SECURED RELEASE OF SEARCH QUERIES, CLICKS AND COUNTS

M.M.Ramya*

Mrs. R.Medona Selin, M.E**

***Abstract*:**

Text mining refers to extracting or mining information from large amounts of text based documents and the security deals with safeguarding the useful informations. Search engines such as Bing, Google, or Yahoo log interactions with their users. When a user submits a query and clicks on one or more results, a new entry is added to the search log. However these search engine companies are aware of publishing search logs in order not to disclose sensitive information. The goal of this project is to publish frequent items (utility) without disclosing sensitive information about the users (privacy).The existing proposals to achieve k-anonymity in search logs are insufficient in the light of attackers who can actively influence the search log and its impossibility to achieve good utility with differential privacy. This project propose a novel algorithm ZEALOUS and show how to set its parameters to guarantee probabilistic differential privacy. It compares the probabilistic differential privacy with indistinguishablity. This project concludes with an extensive experimental evaluation, by comparing the utility of various algorithms that guarantee anonymity or privacy in search log publishing. This evaluation includes applications that use search logs for improving both search experience and search performance, and the results show that 'ZEALOUS' output is sufficient for these applications by achieving strong formal privacy guarantees.

General Terms: Security

Keywords: Privacy

* M.M.Ramya doing final ME-CSE in Vins Christian College Of Engineering, Chunkankadai

** Mrs R. Medona Selin, M.E, Assistant Professor, Vins Christian College Of Engineering, Chunkankadai.

I.INTRODUCTION

Web search logs collect queries and clicks of users as they interact with a search engine. These logs have been successfully used by search engines in order to improve the quality of search results. Of all the data collected by search engines, [1]the queries submitted by users are among the most valuable. Indeed, query logs can give great insight into human intent and have numerous diverse uses. However, they also have immense potential for misuse. Publishing of user query logs has become a sensitive issue. Similar to other data releases involving individual records such as microdata, the potentially personal content of query logs raises genuine privacy concerns.

In particular, the recent incident involving AOL has increased the public awareness of how the information in the query log file can be used to profile a single user without their knowledge. On the other hand, these search data, if published are, invaluable for researchers and law enforcement Thus, the challenge is to develop anonymization methods to publish query log data without breaching privacy or diminishing utility.

The open question to date is if there even exists a way to publish search logs in a perturbed fashion in a manner that is simultaneously useful and private. At first blush, the problem seems deceptively easy: why not just replace usernames with random identifiers? This simplistic view led to an AOL data release in 2006 in which the searches of an innocent citizen were quickly identified by a newspaper journalist . As a consequence of releasing this private data set the CTO of AOL resigned, two employees were fired, a class action lawsuit is pending.

In October of 2006, Netflix announced the $1-million Netflix Prize for improving their movie recommendation system. As a part of the contest Netflix publicly released a dataset containing 100 million movie ratings created by 500, 000 Netflix subscribers over a period of 6 years. Once again, a simplistic anonymization procedure of replacing usernames with random identifiers was used prior to the release. Nevertheless, it was shown that 84% of the subscribers could be uniquely identified by an attacker who knew 6 out of 8 movies that the subscriber had rated outside of the top 500 .The commonality between the AOL and Netflix datasets is that each individual's data is essentially a set of items. Further this set of items is both identifying of the individual as well as private information about the individual, and partial knowledge of this set of items is used in the privacy attack. In the case of the Netflix data (representative of market-

basket data), for instance, it is the set of movies that a subscriber rated, and in the case of the AOL data, it is the set of queries that a user posed, also called the *user session*

Following the AOL query log incident, recently there has been work on anonymizing query logs. Kumar et al. propose a token- based hashing approach to anonymize query logs. In particular, each query in the log is tokenized, and then a secure hash function applied to produce hashes that are inverted. However, inversion cannot be done using just the token frequencies. It turns out that their token-based approach does not work since serious leaks are possible even when the order of tokens is hidden

This paper first describe two negative results. The existing proposals to achieve k-*anonymity* in search logs are insufficient in the light of attackers who can actively influence the search log. and *differential privacy* ,a much stronger privacy guarantee ;however ,it is impossible to achieve good utility with differential privacy.

