

Secure Anonymous Database Search

Mariana Raykova, Binh Vo, Steven Bellovin, Tal Malkin

Columbia University

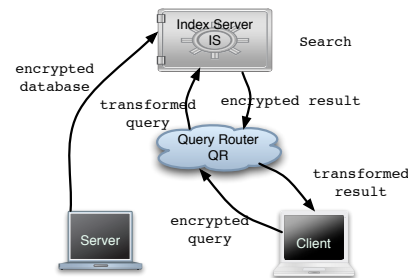


Problem Statement

Secure Anonymous Database Search – controlled data sharing between untrusting parties. Applications:

- Intelligence agencies sharing information
- Police investigation on sensitive information
- Medical records search
- Detecting attack behavior from log files
- Automatic email filtering

Security Architecture



Re-routable Encryption

(GEN, ENC, TRANS, ENC-TP, DEC-R) :

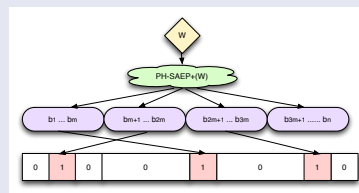
- GEN - generates keys for sender, receiver, and third party
- ENC - encrypts message from sender to third party
- TRANS - identifies receiver
- ENC-TP - transforms message received from user to ciphertext for the receiver, optionally performs a privacy preserving operation on receiver's ciphertext
- DEC-R - extracts information on the receiver side

Private Key Deterministic Encryption

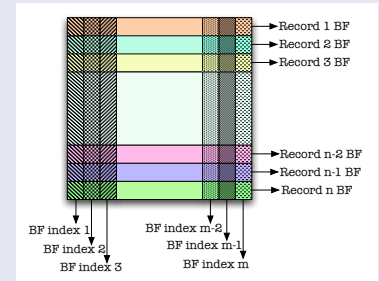
PH-DSAEP+ - private key deterministic encryption

- *Efficient search* – sublinear complexity in the number of ciphertexts, requires deterministic encryption. *Bellare et. al 2007* construction – replace randomness with hash of inputs.
- *Group property* – reencrypt under new key, allows user authorization and revocation.
- *PH-SAEP+ Construction* - Pohlig-Hellman function (*group property*) + SAEP padding (*security guarantee*); deterministic transformation *PH-DSAEP+* (*Bellare et. al 2007*)

Bloom Filter Search and Storage

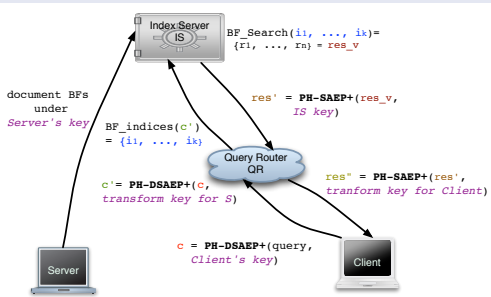


a) Bloom filter per document containing the word stems of the document



b) Multiple Bloom filters storage that allows efficient parallel search

SADS



Privacy

Server's database:

- IS cannot link BFs to documents; QR enforces client authorization; IS and QR cannot search;
- C receives only relevant results (adjustable FP rate).

Client's query:

- IS cannot link results to documents and queries of the same client;
- QR learns only equality of queries, QR does not learn anything about results.

Boolean Queries

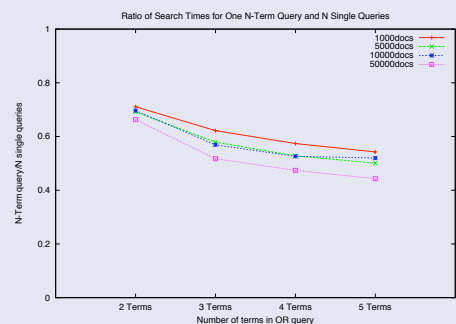
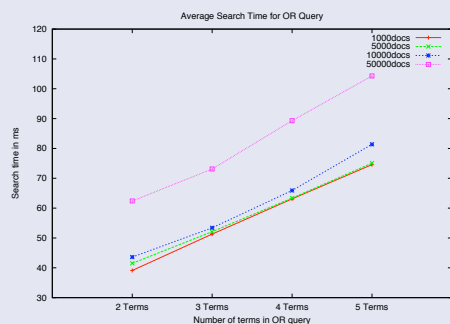
- **AND** queries *unioned* in query indices.
- **OR** queries processed in parallel; query indices are handled *in order of frequency in queries*, improves cache behavior.
- Efficient search for boolean queries representable in *monotone disjunctive normal form*.

Performance

Performance Results for Different Corpus Sizes



Fixed false positive rate = 0.001



Capacity	1k	2k	3k	5k	10k
Size(bits)	16384	32768	65536	131072	262144