Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Differential Privacy – A Privacy Guaranteed Publishing of Search Logs

R.Ashwini[*], K.Praveen, R.V.Krishnaiah

Department of CSE & JNTUH

India

Abstract - Real time search engines like Google collect users' search histories with respect to their queries and clicks. The search logs are valuable to the researchers in data mining domain. However, search engines generally do not publish search logs as they are concerned about the disclosure of the sensitive information. This paper focuses on search logs and analyzes clicks, queries, and frequent keywords. The experiments revealed that k-anonymity variants are vulnerable to various attacks;stronger privacy is guaranteed with e-differential privacy which doesn't have utility for search logs. Therefore we proposed a new algorithm for achieving guaranteed privacy. We have developed a prototype web application to test the effectiveness of the proposed algorithm. The results revealed that the proposed application is useful in the real world publication of search logs.

Index Terms –Integrity, privacy, web search engines, k-anonymity, search logs

I. Introduction

Privacy to the published search logs has got so much importance though such data is very useful for researchers. The security concern in publishing such search logs is that the sensitive data might be disclosed. There are many real world search engines that constantly generate search logs. However, publishing such search logs is not generally done by them due to privacy concern. Today's search engines have lot of information associated with search queries. There might be sensitive information that can be obtained from the logs which are published. Search logs are valuable sources for researchers as they can obtain trends or patterns from the search logs using data mining techniques. Such research also helps in improving search performance and quality of the search. For this reason the researchers across the world treat the search logs as goldmine for their research. However, as said earlier, the search engines are not in favor of publishing their search logs due to the security of the sensitive information present there in the search logs. As an exception to this in 2006, AOL published its search logs. The AOL could not provide complete security to the published search logs. Only precaution to show was to replace some sensitive fields with random numbers. That protection was no sufficient and many users' privacy was lost. For instance Georgia [1] has lost his valuable information as it was disclosed to public. Information disclosure was not guaranteed by AOL just by replacing some fields with numbers. The problem with the published search logs is that, hackers can establish the original identity of the user just by using some of the fields like age, zip code etc [2], [3].

This paper compares various methods of limiting the information disclosure when search logs are published containing clicks, queries and frequent keywords. The methods we studies vary in giving protection to search logs. The research revealed that the existing k-anonymity [4] and other techniques [5], [6], [7], [8] are inadequate in stopping information disclosure. When compared to these techniques, the differential privacy [9] is much better but it is not suitable for search logs. We then implement the algorithm originally proposed by Korolova et al. [10], [11] where relaxation of differential privacy is proposed. Out paper have extensive evaluations as it compares k-anonymity, e-differential privacy and also the proposed algorithm. Strong privacy guarantee is achieved with proposed algorithm. After witnessing the results we believed that the proposed algorithm can be used by search engines in case if they want to publish search logs with privacy guarantees. With existing attacks [12] the proposed algorithm was tested. The results revealed that it is robust to such attacks.

The rest of this paper is organized as follows. Section II focuses on review of literature. Section III gives information about the proposed algorithm and other essential details or privacy guarantees. The section IV provides details about the experimental results while section V concludes the paper.

II. Related Work

This section reviews the literature on the relevant topics pertaining to publishing search logs and the security concerns. There has been more research found in literature on achieving anonymity in search logs [5], [7], [6], [8]. The most recent work on this area is from Korolova et al. [10] where a basic algorithm for preventing information disclosure was developed. We continue on this work by refining the algorithm which ensures privacy guarantees. From the literature it is understood that k-anonymity can't give privacy guarantee. Its variants also failed to do so. Therefore we studied e-differentialprivacy algorithm which is much better than that of k-anonymity. However, unfortunately it does not have any utility with publishing search logs. To overcome this problem, in this paper we proposed a new algorithm for publishing search logs with privacy guarantees. The results are also compared with k-anonymity and e-differential privacy concept.

III. Proposed Algorithm and Comparison with Others

Before implementing a new algorithm, we studied both k-anonymity variants and also the e-differential privacy. The negative results we obtained are described here. The problem with k-anonymity and its variants is that they can prevent an attacker from identifying a user uniquely. However, they fail in preventing such attackers to get information such as key words that can be used to establish original identity of the users. On the other hand the differential privacy has been proved to be much more secure. Unfortunately, it does not have the utility to serve the publishing search logs with privacy guarantees. Therefore we proposed a new algorithm which is presented in listing 1.

