Certifying Data from Multiple Sources

[Extended Abstract][*]

Glen Nuckolls
Dpt. of Computer Science
UC Davis
Davis, California
nuckolls@cs.ucdavis.edu

Chip Martel
Dpt. of Computer Science
UC Davis
Davis, California
martel@cs.ucdavis.edu

Stuart Stubblebine
Stubblebine Research Labs
8 Wayne Blvd.
Madison, New Jersey
stuart@stubblebine.com

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Information Storage and Retrieval—*Online Information Services*; E.1 [**Data Structures**]: General

General Terms: Algorithms, Reliability, Security

Keywords: Authentication, Hash, Database, Distributed

Introduction

There are many settings where data from multiple sources is combined into a single online database. A combined data set offers the convenience of a single source for queries and the addition of query types that are dependent on the overall data set. These benefits come at the cost of introducing new security concerns. Principal among them is the need to ensure the honesty of the party who collects the data and provides answers to users' queries. In order to assure accurate answers to queries, we need to prevent data corruption whether it is inadvertent or due to malicious intent.

We provide a formal framework and specific implementations which address the problems that arise when an untrusted third party publisher collects and organizes data from many different data owners and then provides answers to user queries on the combined data set. We focus on achieving the two closely related goals of providing a third-party publisher with **efficient** mechanisms to:

1. Assure data owners that their data will be accurately represented in answers to user queries.

2. Assure users that query answers will be accurate.

In our scheme, each owner gets a proof from the publisher that his data is properly represented, and each user gets a proof that the answer given to them is correct. Thus owners can be confident their data is properly represented and users can be confident they have correct answers. We show that a group of data owners can efficiently certify that an untrusted third party publisher has computed the correct digest of the owners' collected data sets. Users can then verify that the answers they get from the publisher are the same as a fully trusted publisher would provide, or detect if they are not. The results presented support selection and range queries on multi-attribute data sets and are an extension of earlier work on Authentic Publication which assumed that a single trusted owner certified all of the data.

Sites which provide provably accurate data have many important applications including consumer, financial, medical and government settings. Many of these applications draw from multiple sources for their content and we want to provide ways to prevent dishonest data representation. A product pricing engine provides a simple and familiar example where a number of data owners, here retailers, contribute data to an untrusted publisher, the pricing engine.

Extending Authentic Publication

Since our results extend Authentic Publication [3] to multiple owners, we briefly describe the basic single Owner protocol. The initial setup has three steps: (1) A trusted owner digests the data (e.g. with a binary search tree); (2) the data is given to one or more untrusted publishers; (3) the digest is distributed to the users. Once a Users has the digest, they can send queries to an untrusted publisher who returns an answer along with a proof that the answer is the same as would be given by the trusted owner. Merkle trees [8] provided the original basis for this work and so we give an example using a binary search tree over data set D, with the values stored at the leaves. The tree is digested using a cryptographic hash function as follows: The digest value of a leaf is the hash of its data value. The digest value of an internal node is the hash of the digest values of its (two) children. The overall digest value of the tree is just the digest value at the root. With this digest value, an efficient proof, of size $O(\log |D|)$ can be given that a data item is or is not in the set.

Efficient protocols based on the Merkle tree approach [3, 4] and general methods for producing them [7] exist for a number of settings [5, 9] and were proved to be secure in [7] as long as the publisher is unable to find a collision in the hash function. This approach has many advantages including scalability, since anyone can be a publisher, efficiency, since it uses efficient hash computations, and security, since there are no secret keys used in the proofs to be compromised. Also, in [1] they show how to guarantee data owners that individual data items are properly included in a digest.

A number of structures including authenticated dictionaries [4], and B-trees for data repositories [6] have been shown

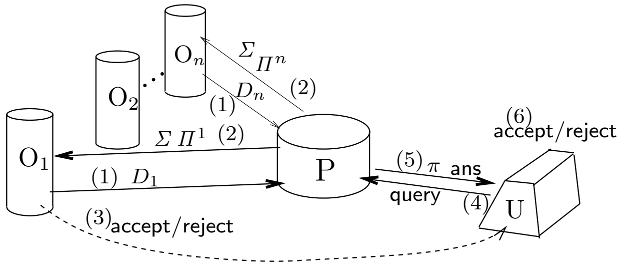
[*] Supported by NSF grant CCR 85961. A full version is available at `http://www.db.cs.ucdavis.edu/truthsayer/`

Copyright is held by the author/owner.
EC'03, June 9–12, 2003, San Diego, California, USA.
ACM 1-58113-679-X/03/0006.

to have authenticated versions. The work in [3] introduces Authentic Publication and broadened the range of authentication query types and in [2] this was extended to XML document collections.

