Technical perspective on 'Differentially Private Substring and Document Counting'

Grigorios Loukides
King's College London
London, United Kingdom
grigorios.loukides@kcl.ac.uk

ACM Reference Format:
Grigorios Loukides. 2026. Technical perspective on 'Differentially Private Substring and Document Counting'. In . ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Differential privacy has become the de facto privacy standard, as it is widely employed by various organizations. However, only a few research works have focused on strings (e.g., [1, 2, 5]). This is surprising, given that strings are fundamental in modeling, for example, genomic sequences, mobility traces, or text logs. One of the key tasks in string database analysis is to count the occurrences of string fragments in the database: for instance, extracting frequent patterns or q-grams, or publishing word frequency statistics. Most existing differentially private algorithms for string mining are largely heuristic and come with little or no worst-case error analysis, which is a major limitation given the fact that strings appear in key domains for decision making.

The paper "Differentially Private Substring and Document Counting" by Bernardini, Bille, Gørtz, and Steiner addresses this limitation by providing the first near-tight accuracy bounds and efficient data structures for the Substring Count and Document Count problems under differential privacy. The Substring Count problem asks to count the total number of occurrences of a given string P (referred to as *pattern*) in a given collection of strings $\mathcal{D}$ (referred to as database). The Document Count problem asks to count the total number of strings in $\mathcal{D}$ that contain P.

Both problems could intuitively be solved by building a data structure on $\mathcal{D}$, answering a user query on $\mathcal{D}$, and then protecting the query answer by making it differentially private before returning it to the user. However, if one were to answer many queries, then the privacy loss would grow with the number of queries answered, significantly weakening the offered privacy and making the data structure less useful over time. In contrast, the approach by Bernardini et al. is to design a data structure for these problems, which itself (rather than each query answer) is differentially private and can be queried for an arbitrary number of patterns without any privacy loss. To enforce privacy, this approach hides the presence or absence of any string in $\mathcal{D}$, as it requires the answer to be (almost) the same in any two neighboring databases that differ by a single string. To preserve data utility, the data structure outputs a noisy (differentially private) count for any query, while minimizing the additive error.

Such a data structure could be applied, for instance, to analyze DNA databases by counting how often some mutation patterns appear in a population (Substring Count queries) without exposing highly identifying data like ethnicity or family relationships; or to find out how many patients followed specific treatment patterns in a corpus of treatment histories (Document Count queries) while satisfying legal privacy requirements like HIPAA and GDPR.

While the most general data structure for the strictest model of differential privacy proposed by Bernardini et al. is arguably not yet practical because of the high construction time, their data structure to answer queries for fixed-length q-grams under the less strict model of approximate differential privacy can be constructed in almost linear time in the database size, and it is thus of practical relevance.

Such a practical data structure could be successfully applied, for instance, to publish top-trending query q-grams in corpora of search query sequences without leaking users' private information like political beliefs or health concerns. Another important and impactful application would be in releasing q-gram frequency statistics in natural language collections, a task that is particularly relevant for training LLMs on private texts.

Interestingly, the techniques in Bernardini et al. can be easily adapted to enforce differential privacy or approximate differential privacy on popular counting problems on trees [3, 4].

References

[1] Huiping Chen, Changyu Dong, Liyue Fan, Grigorios Loukides, Solon P. Pissis, and Leen Stougie. Differentially private string sanitization for frequency-based mining tasks. In *IEEE International Conference on Data Mining, ICDM 2021, Auckland, New Zealand, December 7-10, 2021*, pages 41–50, 2021.
[2] R. Chen, B. CM Fung, B. C Desai, and N. M Sossou. Differentially private transit data publication: a case study on the montreal transportation system. In *KDD*, pages 213–221, 2012.
[3] Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, and Kewen Wu. On differentially private counting on trees. In *Proc. 50th ICALP*, volume 261, pages 66:1–66:18, 2023.
[4] Badih Ghazi, Ravi Kumar, Jelani Nelson, and Pasin Manurangsi. Private counting of distinct and k-occurring items in time windows. In *Proc. 14th ITCS*, pages 55:1–55:24, 2023.
[5] S. Xu, X. Cheng, S. Su, K. Xiao, and L. Xiong. Differentially private frequent sequence mining. *TKDE*, 28(11):2910–2926, 2016.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn