IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 24, No. 4, August, 2019, pp. 433-437
Journal homepage: http://iieta.org/journals/isi

Design of a Privacy-Preserving Algorithm for Peer-to-Peer Network Based on Differential Privacy

Jian Yu

School of Electronic Information Engineering, Liuzhou Vocational and Technical College, Liuzhou 545005, China

Corresponding Author Email: alex.yjian@gmail.com

https://doi.org/10.18280/isi.240411

Received: 17 February 2019
Accepted: 20 June 2019

Keywords:
peer-to-peer network (P2P), privacy preserving, differential privacy, sensitivity, privacy budget

ABSTRACT

In the peer-to-peer network (P2P), the private information of individual users faces the risks of being tracked, identified or leaked. This paper attempts to develop a security technique that fully protects the privacy of P2P users. Firstly, the node data were collected and analyzed from the P2P. Then, a privacy-preserving algorithm was proposed for the P2P based on the differential privacy model. The proposed algorithm adds noise to the degree distribution of the nodes, and solves the high sensitivity under the Laplace mechanism, which arises from the unique structure of the P2P. Finally, the proposed algorithm was verified through experiments. The results show that our algorithm protected the privacy of individual data in the storage and dissemination process, controlled the sensitivity to a low level through noise addition, and allocated the privacy budget rationally. Thus, the proposed algorithm is available and reliable for P2P operations like credit verification, transmission and storage, and enjoys excellent effectiveness and robustness.

1. INTRODUCTION

In the 40th China Statistical Report on Internet Development, it's reported that China's Internet penetration rate has reached nearly 60%, wherein the P2P file sharing, transmission and video traffic account for up to 70% [1, 2]. The P2P is a distributed trust mechanism adopted by the decentralized networks, the nodes in the network are random and free joining or exiting the network is allowed, and meanwhile the nodes follow the power-law distribution which has the characteristics of small-world property, scale-free, and high clustering [3-5]. Although users want to realize individual anonymity in P2P, in actual applications, the texts, pictures and videos shared by the users contain a large amount of individual private information, as long as P2P is used, the information might be leaked. Another major security threat of P2P comes from the exchange, transmission and storage of a large amount of information between nodes. Even if the large amount of user information has been encrypted, with certain background knowledge, one can still attack it and obtain the privacy of individual users [6]. Therefore, how to prevent the leakage of user individual privacy information in P2P applications has become one of the hot topics for many researchers.

At present, the main privacy-preserving scheme of P2P is anonymous processing of nodes and sensitive edges. It mainly includes the k-1 algorithm proposed by Han et al. [7], the anonymous node algorithm proposed by Yuan et al. [8] and the anonymization privacy-preserving algorithm based on clustering proposed by Zhang et al. [9]. But all these algorithms assume the attacker has not fully mastered the sufficient background knowledge. Targeting on the various P2P applications, the attacker can infer the sensitive attributes of the nodes according to the edge weight of the nodes, so that the private information of the nodes is tracked and identified. Based on differential privacy model, this paper proposes a P2P privacy-preserving algorithm which adds noise to the distribution of nodes and solves the high-sensitivity problem caused by P2P itself under the Laplace mechanism. The algorithm guarantees the privacy of the user personal data during the storage and distribution process in P2P, and after the addition of the noise, the sensitivity of the algorithm is controllable and lower, and it is available and reliable when performing P2P credit verification, transmission and storage; moreover, the algorithm guarantees reasonable allocation of privacy budget, while enjoying effectiveness and robustness.

This paper proposed a method utilized differential privacy-preserving model to ensure PNN node privacy, which essentially performs privacy preserving based on the disturbance of node degree distribution. The related work of this paper mainly includes:

(1) According the special structure of P2P, this paper presents a clear definition of differential privacy;

(2) Based on the differential privacy-preserving mechanism, this paper presents mathematical formulas for the differential privacy of P2P;

(3) This paper designs a noise-adding algorithm based on node degree distribution. The key problem of the algorithm is to find a suitable probability distribution to add the noise to control the abnormal query results caused by too-high sensitivity of the mechanism; and this paper proposes the noise distribution and verifies it by experiment.

