A Survey Study on the Inference Problem in Distributed Environment

Adel Jebali
Tunis El Manar University,
Faculty of Mathematical Physical and Natural Sciences of Tunis,Tunisia
VPNC Laboratory
adel.jbali@fst.utm.tn

Salma Sassi
Jendouba University, Faculty of Law Economics and Management of Jendouba, Tunisia
VPNC Laboratory
salma.sassi@fsjegj.rnu.tn

Abderrazak JEMAI
Carthage University, Polytechnic School of Tunisia,
SERCOM Laboratory,
INSAT, 1080, Tunis, Tunisia
abderrazekjemai@yahoo.co.uk

***Abstract*— Traditional access control models aim to prevent data leakage via direct accesses. A direct access occurs when a requester poses his query directly on the desired object. However, these models fail to protect sensitive data from being accessed with inference channels. An inference channel is produced by the combination of the legitimate response which a user receives from the system and metadata. Detecting and removing inference in database systems guarantee a high-quality design in terms of data secrecy and privacy. Parting from the fact that data distribution exacerbates inference problem, we give in this paper a survey of the current and emerging research on the inference problem in both centralized and distributed database systems and highlighting research directions in this field.**

Keywords- Access Control, Inference Control, External Knowledge, Data Distribution, Secrecy and Privacy

I. INTRODUCTION

Access control models protect sensitive data from direct disclosure via direct accesses, however they fail to prevent indirect accesses [10]. Indirect accesses via inference channels occur when a malicious user combines the legitimate response that he received from the system with metadata. According to [11], external information to be combined with data in order to produce an inference channel could be database schema, system's semantics, statistical information, exceptions, error messages, user-defined functions and data dependencies. Detecting and removing inference in database systems guarantee a high-quality design in terms of data secrecy and privacy since this latter is considered as a new vision of the inference problem. Absolutely, this diversity of techniques to bypass access control mechanisms with inference channels has attracted considerable attention in recent years. A growing body of literature has examined the inference problem but no one of the proposed solutions seems to be the universal one. In reality, for each of the underlying techniques a specific solution has been proposed for handling each particular attack. There is consensus among security community that data distribution exacerbates inference problem. This is why several attempts have been done in the last two decades to address this problem. This paper investigates current and emerging research on the inference control in centralized database systems, then it highlights inference in distributed environment. The reminder of this paper is organized as follows: Section 2 provides a brief description of research efforts on controlling inference in centralized database systems, section 3 review works on the inference control in distributed environment. Research directions are given in section 4. Finally, we conclude in section 5.

DOI reference number: 10.18293/SEKE2019-062

II. INFERENCE CONTROL IN CENTRALIZED DATABASE SYSTEMS

Traditional access control models aim to prevent data leakage via direct accesses. A direct access occurs when a requester poses his query directly on the desired object. However, these models fail to protect sensitive data from being accessed via indirect accesses [10]. An inference problem (also called inference aggregation problem) occurs when a user deduces sensitive information from a sequence of innocuous information in the database. It has been widely investigated in the literature since 1987 with the emergence of multilevel database systems. The first works in this field are presented in [16, 20, 22].

A. Inference Attacks and Prevention Methods

According to [10], there are three types of inference attack: Statistical attacks, semantic attacks and inference due to data mining. For each of the mentioned techniques, researchers have devoted a lot of efforts to deal with inference problem. For statistical attacks, techniques like Anonymization and Data-perturbation have been developed to protect data from indirect access. For security threats based on data mining, techniques like privacy-preserving data mining and Privacy-preserving data publishing was carried out. Furthermore, a lot of works have investigated the semantic attacks [3, 15, 20].

There exist in the literature more than one criteria to classify approaches that deal with inference. One proposed criteria is to classify these approaches according to data level and schema level [28]. In such classification, inference constraints are classified into schema constraints level and data constraints level. Other criteria could be according to

the time when the inference control techniques are performed. According to this criteria, the proposed approaches are classified in two categories: design time [7, 13, 14, 15, 20, 26] and query run time [1, 3, 11, 21, 22].

B. Discussion of the Inference Prevention Methods

The purpose of inference control at design time is to detect inference channels from earliest stage and eliminate them. These approaches provide a better performance for the system since no monitoring module is needed when the users query the database, by consequence improving query execution time. Nevertheless, design time approaches are too restrictive and may lead to over classification of the data. Besides, it requires that the designer has a good concept of how the system will be utilized. On the other hand, run time approaches provide data availability since they monitor the suspicious queries at run time. However, run time approaches lead to performance degradation of the database server since every query needs to be checked by the inference engine. Furthermore, the inference engine needs to manage a huge number of log files and users. As a result, this could induce slowing down query processing. In addition, run time approaches could induce a non deterministic access control behavior (users with the same privileges may not get the same response).

