ARES project CONSOLIDER-INGENIO 2010 CSD2007-00004 Report on conciliation of database inference control and private information retrieval Deliverable Report

Jesús A. Manjón Paniagua and Constantinos Patsakis

Universitat Rovira i Virgili
UNESCO Chair in Data Privacy,
Department of Computer Engineering and Mathematics,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia

1 Introduction

Public access databases are an indispensable source of information, especially when the information needs to be constantly updated. But the databases also imply a high risk for the privacy of the users, since a curious administrator of a database can track the queries of a user and deduce her interests and necessities. There are several disciplines studying different aspects of privacy in communications.

1.1 Private Information Retrieval

In private information retrieval (PIR), a user wants to retrieve an item from a database or search engine without the latter learning which item the user is interested in. PIR was invented in 1995 by Chor et al. [1],[3] with the assumption that there are at least two copies of the same database, which do not communicate with each other. In the same paper, Chor et al. showed that single-database PIR (that is, with a single copy) is infeasible in the information-theoretic sense. However, two years later, Kushilevitz and Ostrovsky [4] presented a method for constructing single-database PIR based on the algebraic properties of the GoldwasserMicali public-key encryption scheme [5]. Subsequent developments in PIR are surveyed in [7]. In the PIR literature the database is usually modeled as a vector. The user wishes to retrieve the value of the ith component of the vector while keeping the index i hidden from the database. Thus, it is assumed that the user knows the physical address of the sought item, which might be too strong an assumption in many practical situations.

Keyword PIR [2] is a more flexible form of PIR: the user can submit a query consisting of a keyword and no modification in the structure of the database is needed. We claim that PIR protocols proposed so far have two fundamental shortcomings which hinder their practical deployment:

- (1) The database is assumed to contain n items and PIR protocols attempt to guarantee maximum privacy, that is, maximum server uncertainty on the index i of the record retrieved by the user. Thus, the computational complexity of such PIR protocols is $O(n)$, as proven in [1],[3]. Intuitively, all records in the database must be "touched"; otherwise, the server could rule out some of the records when trying to discover i. For large databases, an $O(n)$ computational cost is unaffordable [2].
- (2) It is assumed that the database server cooperates in the PIR protocol. However, it is the user who is interested in her own privacy, whereas the motivation for the database server is dubious. Actually, PIR is likely to be unattractive to most companies running queryable databases, as it limits their pro?ling ability. This probably explains why no real instances of PIR-enabled databases can be mentioned.

If one wishes to run PIR against a search engine, there is another shortcoming beyond the lack of server cooperation: the database cannot be modeled as a vector in which the user can be assumed to know the physical location of the keyword sought. Even keyword PIR does not really fit, as it still assumes a mapping between individual keywords and physical addresses (in fact, each keyword is used as an alias of a physical address). A search engine allowing only searches of individual keywords stored in this way would be much more limited than real engines like Google or Yahoo. In view of the above, relaxations of PIR seem necessary in order to attain practical systems offering some privacy in information retrieval. TrackMeNot [10] is a practical system based on a browser extension installed in the users computer that hides the users actual queries in a cloud of automatic ghost queries submitted to popular search engines at different time intervals. While practical at a small scale, if the use of TrackMeNot became generalized, the overhead introduced by ghost queries would significantly degrade the performance of search engines and communications networks. Also, the submission timing of automatic ghost queries may be distinguishable from the submission timing of actual queries, which could provide an intruder with clues to identify the latter type of queries.

2 ARES Contribution

2.1 Web Search Engines

Web search engines (WSE, e.g. Google, Bing,...) are widely used to find certain data among a huge amount of information in a minimal amount of time. However, these useful tools also pose a privacy threat to the users: web search engines profile their users by storing and analyzing past searches submitted by them. To address this privacy threat, current solutions propose new mechanisms that introduce a high cost in terms of computation and communication.