The proposed algorithm ZEALOUS achieves more privacy by providing (δ)-indistinguishability and probabilistic differential privacy.

II LIMITATIONS FOR PUBLISHING SEARCH LOGS IN EXISTING SYSTEM:

k-anonymity:- .

A search log is k-anonymous if the search history of every individual is indistinguishable from the history of at least k - 1 other individuals in the published search log.

ϵ −differential privacy:-

An algorithm A is differentially private if for all search logs S and S_0 differing in the search history of a single user and for all output search logs O:

$pr[A(S)=O] \leq e^{\epsilon} pr[A(S_0)=O]$.

This means that for any two datasets which are close to one another (that is, which differ on a single element) a given differentially private algorithm A will behave approximately the same on both data sets. The definition gives a strong guarantee that presence or absence of an individual will not affect the final output of the query significantly.There are infeasibility of differential privacy in search log publication. In particular, under realistic settings, no differentially private algorithm can produce a sanitized search log with reasonable utility.

III PROPOSED SYSTEM:

In the proposed system a search log publishing algorithm called ZEALOUS has been introduced. ZEALOUS ensures probabilistic differential privacy, and it follows a simple two-phase framework. In the first phase, ZEALOUS generates a histogram of items in the input search log, and then removes from the histogram the items with frequencies below a threshold. In the second phase, ZEALOUS adds noise to the histogram counts, and eliminates the items whose noisy frequencies are smaller than another threshold. The resulting histogram (referred to as the *sanitized* histogram) is then returned as the output. Figure 1 depicts the steps of ZEALOUS.

ZEALOUS Algorithm works as follows:

- For each user u select a set of distinct items from search history.
- Based on the selected items, create a histogram consisting of pairs (k, ck), where k denotes an item and ck denotes the number of users that have . This histogram is called as the original histogram.
- Delete from the histogram the pairs (k, c_k) with count c_k smaller than a threshold τ.
- For each pair (k, c_k) in the histogram, sample a random number ηk from the Laplace distribution , and add ηk to the count c_k, resulting in a noisy count
- Delete from the histogram the pairs with noisy counts less than second threshold τ'
- Publish the remaining items.

To understand the purpose of the various steps one has to keep in mind the privacy guarantee we would like to achieve. Step 1., 2. and 4. of the algorithm are fairly standard. It is known that adding Laplacian noise to histogram counts achieves differential privacy .However, these steps alone result in poor utility because for large domains many infrequent items will have high noisy counts.

To deal better with large domains we restrict the histogram to items with counts at least τ in Step 2. This restriction leaks information and thus the output after Step 4 is not $\in$differentially

private. One can show that it is not even probabilistic differentially private (for $\delta < 1/2$).Step 5. disguises the information leaked in Step 3. In order to achieve probabilistic differential privacy.

The performance of ZEALOUS is measured in terms of both privacy and utility. The ZEALOUS algorithm provides both indistinguishability and probabilistic differential privacy.

The project allows only the authorized users to view the sensitive information's. The following section explains about:

- Query Substitution
- Index Caching
- Item Set Generation and Ranking

Query Substitution:

Query substitutions are suggestions to rephrase a user query to match it to documents or advertisements that do not contain the actual keywords of the query. Query substitutions can be applied in query refinement, sponsored search, and spelling error correction.

Query substitution as a representative application for search quality. Query substitution requires two steps. First, the query is partitioned into subsets of keywords, called phrases, based on their mutual information.

Next, for Each phrase, candidate query substitutions are determined based on the distribution of queries.

Quality of query substitution is measured using:

- Precision
- Recall
- Mean average precision

Precision:

The precision of a query q is the fraction of substitutions from the sanitized search log that are also contained in ground truth ranking.

