Aplicação de Top-k Reverso com Privacidade sobre os Dados Públicos de COVID-19 no Estado do Ceará

Maria de Lourdes M. Silva[1], **Iago Chaves**[1], **Javam Machado**[1]

[1]Laboratório de Sistemas e Bancos de Dados
Departamento de Computação – Universidade Federal do Ceará (UFC)
60.455-760 – Fortaleza – CE – Brazil

{malu.maia,iago.chaves,javam.machado}@lsbd.ufc.br

Abstract. *In this article we propose a differentially private reverse top-k query. Our strategy allows the researcher to obtain less frequent data according to his search criteria, with a high guarantee of privacy of the individuals who contributed with the personal data in the original database. We apply our strategy on public data for COVID-19 in the State of Ceará. Our experimental results show that the result of the proposed top-k query returns a high degree of similarity to the result of a conventional top-k query.*

Resumo. *Neste artigo propomos uma consulta top-k reverso diferencialmente privada. Nossa estratégia permite ao pesquisador obter dados menos frequentes de acordo com o seu critério de busca, com alta garantia de privacidade dos indivíduos que contribuíram com os dados pessoais no banco de dados original. Aplicamos a nossa estratégia sobre os dados públicos da COVID-19 do Estado do Ceará. Nossos resultados experimentais mostram que o resultado da consulta top-k proposta retorna alto grau de semelhança ao resultado de uma consulta top-k convencional.*

1. Introdução

Com o objetivo de facilitar o desenvolvimento científico e de dar transparência à informação sobre a evolução da doença, inúmeras instituições publicam dados sobre os pacientes com COVID-19 [SUS 2020], dentre elas, o governo do estado do Ceará. Informações sobre os pacientes são publicadas no nível de microdados, com o cuidado de aplicar técnicas de supressão de alguns atributos como o nome e o CPF dos pacientes para que estes não sejam identificados e a privacidade de cada um seja assegurada.

Muitos trabalhos mostram que a supressão de identificadores ou atributos não é suficiente para manter a privacidade dos indivíduos [Narayanan and Shmatikov 2006]. Técnicas de supressão ou ofuscação não são suficientes para garantir de fato a privacidade do paciente, principalmente em indivíduos que apresentam características ímpares. Por exemplo, pacientes com endereço em regiões geográficas com baixa incidência da doença podem ser identificados por processo de mineração no *dataset* publicado. A privacidade diferencial [Dwork 2011] tem sido usada com sucesso na publicação de dados que contêm informações sensíveis de indivíduos. Suas propriedades formais dão garantia da privacidade dos indivíduos que contribuem com os seus dados, enquanto que mantém um nível aceitável de utilidade dos dados à pesquisa científica.

Consultas do tipo top-k são utilizadas para a descoberta dos k elementos mais frequentes em um conjunto de dados. São consultas que discriminam pois ordenam os dados segundo suas propriedades. Neste sentido apresentam grande potencial para aprender sobre os dados, incluindo descobrir indivíduos que porventura tenham contribuído com os dados consultados, afetando a privacidade dos mesmos. Consultas do tipo top-k reverso tem ainda maior potencial, pois resultam nos k elementos menos frequentes [Vlachou et al. 2010]. Diversos trabalhos abordam o problema do top-k de forma privada, como o DP-Apriori [Cheng et al. 2015], entretanto a literatura carece de trabalhos no contexto de top-k reverso privado.

Este trabalho estuda a abordagem de consulta top-k reverso, buscando estendê-la com os mecanismos de Laplace e exponencial da privacidade diferencial para permitir o seu uso na publicação de dados de saúde. Na Seção 3 descrevemos os algoritmos e analisamos seus resultados quando aplicados aos dados de COVID-19 do Estado do Ceará. Com base nos resultados, a Seção 4 fornece indicações de escolha do algoritmo privado para essa consulta e a sua calibração a fim de alcançar um bom nível de utilidade e assegurar a privacidade dos pacientes.

2. Privacidade Diferencial e Consultas Top-k

Privacidade diferencial é um processo, normalmente chamado de mecanismo, que, através da aleatoriedade no resultado, garante privacidade [Dwork et al. 2014]. Mais intuitivamente, a privacidade diferencial assegura que a presença ou ausência de cada um dos indivíduos do conjunto de dados não impactará na resposta do processo, impedindo que se aprenda sobre um indivíduo algo além do que já se sabe sobre ele. Os mecanismos de Laplace, para consultas numéricas ou de contagem, e exponencial, para consultas que retornam valores não-numéricos, são os mais conhecidos.

