GRUPO DE TRABALHO DE PROTEÇÃO DE DADOS DO ARTIGO 29.º

0829/14/PT
GT216

Parecer 05/2014 sobre técnicas de anonimização

Adotado em 10 de abril de 2014

GRUPO DE TRABALHO PARA A PROTEÇÃO DAS PESSOAS NO QUE DIZ RESPEITO AO TRATAMENTO DE DADOS PESSOAIS

instituído pela Diretiva 95/46/CE do Parlamento Europeu e do Conselho, de 24 de outubro de 1995,

Tendo em conta os artigos 29.º e 30.º da referida diretiva,

Tendo em conta o seu regulamento interno,

ADOTOU O PRESENTE PARECER:

RESUMO

Neste parecer, o GT analisa a eficácia e os limites das técnicas de anonimização existentes no contexto jurídico comunitário de proteção de dados e apresenta recomendações para lidar com essas técnicas, tendo em conta o risco residual de identificação inerente a cada uma delas.

O GT reconhece o valor potencial da anonimização de dados pessoais, em particular enquanto estratégia para colher os benefícios dos «dados abertos» para as pessoas e a sociedade em geral, reduzindo, simultaneamente, os riscos para as pessoas em causa. No entanto, alguns estudos de caso e publicações científicas demonstraram a dificuldade de criar um conjunto de dados verdadeiramente anónimo que mantenha simultaneamente as informações subjacentes suficientes exigidas para a tarefa em questão.

Tendo em conta a Diretiva 95/46/CE e outros instrumentos jurídicos aplicáveis da UE, a anonimização resulta do tratamento de dados pessoais a fim de evitar irreversivelmente a identificação. Ao fazê-lo, devem ser tidos em conta vários elementos pelos responsáveis pelo tratamento de dados, considerando o conjunto dos meios «suscetíveis de serem razoavelmente» utilizados para identificação (seja pelo responsável pelo tratamento, seja por terceiros).

A anonimização constitui um tratamento posterior de dados pessoais; como tal, deve satisfazer o requisito de compatibilidade em função dos fundamentos jurídicos e circunstâncias do tratamento posterior. Além disso, os dados anónimos são, de facto, abrangidos pelo âmbito de aplicação da legislação relativa à proteção de dados, mas os titulares dos dados podem ainda ter direito à proteção ao abrigo de outras disposições (tais como as relativas à proteção da confidencialidade das comunicações).

As principais técnicas de anonimização de dados pessoais, designadamente a aleatorização e a generalização, encontram-se descritas no presente parecer. Em especial, o parecer aborda a adição de ruído, a permuta, a privacidade diferencial, a agregação, o k-anonimato, a l-diversidade e a t-proximidade. Explica os seus princípios, os seus pontos fortes e fracos, bem como os erros e deficiências comuns relacionados com a utilização de cada técnica.

O parecer analisa a solidez da cada técnica com base em três critérios:

(i) se ainda é possível identificar uma pessoa,

(ii) se ainda é possível estabelecer a ligação entre registos relativos a uma pessoa singular, e

(iii) podem ser inferidas informações relativamente a um indivíduo?

O facto de conhecer os principais pontos fortes e fracos de cada técnica ajuda a escolher a forma de criar um processo adequado de anonimização num determinado contexto.

A utilização de pseudónimos é igualmente abordada, a fim de clarificar alguns armadilhas e equívocos: a utilização de pseudónimos não é um método de anonimização de dados pessoais. Apenas dificulta a possibilidade de correspondência de um conjunto de dados à identidade original de um titular de dados e é, por conseguinte, uma medida de segurança útil.

O parecer conclui que as técnicas de anonimização podem fornecer garantias de privacidade e podem ser utilizadas para gerar processos eficazes de anonimização, mas apenas se a sua

aplicação for adequadamente construída – o que significa que os requisitos prévios (âmbito) e os objetivos do processo de anonimização devem ser claramente definidos a fim de obter a anonimização pretendida, ao mesmo tempo que produzem alguns dados úteis. A melhor solução deve ser decidida caso a caso, eventualmente por meio de uma combinação de técnicas diferentes e tendo em conta as recomendações práticas desenvolvidas no presente parecer.

Por último, os responsáveis pelo tratamento de dados devem ter em conta que um conjunto de dados anonimizado ainda é passível de apresentar riscos residuais para os titulares dos dados. Com efeito, se por um lado a anonimização de dados pessoais e a reidentificação são domínios de investigação ativos e são publicadas regularmente novas descobertas, por outro lado, até os dados anonimizados, como estatísticas, podem ser utilizados para enriquecer os perfis existentes das pessoas singulares, criando assim novas questões relativas à proteção de dados. Assim, a anonimização de dados pessoais não deve ser considerada um exercício pontual e os riscos inerentes devem ser reavaliados regularmente por responsáveis pelo tratamento de dados.

1 Introdução

Enquanto os dispositivos, sensores e redes criam grandes volumes e novos tipos de dados e os custos de armazenamento de dados se estão a tornar negligenciáveis, existe um crescente interesse e procura do público pela reutilização desses dados. Os «dados abertos» podem oferecer vantagens evidentes para a sociedade, as pessoas e organizações, mas apenas se forem respeitados os direitos de todos em matéria de proteção dos dados pessoais e da vida privada de cada um.

A anonimização de dados pessoais pode ser uma boa estratégia para manter os benefícios e atenuar os riscos. Quando um conjunto de dados se encontra verdadeiramente anonimizado e as pessoas deixam de ser identificáveis, a legislação europeia de proteção de dados deixa de ser aplicável. No entanto, estudos de casos e publicações de investigação evidenciam que criar um conjunto de dados verdadeiramente anónimo a partir de um conjunto substancial de dados pessoais mantendo, simultaneamente, as informações subjacentes exigidas para a tarefa não é um desafio simples. Por exemplo, um conjunto de dados considerado anónimo pode ser combinado com outro conjunto de dados de modo a que uma ou mais pessoas sejam passíveis de ser identificadas.

No presente parecer, o GT analisa a eficácia e os limites de técnicas existentes para a anonimização de dados pessoais no contexto jurídico da UE em matéria proteção de dados e apresenta recomendações para uma utilização prudente e responsável dessas técnicas para a construção de um processo de anonimização de dados pessoais.

2 Definições e análise jurídica

2.1. Definições constantes do quadro jurídico da UE

A Diretiva 95/46/CE refere a anonimização de dados pessoais no considerando 26 para excluir os dados anonimizados do âmbito de aplicação da legislação relativa à proteção de dados:

> *«Considerando que os princípios de proteção devem aplicar-se a qualquer informação relativa a uma pessoa identificada ou identificável; que, para determinar se uma pessoa é identificável, importa considerar o conjunto dos meios suscetíveis de serem razoavelmente utilizados, seja pelo responsável pelo tratamento, seja por qualquer outra pessoa, para identificar a referida pessoa; que os princípios da proteção não se aplicam a dados tornados anónimos de modo tal que a pessoa já não possa ser identificável; que os códigos de conduta na aceção do artigo 27.º podem ser um instrumento útil para fornecer indicações sobre os meios através dos quais os dados podem ser tornados anónimos e conservados sob uma forma que já não permita a identificação da pessoa em causa;»*[1].

[1] Salienta-se, além disso, que se trata da abordagem igualmente seguida para o projeto de regulamento da EU relativo à proteção de dados, nos termos do considerando 23, «para determinar se uma pessoa é identificável, importa considerar o conjunto dos meios suscetíveis de serem razoavelmente utilizados, quer pelo responsável pelo tratamento dos dados quer por qualquer outra pessoa para identificar a referida pessoa».

Uma leitura atenta do considerando 26 oferece uma definição conceptual da anonimização. O considerando 26 significa que, para anonimizar quaisquer dados, têm de lhes ser retirados elementos suficientes para que deixe de ser possível identificar o titular dos dados. Mais precisamente, os dados têm de ser tratados de forma a que já não possam ser utilizados para identificar uma pessoa singular utilizando «o conjunto dos meios suscetíveis de serem razoavelmente utilizados», seja pelo responsável pelo tratamento, seja por terceiros. Um fator importante que convém assinalar é que o tratamento tem de ser irreversível. A diretiva não esclarece no atinente a como o processo de desidentificação deve ou pode ser realizado[2]. Centra-se nos resultados: que os dados devem ser de molde a não permitir que o titular dos dados seja identificado através de «todos» os meios, «prováveis» e «razoáveis». É feita referência aos códigos de conduta enquanto ferramenta para estabelecer mecanismos possíveis de anonimização de dados pessoais, bem como a retenção sob uma forma em que a identificação do titular dos dados «já não seja possível». Assim, a diretiva estabelece claramente um padrão bastante elevado.

A Diretiva relativa à Privacidade e Comunicações Eletrónicas (Diretiva 2002/58/CE) também se refere à «anonimização de dados pessoais» e a «dados anónimos» muito similarmente. O considerando 26 refere que:

> *«Dados de tráfego utilizados para comercialização de serviços ou para a prestação de serviços de valor acrescentado devem igualmente ser eliminados ou tornados anónimos após o fornecimento do serviço».*

Por conseguinte, o artigo 6.º, n.º 1, estabelece que:

> *«Sem prejuízo do disposto nos n.os 2,3 e 5 do presente artigo e no n.º 1 do artigo 15.º, os dados de tráfego relativos a assinantes e utilizadores tratados e armazenados pelo fornecedor de uma rede pública de comunicações ou de um serviço de comunicações eletrónicas publicamente disponíveis devem ser eliminados ou tornados anónimos quando deixem de ser necessários para efeitos da transmissão da comunicação.»*

Além disso, nos termos do artigo 9.º, n.º 1:

> *«Nos casos em que são processados dados de localização, para além dos dados de tráfego relativos a utilizadores ou assinantes de redes públicas de comunicação ou de serviços de comunicações eletrónicas publicamente disponíveis, esses dados só podem ser tratados se forem tornados anónimos com o consentimento dos utilizadores ou assinantes, na medida do necessário e pelo tempo necessário para prestação de um serviço de valor acrescentado.»*

O princípio subjacente é de que o resultado da anonimização como técnica aplicada aos dados pessoais deve ser, no estado atual da tecnologia, tão permanente quanto a eliminação, ou seja, tornando impossível o tratamento de dados pessoais.[3]

---

[2] Este conceito é aprofundado mais adiante, na p. 8. do presente parecer.

[3] Convém aqui recordar que a anonimização se encontra também definida nas normas internacionais, tais como a norma ISO 29100 – consistindo no «processo pelo qual as informações pessoais identificáveis (IPI) são alteradas irreversivelmente de modo que uma entidade IPI já não possa ser identificada direta ou indiretamente, quer pelo responsável pelo tratamento de IPI por si só ou em colaboração com qualquer outra parte» (ISO 29100:2011). A irreversibilidade da modificação sofrida pelos dados pessoais para permitir a identificação direta ou indireta também é a chave para a norma ISO. Nesta perspetiva, existe uma grande convergência com os princípios e conceitos em que assenta a Diretiva 95/46/CE. O mesmo se aplica às definições existentes em algumas legislações nacionais (por exemplo, em Itália, Alemanha e Eslovénia), centradas na não identificabilidade e sendo feita alusão ao «esforço desproporcional» de voltar a identificar (D, SI). No entanto, a legislação francesa de proteção de dados estabelece que os dados permanecem dados pessoais, mesmo que seja extremamente difícil

2.2. Análise jurídica

A análise da redação referente à anonimização de dados pessoais constante nos principais instrumentos de proteção de dados da UE permite destacar quatro características fundamentais:

– A anonimização pode ser um resultado do tratamento de dados pessoais, com o objetivo de evitar irreversivelmente a identificação do titular dos dados.

– Podem ser previstas várias técnicas de anonimização, não existe qualquer norma prescritiva na legislação europeia.

– Importa reconhecer a devida importância dos elementos contextuais: deve ter-se em conta «o conjunto» dos meios «suscetíveis de serem razoavelmente» utilizados para identificação pelo responsável pelo tratamento e por terceiros, com especial atenção para os que recentemente se tornaram, com a atual tecnologia, «suscetíveis de serem razoavelmente» utilizados (tendo em conta a evolução da capacidade computacional e das ferramentas disponíveis).

– É inerente à anonimização um fator de risco: este fator de risco deve ser tido em conta ao avaliar a validade de qualquer técnica de anonimização – incluindo as possíveis utilizações de quaisquer dados que sejam «anonimizados» através dessa técnica – e a gravidade e probabilidade deste risco devem ser avaliadas.

Neste parecer, é utilizada a notação «técnica de anonimização», em vez de «anonimato» ou «dados anónimos», para salientar o risco residual inerente de reidentificação que qualquer medida técnica e organizativa destinada a tornar os dados anónimos comporta.

2.2.1. Licitude do processo de anonimização

Em primeiro lugar, a anonimização é uma técnica aplicada aos dados pessoais a fim de atingir uma desidentificação irreversível. Por conseguinte, o pressuposto inicial é que os dados pessoais têm de ter sido recolhidos e tratados em conformidade com a legislação aplicável relativa à conservação de dados num formato identificável.

Neste contexto, o processo de anonimização, ou seja, o tratamento de dados pessoais para atingir a respetiva anonimização, constitui um «tratamento ulterior». Como tal, este tratamento deve cumprir o teste de compatibilidade em conformidade com as orientações fornecidas pelo grupo de trabalho no seu parecer 03/2013 relativo à limitação de finalidades[4].

Isto significa que, em princípio, a base jurídica para a anonimização pode ser encontrada em qualquer um dos motivos mencionados no artigo 7.º (incluindo o interesse legítimo do responsável pelo tratamento de dados), desde que os requisitos de qualidade de dados constantes do artigo 6.º da diretiva sejam igualmente preenchidos e tendo devidamente em

---

e pouco provável voltar a identificar o titular dos dados – isto é, não existe qualquer disposição referente ao teste de «razoabilidade».