Proposed Algorithm

The listing 1 shows the pseudocode for the proposed algorithm which is aimed at publishing search logs with privacy guarantees.

```
Input: Search log S, positive numbers m, $\lambda$,T,T'
1. For each user u select a set $s_u$ of up to m distinct items from u's search history in $S_3$.
2. Based on the selected items, create a histogramconsisting of pairs (k; $c_k$), where k denotes an item and$c_k$ denotes the number of users u that have k in theirsearch history $s_u$. We call this histogram the original histogram.
3. Delete from the histogram the pairs (k; $c_k$) with count $c_k$smaller than T.
4. For each pair (k; $c_k$) in the histogram, sample a randomnumber _k from the Laplace distribution Lapð_Þ,4 andadd nk to the count $c_k$, resulting in a noisy count:$c_k$          $c_k$+$n_k$
5. Delete from the histogram the pairs (k; $c_k$) with noisy counts ~ck <T'.
6. Publish the remaining items and their noisy counts.
```

Listing 1 –Algorithm for Publishing Search Logs

The algorithm presented in listing 1 ensures privacy. To achive this it follows two phases. In the first phase it generates histogram of items and prunes the items which are below given thrershold. In the second phase, the algorithm intentionally adds noise and eliminates the items which are having noise which is less than the given threshold. Finally it retuns the resultant histogram. The search logs published with this algortim provide higher ssecurity as they can't be attacked by hackers. They are robust to attacks to reveal sensitive information.

IV. Implementation and Results

The implementation has been done using Microsoft .NET framework where the technologies like AJAX and ASP.NET are used to build web interface which is very dynamic in nature. The functionality of the algorithm is done using C# programming language. The environment used for this application development is a PC with 4 GB or RAM with Core 2 Dual processor. The proposed algorithm just relaxes the original e-differential privacy to make it suitable for publishing search logs. The experimental are done with search logs in terms of query pairs, clicks, queries and frequent keywords. The summary of them are presented in table 1.

Table 1- Distinct Item Counts

M	1	4	8	20	40
Keywords	6663	6041	5370	4061	2963
Queries	333o	2082	1425	748	405
Clicks	2812	1574	1001	482	248
Query pairs	329	164	100	38	12

As can be seen in table 1, the distinct item counts in the given dataset with respect to keywords, clicks, queries and query pairs. The following figures shows the evaluation of the results of k-anonymity and the proposed algoirtm which is an improved form of differential privacy.

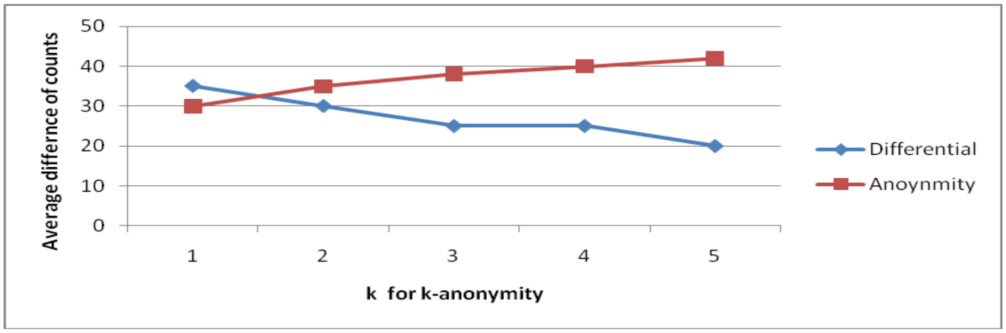

Fig. 1 – Average difference between counts with respect to keywords

As can be seen in fig. 1, the horizontal axis repreents k value used in k-anonymity while the vertical axis represents average difference of count. The results reveal that the proposed algorithm outperforms the k-anoymity.

