SCIENCE CHINA
Information Sciences

• LETTER •

November 2017, Vol. 60 119103:1–119103:3
doi: 10.1007/s11432-017-9222-7

APRS: a privacy-preserving location-aware recommender system based on differentially private histogram

Sheng GAO[1*], Xindi MA[2,3], Jianming ZHU[1] & Jianfeng MA[2,3]

[1]*School of Information, Central University of Finance and Economics, Beijing* 100081*, China;*
[2]*School of Computer Science and Technology, Xidian University, Xi'an* 710071*, China;*
[3]*School of Cyber Engineering, Xidian University, Xi'an* 710071*, China*

Received March 10, 2017; accepted July 19, 2017; published online September 12, 2017

Citation Gao S, Ma X D, Zhu J M, et al. APRS: a privacy-preserving location-aware recommender system based on differentially private histogram. Sci China Inf Sci, 2017, 60(11): 119103, doi: 10.1007/s11432-017-9222-7

Dear editor,

Recently, a new paradigm of location-aware recommender system (LARS) has become increasingly popular. Compared with traditional recommendation systems such as collaborative filtering and content-based recommendation, LARS exploits the spatial aspect of ratings for recommendations [1]. However, these rating data contain privacy information such as users' locations and preferences, which enable recommender servers (RS) to easily infer the points of interest of users. Thus, RS can track users or provide users' preferences to advertisers, which would seriously threaten their personal safety.

To solve the privacy issue of LARS, existing studies [2–4] mainly focused on protecting users' location privacy, neglecting the history footprint privacy. However, the reveal of users' history footprints to RS would still cause privacy leakage. Shen and Jin [5] first proposed to perturb users' data by differential privacy on their private devices, which can protect the privacy of both users' location and history data. However, this scheme cannot be used in practice because it is not user friendly. Subsequently, they designed another practical differentially private framework to obfuscate users' privacy data on their own devices [6]. However, the perturbed data retained the category information of users' history footprints, which still could disclose a user's preference.

In this letter, we propose a novel built-in-client mechanism, namely APRS, to obfuscate both users' locations and history data under unreliable RS. In APRS, we introduce the notion of geo-indistinguishability [3] to perturb users' locations to achieve ε_1-differential privacy. To protect users' history data, we first aggregate these data to generate a category histogram and then perturb it to achieve ε_2-differential privacy. By using APRS, RS can efficiently provide service for users without knowing their raw data. Finally, we theoretically prove that our APRS can achieve ε-differential privacy and conduct experiments over a real-world dataset. The evaluation results demonstrate that our APRS can not only strengthen users privacy but also improve the recommendation efficiency without reducing recommendation accuracy.

APRS overview. Our APRS contains two parts: location privacy preservation and history data privacy preservation, depicted in Figure 1. Some preliminaries can be found in Appendix A. To achieve ε-differential privacy, we divide privacy budget

* Corresponding author (email: sgao@cufe.edu.cn)

The authors declare that they have no conflict of interest.

into ε_1 and ε_2, $\varepsilon_1 + \varepsilon_2 = \varepsilon$, for location perturbation and history data perturbations, respectively.

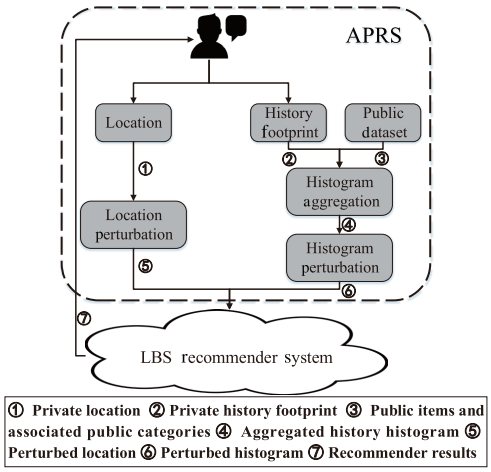

Figure 1 Overall APRS procedure.

Location privacy preservation. We introduce the geo-indistinguishability mechanism [3] to sanitize a user's location. Let the user's current location be $x_u \in \mathbb{R}^2$; a sanitized point $\widetilde{x_u} \in \mathbb{R}^2$ will be generated by the Laplace mechanism and then be sent to RS. Specifically, given the privacy budget $\varepsilon_1 \in \mathbb{R}^+$ and the actual location $x_u \in \mathbb{R}^2$, the probability density function of our noise mechanism in polar coordinates with origin at x_u is $D_{\varepsilon_1}(r,\theta) = \frac{\varepsilon_1^2}{2\pi} r \mathrm{e}^{-\varepsilon_1 r}$, where $\varepsilon_1^2/2\pi$ is the normalization factor, r is the distance $d(x_u, \widetilde{x_u})$ of sanitized location $\widetilde{x_u}$ from x_u, and θ is the angle that the line $x_u\widetilde{x_u}$ creates with respect to the horizontal axis of the cartesian system.