Definição 2.1 (Privacidade Diferencial)**.** Um mecanismo *M* é ϵ-diferencialmente privado se para todos os *datasets* vizinhos D_1 e D_2, onde *datasets* vizinhos são conjuntos de dados que se diferem em apenas um registro; e para todo conjunto *S* contido na variação de resultados de *M*, isto é, para todo $S \subset Range(M)$, a seguinte condição é satisfeita:

$$Pr[M(D_1) \in S] \leq \exp(\varepsilon) \times Pr[M(D_2) \in S]. \quad (1)$$

Onde ε, chamado de *budget*, é o limite para a perda de privacidade no resultado de uma consulta. Observe que $\varepsilon \in (0, \infty)$, com $\varepsilon = 0$ para a máxima garantia de privacidade e baixa utilidade dos dados e $\varepsilon = \infty$ para ausência de garantia de privacidade e máxima utilidade dos dados.

Definição 2.2 (Mecanismo de Laplace)**.** Dada uma consulta de agregação f, o mecanismo de Laplace [Dwork et al. 2006] é definido como:

$$M_L(x, f, \epsilon) = f(x) + (Y_1, ..., Y_k) \quad (2)$$

tal que Y_i seja uma variável aleatória que segue a distribuição de Laplace dada por $Lap(\frac{\Delta f}{\epsilon})$, onde $\Delta f = \max ||f(D_1) - f(D_2)||_1$ representa a sensibilidade da consulta, mais especificamente a máxima variação de todas as respostas possíveis.

Definição 2.3 (Mecanismo Exponencial)**.** Seja $M_E(x, u, R)$ um mecanismo exponencial [McSherry and Talwar 2007], a probabilidade da resposta $r \in R$ é proporcional a

$\exp(\frac{\epsilon \times u(x,r)}{2 \times \Delta u})$, onde u representa uma função de pontuação que captura, para cada registro do conjunto de dados, uma pontuação para uma possível resposta $r \in R$. R é um *range* arbitrário que representa os possíveis valores de saída. A sensibilidade Δu é definida como:

$$\Delta u = \max_{r \in R} \max_{D_1, D_2 : ||D_1 - D_2||_1 \leq 1} |u(D_1, r) - u(D_2, r)| \tag{3}$$

Consultas top-k são definidas com base em uma função de pontuação f que calcula a relevância de cada registro para uma determinada consulta ou tarefa e produz como resultado uma lista ordenada com os k registros mais relevantes segundo f,

Definição 2.4 (Top-k)**.** Dado um inteiro positivo k, um conjunto de dados D e uma função de pontuação f.

$$\text{TOP}(D, k, f) = \underset{D' \subseteq D, |D'| = k}{\arg\max} \sum_{d \in D'} f(d) \tag{4}$$

Para consultas top-k reverso utilizaremos uma função de pontuação f de forma análoga ao top-k, entretanto buscaremos os resultados de menor pontuação. Portanto, a consulta resultará em uma lista inversamente ordenada com k registros de menor pontuação.

Definição 2.5 (Top-k Reverso)**.** Dado um inteiro positivo k, um conjunto de dados D e uma função de pontuação f.

$$\text{RTOP}(D, k, f) = \underset{D' \subseteq D, |D'| = k}{\arg\min} \sum_{d \in D'} f(d) \tag{5}$$

3. Top-k Reverso Privado

Neste trabalho propomos dois algoritmos privados para a abordagem top-k reverso. Ambos fazem uso de mecanismos diferencialmente privados no processo interno de execução da consulta. Utilizamos da aleatoriedade do mecanismo de Laplace no primeiro e do mecanismo exponencial no segundo para garantir a privacidade diferencial. Como vimos na Seção 2, esses mecanismos tem processos distintos. O mecanismo de Laplace adiciona ruído ao resultado de consultas de agregação, como uma consulta que busca os 50 bairros com menor número de moradores. O algoritmo da consulta processa a quantidade de moradores por bairro e, para cada bairro, adiciona um ruido aleatório retirado da distribuição de Laplace como estabelecido na **Definição 2.2**. A versão top-k reverso com o mecanismo exponencial tem resultado similar, todavia o processo é diferente. Se desejamos os 50 bairros com o menor número de moradores, nossas possíveis respostas são todos os subconjuntos de tamanho 50 do conjunto de bairros. Para cada um desses subconjuntos precisamos calcular uma pontuação que representa a similaridade do subconjunto com o resultado da resposta real (ver **Definição 2.3**).