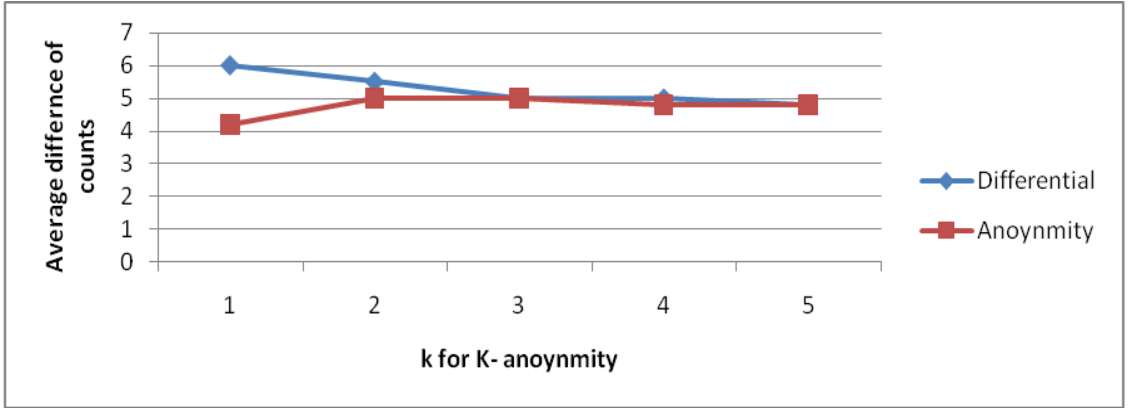

Fig. 2 – Average difference between counts with respect to queries

As can be seen in fig. 2, the horizontal axis repreents k value used in k-anonymity while the vertical axis represents average difference of count. The results reveal that the proposed algorithm outperforms the k-anoymity.

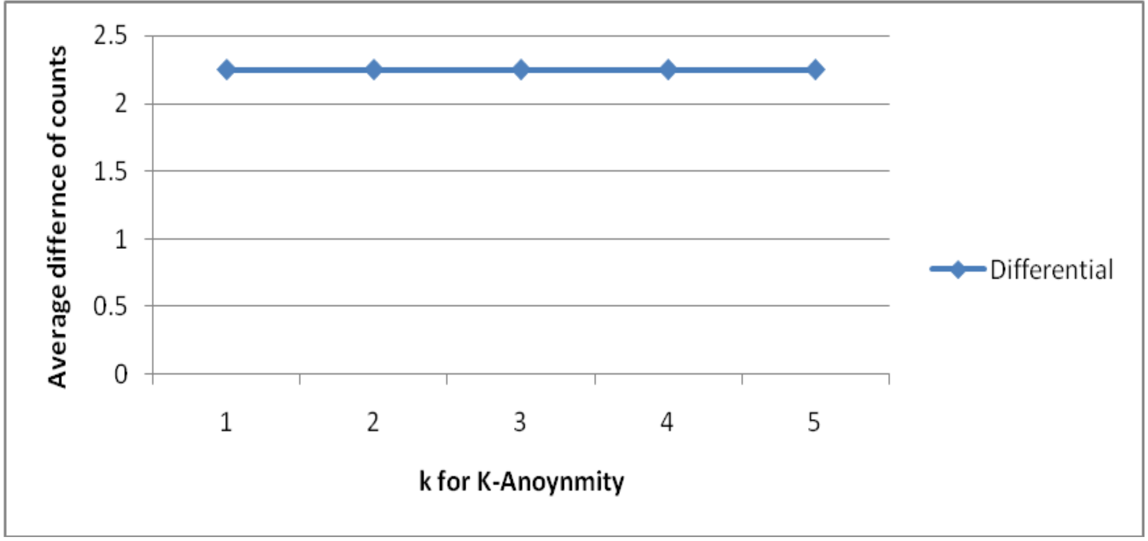

Fig. 3 – Average difference between counts with respect to clicks

As can be seen in fig. 3, the horizontal axis repreents k value used in k-anonymity while the vertical axis represents average difference of count. The results reveal that the proposed algorithm outperforms the k-anoymity.