To generate a point (r,θ) from $D_{\varepsilon_1}(r,\theta)$, according to [3], two random variables R and Θ that represent the radius and the angle are independent. Thus, we only need to separately generate r and θ from the marginal probability $D_{\varepsilon_1,R}(r)$ and $D_{\varepsilon_1,\Theta}(\theta)$, respectively. Overall, the Laplacian noise generated in [3] is described as follows:

- Generate θ uniformly in $[0, 2\pi)$.
- Generate p uniformly in $[0,1)$ and set $r = C_{\varepsilon_1}^{-1}(p)$, where $C_{\varepsilon_1}^{-1}(p) = -\frac{1}{\varepsilon_1}(W_{-1}(\frac{p-1}{\mathrm{e}}) + 1)$ and W_{-1} is the Lambert W function (the -1 branch).

Finally, the perturbed location $\widetilde{x_u} = x_u + \langle r\cos(\theta), r\sin(\theta)\rangle$ is sent to RS for recommendation.

History data privacy preservation. To protect users' history data privacy, we first convert these data into a histogram and then adopt a differentially private mechanism to sanitize it. Given a user's raw history item vector d_r, $d_r \in \mathbb{R}^m$, and the public item-category correlation matrix M, we aggregate the raw history data histogram as $H_r = d_r \cdot M$, $H_r \in \mathbb{R}^n$, which represents the user's preference for different categories of items, and each bin H_{ri} represents the user's visit counts for category i. Then, we generate some noise for each bin and make it satisfy ε_2-differential privacy. We first split the privacy budget ε_2 into two portions ε_{2a} and ε_{2b} to sanitize the histogram bins and histogram clusters, respectively.

Histogram bins perturbation. To cluster these histogram bins, we should first sort them based on their counts. However, if we sort them based on their raw counts, the notion of differential privacy will be violated [7]. Therefore, we should first perturb the raw histogram bins by the Laplace mechanism with privacy budget ε_{2a}.

To mitigate the impact of noise, we adopt the row sampling technique [7] to sanitize these raw history footprints and achieve ε_{2a}-differential privacy. To reliably sort two sanitized bins $\widehat{H_{ri}}$ and $\widehat{H_{rj}}$, we should ensure that their true difference must be larger than the magnitude of introduced noise. Therefore, we adopt the ratio of the raw bins' difference $|H_{ri} - H_{rj}|$ to that of the introduced noise, $\frac{|H_{ri}-H_{rj}|}{\sqrt{2}/\varepsilon_{2a}}$, which is treated as an indicator of the sorting quality [8]. According to Corollary 1 in [7], after performing the row sampling technique, the ratio becomes $\Delta = \frac{|H_{ri}-H_{rj}|}{\sqrt{2}/\ln(1+\beta(\exp(\varepsilon_{2a})-1))}$, where β denotes the row sampling probability. Because $\frac{|H_{ri}-H_{rj}|}{\sqrt{2}/\varepsilon_{2a}} > \Delta$ for any $0 < \beta < 1$, we conclude that the sampling leads to more precise sorting.

Now we get the perturbed histogram $\widehat{H_r}$, from which adversaries cannot guess the user's accurate history data under any background knowledge. However, the addition of Laplace noise to the raw histograms will seriously disturb the sorting process. For example, those counts of bins with zero in the raw histograms H_r will be artificially changed into non-zero in $\widehat{H_r}$. To alleviate the influence of noise, we adopt a threshold strategy, defined as $\widehat{H_{ri}} = \begin{cases} \widehat{H_{ri}}, & \text{if } \widehat{H_{ri}} \geqslant \lambda \\ 0, & \text{otherwise} \end{cases}$, where $\lambda = \eta\log(n)/\varepsilon_{2a}$ and η is a revision parameter. By using the threshold strategy, the introduced noise for relatively small counts can be smoothed and a more accurate sorting result can be obtained.

Histogram bins clustering. After sanitizing the histogram bins, we sort and cluster the perturbed histogram on the smoothed $\widehat{H_r}$. Because the sorting operation is conducted on differentially private histogram bins, it does not violate the differential privacy and reveal any extra privacy any more.

