# Trabalho 1
# Robust De-anonymization of Large Sparse Datasets

Matilde Simões
up202108782
Faculdade de Ciências
Universidade do Porto

***Resumo*—Este estudo analisa os riscos da re-identificação em bases de dados anónimas, evidenciando que a remoção de atributos identificadores não garante privacidade quando há fontes externas disponíveis.**

## I. Descrição do Problema

A utilização de base de dados anónimas, que contêm micro-dados sobre preferências individuais, tem sido associada a um aumento significativo do risco da exposição dos utilizadores [1], ou seja, à perda de privacidade. A remoção dos atributos identificadores, como o nome, deixou de ser suficiente para eliminar o risco de re-identificação. Atualmente, a interligação dos dados de diversas fontes externas possibilitam a identificação de utilizadores de forma mais acessível e precisa. O estudo [1] foi realizado com base no risco da exposição de uma base de dados que contém dados de alta dimensionalidade e esparsidade. Foi utilizado um conjunto de dados disponibilizado pelo concurso *Netflix Prize*, que inclui preferências de filmes de 500.000 utilizadores.

Para mostrar a vulnerabilidade da anonimização da base de dados, os autores do artigo implementaram um algoritmo que comprovou a viabilidade da re-identificação de utilizadores, ou seja, a capacidade de associar dados anónimos a identidades específicas, sempre que existir informação externa disponível que possa ser correlacionada. Com isto, concluiram que, apenas com a remoção dos atributos identificadores, aliada a pouca perturbação nos dados, não é suficiente para garantir privacidade. Destacaram o perigo inerente à publicação de bases de dados anónimas e sensibilizaram para a possibilidade de identificação irreversível.

## II. Abordagem e Relevância da informação prévia para a re-identificação

A abordagem baseou-se na implementação de um algoritmo para re-identificar utilizadores num conjunto de dados caracterizado por alta dimensionalidade e esparsidade. Uma base de dados é considerada com alta dimensionalidade quando cada caso contém um grande número de atributos. Já uma base de dados esparsa é aquela em que a maioria dos casos apresenta poucos valores preenchidos, predominando os valores nulos ou ausentes. No contexto do artigo, os dados do *Netflix Prize* apresentaram alta dimensionalidade, uma vez que cada utilizador tinha diversos filmes avaliados, juntamente com as suas preferências, como a classificação atribuída. Por outro lado, a esparsidade dos dados refletiu-se no facto de cada utilizador ter avaliado apenas uma pequena fração do total de filmes disponíveis na base de dados.

Para explorar essas características, os autores desenvolveram o algoritmo *Scoreboard-RH* que não dependia de atributos identificadores explícitos, mas sim da análise de padrões únicos do comportamento dos utilizadores. Foi referido que, inicialmente, o atacante constrói um perfil da vítima, utilizando informações prévias obtidas a partir de fontes externas, como avaliações públicas dos filmes no IMDb. Em seguida, o algoritmo calcula uma pontuação de similaridade entre esse perfil o os casos da base de dados anónima. A similaridade considera a correspondência exata de atributos e, também, pequenas variações, como pequenas diferenças nas datas das avaliações. Por último, é aplicado um critério estatístico para determinar se há correlação suficiente para identificar a vítima, isolando um subconjunto de casos ou até mesmo um único caso.

A apropriação de informação prévia é um fator essencial para se obter resultados positivos na re-identificação. O estudo demonstrou que, com o conhecimento prévio de apenas 5 a 10 avaliações de filmes é possível identificar utilizadores com alta probabilidade. Não é só a quantidade de informação que influencia a precisão do ataque, mas também a sua especificidade e originalidade. Se a informação prévia for pouco comum, o conjunto de casos com alta similaridade será significativamente reduzido, tornando a identificação mais precisa. O artigo também evidenciou que os atributos comuns, como a avaliação de filmes populares, são menos relevantes para a re-identificação, uma vez que muitos utilizadores possuem perfis parecidos. Por outro lado, quando os filmes avaliados são pouco conhecidos, o perfil torna-se mais autêntico, fazendo com que o espaço de procura diminua e eficácia do ataque aumente.

Outro aspeto relevante referido é a robustez que o algoritmo oferece em relação à informação ligeiramente incorreta proveniente de fontes externas. O estudo provou que a re-identificação é possível mesmo quando a informação prévia contém pequenos erros ou imprecisões. Este resultado reforça a importância de melhorar as técnicas de anonimização para disponibilizar bases de dados anónimas sem comprometer a privacidade dos utilizadores.