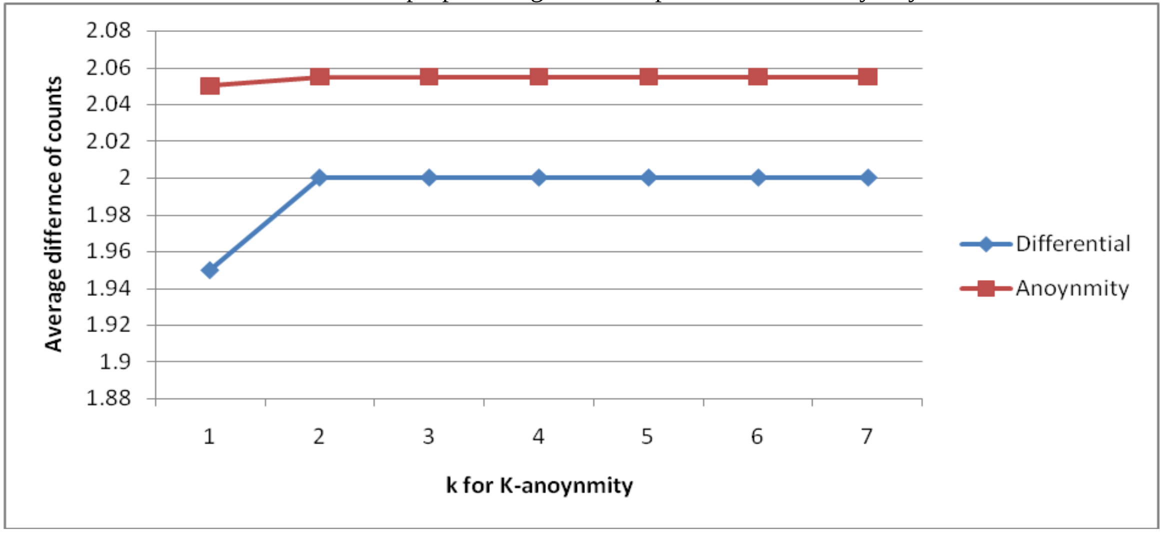

Fig. 4 – Average difference between counts with respect to query pairs

As can be seen in fig. 4, the horizontal axis represents k value used in k-anonymity while the vertical axis represents average difference of count. The results reveal that the proposed algorithm outperforms the k-anoymity.

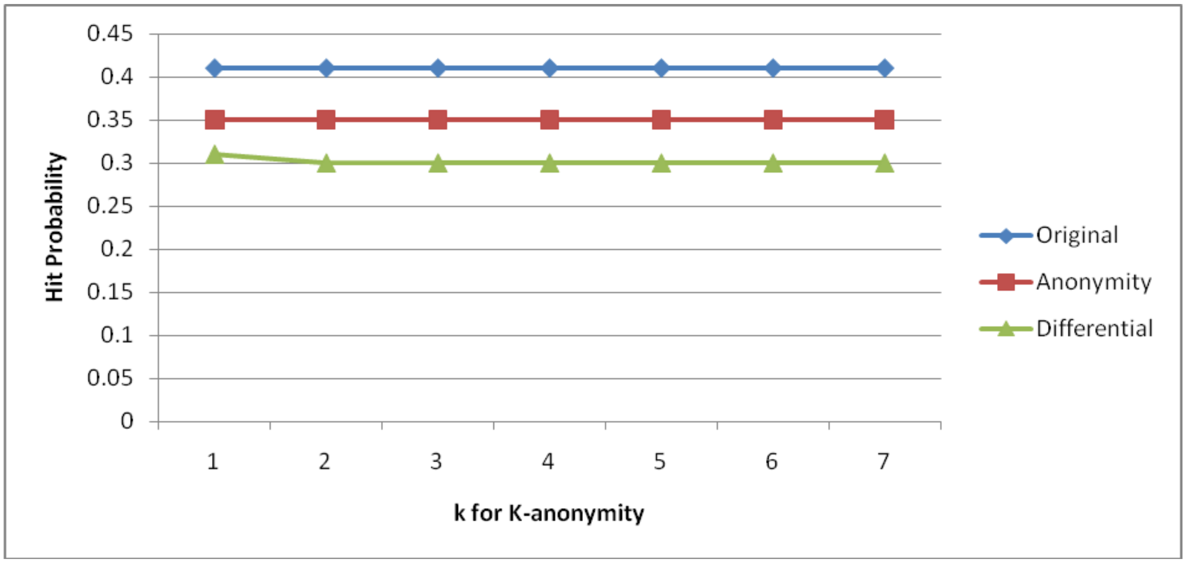

Fig. 5 –Hit Probabilities

As can be seen in fig. 5, the horizontal axis represents k value used in k-anonymity while the vertical axis represents hit probability. The results reveal that the proposed algorithm outperforms the other algorithms.

