．

また，確率的な暗号化による暗号文を $E_r(p)$ と表す．同一の平文 p, p' に対し，暗号文 $E_r(p)$, $E_r(p')$ は同一にならない．暗号文 c について，復号は $D(c)$ と表し，$D(E_r(p)) = p$ である．

2.3 転置索引，基本的な SSE

転置索引は，キーワード w' から，それを含む文書 ID 集合 $\{I(M)|w' \in M\}$ を高速に取得できるようにした構造である．

転置索引のキーワードを暗号化したものを，本稿では基本的な SSE と呼ぶ．図 1 に基本的な SSE の構成を示す．文書表は，文書 ID から，それに対応した文書の暗号文を高速に取得できるようにした構造である．基本的な SSE での検索処理は次のようになる．

1. クライアントは，検索キーワード w' を暗号化し，検索クエリー $E(w')$ としてサーバに送る．
2. サーバは，転置索引からクエリー $E(w')$ に紐付けられた全文書 ID $\{I(M)|w' \in M\}$ を取得し，さらに文書表からそれら文書 ID に該当する各文書 $\{E_r(M)|w' \in M\}$ を取得し，検索結果 R としてクライアントに送る．
3. クライアントは，R の各文書を復号し，最終的な検索結果 $\{M|w' \in M\}$ を得る．

2.4 k-匿名性

k-匿名性とは，匿名化技術の分野において提案された匿名性の指標であり，各個体が同じデータを持つ k 個以上の個体で構成されるグループに属しているという性質である [8]．k-匿名性の意義から，$k \geq 2$ とする．k-匿名性を満たす場合，どの個体も他に同じデータを持つ個体が $k-1$ 個以上あるため，一意に識別されなくなる．

3 検索可能暗号への攻撃

検索可能暗号への主な攻撃である頻度分析と選択平文攻撃について説明する．

頻度分析とは，主に値の頻度の情報を使い，暗号文の元の平文を推定する攻撃である．元々は単一換字式暗号の代表的な解読手法であり，たとえば暗号文で一番頻度が高い文字の元の文字は「e」ではないか，などと推定していく手法である．

検索可能暗号での頻度分析はキーワード単位でおこなわれ，攻撃者はサーバ管理者とする．キーワードが暗号化されていても，多くの文書に含まれていることや，ある時期だけそのキーワードによる検索が急増したことなどがわかると，攻撃者はその平文を推定できる可能性がある．

本稿では，頻度分析の内容を，対応分析と同一性分析という 2 つに細分化する．

一方，選択平文攻撃とは，攻撃者が任意の平文の暗号文を得られるとし，それにより暗号文の元の平文を推定する攻撃である．

以下，これらについて説明する．

3.1 対応分析

本稿での対応分析は，クエリーから，検索キーワードを含む文書 ID を特定することである．すなわち，検索キーワードを w'，それに対応するクエリーを q とすると，q から $\{I(M)|w' \in M\}$ を得ることである．

基本的な SSE では，$q = E(w')$ であり，検索処理そのものが $E(w')$ から $\{I(M)|w' \in M\}$ を得ることであるため，必ず対応分析できる．

対応分析ができる場合，文書における検索キーワードの出現頻度がわかる．さらにクエリーされた検索キーワードが何かまでわかると，どの文書にそのキーワードが含まれているかもわかり，文書に含まれるキーワードが分かってくると文書全体の内容まで推定できる．そのため，クエリーされた検索キーワードが頻度からの推定やその他の理由で攻撃者にわかってしまうことを想定すると，対応分析への耐性は重要である．

