Searchable Encryption for Sensitive Data in Social Network

A Confirmation Report submitted in fulfilment of the requirements for the candidature of

Doctor of Philosophy

By

Shangqi Lai

Supervisor: Dr. Joseph Liu

Co-supervisor: Dr. Ron Steinfeld, Dr. Dongxi Liu

Faculty of Information Technology

Monash University

20/11/2017

Contents

Chapter 1

Research Problem

This report outlines the doctoral research on solving sensitive data privacy issue in Online Social Network (OSN) by using 'searchable encryption' technique.

1.1 Background of Problem

OSN is highly popular for many years, and more and more Internet users start to use OSN as their medium of communication: The monthly active user of Facebook is reached 2 billion in 2017, which means over one half Internet users are using the services from Facebook [1, 2], and this is just an epitome of the today's OSN dependency.

A result of the huge growth of OSN is more and more sensitive personal data is revealed while OSN users enjoy using it. In 2005, two researchers from Carnegie Mellon University conducted a study in the online behaviour of more than 4,000 students in the university who had signed up in Facebook [3]. In their study, they found that the majority of Facebook users at CMU provided astonishing amount of sensitive information in their Facebook profiles: it may contain their real names (over 90%), date of birth (87%), personal images (80%) and so on; actually, privacy preference is provided to OSN user for controlling the visibility and access privilege of their profiles, but it is sparingly used, which makes most of the users searchable and identifiable to the vast majority of anybody else in the network.

Due to the richness and variety of sensitive information in OSN, the privacy of these data is of critical importance to OSN users, because the permissive and unconcerned uses of sensitive information will put themselves at risk for many attacks from the cyberspace or even the real world: The pre-

cise personal information collecting from OSN helps potential adversaries to construct more deceitful fraud messages (e.g. Context-Aware Spam [4]); it also can be used to re-identify some anonymous datasets like hospital medical information by matching the common attributes [5]; additionally, OSN users may be easily stalked as they revealed their location or timetable when they shared their activities with their friends.

Although the main purposes of OSN are to help its members to communicate with others and maintain social relations in a more convenient way, OSN service providers also design diverse Social Network Services (SNS) for its user interaction virtually: it includes updating activities and location information, sharing multimedia (e.g. photo, video) during some events, getting updates and comments on activities by friends, etc.. All of these services retain a huge amount of data about its users: Until 2013, there were 1.11 billion monthly active users and they put 4.75 billion shared items in Facebook, these contents received over 4.5 billion tags such as 'Likes' from their friends [6].

Making these data searchable to further enhance to capabilities of OSN had become a prevailing trend in recent years. In 2013, Facebook introduced their social data search engine – Graph Search [7]. This search engine aims to return the information based on the content from user's social network of friends and connections (their social circle): For example, users may search "The restaurants visited by friends live in Melbourne" to get restaurants recommendation from their friends in Melbourne. The search results are more personally relevant and satisfactory because the results are generated and refined by users' social circle where the people within always have some similar or the same attributes (e.g. geographical location, hobbies, education background etc.) according to "Homophily" theory [8]. The concept of Graph Search is also introduced by other OSN service provider: LinkedIn has Job Search Engine which is based on users' location, skill as well as the

information from their colleague and alumni; WeChat also can search user-specific content from Moments and Official Accounts.

In a nutshell, Graph Search-like search engines are powerful tools from the aspect of search because the search results are from a more reliable source (i.e. social circle) and more user-centric. However, they make user data searchable and then provide an extremely simple way to unearth user sensitive data: For instance, potential adversaries now can easily construct a query like "Friends in Melbourne working/traveling outsides" to find break-in crime targets. As a result, these new social data search engines ("Graph Search" is used to represent them) cause increasing privacy concerns about OSN to the public, and, to protect user's privacy in the context of Graph Search becomes the main motivation of this research.

1.2 Research Questions

In general, this research is going to answer following question:

How can we design a privacy preserving scheme to protect user's privacy in OSN?

in the context of today's OSN dependency and the inevitable trend of Graph Search wide deployment.

1.3 Research Scope

After providing the research question of this research, the scope of this research will be defined in this section. Indeed, there are multiple solutions to secure users' privacy in OSN, since the limited time and resource for this research, it will focus on solving privacy issue of Graph Search by addressing the following key technical research points:

because it can stop the privacy invasion not only from malicious adversaries, but also from untrusted service providers in such an , and try to

- **Cryptography in OSN.** Cryptographic solutions for privacy issue are

preferred in outsourced cloud service scenario such as OSN.