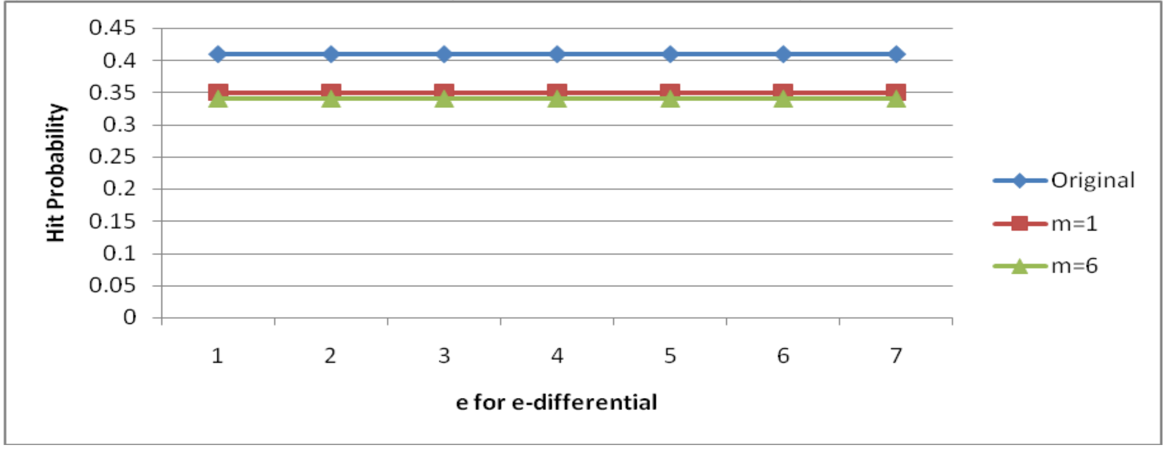

Fig. 6 – Hit Probabilities

As can be seen in fig. 6, the horizontal axis represents e value used in e-differential while the vertical axis represents hit probability. The results reveal that the proposed algorithm outperforms the other algorithms.

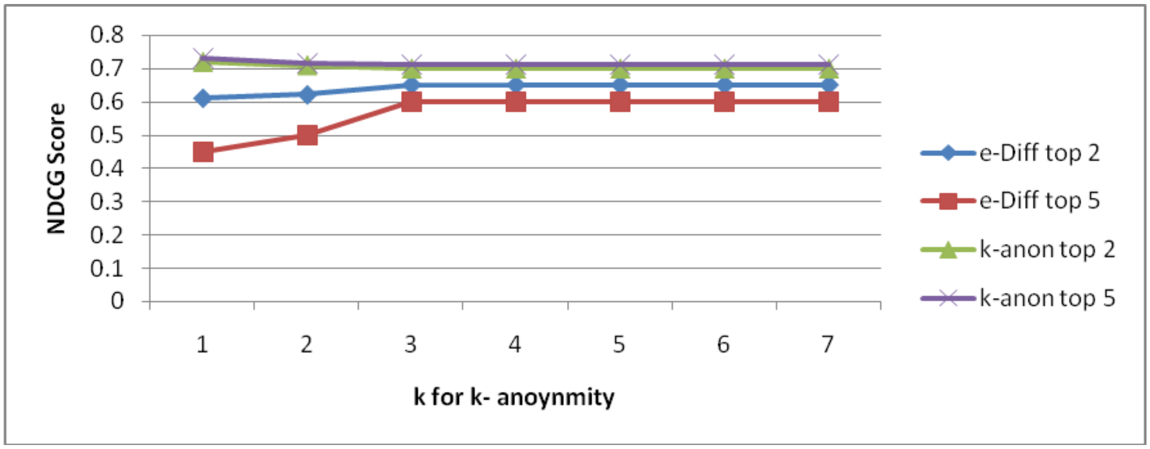

Fig. 7 – Quality measured with NDCG

As can be seen in fig. 7, the horizontal axis represents k value used in k-anonymity while the vertical axis represents NDCG score. The results reveal that the proposed algorithm outperforms the other algorithms.