Precision(q)=$\frac{|\{q_0,\ldots\ldots q_{j-1}\}\cap\{q'_0\ldots\ldots,q'_{j-1}\}|}{|\{q'_0,\ldots\ldots\ldots\ldots q'_{j-1}\}|}$

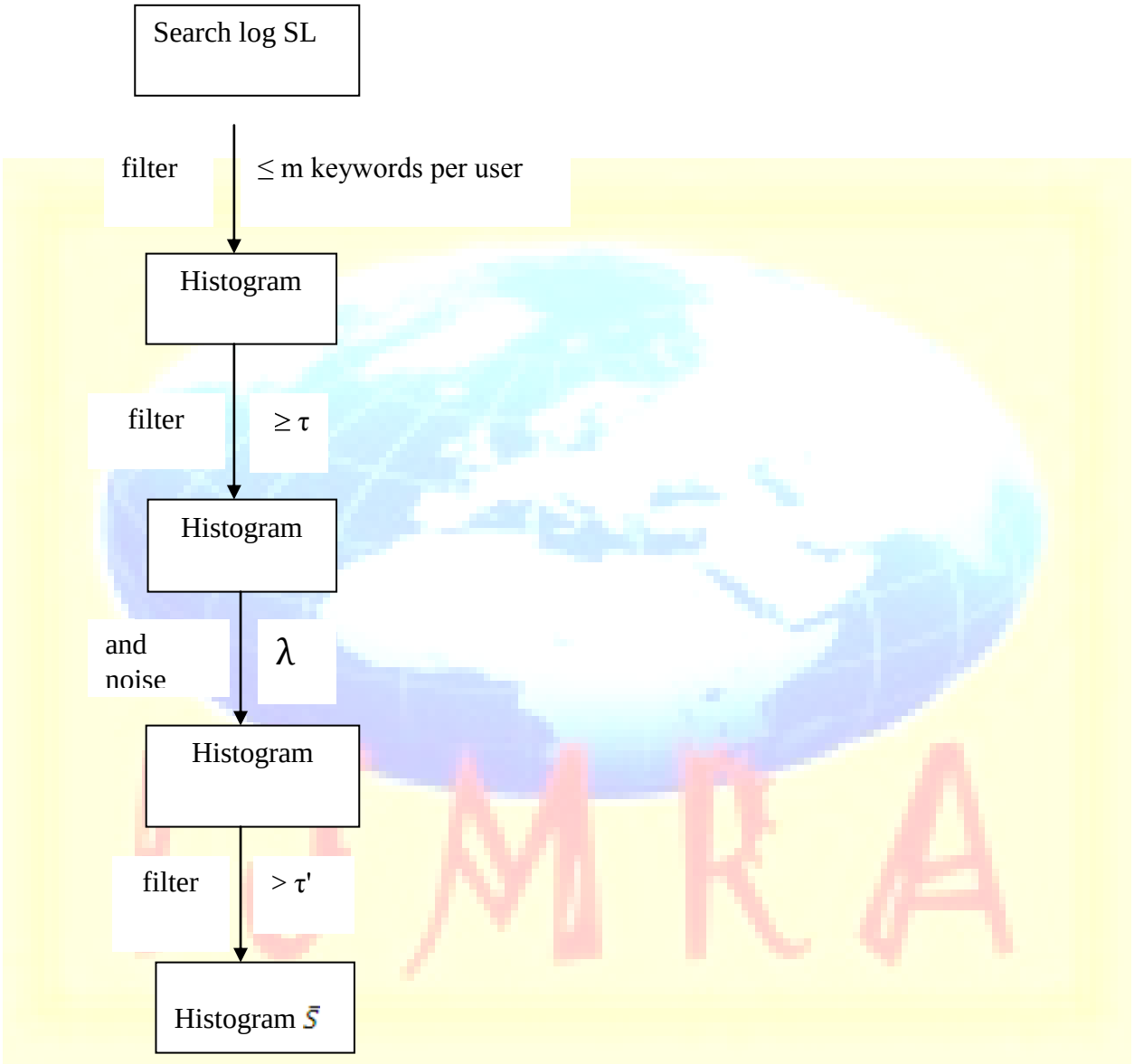


FIG 1. PRIVACY PRESERVING ALGORITHM

Recall:

The recall of substitutions in ground truth that are contained in the substitutions from the sanitized search log query q is the fraction.

