Efficient Fuzzy-Searchable Encryption

ARC Fellowship SP12 Proposal
Nathan Chenette, ACO (Advisor: Boldyreva)

Background

Motivation. Current worldwide trends towards remote data storage solutions [1] such as cloud computing [5], in which large stores of data on insecure networks must be efficiently accessible, have emphasized the need to support efficient search functionality on large encrypted databases, or *(efficient) searchable encryption*, even at the expense of strong security.

Past work in searchable encryption. Searchable encryption is a balancing act of efficiency, functionality, and security. Past results [13, 14] have shown that the strong security cryptographers typically target is unattainable if efficient functionality—specifically, sub-linear search time—is required. Many past results in fact build constructions with efficient functionality but fail to address security other than in an ad-hoc manner. An emerging line of research bridges the gap by seeking formal *provable security*. The canonical task, for a given search functionality, is to (1) develop security notions that allow *only* the necessary security leakage, and (2) construct schemes that have the desired efficient functionality and provably satisfy the security definition.

Progress has been made for certain types of functionality, particularly exact-match queries [11, 2, 6, 7, 10]. In 2009, we (myself, A. Boldyreva, Y. Lee, and A. O'Neill) published the first cryptographic study of encryption allowing efficient range queries [8] through *order-preserving encryption* (OPE). We (myself, A. Boldyreva, and A. O'Neill) followed this work with a more in-depth study of the security guarantees of OPE [9]. Our work on OPE represents the total progress in the field and has received significant attention from both the crypto community and the database community as well as companies such as JP Morgan, Symantec, and SmartForce.

Fuzzy searchable encryption (FSE). One type of query that has been less studied in the context of efficiently searchable encryption is fuzzy search. In general, a fuzzy query returns database elements that correspond to a message "close to" the underlying queried message. Here, closeness is an abstract concept associated with the message space (we call the combined concept and space a "closeness domain"), and can be defined in a way appropriate to the application, e.g. edit distance for text or Hamming distance for binary values. We know of only one (ad-hoc) prior attempt [15] to construct an efficient FSE scheme, and it shows at least one crucial flaw (see below.)

Our goal. As we did with OPE, we initiate the first formal cryptographic study of FSE. In particular, we focus on developing appropriate security definitions and building the first provably-secure FSE schemes. We anticipate much academic and industrial interest in our work as fuzzy search is even more practicable (e.g., for searching biometric data) than range query search.

Current preliminary results in FSE

A security definition for fuzzy-searchable encryption schemes. We propose a novel definition abbreviated IND-FS-CPA. An adversary is given an encryption oracle and provides two challenge messages with the same *closeness pattern* versus all messages queried to the oracle. The scheme is insecure if an efficient adversary can correctly determine whether a ciphertext is the encryption of the left or right message with better than probability 1/2.

Analysis of the scheme of [15]. We describe a straightforward, efficient attack that shows the fuzzy-searchable scheme of [15] is IND-FS-CPA-insecure.

A new IND-FS-CPA-secure fuzzy-searchable encryption scheme. With the failure of the [15] scheme, no IND-FS-CPA-secure fuzzy-searchable encryption scheme is known. To rectify this, we design an efficient scheme that is IND-FS-CPA-secure, though (like the [15] scheme) it requires a large ciphertext length.

Optimality of the scheme's space-complexity for "perfect" functionality. We show (via a combinatorial argument) that the above scheme's ciphertext length is in fact asymptotically best-possible for "perfect" (no false positives) query support on an arbitrary closeness domain.

Space-efficient schemes. We design a general "bucketing" adaptation of the above scheme that can make it more space-efficient. The specific bucketing function and closeness domain chosen determines the security properties and space-efficiency of the scheme. Using this model we instantiate several application-relevant schemes that are not IND-FS-CPA-secure but whose insecure aspects (which we attempt to characterize) may be acceptable in applications. One bucketing instantiation uses locality-sensitive hashing (LSH), a topic receiving much attention in recent years [4, 3, 12]. (Thanks to Santosh Vempala for tuning us in to LSH.)

