# Avaliação Empírica de Técnicas de Comparação Privada Aplicadas na Resolução de Entidades

**Thiago Pereira da Nóbrega**[1,2], **Carlos Eduardo Santos Pires**[2], **Tiago Brasileiro Araújo**[2]

[1]Universidade Estadual da Paraíba, Brazil

thiagonobrega@uepb.edu.br

[2]Universidade Federal de Campina Grande, Brazil

cesp@dsc.ufcg.edu.br, tiagobrasileiro@copin.ufcg.edu.br

***Abstract.*** *Privacy-Preserving Record Linkage (PPRL) consists in identifying which records in two or more databases correspond to the same entity. In this context, different private data comparison techniques have been used, e.g., Bloom filters. However, Bloom filters do not perform well when numeric data or dates are compared. This work aims to evaluate if Homomorphic Asymmetric Cryptography (HAC) can improve the accuracy of the comparison involving non-textual private data. The results indicate that the use of HAC in non-textual data comparison can improve the accuracy of PPRL*

***Resumo.*** *A Resolução de Entidades com Preservação de Privacidade (REPP) consiste em identificar entidades que representam o mesmo objeto no mundo real, mantendo a privacidade dos dados. Nesse contexto, diferentes técnicas de comparação de dados privados vêm sendo utilizadas como, por exemplo, Filtros de Bloom. Contudo, o Filtro de Bloom, não apresenta uma boa precisão quando dados numéricos ou datas são comparados. Este trabalho tem por objetivo avaliar empiricamente se a Criptografia Assimétrica Homomórfica (CAH) pode melhorar a precisão da comparação envolvendo dados privados não textuais. Os resultados apontam que o uso de CAH para comparar dados não textuais pode melhorar a precisão da REPP.*

## 1. Introdução

Com frequência, os dados dos governos e empresas precisam ser integrados e combinados a fim viabilizar análises como, por exemplo, a identificação de reações adversas de medicamento sobre um grupo de pessoas com a mesma característica genética ou a detecção de fraudes nos programas de assistência social [Christen 2012]. Uma tarefa recorrente na integração de dados é a Resolução de Entidades (RE) cujo objetivo é identificar as entidades armazenadas em fontes de dados que representam o mesmo objeto no mundo real.

A tarefa de RE enfrenta diversos desafios como, por exemplo, escalabilidade (para grandes bases de dados), precisão (visto que os dados podem apresentar problemas de qualidade) e privacidade. Este último problema surgiu com a necessidade de reconciliar dados privados (dados que legalmente não podem ser compartilhados). Como exemplo, temos os dados médicos que podem ser integrados para diversos propósitos incluindo

análise de perfil de pacientes e doenças, avaliação de políticas de saúde pública, estudos de novas drogas, entre outros [Pita et al. 2015].

Nesse contexto, a Resolução de Entidades com Preservação de Privacidade (REPP) é o processo utilizado quando há necessidade de identificar entidades que representam o mesmo objeto no mundo real em fontes de dados privados, assegurando que o sigilo e confidencialidade dos dados sejam mantidos durante todo o processo [Vatsalan et al. 2013]. Com o intuito de garantir o sigilo na REPP, se faz necessária a utilização de técnicas para mascarar os dados (*data anonymization*) de modo que: (1) um dado mascarado não possa ser mapeado para o dado original e (2) ao se calcular a similaridade entre dois dados mascarados, o resultado seja o mesmo ao se calcular a similaridade entre seus respectivos valores originais [Vatsalan et al. 2013].

Dentre as técnicas utilizadas para mascarar dados, o Filtro de Bloom [Schnell et al. 2009] se destaca pois permite a comparação de dados textuais com uma boa precisão. No entanto, esta técnica não apresenta a mesma precisão quando aplicada em dados numéricos ou datas, pois não considera as características específicas destes tipos de dados como ordem e quantidade. Por exemplo, a comparação privada de duas datas (01/01/1994 e 01/01/1944) gera um alto valor de similaridade, pois o Filtro de Bloom trata as datas como dados textuais, desconsiderando que alterações nos dígitos do ano representam uma diferença de 50 anos entre as datas.

Por sua vez, a Criptografia Assimétrica Homomórfica (CAH) permite que operações algébricas (adições e multiplicações) possam ser realizadas em dados numéricos e dados mascarados e, consequentemente, que funções de comparação específicas sejam utilizadas. Considerando as datas anteriores, o valor de similaridade entre as mesmas calculado por esta técnica seria baixo, pois os 50 anos de diferença entre as datas seriam considerados.

Este trabalho apresenta uma investigação empírica do impacto da aplicação de Filtro de Bloom e CAH na comparação de dados não textuais no processo de REPP. As técnicas foram avaliadas utilizando métricas de qualidade (Precision, Recall e F-measure) em bases de dados sintéticas. A utilização de CAH apresentou um aumento na precisão da resolução nos experimentos realizados.