[4] Parecer 03/2013 do Grupo de Trabalho do artigo 29.º, disponível em:http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

conta as circunstâncias específicas e todos os fatores referidos pelo grupo de trabalho no parecer relativo à limitação de finalidades[5].

Por outro lado, as disposições constantes do artigo 6.º, n.º 1, da Diretiva 95/46/CE (mas também nos artigos 6.º, n.º 1, e 9.º, n.º 1, da diretiva relativa à privacidade e comunicações eletrónicas) devem ser salientadas, dado que demonstram a necessidade de manter os dados pessoais «de forma a permitir a identificação» por um período não superior ao necessário para efeitos de recolha ou tratamento subsequentes.

Por si só, esta disposição sublinha firmemente que os dados pessoais devem, pelo menos, ser anonimizados «por defeito» (sem prejuízo de requisitos jurídicos diferentes, tais como os constantes da diretiva relativa à privacidade de comunicações eletrónicas em matéria de dados de tráfego). Se o responsável pelo tratamento de dados desejar conservar esses dados pessoais depois de atingir os fins do tratamento inicial ou na sequência do tratamento de dados, deve recorrer a técnicas de anonimização a fim de evitar irreversivelmente a identificação.

Por conseguinte, o grupo de trabalho considera que a anonimização como tratamento ulterior de dados pessoais pode ser considerada compatível com os objetivos iniciais do tratamento de dados, mas apenas na condição do processo de anonimização consistir na produção fiável de informações anonimizadas no sentido referido no presente documento.

Importa igualmente realçar que a anonimização tem de ser realizada em conformidade com as restrições legais recordadas pelo Tribunal de Justiça da União Europeia no processo C-553/07 (*College van burgemeester en wethouders van Rotterdam contra M.E.E. Rijkeboer*), relativas à necessidade de conservar os dados em formato identificável, para permitir, por exemplo, o exercício dos direitos de acesso pelos titulares dos dados. O Tribunal de Justiça europeu deliberou que o *«artigo 12.º, alínea a), da Diretiva [95/46/CE] exige que os Estados-Membros garantam um direito de acesso à informação sobre os destinatários ou categorias de destinatários dos dados pessoais e sobre o conteúdo dos dados divulgados não apenas relativamente ao presente mas também no que respeita ao passado. Cabe aos Estados-Membros fixar o prazo durante o qual essa informação deve ser conservada e o acesso correlativo a esta que representem um equilíbrio entre, por um lado, o interesse da pessoa em causa em proteger a sua privacidade, designadamente através das vias judiciais de intervenção e de recurso e, por outro, o ónus que a obrigação de conservar essa informação representa para o responsável pelo tratamento»*.

Tal é especialmente relevante se o artigo 7.º, alínea f), da Diretiva 95/46/CE for invocado por um responsável pelo tratamento de dados, no que diz respeito à anonimização de dados pessoais: o interesse legítimo do responsável pelo tratamento de dados deve ser sempre avaliado tendo em conta os direitos e liberdades fundamentais dos titulares dos dados.

Por exemplo, uma investigação realizada pela APD neerlandesa entre 2012 e 2013 sobre a utilização de tecnologias de inspeção exaustiva de pacotes de dados por quatro operadores das

[5] Tal significa que, em especial, deve ser efetuada uma avaliação de fundo, tendo em conta todas as circunstâncias relevantes, dando especial atenção aos seguintes fatores-chave:
a) a relação entre a finalidade para a qual os dados pessoais foram recolhidos e a finalidade do tratamento ulterior;
b) a contexto em que os dados pessoais foram recolhidos e as expectativas razoáveis dos titulares dos dados quanto à sua futura utilização;
c) a natureza dos dados pessoais e o impacto do tratamento ulterior sobre os titulares dos dados;
d) as garantias adotadas pelo responsável pelo tratamento de dados para assegurar o tratamento fiel dos dados e evitar qualquer impacto negativo sobre os titulares dos dados.

redes móveis revelou um fundamento legal ao abrigo do artigo 7.º, alínea f), da Diretiva 95/46/CE para a anonimização do conteúdo dos dados de tráfego o mais rapidamente possível após a recolha desses dados. Com efeito, o artigo 6.° da diretiva relativa à privacidade e comunicações eletrónicas estipula que os dados de tráfego relativos a assinantes e utilizadores tratados e armazenados pelo fornecedor de uma rede de comunicações públicas ou de serviços de comunicações eletrónicas publicamente disponíveis devem ser apagados ou tornados anónimos logo que possível. Neste caso, dado que é permitido ao abrigo do artigo 6.º da diretiva relativa à privacidade e comunicações eletrónicas, existe um fundamento legal correspondente no artigo 7.º da diretiva relativa à proteção dos dados. Tal pode igualmente apresentar-se de forma inversa: se um tipo de tratamento de dados não for permitido pelo artigo 6.º da diretiva relativa à privacidade e comunicações eletrónicas, não pode haver um fundamento legal no artigo 7.º da diretiva relativa à proteção dos dados.

2.2.2. Identificabilidade potencial dos dados anonimizados

O grupo de trabalho já abordou o conceito de dados pessoais em pormenor no parecer 4/2007 relativo a dados pessoais, centrando-se nos elementos constitutivos da definição contida no artigo 2.º, alínea a), da Diretiva 95/46/CE, incluindo a parte «identificada ou identificável» de tal definição. Neste contexto, o grupo de trabalho concluiu também que «dados anonimizados seriam dados anónimos que anteriormente se referiam a uma pessoa identificável cuja identificação deixou de ser possível».

O grupo de trabalho já esclareceu, por conseguinte, que o teste dos «meios (...) a utilizar razoavelmente» é recomendado pela diretiva como um critério a aplicar a fim de avaliar se o processo de anonimização de dados pessoais é suficientemente sólido, ou seja, se a identificação se tornou «razoavelmente» impossível. O contexto concreto e as circunstâncias de um caso específico influenciam diretamente a identificabilidade. No anexo técnico do presente parecer, é apresentada uma análise sobre o impacto da escolha da técnica mais adequada.

Conforme já destacado, a investigação, as ferramentas e a capacidade informática sofrem evoluções. Por conseguinte, não é possível nem útil fornecer uma enumeração exaustiva das circunstâncias em que a identificação deixa de ser possível. No entanto, alguns fatores-chave merecem ser considerados e exemplificados.

Em primeiro lugar, pode afirmar-se que os responsáveis pelo tratamento de dados se devem centrar nos meios concretos que seriam necessários para inverter a técnica de anonimização de dados pessoais, nomeadamente no que respeita ao custo e ao saber-fazer necessários para executar esses meios e à avaliação da sua probabilidade e gravidade. Por exemplo, devem estabelecer o equilíbrio entre o seu esforço anonimização de dados pessoais e os custos (em termos de tempo e recursos necessários) e a crescente disponibilidade no que se refere a meios técnicos de baixo custo para identificar as pessoas singulares nos conjuntos de dados, a crescente disponibilidade pública de outros conjuntos de dados (como os disponibilizados no âmbito de políticas de «dados abertos») e os muitos exemplos de anonimização incompleta de dados pessoais que comportam consequências adversas, e por vezes irreparáveis, para os titulares dos dados[6]. Salienta-se que o risco de identificação pode aumentar ao longo do tempo e que depende também do desenvolvimento das tecnologias da informação e comunicação.

[6] Curiosamente, as alterações do Parlamento Europeu ao projeto de regulamento geral sobre a proteção de dados apresentadas recentemente (21 de outubro de 2013) mencionam especificamente, no considerando 23, que «para verificar se os meios são suscetíveis de ser utilizados para identificar o indivíduo, há que tomar em consideração todos os fatores objetivos, tais como os custos e a quantidade de tempo necessários para a identificação, tendo em conta tanto a tecnologia disponível no momento do tratamento como a evolução tecnológica».

A regulamentação legal, se existir, deve, assim, ser formulada de uma forma tecnologicamente neutra e, idealmente, ter em conta as alterações nas potencialidades de desenvolvimento das tecnologias da informação.[7]

Em segundo lugar, «os meios suscetíveis de serem razoavelmente utilizados para determinar se uma pessoa é identificável» são os utilizados «pelo responsável pelo tratamento dos dados ou por qualquer outra pessoa». Por conseguinte, é fundamental perceber que quando um responsável pelo tratamento dos dados não elimina os dados originais (identificáveis) a nível do evento e entrega parte deste conjunto de dados (por exemplo, após remoção ou encobrimento de dados de identificação), o conjunto de dados daí resultante constitui, ainda assim, dados pessoais. Apenas quando o responsável pelo tratamento de dados agrega os dados para um nível em que cada evento deixa de ser identificável é que o conjunto de dados daí resultante pode ser classificado como anónimo. Por exemplo: se uma organização recolher dados sobre os movimentos de viagem de uma pessoa singular, os padrões de viagem dessa pessoa a nível de evento continuariam a ser classificados como dados pessoais para qualquer das partes enquanto o responsável pelo tratamento dos dados (ou qualquer outra parte) continuar a ter acesso aos dados brutos originais, mesmo que os identificadores diretos tenham sido retirados do conjunto fornecido a terceiros. Mas se o responsável pelo tratamento de dados excluir os dados brutos e apenas fornecer estatísticas agregadas a terceiros a um nível elevado, tais como «na segunda-feira, na trajetória X há mais 160 % de passageiros do que na terça-feira», estas informações serão consideradas como dados anónimos.

Uma solução eficaz de anonimização impede que qualquer uma das partes identifique uma pessoa num conjunto de dados, relacione dois registos num conjunto de dados (ou entre dois conjuntos de dados separados) e deduza quaisquer informações desse conjunto de dados. De um modo geral, por conseguinte, a eliminação de elementos identificadores diretos não é suficiente, por si só, para garantir que a identificação do titular dos dados deixa de possível. Será frequentemente necessário tomar medidas adicionais a fim de evitar a identificação, consoante, uma vez mais, o contexto e a finalidade do tratamento a que se destinam os dados anonimizados.

EXEMPLO:
Os perfis de dados genéticos são um exemplo de dados pessoais que podem correr um risco de identificação se a única técnica utilizada for a remoção da identidade do dador, dada a natureza singular de determinados perfis. A literatura existente já revelou[8] que a combinação de recursos genéticos publicamente disponíveis (por exemplo, registos genealógicos, obituários, resultados de consultas em motores de pesquisa) e os metadados relativos a dadores de ADN (momento da doação, idade, local de residência) é passível de revelar a identidade de determinadas pessoas, mesmo que o ADN em causa tinha sido doado de forma «anónima».

As duas famílias de técnicas de anonimização – aleatorização e generalização de dados[9] – têm limitações. No entanto, cada uma delas poderá ser adequada consoante as circunstâncias e o contexto para alcançar o objetivo desejado sem colocar em risco a privacidade das pessoas em causa. É importante esclarecer que se entende por «identificação» não só a possibilidade de obter o nome e/ou morada da pessoa, mas também a identificabilidade potencial por meio de individuação, ligação e inferência. Além disso, para que a legislação relativa à proteção de dados seja aplicável, são indiferentes as intenções do responsável pelo tratamento dos dados

[7] Ver parecer 4/2007 do Grupo de Trabalho do artigo 29.º, p. 15.
[8] Ver John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, Vol. 339, n.º 6117 (18 de janeiro de 2013), p. 262.
[9] As principais características e diferenças destas duas técnicas de anonimização de dados pessoais encontram-se descritas no ponto 3 («Análise Técnica»).

ou do destinatário. Desde que os dados sejam identificáveis, aplicam-se as regras de proteção de dados.

Sempre que um terceiro procede ao tratamento de um conjunto de dados processados com uma técnica de anonimização de dados pessoais (anonimizados e divulgados pelo responsável pelo tratamento dos dados original), pode efetuá-lo licitamente sem a necessidade de ter em consideração as exigências relativas à proteção de dados, desde que não lhe seja possível identificar (direta ou indiretamente) os titulares dos dados no conjunto de dados original. No entanto, os terceiros são obrigados a ter em conta quaisquer fatores contextuais e circunstanciais anteriormente referidos (incluindo as características específicas das técnicas de anonimização de dados pessoais aplicadas pelo responsável pelo tratamento de dados inicial) ao decidir como utilizar e, em especial, combinar tais dados anonimizados para fins próprios – dado que as consequências daí decorrentes podem implicar diferentes tipos de responsabilidade da sua parte. Sempre que tais fatores e características sejam suscetíveis de implicar um risco inaceitável de identificação dos titulares dos dados, o tratamento está sujeito, uma vez mais, à legislação de proteção de dados.

A lista anterior não pretende de forma alguma ser exaustiva, mas antes fornecer orientações gerais sobre o método de avaliação da eventual identificabilidade de um determinado conjunto de dados que seja submetido a anonimização de acordo com as diferentes técnicas disponíveis. Os fatores acima referidos podem ser considerados os vários fatores de risco a ponderar, tanto pelos responsáveis pelo tratamento de dados na anonimização de conjuntos de dados, como por terceiros na utilização desses conjuntos de dados «anonimizados» para os seus próprios fins.

2.2.3. Riscos da utilização de dados anonimizados

Ao ponderar a utilização de técnicas de anonimização, os responsáveis pelo tratamento dos dados devem ter em conta os seguintes riscos:

– Uma armadilha específica é considerar os dados sob pseudónimo equivalentes a dados anonimizados. O ponto relativo à análise técnica explica que os dados sob pseudónimo não podem ser equiparados a informações anónimas, uma vez que continuam a permitir que um titular de dados seja distinguido e passível de ser ligado entre diferentes conjuntos de dados. O uso de pseudónimos é suscetível de permitir a identificação e, por conseguinte, permanece dentro do âmbito de aplicação do regime jurídico de proteção de dados. Tal é especialmente relevante no contexto da investigação científica, estatística ou histórica[10].