Stand-alone Solutions [16] define $h(k)$-private information retrieval ($h(k)$-PIR) as a practical compromise between computational efficiency and privacy. Also presented are $h(k)$-PIR protocols that can be used to query any database, which does not even need to know that the user is trying to preserve his or her privacy. The proposed methods are able to properly protect the privacy of users' queries. When internet users apply the protocols, search engines (e.g. Google) are not able to determine unequivocally the real interests of their users. The quality of the results decreases with the increase in privacy, but the obtained trade-off is excellent. A prototype called GooPIR has been developed in Java JDK 6.0 Standard Edition to implement this scheme (`http://crises2-deim.urv.cat/technology/get/id/1`). The prototype accepts queries consisting of single keywords or queries consisting of a logical AND of several keywords (with the limitation that independence between the keywords must be a plausible assumption). GooPIR locally masks the target keyword(s), submits the masked query to the Google search engine and then locally filters the results relevant to the target keyword(s).

Collaborative Solutions In our first approach to this theme [15], we presented a novel protocol specially designed to protect the users privacy in front of web search profiling. The system provides a distorted user profile to the web search engine. We offered implementation details and computational and communication results that show that the proposed protocol improves the existing solutions in terms of query delay. The scheme provides an affordable overhead while offering privacy benefits to the users.

In [24] we used social network for the first time. We proposed a new scheme designed to protect the privacy of the users from a web search

engine that tries to profule them. Our system provides a distorted user profile to the web search engine. The proposed protocol submits standard queries to the web search engine, thus it does not require any change in the server side. In addition to that, this scheme does not require the server to collaborate with the users. Our proposal is based on the two following assumptions: (i) the proliferation of home computers equipped with flat-rate broadband connection to the Internet. This implies that computers can be permanently connected to the network; and (ii) the gradual introduction of social networks.

[24] offers unique and very interesting features:

- It uses existing social networks in order to provide already-generated groups of users.
- These fixed groups are made of friends in real life. Therefore, the users of a certain group are very likely to share similar interests. As a result, [24] generates a distorted profile which is a trade-off between the privacy level achieved and the quality of the service.
- It outperforms the rest of proposals in the literature in terms of query delay. A system that provides a small query delay is more likely to be used by the users.

Nevertheless, the following research questions appear when considering this scheme:

- The privacy level achieved by the users of this proposal depends on the function that calculates the probability of submitting a query. Can this function be re-designed to improve the current results?
- Mechanisms to measure the privacy level achieved by the users are needed in order to compare different proposals. Is there a standard measure that can be used for this purpose?
- The simulations which are shown in [24] have been performed using synthetic queries (queries which are generated at random by a computer) and each user is always submitting the same one towards the WSE. Will the use of real queries (queries which are generated by humans) influence the behavior of this scheme in terms of privacy protection?

[25] addressed these research questions:

- The function used to decide which user must submit a certain query to the WSE was studied and re-designed. As a result, the privacy level achieved by the users was improved.

– A new measure to estimate the privacy achieved by the users, the *Profile Exposure Level (PEL)*, was proposed.
– For the first time the tests were performed using real data extracted from the AOL file [12]. In this way, the correct behavior of the proposed system was tested with queries which have been generated by real users.

These changes improved the privacy achieved by the users in the previous version while keeping its usability.

Previous proposals of privacy-preserving web search protocols increased significantly the query delay. This is the time that users have to wait in order to obtain the search results for their queries. For this reason, the protocol presented in [26] focused on reducing the query delay. The resulting scheme was implemented and tested in an open environment and the results showed that it achieves the lowest query delay which had been reported in the literature. On the other hand, the work presented in [27] focuses on improving the level of security of previous proposals. More specifically, this work proposed a multi-party protocol that protected the privacy of the user not only in front of the web search engine, but also in front of other members of her own group. The results showed that this scheme outperforms similar proposals in terms of computation and communication.

[28] was developed in collaboration with the Distributed Computation Group of the University of Lleida. This work focused on the development of a P2P network that groups users according to their search preferences. Once the users are classified, they execute a protocol that protects their privacy in front of the web search engine.

