Secure Distributed Data Aggregation

Secure Distributed Data Aggregation

Haowen Chan

Carnegie Mellon University
USA
haowenchan@cmu.edu

Hsu-Chun Hsiao

CyLab/Carnegie Mellon University
USA
hchsiao@cmu.edu

Adrian Perrig

CyLab/Carnegie Mellon University
USA
perrig@cmu.edu

Dawn Song

University of California, Berkeley
USA
dawnsong@cs.berkeley.edu

the essence of knowledge

Boston – Delft

Foundations and Trends® in Databases

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is H. Chan, H.-C. Hsiao, A. Perrig and D. Song, Secure Distributed Data Aggregation, Foundation and Trends® in Databases, vol 3, no 3, pp 149–201, 2010

ISBN: 978-1-60198-450-0

Foundations and Trends® in Databases

Volume 3 Issue 3, 2010

Editorial Scope

Foundations and Trends® in Databases covers a breadth of topics relating to the management of large volumes of data. The journal targets the full scope of issues in data management, from theoretical foundations, to languages and modeling, to algorithms, system architecture, and applications. The list of topics below illustrates some of the intended coverage, though it is by no means exhaustive:

- Data Models and Query Languages
- Query Processing and Optimization
- Storage, Access Methods, and Indexing
- Transaction Management, Concurrency Control and Recovery
- Deductive Databases
- Parallel and Distributed Database Systems
- Database Design and Tuning
- Metadata Management
- Object Management
- Trigger Processing and Active Databases
- Data Mining and OLAP
- Approximate and Interactive Query Processing
- Data Warehousing
- Adaptive Query Processing
- Data Stream Management
- Search and Query Integration
- XML and Semi-Structured Data
- Web Services and Middleware
- Data Integration and Exchange
- Private and Secure Data Management
- Peer-to-Peer, Sensornet and Mobile Data Management
- Scientific and Spatial Data Management
- Data Brokering and Publish/Subscribe
- Data Cleaning and Information Extraction
- Probabilistic Data Management

Information for Librarians

Foundations and Trends® in Databases, 2010, Volume 3, 4 issues. ISSN paper version 1931-7883. ISSN online version 1931-7891. Also available as a combined paper and online subscription.

Foundations and Trends® in
Databases
Vol. 3, No. 3 (2010) 149–201

DOI: 10.1561/1900000025

Secure Distributed Data Aggregation

Haowen Chan[1], Hsu-Chun Hsiao[2], Adrian Perrig[3], and Dawn Song[4]

[1] *Carnegie Mellon University, USA, haowenchan@cmu.edu*
[2] *CyLab/Carnegie Mellon University, USA, hchsiao@cmu.edu*
[3] *CyLab/Carnegie Mellon University, USA, perrig@cmu.edu*
[4] *University of California, Berkeley, USA, dawnsong@cs.berkeley.edu*

Abstract

We present a survey of the various families of approaches to secure aggregation in distributed networks such as sensor networks. In our survey, we focus on the important algorithmic features of each approach, and provide an overview of a family of secure aggregation protocols which use resilient distributed estimation to retrieve an approximate query result that is guaranteed to be resistant against malicious tampering; we then cover a second family, the commitment-based techniques, in which the query result is exact but the chances of detecting malicious computation tampering is probabilistic. Finally, we describe a hash-tree based approach that can both give an exact query result and is fully resistant against malicious computation tampering.

Contents

1

Introduction

Recent advances in technology have made the application area of highly distributed data collection networks increasingly important. One example of this is sensor networks [39], which are wireless multihop networks composed of a large number of low cost, resource constrained nodes. Another example occurs in distributed systems such as distributed databases [34], peer-to-peer networks [58] or "grid" computing, where a large number of nodes are distributed over the Internet while engaging in some shared data collection or processing task.

A common task in these systems is the transmission of data towards a designated collection point, or *sink*. In sensor networks, the sink is typically a wireless base station that relays the collected data to an off-site server; in other distributed systems the sink may be a designated coordinating server that is responsible for archiving the data and answering user queries. The most straightforward method for collecting data is for each node in the network to send their raw data directly to the sink, via multi-hop routes in which intermediate nodes act as passive message forwarders and neither inspect nor modify the data. However, this approach is communication inefficient since not all of the collected data may be relevant or necessary for the application.