Challenges of the Multiple Owner Setting

Extending Authentic Publication to multiple owners presents new challenges since there is no single trusted source. Now, the Publisher must convince each Owner that the digest is valid as well as convincing users that answers are accurate. This is difficult since each owner only sees a part of the data and digest structure being verified. Since the publisher is not trusted, we must also prevent him from creating and including his own values in the data set and digest. We now give a high level view of our protocols for multiple owners:



1. Owners O_i send their data sets D_i to the publisher P.
2. Publisher P collects the data sets and computes a digest Σ of the combined data set $D = \bigcup_{i=1}^{n} D_i$, and for each owner, a proof Π^i that the data set D_i is correctly included in Σ. Π^i and Σ are sent to each owner.
3. Each owner evaluates his proof and either accepts or rejects, notifying any users of this result and Σ.
4. A user U sends query to P.
5. P computes an answer ans and a proof π that ans is correct, returning these to U.
6. U evaluates ans and π, and either accepts or rejects.

We consider two cases: 1) all owners are trusted to complete the protocol, and 2) some owners are untrusted and might conspire with the publisher or other untrusted owners. We also consider settings where a trusted third party collects owner approvals of the digest. This reduces owner-user interactions and the user's need to know about participating owners.

Trusted owners

For a collection of n trusted owners, we provide an efficient protocol that an untrusted publisher can use to compute a digest value Σ for a binary search tree storing the owners' combined data set. The publisher provides a proof to each owner, proportional in size to each owner's data set (and the log of the combined set), that Σ is valid. If each owner accepts their proof, and there is no collision found in the hash function, then Σ is exactly what a fully trusted publisher would have produced for the owners' combined data. We achieve this using a *count certified search tree* which we define and use to support the protocol. A count certified search tree uses an enhanced version of the Merkle tree digest scheme. The digest value of a node now also uses the node's split value and the number of leaves in each of its subtrees. This provides support for, among other query types, authentication of answers to selection and range queries. Users can query the Publisher many times based on Σ, just as in the single owner Authentic Publication setting. We also extend this result to digests of multi-dimensional range trees which allows a richer set of multi-attribute queries (e.g. return all digital cameras which have 1-3 mega-pixels, cost $250-$400 and weigh at most 6 ounces).

Untrusted owners

If some owners are not trusted to follow the protocol, we still want to guarantee each honest owner who approves the digest that his data will be properly included in answers to queries. As before, each owner gets a proof of size proportional to his individual data set (and the log of the entire data set). If an owner approves this proof, then he can be confident that any answer a user accepts will contain exactly those data items from his data set which satisfy the user's query. This works for selection, range, and multi-dimensional range queries (e.g. in our digital camera example above, a retailer who approved the digest value of a publisher could be confident that all of his cameras which fit the user's criteria would be returned). We also provide new, efficient mechanisms to prevent false data attribution. An honest owner O can be sure that the publisher cannot return a data item which is claimed to belong to O, but is not in O's data set.

1. REFERENCES

[1] A. Buldas, P. Laud, and H. Lipmaa. Eliminating counterevidence with applications to accountable certificate management. *Journal of Computer Security*, 10:273–296, 2002.

[2] P. Devanbu, M. Gertz, A. Kwong, C. Martel, G. Nuckolls, and S. G. Stubblebine. Flexible authentication of xml documents. In *Proceedings of the 8th ACM Conference on Computer and Communications Security (CCS-8)*, pages 136–145, 2001.

[3] P. Devanbu, M. Gertz, C. Martel, and S. G. Stubblebine. Authentic third-party data publication. *14th IFIP 11.3 Working Conference in Database Security (DBSec 2000)*, 2000.

[4] M. Goodrich, R. Tamassia, and A. Schwerin. Implementation of an authenticated dictionary with skip lists and commutative hashing. *DISCEX II*, 2001.

[5] S. Haber and W. S. Stornetta. How to timestamp a digital document. *J. of Cryptology*, 3(2), 1991.

[6] P. Maniatis and M. Baker. Secure history preservation through timeline entanglement. In *Proceedings of the 11th USENIX Security Symposium*, San Francisco, CA, USA, Aug. 2002.

[7] C. Martel, G. Nuckolls, P. Devanbu, M. Gertz, A. Kwong, and S. Stubblebine. A general model for authentic data publication. `http://truthsayer.cs.ucdavis.edu/pubs.html`.

[8] R. Merkle. A certified digital signature. Advances in Cryptology–Crypto '89*, Lecture Notes in Computer Science*, 435:218–238, 1990.

[9] M. Naor and K. Nissim. Certificate revocation and certificate update. *Proceedings of the 7th USENIX Security Symposium*, 1998.