2.2 uPIR

Like [16] and [10], [17] and [18] propose to relax strict PIR in order to obtain a practical system. However, rather than cloaking a query in a set of queries in a standalone fashion, we propose here to cloak the users query history in a peer-to-peer user community: a user gets her queries submitted on her behalf by other users in the P2P community. In this way, the database still learns which item is being retrieved (which deviates from strict PIR), but it cannot obtain the real query histories of users, which become diffused among the peer users. We name the resulting PIR relaxation user-private information retrieval (UPIR). This approach certainly requires the availability of peers, not needed in the standalone

systems [16],[10], but it has some advantages: unlike [16], it does not require knowledge of the frequencies of all possible keywords and phrases that can be queried; unlike [10], it avoids the overhead of ghost query submission. Note that what we offer is different from what can be achieved using anonymization systems based on onion routing, like Tor [11]. In an onion routing system, the transport of data is protected by bouncing the communication between a user and a server around a distributed network of volunteer relays, with a view to protecting against traffic analysis. However, such systems give no end-to-end protection (at the application level). Specifically, as long as a search engine (or a database server) can link the successive queries submitted by the same user (e.g. by using cookies or some other mechanism), the profiling and the re-identification capabilities of the search engine are unaffected even if the user is submitting her queries through Tor: the user still submits all of her queries herself (the relays merely relay them), so her query history is unaltered and a query history may suffice for re-identification, as illustrated by the AOL query disclosure scandal in August 2006 [12]. What [17] and [18] propose is to diffuse a users query profile among the peers in a peer-to-peer community. However, onion routing systems can indeed complement our solution and be used for peers to communicate among themselves and hide their identity from each other at the transport level. The new scheme uses a type of combinatorial design called configuration to increase service availability and reduce the number of required keys (see [8],[9] for background on designs and configurations). The use of configurations in cryptographic key management is not new (e.g. see [9]), but their use in private information retrieval is.

2.3 Configurations

As we could see in the previous section, configurations are a very important point of a uPIR system. We have focus our research in this combinatorial research in order to improve our uPIR protocols.

We presented a first paper about this theme [19] but was [20] the first serious approach. In this paper we proved that the optimal configurations for the P2P UPIR protocol presented in [17] and [18] are the finite projective planes. We also presented an efficient and explicit algorithm for the construction of finite projective planes. Finally we gave another aspect of the optimality of finite projective planes; giving a short proof of the fact that they are Ramanujan graphs.

We have presented other advanced papers about configurations that can be looked up in [21], [22] and [23].

2.4 Query Logs

The search logs generated by a web search engine is a great source of information for researchers or marketing companies, but at the same time their publication may expose the privacy of the users from which the logs were generated [13]. There is at least one well known case of released search logs with poor anonymization, which have been shown to reveal enough information to re-identify some users. The release was done by AOL in an attempt to help the information retrieval research community, and ended up with not only important damage to AOL users privacy, but also a major damage to AOL itself with several class actions suits and complaints against the company [12]. Ideally, the search logs should be properly anonymized before they become public. The problem is that achieving a desirable degree of privacy in search logs is not easy, and presents an important trade-off between privacy and the usefulness of the data.

In [29] we addressed the privacy problem exposed by the WSE query logs, which can be made publicly available without risking the privacy of their users. To that end we followed the same ideas found in statistical disclosure control, proposing a novel microaggregation method to anonymize query logs. This approach ensures a high degree of privacy, providing k-anonymity at user level, while preserving some of the data usefulness. Moreover, and unlike most of the previous work, our approach took into account the semantics of the queries made by the user in the anonymization process making use of information obtained from the Open Directory Project [14]. A more extended version was presented in [30].