- **Search with Secure Index.** Existing research works had proposed privacy preserving schemes for sensitive data in OSN by using cryptographic solutions [9–16]. However, they are unsearchable after encryption and not compatible with Graph Search while this research attempts to introduce searchable encryption to satisfy the searchable requirement of privacy preserving scheme over Graph Search.

- **Extended Search Function for Encrypted Graph.** Previous works proposed several constructions based on searchable encryption to answer the limited queries such as adjacency queries, neighbour queries [17] and short distance queries [18] in encrypted graph. To fulfil the requirement of rich search functions in Graph Search, this research will design privacy preserving graph search scheme with boolean queries, spatial queries and ranked queries.

- **Large Scale Deployment.** Deploying the privacy preserving scheme into a extensive graph to enable private graph search on it is the final objective of this research, a formal security analysis and evaluation in big dataset will be presented to verify its performance in security as well as in efficiency over Big Data.

As a practical privacy preserving scheme is the main targeted artefact of this research, the research will not cover the modelling and algorithm design of searchable encryption scheme itself.

Chapter 2

Contribution to Knowledge

This research focuses on the research questions about the privacy issues on Graph Search and solve it by designing a cryptography based privacy preserving scheme for Graph Search with rich search functions. To the best of our knowledge, this is the first work that studies the above issue and provide a feasible solution for it. To enable the search functions over a encrypted graph, a service that supports secure and efficient boolean queries is proposed, and this service will be extended to support necessary functions for Graph Search. A formal security analysis will be given to verify the security performance of this scheme. Once the new scheme is fully defined, the scheme will be deployed in a distributed system, and subjected to intensive evaluation. The evaluation assesses the scheme's efficiency by collecting the queries delays and measuring the overhead from cryptography under a large dataset. The evaluation will play a important role to justify the practicability of the proposed scheme.

Chapter 3

Literature Review

3.1 Privacy Preserving Scheme for OSN

Numerous works have proposed to preserve the sensitive data privacy in OSN, and two strategies are mainly applied in these works: Data Decentralisation and Cryptography

3.1.1 Data Decentralised

Decentralised solutions attempt to provide social services through a collection of independent nodes and users can control their data in a flexible way:

Decentralised OSN in Peer-to-Peer overlays: PeerSoN [19] and Safebook [20] build a decentralised social network in Peer-to-Peer (P2P) overlays. Users can store their data in their local device and on the devices of their trusted friends, they consist of two parts: a distributed hash table (DHT) is used provide lookup service to find users and the data they store; and the P2P network offers direct data exchange between users' devices. In these social network, users are granted full privilege to control their data as well as the communication channel over P2P network: they can choose to either establish communication via their real-life trust [20], or the encrypted channel [19].

Decentralised OSN in outsourced storage: Diaspora [21] and Confidant [22] create the decentralised social network by decoupling users's data with the SNS. Users are asked for keeping their data in some trusted server, and then, the scheme can generate some data descriptors and use them to request the SNS from existing OSN. Others can read and retrieve the data

from trusted server via a valid descriptor and an appropriate access privilege. In comparison, Persona [9] allows its user keeping their data in untrusted server by using cryptographic techniques, and Cachet [16] uses the idea from P2P network to further address the performance issue in decentralised system by introducing social cache over DHT.

However, decentralised solutions without cryptography are considered as insufficient approaches for solving the privacy issue of OSN. The reason for it is because of the main purpose of these decentralised solutions is to provide access control by users' preference, and this functionality is overlap with OSN access control settings [23]. In addition, existing OSN service providers have ability to enforce the access control policy in a better way: For example, it is easier for a centralised OSN service provider to stop crawlers and makes their users' data invisible to search engines. Also, some OSN give "nudge" when users intent to reveal their sensitive data [24], but a decentralised system may not work properly in above cases, because its implementation only needs to fit the protocol but does not consider such details.

3.1.2 Cryptography

Cryptographic solutions protect the sensitive data via encryption, these solutions are preferred in public social network because it achieve access control by making data unreadable instead of enforcing some protocols, which is considered as more reliable way to stop unauthorised access.