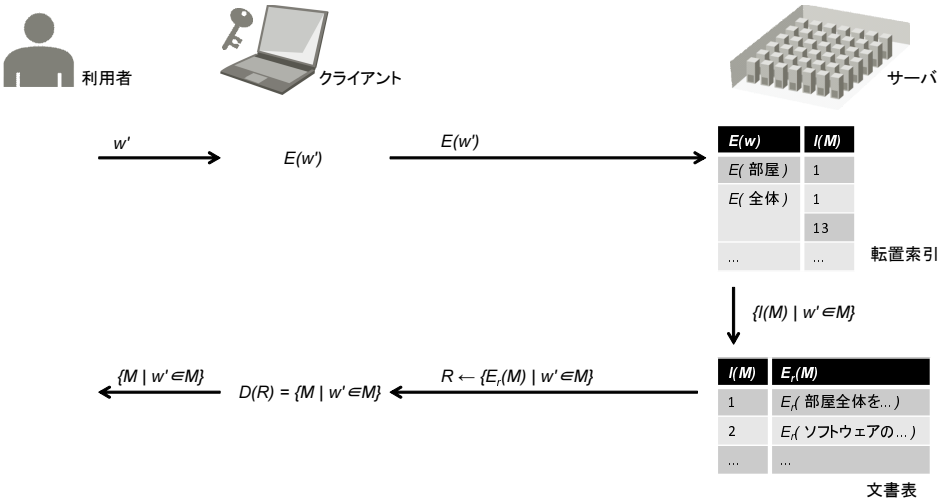

図 1: 基本的な SSE の構成

3.2 同一性分析

本稿での同一性分析は，あるクエリーと別のクエリーが同一キーワードであるかどうか特定することである．

基本的な SSE では，クエリーの同一性でキーワードの同一性が判断できるので，必ず同一性分析できる．また，文書数が膨大で，特に対応分析もできる場合，検索結果が同じになったクエリー同士は同じ検索キーワードである可能性が高い，つまり同一性分析しやすいと考えられる．

同一性分析ができる場合，検索キーワードの検索頻度がわかる．検索頻度がわかると，そのキーワードを推定できる可能性が高まる．たとえば，X 社の不祥事がニュースで大きく取り上げられた後に，あるキーワードでの検索が急増したことがわかった場合，そのキーワードは「X 社」かそれに関係するものと推定できるかもしれない．上述のように対応分析ができると同一性分析もしやすいため，同一性分析への耐性を持たせる観点でも対応分析への耐性は重要である．

3.3 選択平文攻撃

本稿での選択平文攻撃は，頻度分析の効果を高めるために，任意の文書を利用者に暗号化させ，その暗号文がどれかわかる形でサーバに置かせる攻撃である．

これができる状況として，たとえば次が考えられる．

- 利用者は，公開しているあるメールアドレス宛のメールは，読み次第サーバに置くようにしている．
- 攻撃者は，そのメールアドレス宛に任意の文面でメールを送り，サーバに暗号文が追加で置かれることを監視する．メール送信後，ある程度の期間内にサーバに追加された暗号文が 1 つだけな場合，その暗号文が送ったメールである蓋然性が高いと考える．メール送信とサーバ監視を繰り返すことで，その確信を強めることができる．

この状況では，攻撃者は任意の文書（メール）をそれがどれかわかる形でサーバに置かせることができる．

選択平文攻撃ができ，対応分析もできる場合，攻撃者は任意のクエリーの検索キーワードがわかる．たとえば，利用者に $M' = \{w'\}$ という文書を置かせ，その後 M' が検索結果になるようなクエリーを待つことで，そのクエリーの検索キーワードが w' であることがわかる．さらにこの場合，対応分析できるため，w' を含む全ての文書もわかる．選択平文攻撃は上記のようにサーバの利用の仕方によっては可能となるものなので技術的な対策が難しい．よって，この観点でも対応分析への耐性が重要である．

4　関連研究

関連研究として，検索可能暗号の安全性モデルと方式について説明する．

4.1　安全性モデル

検索可能暗号の安全性モデルとして最も使用されているものに，IND-CKA1 および IND-CKA2[4] がある．これらは，検索がおこなわれていない状態で索引から得られる情報を最小化することを目的とした，索引の安全性モデルである．選択平文攻撃を仮定しているが，あくまで文書により索引がどう変わるかに着目しており，本稿のような頻度分析の観点は対象にしていない．

本稿が対象とする頻度分析に対する安全性モデルは確立されていない [9].