EXEMPLO:
Um exemplo típico de ideias erradas sobre o uso de pseudónimos é o famoso «incidente AOL (América em linha)». Em 2006, foi divulgada ao público uma base de dados que continha vinte milhões de palavras-chave para efeitos de pesquisa para mais de 650 000 utilizadores durante um período de três meses, sendo a única medida de proteção de privacidade a substituição da identificação do utilizador AOL por um atributo numérico. Tal conduziu à identificação e localização públicas de algumas delas. As sequências de consultas num motor de pesquisa colocadas sob pseudónimo, especialmente se em conjunto com outros atributos, tais como endereços IP ou outros parâmetros de configuração do cliente, possuem uma capacidade de identificação muito elevada.

– Um segundo erro é considerar que os dados devidamente anonimizados (que tenham cumprido todas as condições e critérios acima referidos e que, por definição, não se encontrem abrangidos no âmbito de aplicação da Diretiva relativa à proteção de dados)

[10] Ver também parecer 4/2007, do Grupo de Trabalho do artigo 29.º, p. 18 a 20.

privam as pessoas de todas as garantias – acima de tudo, porque outros atos legislativos podem ser aplicáveis à utilização destes dados. Por exemplo, o artigo 5.º, n.º 3, da Diretiva relativa à privacidade e comunicações eletrónicas impede o armazenamento e acesso a «informações» de qualquer tipo (incluindo informações não pessoais) em equipamentos terminais sem o consentimento do assinante/utilizador, pois tal faz parte do princípio mais amplo da confidencialidade das comunicações.

– Pode também ocorrer uma terceira negligência se não for tido em conta o impacto nas pessoas, em determinadas circunstâncias, de dados devidamente anonimizados, em especial no caso da criação de perfis. A esfera da vida privada de uma pessoa singular encontra-se protegida pelo artigo 8.º da CEDH e pelo artigo 7.º da Carta dos Direitos Fundamentais da União Europeia. Deste modo, embora a legislação relativa à proteção de dados possa já não ser aplicável a este tipo de dados, a utilização dada a conjuntos de dados anonimizados e divulgados para utilização por terceiros é passível de originar a perda de privacidade. Há que proceder com especial prudência no tratamento de informações anonimizadas, em especial quando tais informações são utilizadas (frequentemente em conjunto com outros dados) para a tomada de decisões que produzem efeitos sobre as pessoas (mesmo que indiretamente). Tal como já referido no presente parecer e clarificado pelo grupo de trabalho, nomeadamente no seu parecer sobre o conceito de «limitação de finalidades» (parecer 03/2013)[11], devem ser avaliadas as expectativas legítimas dos titulares dos dados relativamente ao tratamento posterior dos dados que lhes dizem respeito tendo em conta os fatores contextuais relevantes – como a natureza da relação entre os titulares dos dados e os responsáveis pelo tratamento dos dados, as obrigações legais aplicáveis, a transparência das operações de tratamento.

3 Análise técnica, robustez das tecnologias e erros típicos

Existem diferentes técnicas e práticas de anonimização, com graus variáveis de robustez. O presente ponto incide sobre os principais elementos a considerar pelos responsáveis pelo tratamento de dados na sua aplicação, tendo em conta, nomeadamente, a garantia possível oferecida por determinada técnica atendendo ao estado da tecnologia atual e tendo em conta três riscos que são fundamentais para a anonimização:

- *Identificação*, que corresponde à possibilidade de isolar alguns ou todos os registos que identifiquem uma pessoa num conjunto de dados;
- *Possibilidade de ligação*, que representa a capacidade de ligar pelo menos dois registos sobre a mesma pessoa ou um grupo de pessoas em causa (tanto na mesma base de dados, como em duas bases de dados diferentes). Se um intruso conseguir estabelecer (por exemplo, através da análise de correlação) que dois registos se encontram atribuídos a um mesmo grupo de pessoas, mas não conseguir selecionar pessoas desse grupo, a técnica fornece resistência contra «identificação», mas não contra a possibilidade de ligação;
- *Inferência*, que é a possibilidade de deduzir, com uma probabilidade significativa, o valor de um atributo a partir dos valores de um conjunto de outros atributos.

Assim, uma solução contra estes três riscos seria robusta face a uma reidentificação efetuada pela via mais provável e razoável passível de ser utilizada pelo responsável pelo tratamento de dados e por terceiros. O grupo de trabalho salienta, a este respeito, que as técnicas de

[11] Disponível em http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

desidentificação e anonimização ainda estão a ser objeto de investigação e que esta investigação tem vindo a mostrar sistematicamente que nenhuma técnica é, por si só, desprovida de lacunas. Em termos gerais, existem duas abordagens distintas de anonimização de dados pessoais: a primeira tem por base a ***aleatorização***, enquanto a segunda se baseia na ***generalização***. O parecer também aborda outros conceitos, como a *utilização de pseudónimos*, *privacidade diferencial*, *l-diversidade* e *t-proximidade*.

O presente parecer utiliza o seguinte vocabulário nesta secção: um conjunto de dados é composto por registos diferentes relativos a pessoas singulares (os titulares dos dados). Cada registo está relacionado com um titular de dados e é constituído por um conjunto de valores (ou «entradas», por exemplo: 2013) para cada atributo (por exemplo, ano). Um conjunto de dados é um conjunto de registos que pode assumir a forma duma tabela (ou um conjunto de tabelas) ou de um gráfico anotado/ponderado, como é cada vez mais frequente hoje em dia. Os exemplos no parecer referem-se a tabelas, mas são igualmente aplicáveis a outras representações gráficas de registos. As combinações de atributos relacionados com um titular dos dados ou um grupo de titulares dos dados podem ser designadas por quase-identificadores. Em alguns casos, um conjunto de dados pode ter vários registos sobre a mesma pessoa. Um «intruso» é um terceiro (isto é, outro que não o responsável pelo tratamento de dados nem o subcontratante) que aceda aos registos originais, quer acidental quer intencionalmente.

3.1. Aleatorização

A aleatorização é uma família de técnicas que altera a veracidade dos dados a fim de eliminar a estreita ligação entre os dados e a pessoa. Se os dados forem suficientemente imprecisos já não poderão ser relacionados com uma pessoa específica. A aleatorização não reduz, por si só, a singularidade de cada registo, uma vez que cada registo continua a ser proveniente de um único titular dos dados, mas é passível de proteger contra ataques ou riscos de inferência e pode ser combinada com técnicas de generalização a fim de fornecer garantias de privacidade mais sólidas. Pode ser necessária a aplicação de técnicas suplementares para garantir que um registo não é passível de identificar um indivíduo em particular.

3.1.1. Adição de ruído

A técnica de adição de ruído é especialmente útil quando os atributos são passíveis de ter um grande efeito adverso sobre as pessoas e consiste em modificar atributos no conjunto de dados de modo a este serem menos precisos, enquanto se mantém a distribuição global. Ao efetuar o tratamento de um conjunto de dados, um observador irá presumir que os valores são exatos, o que só será verdade até um certo nível. Por exemplo, se a altura de uma pessoa tiver sido originalmente medida até ao centímetro mais próximo, o conjunto de dados anonimizados pode conter uma altura com uma precisão arredondada ao intervalo de 10 cm mais próximo. Se esta técnica for aplicada eficazmente, um terceiro não conseguirá identificar uma determinada pessoa, nem tão pouco conseguirá reparar os dados ou detetar de que forma estes foram alterados.

Frequentemente a adição de ruído necessita de ser combinada com outras técnicas de anonimização de dados pessoais, tais como a remoção de atributos evidentes e de quase-identificadores. O nível de ruído deve depender da necessidade do nível de informação exigido e do impacto na privacidade das pessoas em resultado da divulgação dos atributos protegidos.

3.1.1.1. Garantias

- *Identificação:* Continua a ser possível selecionar os registos de uma pessoa (eventualmente de uma forma não identificável), embora os registos sejam menos fiáveis.
- *Possibilidade de ligação:* Continua a ser possível efetuar a ligação entre os registos de uma mesma pessoa, mas os registos são menos fiáveis e, por conseguinte, um registo real é passível de ser associado a um registo alterado (ou seja, ao «ruído»). Em alguns casos, uma atribuição incorreta é suscetível de expor um titular dos dados a um nível de risco significativo e até mesmo mais elevado do que uma atribuição correta.
- *Inferência:* Os ataques de inferência são eventualmente possíveis, mas a taxa de sucesso será inferior e são plausíveis alguns falsos positivos (e falsos negativos).

3.1.1.2. Erros comuns

- Adição de ruído incoerente: Se o ruído não for semanticamente viável (ou seja, se estiver «fora de escala» e não respeitar a lógica entre atributos num conjunto), um intruso que tenha acesso à base de dados conseguirá filtrá-lo e, em alguns casos, gerar novamente as entradas em falta. Além disso, se o conjunto de dados for demasiado disperso[12], a ligação entre as entradas de dados com ruído e uma fonte externa continuará a ser possível.
- Partir do pressuposto que a adição de ruído é suficiente: a adição de ruído é uma medida complementar que dificulta a um intruso a recuperação de dados pessoais. A menos que o ruído seja mais elevado do que a informação contida no conjunto de dados, não se deve presumir que a adição de ruído representa uma solução autónoma para a anonimização de dados pessoais.

3.1.1.3. Falhas da adição de ruído

Uma experiência muito conhecida de reidentificação foi a efetuada com a base de dados dos clientes do fornecedor de conteúdos vídeo Netflix. Os investigadores analisaram as propriedades geométricas dessa base de dados, constituída por mais de 100 milhões de classificações numa escala de 1 a 5 sobre mais de 18 000 filmes, expressas por quase 500 000 utilizadores, divulgada publicamente pela empresa após terem sido «anonimizadas» segundo uma política interna de privacidade, tendo sido removidas todas as informações identificativas dos clientes, à exceção das classificações e datas. Foi adicionado ruído aumentando ou diminuindo ligeiramente as classificações.

Apesar disso, verificou-se que 99 % dos registos dos utilizadores eram passíveis de serem identificados singularmente utilizando como critérios de seleção 8 classificações e datas com erros de 14 dias, ao passo que a diminuição do critério de seleção (2 classificações e um erro de 3 dias) continuava a permitir identificar 68 % dos utilizadores[13].

[12] Este conceito é mais aprofundado no Anexo, p. 30.

[13] Narayanan, A., e Shmatikov, V. (2008, maio). Robust de-anonymization of large sparse datasets. Em *Segurança e Privacidade, 2008. SP 2008. Simpósio do IEEE* (p. 111-125). IEEE.

3.1.2. Permutação

Esta técnica consiste em misturar aleatoriamente os valores dos atributos numa tabela, de modo a que alguns destes sejam ligados artificialmente a titulares de dados diferentes. É útil quando é importante manter a distribuição exata de cada atributo no conjunto de dados.

A permutação pode ser considerada como uma forma especial de adição de ruído. Numa técnica de ruído clássica, os atributos são modificados com valores aleatórios. A geração de ruído consistente pode ser uma tarefa difícil e a alteração ligeira dos valores dos atributos pode não conceder a devida privacidade. Em alternativa, as técnicas de permuta alteram valores existentes no conjunto de dados através da sua simples troca de um registo para o outro. Essa troca irá garantir que o alcance e a distribuição dos valores permanecem iguais, mas que as correlações entre os valores e as pessoas não. Se dois ou mais atributos tiverem uma relação lógica ou uma correlação estatística e forem permutados independentemente, tal relação será destruída. Por conseguinte, pode ser importante permutar um conjunto de atributos conexos de modo a não romper a relação lógica, caso contrário, um intruso poderá identificar os atributos permutados e inverter a permutação.

Por exemplo, se considerarmos um subconjunto de atributos de um conjunto de dados médicos, tais como «causas de hospitalização/sintomas/serviço responsável», uma forte relação lógica irá ligar os valores na maioria dos casos e a permutação de apenas um dos valores será, por conseguinte, detetada e será até passível de ser revertida.

À semelhança da adição de ruído, a permutação é suscetível de não fornecer, por si só, a anonimização de dados pessoais e deve ser sempre combinada com a remoção dos atributos e quase-identificadores óbvios.

3.1.2.1. Garantias

- *Identificação:* Tal como acontece com a adição de ruído, continua a ser possível selecionar os registos de uma pessoa, mas os registos são menos fiáveis.
- *Possibilidade de ligação:* Se a permutação afetar atributos e quase-identificadores, pode evitar a ligação «correta» de atributos a um conjunto de dados tanto a nível interno como externo, mas ainda assim possibilitar uma ligação «incorreta», dado que uma entrada real pode ser associada a um titular de dados diferente.
- *Inferência:* Continua a ser possível formular inferências a partir do conjunto de dados, especialmente se os atributos estiverem correlacionados ou tiverem relações lógicas fortes. Todavia, sem saber quais atributos foram permutados, o intruso tem de considerar que a sua inferência se baseia numa hipótese errada e, por conseguinte, apenas permanece a possibilidade de formular uma inferência probabilística.

3.1.2.2. Erros comuns

- Selecionar o atributo errado: a permutação dos atributos não sensíveis ou não arriscados não irá conduzir a um ganho significativo em termos de proteção de dados pessoais. Com efeito, se os atributos confidenciais ou arriscados permanecessem associados ao atributo inicial, um intruso ainda conseguiria extrair informações confidenciais sobre indivíduos.
- Permutar aleatoriamente atributos: Se dois atributos se encontrarem fortemente correlacionados, a permutação aleatória dos atributos não proporciona garantias sólidas. Este erro comum é exemplificado na Tabela 1.

- Pressupor que a permutação é suficiente: À semelhança da adição de ruído, a permutação não proporciona, por si só, anonimato e deve ser combinada com outras técnicas, como a remoção dos atributos óbvios.

3.1.2.3. Falhas da permutação

Este exemplo mostra como a permutação aleatória de atributos origina poucas garantias de privacidade quando existem ligações lógicas entre diferentes atributos. Na sequência da tentativa de anonimização, é fácil deduzir o rendimento de cada pessoa consoante o trabalho (e o ano de nascimento). Por exemplo, é possível afirmar, através da verificação direta dos dados, que o diretor-geral constante da tabela nasceu muito provavelmente em 1957 e aufere do salário mais elevado, enquanto o desempregado nasceu em 1964 e aufere o rendimento mais baixo.