Index Obfuscation: NOYB [10] stores users' data with untrusted service providers, but the index of these data is obfuscated by cryptographic technique. In this system, authorised users possess a part of secret can recover the real index and personal information associated with specific users while unauthorised users only get a mixture of arbitrary personal data from the crowd. NOYB is able to protect the privacy of user profile but it doesn't

work for user relations and interactions.

Attribute Based Encryption: Persona [9] uses cryptographic primitive (i.e. Attribute Based Encryption (ABE)) to enforce access control over a group of users in OSN. The ability to associate the attributes of user with a key provides a convenient way for group manager to grant different access privileges to different groups. Users within those groups use the key to access the secret key of group manager as well as his/her sensitive data, so each group's access to the sensitive data is properly restricted, which guarantees the privacy of group manager. In Persona, key revocation is not efficient because it needs to re-issue a new key to all remaining users in the group. To address this issue, Sun et al. [12] and Jahid et al. [15] achieve efficient key revocation by using broadcast encryption.

Private Social Relationship: Lockr [11] separates social relationship from OSN service providers by providing Social Attestation mechanism and Social Access Control Lists. Furthermore, Social Attestation mechanism makes it easier for key revocation as it has a explicit expire date. Another benefit for using Lockr applies proof of knowledge protocol to ensures the non-transferability of sensitive data – third parties sites cannot abuse these sensitive data because they will not receive the actual identifier of users under the protocol. Some other scheme also proposed private social relationship based on privacy homomorphisms [25] or Blind Signature [26].

Above cryptographic solutions protect users' privacy generally following two steps:

1. Encrypt data and put it into some public accessible storages.

2. Authorised users use their key to access, decrypt and use it.

However, this procedure infers the dilemma when we want to integrate OSN services with the schemes: Most of the OSN services are deployed in server side, if the data in server side is encrypted, it can't be used to support the OSN services.

3.2 Graph Search

Graph Search [7] is a typical OSN service to enrich the users' experience of using OSN. Nonetheless, it significantly improves users' search ability over user-generated data in OSN. To reach this goal, Facebook conducted a research about the social graph based on the large amount of the users' data they possessed, and they concluded that although there are billions of nodes representing the entities (e.g. users and user-generated contents like photos, pages) in social graph, the graph is very sparse because a typical node only have less than a thousand edges describing relations (e.g. friends, likes, tags) between nodes. Intuitively, it is better to represent the graph as a set of adjacency lists of users' id, and an inverted index it by edges [27]. Hence, Graph Search service is built upon the backend inverted index service for edges, and achieve a very low query delay (11 ms in average for finishing a query with 6 million results) for querying the entities such as friends, places, pages, etc.. Furthermore, Graph Search allows more complex queries such as graph traversal to find entities which are more than one edge away from source and it only takes 2000 ms at most for a 100 thousand results query.

Graph Search highly depends on users' data especially their relations to build its inverted index system. It is obvious that it can not be deployed in decentralised system as the relations are distributed in different network. Prior cryptographic solutions also have met the same dilemma that the system can not build inverted index to support Graph Search with encrypted data. To compatible with Graph Search, searchable encryption is introduced to guarantee the usability of Graph Search service by using cryptographic encrypted data.

3.3 Searchable Encryption

Graph Search makes use of users' data storing on their server to generate its inverted indexes, Searchable Symmetric Encryption (SSE) can offer a potential better solution for privacy issues in this data outsourcing case.

SSE with Index: The first SSE scheme is proposed by Song et al. [28], but his construction doesn't have an index to accelerate keyword search. However, index based SSE protocols have been proposed to support secure index, each more efficient than its predecessor: The first secure encrypted index was proposed in [29], based on form of forward index, storing for each document a Bloom filter containing all the document's keywords. This allows a single document to be searched in a constant time but requires each document to be checked in turn which gains a search complexity proportional to the number of documents.

Curtmola et al. [30] are the first to propose to use inverted-index data structure, storing in a hash table for each keyword, the encrypted IDs of the documents that contain it (while hiding the number of documents matching each keyword), resulting in complexity proportional to the number of matching results, even for searching the whole document collection. However, [30] does not support multiple keyword conjunctive queries efficiently; it has complexity proportional to the number of documents matching the most frequent queried keyword. Later, [31] presented the OXT protocol, extending [30] by adding a 'Cross-Tag Set' (XSet) data structure, which lists hashed pairs of keywords and IDs of documents containing them, and reducing search complexity to be proportional to the number of results matching the least frequent queried keyword.