O artigo encontra-se estruturado como a seguir. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta as duas técnicas de comparação analisadas: Filtros de Bloom e Criptografia Assimétrica Homomórfica. A Seção 4 apresenta os experimentos realizados e uma análise dos resultados obtidos. Por fim, a Seção 5 conclui o artigo e apresenta sugestões de trabalhos futuros.

## 2. Trabalhos Relacionados

A Resolução de Entidades com Preservação de Privacidade (REPP) vem sendo abordada em vários trabalhos. Em [Schnell et al. 2009], os autores propõem a utilização de Filtros de Bloom para comparação privada de dados médicos. Em [Randall et al. 2014, Schmidlin et al. 2015], os autores propõem diferentes estratégias para utilização de Filtros de Bloom para tipos de dados não textuais.

## 3. Comparação de Dados em REPP

Nesta seção, são descritos alguns dos principais conceitos referentes à comparação de dados privados.

### 3.1. Filtros de Bloom

O Filtro de Bloom [Agarwal and Trachtenberg 2006] é uma estrutura de dados probabilística que utiliza funções de dispersão criptográficas (*hash*) para verificar se um determinado elemento pertence a um conjunto. O filtro é composto por um array de bits no qual todos os bits têm o valor zero no momento da criação. Os elementos são inseridos por meio de funções de dispersão criptográficas que indicam quais posições do array devem ser alteradas para representar o elemento. Schnell (2009) propôs que os Filtros de Bloom fossem utilizados para comparação privada de dados textuais. Para tal, é necessário transformar o dado original em um conjunto de *substrings* (*q-grams*) e inseri-los em um Filtro de Bloom. Por exemplo, a inserção dos nomes *SMITH* e *SMYTH* nos Filtros de Bloom A e B é ilustrada na Figura 1. Primeiramente, os nomes são transformados em *bi-grams* e, em seguida, cada bigrama é mapeado por uma função de dispersão para uma posição do filtro e o valor da posição alterado para *1*.

A similaridade entre dados armazenados em Filtros de Bloom é calculada utilizando funções de distância baseadas em *Token* como a *DICE*, cuja similaridade é dada pela Equação 1:

$$DICE = \frac{2h}{a+b} \tag{1}$$

*h* é o número de posições com o valor *1* que coincidem nos dois filtros e *a* e *b* representam o número de total de *1s* em cada filtro. A comparação dos nomes SMITH e SMYTH é ilustrada na Figura 1. Os nomes são inseridos nos filtros A e B e, em seguida, as posições com o valor 1 são computadas em cada filtro (*a*=11 e *b*=10). Por fim, o número de posições com o valor *1*, que coincidem nos dois filtros, são contabilizadas (*h*=8) e a Equação 1 é aplicada.

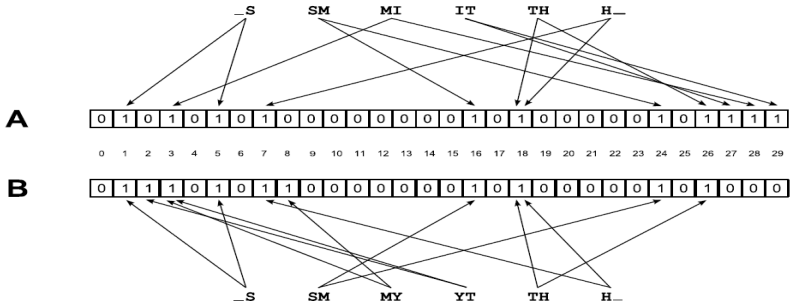


**Figura 1. Exemplo da inserção e cálculo de similaridade entre dados textuais armazenados em Filtros de Bloom. O coeficiente Dice é de 0,76.**

### 3.2. Criptografia Assimétrica Homomórfica (CAH)

A Criptografia Assimétrica Homomórfica (CAH) [Parmar et al. 2014] é uma função criptográfica que utiliza duas chaves, uma para cifrar os dados e outra para a operação oposta. A função possibilita a realização de operações algébricas como adição privada ($\oplus$) e produto privado ($\otimes$), em dados numéricos criptografados, permitindo assim ordenar e calcular (com precisão) a diferença dos dados. Neste trabalho, será utilizada a Criptografia

Parcialmente Homomórfica (CPH) que permite a adição privada de dados cifrados e o produto de dados cifrados por constantes não cifradas. Seja $E_k$ uma função CPH que cifra os dados, então $E_k$ possui as seguintes propriedades:

$$\begin{aligned} E_k(x \oplus y) &= E_k(x) \oplus E_k(y) \\ E_k(x) \otimes c &= E_k(x \times c) \end{aligned} \tag{2}$$

onde *c* é uma constante não criptografada. A utilização de funções CPH permite que a similaridade entre dados não textuais seja calculada utilizando funções especializadas, mantendo a privacidade dos dados e do resultado da comparação [Parmar et al. 2014].