Ano	Sexo	Cargo	Rendimento (permutado)
1957	H	Engenheiro	70 mil
1957	H	Diretor Geral	5 mil
1957	H	Desempregado	43 mil
1964	H	Engenheiro	100 mil
1964	H	Administrador	45 mil

Tabela 1. Um exemplo de anonimização ineficaz de dados pessoais através da permutação de atributos correlacionados

3.1.3. Privacidade diferencial

A privacidade diferencial[14] inclui-se na família de técnicas de aleatorização, com uma abordagem diferente: enquanto, na verdade, a inserção de ruído se aplica previamente à divulgação do conjunto de dados, a privacidade diferencial é passível de ser utilizada quando o responsável pelo tratamento de dados gera visualizações anonimizadas de um conjunto de dados, conservando uma cópia dos dados originais. Essas visualizações anonimizadas seriam normalmente geradas através de um subconjunto de consultas para um terceiro em especial. O subconjunto inclui algum ruído aleatório *ex post*, deliberadamente adicionado. A privacidade diferencial informa o responsável pelo tratamento de dados quanto e qual a forma de ruído que este tem de acrescentar para obter as garantias de privacidade necessárias[15]. Neste contexto, será especialmente importante controlar continuamente (pelo menos a cada nova consulta), qualquer possibilidade de identificação de uma pessoa no conjunto de resultados da consulta. Há que esclarecer, no entanto, que as técnicas de privacidade diferencial não irão alterar os dados originais e, por conseguinte, desde que os dados originais se mantenham, o responsável pelo tratamento de dados consegue identificar as pessoas singulares em resultados de consultas de privacidade diferencial, tendo em conta todos os meios que possam ser razoavelmente utilizados. Esses resultados devem igualmente ser considerados dados pessoais.

Uma das vantagens de uma abordagem baseada na privacidade diferencial assenta no facto de os conjuntos de dados serem fornecidos a terceiros autorizados, em resposta a uma consulta específica e não através da divulgação de um conjunto de dados único. Para facilitar a

[14] Dwork, C. (2006). Privacidade diferencial. Em *Automata, languages and programming* (p. 1-12). Springer Berlin Heidelberg.
[15] Cf. Ed Felten (2012), Protecting privacy by adding noise. URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

auditoria, o responsável pelo tratamento de dados pode conservar uma lista de todas as consultas e pedidos de informação para se certificar de que não há acesso de terceiros não autorizados aos dados. Uma consulta pode igualmente ser submetida a técnicas de anonimização, incluindo a adição de ruídos ou a substituição, para uma maior proteção de privacidade. Ainda se encontra em processo de investigação a procura de um mecanismo de consulta satisfatório e interativo que consiga responder a quaisquer questões com razoável precisão (ou seja, de um modo menos ruidoso) e preserve, simultaneamente, a privacidade.

Para limitar ataques de inferência e tentativas de estabelecer ligações, é necessário acompanhar de perto as consultas efetuadas por uma entidade e observar as informações obtidas sobre os titulares dos dados. Por conseguinte, as bases de dados de «privacidade diferencial» não devem ser implantadas em motores de busca abertos que não ofereçam qualquer possibilidade de rastreio das entidades consultoras.

3.1.3.1 Garantias

- *Identificação:* Se apenas forem extraídas estatísticas e se as regras aplicadas ao conjunto forem devidamente escolhidas, não deverá ser possível utilizar as respostas para identificar uma pessoa.
- *Possibilidade de ligação:* Ao utilizar vários pedidos, poderá ser possível ligar as entradas relativas a uma pessoa específica a partir de duas respostas.
- *Inferência:* É possível inferir informações sobre pessoas ou grupos através da utilização de pedidos múltiplos.

3.1.3.2. Erros comuns

- Não inserir ruído suficiente: Para evitar a possibilidade de ligação com informações de base, o desafio é fornecer o mínimo de elementos sobre a hipótese de um titular dos dados ou um grupo de titulares dos dados específicos terem ou não contribuído para o conjunto de dados. De uma perspetiva de proteção de dados, a maior dificuldade é conseguir produzir a quantidade adequada de ruído a adicionar às respostas verdadeiras, de modo a proteger a privacidade das pessoas, e ainda assim preservar a utilidade das respostas divulgadas.

3.1.3.3 Falhas da privacidade diferencial

Tratar cada consulta de forma independente: Uma combinação de resultados de consulta é suscetível de permitir a divulgação de informações que se queriam confidenciais. Se não for conservado um histórico de consulta, um intruso poderá elaborar múltiplas questões a uma base de dados de «privacidade diferencial» que reduzam progressivamente a amplitude da amostra extraída até que, eventualmente, surja uma especificidade de uma única pessoa ou de um grupo de pessoas, deterministicamente ou com uma probabilidade muito elevada. Além disso, adverte-se ainda para o erro de pensar que os dados são anónimos para terceiros, quando o responsável pelo tratamento dos dados continua a poder identificar o titular dos dados na base de dados original, tendo em conta todos os meios que possam razoavelmente ser utilizados.

3.2. Generalização

A generalização é a segunda família de técnicas de anonimização. Esta abordagem consiste em generalizar, ou diluir, os atributos dos titulares dos dados através da alteração da respetiva escala ou ordem de grandeza (isto é, uma região em vez de uma cidade, um mês em vez de

uma semana). Embora a generalização possa ser eficaz para impedir a identificação, não permite a anonimização efetiva em todos os casos; requer, em particular, abordagens quantitativas específicas e sofisticadas para evitar a possibilidade de ligação e inferência.

3.2.1. Agregação e k-anonimato

As técnicas de agregação e k-anonimato visam impedir que um titular dos dados seja selecionado através do agrupamento com, pelo menos, outras k pessoas. Para este efeito, os valores dos atributos são generalizados de modo a que cada pessoa partilhe o mesmo valor. Por exemplo, ao reduzir a granularidade de um local de uma cidade para um país é incluído um maior número de pessoas. As datas de nascimento individuais podem ser generalizadas num intervalo de datas ou agrupadas por mês ou ano. Outros atributos numéricos (por exemplo, salários, peso, altura ou a dosagem de um medicamento) podem ser generalizados por intervalos de valores (por exemplo, salário de 20 000 a 30 000 €). Estes métodos podem ser utilizados quando a correlação entre valores pontuais de atributos é passível de criar quase-identificadores.

3.2.1.1. Garantias

- *Identificação:* Dado que os mesmos atributos são agora partilhados por k utilizadores, deixa de ser possível selecionar uma pessoa dentro de um grupo de k utilizadores.
- *Possibilidade de ligação:* Embora a possibilidade de ligação seja limitada, continua a ser possível estabelecer a ligação entre registos através de grupos de k utilizadores. Então, dentro deste grupo, a probabilidade de que dois registos correspondam aos mesmos pseudo-identificadores é de 1/k (que pode ser significativamente mais elevada do que a probabilidade de tais entradas não serem passíveis de ligação).
- *Inferência:* O principal defeito do modelo k-anonimato é não impedir qualquer tipo de ataques de inferência. Com efeito, se todos os indivíduos k se encontrarem num mesmo grupo, se for sabido a que grupo pertence uma determinada pessoa, é simples recuperar o valor desta propriedade.

3.2.1.2. Erros comuns

- Falta de alguns quase-identificadores: Um parâmetro fundamental ao considerar o k-anonimato é o limite de k. Quanto maior for o valor de k, mais elevadas serão as garantias de privacidade. Um erro comum é aumentar artificialmente o valor k reduzindo o conjunto considerado de quase-identificadores. A redução de quase-identificadores facilita a criação de grupos de utilizadores k devido ao poder inerente de identificação associado aos outros atributos (especialmente se alguns deles forem confidenciais ou possuírem uma entropia muito elevada, como no caso de atributos muito raros). O facto de não considerar todos os quase-identificadores ao selecionar o atributo para generalizar é um erro grave. Se alguns atributos puderem ser utilizados para identificar uma pessoa num agrupamento de k, então a generalização não protegerá algumas pessoas (ver exemplo na Tabela 2).
- Valor baixo de k: Visar um valor baixo de k é igualmente problemático. Se k for demasiado pequeno, o peso de cada pessoa num conjunto é demasiado significativo e os ataques de inferência terão uma maior taxa de sucesso. Por exemplo, se $k = 2$, a probabilidade das duas pessoas partilharem a mesma propriedade é maior do que se $k > 10$.

- Não agrupar pessoas com o mesmo peso: O agrupamento de um conjunto de pessoas com uma distribuição de atributos desigual também pode ser problemático. O impacto do registo de uma pessoa num conjunto de dados irá variar: algumas representarão uma parte significativa das entradas, enquanto as contribuições de outros se irão manter relativamente insignificantes. Por conseguinte, é importante garantir que k seja suficientemente elevado para que nenhuns indivíduos representem uma fração demasiado importante das entradas num grupo.

3.1.3.3. Falhas do k-anonimato

O principal problema do k-anonimato é não impedir ataques de inferência. No exemplo que se segue, se o intruso souber que uma pessoa específica se encontra no conjunto de dados e que nasceu em 1964, saberá também que essa pessoa teve um enfarte. Mais ainda, se for sabido que esse conjunto de dados foi obtido de uma organização francesa, então cada pessoa reside em Paris, já que os três primeiros números dos códigos postais parisienses são 750*).

Ano	Sexo	Código Postal	Diagnóstico
1957	H	750*	Enfarte
1957	H	750*	Colesterol
1957	H	750*	Colesterol
1964	H	750*	Enfarte
1964	H	750*	Enfarte

Tabela 2: Um exemplo de k-anonimização fracamente elaborada

3.2.2. L-diversidade/t-proximidade

A l-diversidade amplia o k-anonimato para garantir que os ataques determinísticos de inferência deixam de ser possíveis, ao garantir que em cada classe de equivalência cada atributo tem, pelo menos, *l* valores diferentes.

Um objetivo fundamental a alcançar é limitar a ocorrência de classes de equivalência com fraca variabilidade do atributo, para que o intruso que tenha conhecimentos de base sobre um titular de dados específico permaneça sempre com um grau significativo de incerteza.

A l-diversidade é útil para proteger os dados contra ataques de inferência quando os valores dos atributos se encontram bem distribuídos. É de salientar, no entanto, que esta técnica não consegue impedir a fuga de informações se os atributos existentes numa segmentação forem distribuídos de forma desigual ou pertencerem a uma diversidade reduzida de valores ou de significados semânticos. Em suma, a l-diversidade está sujeita a ataques de inferência probabilísticos.

A t-proximidade é um refinamento da l-diversidade, na medida em que visa criar classes equivalentes que se assemelhem à distribuição inicial de atributos na tabela. Esta técnica é útil quando é importante manter os dados tão próximo quanto possível do original. Para esse efeito, é colocada uma nova restrição na classe de equivalência, designadamente, que devem existir não só pelo menos *l* valores diferentes em cada classe de equivalência, mas também

que cada valor é representado as vezes que forem necessárias para refletir a distribuição inicial de cada atributo.

3.2.2.1. Garantias

- *Identificação:* Tal como o k-anonimato, a l-diversidade e a t-proximidade podem garantir que os dados relativos a uma pessoa não são passíveis de serem selecionados na base de dados.
- *Possibilidade de ligação:* a l-diversidade e a t-proximidade não são uma melhoria do k-anonimato no que respeita à impossibilidade de ligação. O problema é o mesmo que o que ocorre com qualquer grupo: a probabilidade de que as mesmas entradas pertençam ao mesmo titular dos dados é superior a 1/n (em que n é o número dos titulares dos dados na base de dados).
- *Inferência:* A melhoria principal da l-diversidade e da t-proximidade em relação ao k-anonimato é que deixa de ser possível criar ataques de inferência contra uma base de dados «l-diversificada» ou «t-próxima» com um grau de 100 % de certeza.

3.2.2.2. Erros comuns

- Proteger os valores dos atributos confidenciais misturando-os com os outros atributos confidenciais: Não basta ter dois valores de um atributo num grupo para oferecer garantias de privacidade. Com efeito, a distribuição de valores confidenciais em cada grupo deve assemelhar-se à distribuição desses valores na população total ou, pelo menos, deve ser uniforme em todo o grupo.

3.2.2.3. Falhas da l-diversidade

Na tabela que se segue, é aplicada l-diversidade no que respeita ao atributo «Diagnóstico». No entanto, sabendo que um indivíduo nascido em 1964 se encontra nesta tabela, é ainda possível presumir-se com um grau de probabilidade muito elevado que ele terá sido vítima de um enfarte.

Ano	Sexo	Código Postal	Diagnóstico
1957	H	750*	**Enfarte**
1957	H	750*	**Colesterol**
1957	H	750*	**Colesterol**
1957	H	750*	**Colesterol**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**
1964	H	750*	**Colesterol**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**
1964	H	750*	**Enfarte**

Tabela 3 Uma tabela l-diversificada em que os valores «Diagnóstico» não se encontram distribuídos de modo uniforme

Nome	Data de nascimento	Sexo
Smith	1964	H
Rossi	1964	H
Dupont	1964	H
Jansen	1964	H
Garcia	1964	H

Tabela 4. Sabendo que essas pessoas se encontram na Tabela 3, um intruso poderia inferir que foram vítimas de um enfarte

4. Utilização de pseudónimos

A utilização de pseudónimos consiste em substituir um atributo de um registo (normalmente um atributo único) por outro. A pessoa singular continua, portanto, suscetível de ser indiretamente identificada; por conseguinte, a utilização de pseudónimos, quando usada isoladamente, não irá originar um conjunto de dados anónimo. No entanto, é abordada no presente parecer devido às numerosas ideias erradas e equívocos relacionados com a sua utilização.

A utilização de pseudónimos reduz a possibilidade de ligação entre um conjunto de dados e a identidade original de um titular; por conseguinte, é uma medida de segurança útil, mas não um método de anonimização de dados pessoais.