$$\text{Recall}(q)=\frac{|\{q_0,\ldots\ldots\ldots\ldots q_{j-1}\}\cap\{q_0'\ldots\ldots\ldots\ldots q_{j-1}'\}|}{|\{q_0,\ldots\ldots\ldots\ldots q_{j-1}\}|}$$

Mean average precision(MAP):

MAP measures the precision of the ranked items for a query as the ratio of true rank and assigned rank.

$$\text{MAP}(q)=\sum_{i=0}^{j-1}\frac{i+1}{rank\ of\ q_i in\left[q_0',\ldots\ldots\ldots q_{j-1}'\right]+1}$$

Index Caching:

The aim of index caching problem is to cache in-memory a set of posting lists that maximizes the hit probability over all keywords. The algorithm first assigns each keyword a score, which equals its frequency in the search log divided by the number of documents that contain the keyword.

Keywords are chosen using a greedy bin-packing strategy where sequentially added posting lists from the keywords with the highest score until the memory is filled. The fixed memory size is to be 1 GB, and each document posting to be 8 Bytes. The inverted index stores the document posting list for each keyword sorted according to their relevance which allows to retrieve the documents in the order of their relevance.

The fact is that it requires only a few very frequent keywords to achieve a high hit probability. Keywords with a big positive impact on the hit probability are less likely to be filtered out by ZEALOUS than keywords with a small positive impact. This explains the marginal decrease in utility for increased privacy.

Item Set Generation and Ranking:

The ideal ranking for query places all good documents at top ranks and all bad documents after that. The algorithm runs to generate ranked substitution on the sanitized search logs. Then these rankings are compared with the rankings produced by the original search log which serve as ground truth.

In the index caching problem, we aim to cache in memory a set of posting lists that maximizes the hit probability over all keywords.

In our experiments, we use an improved version of the algorithm developed by Baeza–Yates to decide which postinglists should be kept in memory . Our algorithm first assigns each keyword a score, which equals its frequency in the search log divided by the number of documents that contain the keyword.

Keywords are chosen using agreedy bin-packing strategy where we sequentially add posting lists from the keywords with the highest score until the memory is filled. In our experiments we fixed the memory size to be 1 GB, and each document posting to be 8 Bytes .

The inverted index stores the document posting list for each keyword sorted according to their relevance which allows to retrieve the documents in the order of their relevance. We truncate this list in memory to contain at most 200,000 documents. Hence, for an incoming query the search engine retrieves the posting list for each keyword in the query either from memory or from disk. If the intersection of the posting lists happens to be empty, then less relevant documents are retrieved from disk for those keywords for which only the truncated posting list is kept on memory.

Figure 2 shows the hit–probabilities of the inverted index constructed using the original search log, the kanonymous search log, and the ZEALOUS histogram (for m = 6) with our greedy approximation algorithm.

We observe that our ZEALOUS histogram achieves better utility than the k-query anonymous search log for a range of parameters. We note that the utility suffers only marginally when increasing the privacy parameter or the anonymity parameter (at least in the range that we have considered). This can be explained by the fact that it requires only a few very frequent keywords to achieve a high hit–probability.

Keywords with a big positive impact on the hit-probability are less likely to be filtered out by ZEALOUS than keywords with a small positive impact. This explains the marginal decrease in utility for increased privacy. As a last experiment we study the effect of varying m on the hit-probability in Figure 2.

We observe that the hit probability for m = 6 is above 0.36 whereas the hit probability for m = 1 is less than 0.33. As discussed a higher value for m increases the accuracy,but reduces the coverage.

Index caching really requires roughly the top 85 most frequent keywords that are still covered when setting m=6. We also experimented with higher values of m and observed that the hit-probability decreases at some point.

