Fast top-k frequent itemset mining under Local Differential Privacy*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Wang JiaLi
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This is the abstract.**

***Index Terms*—This is the keywords**

I. Introduction

Differential privacy (DP) [7] is the state-of-the-art approach that is used to protect individual privacy in the process of data collection, which has been named one of the world's top 10 breakthrough technologies in 2020 by the MIT technology review. It is a means in cryptography that aims to provide a way to maximize the accuracy of data queries when querying from statistical databases while minimizing the chances of identifying their records. As a mathematical technique, it can add noise to the data while quantifying the extent of the increase in privacy, thus making the process of adding "noise" more rigorous.

Due to its unique advantages, DP has been widely studied by the academia and industry. For example, Google, Microsoft, apple and other companies use this technology to protect users' privacy, and at the same time, mobile phones aggregate data, so as to improve service quality. And the U.S. government is to complete a census of 330 million U.S. residents by 2020, keeping their identities secret, in what would be the largest census ever.

There are two types of differential privacy - Centralized differential privacy (CDP) and Local differential privacy (LDP). Compared with CDP, the LDP does not require the assumptions of a trusted third party and provides stronger privacy guarantees. DP's research has involved many aspects, in recent years, the work in frequent itemset mining (FIM) has attracted the attention, which is one of the most important techniques because of its ability to locate the repeating relationships between different items in a data set and plays an essential role in mining association rules [9]. A lot of work [3]–[6] has been done to solve DM problems in CDP. However, since the analyst holds the user's raw data in CDP setting, its main job is to add noise to the results to satisfy the DP definition.

The LDP setting, by contrast, has no reliance on third party assumptions. The main challenge with the FIM task is that the data analyst does not hold the user's original sensitive information, so it is quite difficult to mine useful information with noise data. In this paper, we consider the FIM problem in transaction databases under LDP setting. Formally, let $I = \{x_1, x_2, ..., x_d\}$ be a set of items and $D =< T_1, T_2, ..., T_n >$ denote a sensitive transaction database, where $T_i (i \in [1...n])$ is a sensitive transaction that is a subset of I. The support of an itemset X, where $X \subset I$ is a set of items, is the number of transactions containing X in D. Given a minimum support threshold δ, the data analyst wants to find the complete set of itemsets that its support is no less than δ; users are sensitive and unwilling to answer their real infomation.

Specifically for FIM in the local setting, Qin et al. [1] leave it as a future work but there is no clear solution. Wang et al. [2] solves the $top-k$ frequent itemset mining (FIM) task for the first time with **padding-and-sampling-based frequency oracle** (PSFO). In [2], the Set-Value Item Mining (SVIM) protocol has been proposed to handles set values under the LDP setting, with the purpose of finding the k most frequent items and their frequencies. To mine frequent itemsets , a core technique is "Guessing Frequency (GF)". That is, the analyst first calculated the frequency of a given itemset X for all candidate itemsets by (1),

$$\varphi(X) = \prod_{x \in X} \mu(x), \mu(x) = \frac{0.9 \times \tilde{\theta}(x)}{\max\limits_{x \in S'} \tilde{\theta}(x)} \tag{1}$$

where $\varphi(X)$ represents the speculative frequency of itemset X, S' and $\tilde{\theta}(x)$ are denoted separately the $top-k$ frequent items set and the frequency of a given item x. Then $2k$ itemsets with highest guessing frequencies are selected to construct candidate set IS. Finally, reference [2] utilized SVIM protocol again with the domain IS to mine $top-k$ itemsets. We observe that, the size of candidate set to construct IS increase significantly with k. As a result, it is computationally expensive when k is large (e.g., $k = 100$).

Inspiringly, we propose minefp protocol, which aims at finding $top-k$ itemsets under the LDP setting and provides similar accuracy while providing lower overhead than existing SVSM protocol within the same privacy constraints. First,

Identify applicable funding agency here. If none, delete this.

TABLE I

SAMPLE OF TRANSACTIONAL DATA.

TID	List of items
T01	a, f, c, g, p
T02	a, b, c, f, l, o
T03	b, f, h, o
T04	b, c, p
T05	f, a, c, l, p, n

the SVIM protocol is used to estimate the k most frequent items and their frequencies. Second, users report the number of frequent items they have; the analyst estimates the distribution user reported and figure out the right M as the maximum iteration of the tree. Third, users interact with the analyst to build effectively the FP-tree [8]. Fourth, the analyst optimizes and mines the FP-tree. Fifth, the analyst publishes $top-k$ itemsets.Experimental results how that minefp outperforms SVSM in that it identifies quickly frequent itemsets as well as estimates the frequencies more accurately.

To summarize, the main contributions of this paper are:

- We study the application of fp-growth algorithm and design the FP-tree-based-mine (minefp) protocol to find frequent itemsets as well as their frequencies in the LDP setting. Experimental results on real-world datasets show the significant improvement over previous techniques.
- We investigate GF to construct candidate set and point out that it is beneficial to build hierarchically FP-tree.

Roadmap.

II. Background

A. Local Differential Privacy (LDP)

In the local setting, there is no trusted third party and an aggregator wants to gather information from users, where each user possesses an input v. The privacy of the data contributor is protected by perturbing her/his original data at the data contributor's side; thus, the agregator cannot access the original data, but is still able to obtain population statistics.