O resultado da utilização de pseudónimos pode ser independente do valor inicial (como é o caso de um número aleatório gerado pelo responsável pelo tratamento de dados ou de um apelido escolhido pelo titular de dados) ou derivar dos valores iniciais de um atributo ou conjunto de atributos, por exemplo, uma função *hash* ou um sistema de cifragem.

As técnicas de utilização de pseudónimos mais utilizadas são as seguintes:

- Cifragem com chave secreta: neste caso, o titular da chave pode voltar a identificar cada titular de dados de forma simples através da descodificação do conjunto de dados, pois os dados pessoais permanecem no conjunto de dados, embora de forma codificada. Partindo do princípio de que foi aplicado um sistema de cifragem de última geração, a descodificação só é possível mediante o conhecimento da chave.
- Função *hash*: corresponde a uma função que devolve um resultado de dimensão fixa a partir de uma entrada de qualquer dimensão (a entrada pode consistir num único atributo ou num conjunto de atributos) e não pode ser invertida. Isto significa que o risco de inversão verificado na cifragem deixa de existir. No entanto, se a gama de valores de entrada da função *hash* for conhecida, estes poderão ser novamente submetidos à função *hash* a fim de obter o valor correto de um registo em particular. Por exemplo, se um conjunto de dados tiver sido objeto da utilização de pseudónimos através da aplicação da função *hash* ao número de identificação nacional, o valor pode ser derivado através da simples utilização da função *hash* em todos os valores possíveis de entrada e da comparação do resultado com os valores constantes no conjunto de dados. As funções *hash* são habitualmente concebidas para serem de cálculo rápido e estão sujeitas a ataques de força bruta[16]. É também possível criar tabelas pré-computorizadas para permitir a inversão em massa de um amplo conjunto de valores de *hash*.

  A utilização de uma função *hash* com uma variável criptográfica (em que um valor aleatório, designado por «variável criptográfica», é adicionado ao atributo a ser dividido [*hashed*]) é passível de reduzir a probabilidade da determinação do valor de entrada mas, ainda assim, continua a ser possível de efetuar, mediante os meios razoáveis, o cálculo do valor original do atributo escondido por detrás do resultado de uma função *hash* com uma variável criptográfica[17].

[16] Tais ataques consistem em tentar todos os dados de entrada plausíveis a fim de construir tabelas de correspondência.

[17] Especialmente se o tipo do atributo for conhecido (nome, número de segurança social, data de nascimento, etc.). Para adicionar um requisito computacional, seria possível contar com uma função de *hash* de derivação de

- Função *hash* com chave armazenada: corresponde a uma função *hash* especial que utiliza uma chave secreta como entrada suplementar (o que difere da função *hash* com uma variável criptográfica, já que a variável criptográfica habitualmente não é secreta). Um responsável pelo tratamento de dados pode voltar a reproduzir a função no atributo utilizando a chave secreta, mas é muito mais difícil para um intruso voltar a reproduzir a função sem conhecer a chave, já que o número de hipóteses a tentar é suficientemente grande para ser inexequível.
- Cifragem determinística ou função *hash* com chave, com eliminação da chave: esta técnica pode ser vista como a seleção de um número aleatório como pseudónimo para cada atributo na base de dados, eliminando-se, de seguida, a tabela de correspondência. Esta solução permite[18] diminuir o risco de possibilidade de ligação entre os dados pessoais no conjunto de dados e os relativos à mesma pessoa noutro conjunto de dados em que é utilizado um pseudónimo diferente. Tendo em conta um algoritmo de última geração, será difícil em termos informáticos que um intruso decifre ou reproduza novamente a função, pois isso implicaria tentar todas as chaves possíveis, dado que a chave não se encontra disponível.
- Utilização de dispositivos de autenticação (*tokens*): esta técnica é normalmente aplicada no setor financeiro (mas não só) para substituir o número de identificação de cartões por valores que têm uma utilidade reduzida para um intruso. Deriva dos anteriores e baseia-se habitualmente na aplicação de mecanismos de cifragem unidirecional ou na atribuição, por meio de uma função de indexação, de um número sequencial ou gerado aleatoriamente que não derive matematicamente dos dados originais.

4.1. Garantias

- *Identificação:* Continua a ser possível identificar registos de pessoas, uma vez que a pessoa continua a ser identificada por um atributo único, resultante da função de utilização de pseudónimos (= atributo sob pseudónimo).
- *Possibilidade de ligação:* A possibilidade de ligação continuará a ser simples entre registos que utilizem o mesmo atributo com pseudónimo para se referirem à mesma pessoa. Mesmo que sejam utilizados atributos sob pseudónimos diferentes para o mesmo titular de dados, continua a ser possível a ligação por meio de outros atributos. Apenas se mais nenhum atributo no conjunto de dados for passível de ser utilizado para identificar a pessoa em causa e se qualquer ligação entre o atributo original e o atributo sob pseudónimo tiver sido eliminada (incluindo pela supressão dos dados originais), só então não haverá qualquer correlação óbvia entre dois conjuntos de dados que utilizem diferentes atributos sob pseudónimos.
- *Inferência:* Os ataques de inferência à verdadeira identidade do titular dos dados são possíveis no conjunto de dados ou através de diferentes bases de dados que usem o mesmo atributo sob pseudónimo para um indivíduo, ou se os pseudónimos forem por si só esclarecedores e não disfarçarem devidamente a identidade original do titular dos dados.

---

chaves, sempre que o valor calculado é codificado (*hashed*) várias vezes com uma variável criptográfica (salga) curta.

[18] Dependendo dos outros atributos no conjunto de dados e da supressão dos dados originais.

4.2. Erros comuns

- Supor que um conjunto de dados sob pseudónimo se encontra anonimizado: Os responsáveis pelo tratamento de dados pressupõem frequentemente que a remoção ou substituição de um ou mais atributos é suficiente para tornar o conjunto de dados anónimo. Muitos exemplos têm vindo a demonstrar que tal não ocorre. A simples alteração da identificação não impede que alguém identifique um titular de dados se permanecerem quase-identificadores no conjunto de dados, ou se os valores de outros atributos ainda forem passíveis de identificar uma pessoa. Em muitos casos, pode ser tão fácil identificar uma pessoa a partir de um conjunto de dados que tenha sido objeto da utilização de pseudónimos como o seria a partir dos dados originais. Há que tomar medidas adicionais para poder considerar o conjunto de dados anonimizado, nomeadamente remover e generalizar atributos ou excluir os dados originais ou, pelo menos, reuni-los num nível mais agregado.
- Erros comuns na utilização de pseudónimos como técnica de redução da possibilidade de ligação:
    - Utilizar a mesma chave em bases de dados diferentes: a eliminação da possibilidade de ligação de diferentes conjuntos de dados depende fortemente da utilização de um algoritmo com chave e do facto de um único indivíduo corresponder a diferentes atributos sob pseudónimo em contextos diferentes. Assim, é importante evitar a utilização de uma mesma chave em bases de dados diferentes para conseguir reduzir a possibilidade de ligação.
    - Utilização de chaves diferentes («chaves alternadas») para utilizadores diferentes: pode ser tentador utilizar chaves diferentes para diferentes grupos de utilizadores e alterar a chave com base na utilização (por exemplo, utilizar a mesma chave para registar 10 entradas relativas ao mesmo utilizador). No entanto, se esta operação não for devidamente executada, será passível de desencadear a ocorrência de padrões, o que reduzirá parcialmente os benefícios a que se propõe. Por exemplo, alternar a chave através de regras específicas para pessoas específicas poderá facilitar a possibilidade de ligação das entradas correspondentes a determinadas pessoas. Além disso, o desaparecimento de um dado sob pseudónimo repetido na base de dados no momento em que surge um novo dado pode indicar que ambos os registos se referem à mesma pessoa singular.
    - Guardar a chave: se a chave secreta for armazenada junto aos dados com pseudónimo e os dados forem comprometidos, o intruso poderá facilmente ligar os dados sob pseudónimo ao respetivo atributo inicial. O mesmo é verdade se a chave for armazenada separadamente dos dados mas não de uma forma segura.

4.3. Limitações da utilização de pseudónimos

- *Cuidados de saúde*

1. Nome, morada e data de nascimento	2. Período de benefício de assistência especial	3. Índice de massa corporal	6. N.º de referência do grupo de investigação.
[redacted]	< 2 anos	15	QA5FRD4
	> 5 anos	14	2B48HFG
	< 2 anos	16	RC3URPQ
	> 5 anos	18	SD289K9
	< 2 anos	20	5E1FL7Q

Tabela 5. Um exemplo de utilização de pseudónimos através da função *hash* (nome, morada e data de nascimento) passível de ser facilmente revertida

Foi criado um conjunto de dados a fim de analisar a relação entre o peso de uma pessoa e o pagamento de um benefício de assistência especial. O conjunto de dados inicial incluía o nome, morada e data de nascimento dos titulares dos dados, que foram suprimidos. O número de referência do grupo de investigação foi gerado a partir dos dados eliminados através da utilização de uma função *hash*. Embora o nome, morada e data de nascimento tenham sido excluídos da tabela, se o nome, morada e data de nascimento de um titular de dados forem conhecidos a par da função *hash* utilizada, é fácil calcular os números de referência do grupo de investigação.

- *Redes sociais*

Foi demonstrado[19] que podem ser extraídas informações confidenciais sobre indivíduos específicos a partir de gráficos de rede social, apesar das técnicas de «utilização de pseudónimos» aplicadas a esses dados. Um fornecedor de uma rede social presumiu erradamente que a utilização de pseudónimos tinha robustez para impedir a identificação após a venda dos dados a outras empresas para fins comerciais e publicitários. Em vez dos nomes reais, o prestador utilizou alcunhas, mas esta ação não foi evidentemente suficiente para tornar anónimos os perfis de utilizador, uma vez que as relações entre as diferentes pessoas são únicas e podem ser utilizadas como agente identificador.

- *Localizações*

Os investigadores do MIT[20] analisaram recentemente um conjunto de dados sob pseudónimos constituído por 15 meses de coordenadas de mobilidade espácio-temporal de 1,5 milhões de pessoas num território num raio de 100 km. Esta análise revelou que 95 % da população era passível de ser identificada com quatro pontos de localização e que bastavam apenas dois pontos para identificar mais de 50 % dos titulares dos dados (um desses pontos é conhecido, sendo muito provavelmente o «domicílio» ou «escritório») com pouca margem para a proteção de privacidade, mesmo que a identidade dos indivíduos fosse objeto da utilização de pseudónimos através da substituição dos seus verdadeiros atributos [...] por outras classificações.

[19] A. Narayanan e V. Shmatikov, «De-anonymizing social networks», trigésimo Simpósio IEEE sobre Segurança e Privacidade, 2009.
[20] Y.-A. de Montjoye, C. Hidalgo, Verleysen M. e V. Blondel, «Unique in the Crowd: The privacy bounds of human mobility» Nature, n.º 1376, 2013.

5. Conclusões e recomendações

5.1. Conclusões

As técnicas de desidentificação e anonimização são objeto de intensa investigação e o presente documento demonstrou sistematicamente que cada técnica tem as suas vantagens e desvantagens. Na maioria dos casos, não é possível formular recomendações mínimas para os parâmetros a utilizar, pois há que considerar cada conjunto de dados caso a caso.

Em muitas situações, um conjunto de dados anonimizado pode ainda apresentar um risco residual para os titulares dos dados. Com efeito, mesmo quando já não é possível recuperar com precisão os dados de uma pessoa, pode continuar a ser possível recolher informações sobre essa pessoa recorrendo a outras fontes de informação disponíveis (ao público ou não). Há que salientar que para além do impacto direto sobre os titulares dos dados produzido pelas consequências de um processo de anonimização débil (incómodo, morosidade e sentimento de perda de controlo por serem incluídos num agregado sem conhecimento ou consentimento prévio), podem ocorrer outros efeitos indiretos decorrentes de uma fraca anonimização sempre que um titular dos dados seja erradamente incluído num alvo por algum intruso, na sequência do tratamento de dados anonimizados – especialmente se o intruso tiver intenções maliciosas. Por conseguinte, o grupo de trabalho sublinha que as técnicas de anonimização de dados pessoais podem fornecer garantias de privacidade, mas apenas se a sua aplicação for devidamente elaborada – o que significa que os requisitos prévios (âmbito) e o ou os objetivos do processo de anonimização têm de ser claramente definidos a fim de se atingir o nível de anonimização pretendido.

5.2. Recomendações

- Algumas técnicas de anonimização mostram limitações inerentes. Estas limitações devem ser seriamente ponderadas antes da utilização de uma determinada técnica na criação de um processo de anonimização pelos responsáveis pelo tratamento de dados. Devem ter em conta os fins a atingir através da anonimização – tais como proteger a privacidade dos indivíduos aquando da publicação de um conjunto de dados, ou permitir que um elemento de informação seja recuperado de um conjunto de dados.
- Nenhuma das técnicas descritas no presente documento satisfaz plenamente os critérios de anonimização eficaz (isto é, a não identificação de uma pessoa, a impossibilidade de ligação entre os registos referentes a um indivíduo e a não inferência a respeito de um indivíduo). No entanto, uma vez que alguns destes riscos podem ser encontrados total ou parcialmente numa determinada técnica, é necessário um planeamento meticuloso na definição da aplicação de uma técnica individual à situação específica, bem como na aplicação de uma combinação dessas técnicas de modo a reforçar a robustez dos resultados.