INDISTINGUISHABILITY ANALYSIS

This section states how the parameters of ZEALOUS can be set to obtain a sanitized histogram that provides (∈ ',δ')-indistinguishability. Given a search log S and positive numbers m, τ, τ', and λ, ZEALOUS achieves
(∈ ',δ')-indistinguishability, if

$$\lambda \geq {}^{2m}/_{\epsilon'}$$

$$\tau = 1$$

$$\tau' \geq m\left(1 - \frac{\log\left(\frac{2\delta}{m}\right)}{\epsilon'}\right)$$

PROBABILISTICDIFFERENTIAL PRIVACY ANALYSIS

Given values for ∈, δ, τ and m, the following theorem tells us how to set the parameters λ and τ' to ensure that ZEALOUS achieves (∈ ',δ')-probabilistic differential privacy. Given a search log S and positive numbers m,τ, τ', and λ, ZEALOUS achieves (∈, δ)-probabilistic differential privacy, if

$\lambda \geq 2m/\epsilon$, and

$$\tau' - \tau \max\left(-\lambda \ln\left(2 - 2e^{-\frac{1}{\lambda}}\right), \lambda \ln\left(\frac{2\delta}{U.{}^{m}/_{T}}\right)\right)$$

where U denotes the number of users in S.

Choosing Threshold τ

It is necessary to retain as much information as possible in the published search log. A smaller value for τ' immediately leads to a histogram with higher utility because fewer items and

their noisy counts are filtered out in the last step of ZEALOUS. Thus if we choose τ in a way that minimizes τ' we maximize the utility of the resulting histogram.

The utility of ZEALOUS against a representative k-anonymity algorithm by Adar for publishing search logs Recall that Adar's Algorithm creates a k-query anonymous search log as follows: First all queries that are posed by fewer than k distinct users are eliminated. Then histograms of keywords, queries, and query pairs from the k-query anonymous search log are computed.

Thus ZEALOUS can be used to achieve (∈ ',δ')-indistinguishability as well as (∈,δ)-probabilistic differential privacy.

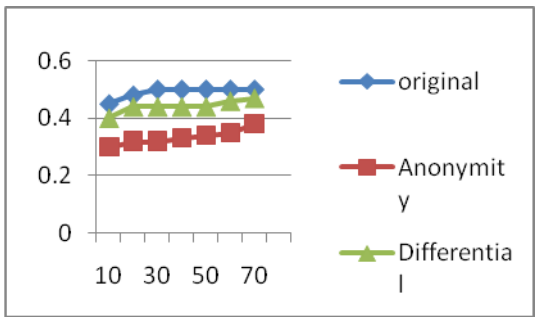
Choosing the Number of Contributions m:

The value of m chosen must be smaller than the average number of items, because it offers better coverage, higher total counts, and reduces the noise compared to higher values of m. This trend has two reasons. First, threshold τ' increases super-linearly in m. Second, as m increases the number of keywords contributed by the users increases only sub-linearly in m; fewer users are able to supply m items for increasing values of m.

IV EXPERIMENTAL RESULTS

The performance of the algorithm can be measured by showing how well the output of the algorithms preserves selected statistics of the original search log.

The different statistics that measure the difference of sanitized histograms to the histograms computed using the original search log is explored. Then analyze the histograms of keywords, queries, and query pairs for both sanitization methods.



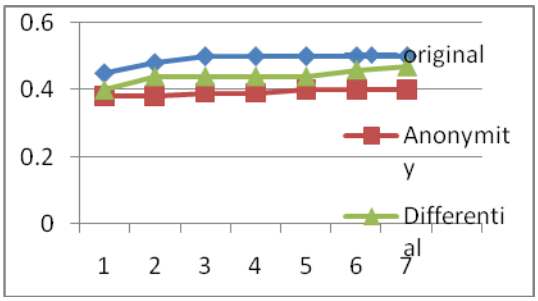


Fig2:Hit probabilities.

The total frequencies are lower for the sanitized search logs than the frequencies in the original histogram because the algorithms filter out a large number of times.

We observe that varying ε and k has hardly any influence on performance. On all precision measures, ZEALOUS provides utility comparable to k-query-anonymity. However, the coverage provided by ZEALOUS is not good. This is because the computation of query substitutions relies not only on the frequent query pairs but also on the countof phrase pairs which record for two sets of keywords how often a query containing the first set was followed by another query containing the second set.