2. DIFFERENTIAL PRIVACY

Differential privacy realizes data addition or deletion to the dataset by adding random noise to the aggregate query results without affecting the output of the query. Even in worst-case scenario, the attacker knows all sensitive data but one of the records, it could still guarantee that the sensitive information of this record will not be leaked.

Definition 1.1: ϵ-Differential Privacy: $\mathcal{D}$ is a data set, privacy mechanism $\mathcal{M}$ satisfies differential privacy, if and only if for any pair of data sets $\mathcal{D}_1, \mathcal{D}_2 \in \mathcal{D}$, there's one different record at most between datasets $\mathcal{D}_1, \mathcal{D}_2$. For any output, there is $O \subseteq Range(\mathcal{M})$:

$$\Pr[\mathcal{M}(\mathcal{D}_1) \in O] \le \exp(\epsilon) \times \Pr[\mathcal{M}(\mathcal{D}_2) \in O] \tag{1}$$

Namely:

$$\forall O \subseteq Range(\mathcal{M}): \Pr[\mathcal{M}(\mathcal{D}_1) \in O] \le e^{\epsilon} \Pr[\mathcal{M}(\mathcal{D}_2) \in O] \tag{2}$$

Which is equal to:

$$\forall o \subseteq Range(\mathcal{M}): \Pr[\mathcal{M}(\mathcal{D}_1) = o] \le e^{\epsilon} \Pr[\mathcal{M}(\mathcal{D}_2) = o] \tag{3}$$

In Formula 1, Pr represents the privacy risk, namely the probability that the privacy is leaked; ϵ is the privacy budget real number, namely the differential privacy protection strength parameter, and the smaller the value, the greater the privacy protection strength. It can be seen that the ϵ value of the differential privacy controls the similarity of the probability distribution. The smaller the ϵ value, the closer e^{ϵ} is to 1 [10, 11].

Definition 1.2: Global Sensitivity: For query $f: \mathcal{D} \to \mathbb{R}$, then $\Delta f_{GS} = \max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$, Δf_{GS} represents the largest difference between the query results of two adjacent data sets.

Definition 1.3: Local Sensitivity: For query $f: \mathcal{D} \to \mathbb{R}$, then $\Delta f_{LS} = \max_{\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$, the query result set of different records can be corrected by using Δf_{LS}.

Definition 1.4: Smooth Bound: if for $\beta > 0$, function $B: \mathcal{D} \to \mathbb{R}$ is the smooth bound of query f, then Formulas 4 and 5 must be met:

$$\forall \mathcal{D} \in X: B(\mathcal{D}) \ge f_{LS}(\mathcal{D}) \tag{4}$$

$$\forall \mathcal{D}, \mathcal{D}' \in X: B(\mathcal{D}) \ge e^{\beta} S(\mathcal{D}) \tag{5}$$

3. DIFFERENTIAL PRIVACY OF P2P

3.1 Definition and concept

To represent the node individuals in P2P or its networks more conveniently, the P2P in this paper is represented by $G(V,E)$, which denotes an undirected graph. The node set in a P2P is represented by $V=\{v_1,v_2,...,v_n\}$, the relationship between the nodes, namely the edges, are represented by $E = \{(u,v)|u,v \in V\}$, then the degree of a node is the number of nodes adjacent to it, namely the neighbor is $N(v) = \{u|(u,v) \in E, u \ne v\}$, and the degree is $D(v)=|N(v)|$. When a query function f acquires information from a social network graph G, and the obtained result is $f(G)$. To ensure the privacy of P2P, noise of certain distribution features should be added to the result $f(G)$, so that it guarantees the privacy and satisfies the output results without affecting normal requirement.

Definition 2.1: Differential privacy of P2P. In graph G, by deleting an arbitrary node v_i and all the edges connecting to this node, we can get graph G', which is called a neighbor of G. For the query function f of the node, the query results of $f(G)$ and $f(G')$ are almost the same, then it's considered that node v_i satisfies the differential privacy.