4.2　方式

検索可能暗号には，大別して２つの方式がある．

１つは共通鍵暗号を使った方式，すなわち SSE であり，文書管理のアウトソーシングに向いている．これまでに，IND-CKA2 などの索引の安全性と，検索処理量を両立する方式が提案されている [4, 7, 6, 10]．しかし，本稿が対象とする頻度分析への耐性は，基本的な SSE と同程度に低い．

もう１つは公開鍵暗号による確率的な暗号化を使った方式 [1, 2] で，受信メールサービスのアウトソーシングなどに向いている．公開鍵暗号を使うことで，検索者は秘密鍵を持つ１組織であるのに対し，文書登録者は公開鍵を持つ複数の組織にすることができる．メール送信者が公開鍵で索引をつくりサーバに登録することで，サーバに対し暗号化したまま受信者がメールを検索できる検索可能暗号受信メールサービスを実現できる．この方式はクエリーを確率的に暗号化するため，クエリーだけからの同一性分析に耐性がある．しかし，対応分析まで考慮しておらず，頻度分析耐性があるとは言い難い．また，検索処理量が $\Omega(n)$ になってしまうという大きな課題がある．

5　提案方式

本稿では，頻度分析耐性と検索処理量のトレードオフを選択できる，新しい SSE 方式を提案する．基本的な SSE に比べ，検索処理量を悪化させる代わりに，頻度分析に対する耐性を向上させることができる．利用者は，利用前にパラメーター等を調整することで，そのトレードオフを選択できる．

以下，方式の詳細とその効果について説明する．

5.1　提案方式の構成

提案方式は，基本的な SSE におけるキーワードの暗号文の代わりに，グループ ID を使用する．グループとは，キーワードを分類して作られる，キーワードの集合である．グループ ID g は，利用者が設定した分類規則 f を用い，キーワード w から導出する．

$$g = f(w)$$

転置索引は，グループ ID g' から，それに分類されるいずれかのキーワード w を含む文書 ID 集合 $\{I(M)|\exists w, f(w) = g', w \in M\}$ を高速に取得できるようにする．さらに，転置索引の各グループ ID g について，各文書が含むキーワードのうちグループ ID が g であるキーワードの集合を暗号化し，転置索引に補助情報として持たせる．

図 2 に提案方式の構成を示す．転置索引では，グループ ID g から，対応する文書 ID と補助情報の組 $(I(M), E_r(\{w \in M|f(w) = g\}))$ を高速に取得できる．検索処理は次のようになる．

1. クライアントは，検索キーワード w' のグループ ID $g' = f(w')$ を得て，検索クエリー g' としてサーバに送る．
2. サーバは，転置索引から，クエリー g' に紐付けられた組 $(I(M), E_r(\{w \in M|f(w) =$

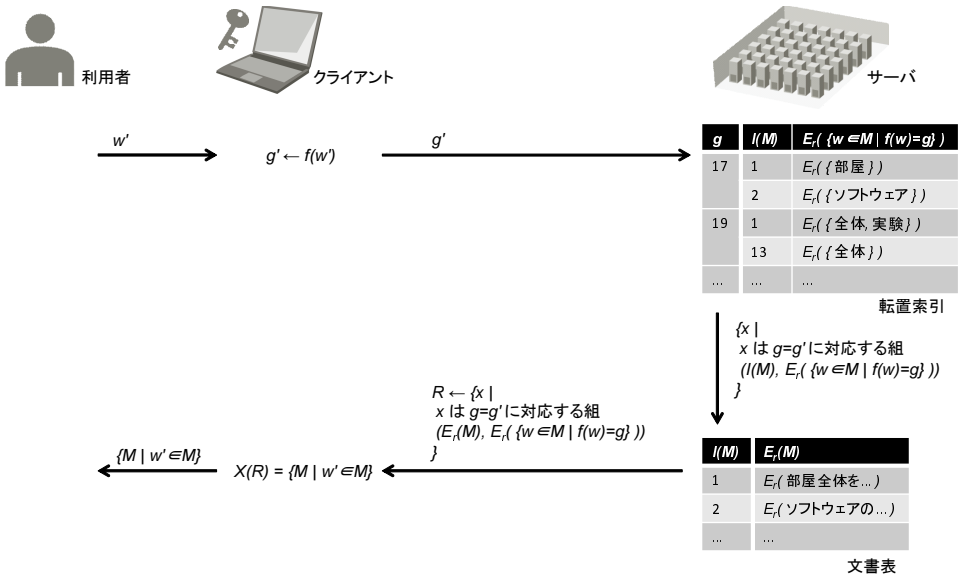

図 2: 提案方式の構成

$g\}))$ を全て取得し，さらに文書表の情報を使って各文書ID $I(M)$ を該当する文書 $E_r(M)$ に置き換え，検索結果 R としてクライアントに送る．

