Master Thesis Expose

Privacy-preserving Targeted Advertising in Peer to Peer Networks

1. Introduction

One of the important activities of web users is Social Networking. websites such as Facebook, Orkut, Instagram captured the interests of millions of people. Current Online Social Networking (OSN) Websites are based on centralized architecture and hence, called Centralised Online Social Networks (COSNs). COSNs provide web services which run on logically centralized infrastructure and provide a central repository for users and applications data. Therefore, service providers in COSNs have control over the users personal information such as posts, comments, photos, likes (possibly sensitive information) [1]. Various operations are performed on users personal data and thus exposing users privacy. Moreover, even after the agreement of legal policies by service providers, users personal information is exposed to third-party agencies for Targeted Advertising [1]. Most recent consequences from a popular service provider: Facebook, reveals how users privacy is oppressed. 87 million users profiles, which were collected over years is handed over to a political firm “Cambridge Analytica” [10]. In turn, this stored data was used to build user profiles and interests for target advertising to gain political interests [10].

However, analysing the privacy problems in current OSN seems to be fruitless and impractical, even if all the users are aware of the legitimate use of Social Networking Services (SNS), imposing appropriate privacy measures [2]. Authority of users information is still in the hands of service providers, which is a potential concern capable of exploiting users privacy. In the current situation, protection of users privacy is the primary objective, which current OSNs are not likely to provide.

The limitations in COSNs are addressed by designing an infrastructure, which decentralises the control of authority from service providers in OSN. In contrast to COSNs, decentralised online Social Network (DOSN) is implemented on a distributed platform, such as a network of trusted peers. In DOSNs, the concept of a single service provider is changed, where a set of peers share the tasks required to run the system. Now, users do not need to register with a single commercial service provider instead they can choose the trusted peer to host their personal information or users themselves can host the services. With this approach, separation of authority and control from service providers is achieved and believed that users have more control over their data, particularly in these aspects [4]:

Privacy: protecting users privacy is considered as the key characteristic of DOSNs. Users are left with capabilities to choose the service host, where to store the information and whom to share the information. Additionally, information can be stored in chunks of data within multiple hosts to address single point of failure problem.

Integrity: Additionally, DOSNs should be able to protect users identity and information from tampering and modification. Users identity is not necessarily provided by a centralised server, but also by trusted parties in a distributed architecture.

Availability: high availability ensures robust infrastructure against failures, exchange of messages. DOSNs should be able to provide continuous services to the users.

As discussed earlier, users of current OSNs are exposed to various privacy risks. Moreover, users information is used by SPs and data analytics companies to learn users behavior and interests for their own benefits. To overcome this drawback, researchers proposed DOSNs which are based on which are based on decentralised architectures, implemented on a network of trusted peers or peer to peer overlays [2]. In DOSNs, control of authority is moved from third-party SPs to users themselves. However, there are still challenges in adopting them. Social networks based on COSNs are widely used and the main challenge could be attracting a permanent user base. For instance, one of the popular DOSN-Diaspora [2], currently has about 669,000 users. Besides, being a popular DOSN, Diaspora is not well-established compared to OSNs such as Facebook, Twitter. Other limitations include, not everyone is interested to host the services on their computer. Moreover, managing DOSNs can be difficult for new users.

Above-mentioned limitations paved way for researchers to propose Hybrid Online Social Networks (HOSNs). Like DOSNs, users in HOSNs have control of their own information, what to share and whom to share, while enabling users to continue with existing SPs. Figure 1 gives a general overview of HOSNs. Users need not register with new OSNs to access services. As users are now able to continue with existing SPs which are based on COSN, their business model exists. TA is one of the important activities of COSN and we should consider their business model prevails. Service providers can play role in TA without exploiting users privacy. Therefore, in HOSNs:
Users have complete control of their data, allows users to perform computations on their own data for TA, in a secure manner.
Service providers have no control over management of users data. Therefore, SPs use data provided by users to perform TA.

The goal of this Thesis is designing and implementing a TA system. TA can be performed by using various data mining techniques such as association rule mining, collaborative filtering. We will work on association rule mining algorithms by adopting decentralised approach to OSN. A decentralised approach to OSN requires all the users to be identified, the integrity of users information and privacy-preserving approaches for secure communication between users. Initial literature reviews showed us that existing privacy-preserving approaches are not scalable considering current user base count. We will also work on privacy-preserving approaches which are scalable and efficient considering user base on current OSN.