## 4. Avaliação

Esta seção tem o objetivo de avaliar as técnicas apresentadas anteriormente e responder as seguintes questões de pesquisa: (1) A utilização de CPH pode aumentar a precisão da REPP, quando aplicada em dados numéricos ou datas?; (2) Qual o custo computacional de utilizar CPH em REPP?

Para responder a primeira questão, foi feita uma comparação entre a precisão obtida com a utilização de Filtros de Bloom e a precisão alcançada com a utilização de Filtros de Bloom e CPH. Para a responder a segunda questão, o tempo de execução da REPP foi medido e comparado.

### 4.1. Experimentos

O acesso a dados sigilosos ou privados é um desafio para pesquisadores. Assim, para gerar bases de dados sintéticos contendo informações pessoais, foi utilizada a ferramenta Generator and Corruptor (GeCo) [Tran et al. 2013]. Esta ferramenta permite que erros sejam inseridos nos dados (por exemplo, erros de digitação, arredondamento numérico, formatação das datas, entre outros) tornando-os mais próximos de dados reais. Foram construídas três bases de dados, cada uma com 2.400 entidades (pessoas) sendo 400 entidades duplicadas[1]. A **base A** é composta por dois atributos textuais (nome e sobrenome), duas datas (nascimento e óbito) e um atributo numérico (salário); a **base B** é composta por um atributo textual (nome), duas datas (nascimento e óbito) e um atributo numérico (salário); por fim, a **base C** é composta por duas datas (nascimento e óbito) e um atributo numérico (salário).

Para avaliar o impacto da utilização de Filtros de Bloom e da CPH em um processo de REPP, duas estratégias foram empregadas no experimento: a primeira consiste em mascarar e comparar os dados utilizando apenas o Filtro de Bloom (estratégia 1), e a segunda (estratégia 2) consiste em mascarar e comparar os dados numéricos e datas usando uma técnica de CPH (*Pailler*) [Parmar et al. 2014]. Como as estratégias utilizadas apresentam resultados determinísticos, cada experimento foi executado uma única vez.

A similaridade entre os dados numéricos, mascarados com a técnica de CPH, foi calculada adaptando a equação de *similaridade numérica absoluta* [Christen 2012] para utilizar CPH (Equações 3 e 4), onde $n_1$ e $n_2$ representam dados numéricos e $d_{max}$ representa o valor da diferença tolerada entre $n_1$ e $n_2$ .

$$\Delta_{n_1 - n_2} = |E_k(n_1) \oplus (E_k(n_2) \otimes (-1))| \tag{3}$$

[1]A configuração utilizada para gerar as bases de dados está disponível em http://github.com/thiagonobrega

$$sim(n_1, n_2) = \begin{cases} 1 - [\Delta_{n_1 - n_2} \otimes (\frac{1}{d_{max}})] & \text{se } \Delta_{n_1 - n_2} \leq d_{max} \\ 0 & \text{se } \Delta_{n_1 - n_2} > d_{max} \end{cases} . \quad (4)$$

As datas foram convertidas de forma que representassem o número de segundos decorridos desde 01 de janeiro de 1970 (POSIC Time). Dessa forma, a similaridade entre as datas foi calculada pela Equação 3 e 4. O valor de $d_{max}$ utilizado para os valores numéricos foi de 10 (que representa 1% do valor médio dos valores numéricos). Já o $d_{max}$ utilizado para as datas foi de 1 dia (86.400 segundos), pois estudos apontam que os entre 17% e 19% dos erros são de 1 dia nas datas, sendo o restante dos erros referentes a problemas de formatação, ausência e inversão entre mês e dia, entre outros [Schmidlin et al. 2015].

A similaridade entre duas entidades foi determinada por um valor global obtido a partir da combinação dos valores gerados para cada atributo [Christen 2012]. Este valor foi comparado com um limiar, valor mínimo para determinar se duas entidades se referiam ao mesmo objeto do mundo real. Os pares de entidades classificados como similares foram comparados com o gabarito (resultado esperado). Por fim, três métricas de qualidade foram calculadas para avaliar as estratégias: **Precision** que indica a fração dos pares que foram classificados corretamente como similares; **Recall** que representa o número total de pares similares sobre o número de pares similares que foram classificados corretamente; e **F-measure** a média harmônica entre *Precision* e *Recall*.

Os experimentos foram executados em uma máquina com processador Intel Core i7-4790 CPU de 3.60GHz, com 8GB RAM, sistema operacional Linux 4.4.0 de 64bits e Python3.5.