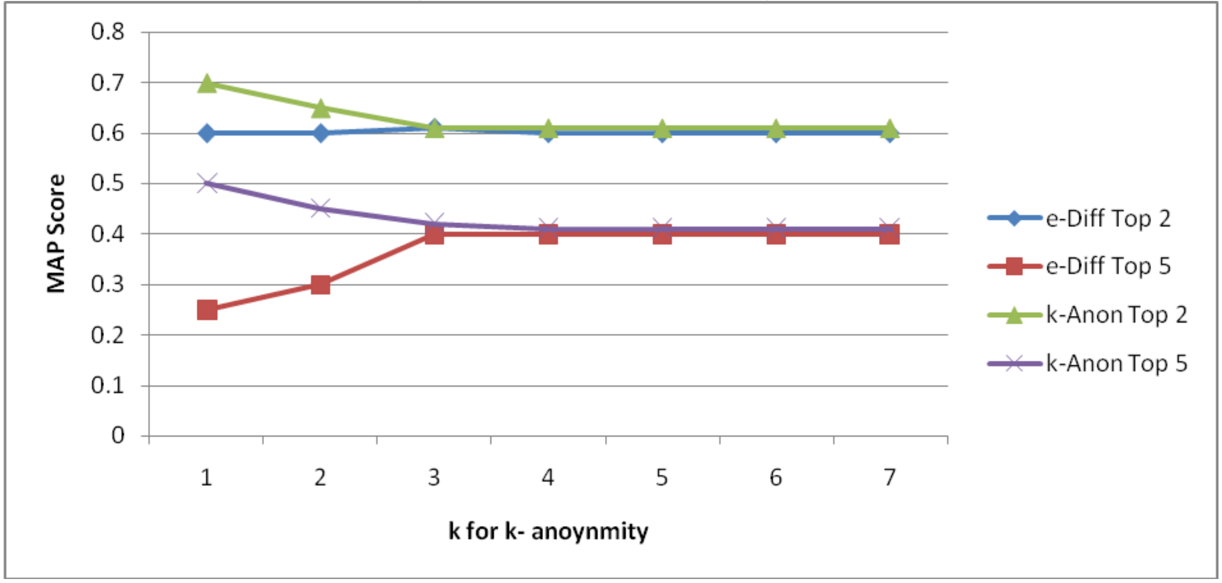

Fig. 8 – Quality measured with MAP

As can be seen in fig. 8, the horizontal axis represents k value used in k-anonymity while the vertical axis represents MAP score. The results reveal that the proposed algorithm outperforms the other algorithms.