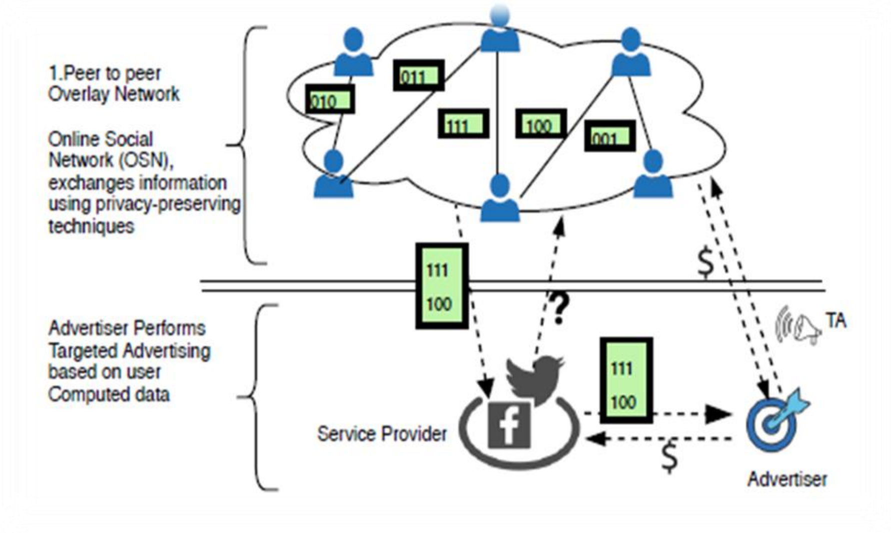

1.Peer to peer Overlay Network
Online Social Network (OSN), exchanges information using privacy-preserving techniques
010
011
111
100
001
111
100
$
?
TA
Advertiser Performs Targeted Advertising based on user Computed data
Service Provider
111
100
$
Advertiser

Figure 1 The general architecture of Hybrid Online Social Networks (HOSNs)

2. Background

In this section, we discuss basics of peer to peer networks and various privacy-preserving techniques for TA.

As discussed earlier, today's Social Networks such as Facebook, Twitter are based on centralised infrastructure also known as web-server based models (client-server). Given a COSN, users have little or no control over the information they post. Therefore, researchers started adopting peer to peer approach, with this approach decentralisation is achieved and combined with appropriate encryption mechanisms, users have freedom whom to share information and who can access their data.

Unlike Client-server model, peer to peer architecture does not require a public-hosted service to provide or receive services. In peer to peer overlay, a user can act as both server and client: given the context and capabilities of a peer. Current peer to peer architectures can be classified into two types [2]:

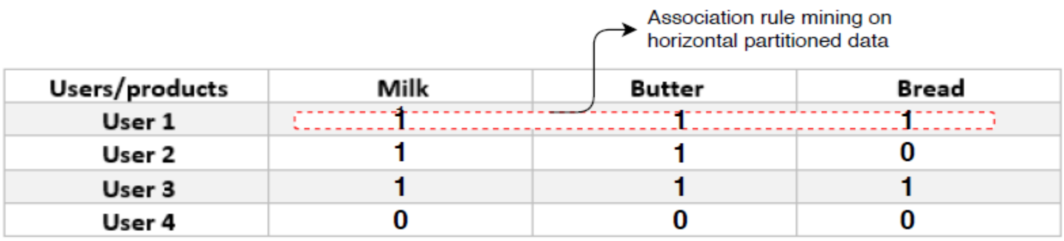
Unstructured: Peers in this architecture style do not follow definite structure and resources are scattered among peers according to the needs. Gnutella is one approach of unstructured peer to peer overlays and communication is carried out by flooding or broadcasting [8]. However, this style of peer to peer model is considered inefficient because of huge communication overhead. Unstructured overlays are suitable for services when there are not many peers to handle.

Structured: In this peer to peer overlay, nodes are arranged in a specific topology to ease the communication overhead and ensure good performance. DHT is an example of structured peer to peer overlays, where hashing techniques are used to identify peers and store contents [8].

Furthermore, peer to peer architecture styles can be classified into three types based on data storage [2]: 1**. Decentralised**: peers do not have the overhead of managing users information such as where to store the data. 2**. Semi-centralised**: super peers take control of managing all users information. 3. **Hybrid**: this peer to peer architecture relies on external service providers to store and manage data.

In HOSNs, users themselves carry out the computation of data, unlike COSNs. The aggregation of users individual data is required to perform targeted advertising. Leakage of data in the peer to peer network exposes users to various security breaches. Privacy-preserving techniques using association rules mining safeguards the sensitive information of users from unauthorized access.