To summarise, the main evaluation criteria of these techniques is a trade-off between availability and system performance. We assert that the distribution of the data exacerbates the inference and privacy problems. In the next section we investigate the inference problem in distributed environment.

III. Inference Control in Distributed Environment

Inference control in distributed environment have been investigated from early 2000 until now. This field of study has received intention from researchers in database security, due to the fact that distribution aggravates inference problems and privacy concerns. In this section, we start by investigating research efforts on inference prevention in distributed database systems, then we review inference in data integration systems and discuss different works for mitigating this later. We survey inference problem in data integration systems through the *Mediator/Wrapper* architecture for the reason that this is the most suitable design to access distributed, heterogeneous and autonomous data sources. Additionally, we highlight inference prevention methods in data outsourcing scenario.

A. Inference Control in Distributed Database Systems

In [4], the authors have considered the inference problem where the data is combined from distributed database and released to the final users. In this situation of data dissemination, problem arises when non-sensitive attributes compromise sensitive attributes. According to presented work, one technique to mitigate inference is by modifying the non-sensitive data in the database. Nevertheless, even with this modification, sensitive attributes still deducible when data from other databases is incorporated. The main idea behind this work is to not release certain non-sensitive information that can lead to probabilistic inference about the sensitive information while minimizing the loss of functionality. Consequently, the outputs are records that have been modified in order to anonymize sensitive attributes.

The authors of [24] have built on [4] to develop a work turning around inference prevention in distributed database systems. They proposed an inference prevention approach that enables each of the database in a distributed system to keep track of probabilistic dependencies with other databases and by consequence use that information to help preserve the confidentiality of sensitive data. The methodology is called "Agent-based" because every node in the distributed system is augmented with an agent to keep track of other nodes so that single point of failure and communication bottleneck are avoided. However, this approach has some limits. It treats the case where the distributed databases are overlapped (similar or have common attributes). Moreover, it assumes that the records in the distributed databases share the same keys constraints.

Inference problem have been also investigated in Peer-to-Peer environment through the work in [6]. The authors pinpoint the inference that occurs in homogeneous peer agent through distributed data mining and call this process peer-to-peer agent-based data mining systems. They assert that performing Distributed Data Mining (DDM) in such extremely open distributed systems exacerbates data privacy and security issues. As a matter of fact, inference occurs in DDM when one or more peer sites learn any confidential information about the dataset owned by other peers during a data mining session. The authors firstly classified inference attacks in DDM in two categories: inside attack scenario and outside attack scenario. After identifying DDM inference attacks, the authors propose an algorithm to control potential attacks (inside and outside attacks) to particular schema for homogeneous distributed clustering, known as *KDEC*. However, the algorithm proposed by the authors need to be improved from an accuracy point to expose further possible weakness of the KDEC schema.

B. Inference Control in Data Integration Systems

Inference control in data integration systems have been investigated in the last decade through the works in [12, 17, 18]. In such systems, a mediator is defined as a unique entry point to the distributed data sources. It provides to the user a unique view of the distributed data. From a security point of view, access control is a major challenge in this situation since the global policy must comply with the source policies. Complying with source policies means that a prohibited access at the source level should be also prohibited at the global level. [12, 17, 18] have demonstrated that despite the generation of a global policy at the mediator level that synthesizes and enforces the back-end data sources policies,

security breaches still possible via inference channel produced by semantic constraints. The problem is that the designer of the system cannot anticipates the inference channels that arise due to the dependencies that appear at the mediator level.

The first work attempting to control inference in data integration systems was introduced in [12]. The authors propose an incremental approach to prevent inference with functional dependencies. The proposed methodology includes three steps: synthesizing global policies, detection phase and Reconfiguration phase. In this work, authors have discussed only semantic constraints due to functional dependencies. Neither inclusion nor multivalued dependencies was investigated. Besides, other mapping approaches need to be discussed such as LAV and GLAV approaches.

The authors of [18] have inspired from [12] to propose an approach aiming to control inference in data integration systems. The proposed methodology resort to formal concept analysis as a formal framework to reason about authorization rules and functional dependencies as a source of inference. The authors adopt an access control model with authorization views and propose an incremental approach with three steps: generation of the global policy, global schema and global FD, Identifying disclosure transactions and Reconfiguration phase.

In [17] the authors have examined inference that arise in the web through RDF store. They propose a fine-grained framework for RDF data, then they exploit close graph to verify the consistency propriety of an access control policy when inference rules and authorization rules interact. Without accessing the data (at policy design-time), the authors propose an algorithm to verify if an information leakage will arise given a policy P and a set of inference rules R. Furthermore, the authors demonstrate the applicability of the access control model using a conflict resolution strategy (most specific takes precedence).