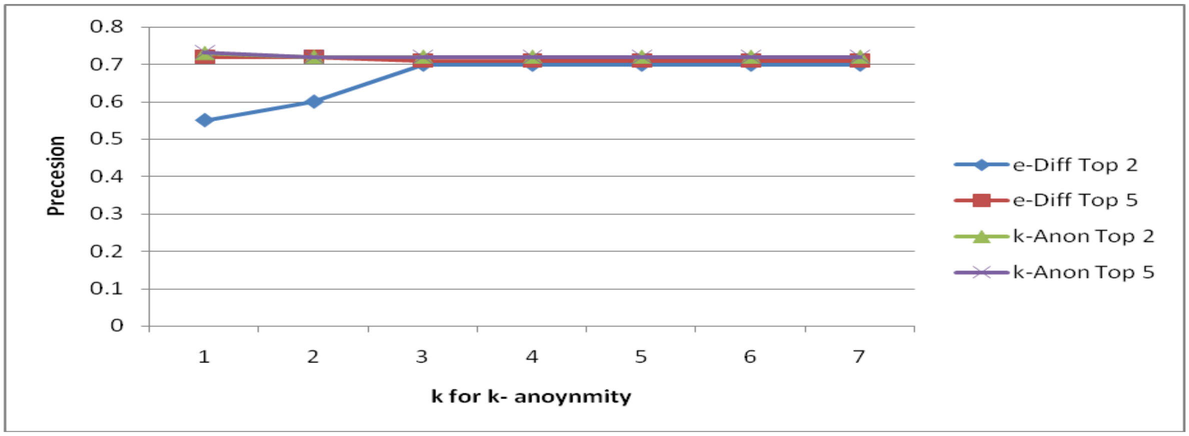

Fig. 9 – Quality measured with Precision

As can be seen in fig. 9, the horizontal axis represents k value used in k-anonymity while the vertical axis represents Precision score. The results reveal that the proposed algorithm outperforms the other algorithms.

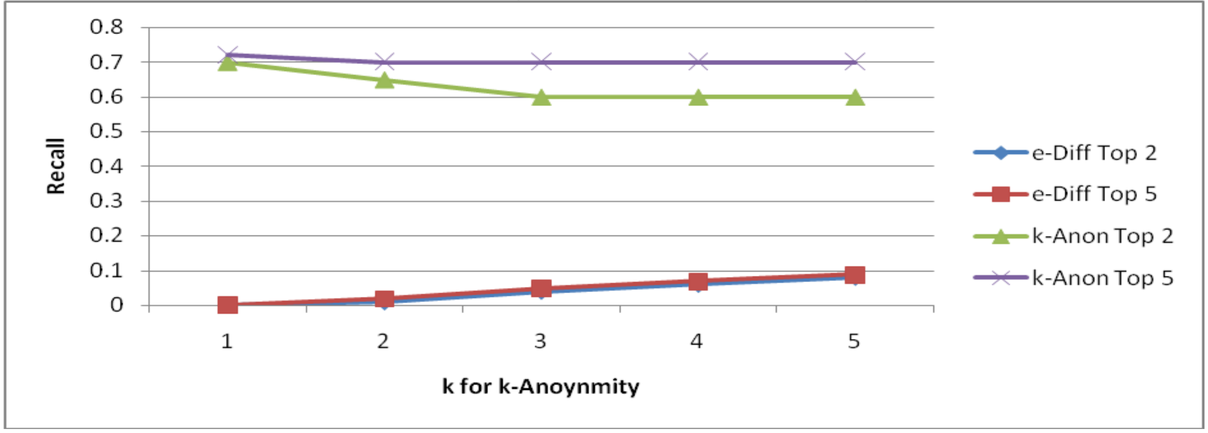

Fig. 10 – Quality measured with Recall

As can be seen in fig. 10, the horizontal axis represents k value used in k-anonymity while the vertical axis represents Recall score. The results reveal that the proposed algorithm outperforms the other algorithms.