If in P2P, f_1 is a method for querying the number of users, then Δf_1 is 1, and the noise of f_1 satisfies the Laplace distribution lap $\left(\frac{1}{\epsilon}\right)$; if in P2P f_2 is a method for querying the number of associated nodes, then Δf_2 is 5, namely the maximum value of the degree of the node; Δf_1 and Δf_2 are quite different, as shown in Figure 1:

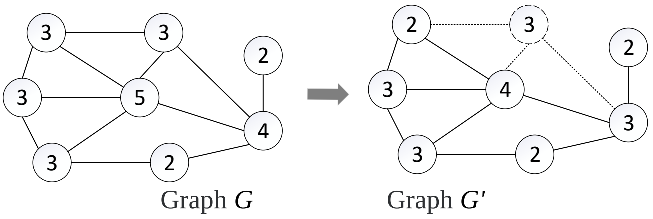


Figure 1. Differential privacy of P2P

Definition 2.2: G is an undirected graph, θ is a given threshold and $\theta<Degree_{max}$, then G_θ is a cut-out set with a degree G of θ.

As shown in Figure 2, graph G_3 is a cut-out graph of G, wherein a node with a degree of 5 is cut out so as to ensure the degrees of all nodes are not greater than 3. In this way, the sensitivity Δf_2 of the query of edges will decrease from 5 to 3.

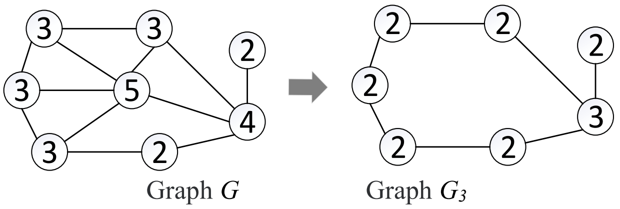


Figure 2. Truncation of graph

3.2 Noise-adding algorithm based on node degree distribution

Hay and Kasiviswanathan et al. first proposed the concept of node difference privacy, and pointed out that deleting a node or an edge affects not only the independent node and the edge, but would result in larger noise and affect the query results [10]. The key problem for P2P to satisfy differential privacy is that when calculating the sensitivity of the original graph and neighbor graph, theoretically the highest sensitivity is $n-1$, but the actual noise is much larger than expected, so it results in abnormal query results [12, 13]. Therefore, this paper proposes a solution for the reduction of Δf_2.

As shown in the above figure, there is a given query f on the cut-out graph G_θ. For a node whose degree changes due to truncation, the local sensitivity depends entirely on the node degree distribution of graph G, for the local sensitivity of each node, selecting properly distributed noise would make it approach to the global sensitivity, which would greatly reduce the local sensitivity. Algorithm 1 is shown as follows.

The algorithm is mainly divided into three steps. First, determine the truncation threshold θ, because the maximum value of the degree of graph G cannot be obtained, and the maximum degree may be very large, θ usually starts with a random value $\hat{\theta}$, therefore, truncation is carried out randomly, and an initial boundary is obtained on this basis. For a given target parameter θ, the algorithm selects a random parameter within the range of bounded constant multiples of θ. And then, the cut-out graph is created by node with a discard degree greater than θ. At last, a certain kind of distributed noise is added to the node degree of the cut-out graph. The truncation of G and conversion of G_θ are relatively difficult, it's because when deleting all nodes whose node degrees are larger than θ, the nodes and edges that are finally deleted are more than expected. Therefore, choosing an appropriate threshold θ for the calculation of degree distribution cut-out graph $G_{\hat{\theta}}$ and smooth bound $S(G_\theta)$ is the key of the algorithm. According to formula 5 of definition 1.4, the modified Algorithm 2 is as follows.

```
Algorithm 1: Noise-adding algorithm based on degree distribution
If there is graph G, privacy budget $\epsilon$, distribution query function f, then:
1. Determine the random truncation threshold $\theta$;
2. Select a distribution to calculate the cut-out graph $G_{\hat{\theta}}$ and the smooth bound $S(G_o)$;
3. Output $\hat{f} = f(G_{\hat{\theta}})$+degree distributions.
```

```
Algorithm 2: Noise-adding algorithm based on degree distribution
If there is graph G, privacy budget $\epsilon$, distribution query function f, then:
1. Determine the random truncation threshold $\theta$, Select $\hat{\theta} \in \left\{D + \frac{\log n}{\beta} + 1, \dots, 2D + \frac{\log n}{\beta}\right\}$;
2. Use $\beta = \frac{\epsilon}{\sqrt{2(\hat{\theta}+1)}}$ to calculate the cut-out graph $G_{\hat{\theta}}$ and the smooth bound $S(G_\theta)$.
3. Output $\hat{f} = f(G_{\hat{\theta}}) + Cauchy\left(\frac{2\sqrt{2}}{\epsilon}\hat{\theta}S(G_\theta)\right)^{\hat{\theta}+1}$.
```