The goal of distributed methods of privacy-preserving is to allow computation on aggregate datasets without compromising the privacy of individual datasets within different participants [17]. Thus, participants are not interested to share their own datasets but wish to obtain aggregate results. For this, data of individual users is arranged horizontally or vertically to perform computations on them [17]. Here, we consider partitioning data horizontally and perform data mining approaches such as filtering, association rule mining. Below are few of them explained. Figure2 gives a general idea of association rule mining on horizontal partitioned binary data.



Association rule mining on horizontal partitioned data

Users/products	Milk	Butter	Bread
User 1	1	1	1
User 2	1	1	0
User 3	1	1	1
User 4	0	0	0

Figure2. Association rule mining applied on horizontal portioned data

Association rule mining [20] is a technique widely used in data mining to determine the frequent itemset in large transactions. The main purpose of this technique is to find the interesting relations between products in a large-scale transaction database. For example, an association rule found in supermarket sales data {Bread, Butter} => {Milk}, denotes that if a customer buys bread and butter together he/she is also likely to buy milk. Such kind of information can be used to drive the sales of products using marketing policies.

Apriori Algorithm [20] is used for mining frequent itemsets for Boolean association rules. An Association rule has IF component is known as antecedent and THEN component is known as consequent. Apriori algorithm follows key concepts such as,

- **Support**: number of transactions which contain itemset X to the total number of transactions
- **Confidence**: rules have associated confidence, the conditional probability that consequent will occur given the occurrence of the antecedent.
- **Frequent itemsets**: The sets of items which has minimum support, if {XY} is frequent itemset both {X}{Y} should be frequent itemset. These frequent itemsets and the minimum confidence constraint are used to form association rules

Collaborative filtering (CF) is widely used for many applications as a recommender system. Various companies such as Amazon, Netflix use this approach to recommend various products, movies to the users based on their interests. CF uses large amounts of users historic or previous data and predicts users data to recommend products based on their likes and interests [16]. However, a user receives good recommendations only when users themselves actively participate in providing sufficient data.

3. Related Work

In this section, we give brief overview of previous researches on PPARM approaches and distributed approaches of OSN,

Murat Kantarcioglu and Chris Clifton [11] followed the approach of private association rule mining in two phases:

In **first phase**, each party encrypts their own itemsets, then encrypts itemsets of other parties (commutative encryption). These itemsets are passed around, with each site decrypting thus, obtaining the complete set.

In **second phase**, initial party passes its support count adding a random value, to its neighbour. This continues till the last party, and final party determines with initial party if the result is greater than random value plus the threshold. This is the basis of secure multi-party computation used in this paper.

Further, this paper also gave good overview of distributed association rule mining on horizontally partitioned data and distributed association rule mining under reasonable security assumptions.

Another study by Rachit V. Adhvaryu, Nikunj H. Domadiya [14] proposed a new algorithm for PPARM on horizontally-partitioned data. This algorithm was implemented in 3 phases.

In first phase, RSA cryptosystem was used and in second and third phase Homomorphic Pailier cryptosystem was used. The basic concepts of this algorithm are no involving party should know contents of other parties and adversaries should not be able to affect the privacy of communication between involving parties. The experimental results in this paper concluded that proposed algorithm has better performance with dense datasets than EMHS with increase in number of sites.

A paper by Pattnaik Dr. Prashant Kumar [21] adopted a hash based secure sum cryptography technique to find global association rules in a distributed database considering the complexity of PP in rule mining. Further, double hashing function is used to enhance privacy. Also results of global computation with trusted and without trusted-party are compared and fond that data leakage with trusted party is more as compared to without trusted-party.

With respect to the peer to peer overlays, Alexandra Olteanu, Guillaume Pierre [3] proposed an algorithm to build a robust and scalable peer to peer OSNs using simple techniques such as SpiderCast, high cluster coefficient. However, privacy and security are left for future research and is not integrated into their system. Another peer to peer implementations such as Safebook [4], PeerSoN assumes access control through encryption and traditional public-key cryptographies [4][9]. Unfortunately, these systems suffer from delays and incur communication costs [4] [9].

4. Research Problem

In this section, we present few limitations and drawbacks of previous researches in case of PPARM and in peer to peer to peer overlay networks.

Previous researches proposed many algorithms such as Secure Multi-Party Computation SMC with a trusted-party and with semi-honest model, PPDM-ARBSM algorithm, cryptography algorithms for preserving privacy in distributed approaches [14]. Disadvantages of the above-mentioned algorithms are,

- Failure of the trusted third party in a PPDM-ARBSM algorithm can lead to failure of communication and loss of data [13] [15].
- SMC with semi-honest model leads to an increase in computation time with an increase in the number of sites [14] [15]. Moreover, all the sites need to encrypt and decrypt the data to compute the global result which is a huge communication overhead with millions of sites [14].
- Homo-morphic encryption is used in previous PPARM approaches which is inefficient due to bulk of computations on the on the ciphertext.