A tabela seguinte apresenta uma visão geral dos pontos fortes e fracos das técnicas quando consideradas em termos dos três requisitos básicos:

	A identificação ainda é um risco?	**A possibilidade de ligação ainda é um risco?**	**A inferência ainda é um risco?**
Utilização de pseudónimos	Sim	Sim	Sim
Adição de ruído	Sim	Não pode	Não pode
Substituição	Sim	Sim	Não pode
Agregação ou k-anonimato	Não	Sim	Sim
L-diversidade	Não	Sim	Não pode
Privacidade diferencial	Não pode	Não pode	Não pode
Utilização da função *hash*/tokens	Sim	Sim	Não pode

Quadro 6. Pontos fortes e fracos das técnicas consideradas

- Há que decidir a melhor solução caso a caso. Uma solução (ou seja, um processo completo de anonimização) que satisfaça os três requisitos deverá ser sólida face à identificação realizada pelos meios mais prováveis e razoáveis suscetíveis de serem utilizados pelo responsável pelo tratamento de dados ou por qualquer terceiro.
- Sempre que uma proposta não satisfaça um dos critérios, deve ser efetuada uma avaliação minuciosa dos riscos de identificação. Esta avaliação deve ser fornecida ao órgão de fiscalização se a legislação nacional exigir que o órgão de fiscalização avalie ou autorize o processo de anonimização de dados pessoais.

A fim de reduzir os riscos de identificação, devem ser tidas em conta as seguintes boas práticas:

Boas práticas para a anonimização de dados pessoais

De um modo geral:

- Não depender da abordagem «divulgar e esquecer». Tendo em conta o risco residual de identificação, os responsáveis pelo tratamento devem:
    - 1. Identificar novos riscos e reavaliar os riscos residuais regularmente;
    - 2. Avaliar se os controlos para os riscos identificados são suficientes e ajustá-los em conformidade;
    - 3. Acompanhar e controlar os riscos.
- No âmbito de tais riscos residuais, ter em conta o potencial de identificação da parte não anonimizada de um conjunto de dados (se for caso disso), nomeadamente quando combinada com a parte anonimizada, além de eventuais correlações entre atributos (por exemplo, entre dados de localização geográfica e níveis de riqueza).

Elementos contextuais:

- Os objetivos a alcançar por meio do conjunto de dados anonimizado devem ser claramente estabelecidos, dado que representam um papel fundamental na determinação do risco de identificação.

- Paralelamente, devem ser considerados todos os elementos contextuais pertinentes – por exemplo, a natureza dos dados originais, mecanismos de controlo implantados (incluindo medidas de segurança a fim de restringir o acesso aos conjuntos de dados), a dimensão da amostra (elementos quantitativos), disponibilidade de recursos de informação pública (a ser invocados pelos recetores), divulgação prevista de dados a terceiros (limitada, ilimitada, por exemplo, na Internet, etc.).
- Há que considerar os intrusos possíveis, tendo em conta o grau de apelabilidade dos dados para ataques propositados (também aqui, a sensibilidade das informações e a natureza dos dados constituem fatores-chave).

Elementos técnicos:

- Os responsáveis pelo tratamento de dados devem divulgar a técnica de anonimização/combinação de técnicas a executar, em especial se previrem a divulgação do conjunto de dados anónimos.
- Os atributos óbvios (por exemplo, os raros)/quase-identificadores devem ser retirados do conjunto de dados.
- Se forem utilizadas técnicas de adição de ruído (na seleção aleatória), o nível de ruído adicionado aos registos deve ser determinado em função do valor de um atributo (ou seja, não deverá ser injetado ruído desproporcional), do impacto nos titulares dos dados dos atributos a proteger e/ou da dispersão do conjunto de dados.
- Ao utilizar a privacidade diferencial (seleção aleatória), há que tomar em consideração a necessidade de manter um registo de consultas, a fim de detetar consultas intrusivas para a privacidade, dado que o caráter intrusivo das consultas é cumulativo.
- Se forem implementadas técnicas de generalização, é fundamental que o responsável pelo tratamento de dados não as limite a um critério de generalização, mesmo para o mesmo atributo. Ou seja, devem ser selecionadas granularidades de localização diferentes ou intervalos de tempo diferentes. A seleção do critério a aplicar deve ser conduzida em função da distribuição dos valores de atributo na população em causa. Nem todas as distribuições permitem uma generalização, ou seja, não pode ser seguida uma abordagem única na generalização. Deve ser assegurada a variabilidade em classes de equivalência; por exemplo, devem ser selecionados limites específicos consoante os «elementos contextuais» acima referidos (dimensão da amostra, etc.) e, se esse limite não for alcançado, então a amostra específica deve ser eliminada (ou deve ser estabelecido um critério de generalização diferente).

ANEXO
Resumo das técnicas de anonimização

A.1. Introdução

O anonimato é interpretado de forma diferente em toda a UE – sendo que em alguns Estados-Membros corresponde a anonimato computacional (ou seja, deverá ser informaticamente difícil, mesmo para o responsável pelo tratamento de dados em colaboração com terceiros, identificar direta ou indiretamente um dos titulares de dados) e em outros corresponde ao anonimato perfeito (ou seja, deverá ser impossível, mesmo para o responsável pelo tratamento de dados em colaboração com terceiros, identificar direta ou indiretamente um dos titulares de dados). No entanto, a «anonimização» corresponde, em ambos os casos, ao processo através do qual os dados são tornados anónimos. A diferença reside naquele que é considerado um nível de risco de reidentificação aceitável.

Podem prever-se várias utilidades para os dados anonimizados, desde inquéritos sociais e análises estatísticas, ao desenvolvimento de novos serviços e produtos. Por vezes, até essas atividades com objetivo geral podem ter um impacto em titulares de dados específicos, anulando a natureza supostamente anónima dos dados tratados. Muitos exemplos podem ser dados, desde o lançamento de iniciativas de comercialização seletiva, à aplicação de medidas públicas tendo por base a elaboração de perfis do utilizador, ou os seus comportamentos ou padrões de mobilidade[21].

Infelizmente, para além de afirmações empíricas, não existe qualquer unidade de medida sustentável para avaliar previamente o tempo ou o esforço necessários para a reidentificação após o tratamento dos dados, ou, em alternativa, para selecionar o procedimento mais adequado a seguir no caso de se pretender reduzir a probabilidade de uma base de dados divulgada referir um conjunto identificado de titulares dos dados.

A «arte de anonimização», conforme são por vezes designadas estas práticas na literatura científica[22], é um novo ramo científico que ainda se encontra em fase inicial, existindo muitas práticas para comprometer as possibilidades de identificação de conjuntos de dados. No entanto, há que afirmar claramente que a maioria destas práticas não impede a ligação entre os dados tratados e os titulares dos dados. Em algumas circunstâncias, a identificação de conjuntos de dados considerados anónimos revelou-se muito eficaz, mas noutras situações verificaram-se falsos positivos.

De um modo geral, existem duas abordagens diferentes: uma baseia-se na generalização do atributo, outra na aleatorização. A análise dos pormenores e subtilezas destas práticas irá conduzir-nos a uma nova perspetiva sobre a possibilidade de identificação de dados e irá trazer novos pontos de vista sobre a própria noção de dados pessoais.

A.2. «Anonimização» por aleatorização

Uma opção de anonimização de dados pessoais consiste em alterar os valores reais a fim de impedir a ligação entre os dados anonimizados e os valores originais. Este objetivo pode ser alcançado através de um vasto número de métodos que vão desde a injeção de ruído à troca de

[21] Por exemplo, o processo da TomTom nos Países Baixos (ver o exemplo referido no ponto 2.2.3).
[22] Jun Gu, Yuexian Chen, Junning Fu, HuanchunPeng, Xiaojun Ye, Synthesizing: Art of Anonymization, Database and Expert Systems Applications Lecture Notes in Computer Science - Springer - Volume 6261, 2010, pp 385-399.

dados (permutação). É de salientar que a remoção de um atributo equivale a uma forma extrema de aleatorização deste atributo (em que o atributo é totalmente coberto pelo ruído).

Em algumas circunstâncias, o objetivo do tratamento global de dados não é tanto a divulgação de um conjunto de dados aleatório, mas sim a concessão do acesso aos dados através de consultas. Neste caso, o risco para o titular dos dados decorre da probabilidade de um intruso conseguir retirar informações de uma série de consultas diferentes sem o conhecimento do responsável pelo tratamento de dados. A fim de garantir o anonimato de indivíduos no conjunto de dados, não deve ser possível concluir que um titular de dados contribuiu para o conjunto de dados, quebrando assim a ligação com qualquer tipo de informações de base que um intruso possa ter.

A adição de ruído adequada à resposta da consulta pode reduzir ainda mais o risco de reidentificação. Esta abordagem, também conhecida na literatura como privacidade diferencial[23], desvia-se das descritas anteriormente, na medida em que cede aos divulgadores de dados um maior controlo sobre o acesso aos dados em comparação com a divulgação ao público. A adição de ruído tem dois objetivos principais: proteger a privacidade dos titulares de dados no conjunto de dados e manter a utilidade da informação divulgada. Em especial, a extensão do ruído deve ser proporcional ao nível de consulta (demasiadas consultas a pessoas com respostas demasiadamente precisas resultam numa maior probabilidade de identificação). Hoje em dia a aplicação de uma aleatorização eficaz deve ser considerada caso a caso, já que nenhuma técnica oferece uma metodologia infalível uma vez que existem exemplos de fugas de informação sobre os atributos de um titular de dados (quer este esteja ou não incluído no conjunto de dados), mesmo quando o conjunto de dados havia sido considerado aleatorizado pelo responsável pelo tratamento de dados.

Pode ser útil discutir exemplos específicos para esclarecer as possíveis deficiências da aleatorização enquanto meio para fornecer a anonimização de dados pessoais. Por exemplo, no contexto do acesso interativo, as consultas consideradas respeitadoras da privacidade são passíveis de constituir um risco para os titulares de dados. De facto, se o intruso souber que um subgrupo S de pessoas singulares se encontra num conjunto de dados que contém informações sobre a incidência do atributo A dentro de uma população P, com a simples consulta das duas questões «Quantas pessoas na população P possuem o atributo A?» e «Quantas pessoas na população P, à exceção das que pertencem ao subgrupo S, possuem o atributo A?», poderá ser possível determinar (por diferença), o número de pessoas em S que efetivamente possuem o atributo A – seja deterministicamente ou por inferência de probabilidade. Em qualquer caso, a privacidade dos indivíduos no subgrupo S pode ficar gravemente ameaçada, sobretudo em função da natureza do atributo A.

Pode igualmente considerar-se que se um titular de dados não fizer parte do conjunto de dados mas a sua relação com os dados constantes no conjunto de dados for conhecida, então a divulgação do conjunto de dados pode colocar em risco a sua privacidade. Por exemplo, se for sabido que «o valor do alvo no que se refere ao atributo A difere numa quantidade de X do valor médio da população», ao solicitar meramente ao responsável pela base de dados para executar a operação respeitadora de privacidade de extração do valor médio do atributo A, o intruso pode inferir com exatidão dados pessoais relativos a uma pessoa específica.

A injeção de algumas imprecisões relativas nos valores reais numa base de dados é uma operação que deve ser adequadamente elaborada. É necessário adicionar ruído suficiente para

[23] Cynthia Dwork, Differential Privacy, International Colloquium on Automata, Languages and Programming (ICALP) 2006, p. 1–12.

proteger a privacidade, mas que seja também suficientemente baixo para preservar a utilidade dos dados. Por exemplo, se o número de titulares de dados com um atributo peculiar for muito reduzido, ou se a sensibilidade do atributo for elevada, pode ser preferível referir um intervalo ou uma frase genérica como «um pequeno número de casos, possivelmente até zero», em vez de referir o número real. Deste modo, mesmo que o mecanismo de revelação do ruído seja conhecido de antemão, a privacidade do titular de dados será preservada, uma vez que se mantém um grau de incerteza. Numa perspetiva de utilidade, se a imprecisão for corretamente elaborada, os resultados permanecerão úteis para efeitos estatísticos ou de tomada de decisões.

A aleatorização da base de dados e o acesso à privacidade diferencial exigem maior reflexão. Em primeiro lugar, a quantidade certa de distorção pode variar significativamente consoante o contexto (tipo de consulta, dimensão da população na base de dados, natureza do atributo e a sua capacidade de identificação inerente) e não pode ser prevista uma solução *ad omnia*. Além disso, o contexto pode alterar-se ao longo do tempo e o mecanismo interativo deve ser alterado em conformidade. A calibração do ruído requer a monitorização dos riscos de privacidade cumulativos que qualquer mecanismo interativo constitui para os titulares de dados. O mecanismo de acesso aos dados deve, então, ser equipado com alertas sempre que um orçamento de «custo de privacidade» seja alcançado e os titulares dos dados sejam suscetíveis de serem expostos a riscos específicos caso seja executada uma nova consulta, a fim de auxiliar o responsável pelo tratamento de dados na determinação do nível adequado de distorção a injetar de cada uma das vezes nos dados pessoais originais.

Por outro lado, também se deve considerar o caso em que os valores dos atributos são eliminados (ou modificados). Uma solução frequentemente utilizada para fazer face a alguns valores atípicos de atributos é a eliminação do conjunto de dados relacionados com os indivíduos atípicos ou dos valores atípicos. Neste último caso, é então importante garantir que a própria falta de valor não se torna um elemento de identificação da pessoa em causa.

Consideremos agora a aleatorização por substituição de atributo. Um grande equívoco cometido em matéria de anonimização de dados pessoais é equipará-la à cifragem ou à codificação com chave. Este equívoco tem por base dois pressupostos, designadamente: a) que quando é aplicada uma cifragem a alguns atributos de um registo numa base de dados (por exemplo, nome, morada, data de nascimento), ou esses atributos são substituídos por uma sequência aparentemente aleatória resultante de uma operação de codificação com chave, como a função *hash* com chave, o registo se encontra «anonimizado», e b) que a anonimização é mais eficaz se o tamanho da chave for adequado e o algoritmo de cifragem for de última geração. Este equívoco é generalizado entre os responsáveis pelo tratamento de dados e merece clarificação, dado que se encontra também relacionado com a utilização de pseudónimos e os seus riscos alegadamente mais baixos.