3.3 Selection of random truncation threshold θ

Because the P2P does not take isolated nodes or rings into consideration, most network interactions come from the community structures formed by super-nodes, and the scale of P2P could expand and contract dynamically according to the real-time joining and exiting of nodes, therefore, its degree distribution is significantly dynamic, and most nodes tend to short-term or one-time resource sharing. The interaction between nodes is closely related to the node properties and the network distance between the nodes, thereby a potential field is determined across the whole network.

The topology potential of node v_i in graph G is:

$$\Phi(v_i) = \sum_{j=1}^{n}\left(Resource_j \times e^{-\left(\frac{d(i,j)}{\tau}\right)^2}\right) \tag{6}$$

d(i,j) represents the shortest path value of nodes v_i and v_j, $Resource_j$ represents the health intensity or activity of the node itself, and τ represents the influence of the node. It can be seen that when the distance is greater than $\frac{3\tau}{\sqrt{2}}$, the function decays rapidly to 0.

If B_j is the trust vector of v_i to v_j, then the super-node distribution of the P2P based on topology potential can be expressed as:

$$\Pr(V_i) = \arg\max\left(\sum_{j=1}^{n}\left(B_j \times e^{-\left(\frac{d(i,j)}{\tau}\right)^2}\right)\right) \tag{7}$$

According to formula 6 and formula 7, the modified Algorithm 3 is as follows:

```
Algorithm 3: Noise-adding algorithm based on degree distribution
If there is graph G, privacy budget $\epsilon$, distribution query function f, and trust degree vector B(V), then:
1. Calculate formula 2.4:
2. Determine the random truncation threshold $\theta \in \Pr(V_i)$,  select $\hat{\theta} \in \left\{D + \frac{\log n}{\beta} + 1, \dots, 2D + \frac{\log n}{\beta}\right\}$;
3. Use $\beta = \frac{\epsilon}{\sqrt{2(\hat{\theta}+1)}}$to calculate cut-out graph $G_{\hat{\theta}}$ and smooth bound $S(G_\theta)$
4. Output: $\hat{f} = f(G_{\hat{\theta}}) + Cauchy\left(\frac{2\sqrt{2}}{\epsilon}\hat{\theta}S(G_\theta)\right)^{\hat{\theta}+1}$.
```

4. EXPERIMENT AND ANALYSIS

The experimental computing environment of this paper is a Pentium Xeon E5-2690 V4 CPU with 64GB memory and a NVidia GTX 1080 TI GPU. The experimental data preprocessing was implemented in Python, and the specific algorithm was implemented using C++, mainly ran on a single machine. The data in this paper adopted the P2P dataset in the Stanford Large Network Dataset Collection [14], which contains 4039 nodes and 88234 edges. Figure 3 shows the degree distribution of a data set with a maximum degree of 1045.

