Alinteri J. of Agr. Sci. (2024) 39(2): 111-116
e-ISSN: 2587-2249
info@alinteridergisi.com

http://dergipark.gov.tr/alinterizbd
http://www.alinteridergisi.com/

Online Pricing with Reserve Price Constraint forPersonalData Markets

Mr.S.Kiran Kumar[1], D. Nikitha[2], A. Vinay Rao[2], J. Ravi Kiran[2], R. Chandra Lekha[2]
[1]Asst.Professor,Computer Science and Engineering, CMR Engineering College, medchal, T.S, India [2]B.Tech, Computer Science and Engineering, CMR Engineering College, medchal, T.S, India

Abstract - The society's insatiable appetites for personal data are driving the emergency of data markets, allowing data con- sumers to launch customized queries over the datasets collected by a data broker from data owners. In this paper, we studyhow the data broker can maximize her cumulative revenue by posting reasonable prices for sequential queries. We thus propose a contextual dynamic pricing mechanism with the reserve price constraint, which features the properties of ellipsoid for efficient online optimization, and can support linear and non-linear market value models with uncertainty. In particular, under low uncertainty, our pricing mechanism provides a worst-case regret logarithmic in the number of queries. We further extend to other similar application scenarios, including hospitality service and online advertising, and extensively evaluate all three application instances over MovieLens 20M dataset, Airbnb listings in U.S. major cities, and Avazu mobile ad click dataset, respectively.The analysis and evaluation results reveal that our proposed pricing mechanism incurs low practical regret, online latency, and memory overhead, and also demonstrate that the existenceof reserve price can mitigate the cold-start problem in a posted price mechanism, and thus can reduce the cumulative regret.

Index Terms—personal data market, revenue maximization, contextual dynamic pricing, reserve price.

I. INTRODUCTION

With the proliferation of Internet of Things (IoTs), tremen- dous volumes of data are collected to monitor human behaviorsin daily life. However, for the sake of security, privacy, or busi- ness competition, most of data owners are reluctant to share their data, resulting in a large number of data islands. Thedata isolation status locks the value of personal data against potential data consumers, such as commercial companies, financial institutions, medical practitioners, and researchers. Tofacilitate personal data circulation, more and more data brokershave emerged to build bridges between the data owners and the data consumers. Typical data brokers in industry include Factual, DataSift, Datacoup, CitizenMe, and CoverUS. Onone hand, a data broker needs to adequately compensate the privacy leakages of data owners during the usage of theirdata, and thus incentivize them to contribute private data. On the other hand, the data broker should properly charge the online data consumers for their sequential queries over the

This work was supported in part by Science and Technology Innova- tion 2030 - "New Generation Artificial Intelligence" Major Project No. 2018AAA0100905, in part by China NSF grant 61972252, 61972254, 61672348, and 61672353, in part by the Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing 2018A09, and in part by Alibaba Group through Alibaba Innovation Research (AIR) Program. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies or the government.
Fan Wu is the corresponding author.

collected datasets, since the behaviors of both underpricingand overpricing can incur the loss of revenue at the databroker. Such a data circulation ecosystem is conventionally called "data market" in the literature [1].

In this paper, we study how to trade personal data for revenue maximization from the data broker's standpoint in online data markets. We summarize three major design chal- lenges as follows. The first and the thorniest challenge is that the objective function for optimization is quite complicated. The principal goal of a data broker in data markets is tomaximize her cumulative revenue, which is defined as thedifference between the prices of queries charged from thedata consumers and the privacy compensations allocated tothe data owners. Let's examine one round of data trading as follows. Given a query, the privacy leakages together with the total privacy compensation, regarded as the reserve price ofthe query, are virtually fixed. Thus, for revenue maximization,an ideal way for the data broker is to post a price, whichtakes the larger value of the query's reserve price and market value. However, the reality is that the data broker does not know the exact market value, and can only estimate it fromthe context of the current query and the historical transaction records. Of course, loose estimations will lead to different levels of regret: if the reserve price is higher than the market value, the query definitely cannot be sold, and the regret is zero; if the reserve price is no more than the market value,a slight underestimation of the market value incurs a lowregret, whereas a slight overestimation causes the query notto be sold, generating a high regret. Therefore, the initialgoal of revenue maximization can be equivalently converted toregret minimization. Considering even the single-round regret function is piecewise and highly asymmetric, it is nontrivial forthe data broker to perform optimization for multiple rounds.