An alternative method for data collection is to observe that, commonly, only very simple *aggregation functions* are queried on the data. An aggregation function takes as inputs all the data values of the nodes in the network, but outputs only a single scalar. Examples of common aggregation functions include the sum of all data values; the count of the number of nodes fulfilling a given predicate, the minimum or maximum data value over the nodes in the network, and various measures such as mean and median. Since the result of computing an aggregation function is only a single value, this computation can be efficiently distributed in the network by having intermediate nodes compute sub-aggregates, leading to an extremely communication-efficient protocol. This technique is known as *in-network aggregation* [39] and is briefly described in Section 2.4.

The efficiency of in-network aggregation comes at a price to resilience, however, since it relies on the honest behavior of intermediate nodes in terms of computing accurate sub-aggregates. For example, for a sum computation, a malicious intermediate node with two children each reporting a data value of '1', could report an inaccurate sub-aggregate value of '100' instead of the correct value of '2', thus skewing the final result by a large amount. Such attacks are not easily preventable since the efficiency of in-network aggregation relies on intermediate sub-aggregators reporting only concise summaries of their received values; since a large fraction of the input data is hidden by necessity, this exposes the network to greater opportunities for attack.

In this article, we provide an overview of the known approaches towards combating such malicious mis-aggregation attacks. Ideally, a secure aggregation protocol should offer three key features: it should (1) produce accurate answers (typically, an accuracy guarantee bounded by some function of the number of malicious nodes in the network), (2) require only low communication overhead, and (3) be resilient against general node compromise models. We present a brief summary of several approaches drawn from a selection of the current literature as well as a more in-depth tutorial in one of the more important frameworks. In our selection of covered literature, our goal is to provide the reader with a general intuitive understanding of the field, rather than to bring the reader exhaustively up to date with all algorithms for the area. Towards this end, we have opted towards a more tutorial approach in terms of selecting the publications that most clearly exemplify a certain class of approaches (or which have been most influential historically), rather than focusing on breadth or depth of coverage in terms of the most effective or the most recent algorithms.

The remainder of the article is organized as follows. In Section 2 we define the problem and introduce the notion of in-network aggregation more rigorously. In particular, in this article we focus only on aggregation computations for which the secure aggregation problem is feasible: the family of such functions is examined and defined. In Section 3 we highlight some earlier work which show the basic flavors of integrity verification and result checking for secure aggregation. The existing literature on secure aggregation can be broadly divided into two categories: the first category uses *verifiable sampling* to provide resilient probabilistic estimates of the aggregate result; the second category uses *commitment verification*, which, unlike the first category, can provide highly precise results for which any malicious tampering is immediately evident, but at the cost of availability. We cover these approaches in Sections 4 and 5, respectively.

References

[1] D. J. Abadi, S. Madden, and W. Lindner, "Reed: Robust, efficient filtering and event detection in sensor networks," in *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 2005.

[2] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," *Journal of Computer and System Sciences*, vol. 58, no. 1, pp. 137–147, 1999.

[3] R. Anderson, F. Bergadano, B. Crispo, J.-H. Lee, C. Manifavas, and R. Needham, "A new family of authentication protocols," *ACM SIGOPS Operating Systems Review*, vol. 32, no. 4, pp. 9–20, 1998.

[4] B. Arai, G. Das, D. Gunopulos, and V. Kalogeraki, "Efficient approximate query processing in peer-to-peer networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 919–933, 2007.

[5] B. Babcock, S. Chaudhuri, and G. Das, "Dynamic sample selection for approximate query processing," in *Proceedings of ACM SIGMOD*, 2003.

[6] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan, "Counting distinct elements in a data stream," in *Proceedings of International Workshop on Randomization and Approximation Techniques*, 2002.

[7] V. Barnet and T. Lewis, *Outliers in Statistical Data*. John Wiley and Sons, Inc., 1994.

[8] J. Branch, B. Szymanski, C. Giannella, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," in *Proceedings of International Conference on Distributed Computing Systems (ICDCS)*, 2006.

[9] C. Castelluccia, E. Mykletun, and G. Tsudik, "Efficient aggregation of encrypted data in wireless sensor networks," in *Proceedings of MobiQuitous*, 2005.

[10] H. Chan and A. Perrig, "Efficient security primitives derived from a secure aggregation algorithm," in *Proceedings of ACM Conference on Computer and Communications Security (CCS)*, 2008.

[11] H. Chan, A. Perrig, B. Przydatek, and D. Song, "SIA: Secure information aggregation in sensor networks," *Journal of Computer Security*, vol. 15, no. 1, pp. 69–102, 2007.