3. クライアントは，R の各補助情報を復号し w' が含まれている要素だけを抽出し，抽出した要素の文書を復号し，最終的な検索結果 $\{M|w' \in M\}$ を得る．図2の $X(R)$ はこの処理に相当する．

たとえば，図2で検索キーワード w' が「部屋」である場合，検索結果 R には w' を含まない文書IDが2のような文書も含まれるが，$X(R)$ でそのような w' を含まない文書は除外される．

補助情報 $E_r(w \in M|f(w) = g)$ は，文書 M が大きい場合にクライアントの処理負担を減らすために使用している．もし，クライアントで補助情報を復号するより，直接 M から w' を検索する方が効率的な場合は，補助情報は不要である．

5.2 分類規則

分類規則 f は，各グループに k 個以上のキーワードが分類され，各キーワードがそれぞれいずれか1つのグループに分類されるように設定する．それにより，キーワードが k-匿名性を満たすようになり，キーワードが一意に識別されなくなる．なお，キーワードとグループIDの関係は，攻撃者に推定できないようにすべきである．

w	g
部屋	17
ソフトウェア	17
全体	19
実験	19
…	…

図 3: 対応表の例

分類規則 f はクライアントで適用するため，そのデータ量が少ないことが望ましい．クライアントで必要なデータの量は最小限でないと，文書管理をアウトソースする意義が薄れるためである．

最も簡明な f は，キーワードとグループIDの対応表 f_t である．図3に f_t の例を示す．この方式の特長は，分類を利用者が完全に制御できる点にある．どのキーワード同士を同じグループに分類するかといったことを自由に決められる．たとえばキーワードの出現頻度があらかじめわかっているなら，グループ内キーワードの頻度の和が均一になるようにグループを作ることで，頻度分析を難しくすることができる．対応表 f_t を少ないデータ量で実現する方法とし

て，たとえば最小完全ハッシュ関数 [5] がある．

より少ないデータ量で f を実現する方法として，次のようにハッシュ値の剰余を導出する関数 f_h が考えられる．

$$f_h(w) := h(w) \bmod n/k$$

ここで，h は鍵付ハッシュ関数である．これにより，各グループは平均 k 個のキーワードを含むようになる．ただし，この方式では分類の制御はできず，k-匿名性も保証されない．

5.3 効果

提案方式の主な良い効果は，頻度分析耐性の向上である．

まず，対応分析への耐性が向上する．クエリーはグループ ID g' なので，検索結果には検索キーワード w' を含まない文書も含まれる．攻撃者は，そのうちどれが w' を含む文書なのかわからない．そのため，文書における検索キーワードの出現頻度もわからないし，たとえ検索キーワード w' がわかってもそのキーワードを含む文書がどれかはわからない．

また，同一性分析への耐性も向上する．複数のクエリーについて，同じグループ ID でのクエリーなことは容易にわかるが，利用者が同一の検索キーワードで検索したのかはわからない．そのため，検索キーワードの検索頻度もわからない．ただし，全体の検索数がある時期に急増したことはわかってしまうため，それにより検索キーワードが何か推定できる可能性がある．その対策は今後の課題である．

k の値が大きい程，検索キーワードを識別しづらくなるため，頻度分析の耐性は向上する．

一方，主な悪い効果は，検索処理量およびクライアントの処理量の増加と，クライアントで分類規則 f の管理が必要になることである．

検索キーワードのグループの要素数（キーワード数）を k' とすると，検索処理量，クライアントの処理量，検索結果通信量は，いずれも基本的な SSE に比べ約 k' 倍になる．これは，サーバでの検索結果が，基本的な SSE に比べ約 k' 倍になるためである．なお，補助情報は文書に比べデータ量が小さい場合が多く，それによる影響は少ないと考える．

分類規則 f はサーバに知られないようにすべきで，クライアントで管理しなければならないため，そのデータ量が多いと管理コストが大きくなる．たとえば，分類規則 f を暗号化してサーバに置いておき，文書検索時など必要なときに取得し復号する運用が考えられ，その運用コストは f のデータ量に比例する．そのため，分類規則 f のデータ量は少ないことが望ましい．

分類規則 f として関数 f_h を使う場合，文書本体の復号などに必要なデータと比べ，問題にならないほど小さい．ただし，前述の通り関数 f_h には分類の制御ができないなどの欠点がある．