C. Inference and Data Outsourcing

Inference problem was not only investigated in previous distribution scenarios but also in data outsourcing. In this case, data owners place their data among cloud service providers in order to increase flexibility, optimize storage, enhance data manipulation and decrease processing time. Nonetheless, data security is widely recognized as a major barrier to cloud computing and other data outsourcing or Database-As-a-Service arrangements. Users are reluctant to place their sensitive data in the cloud due to concerns about data disclosure to potentially untrusted cloud providers and other malicious part [27]. It is from this perspective that inference problem was investigated in [2, 8].

In [2] authors resort to a controlled query evaluation strategy (CQE) to detect inference based on the knowledge of non-confidential information contained in the outsourced fragments and priori knowledge that a malicious user might have. Regarding that CQE relies on logic-oriented view on database systems, the main idea of this approach is to model fragmentation logic-oriented too allowing for inference proofness to be proved formally even the semantic database constraints that an attacker may hold. Besides, vertical database fragmentation technique was considered by authors in [8] to ensure data confidentiality in presence of data dependencies among attributes. Those dependencies allow unauthorized users to deduce information about sensitive attributes. To tackle this issue, authors reformulate the problem graphically through an hypergraph representation and then compute the closure of a fragmentation by deducing all information derivable from its fragments via dependencies to identify indirect access. Nevertheless, the major limit of this approach is that it explores the problem only in single relational database.

IV. RESEARCH DIRECTIONS

Since the discussed works are recent, there are a number of concepts associated to security policies, privacy, data distribution and semantic constraints which could be considered to ensure better security and prevent inference from occurring in distributed environment. Hence, there are many research directions to pursue:

- Absence of modularity in data integration systems: in the case where a new source joins the system it is necessary to revise the global schema and the global policy. This is not suitable for distributed environment where the source joins and leaves the system continuously (e.g. Mobile environment).
- Authors deal only with semantic constraints represented by functional dependencies and probabilistic dependencies as a source of inference. However, other semantic constraints, example inclusion dependencies, join dependencies and multivalued dependencies should be considered as sources of inference.
- In data integration scenario, all approaches aim to handle inference at query run time by keeping track of the history of user queries and the current query. In the case where the system deals with a large volume of data and users number, run time approaches will lead to performance degradation by slowing down query processing, consequently, this may push the server (mediator) to bottleneck. Hence, design time approach should be adopted to overcome these problems since it is performed offline.
- Another weakness in these approaches is the negligence of collaborative inference. In fact, authors propose to block a sequence of violating transaction from being achieved to prevent the inference channel, but, what if this violating transactions results from a

combination of a set of queries from more than one user?

- Functional dependencies should be considered as a source of inference in data outsourcing scenario. although data dependencies may resemble functional dependencies, they model a different concept. In future work, we will present a study aiming to prevent inference from occurring in distributed cloud database. Our approach is graph-based that firstly detects inference channels caused by functional dependencies and secondly breaks those channels by exploiting vertical database fragmentation while minimizing dependencies loss.

V. CONCLUSION

This paper has surveyed the inference problem from two perspectives: centralized and distributed design. We first gave a review of current and emerging research about the inference control in centralized database systems, we have introduced different inference attacks and their prevention methods and discussed the trade-off between them. Furthermore, an insightful discussion about inference control in distributed environment was provided. We also pinpoint potential issues that are still unresolved. These issues are expected to be addressed in future work.

REFERENCES

[1] Xiangdong An, Dawn Jutla, and Nick Cercone. 2006. Dynamic inference control in privacy preference enforcement. In Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services. ACM, 24.

[2] Joachim Biskup, Marcel Preuß, and Lena Wiese. 2011. On the inference-proofness of database fragmentation satisfying confidentiality constraints. In International Conference on Information Security. Springer, 246–261.

[3] Alexander Brodsky, Csilla Farkas, and Sushil Jajodia. 2000. Secure databases: Constraints, inference channels, and monitoring disclosures. IEEE Transactions on Knowledge and Data Engineering 12, 6 (2000), 900–919.

[4] LiWu Chang and Ira Moskowitz. 2003. A study of inference problems in distributed databases. In Research Directions in Data and Applications Security. Springer, 191–204.

[5] Yu Chen and Wesley W Chu. 2008. Protection of database security via collaborative inference detection. In Intelligence and Security Informatics. Springer, 275–303.