Yet, another challenge lies in how to model the market values of the customized queries from the data consumers.To minimize the regret in pricing online queries, the piv-otal step for the data broker is to gain a good knowledgeof their market values. However, markets for personal data significantly differ from conventional markets in that eachdata consumer as a buyer, rather than the data broker as a seller, can determine the product, namely a query. In general, each query involves a concrete data analysis method and a tolerable level of noise added to the true answer, which are both customized by a data consumer [2]. Hence, the queries from different data consumers are highly differentiated, and are uncontrollable by the data broker. This striking property

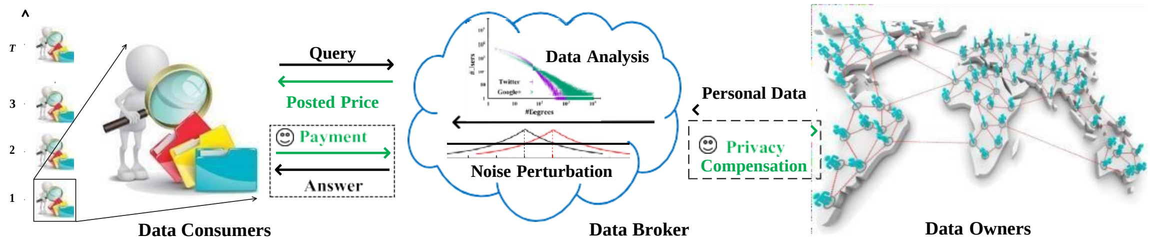

Fig. 1. A general system model of online personal data markets. (The smile indicates that the posted price is accepted and a deal is made.)

further implies that most of the dynamic pricing mechanisms, which target identical products or a manageable number of distinct products, cannot apply here. Besides, existing works on data pricing, which either considered a single query [3] or investigated the determinacy relation among multiple queries but ignored whether the data consumers acceptor reject the marked prices, and thus omitted modeling the market values of queries, are parallel to this work.

The ultimate challenge comes from the novel online pricingwith reserve price setting. For the market value estimationof a query, the data broker can only exploit the current and historical queries. Thus, the pricing of sequential queries can be viewed as an online learning process. In addition to the usual tension between exploitation and exploration, our pricingproblem also needs to incorporate three atypical aspects. First, the feedback after trading one query is very limited.In particular, the data broker can only observe whether the posted price for the query is higher than its market value or not, but cannot obtain the exact market value, which makes standard online learning algorithms inapplicable. Second, the reserve price essentially imposes a lower bound on the posted price beyond the market value estimation, while the ordering between the reserve price and the market value is unknown. Besides, the impact of such a lower bound on the whole learning process has not been studied. Last but not least, the online mode requires our design of the posted price mechanismto be quite efficient. In other words, the data broker needs to choose each posted price and further update her knowledge about the market value model with low latency.

Jointly considering the above three challenges, we propose a contextual dynamic pricing mechanism with the reserve price constraint for the data broker to maximize her revenue inonline personal data markets. For problem formulation, we first adopt contextual/hedonic pricing to model the marketvalues of different queries, which are a certain linear or non- linear function of their features plus some uncertainty. Besides,we choose the state of the privacy compensations under a query as its feature vector. In fact, such a feature representationinherits the key principle of cost-plus pricing. For posted price mechanism design, we start with the fundamental linear model, and covert the market value estimation problem todynamically exploiting and exploring the market values of different features, *i.e.*, the weight vector in the linear model. Specifically, depending on whether a sale occurs or not ineach round, the data broker can introduce a linear inequalityto update her knowledge set about the weight vector. Thus, the raw knowledge set is kept in the shape of polytope, which makes the real-time task of predicting the range of a query's market value computationally infeasible. To handle this prob- lem, we replaces the raw knowledge set with its smallest enclosing ellipsoid, namely Löwner-John ellipsoid. Under the ellipsoid-shaped knowledge set, it only requires a few matrix- vector and vector-vector multiplications to obtain a lower bound and an upper bound on each query's market value. By further incorporating the total privacy compensation, namely the reserve price, as an additional lower bound, we define a conservative posted price and an exploratory posted price fora query. These two kinds of posted prices give different biasesto the immediate rewards (exploitation) and the future rewards(exploration). Besides, the choice of which price in a certain round hinges on the size measure of the latest knowledge set. We further investigate how to tolerate uncertainty, and mainlyintroduce a "buffer" in posting the price and updating theknowledge set. We finally extend to several non-linear modelscommonly used in interpreting market values, including log- linear, log-log, logistic, and kernelized models.