Although many secure solutions exist, achieving efficient secure solutions for privacy-preserving is still open.

A scalable and robust peer to peer overlay has been difficult to choose with respect to availability, reliability as the OSN users change the status dynamically.

- In P2P OSNs, as the number of OSN users increase the active connections between the users of OSNs grows super-linearly.
- In case of node joins, failure needs additional computation time to recover to stable configuration.

5. Requirements

In this section, we discuss the requirements of both PPARM techniques and P2P overlay networks.

The aim of PPARM algorithms is to obtain relevant knowledge from a large amount of data and at the same time ensure the privacy of users information [19].

- A privacy-preserving algorithm should prevent the discovery of sensitive information to other parties in the network. It should not compromise in case of security vulnerabilities.
- It should be resistant to various data mining techniques, association mining rules, encryption and decryption mechanisms. The computational complexity should be kept as low as possible [12].
- After application of the privacy-preserving techniques and association rule mining algorithms, quality of data should be preserved without any alterations or modification of data [19].

- A privacy-preserving algorithm should be able to execute with good performance, given large sets of data over multiple sites in a distributed environment. Privacy-preserving algorithm also should adapt to dynamic changes to the data as well as network infrastructure.

Therefore, global knowledge extracted after application of privacy-preserving techniques can be used for targeted advertising by advertisers.

In addition to privacy-preserving techniques, a scalable and robust peer to peer overlay network should be established. Here are few requirements with respect to the peer to peer overlay networks [3][7]:

- **Scalability with respect to Network connectivity**: a peer to peer overlay should be able to connect millions of users without any performance degradation.
- **Cost effective packet processing**: packet transmission between the nodes should be carried out in an effective manner. A peer to peer overlay should be able to carry out the propagation of messages through a small number of edges and should be able to minimize encryption costs.
- **A tailored peer to peer architecture**: A peer to peer overlay may have millions of users, establishing active connections between all the nodes is not an efficient and reliable solution. Periodic updates between the nodes may incur huge communication costs and therefore, it is necessary to reduce or trim the number of edges between the nodes.

And finally, as discussed in the previous sections **privacy, integrity and availability** [7] are other important requirements a peer to peer overlay should ensure.

6.Methodology

In this Master Thesis, first we go through current established PP techniques and the association rule mining algorithms such as Apriori, FP-growth and generate rules for mining frequent datasets. After the survey of various PPARM techniques in previous and current ongoing researches, in our research we implement a PPARM method which can adapt to HOSNs for TA.

Also, we study various peer to peer overlays and choose the best-suited one for our implementation. Table 1 gives the brief overview of variously structured peer to peer overlays [6][8]

	Chord	**Kademlia**	**CAN**	**Pastry**
Architecture	Uni-directional and circular NodeID space	XOR metric for Distance between points in the key space.	Multidimensional ID coordinate space.	Plaxton-style global mesh network.
Lookup protocol	Matching key and NodeID.	Matching key and NodeID-based routing.	{key, value} pairs to map a point P in the coordinate space using uniform hash function.	Matching key and prefix in NodeID.

Reliability/ fault resiliency	Failure of peers will not cause networkwide failure. Replicate data on multiple consecutive peers. On failures, application retries.	Failure of peers will not cause network-wide failure. Replicate data across multiple peers	Failure of peers will not cause network-wide failure. Multiple peers responsible for each data item. On failures, application retries.	Failure of peers will not cause network-wide failure. Replicate data across multiple peers. Keep track of multiple paths to each peer.
Routing performance	O(logN)	O(logBN)+c where c = small constant	O(d.N1/d)	O(logBN)

Table1 [6]. Overview of structured peer to peer overlays

Also, we study various unstructured peer to peer overlays such as Gnutella, Freenet and analyse network if we can trim the nodes by reducing the active connections between each node, therefore achieving performance. Furthermore, we also consider software-defined networking WAN in terms of OSN, analyse if we can improve performance using OpenFlow protocol and various SDN controllers.