Para avaliar os algoritmos que propomos, fazemos uso dos dados fornecidos pelo governo do Estado do Ceará sobre casos de COVID-19, atualizados no dia 18 de junho de 2020. O dataset contem a lista de indivíduos testados para a doença no Estado, tendo sido originalmente retirados os atributos identificadores dos pacientes. Nesta data, o dataset continha 237854 registros e após processo de limpeza ("cleaning") de dados, ele foi reduzido a 177125 registros. Agrupamos as idades em classes para ativarmos a generalização, dificultando a identificação de um indivíduo. A consulta top-k reverso executada busca

obter a relação de classes de indivíduos agrupados por idade com o menor número de casos positivos para o teste da COVID-19. Suas versões diferencialmente privadas descritas nos Algoritmos 1 e 2 buscam obter resultado semelhante, todavia suas respostas são aproximadas de acordo com o ruído introduzido pelo mecanismo aleatório utilizado.

Para determinarmos a qualidade da saída das versões diferencialmente privadas do top-k reverso, optamos por utilizar a métrica F1-Score, visto que esta é a média harmônica entre precisão e revocação. Os *budgets* escolhidos variam entre 0.01 e 2, para que seja possível acompanhar o crescimento do valor do F1-Score, a medida que os valores de ϵ aumentam. A sensibilidade da consulta top-k é 1, dado que esta é a máxima diferença que a alteração, remoção ou inserção de um registro pode impactar na contagem de infectados. Ademais fizemos uso da composição sequencial [McSherry 2009], que diz que para garantirmos ϵ-privacidade diferencial, $\sum_{i=0...k} \epsilon_i = \epsilon$, sendo ϵ_i o *budget* da i-ésima consulta. Como consultas top-k retornam k valores, dizemos que são feitas k consultas dentro de um top-k, com isso, devemos dividir ϵ por k de forma que nossa consulta seja ϵ-diferencialmente privada.

3.1. Top-k Reverso Privado via Mecanismo de Laplace

O Algoritmo 1 abaixo descreve a consulta com o mecanismo de Laplace.

Algoritmo 1: Top-k Reverso Via Laplace

Input: ϵ, Δf, consultaOriginal, dataset, k.
Resultado: saída ordenada com base em ruidosLaplace.
contagemOriginal ← consultaOriginal[‘count’];
classesIdades ← consultaOriginal[‘idade’];
ruidosLaplace ← vetor de zeros de dimensão k;
para *i ← 1...k* **faça**
    ruido ← variavelLaplace(loc=0, scale=$\frac{\Delta f}{(\frac{\epsilon}{k})}$);
    ruidosLaplace[i] ← contagemOriginal + ruido
fim
saida ← classesIdades, ruidosLaplace;

No algoritmo, armazenamos os valores das classes de idades e contagens de infectados obtidas na consulta não privada e somamos ao valor de contagem uma variável aleatória que segue a distribuição de Laplace, onde o parâmetro de localização é 0 e o de escala é $\frac{\Delta f}{(\frac{\epsilon}{k})}$, devido à composição sequencial. Por fim ordenamos o resultado, crescentemente, com base na contagem de infectados de cada classe. A Figura 1 mostra a acurácia dos resultados desta consulta variando o *budget* de 0.01 a 2. O tempo de execução deste algoritmo foi de 0.44525 segundos.

Para a construção do gráfico da Figura 1, executamos o algoritmo 10 vezes e calculamos a média do F1-Score para todos os ϵ. A consulta é `SELECIONAR CONTAGEM(infectado) TOP 10 REVERSO EM classe`. O resultado da consulta privada com $\varepsilon = 0.01$ apresentou o pior F1-Score, visto que o limite de perda de privacidade é muito próximo de zero, adicionando mais ruído na resposta da consulta. Nos *budgets* 0.1, 0.5 e 1, é possível notar o crescimento progressivo do valor do F1-Score, o ruído adicionado na resposta diminui conforme o valor de ϵ aumenta.

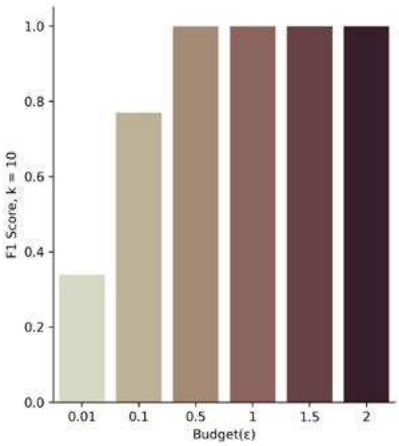

Figura 1. Top-k com Laplace.

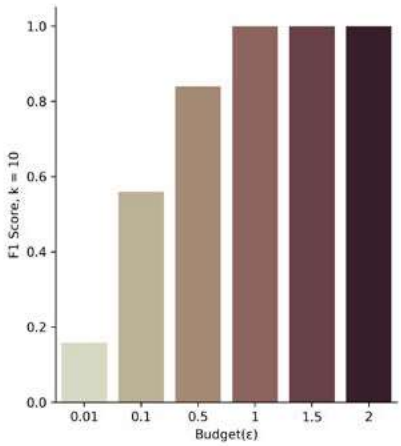

Figura 2. Top-k com exponencial.