We outline our key contributions in this paper as follows.

- To the best of our knowledge, we are the first to study trading personal data for revenue maximization, from the data broker's point of view in online data markets. Additionally,we formulate this problem into a contextual dynamic pricing problem with the reserve price constraint.
- Our proposed pricing mechanism features the properties of ellipsoid to exploit and explore the market values of sequential queries effectively and efficiently. It facilitates both linearand non-linear market value models, and is robust to some uncertainty. In particular, the worst-case regret under low uncertainty is $O(\max(n^2 \log(T/n), n^3 \log(T/n)/T))$, where$n$ is the dimension of feature vector and T is the total number of rounds. Besides, the time and space complexities are $O(n^2)$. Furthermore, our market framework can also support trading other similar products, which share customization, existenceof reserve price, and timeliness with online queries.
- We extensively evaluate three application instances over three real-world datasets. The analysis and evaluation results reveal that our pricing mechanism incurs low practical regret, online latency, and memory overhead, under both linear and non-linear market value models and over both sparse and densefeature vectors. In particular, (1) for the pricing of noisy linearquery under the linear model, when $n = 100$ and the number of rounds t is 10^5, the regret ratio of our pricing mechanism with reserve price (*resp.*, with reserve price and uncertainty) is7.77% (*resp.*, 9.87%), reducing 57.19% (*resp.*, 45.64%) of the

regret ratio than a risk-averse baseline, where the reserve priceis posted in each round; (2) for the pricing of accommodation rental under the log-linear model, when n = 55, t = 74, 111, and the ratio between the natural logarithms of market value and reserve price is set to 0.6, the regret ratio of our pricing mechanism is 3.83%, reducing 77.46% of the regret ratio compared with the risk-averse baseline; (3) for the pricingof impression under the logistic model, when n = 1024 and$t$ = 10^5, the regret ratios of our pure pricing mechanism are 8.04% and 0.89% in the spare and dense cases, respectively. Furthermore, the online latencies of three applications perround are in the magnitude of millisecond, and the memory overheads are less than 160MB.

. We instructively demonstrate that the reserve price can mitigate the cold-start problem in a posted price mechanism, and thus can reduce the cumulative regret. Specifically, forthe pricing of noisy linear query, when n = 20 and t = 10^4,our pricing mechanism with reserve price (*resp.*, with reserve price and uncertainty) reduces 13.16% (*resp.*, 10.92%) of the cumulative regret than without reserve price; for the pricing of accommodation rental, as the reserve price is approaching the market value, its impact on mitigating cold start is more evident. These findings may be of independent interest.

II. TECHNICAL OVERVIEW

In this section, we introduce system model and problem formulation, and also sketch the fundamental design.

A. System Model

As shown in Fig. 1, we consider a general system model for online personal data markets. There are three kinds of entities:data owners, a data broker, and data consumers.

The data broker first collects massive personal data from data owners. Then, the data consumers comes to the data market in an online fashion. In round t $[T]$, a data consumer arrives, and makes her customized query Q_t over $\in$the collected dataset. Specifically, Q_t comprises a concrete data analysis method and a tolerable level of noise added to the true answer [2]. Here, the noise perturbation can not only allow thedata consumer to control the accuracy of a returned answer, but also preserve the privacies of data owners.

Depending on the query Q_t and the underlying dataset, the data broker quantifies the privacy leakage of each data owner, and needs to compensate her if a deal occurs. The data broker then offers a price p_t to the data consumer. If p_t is nomore than the market value v_t of Q_t, this posted price will be accepted. The data broker charges the data consumer p_t, returns the noisy answer, and compensates the data owners as planned. Otherwise, this deal is aborted, and the data consumergoes away. We note that to guarantee non-negative utility atthe data broker no matter whether a deal occurs in round t ornot, the posted price p_t should be no less than the total privacycompensation q_t, where q_t functions as the *reserve price*, and can be pre-computed when given Q_t.