7. Timetable

Month	**Tasks**
	➢ Inform myself about various privacy-preserving techniques, association rule mining and peer to peer networks
1	➢ Analyse the state of the art ➢ Get familiarized with various peer to peer network overlays and pp techniques, AR mining rules ➢ Thesis Writing (State of the art)
2	➢ Design a concrete methodology ➢ Implementing new peer to peer network overlays , PP and AR techniques and alogorithms ➢ Thesis Writing (Background)
3	➢ Implementing algorithms
4	➢ Implementing peer to peer network evaluation process ➢ Develop a communication prototype
5	➢ Experiments & evaluation ➢ Thesis Writing (Methodology)
6	➢ Thesis Writing (Experiments & Results) ➢ Thesis Writing (Finishing) ➢ Presentation preparation

Table 2. Timetable for six months

8.References

[1] Decentralization: The Future of Online Social Networking Ching-man Au Yeung1, Ilaria Liccardi1, Kanghao Lu2, Oshani Seneviratne2, Tim Berners-Lee2 1 School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK 2 Decentralized Information Group, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] De Salve, Andrea, Paolo Mori, and Laura Ricci. "A survey on privacy in decentralized online social networks." *Computer Science Review* 27 (2018): 154-176.

[3] Olteanu, Alexandra, and Guillaume Pierre. "Towards robust and scalable peer-to-peer social networks." *Proceedings of the Fifth Workshop on Social Network Systems*. ACM, 2012.

[4] Cutillo, Leucio Antonio, Refik Molva, and Thorsten Strufe. "Safebook: A privacy-preserving online social network leveraging on real-life trust." *IEEE Communications Magazine*47.12 (2009): 94-101.

[5] Datta, Anwitaman, et al. "Decentralized online social networks." *Handbook of Social Network Technologies and Applications*. Springer, Boston, MA, 2010. 349-378.

[6] Lua, Eng Keong, et al. "A survey and comparison of peer-to-peer overlay network schemes." *IEEE Communications Surveys & Tutorials* 7.2 (2005): 72-93.

[7] Zhang C, Sun J, Zhu X, Fang Y. Privacy and security for online social networks: challenges and opportunities. IEEE network. 2010 Jul;24(4).

[8] A. Anitha, J. JayaKumari and G. Venifa Mini, "A survey of P2P overlays in various networks," *2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies*, Thuckafay, 2011, pp. 277-281.

[9] O. Bodriagov, G. Kreitz and S. Buchegger, "Access control in decentralized online social networks: Applying a policy-hiding cryptographic scheme and evaluating its performance," *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, Budapest, 2014, pp. 622-628.

[10] Tuttle, Hilary. "Facebook Scandal Raises Data Privacy Concerns." *Risk Management* 65.5 (2018): 6-9.

[11] Kantarcioglu, Murat, and Chris Clifton. "Privacy-preserving distributed mining of association rules on horizontally partitioned data." *IEEE transactions on knowledge and data engineering* 16.9 (2004): 1026-1037.

[12] Tassa, Tamir. "Secure mining of association rules in horizontally distributed databases." *IEEE Transactions on Knowledge and Data Engineering* 26.4 (2014): 970-983.

[13] Shmueli, Erez, and Tamir Tassa. "Secure multi-party protocols for item-based collaborative filtering." *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 2017.

[14] Adhvaryu, Rachit V., and Nikunj H. Domadiya. "Privacy Preserving in Association Rule Mining On Horizontally Partitioned Database." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*3.5 (2014).

[15] Danasana, Jayanti, Raghvendra Kumar, and Debadutta Dey. "Mining Association Rule For Horizontally Partitioned Databases Using Ck Secure Sum Technique." *International Journal of Distributed and Parallel Systems* 3.6 (2012): 149.

[16] Xu, Lei, et al. "User participation game in collaborative filtering." *Data Privacy Games* (2018): 119-149

[17] Jain, Nidhi, and Angad Singh. "A SURVEY ON PRIVACY PRESERVING MINING VARIOUS TECHNIQUES WITH ATTACKS." (2017).

[18] Sriramoju, Shoban Babu. "Analysis and Comparison of Anonymous Techniques for Privacy Preserving in Big Data." Analysis 6.12 (2017).

[19] Bertino, Elisa, Dan Lin, and Wei Jiang. "A survey of quantification of privacy preserving data mining algorithms." *Privacy-preserving data mining*. Springer, Boston, MA, 2008. 183-205.

[20] Danasana, Jayanti, Raghvendra Kumar, and Debadutta Dey. "Mining Association Rule For Horizontally Partitioned Databases Using Ck Secure Sum Technique." *International Journal of Distributed and Parallel Systems* 3.6 (2012): 149.

[21] Kumar, Pattnaik Dr Prasant, Kumar Raghvendra, and SharmaDr Yogesh. "Privacy preservation in distributed database." European Journal of Academic Essays 1.2 (2014): 35-39.