Distance-preserving encryption (DPE). Severely strengthening the FSE paradigm, suppose we require distances between ciphertexts to *equal* distances of underlying plaintexts, and define an analogous security definition, IND-DP-CPA. We have an interesting result: consider a situation where plaintext and ciphertext spaces are equal (which seems natural to allow distance-preservation.) By a algorithmic reduction argument, IND-DP-CPA-secure DPE is possible on such a space only if the message/ciphertext space is *fully homogeneous*. This rules out many useful metric spaces such as the discrete hypercube under Hamming distance. However, a weakened version of the security definition for Hamming distance can be achieved by a scheme that we construct.

Open problems for ongoing FSE research

Targeted instantiations of the bucketing scheme. A bucketing scheme's functionality and security depends on the bucketing function chosen. We should investigate bucketing functions to target various applications. Studying recent LSH function constructions is likely to be helpful here.

Improving security analysis of the space-efficient schemes. Our security analysis of the space-efficient schemes is currently somewhat non-intuitive, and we believe we can do better. In particular, it seems that if the closeness domain and bucketing scheme interact in a "regular" way across the message space, we can say something about how the scheme hides the relative locations of clusters of known ciphertexts across the space, which would be a very useful notion for practical applications. We need to formalize and prove this.

Metric and threshold-based solutions. Efficient fuzzy-searchable encryption seems naturally achievable using a ciphertext space with an associated metric, where distance under some threshold indicates closeness of underlying plaintexts. Such a solution, though, cannot be IND-FS-CPA-secure, and a straightforward weakening of the definition is appropriate. However, we conjecture that even the weakened security definition is unachievable for such a scheme assuming the scheme avoids false positives in some way. We would like to prove this conjecture.

Further distance-preserving encryption questions. Can we build a secure DPE scheme from a fully homogenous space to itself, or from a non-fully homogeneous space to some other space? Besides IND-DP-CPA, what secondary security guarantees would be desired for DPE schemes and can we achieve them?

References

[1] AccountingWEB. White paper: 10 reasons to outsource remote data protection. `http://www.accountingweb.com/print/node/135383`, February 2007.

[2] G. Amanatidis, A. Boldyreva, and A. O'Neill. Provably-secure schemes for basic query support in outsourced databases. In S. Barker and G.-J. Ahn, editors, *DBSec*, volume 4602 of *Lecture Notes in Computer Science*, pages 14–30. Springer, 2007.

[3] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:459–468, 2006.

[4] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117, 2008.

[5] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, and M. Zaharia. Above the clouds: A berkeley view of cloud computing. Technical report, 2009.

[6] M. Bellare, A. Boldyreva, and A. O'Neill. Deterministic and efficiently searchable encryption. In A. Menezes, editor, *CRYPTO*, volume 4622 of *Lecture Notes in Computer Science*, pages 535–552. Springer, 2007.

[7] M. Bellare, M. Fischlin, A. O'Neill, and T. Ristenpart. Deterministic encryption: Definitional equivalences and constructions without random oracles. In D. Wagner, editor, *CRYPTO*, volume 5157 of *Lecture Notes in Computer Science*, pages 360–378. Springer, 2008.

[8] A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill. Order-preserving symmetric encryption. In *EUROCRYPT*, pages 224–241, 2009.

[9] A. Boldyreva, N. Chenette, and A. O'Neill. Order-preserving encryption revisited: Improved security analysis and alternative solutions. In *CRYPTO*, pages 578–595, 2011.

[10] A. Boldyreva, S. Fehr, and A. O'Neill. On notions of security for deterministic encryption, and efficient constructions without random oracles. In D. Wagner, editor, *CRYPTO*, volume 5157 of *Lecture Notes in Computer Science*, pages 335–359. Springer, 2008.

[11] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky. Searchable symmetric encryption: improved definitions and efficient constructions. In *CCS '06: Proceedings of the 13th ACM conference on Computer and communications security*, pages 79–88, New York, NY, USA, 2006. ACM.

[12] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie, editors, *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases*, pages 518–529. Morgan Kaufmann, 1999.

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

[14] M. Kantarcioglu and C. Clifton. Security issues in querying encrypted data. In *Data and Applications Security XIX. LNCS, vol. 3654*, pages 325–337, Heidelberg, 2005. Springer.

[15] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou. Fuzzy keyword search over encrypted data in cloud computing. In *INFOCOM, 2010 Proceedings IEEE*, 2010.