B. Problem Formulation

We now formulate the regret minimization problem forpricing sequential queries in online personal data markets.

We first model the market values of queries. We use an elementary assumption from *contextual pricing* in computational economics [11]–[13] and *hedonic pricing* in marketing [14], [15], which states that the market value of a product is a deterministic function of its features. Here, the product is a query, and the function can be linear or non-linear. Besides,to make the pricing model more robust, we allow for some uncertainty in the market value of each query. In particular,for a query Q_t, we let $\mathbf{x}_t$ R^n denote its n-dimensional feature$\in$ vector, let $f : \mathrm{R}^n \quad \mathrm{R}$ denote the mapping from thefeature vector $\mathbf{x}_t$ to the deterministic part in its market value,and let δ_t R denote the random variable in its market value$\in$,which is independent of $\mathbf{x}_t$. In a nutshell, $v_t = f(\mathbf{x}_t) + \delta_t$.

We next identify the features of a query for measuringits market value. One naive way is to directly encode thecontents of the query, including the data analysis method and the noise level. However, the query alone, especially the data analysis method, is hard to embody its economicvalue. Thus, we turn to utilizing the underlying valuations from massive data owners about the query, namely the privacycompensations, as the feature vector. We give some commentson such a feature representation: (1) The market value of a query depending on the privacy compensations inherits the core principle of *cost-plus pricing* [16], [17], and has been widely used in personal data pricing [2], [9], [10]. In particular,cost-plus pricing states that the market value of a productis determined by adding a specific amount of markup to its cost. Here, the cost is the total privacy compensation, the determinacy is reflected in the feature representation, and the markup is realized by setting the reserve price constraint.
(2) The privacy compensations are observable by the databroker, and can help her to discriminate the economic values of distinct queries. For example, the privacy compensationsare higher, which implies that the privacy leakages to thedata owners are larger, the knowledge discovered by the data consumer is richer, and thus the market value of the queryto the data consumer should be higher. (3) Considering the large scale of data owners, the dimension of feature vectorcan be prohibitively high. Under such circumstance, we can apply some celebrated dimensionality reduction techniques, *e.g.*, Principal Components Analysis (PCA). Yet, we can also apply aggregation/clustering to the privacy compensations, and regard the aggregate results as the feature vector, where its dimension n controls the granularity of aggregation. For exam-ple, we can sort the privacy compensations, and evenly divide them into n partitions. We sum the privacy compensations falling into a certain partition, and thus obtain a feature. Inthis aggregation pattern, one extreme case is n = 1, wherethe only feature is the total privacy compensation. Another extreme case is n equal to the number of data owners, where every feature corresponds to a data owner's individual privacy compensation.

We finally define the cumulative regret of the data broker due to her limited knowledge of market values. We considera game between the data broker and an adversary. Duringthis game, the adversary chooses the sequence of queries $Q_1, Q_2, \ldots, Q_T$, selects the mapping f, but cannot controlthe uncertainty δ_t in each round t, *i.e.*, she can determine the part $f(\mathbf{x}_t)$ in the market value v_t. In contrast, the data broker can only passively receive each query Q_t, and then post a price p_t. If the posted price is no more than the market value, *i.e.*, $p_t \le v_t$, a deal occurs, and the data broker earns a revenue of p_t. Otherwise, the deal is aborted, and the data broker gainsno revenue. We define the regret in round t as the difference between the adversary's revenue and the data broker's revenue for trading the query Q_t, *i.e.*,

$$R_t = \begin{cases} 0 & \text{if } q_t > v_t, \\ \max_{p_t^*} p_t^* \Pr_{\delta_t}(p_t^* \le v_t) - p_t \mathbf{1}\{p_t \le v_t\} & \text{otherwise.} \end{cases}$$