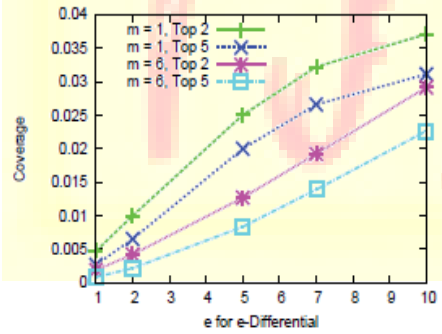


Fig3:Coverage of the privacy-preserving histograms for m = 1 and m = 6.

Thus a phrase pair can have a high frequency even though all query pairs it is contained in have very low frequency. ZEALOUS filters out these low frequency query pairs and thus loses many frequent phrase pairs. As a last experiment, we study the effect of increasing m for query

substitutions. Figure 3 plots the average coverage of the top-2 and top-5 substitutions produced by ZEALOUS for m = 1 and m = 6 for various values of ε. It is clear that across the board larger values of m lead to smaller coverage.

V CONCLUSION

This paper involves the development of algorithm to release useful information about infrequent keywords, queries and clicks in a search log preserving user privacy. In this paper the formal methods of limiting disclosure when publishing frequent keywords, queries, and clicks of a search log is compared. The existing proposals to achieve k-anonymity and differential privacy in search logs are insufficient in the light of attackers who can actively influence the search log. The algorithm ZEALOUS, developed with the goal to achieve relaxations of differential privacy. The project shows how to set the parameters of ZEALOUS to guarantee indistinguishability ,and offers a new analysis that shows how to set the parameters of ZEALOUS to guarantee probabilistic differential privacy, a much stronger privacy guarantee . This evaluation includes applications that use search logs for improving both search experience and search performance, and the results show that ZEALOUS output is sufficient for these applications while achieving strong formal privacy guarantees.

REFERENCES:

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

[2] M. Barbaro and T. Zeller, "A Face is Exposed for AOL Searcher No. 4417749," New York Times, http://www.nytimes.com

[3] A. Blum, K. Ligett, and A. Roth, "A Learning Theory Approach to Non-Interactive Database Privacy," Proc. 40th Ann. ACM Symp. Theory of Computing (STOC).

[4] S. Chakrabarti, R. Khanna, U. Sawant, and C. Bhattacharyya, "Structured Learning for Non-Smooth Ranking Losses. *Knowledge Discovery and Data Mining (KDD).*

[5] M Gotz, A Machanavajjhala, G Wang, X Xiao, and J Gehrke," Publishing Search Logs – A Comparative Study of Privacy Guarantees",July 2012.

[6] M. Gotz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke, "Privacy in Search Logs," *CoRR, abs/0904.0682v2, 2009*.

[7] J. Han and M. Kamber, Data Mining: Concepts and Techniques, first ed. *Morgan Kaufmann, Sept. 2000.*

[8] 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.*

[9] Y. Hong, X. He, J. Vaidya, N. Adam, and V. Atluri, "Effective Anonymization of Query Logs," *Proc. ACM Conf. Information and Knowledge Management (CIKM), 2009.*

[10] R. Jones, R. Kumar, B. Pang, and A. Tomkins, "I Know What You Did Last Summer: Query Logs and User Privacy," *Proc. ACM Conf. Information and Knowledge Management (CIKM).*

[11] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," *Proc. 15th Int'l Conf. World Wide Web (WWW), 2006*

BIBLIOGRAPHY

Ms.M.M.Ramya received her B.E degree Vins Christian College Of Engineering, Chunkankadai , Tamil Nadu ,India and is currently pursuing her M.E degree in Vins Christian College Of Engineering , Chunkankadai , Tamil Nadu ,India.

Her area of interests are Data Mining, Networks .She had attended International Conference at Maria College Of Engineering, Attoor, Tamil Nadu, India and at Bharathiyar Institute Of Technology For Women, Salem, India.

Mrs. R. Medona Selin, M.E received her B.E degree at C.S.I Institute Of Technology , Thovalai , Tamil Nadu ,India and her M.E Karunya Institute Of Technology And Science, Coimbatore, India

Her area of interest is Image Processing . She is currently working as an Assistant Professor in Vins Christian College Of Engineering , Chunkankadai ,Tamil Nadu, India. She has an experience of 8 years in lecturing.