## III. Formulação das Questões

### *A. As base de dados anónimas criam uma nova discriminação?*

O artigo enfatizou fortemente a importância dos atributos raros na re-identificação, demonstrando que os utilizadores que têm preferências e características menos comuns são mais vulneráveis à perda da privacidade. Isto deve-se ao facto de que a singularidade pode ser facilmente correlacionada com informações de fontes externas, o que aumenta significativamente a probabilidade de re-identificação. Em outras palavras, podemos concluir que a exclusividade e a originalidade constituem um risco para a privacidade.

Contudo, esta ideia parece contraditória. Não há nada de errado em ter preferências diferentes, pelo contrário. A sociedade tende a valorizar a individualidade, a estimular o pensamento crítico e as premiar as pessoas que se destacam nos seus interesses, estilo de vida, entre outros. No entanto, quando falamos de privacidade, essa autenticidade torna-se numa vulnerabilidade, pois as técnicas de anonimização baseiam-se na homogeneidade para serem eficazes. Esta discrepância cria uma nova forma de discriminação, comprometendo a equidade na proteção dos dados. Na prática, significa que alguns utilizadores estão inevitavelmente mais vulneráveis do que outros por terem personalidades e gostos que se distanciam do padrão geral da sociedade em que vivem, muitas vezes sem qualquer controlo ou escolha consciente. Além disso, a consciencialização deste risco pode limitar a liberdade pessoal, fazendo com que os utilizadores se sintam pressionados a adotar comportamentos mais comuns e interesses e preferências mais populares.

Uma possível solução é a utilização de dados sintéticos ou de técnicas que não se baseiam na homogeneidade. Sem mudanças nesse sentido, vamos continuar a expor a singularidade, em vez de a reconhecer como uma qualidade ímpar.

### *B. Manter a privacidade e evitar a exposição das identidades ainda é viável, ou obter informação prévia já é inevitável?*

De acordo com o artigo [1], a exposição de identidades e a perda da privacidade estão associadas à informação prévia disponível sobre cada utilizador. No dia a dia, a pegada digital tornou-se inevitável, seja ao aceitar termos que não lemos, ao permitir a utilização de *cookies*, entre outros. Este comportamento demonstra iliteracia de privacidade digital, onde deixa de ser vista como um direito fundamental e passa a ser cedida de forma automática e por conveniência. Cada uma dessas ações contribui para um sistema que guarda e processa informações detalhadas das nossas preferências, padrões e comportamentos. Independentemente do cuidado que tenhamos para não sermos re-identificados, irá sempre existir informação prévia disponível sobre nós e essa quantidade de dados só tende a aumentar.

Atualmente, a pegada digital é um fator decisivo em processos de *backgroud check*. As empresas não avaliam só o currículo, mas também têm interesse nas atividades que os seus colaboradores realizam *online*, bem como as suas interações sociais, entre outros aspetos. Ter histórico digital deixou de ser opcional e tornou-se numa necessidade, sendo a sua ausência frequentemente interpretada como algo incomum.

Por outro lado, existem os *hackers* e pessoas que querem trabalhar sem deixar rasto. Hoje em dia, minimizar ou eliminar a pegada digital tornou-se quase impossível, uma vez que praticamente tudo o que fazemos digitalmente fica registado e cria dados. Portanto, irão sempre existir base de dados, que não são apenas guardadas, mas também, expostas e vendidas. Na minha opinião e pelas notícias que vejo, a exposição dos dados tornou-se um problema diário que é, muitas vezes, silenciado e encoberto.

Para concluir, parece-me inevitável existir informação prévia sobre qualquer utilizador. Depois desta reflexão, questiono-me se a técnica de tornar base de dados anónimas faz sentido, ou se precisamos de uma técnica completamente inovadora, tendo em conta que a quantidade de informação disponível só vai aumentar e vai ser sempre possível identificar utilizadores ao correlacionar dados de outras fontes, e ao utilizar *machine learning* e *big data*. Apesar disso, acredito que manter a privacidade e evitar a exposição das identidades ainda é viável. A criptografia é uma área em constante evolução e, se for combinada com técnicas de anonimização, pode oferecer melhores garantias. Também considero essencial consciencializar a sociedade sobre este tema, pois parece-me que a perceção sobre os dados expostos e ataques informáticos é encarado como um problema de terceiros.

A privacidade é um tema bastante atual e relevante, logo é fundamental implementar medidas para mitigar as vulnerabilidades e os riscos conhecidos. Talvez começar por reavaliar criticamente se estamos a proteger os dados de forma eficaz, ou só a criar uma falsa sensação de segurança nos dados.

## Referências

[1] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. *Proceedings - IEEE Symposium on Security and Privacy*, pages 111–125, 2008.