Here, in the first branch, if the reserve price and thus the posted price are higher than the market value, there is no regret. This is because under such circumstance, no matter whether the adversary knows the market value in advance or the data broker does not, there is definitely no deal/revenue. Besides, p_t^* is the adversary's optimal posted price to maximize her expected revenue in round t, where the expectation is taken over δ_t. When δ_t is omitted, the adversary will just post the market value, if the reserve price is no more than the market value, *i.e.*, $q_t \le p_t^* = v_t$, and R_t will change to:

$$R_t = \begin{cases} 0 & \text{if } q_t > v_t, \\ v_t - p_t \mathbf{1}\{p_t \le v_t\} & \text{otherwise.} \end{cases} \quad (1)$$

At last, considering the queries can be chosen adversarially, *e.g.*, by other competitive data brokers or malicious data consumers, our design goal is to minimize the total worst-case regret accumulated over T rounds.

C. Fundamental Design Under Linear Market Value Model

Due to space limitations, we sketch our proposed pricing mechanism under the linear market value model with σ-subGaussian uncertainty in Algorithm 1. Interested readers can refer to our full article in [18] for design principles, design details, analyses of complexities and worst-case regret, extensions to non-linear market value models, application scenarios, evaluation results, and related work.

Algorithm 1: Online Personal Data Pricing

Input: $\mathbf{A}_1 = R^2\mathbf{I}_{n\times n}$, $\mathbf{c}_1 = \mathbf{0}_{n\times 1}$, an uncertainty parameter $\delta = \sqrt{2\log C\sigma}\log T$, a threshold G

Output: Posted price p_t in each round $t \in [T]$

for $t = 1, 2, \ldots, T$ **do**
- $E_t = \{\vartheta \in \mathbb{R}^n \mid (\vartheta - \mathbf{c}_t)^T \mathbf{A}_t^{-1} (\vartheta - \mathbf{c}_t) \le 1\}$;
- Receive a query Q_t with the feature vector $\mathbf{x}_t \in \mathbb{R}^n$;
- Determine the reserve price q_t of Q_t;
- $\mathbf{b}_t = \frac{\mathbf{A}_t\mathbf{x}_t}{\sqrt{\mathbf{x}_t^T\mathbf{A}_t\mathbf{x}_t}}$;
- $\underline{p}_t = \min_{\vartheta\in E_t} \mathbf{x}_t^T\vartheta = \mathbf{x}_t^T(\mathbf{c}_t - \mathbf{b}_t)$;
- $\bar{p}_t = \max_{\vartheta\in E_t} \mathbf{x}_t^T\vartheta = \mathbf{x}_t^T(\mathbf{c}_t + \mathbf{b}_t)$;
- **if** $q_t \ge \bar{p}_t + \delta$ **then**
  - $\mathbf{A}_{t+1} = \mathbf{A}_t$; $\mathbf{c}_{t+1} = \mathbf{c}_t$;
- **else continue**;
- **if** $\bar{p}_t - \underline{p}_t = 2\sqrt{\mathbf{x}_t^T\mathbf{A}_t\mathbf{x}_t} >$ **then**
  - Post the price $p_t = \max\left\{q_t, \frac{\underline{p}_t+\bar{p}_t}{2} = \mathbf{x}_t^T\mathbf{c}_t\right\}$;
  - **if** p_t is rejected **then**
    - $\alpha_t = \frac{\frac{\underline{p}_t+\bar{p}_t}{2} - (p_t+\delta)}{\sqrt{\mathbf{x}_t^T\mathbf{A}_t\mathbf{x}_t}} = \frac{\mathbf{x}_t^T\mathbf{c}_t - p_t - \delta}{\sqrt{\mathbf{x}_t^T\mathbf{A}_t\mathbf{x}_t}}$;
    - **if** $-\frac{1}{n} \le \alpha_t \le 1$ **then**
      - $\mathbf{A}_{t+1} = \frac{n^2(1-\alpha_t^2)}{n^2-1}\left(\mathbf{A}_t - \frac{2(1+n\alpha_t)}{(n+1)(1+\alpha_t)}\mathbf{b}_t\mathbf{b}_t^T\right)$;
      - $\mathbf{c}_{t+1} = \mathbf{c}_t - \frac{1+n\alpha_t}{n+1}\mathbf{b}_t$;
    - **else**
      - $\mathbf{A}_{t+1} = \mathbf{A}_t$; $\mathbf{c}_{t+1} = \mathbf{c}_t$;
  - **else**
    - $\alpha_t = \frac{\frac{\underline{p}_t+\bar{p}_t}{2} - (p_t-\delta)}{\sqrt{\mathbf{x}_t^T\mathbf{A}_t\mathbf{x}_t}} = \frac{\mathbf{x}_t^T\mathbf{c}_t - p_t + \delta}{\sqrt{\mathbf{x}_t^T\mathbf{A}_t\mathbf{x}_t}}$;
    - **if** $-\frac{1}{n} \le -\alpha_t \le 1$ **then**
      - $\mathbf{A}_{t+1} = \frac{n^2(1-\alpha_t^2)}{n^2-1}\left(\mathbf{A}_t - \frac{2(1-n\alpha_t)}{(n+1)(1-\alpha_t)}\mathbf{b}_t\mathbf{b}_t^T\right)$;
      - $\mathbf{c}_{t+1} = \mathbf{c}_t + \frac{1-n\alpha_t}{n+1}\mathbf{b}_t$;
    - **else**
      - $\mathbf{A}_{t+1} = \mathbf{A}_t$; $\mathbf{c}_{t+1} = \mathbf{c}_t$;
- **else** Post the price $p_t = \max\{q_t, \underline{p}_t - \delta\}$;
  - $\mathbf{A}_{t+1} = \mathbf{A}_t$; $\mathbf{c}_{t+1} = \mathbf{c}_t$;