Em primeiro lugar, os objetivos destas técnicas são radicalmente diferentes: a cifragem enquanto prática de segurança destina-se a fornecer a confidencialidade de um canal de comunicação entre as partes identificadas (seres humanos, dispositivos ou partes de *hardware*/programas informáticos), a fim de evitar a escuta não autorizada ou a divulgação não intencional. A codificação com chave corresponde a uma tradução semântica dos dados em função de uma chave secreta. Por outro lado, o objetivo da anonimização de dados pessoais é evitar a identificação das pessoas singulares ao impossibilitar a ligação oculta de atributos a um titular de dados.

A cifragem e a codificação por chave não permitem, por si só, tornar o titular de dados não identificável, uma vez que, pelo menos nas mãos do responsável pelo tratamento de dados, os

dados originais permanecem ainda disponíveis ou dedutíveis. A simples aplicação de uma tradução semântica de dados pessoais, como é o caso da codificação por chave, não elimina a possibilidade de reverter os dados à sua estrutura inicial – seja através da aplicação do algoritmo na direção oposta, seja de ataques de «força bruta», consoante a natureza dos regimes, ou ainda na sequência de uma violação de dados. A cifragem de última geração consegue garantir que dados estão protegidos a um nível mais elevado, ou seja, que são ininteligíveis para as entidades que desconheçam a chave de decifração, mas não resulta necessariamente numa anonimização de dados pessoais. Enquanto a chave ou os dados originais permanecerem disponíveis (mesmo no caso de um terceiro de confiança contratualmente vinculado a fornecer um serviço de depósito de chave segura), subsiste a possibilidade de identificação do titular dos dados.

Centrar-se apenas na solidez do mecanismo de cifragem como medição do grau de «anonimização» de um conjunto de dados é enganador, já que muitos outros fatores técnicos e organizacionais afetam a segurança global de um mecanismo de cifragem ou de uma função *hash*. A literatura já divulgou vários ataques bem-sucedidos que contornam totalmente o algoritmo, seja porque se aproveitam de debilidades na conservação das chaves (por exemplo, a existência de um modo predefinido menos seguro) ou de outros fatores humanos (por exemplo, palavras-chave fracas para a recuperação da chave). Por último, um regime escolhido de cifragem com um determinado tamanho de chave destina-se a garantir a confidencialidade durante um determinado período (a maioria das chaves atuais terá de ser redimensionada por volta de 2020), ao passo que um processo de anonimização de dados pessoais não deve ser limitado no tempo.

Importa agora aprofundar os limites de aleatorização do atributo (ou de substituição e exclusão) tendo em conta vários maus exemplos de anonimização por aleatorização que ocorreram nos últimos anos e os motivos subjacentes a tais falhas.

Um caso famoso que envolveu a divulgação de um conjunto de dados mal anonimizado é o do prémio Netflix[24]. Ao observar um registo genérico numa base de dados em que uma série de atributos relativos a um titular de dados foi aleatorizada, cada registo pode ainda ser dividido em dois subgrupos de registos do seguinte modo: {atributos aleatórios, atributos claros}, em que os atributos claros podem ser qualquer combinação de dados supostamente não pessoais. Uma observação específica que pode ser feita do conjunto de dados do prémio Netflix resulta do princípio de que cada registo pode ser representado por um ponto num espaço multidimensional, onde cada atributo claro é uma coordenada. Segundo esta técnica, qualquer conjunto de dados pode ser considerado uma constelação de pontos num determinado espaço multidimensional que pode apresentar um elevado grau de dispersão, o que significa que os pontos podem encontrar-se distantes entre si. Com efeito, podem encontrar-se de tal forma afastados que após a divisão do espaço em regiões amplas, cada região contém apenas um registo. Nem a injeção de ruído consegue aproximar suficientemente os registos entre eles de forma a partilharem a mesma região multidimensional. Por exemplo, na experiência Netflix, os registos eram suficientemente únicos com apenas 8 classificações de filmes atribuídas num período de 14 dias. Após a adição de ruído às classificações e às datas não houve sobreposição das regiões. Por outras palavras, a mesma seleção de apenas 8 filmes classificados constituiu uma impressão digital das classificações expressas, as quais não eram partilhadas entre dois titulares de dados na base de dados. Com base nesta observação geométrica, os investigadores fizeram corresponder o conjunto de dados da Netflix, supostamente anónimo, com outra base de dados pública de classificações de filmes

[24] Arvind Narayanan, Vitaly Shmatikov: Robust De-anonymization of Large Sparse Datasets. Simpósio IEEE sobre a segurança e privacidade 2008: 111-125

(IMDB), encontrando assim utilizadores que tinham classificado os mesmos filmes dentro dos mesmos intervalos de tempo. Uma vez que a maioria dos utilizadores mostrou uma correspondência exata, a informação auxiliar obtida na base de dados da IMDB pôde ser importada para o conjunto de dados da Netflix divulgado, enriquecendo assim, com identidades, todos os registos supostamente anonimizados.

É importante salientar que se trata de uma propriedade geral: a parte residual de qualquer base de dados «aleatória» continua a possuir um poder de identificação muito elevado, consoante a raridade da combinação dos atributos residuais. Trata-se de uma ressalva que os responsáveis pelo tratamento de dados devem ter sempre em conta ao selecionar a aleatorização como forma de obtenção da anonimização pretendida.

Muitas experiências de reidentificação deste tipo seguiram uma abordagem semelhante de projeção de duas bases de dados num mesmo subespaço. Trata-se de uma metodologia de identificação muito forte, que teve recentemente muitas aplicações em vários domínios. Por exemplo, uma experiência de identificação efetuada em relação a uma rede social[25] explorou o gráfico social de utilizadores sob pseudónimos por meio de etiquetas. Neste caso, o atributo utilizado para identificação foi a lista de contactos de cada utilizador, dado que foi demonstrado que a probabilidade de existência de uma lista de contactos idêntica entre duas pessoas singulares é muito baixa. Com base neste pressuposto intuitivo, constatou-se que um subgráfico das ligações internas de um número muito limitado de nós constitui uma impressão digital topológica passível de ser encontrada, oculta na rede, e que uma grande parte de toda a rede social pode ser identificada assim que esta sub-rede for identificada. Com o mero intuito de fornecer alguns dados acerca do desempenho de um ataque semelhante, verificou-se que através da utilização de menos de 10 nós (suscetíveis de dar origem a milhões de configurações diferentes de sub-redes, cada uma podendo constituir uma impressão digital topológica), uma rede social de mais de 4 milhões de nós sob pseudónimo e de 70 milhões de ligações pode ser sujeita a ataques de reidentificação e a privacidade de um elevado número de ligações pode ser comprometida. Deve salientar-se que esta abordagem de reidentificação não é adaptada em função do contexto específico das redes sociais, mas é suficientemente genérica para poder ser adaptada a outras bases de dados em que sejam registadas as relações entre os utilizadores (por exemplo, contactos telefónicos, correspondência eletrónica, sítios de encontros, etc.).

Uma outra forma de identificar um registo supostamente anónimo tem por base a análise do estilo de escrita (estilometria)[26]. Já foi desenvolvida uma série de algoritmos para extrair parâmetros do texto analisado, incluindo a frequência da utilização de uma palavra em especial, a ocorrência de padrões gramaticais específicos e o tipo de pontuação. Todas estas propriedades podem ser utilizadas para atribuir um texto supostamente anónimo ao estilo de escrita de um autor identificado. Os investigadores recolheram o estilo de escrita de mais de 100 000 blogues e têm atualmente a capacidade de identificar automaticamente o autor de uma publicação com um grau de precisão já próximo dos 80 %. Prevê-se que a precisão desta técnica venha a aumentar, explorando também outros sinais, tais como a localização ou outros metadados contidos no texto.

O poder de identificação através da utilização da semântica de um registo (ou seja, a parte residual não aleatória de um registo) é uma questão que merece maior reflexão por parte da

[25] L. Backstrom, C. Dwork, e J. M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography, Proceedings of the 16th International Conference on World Wide Web WWW'07, page 181-190. (2007)

[26] http://33bits.org/2012/02/20/is-Writing-Style-sufficient-to-deanonymize-material-Posted-online/

comunidade de investigação e da indústria. A inversão recente da identidade dos dadores de ADN (2013)[27] revela que muito pouco se progrediu desde o famoso incidente do AOL (2006) – quando foi divulgada ao público uma base de dados que continha vinte milhões de palavras-chave de pesquisa de mais de 650 000 utilizadores durante um período de três meses. Tal originou a identificação e localização de vários utilizadores do AOL.

Outra família de dados que raramente é anonimizada pela simples remoção da identidade dos titulares de dados ou pela cifragem parcial de alguns atributos são os dados de localização. Os padrões de mobilidade dos seres humanos são de tal forma suficientemente únicos que a parte semântica dos dados de localização (os locais onde o titular dos dados se encontrava num determinado momento), mesmo sem outros atributos, consegue revelar muitos traços do titular dos dados[28]. Tal facto foi provado por diversas vezes em estudos académicos representativos[29].

A este respeito, é necessário alertar sobre a utilização de pseudónimos como uma forma de assegurar proteção adequada aos titulares de dados contra fugas de identidade ou de atributo. Se a utilização de pseudónimos tem por base a substituição de uma identidade por um outro código único, a presunção de que este facto constitui uma desidentificação sólida é ingénua e não tem em conta a complexidade das metodologias de identificação e os múltiplos contextos em que estas são passíveis de ser aplicadas.

A.3. «Anonimização» por generalização

Um exemplo simples pode ajudar a clarificar a abordagem baseada na generalização do atributo.

Consideremos o caso em que um responsável pelo tratamento de dados decide divulgar um simples quadro que contém três elementos de informação ou atributos: um número de identificação, único para cada registo, uma identificação de localização, que relaciona o titular dos dados ao local em que habita e uma identificação de propriedade, que indica a propriedade desse titular dos dados. Assumamos ainda que esta propriedade é um dos dois valores distintos, indicados genericamente por {P1, P2}:

[27] Os dados genéticos são um exemplo especialmente importante de dados confidenciais que podem estar em risco de nova identificação se o único mecanismo que os «anonimiza» for a remoção da identidade dos doadores. Ver o exemplo citado acima no ponto 2.2.2. Ver John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science 339 Vol., n.° 6117 (18 de janeiro de 2013), p. 262.

[28] Esta questão tem sido abordada em algumas legislações nacionais. Por exemplo, em França, as estatísticas de localização publicadas são anonimizadas através de generalização e permutação. Por conseguinte, o INSEE publica dados estatísticos que são generalizados através da agregação de todos os dados numa área de 40 000 m^2. A granularidade do conjunto de dados é suficiente para preservar a utilidade dos dados e as permutações evitam ataques de desanonimização em áreas dispersas. De um modo mais geral, agregar esta família de dados e permutá-la oferece garantias sólidas contra ataques de inferência e de desanonimização de dados pessoais (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. e Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. Nature. 3, 1376 (2013)

Número de identificação	Identificação de Localização	Propriedade
#1	Roma	P1
#2	Madrid	P1
#3	Londres	P2
#4	Paris	P1
#5	Barcelona	P1
#6	Milão	P2
#7	Nova Iorque	P2
#8	Berlim	P1

Quadro A1. Amostra de titulares de dados recolhida por localização e propriedades P1 e P2

Se alguém, designado por intruso, souber de antemão que um titular de dados específico (o alvo) vive em Milão e está incluído no quadro, então, ao analisar o quadro pode concluir que, sendo o #6 o único titular de dados com essa identificação de localização, também possui a propriedade P2.

Este simples exemplo revela os elementos principais de qualquer procedimento de identificação aplicados a um conjunto de dados que tenha sido objeto de um suposto processo de anonimização de dados pessoais. Designadamente, existe um intruso que (acidental ou intencionalmente) tinha conhecimentos prévios sobre todos ou alguns titulares de dados pertencentes a um conjunto de dados. O intruso visa estabelecer a ligação entre estes conhecimentos de base que detém de antemão e os dados constantes no conjunto de dados divulgado para obter uma ideia mais clara das características desses titulares de dados.

A fim de estabelecer a ligação de dados com qualquer tipo de conhecimentos de base menos eficazes ou imediatos, o responsável pelo tratamento de dados poderia centrar-se na identificação de localização, substituindo a cidade em que os titulares dos dados vivem por uma zona mais ampla, como o país. Assim, o quadro ficaria com a seguinte aparência:

Número de identificação	Identificação de Localização	Propriedade
#1	Itália	P1
#2	Espanha	P1
#3	Reino Unido	P2
#4	França	P1
#5	Espanha	P1
#6	Itália	P2
#7	EUA	P2
#8	Alemanha	P1

Quadro A2. Generalização do quadro A1 por nacionalidade

Com esta nova agregação de dados, os conhecimentos de base do intruso sobre um titular de dados identificado (por exemplo, «o alvo vive em Roma e consta no quadro») não permitem qualquer conclusão mais clara sobre a sua propriedade, pois os dois italianos constantes no quadro possuem propriedades distintas, P1 e P2, respetivamente. O intruso fica com um grau

de incerteza de 50 % sobre a propriedade da entidade alvo. Este exemplo simples mostra o efeito da generalização sobre a prática de anonimização de dados pessoais. Com efeito, embora este truque de generalização possa ser eficaz para reduzir para metade a probabilidade de identificação de um alvo italiano, não é eficaz para um alvo de outras localizações (por exemplo, EUA).

Além disso, um intruso consegue ainda obter informações acerca de um alvo espanhol. Se o conhecimento de base for do género «o alvo vive em Madrid e consta no quadro» ou «o alvo vive em Barcelona e consta no quadro», o intruso consegue inferir com uma certeza de 100 % que o alvo possui a propriedade P1. Por conseguinte, a generalização não produz o mesmo nível de privacidade ou de resistência contra ataques de inferência para toda a população pertencente ao conjunto de dados.