一方，対応表 f_t を使う場合，データ量はキーワード数に比例する．表 1 に，日本語の単語約 40 万語を全キーワードとした場合の，各グループ 3 キーワード（$k = 3$）とする対応表 f_t のデータ量を計測した結果を示す．表の通り，最小完全ハッシュ関数を使うことで 2MB 以下になった．

提案方式のその他の効果として，構成が基本的な SSE と似ているため，基本的な SSE に適用できる技術を適用できる場合がある．たとえば，Kamara ら [7] は基本的な SSE を IND-CKA2 を満たすように拡張したが，同様の方法で提案方式を IND-CKA2 を満たすように拡張できる．

6 まとめ

本稿では，頻度分析耐性と検索処理量のトレードオフを選択できる，新しい SSE 方式を提案した．従来，十分に検索がおこなわれた状態での頻度分析耐性は考慮されていなかった．それに対し，提案方式は検索キーワードに k-匿名性を持たせることによりその耐性を向上させることができる．ただし，検索処理量は基本的な SSE に比べ約 k 倍になる．また，k-匿名性を持たせるための分類規則を，クライアントで管理する必要がある．分類を完全に制御できる対応表でそれを実現する場合，日本語の単語約 40 万語を全キーワードとすると 2MB 以下のデータ量になることを確認した．

表 1: 日本語の単語約 40 万語について作成した，対応表による分類規則のデータ量

方式	サイズ (MB)	zip 圧縮後サイズ (MB)
表形式	5.04	1.47
最小完全ハッシュ関数	1.74	1.19

提案方式でも，全体の検索数がある時期に急増したことはわかってしまうため，それにより検索キーワードが何か推定できる可能性がある．その対策は今後の課題である．

参考文献

[1] Dan Boneh, Giovanni Di Crescenzo, Rafail Ostrovsky, and Giuseppe Persiano. Public key encryption with keyword search. In *Advances in Cryptology - EUROCRYPT 2004*, Vol. 3027 of *Lecture Notes in Computer Science*, pp. 506–522. Springer, 2004.

[2] Dan Boneh, Eyal Kushilevitz, Rafail Ostrovsky, and William E. Skeith, III. Public key encryption that allows pir queries. In *Proceedings of the 27th Annual International Cryptology Conference on Advances in Cryptology*, CRYPTO'07, pp. 50–67, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] Christoph Bösch, Pieter Hartel, Willem Jonker, and Andreas Peter. A survey of provably secure searchable encryption. *ACM Comput. Surv.*, Vol. 47, No. 2, pp. 18:1–18:51, August 2014.

[4] Reza Curtmola, Juan Garay, Seny Kamara, and Rafail Ostrovsky. Searchable symmetric encryption: Improved definitions and efficient constructions. In *Proceedings of the 13th ACM Conference on Computer and Communications Security*, CCS '06, pp. 79–88, New York, NY, USA, 2006. ACM.

[5] Zbigniew J. Czech, George Havas, and Bohdan S. Majewski. An optimal algorithm for generating minimal perfect hash functions. *Inf. Process. Lett.*, Vol. 43, No. 5, pp. 257–264, October 1992.

[6] Seny Kamara and Charalampos Papamanthou. Parallel and dynamic searchable symmetric encryption. In Ahmad-Reza Sadeghi, editor, *Financial Cryptography*, Vol. 7859 of *Lecture Notes in Computer Science*, pp. 258–274. Springer, 2013.

[7] Seny Kamara, Charalampos Papamanthou, and Tom Roeder. Dynamic searchable symmetric encryption. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pp. 965–976, New York, NY, USA, 2012. ACM.

[8] Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Vol. 10, pp. 557–570, October 2002.

[9] 菅孝徳, 西出隆志, 櫻井幸一. 検索可能暗号の安全性再考. コンピュータセキュリティシンポジウム 2011 論文集, 第 2011 巻 of *3*, pp. 125–130, oct 2011.

[10] 伊藤隆, 服部充洋, 松田規, 坂井祐介, 太田和夫. 頻度分析耐性を持つ高速秘匿検索方式 (情報通信基礎サブソサイエティ合同研究会). 電子情報通信学会技術研究報告. ISEC, 情報セキュリティ, Vol. 110, No. 443, pp. 1–6, feb 2011.