Formally, let D denote the whole transaction databases. ϵ-local differential privacy (or ϵ-LDP) is defined on an algorithm $\mathcal{A}$ and a privacy budget $\epsilon \geq 0$ as follows.

Definition 1: (ϵ-local differential privacy). A randomized algorithm $\mathcal{A}$ satisfies ϵ-local differential privacy (ϵ-LDP), if and only if for (1) all pairs of input $v_i, v_j \in D$, and (2) any possible output $\mathcal{O}$ of $\mathcal{A}$, we have:

$$\frac{\mathbf{Pr}[\mathcal{A}(v_i) = \mathcal{O}]}{\mathbf{Pr}[\mathcal{A}(v_j) = \mathcal{O}]} \leq e^{\epsilon}$$

An important property of differential privacy is *sequential composability* [11], which also applies to the local setting:

Theorem 2.1: *(sequential composability)*. Given m randomized algorithms $\mathcal{A}_i(1 \leq i \leq m)$, each of which satisfies ϵ_i-local differential privacy. Then the sequence of $\mathcal{A}_i$ collectively provides $(\sum_{i=1}^{m} \epsilon_i)$-local differential privacy.

According to sequential composability, given a privacy budget ϵ, we can partition it into multiple portions, each satisfying ϵ_i-LDP, then the whole process provides $\sum_{\epsilon_i}$-LDP.

B. LDP Protocols

A frequency oracle (FO) protocol enables the estimation of the frequency of any given value v

C. FP-growth algorithm

Frequent pattern growth (FP-growth) [8] is an algorithm that mines the complete set of frequent patterns without a costly candidate generation process, which based on the frequent pattern tree (FP-tree) structure that is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns. The FP-Tree is further divided into a set of Conditional FP-Trees for each frequent item so that they can be mined separately. An example of the FP-Tree that represents the frequent items is shown in Fig. 1, where the minimum support count is set to 3.

The FP-growth algorithm solves the problem of identifying long frequent itemsets by searching through smaller conditional FP-tree repeatedly. The conditional pattern base is a "sub-database" which consists of every prefix path in the FP-Tree that co-occurs with every frequent length-1 item. It is used to construct the conditional FP-tree and generate all the frequent patterns related to every frequent length-1 item. In this way, the cost of searching for the frequent patterns is substantially reduced.

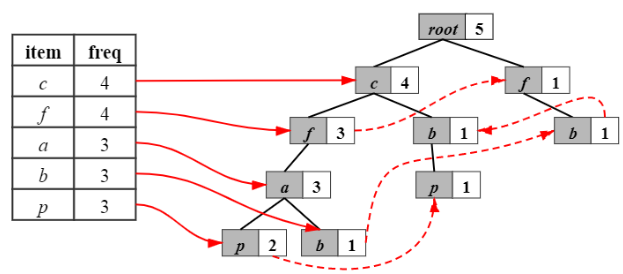

Fig. 1. Frequrnt pattern tree(FP-tree).

III. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IV. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections IV-A–IV-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{2}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(2)", not "Eq. (2)" or "equation (2)", except at the beginning of a sentence: "Equation (2) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.

- There is no period after the “et” in the Latin abbreviation “et al.”.
- The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [18].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 2”, even at the beginning of a sentence.

TABLE II
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

Fig. 2. Example of a figure caption.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [12]. The sentence punctuation follows the bracket [13]. Refer simply to the reference number, as in [14]—do not use “Ref. [14]” or “reference [14]” except at the beginning of a sentence: “Reference [14] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [15]. Papers that have been accepted for publication should be cited as “in press” [16]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [17].

REFERENCES

[1] Qin, Zhan, et al. ”Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy.” computer and communications security (2016): 192-203.

[2] Wang, Tianhao, Ninghui Li, and Somesh Jha. ”Locally Differentially Private Frequent Itemset Mining.” ieee symposium on security and privacy (2018): 127-143.

[3] Bhaskar, Raghav, et al. ”Discovering frequent patterns in sensitive data.” knowledge discovery and data mining (2010): 503-512.

[4] Li, Ninghui, et al. ”PrivBasis: frequent itemset mining with differential privacy.” very large data bases (2012): 1340-1351.

[5] Lee, Jaewoo, and Chris Clifton. ”Top-k frequent itemsets via differentially private FP-trees.” knowledge discovery and data mining (2014): 931-940.
[6] Zeng, Chen, Jeffrey F. Naughton, and Jinyi Cai. ”On differentially private frequent itemset mining.” very large data bases (2012): 25-36.
[7] C. Dwork. Differential privacy. In ICALP, pages 1–12, 2006.
[8] Han J, Pei J, Yin Y, et al. Mining frequent patterns without candidate generation[C]. international conference on management of data, 2000, 29(2): 1-12.
[9] Agrawal R, Imielinski T, Swami A N, et al. Mining association rules between sets of items in large databases[C]. international conference on management of data, 1993, 22(2): 207-216.
[10] Wang T, Blocki J, Li N, et al. Locally Differentially Private Protocols for Frequency Estimation[C]. usenix security symposium, 2017: 729-745.
[11] F. D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In SIGMOD, pages 19–30, 2009.
[12] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[13] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[14] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[15] K. Elissa, “Title of paper if known,” unpublished.
[16] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[17] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[18] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.