[12] H. Chan, A. Perrig, and D. X. Song, "Secure hierarchical in-network aggregation in sensor networks," in *Proceedings of ACM Conference on Computer and Communications Security (CCS)*, 2006.

[13] J. Considine, F. Li, G. Kollios, and J. W. Byers, "Approximate aggregation techniques for sensor databases," in *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, 2004.

[14] G. Cormode, M. N. Garofalakis, S. Muthukrishnan, and R. Rastogi, "Holistic aggregates in a networked world: Distributed tracking of approximate quantiles," in *Proceedings of ACM SIGMOD*, 2005.

[15] G. Cormode and M. Hadjieleftheriou, "Finding frequent items in data streams," in *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 2008.

[16] G. Cormode and S. Muthukrishnan, "An improved data stream summary: The count-min sketch and its applications," *Journal of Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.

[17] E. D. Cristofaro, J.-M. Bohli, and D. Westhoff, "FAIR: Fuzzy-based aggregation providing in-network resilience for real-time wireless sensor networks," in *Proceedings of ACM Conference on Wireless Network Security (WiSec)*, 2009.

[18] J. Domingo-Ferrer, "A provably secure additive and multiplicative privacy homomorphism," in *Proceedings of International Conference on Information Security*, 2002.

[19] W. Du, J. Deng, Y. Han, and P. K. Varshney, "A witness-based approach for data fusion assurance in wireless sensor networks," in *Proceedings of IEEE Global Telecommunications Conference*, 2003.

[20] C. Estan and G. Varghese, "New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice," *ACM Transactions on Computer Systems (TOCS)*, vol. 21, no. 3, pp. 270–313, 2003.

[21] P. Flajolet and G. N. Martin, "Probabilistic counting algorithms for data base applications," *Journal of Computer and System Sciences*, vol. 31, no. 2, pp. 182–209, 1985.

[22] K. B. Frikken and J. A. Dougherty, "An efficient integrity-preserving scheme for hierarchical sensor aggregation," in *Proceedings of ACM Conference on Wireless Network Security (WiSec)*, 2008.

[23] S. Ganeriwal, L. K. Balzano, and M. B. Srivastava, "Reputation-based framework for high integrity sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 4, no. 3, pp. 1–37, 2008.

[24] M. Garofalakis, J. M. Hellerstein, and P. Maniatis, "Proof sketches: Verifiable in-network aggregation," in *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, 2007.

[25] P. B. Gibbons, "Distinct sampling for highly-accurate answers to distinct values queries and event reports," in *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 2001.
[26] M. Greenwald and S. Khanna, "Space-efficient online computation of quantile summaries," in *Proceedings of ACM SIGMOD*, 2001.
[27] M. Greenwald and S. Khanna, "Power-conserving computation of order-statistics over sensor networks," in *Proceedings of PODS*, 2004.
[28] S. Guha and A. McGregor, "Approximate quantiles and the order of the stream," in *Proceedings of PODS*, 2006.
[29] H. Çam, S. Özdemir, P. Nair, D. Muthuavinashiappan, and H. O. Sanli, "Energy-efficient secure pattern based data aggregation for wireless sensor networks," *Computer Communications*, vol. 29, no. 4, pp. 446–455, 2006.
[30] C. Hartung, J. Balasalle, and R. Han, "Node compromise in sensor networks: The need for secure systems," Technical Report CU-CS-990-05, Department of Computer Science, University of Colorado, January 2005.
[31] W. He, X. Liu, H. Nguyen, K. Nahrstedt, and T. F. Abdelzaher, "PDA: Privacy-preserving data aggregation in wireless sensor networks," in *Proceedings of IEEE INFOCOM*, 2007.
[32] L. Hu and D. Evans, "Secure aggregation for wireless networks," in *Proceedings of Symposium on Applications and the Internet Workshops*, 2003.
[33] P. J. Huber, *Robust Statistics.* Wiley, 1981.
[34] R. Huebsch, B. N. Chun, J. M. Hellerstein, B. T. Loo, P. Maniatis, T. Roscoe, S. Shenker, I. Stoica, and A. R. Yumerefendi, "The architecture of PIER: An internet-scale query processor," in *Proceedings of Conference on Innovative Data Systems Research (CIDR)*, 2005.
[35] P. Jadia and A. Mathuria, "Efficient secure aggregation in sensor networks," in *Proceedings of International Conference on High Performance Computing*, 2004.
[36] Y. W. Law, J. Doumen, and P. Hartel, "Survey and benchmark of block ciphers for wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 1, pp. 65–93, 2006.
[37] D. Liu and P. Ning, "Multilevel μTESLA: Broadcast authentication for distributed sensor networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 3, no. 4, pp. 800–836, 2004.
[38] C. Lochert, B. Scheuermann, and M. Mauve, "Probabilistic aggregation for data dissemination in VANETs," in *Proceedings of ACM International Workshop on Vehicular ad hoc Networks (VANET)*, 2007.
[39] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "TAG: A tiny Aggregation service for ad-hoc sensor networks," in *Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2002.
[40] A. Mahimkar and T. S. Rappaport, "SecureDAV: A secure data aggregation and verification protocol for sensor networks," in *Proceedings of IEEE Global Telecommunications Conference*, 2004.
[41] A. Manjhi, S. Nath, and P. B. Gibbons, "Tributaries and deltas: Efficient and robust aggregation in sensor network streams," in *Proceedings of ACM SIGMOD*, 2005.