Security Definition and Leakage of SSE: The studies of SSE security definition started from [29], their definition can make sure that adversaries can not deduct document's content from its index, but it doesn't consider

the security of search tokens (trapdoors for searching) However, Curtmola et al. [30] pointed out that the security of indexes and the security of trapdoors are inherently linked. In his definition, a SSE is secure if it reveals any information about the documents and the indexes besides the outcome and the pattern of the searches. Chase and Kamara [17] extended their definition to more general structured encryption, as SSE is only a special case of structured encryption.

Oblivious RAM (ORAM) allows SSE with no leakage [32]. However, it requires communication rounds and search complexity poly-logarithmic in the database size.To balance the efficiency and security of SSE, almost all of the practical schemes leak information about documents based on current security definition.

3.4 Searchable Encryption on Graph

The researches about Searchable Graph Encryption are very limited: The notion of Graph Encryption was introduced in [17]. In [17], several construction were proposed to answer Adjacency Queries (give two nodes, do they have a common edge?), Neighbour Queries (get all nodes will a common edge) and Labeled Graph Queries (search graph with some labels in nodes) in encrypted graph.

Meng et al. [18] proposed graph encryption schemes that support shortest distance query on graph encryption. Their construction is based on distance oracle instead of Dijkstra's algorithm so their algorithm can only get an approximation of shortest distance but enabling them to search on a large-scale graph.

3.5 Field of Research

As a conclusion of literature review, The privacy issue in OSN and its solutions are studied for many years, but none of them fit the context of Graph Search due to its searchable requirement. Since the lack of work

about Searchable Encryption on Graph, this research will try to apply the concept of SSE on Graph Search. The new privacy preserving scheme will be designed on the basis of Searchable Graph Encryption, and its performance of security will be further investigated as well as the efficient on large scale graph in distributed environment.

Chapter 4

Construction

Chapter 5

Progress

5.1 Courseworks

To meet the coursework requirements, I've finished two compulsory modules as well as the compulsory events and workshops in FIT 5144 in the first year. Table 5.1 shows the detailed information about my coursework activities.

Course Code	Status
FIT5143	Exempted
FIT6021	Finished

Table 5.1: Coursework Activties and Claimed Hours

5.2 Research

References

[1] Statista. Number of monthly active facebook users worldwide as of 3rd quarter 2017. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ [online], 2017.

[2] Internet World Stats. World internet usage and population statistics. http://www.internetworldstats.com/stats.htm [online], 2017.

[3] Ralph Gross and Alessandro Acquisti. Information Revelation and Privacy in Online Social Networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80. ACM, 2005.

[4] Garrett Brown, Travis Howe, Micheal Ihbe, Atul Prakash, and Kevin Borders. Social Networks and Context-Aware Spam. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 403–412. ACM, 2008.

[5] Latanya Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[6] Facebook. Facebook's Growth In The Past Year. https://www.facebook.com/media/set/?set=a.10151908376636729.1073741825.20531316728&type=3theater [online], 2013.

[7] Facebook. Facebook Graph Search. https://www.facebook.com/graphsearcher/ [online], 2013.

[8] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of A Feather: Homophily in Social Networks. *Annual review of sociology*, 27(1):415–444, 2001.

[9] Randy Baden, Adam Bender, Neil Spring, Bobby Bhattacharjee, and Daniel Starin. Persona: An Online Social Network with User-Defined Privacy. In *ACM SIGCOMM Computer Communication Review*, pages 135–146. ACM, 2009.

[10] Saikat Guha, Kevin Tang, and Paul Francis. NOYB: Privacy in Online Social Networks. In *Proceedings of the first workshop on Online social networks*, pages 49–54. ACM, 2008.

[11] Amin Tootoonchian, Stefan Saroiu, Yashar Ganjali, and Alec Wolman. Lockr: Better Privacy for Social Networks. In *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, pages 169–180. ACM, 2009.