### 4.2. Discussão

Nesta seção são discutidos os resultados dos experimentos e apresentadas as respostas para as questões de pesquisa.

*(1) A utilização de CPH pode aumentar a precisão da REPP quando aplicada em dados numéricos ou datas?* Os resultados dos experimentos estão ilustrados na Figura 2. As duas estratégias foram aplicadas nas três bases de dados.

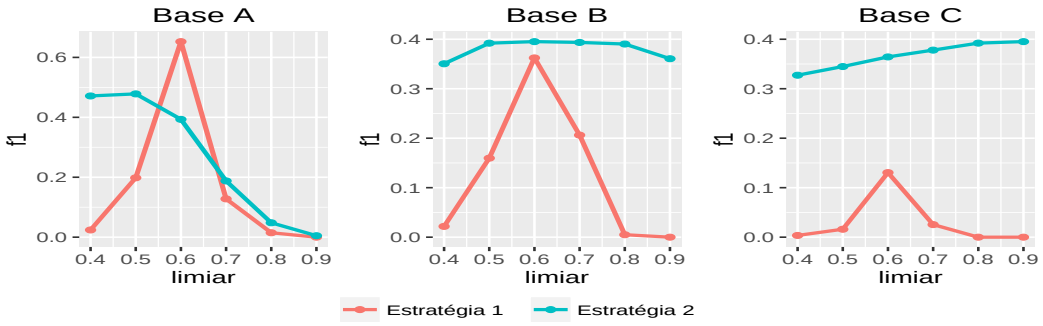


**Figura 2. F-measure (f1) das estratégias por base de dados**

A estratégia 2, quando aplicada na *Base A*, apresentou um resultado superior para todos os limiares avaliados, com exceção do limiar 0.6 para o qual a estratégia 1 apresentou um valor de *f-measure* 25% superior. Esta diferença foi alcançada pela estratégia 1 por ter um *recall* 36% superior ao da estratégia 2. Dois fatores podem explicar esta diferença: a utilização de uma quantidade maior de atributos textuais e o tipo de erros inseridos nos

dados (troca do valor do mês pelo valor do dia nas datas). Para os demais resultados, observa-se que a estratégia 2 apresenta uma melhor precisão à medida que o número de atributos textuais presentes nas bases diminui (*Base B (1)* e *Base C (0)*). Este resultado é alcançado pela utilização de funções de comparação especializadas (Equações 3 e 4) para os dados numéricos e datas.

*(2) Qual o custo computacional de utilizar CPH em REPP?* A estratégia 1 concluiu a REPP em 85 segundos, consumindo 245MB da memória RAM, enquanto que a estratégia 2 concluiu a REPP em 413 segundos, consumindo 164MB da memória RAM. O tempo de execução da estratégia 2 é atribuído ao custo da realização de operações polinomiais para cada comparação realizada, já o maior consumo de memória RAM da estratégia 1 se dá pelo espaço necessário para manter os Filtros de Bloom na memória.

## 5. Conclusão e Trabalhos Futuros

Este artigo avaliou a utilização de CPH e Filtros de Bloom em bases de dados sintéticas para REPP. Embora a aplicação da CPH tenha apresentado um alto custo computacional, aumentando em até 4.8 vezes o tempo de execução, os resultados dos experimentos sugerem que é possível obter um aumento na precisão da REPP quando a CPH é empregada em atributos numéricos e datas. Desse modo, os resultados evidenciam a necessidade de paralelizar a execução das comparações.

## Referências Bibliográficas


Agarwal, S. and Trachtenberg, A. (2006). Approximating the number of differences between remote sets. In *IEEE Information Theory Workshop*, pages 217–221. IEEE.

Christen, P. (2012). *Data Matching*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Parmar, P., B. Padhar, S., N. Patel, S., I. Bhatt, N., and H. Jhaveri, R. (2014). Survey of Various Homomorphic Encryption algorithms and Schemes. *International Journal of Computer Applications*, 91(8):26–32.

Pita, R., Pinto, C., Melo, P., Silva, M., Barreto, M., and Rasella, D. (2015). A Spark-based workflow for probabilistic record linkage of healthcare data. *CEUR Workshop Proceedings*, 1330:17–26.

Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K., and Semmens, J. B. (2014). Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics*, 50:205–212.

Schmidlin, K., Clough-Gorr, K. M., Spoerri, A., and group, f. S. N. C. s. (2015). Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. *BMC Medical Research Methodology*, 15(1):46.

Schnell, R., Bachteler, T., and Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters. *BMC Med Inform.Decis.Mak.*, 9:41.

Tran, K.-n., Vatsalan, D., and Christen, P. (2013). GeCo. *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, pages 2473–2476.

Vatsalan, D., Christen, P., and Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969.