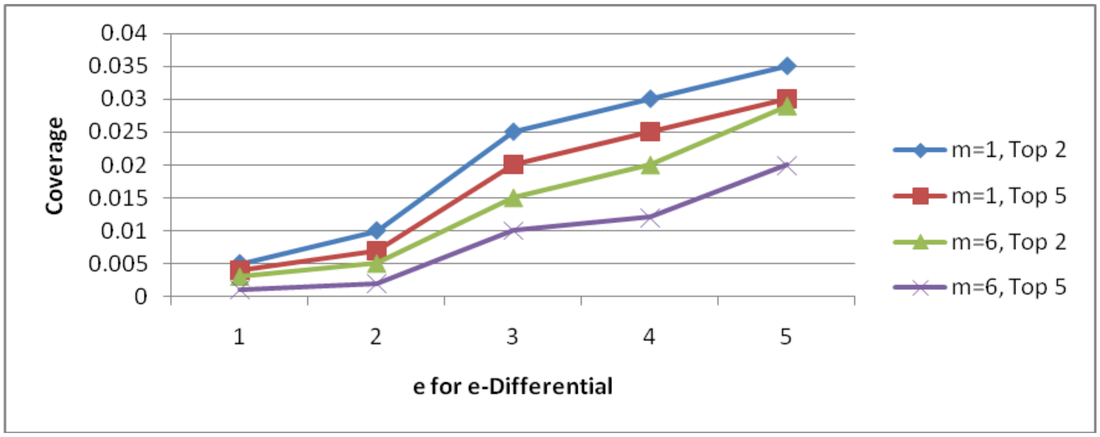

Fig. 11 –Coverage of the privacy-preserving histograms

As can be seen in fig. 11, the horizontal axis represents e value used in e-differential while the vertical axis represents coverage. The results reveal that the larger values of m reflected less coverage which is as expected by this paper.

V. Conclusion

This paper explored the privacy preserving publishing of search logs containing clicks, queries and frequent keywords. Various approaches such as k-anonymity variants and e-differential privacy were compared and analyzed. The experiments revealed that k-anonymity has various vulnerabilities while the e-differential privacy has better privacy guarantees but not suitable for search logs. Then we have proposed our own algorithm for privacy preserving publishing of search logs. The empirical results revealed that the proposed application with new algorithm for publishing search logs with privacy guarantees is able to ensure privacy of sensitive data.

References

[1] M. Barbaro and T. Zeller, "A Face is Exposed for AOL SearcherNo. 4417749," New York Times, http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000en= f6f61949c6da4d38ei=5090, 2006.

[2] R. Jones, R. Kumar, B. Pang, and A. Tomkins, "I Know What YouDid Last Summer: Query Logs and User Privacy," Proc. ACMConf. Information and Knowledge Management (CIKM), 2007.

[3] R. Kumar, J. Novak, B. Pang, and A. Tomkins, "On AnonymizingQuery Logs via Token-Based Hashing," Proc. Int'l Conf. WorldWide Web (WWW), 2007.

[4] P. Samarati, "Protecting Respondents' Identities in MicrodataRelease," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6,pp. 1010-1027, Nov./Dec. 2001.

[5] E. Adar, "User 4xxxxx9: Anonymizing Query Logs," Proc. WorldWide Web (WWW) Workshop Query Log Analysis, 2007.

[6] R. Motwani and S. Nabar, "Anonymizing Unstructured Data,"Corr, abs/0810.5582, 2008.

[7] Y. He and J.F. Naughton, "Anonymization of Set-Valued Data viaTop-Down, Local Generalization," Proc. VLDB Endowment, vol. 2,no. 1, pp. 934-945, 2009.

[8] Y. Hong, X. He, J. Vaidya, N. Adam, and V. Atluri, "EffectiveAnonymization of Query Logs," Proc. ACM Conf. Information andKnowledge Management (CIKM), 2009.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "CalibratingNoise to Sensitivity in Private Data Analysis," Proc. Theory ofCryptography Conf. (TCC), 2006.

[10] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas,"Releasing Search Queries and Clicks Privately," Proc. 18th Int'lConf. World Wide Web (WWW), 2009.

[11] M. Go¨ tz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke,"Privacy in Search Logs," CoRR, abs/0904.0682v2, 2009.

[12] Y. Luo, Y. Zhao, and J. Le, "A Survey on the Privacy PreservingAlgorithm of Association Rule Mining," Proc. Int'l Symp. ElectronicCommerce and Security, vol. 1, pp. 241-245, 2009.

AUTHORS

	Ashwini is student of DRK Institute of Science and Technology, Hyderabad, AP, INDIA. She has received B.Tech Degree computer science and engineering, M.Tech Degree in computer science and engineering. Her main research interest includes data mining, Databases and DWH.
	K.Praveen is working as an Associate Professor in DRK Institute of Science and Technology, JNTUH, Hyderabad, Andhra Pradesh, India.He has completed M.Tech (C.S.E) from Osmania University, Hyderabad. His main research interest includes Databases, Web Methods and Computer Networks.
	Dr.R.V.Krishnaiah (Ph.D) is working as Principal at DRK INSTITUTE OF SCINCE & TECHNOLOGY, Hyderabad, AP, INDIA. He has received M.Tech Degree EIE and CSE. His main research interest includes Data Mining, Software Engineering.