[12] Jinyuan Sun, Xiaoyan Zhu, and Yuguang Fang. A Privacy-Preserving Scheme for Online Social Networks with Efficient Revocation. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[13] Michael Backes, Matteo Maffei, and Kim Pecina. A Security API for Distributed Social Networks. In *NDSS*, volume 11, pages 35–51, 2011.

[14] Ariel J Feldman, Aaron Blankstein, Michael J Freedman, and Edward W Felten. Social Networking with Frientegrity: Privacy and In-

tegrity with an Untrusted Provider. In *USENIX Security Symposium*, pages 647–662, 2012.

[15] Sonia Jahid, Prateek Mittal, and Nikita Borisov. EASiER: Encryption-based Access Control in Social Networks with Efficient Revocation. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pages 411–415. ACM, 2011.

[16] Shirin Nilizadeh, Sonia Jahid, Prateek Mittal, Nikita Borisov, and Apu Kapadia. Cachet: A Decentralized Architecture for Privacy Preserving Social Networking with Caching. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies*, pages 337–348. ACM, 2012.

[17] Melissa Chase and Seny Kamara. Structured encryption and controlled disclosure. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 577–594. Springer, 2010.

[18] Xianrui Meng, Seny Kamara, Kobbi Nissim, and George Kollios. GRECS: Graph Encryption for Approximate Shortest Distance Queries. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 504–517. ACM, 2015.

[19] Sonja Buchegger, Doris Schiöberg, Le-Hung Vu, and Anwitaman Datta. Peerson: P2p social networking: early experiences and insights. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, pages 46–52. ACM, 2009.

[20] Leucio Antonio Cutillo, Refik Molva, and Thorsten Strufe. Safebook: a Privacy Preserving Online Social Network Leveraging on Real-Life Trust. *IEEE Communications Magazine*, 47(12), 2009.

[21] Diaspora Project. diaspora*. `https://diasporafoundation.org` [online].

[22] Dongtao Liu, Amre Shakimov, Ramón Cáceres, Alexander Varshavsky, and Landon P Cox. Confidant: Protecting OSN Data without Locking It Up. In *Proceedings of the 12th International Middleware Conference*, pages 60–79. International Federation for Information Processing, 2011.

[23] Arvind Narayanan, Vincent Toubiana, Solon Barocas, Helen Nissenbaum, and Dan Boneh. A Critical Look at Decentralized Personal Data Architectures. *arXiv preprint arXiv:1202.4503*, 2012.

[24] Dyian Mori. Privacy nudges protect information. `http://thetartan.org/2010/3/22/scitech/privacynudges` [online], May 2010.

[25] Josep Domingo-Ferrer, Alexandre Viejo, Francesc Sebé, and Úrsula González-Nicolás. Privacy homomorphisms for social networks with private relationships. *Computer Networks*, 52(15):3007–3016, 2008.

[26] Emiliano De Cristofaro, Claudio Soriente, Gene Tsudik, and Andrew Williams. Hummingbird: Privacy at the time of twitter. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 285–299. IEEE, 2012.

[27] Michael Curtiss, Iain Becker, Tudor Bosman, Sergey Doroshenko, Lucian Grijincu, Tom Jackson, Sandhya Kunnatur, Soren Lassen, Philip Pronin, Sriram Sankar, et al. Unicorn: A System for Searching the Social Graph. *Proceedings of the VLDB Endowment*, 6(11):1150–1161, 2013.

[28] Dawn Xiaoding Song, David Wagner, and Adrian Perrig. Practical

Techniques for Searches on Encrypted Data. In *Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on*, pages 44–55. IEEE, 2000.

[29] Eu-Jin Goh et al. Secure Indexes. *IACR Cryptology ePrint Archive*, 2003:216, 2003.

[30] Reza Curtmola, Juan Garay, Seny Kamara, and Rafail Ostrovsky. Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions. *Journal of Computer Security*, 19(5):895–934, 2011.

[31] David Cash, Stanislaw Jarecki, Charanjit Jutla, Hugo Krawczyk, Marcel-Cătălin Roşu, and Michael Steiner. Highly-Scalable Searchable Symmetric Encryption with Support for Boolean Queries. In *Advances in cryptology–CRYPTO 2013*, pages 353–373. Springer, 2013.

[32] Oded Goldreich and Rafail Ostrovsky. Software protection and simulation on oblivious rams. *Journal of the ACM (JACM)*, 43(3):431–473, 1996.