[42] G. S. Manku and R. Motwani, "Approximate frequency counts over data streams," in *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 2002.
[43] G. S. Manku, S. Rajagopalan, and B. G. Lindsay, "Approximate medians and other quantiles in one pass and with limited memory," in *Proceedings of ACM SIGMOD*, 1998.
[44] M. Manulis and J. Schwenk, "Provably secure framework for information aggregation in sensor networks," in *Proceedings of the International Conference on Computational Science and Its Applications*, 2007.
[45] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Proceedings of CRYPTO*, 1987.
[46] S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson, "Synopsis diffusion for robust aggregation in sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 4, no. 2, pp. 7:1–7:40, 2008.
[47] A. Perrig, J. D. Tygar, D. Song, and R. Canetti, "Efficient authentication and signing of multicast streams over lossy channels," in *Proceedings of IEEE Symposium on Security and Privacy*, 2000.
[48] F. Picconi, N. Ravi, M. Gruteser, and L. Iftode, "Probabilistic validation of aggregated data in vehicular ad-hoc networks," in *Proceedings of ACM International Workshop on Vehicular ad-hoc Networks (VANET)*, 2006.
[49] B. Przydatek, D. X. Song, and A. Perrig, "SIA: Secure information aggregation in sensor networks," in *Proceedings of ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2003.
[50] R. Roman, C. Alcaraz, and J. Lopez, "A survey of cryptographic primitives and implementations for hardware-constrained sensor network nodes," *Mobile Networks and Applications*, vol. 12, no. 4, pp. 231–244, 2007.
[51] S. Roy, S. Setia, and S. Jajodia, "Attack-resilient hierarchical data aggregation in sensor networks," in *Proceedings of ACM Workshop on Security of Ad-Hoc and Sensor Networks (SASN)*, 2006.
[52] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," in *Proceedings of ACM MobiHoc*, 2007.
[53] N. Shrivastava, C. Buragohain, D. Agrawal, and S. Suri, "Medians and beyond: New aggregation techniques for sensor networks," in *Proceedings of ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2004.
[54] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 2006.
[55] G. Taban and V. D. Gligor, "Efficient handling of adversary attacks in aggregation applications," in *Proceedings of ESORICS*, 2008.
[56] D. Wagner, "Resilient aggregation in sensor networks," in *Proceedings of ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN)*, 2004.
[57] D. Westhoff, J. Girão, and M. Acharya, "Concealed data aggregation for reverse multicast traffic in sensor networks: Encryption, key distribution, and routing adaptation," *IEEE Transactions on Mobile Computing*, vol. 5, no. 10, pp. 1417–1431, 2006.

[58] P. Yalagandula and M. Dahlin, "A scalable distributed information management system," in *Proceedings of ACM SIGCOMM*, pp. 379–390, 2004.
[59] Y. Yang, X. Wang, S. Zhu, and G. Cao, "SDAP: A secure hop-by-Hop data aggregation protocol for sensor networks," in *Proceedings of ACM MobiHoc*, 2006.
[60] H. Yu, "Secure and highly-available aggregation queries in large-scale sensor networks via set sampling," in *Proceedings of International Conference on Information Processing in Sensor Networks*, 2009.
[61] Y. Zhang, X. Lin, Y. Yuan, M. Kitsuregawa, X. Zhou, and J. X. Yu, "Summarizing order statistics over data streams with duplicates," in *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, 2007.