Seguindo esta linha de raciocínio, pode ser tentador concluir que uma generalização mais forte poderia ser útil para impedir qualquer possibilidade de ligação – por exemplo, a generalização por continentes. Assim, o quadro teria a seguinte disposição:

Número de identificação	**Identificação de Localização**	**Propriedade**
#1	Europa	P1
#2	Europa	P1
#3	Europa	P2
#4	Europa	P1
#5	Europa	P1
#6	Europa	P2
#7	América do Norte	P2
#8	Europa	P1

Quadro A3. Generalização do quadro A1 por continente

Com este tipo de agregação, todos os titulares de dados do quadro, à exceção dos que residem nos EUA, estariam protegidos contra ataques de ligação e de identificação e qualquer informação de base do tipo «o alvo vive em Madrid e consta no quadro» ou «o alvo vive em Milão e consta no quadro» conduziria a um determinado nível de probabilidade no que respeita à propriedade aplicável a um determinado titular de dados (P1 com probabilidade de 71,4 % e P2 com probabilidade de 28, 6 %), e não à possibilidade de ligação direta. Além disso, esta generalização é efetuada em detrimento de uma perda evidente e radical de informações: o quadro não permite descobrir eventuais correlações entre as propriedades e a localização, designadamente, saber se uma localização específica é passível de desencadear qualquer uma das duas propriedades com maior probabilidade, dado que apenas produz distribuições designadas por «marginais», nomeadamente a probabilidade absoluta de ocorrência das propriedades P1 e P2 em toda a população (respetivamente, 62,5 % e 37,5 % no nosso exemplo) e dentro de cada continente (respetivamente, como referido, 71,4 % e 28,6 % na Europa e 100 % e 0 % na América do Norte).

O exemplo revela, também, que a prática de generalização afeta a utilidade prática dos dados. Atualmente, estão disponíveis alguns instrumentos de engenharia para identificar antecipadamente (isto é, antes da divulgação de um conjunto de dados) o nível mais adequado de generalização de atributo, por forma a reduzir os riscos de identificação dos titulares de

dados constantes num quadro sem prejudicar excessivamente a utilidade dos dados divulgados.

K-anonimato

Uma das maneiras de impedir ataques de possibilidade de ligação baseados na generalização de atributos é conhecida por k-anonimato. Esta prática decorre de uma experiência de reidentificação realizada em finais dos anos 90, quando uma empresa americana privada, que operava no setor da saúde, divulgou publicamente um conjunto de dados supostamente anonimizado. Esta anonimização consistiu em retirar os nomes dos titulares dos dados, mas o conjunto de dados continuou a conter dados de saúde e outros atributos, tais como o código postal (identificação da localização onde viviam), o sexo e a data de nascimento completa. O mesmo trio de informações {código postal, sexo, data de nascimento completa} foi incluído também em outros registos disponíveis ao público (por exemplo, a lista de eleitores) e pôde, assim, ser utilizado por um investigador académico para estabelecer a ligação entre a identidade de titulares de dados específicos e os atributos no conjunto de dados divulgado. O conhecimento de base detido pelo intruso (o investigador) poderia ser o seguinte: «eu sei que o titular de dados constante na lista de eleitores com um trio de informações específico {código postal, o sexo, data de nascimento completa} é único. Existe um registo no conjunto de dados divulgado com esse trio de informações». Observou-se empiricamente[30] que a grande maioria (mais de 80 %) dos titulares dos dados no registo público utilizados nesta experiência de investigação foi univocamente associada a um trio específico, o que permitiu a identificação. Por conseguinte, os dados não foram devidamente anonimizados neste caso.

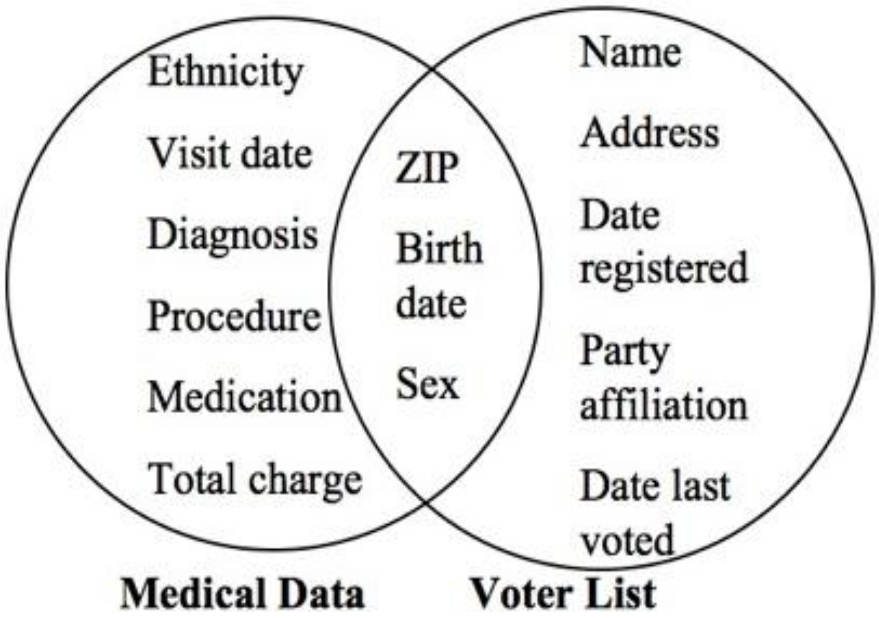

Figura A1. Reidentificação por ligação de dados

A fim de reduzir a eficácia de ataques de ligação semelhantes, tem-se alegado que os responsáveis pelo tratamento devem, antes de tudo, inspecionar o conjunto de dados e agrupar os atributos suscetíveis de serem razoavelmente utilizados por um intruso para a ligação entre o quadro divulgado e outra fonte auxiliar. Cada grupo deve incluir, pelo menos, *k* combinações idênticas de atributos generalizados (ou seja, deve representar uma classe de equivalência de atributos). Os conjuntos de dados só devem, então, ser divulgados após terem

[30] L. Sweeney. Weaving Technology and Policy Together to Maintain Confidentiality. Journal of Law, Medicine &Ethics, 25, n.os 2 e 3 (1997): 98-110

sido divididos nesses grupos homogéneos. Os atributos selecionados para generalização são designados na literatura por quase-identificadores, uma vez que o conhecimento dos mesmos sem qualquer tratamento originaria a identificação imediata dos titulares dos dados.

As várias experiências realizadas demonstraram a fragilidade de quadros k-anonimizados mal realizados. Tal é suscetível de ocorrer, por exemplo, em virtude dos outros atributos de uma classe de equivalência serem idênticos (como acontece para a classe de equivalência de titulares de dados espanhóis no exemplo do quadro A2) ou da sua distribuição ser muito desequilibrada, com uma elevada prevalência de um atributo específico, ou então porque o número de registos numa categoria de equivalência é muito baixo, o que permite, em ambos os casos, a inferência de probabilidade, ou ainda porque não existe uma diferença «semântica» significativa entre os atributos sem tratamento das classes de equivalência (por exemplo, a medida quantitativa desses atributos pode ser efetivamente diferente, mas numericamente muito próxima, ou podem pertencer a um intervalo de atributos semanticamente semelhantes, por exemplo, o mesmo intervalo de risco de crédito, ou a mesma família de patologias), pelo que o conjunto de dados é ainda suscetível de ser alvo de fugas de uma grande quantidade de informações relativas a titulares dos dados em ataques de ligação[31]. É importante aqui notar que sempre que os dados sejam escassos (por exemplo, se existirem poucas ocorrências de uma propriedade específica numa área geográfica) e uma primeira agregação não conseguir agrupar dados com um número suficiente de ocorrências de propriedades diferentes (por exemplo, se ainda for possível localizar um número reduzido de ocorrências de algumas propriedades numa área geográfica), será necessária maior agregação do atributo a fim de alcançar a anonimização pretendida.

L-diversidade

Com base nestas observações, ao longo dos anos têm vindo a ser propostas variantes de k-anonimato e têm vindo a ser desenvolvidos alguns critérios de engenharia para melhorar a prática de anonimização por generalização, com vista a reduzir os riscos de ataques de ligação. Tais medidas assentam em propriedades probabilísticas dos conjuntos de dados. Mais especificamente, é acrescentado um outro constrangimento, designadamente que cada atributo numa categoria de equivalência ocorre pelo menos *l* vezes, para que o intruso fique sempre com um grau significativo de incerteza sobre os atributos, mesmo que em presença de conhecimentos prévios sobre um titular de dados específico. É o mesmo que afirmar que um conjunto de dados (ou uma parte do mesmo) deve possuir um número mínimo de ocorrências de uma propriedade selecionada: este truque poderá atenuar o risco de reidentificação. Este é o objetivo da prática de anonimização por l-diversidade. Nos quadros A4 (dados originais) e A5 (resultado do tratamento) ilustra-se um exemplo desta prática. Como é evidente, ao tratar devidamente a identificação da localização e a idade dos indivíduos no quadro A4, a generalização dos atributos resulta num aumento substancial da incerteza em relação aos atributos reais de qualquer titular de dados no inquérito. Por exemplo, mesmo que o intruso saiba que um titular de dados pertence à primeira classe de equivalência, não consegue verificar se a uma pessoa possui as propriedades X, Y ou Z, uma vez que existe, pelo menos, um registo dessa classe (e em qualquer outra classe de equivalência) com tais propriedades.

[31] Deve salientar-se que é igualmente possível estabelecer correlações depois de os registos de dados terem sido agrupados por atributos. Se o responsável pelo tratamento de dados souber os tipos de correlações que pretende verificar, ele pode selecionar os atributos mais relevantes. Por exemplo, os resultados do inquérito realizado pela PEW não estão sujeitos a ataques de inferência precisos e continuam a ser muito úteis para encontrar correlações entre os dados demográficos e os interesses (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx)

Número de série	Identificação de Localização	Idade	Propriedade
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Quadro A4. Um quadro com indivíduos agrupados por localização, idade e três propriedades X, Y e Z

Número de série	Identificação de Localização	Idade	Propriedade
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Quadro A5. Um exemplo da versão 1-diversa do quadro A4

T-proximidade:

O caso específico de atributos numa partição que se encontram distribuídos de forma desigual ou pertencem a uma gama reduzida de valores ou significados semânticos é tratado pelo método conhecido como t-proximidade. Trata-se de uma melhoria de anonimização por generalização e consiste na prática da organização de dados, a fim de obter classes de equivalência que reflitam o melhor possível a distribuição inicial dos atributos no conjunto de dados original. Para este efeito, é utilizado um procedimento composto por duas fases, essencialmente do modo que se segue. O quadro A6 é a base de dados original, que inclui registos dos titulares de dados sem tratamento, agrupados por localização, idade, salário e duas famílias de propriedades semanticamente semelhantes, respetivamente (X1, X2, X3) e (Y1, Y2, E3) (por exemplo, classes semelhantes de risco de crédito, doenças semelhantes). Em primeiro lugar, o quadro é *l-diversificado* com *l*=1 (quadro A7), mediante o agrupamento dos registos em classes de equivalência semanticamente semelhantes e a fraca anonimização específica. Em seguida, é objeto de tratamento, a fim de obter a t-proximidade (quadro A8) e

maior variabilidade em cada secção. Com efeito, ao aplicar-se a segunda etapa, cada classe de equivalência passa a incluir registos de ambas as famílias de propriedades. Importa referir que a identificação de localização e a idade têm granularidades diferentes nas várias etapas do processo: isto significa que cada atributo pode requerer critérios de generalização diferentes para obter anonimização específica de dados pessoais, o que exige, por sua vez, um tratamento específico e um esforço computacional adequado da parte do responsável pelo tratamento de dados.

Número de série	Identificação de Localização	Idade	Salário	Propriedade
1	1127	29	30 mil	X1
2	1112	22	32 mil	X2
3	1128	27	35 mil	X3
4	1215	43	50 mil	X2
5	1219	52	120 mil	Y1
6	1216	47	60 mil	Y2
7	1115	30	55 mil	Y2
8	1123	36	100 mil	Y3
9	1117	32	110 mil	X3

Quadro A6. Um quadro com indivíduos agrupados por localização, idade, salários e duas famílias de propriedades

Número de série	Identificação de Localização	Idade	Salário	Propriedade
1	11**	2*	30 mil	X1
2	11**	2*	32 mil	X2
3	11**	2*	35 mil	X3
4	121*	>40	50 mil	X2
5	121*	>40	120 mil	Y1
6	121*	>40	60 mil	Y2
7	11**	3*	55 mil	Y2
8	11**	3*	100 mil	Y3
9	11**	3*	110 mil	X3

Quadro A7. Uma versão *l-diversificada* do quadro A6

Número de série	Identificação de Localização	Idade	Salário	Propriedade
1	112*	<40	30 mil	X1
3	112*	<40	35 mil	X3
8	112*	<40	100 mil	Y3
4	121*	>40	50 mil	X2
5	121*	>40	120 mil	Y1
6	121*	>40	60 mil	Y2
2	111*	<40	32 mil	X2
7	111*	<40	55 mil	Y2
9	111*	<40	110 mil	X3

Quadro A8. Uma versão *t*-próxima do quadro A6

Deve ser claramente referido que o objetivo de generalização dos atributos dos titulares com recurso a formas tão elaboradas por vezes só é possível para um pequeno número de registos e não para a sua totalidade. As boas práticas devem velar para que cada classe de equivalência contenha várias pessoas singulares e que já não seja possível qualquer ataque de inferência. Em qualquer dos casos, esta abordagem exige uma apreciação aprofundada dos dados disponíveis por parte dos responsáveis pelo tratamento de dados, bem como uma avaliação combinatória de várias alternativas (por exemplo, amplitudes de intervalo diferentes, granularidades de localização ou idade diferentes, etc.). Por outras palavras, a anonimização por generalização não pode ser o resultado de uma primeira tentativa grosseira dos responsáveis pelo tratamento de dados no sentido de substituir valores analíticos dos atributos de um registo por gamas, uma vez que são necessários métodos quantitativos mais específicos – tal como a avaliação da entropia de atributos dentro de cada secção ou a medição da distância entre as distribuições do atributo inicial e a distribuição em cada classe de equivalência.