Another way to provide microaggregation at user level was presented in [31]. In this paper we defined a new user distance and aggregation operator. The user aggregation was designed in order to be as computationally efficient as possible. Note that the most important part is the aggregation of the queries since it is the information that will be more valuable in future analysis. Note also that queries are aggregated separately. An alternative could be to actually mix the terms of queries from different users to end up with new queries that somehow summarize all the users queries. We opted for the fist approach given the complexity that the second one imposes, and also because it already produced satisfactory. The other parts of the query are aggregated with the most common aggregation operators used in data privacy and statistical disclosure control. Other operators could easily be used, if required.

As always happens with statistical disclosure techniques, there is a trade-off between privacy and usability. We showed that our proposals,

besides providing k-anonymity, maintains to some extend the information of the original logs. Both, in terms of the information regarding the users, and in the use of the data for data mining. Our proposals can be seen as an efficient and relatively simple method to protect query logs, when compared to existing solutions, to ensure a high degree of anonymity and privacy.

References

1. B. Chor, O. Goldreich, E. Kushilevitz, M. Sudan, Private information retrieval. In: IEEE Symposium on Foundations of Computer Science (FOCS), pp. 4150 (1995)
2. Chor, B., Gilboa, N., Naor, M.: Private information retrieval by keywords. Technical Report TR CS0917, Department of Computer Science, Technion (1997)
3. B. Chor, O. Goldreich, E. Kushilevitz, M. Sudan, Private information retrieval. Journal of the ACM 45, 965981 (1998)
4. Kushilevitz, E., Ostrovsky, R.: Replication is not needed: single database, computationally-private information retrieval. In: Proc. of the 38th Annual IEEE Symposium on Foundations of Computer Science, pp. 364373 (1997)
5. Goldwasser, S., Micali, S.: Probabilistic encryption. Journal of Computer and Systems Science 28(1), 270299 (1984)
6. Beimel, A., Ishai, Y., Malkin, T.: Reducing the servers computation in private information retrieval: Pir with preprocessing. Journal of Cryptology 17, 125151 (2004)
7. Ostrovsky, R., Skeith-III, W.E.: A survey of single-database pir: techniques and applications. In: Okamoto, T., Wang, X. (eds.) PKC 2007. LNCS, vol. 4450, pp. 393411. Springer, Heidelberg (2007)
8. Stinson, D.R.: Combinatorial Designs: Constructions and Analysis. Springer, New York (2003)
9. Lee, J., Stinson, D.R.: A combinatorial approach to key predistribution for distributed sensor networks. In: Wireless Communications and Networking Conference-WCNC 2005, vol. 2, pp. 12001205 (2005)
10. D.C. Howe, H. Nissenbaum, TrackMeNot: resisting surveillance in web search, in: I. Kerr, C. Lucock, V. Steeves (Eds.), Lessons from the Identity Trail: Privacy, Anonymity and Identity in a Networked Society, Oxford University Press, Oxford UK, 2009, pp. 409428. Software downloadable from `http://www.mrl.nyu.edu/~dhowe/trackmenot/`
11. The Tor Project, Inc. Tor: Overview". `http://torproject.org/overview.html.en`
12. AOL Search Data Scandal, August 2006. `http://en.wikipedia.org/wiki/AOL_search_data_leak`
13. Jones, R., Kumar, R., Pang, B., & Tomkins, A. (2007). I know what you did last summer: Query logs and user privacy. In Proceedings of the sixteenth ACM conference on conference on information and knowledge management (pp. 909914). ACM.
14. ODP. Open directory project (2010)