[6] R Lopez de Mantaras and L Saina. 2004. Inference attacks in peer-to-peer homogeneous distributed data mining. In ECAI 2004: 16th European Conference on Artificial Intelligence, August 22-27, 2004, Valencia, Spain: Including Prestigious Applicants [sic] of Intelligent Systems (PAIS 2004): Proceedings, Vol. 110. IOS Press, 450.

[7] Harry S. Delugach and Thomas H. Hinke. 1996. Wizard: A database inference analysis and detection system. IEEE Transactions on Knowledge and Data Engineering 8, 1 (1996), 56–66.

[8] Sabrina De Capitani di Vimercati, Sara Foresti, Sushil Jajodia, Giovanni Livraga, Stefano Paraboschi, and Pierangela Samarati. 2014. Fragmentation in presence of data dependencies. IEEE Transactions on Dependable and Secure Computing 11, 6 (2014), 510–523.

[9] Josep Domingo-Ferrer. 2002. Advances in inference control in statistical databases: An overview. In Inference Control in Statistical Databases. Springer, 1–7.

[10] Csilla Farkas and Sushil Jajodia. 2002. The inference problem: a survey. ACM SIGKDD Explorations Newsletter 4, 2 (2002), 6–11.

[11] Marco Guarnieri, Srdjan Marinovic, and David Basin. 2017. Securing Databases from Probabilistic Inference. In Computer Security Foundations Symposium (CSF), 2017 IEEE 30th. IEEE, 343–359.

[12] Mehdi Haddad, Jovan Stevovic, Annamaria Chiasera, Yannis Velegrakis, and Mohand-Saïd Hacid. 2014. Access control for data integration in presence of data dependencies. In International Conference on Database Systems for Advanced Applications. Springer, 203–217.

[13] Thomas H Hinke. 1988. Inference aggregation detection in database management systems. In Security and Privacy, 1988. Proceedings., 1988 IEEE Symposium on. IEEE, 96–106.

[14] Thomas H Hinke, Harry S Delugach, and Randall P Wolf. 1997. Protecting databases from inference attacks. Computers & Security 16, 8 (1997), 687–708.

[15] CE Landwehr and S Jajodia. 1992. The use of conceptual structures for handling the. inference problem. (1992).

[16] Matthew Morgenstern. 1988. Controlling logical inference in multilevel database systems. In Security and Privacy, 1988. Proceedings., 1988 IEEE Symposium on. IEEE, 245–255.

[17] Tarek Sayah, Emmanuel Coquery, Romuald Thion, and Mohand-Saïd Hacid. 2015. Inference leakage detection for authorization policies over RDF data. In IFIP Annual Conference on Data and Applications Security and Privacy. Springer, 346–361.

[18] Mokhtar Sellami, Mohand-Said Hacid, and Mohamed Mohsen Gammoudi. 2015. Inference control in data integration systems. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, 285–302.

[19] Jessica Staddon. 2003. Dynamic inference control. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM, 94–100.

[20] T-A Su and Gultekin Ozsoyoglu. 1991. Controlling FD and MVD inferences in multilevel relational database systems. IEEE Transactions on Knowledge and Data Engineering 3, 4 (1991), 474–485.

[21] Bhavani Thuraisingham, William Ford, Marie Collins, and Jonathan O'Keeffe. 1993. Design and implementation of a database inference controller. Data & knowledge engineering 11, 3 (1993), 271–297.

[22] MB Thuraisingham. 1987. Security checking in relational database management systems augmented with inference engines. Computers & Security 6, 6 (1987), 479–492.

[23] Tyrone S Toland, Csilla Farkas, and Caroline M Eastman. 2010. The inference problem: Maintaining maximal availability in the presence of database updates. Computers & Security 29, 1 (2010), 88–103.

[24] James Tracy, LiWu Chang, and Ira S Moskowitz. 2003. An agent-based approach to inference prevention in distributed database systems. International Journal on Artificial Intelligence Tools 12, 03 (2003), 297–313.

[25] Hui Wang and Ruilin Liu. 2011. Privacy-preserving publishing microdata with full functional dependencies. Data and Knowledge Engineering 70, 3 (2011), 249 – 268.

[26] Jingwen Wang, Jie Yang, Fen Guo, and Huaqing Min. 2017. Resist the Database Intrusion Caused by Functional Dependency. In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2017 International Conference on. IEEE, 54–57.

[27] Xiaofeng Xu, Li Xiong, and Jinfei Liu. 2015. Database fragmentation with confidentiality constraints: A graph search approach. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. ACM, 263–270.

[28] Raymond W Yip and EN Levitt. 1998. Data level inference detection in database systems. In Computer Security FoundationsWorkshop, 1998. Proceedings. 11th IEEE. IEEE, 179–189.