3.2. Top-k Reverso Privado via Mecanismo Exponencial

O Algoritmo 2 descreve a extensão da consulta top-k reverso com o mecanismo Exponencial. Nele as probabilidades das classes escolhidas são proporcionais a $\exp\left(\frac{\frac{\epsilon}{k}\times score}{2\times\Delta f}\right)$.

Algoritmo 2: Top-k Reverso Via Exponencial

```
Input: ϵ, Δf, dataset, scores, k.
Resultado: vetor top.
budget ← ϵ/k;
top ← vetor de dimensão k;
para i ← 1...k faça
    probabilidades ← vetor de zeros de dimensão k;
    para score in scores faça
        probabilidade ← exp((budget×score)/(2×Δf));
        probabilidades[i] ← probabilidade;
    fim
    para j ← 1...tamanho(probabilidades) faça
        probabilidades[j] ← probabilidades[j]/soma(probabilidades);
    fim
    Escolhe uma classe não repetida seguindo as probabilidades calculadas e a
      adiciona ao vetor top;
fim
```

O *budget* é atualizado para $\frac{\epsilon}{k}$, devido à composição sequencial. O algoritmo constrói um vetor de probabilidades em que cada probabilidade é calculada como $\exp\left(\frac{budget\times score}{2\times\Delta f}\right)$, sendo o *score* a pontuação atribuída a cada classe. As probabilidades são atribuídas para uma proporção da soma de todas as probabilidades e uma classe não repetida é escolhida com base na probabilidade associada à ela. Este processo é repetido k vezes, até que tenhamos as classes de idades que compõem o top-k reverso privado. O tempo de execução deste algoritmo foi um pouco menor que o anterior, contabilizando 0.33263 segundos.

Para a construção do gráfico da Figura 2, executamos o algoritmo 10 vezes e calculamos a média do F1-Score para todos os ϵ. Pode-se observar que o F1-Score para os *budgets* 0.01 e 0.1 estão distantes dos outros valores, isso acontece devido ao pequeno valor de ϵ, gerando mais aleatoriedade nas escolhas de classes no algoritmo. Para $\epsilon = 0.5$ e $\epsilon = 1$, a resposta da consulta privada estava bem próxima da resposta da consulta não privada e, por fim, para os demais valores de ϵ, a resposta de ambas as consultas foram idênticas, podendo afetar a privacidade dos pacientes cujos dados alimentaram o *dataset*.

4. Discussão Final

As duas versões privadas da consulta obtiveram resultados semelhantes como se pode observar nas Figuras 1 e 2. Todavia o nível de acurácia na versão com o mecanismo de Laplace é maior com *budget* baixo e atinge alto nível mais rapidamente do que a versão com o mecanismo exponencial. De acordo com os resultados, *budgets* próximos de 0.5 no primeiro caso e próximos de 1.0 no segundo levam as consultas a retornar resultados com alta probabilidade de semelhança ao resultado real. Isto diminui a privacidade dos pacientes mas amplia a utilidade das respostas. O desempenho do mecanismo de Laplace se sobressai nesta consulta, quando garante bom nível de utilidade com *budgets* mais baixos. Podemos afirmar que o top-k reverso com Laplace e *budget* em torno de 0.1 é uma ótima escolha para disponibilizar esta consulta à pesquisadores. Por outro lado, para ambos os mecanismos, a consulta não pode ser disponibilizada com *budgets* acima de 0.7 sob pena de revelar sobremaneira os pacientes que contribuíram com os seus dados na formação do dataset. A avaliação futura de outros mecanismos de privacidade diferencial e de outras consultas vai certamente adicionar novas técnicas de publicação de dados de saúde sem comprometer a privacidade dos pacientes que contribuem com seus dados.

Referências

Cheng, X., Su, S., Xu, S., and Li, Z. (2015). Dp-apriori: A differentially private frequent itemset mining algorithm based on transaction splitting. volume 50, pages 74–90.

Dwork, C. (2011). Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *FOCS*, volume 7, pages 94–103.

McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30.

Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.

SUS (2020). Boletim epidemiológico novo coronavírus (covid-19). https://bit.ly/32yFY7a. Acessado em 19-06-2020.

Vlachou, A., Doulkeridis, C., Kotidis, Y., and Nørvåg, K. (2010). Reverse top-k queries. In *International Conference on Data Engineering*, pages 365–376.