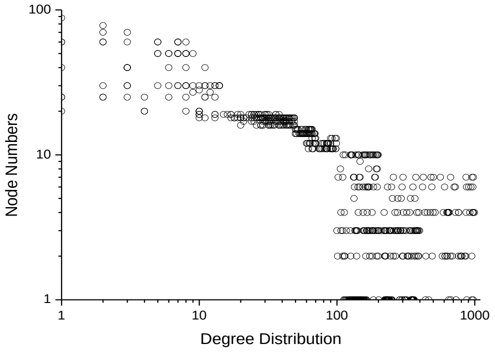
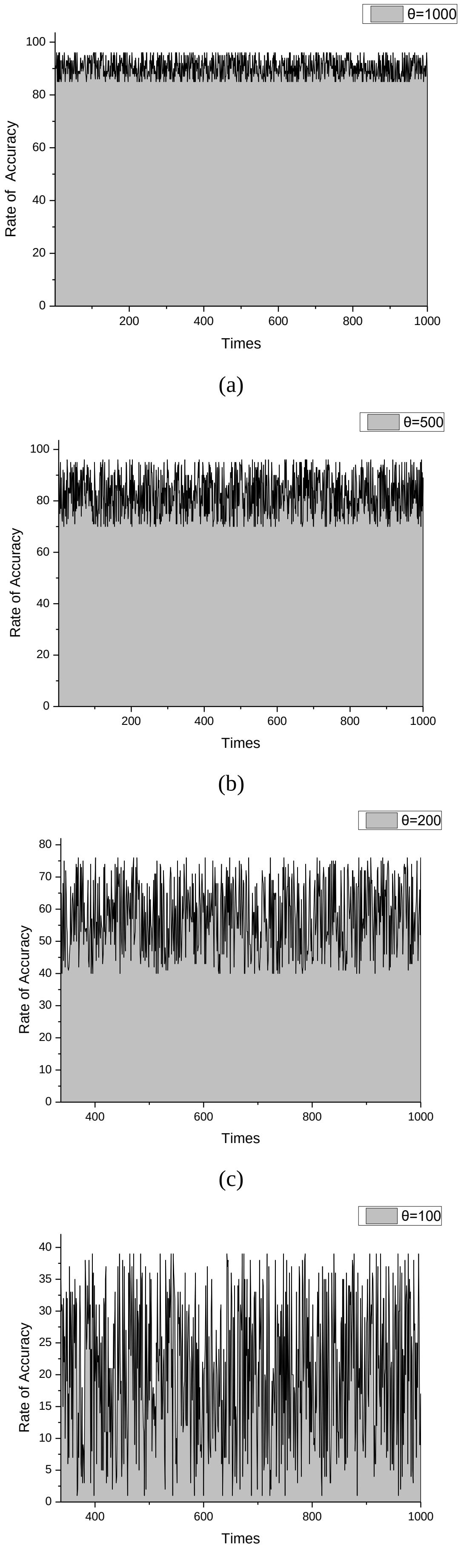


Figure 3. Degree distribution of the graph



(a)

(b)

(c)

(d)

Figure 4. Accuracy rate

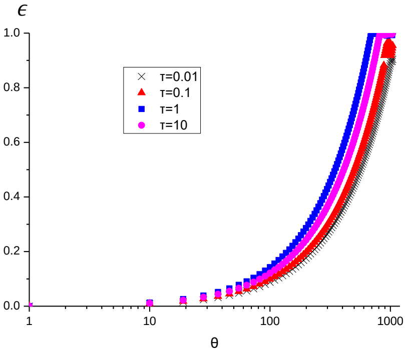


Figure 5. Experimental results of privacy budget

First, for each threshold value θ, the interaction correct rates of P2P were compared, as shown in Figure 4, the selection of θ has a great influence on the accuracy of the algorithm. Then choose *τ=0.01, 0.1, 1, 10* to calculate the corresponding privacy budget ϵ to evaluate the validity and robustness of the privacy preserving of the algorithm, as shown in Figure 5. The experimental results show that the key of the algorithm is to find the optimal threshold θ so that the local sensitivity on the cut-out graph G_θ is closer to the upper smooth bound, and the consumption of privacy budget ϵ of the P2P structure complied with the power-law distribution is minimal.

5. CONCLUSIONS

In various P2P applications, the users' privacy information has been tracked, identified and leaked, to tackle this security problem, this paper analyzed the node data storage and interaction process in P2P, and proposed a P2P privacy-preserving algorithm based on differential privacy model. The algorithm adds noise to the user privacy information and solves the problem that the P2P structure itself has high sensitivity under the Laplace mechanism. The paper used experiment to prove that the algorithm satisfies the differential privacy in the user personal data storage and distribution process in P2P, after the addition of the noise, the data sensitivity becomes lower and the algorithm is available and reliable for the calculation of credit verification, transmission and storage in P2P, and at the same time, the experiment has proved the effectiveness and robustness of the algorithm itself. As the number of interactive applications in P2P is increasing, this paper proposes that the node differential privacy algorithm cannot be applied to interactive and data-related P2P, and our research group will continue to study the differential privacy algorithm for the interactive query in P2P, in the hopes of preventing leakage of privacy data to a greater extent.