REFERENCES

[1] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," *PVLDB*, vol. 4, no. 12, pp. 1482–1485, 2011.

[2] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," *Communications of the ACM*, vol. 60, no. 12, pp. 79–86, 2017.

[3] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. of EC*, 2011, pp. 199–208.

[4] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proc. of PODS*, 2012, pp. 167–178.

[5] ——, "Toward practical query pricing with querymarket," in *Proc. of SIGMOD*, 2013, pp. 613–624.

[6] B. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *PVLDB*, vol. 7, no. 9, pp. 757–768, 2014.

[7] S. Deep and P. Koutris, "QIRANA: A framework for scalable query pricing," in *Proc. of SIGMOD*, 2017, pp. 699–713.

[8] ——, "The design of arbitrage-free data pricing schemes," in *Proc. ofICDT*, 2017, pp. 12:1–12:18.

[9] C. Niu, Z. Zheng, F. Wu, S. Tang, X. Gao, and G. Chen, "Unlockingthe value of privacy: Trading aggregate statistics over private correlated

data," in *Proc. of KDD*, 2018, pp. 2031–2040.

[10] C. Niu, Z. Zheng, S. Tang, X. Gao, and F. Wu, "Making big moneyfrom small sensors: Trading time-series data under pufferfish privacy,"in *Proc. of INFOCOM*, 2019, pp. 568–576.

[11] M. C. Cohen, I. Lobel, and R. P. Leme, "Feature-based dynamic pricing,"in
Proc. of EC, 2016, p. 817.

[12] I. Lobel, R. P. Leme, and A. Vladu, "Multidimensional binary searchfor contextual decision-making," in *Proc. of EC*, 2017, p. 585.
[13] R. P. Leme and J. Schneider, "Contextual search via intrinsic volumes,"in *Proc. of FOCS*, 2018, pp. 268–282.
[14] S. Malpezzi, "Hedonic pricing models: a selective and applied review," *Housing economics and public policy*, pp. 67–89, 2002.
[15] P. Ye, J. Qian, J. Chen, C. Wu, Y. Zhou, S. D. Mars, F. Yang, and L. Zhang, "Customized regression model for airbnb dynamic pricing,"in *Proc. of KDD*, 2018, pp. 932–940.
[16] K. B. Monroe, *Pricing : making profitable decisions*, 3rd ed. McGraw-Hill/Irwin, 2003.
[17] T. T. Nagle and G. Müller, *The strategy and tactics of pricing: A guide to growing more profitably*, 6th ed. Routledge, 2018.
[18] S. Jeyapriyanga,,Dr C.N.Ravi,R. Rathiya," Implementation of A Deep Learning Framework for Intelligent Intrusion Detection in Internet of Things Networks" . Scopus. 1208-1213. 5th International Conference on Inventive Research in Computing Applications (ICIRCA 2023)