15. J. Castellà-Roca, A. Viejo, J. Herrera-Joancomartí, "Preserving userś privacy in web search engines", Computer Communications, Vol. 32, no. 13, pp. 1541-1551, Aug 2009, ISSN: 0140-3664.
16. J. Domingo-Ferrer, A. Solanas, J. Castellà-Roca, $h(k)$-private information retrieval from privacy-uncooperative queryable databases, Online Information Review 33 (4) (2009) 720-744.
17. J. Domingo-Ferrer and M. Bras-Amorós, "Peer-to-peer private information retrieval", Lecture Notes in Computer Science, Vol. 5262 (Privacy in Statistical Databases- PSD 2008), pp. 315-323, Sep 2008, ISSN: 0302-9743.
18. J. Domingo-Ferrer, M. Bras-Amorós, Q. Wu, J. Manjón, "User-Private Information Retrieval Based on a Peer-to-Peer Community", Data & Knowledge Engineering, Vol. 68, no. 11, pp. 1237-1252, Nov 2009, ISSN: 0169-023X.
19. M. Bras-Amorós, J. Domingo-Ferrer and K. Stokes, "Configuraciones combinatricas y recuperacin privada de informacin por pares", Nuevos Avances en Criptografa y Codificacin de la Informacin - RSME 2009, Oviedo (Spain), Feb 2009.
20. K. Stokes and M. Bras-Amorós, "Optimal Configurations for Peer-to-Peer User-Private Information Retrieval", Computers & Mathematics with Applications, Vol. 59, no. 4, pp. 1568-1577, Feb 2010, ISSN: 0898-1221.
21. M. Bras-Amorós and K.Stokes, "The semigroup of combinatorial configurations", Semigroup Forum, Vol. 84, no. 1, pp. 91-96, Sep 2011, ISSN: 0037-1912.
22. K. Stokes and M. Bras-Amorós, "Associating a numerical semigroup to the triangle-free configurations", Advances in Mathemathics of Communication, Vol. 5, no. 2, pp. 351-371, May 2011, ISSN: 1930-5346
23. K.Stokes and O.Farràs, "Linear spaces and transversal designs: k-anonymous combinatorial configurations for anonymous database search", Designs, Codes and Cryptography. To Appear.
24. A. Viejo, J. Castellà-Roca, "Using Social Networks to Distort Users' Profiles Generated by Web Search Engines". Computer Networks. Vol. 54, issue 9, pages 1343-1357. Oct 2010. ISSN: 1389-1286.
25. A. Erola, J. Castellà-Roca, A. Viejo and J.M. Mateo-Sanz, "Exploting Social Networks to Provide Privacy in Personalized Web Search", Journal of Systems and Software, Vol. 84, no. 10, pp. 1734-17445, Oct 2011, ISSN: 0164-1212.
26. C. Romero-Tris, A. Viejo and J. Castellà-Roca, "Improving query delay in private web search", International Workshop on Securing Information in Distributed Environments and Ubiquitous Systems, Barcelona, Spain, Oct 2011.
27. C. Romero-Tris, J. Castellà-Roca and A. Viejo, "Multi-party private web search with untrusted partners", 7th International Conference on Security and Privacy in Communication Networks - SecureComm'11, London, UK, Sep 2011.
28. D. Castellà, C. Romero-Tris, A. Viejo, J. Castellà-Roca, F. Solsona and F. Giné, "Diseño de una red P2P optimizada para la privatización de consultas en WSEs", RECSI 2012, Donostia, Spain, In Actas de la XII Reunión Espaola sobre Criptología y Seguridad de la Información, pp. 273-278, ISBN: 978-84-615-9933, Sep 2012.

29. A. Erola, J. Castellà-Roca, G. Navarro-Arribas and V. Torra, "Semantic Microaggregation for the Anonymization of Query Logs", Lecture Notes in Computer Science, Vol. 6344 (Privacy in Statistical Databases-PSD 2010), pp. 127-137, Sep 2010, ISSN: 0302-9743.
30. A. Erola, J. Castellà-Roca, G. Navarro-Arribas and V. Torra, "Semantic microaggregation for the anonymization of query logs using the open directory project", SORT-Statistics and Operations Research Transactions, Vol. 0, Special issue, pp. 41-58, Sep 2011, ISSN: 1696-2281.
31. G. Navarro-Arribas, V. Torra, A. Erola and J. Castellà-Roca, "User k-anonymity for privacy preserving data mining of query logs", Information Processing and Management, Vol. 48, no. 3, pp. 476-487, May 2012, ISSN: 0306-4573.