Next, we will cluster the bins over the sorted sanitized histogram $\widehat{H_r}$. To obtain an optimal cluster set, we define an error for the cluster

S_i as $\mathrm{err}(S_i) = \mathrm{E}(\sum_{\widehat{H_{rj}} \in S_i} (\widehat{H_{rj}} - \widetilde{H_j})^2)$, where $\widetilde{H_j}$ is the final disclosed histogram bin for category j. If $\widehat{H_{rj}} \in S_i$, we can compute $\widetilde{H_j} = \overline{S_i} + \frac{\mathrm{Lap}(1/\varepsilon_{2b})}{|S_i|}$, where $\overline{S_i} = \frac{\sum_{\widehat{H_{rj}} \in S_i} \widehat{H_{rj}}}{|S_i|}$ and $|S_i|$ is the number of histogram bins in i-th cluster. Thus, the error for cluster S_i can be further derived as $\mathrm{err}(S_i) = \mathrm{E}(\sum_{\widehat{H_{rj}} \in S_i} (\widehat{H_{rj}} - \widetilde{H_j})^2) = \sum_{\widehat{H_{rj}} \in S_i} (\widehat{H_{rj}} - \overline{S_i})^2 + \frac{2}{|S_i|(\varepsilon_{2b})^2}$.

Now, we use a greedy clustering algorithm to obtain the optimal cluster set. The main idea is as follows: during the clustering process, we iteratively judge whether to put the next bin $\widehat{H_{rj}}$ into the current cluster S_i. If adding $\widehat{H_{rj}}$ to S_i results in a lower error, we will merge $\widehat{H_{rj}}$ into S_i; otherwise, a new cluster will be created. Due to space limitations, we provide the detailed algorithm in Appendix B.

Privacy and utility analysis. The privacy of our proposed APRS depends on the privacy of its components. We can theoretically prove that the location perturbation and history data perturbation are ε_1-differential privacy and ε_2-differential privacy, respectively. Thus, our APRS satisfies ε-differential privacy. In addition, we have showed the utilities achieved by location perturbation and history data perturbation in theory, respectively. Detailed analyses are presented in Appendix C.

Experimental evaluation. We conduct a series of experiments to evaluate the effectivity and efficiency of our APRS, and compare our method with the most related scheme S-EpicRec [6] over a real-world dataset. Firstly, we measure the radius of retrieval area and the number of businesses located in the area of retrieval varying with the privacy budget ε_1. The result proves that a less ε_1 will add more noise to the secret location, affecting the system availability. Then, we evaluate the history data perturbation in APRS from the respect of perturbed category aggregates quality and recommendation accuracy. Compared with S-EpicRec [6], our APRS shows a lower expected mean absolute error and a similar recommendation accuracy while using the same privacy budget. Finally, the simulation results also show that our APRS is more efficient than S-EpicRec [6]. Due to space limitations, detailed results and analysis can be found in Appendix D.

Conclusion. In this letter, we propose a novel solution, called APRS, to address the privacy issue in LARS. In APRS, we introduce the notion of geo-indistinguishability to perturb a user's current location and design a differentially private histogram to perturb the user's history data. Theoretical analysis and experimental results demonstrate that our APRS can not only strengthen users' privacy but also improve the recommendation efficiency without reducing recommendation accuracy.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61602537, 61272398, U1509214, U1405255), Beijing Municipal Philosophy and Social Science Foundation (Grant No. 16XCC023), and National High Technology Research and Development Program of China (863 Program) (Grant No. 2015AA016007).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Sarwat M, Levandoski J J, Eldawy A, et al. Lars*: an efficient and scalable location-aware recommender system. IEEE Trans Knowl Data Eng, 2014, 26: 1384–1399
2 Scipioni M P. Towards privacy-aware location-based recommender systems. In: Proceedings of IFIP Summer School, Trento, 2011. 1–8
3 Andrés M E, Bordenabe N E, Chatzikokolakis K, et al. Geo-indistinguishability: differential privacy for location-based systems. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security, Berlin, 2013. 901–914
4 Ma X D, Li H, Ma J F, et al. Applet: a privacy-preserving framework for location-aware recommender system. Sci China Inf Sci, 2017, 60: 092101
5 Shen Y L, Jin H X. Privacy-preserving personalized recommendation: an instance-based approach via differential privacy. In: Proceedings of IEEE International Conference on Data Mining, Shenzhen, 2014. 540–549
6 Shen Y L, Jin H X. Epicrec: towards practical differentially private framework for personalized recommendation. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, 2016. 180–191
7 Kellaris G, Papadopoulos S. Practical differential privacy via grouping and smoothing. Proc VLDB Endowment, 2013, 6: 301–312
8 Zhang X J, Chen R, Xu J L, et al. Towards accurate histogram publication under differential privacy. In: Proceedings of the 2014 SIAM International Conference on Data Mining, Pennsylvania, 2014. 587–595