ACKNOWLEDGMENT

This work is supported by Guangxi Key Laboratory of Cryptography and Information Security (No. GCIS201720) and Foundation of Improving Academic Ability in University for Young Scholars of Guangxi (No. 2019KY1268).

REFERENCES

[1] Gaboardi, M. (2018). Formal verification of differential privacy. PLAS '18 Proceedings of the 13th Workshop on

Programming Languages and Analysis for Security, Toronto, Canada, 1-1. https://doi.org/10.1145/3264820.3264829

[2] Yu, J., Wang, H. (2018). A deep neural network-based algorithm for safe release of big data under random noise disturbance. Ingénierie des Systèmes d'Information, 23(6): 189-200. https://doi.org/10.3166/ISI.23.6.189-200

[3] Tushar, W., Yuen, C., Mohsenian-Rad, H., Saha, T., Poor, H.V., Wood, K.T. (2018). Transforming energy networks via peer to peer energy trading: Potential of game theoretic approaches. IEEE Signal Processing Magazine, 35(4): 38-40. https://doi.org/10.1109/MSP.2018.2818327

[4] Liao, L., Liu, Z.Y. (2014). Recommendation trust model in P2P networks based on fuzzy comprehensive evaluation. Computer Engineering and Design, 35(4): 1183-1187. http://dx.chinadoi.cn/10.3969/j.issn.1000-7024.2014.04.013

[5] Wang, Y., Hou, J., Bai, Y., Xia, Y., Qin, Z.G. (2013). Survey on feedback correlation based dynamic trust model for P2P systems. Computer Science, 40(2): 103-107. http://dx.chinadoi.cn/10.3969/j.issn.1002-137X.2013.02.023

[6] Wang, D., Long, S.G. (2019). Differential privacy algorithm for privacy protection in weighted social network. Computer Engineering, 45(4): 114-118. http://dx.chinadoi.cn/10.19678/j.issn.1000-3428.0049695

[7] Han, Y. (2015). Sensitive-resisting relation social network privacy protection model. International Journal of Security & Its Applications, 9(8): 195-204. http://dx.doi.org/10.14257/ijsia.2015.9.8.16

[8] Chen, C.L., Xiong, J., Chen, L., Yu, H. (2016). Personalized privacy preservation algorithm in weighted social networks. Computer Technology and Development, 26(8): 88-92. http://dx.chinadoi.cn/10.3969/j.issn.1673-629X.2016.08.019

[9] Zhang, F.X., Jiang, C.H. (2015). Privacy-preserving approach in social networks based on DSNPP algorithm. Computer Technology and Development, 25(8): 152-155. http://dx.chinadoi.cn/10.3969/j.issn.1673-629X.2015.08.032

[10] Li, X.G., Li, H., Li, F.H., Zhu, H. (2018). A survey on differential privacy. Journal of Cyber Security, 3(5): 92-104. http://dx.chinadoi.cn/10.19363/J.cnki.cn10-1380/tn.2018.09.08

[11] Macwana, K.R., Patel, S.J. (2018). Node differential privacy in social graph degree publishing. Procedia Computer Science, 143: 786-793. https://doi.org/10.1016/j.procs.2018.10.388

[12] Liu, A., Xia, L.R., Duchowski, A., Bailey, R., Holmqvist, K., Jain, E. (2018). Differential privacy for eye-tracking data. ETRA '19 Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, Denver, Colorado. https://doi.org/10.1145/3314111.3319823

[13] Li, N.H. (2018). Differential privacy in the local setting. IWSPA '18 Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, Tempe, AZ, USA, pp. 42-44. https://doi.org/10.1145/3180445.3190667

[14] Korolova, A. (2019). Privacy-preserving WSDM. WSDM '19 Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne VIC, Australia, p. 4. https://doi.org/10.1145/3289600.3291385