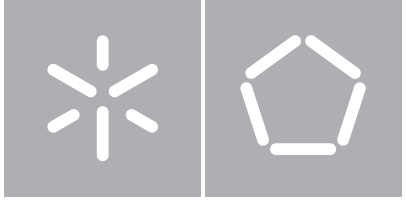
Universidade do Minho
Escola de Engenharia

Ricardo Manuel Amaro Pereira

Proteção da Privacidade em Sistemas de Dados

Janeiro de 2015



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Ricardo Manuel Amaro Pereira

Proteção da Privacidade em Sistemas de Dados

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de

Manuel Bernardo Barbosa
Paulo Jorge Azevedo

Janeiro de 2015

AGRADECIMENTOS

Gostava de agradecer a todos os que me acompanharam ao longo do meu percurso académico.

Agradeço a toda a minha família por todo o apoio e educação dada ao longo dos anos. Pela constante preocupação e interesse na minha vida académica e profissional.

Agradeço a todos os amigos exteriores à Universidade que sempre me acompanharam e contribuíram para aquilo que sou hoje.

Agradeço também em particular a todos os amigos que criei na Universidade do Minho. Por todos os momentos de distração, alegria, e camaradagem. Por estarem presentes em todos os bons e menos bons momentos do percurso universitário.

Queria também agradecer aos meus orientadores, Professor Manuel Bernardo Barbosa e Professor Paulo Jorge Azevedo, por todo o apoio e disponibilidade oferecida ao longo da realização desta dissertação, desde a primeira à ultima semana. Sem o contributo de ambos, esta dissertação não seria possível.

ABSTRACT

Information is nowadays an extremely valuable asset and takes on certain occasions a central role in many organizations. This information however may find itself legally protected (e.g. medical records) or contain data of real people as individual preferences or transactions whose privacy people do not want to see broken. On the other hand, there are sometimes obvious applications for this information as for example to generate statistics for medical research or for simply as a way to provide a better service.

There is thus a conflict of interest between those who hold information and intend to value it somehow and those who do not want to see their privacy compromised.

Two promising technologies arise in this context: techniques based on the notion of differential privacy and data anonymization algorithms. The goal of this dissertation is to explore the interaction between these two technologies and the possibility to utilize them together.

RESUMO

A informação é hoje em dia um bem muito valioso e assume em certos casos um papel central em várias organizações. No entanto esta informação pode encontrar-se protegida legalmente (i.e. registos médicos) ou conter dados de pessoas reais como preferências individuais ou transações cuja privacidade as pessoas não querem ver quebrada. Por outro lado existem por vezes aplicações óbvias para esta informação como por exemplo estatísticas para fins de investigação médica ou simplesmente para oferecer um melhor serviço.

Existe desta forma um conflito de interesses entre aqueles que detêm informação e a pretendem de alguma forma valorizar e aqueles que não querem ver a sua privacidade comprometida.

Surgem neste contexto duas tecnologias promissoras que são as técnicas baseadas na noção de privacidade diferencial e os algoritmos de anonimização de dados. O objetivo desta dissertação é explorar a interação entre estas duas tecnologias e a possibilidade de as utilizar em simultâneo.

CONTEÚDO

LISTA DE FIGURAS

LISTA DE TABELAS

1

INTRODUÇÃO

1.1 CONTEXTO

A importância das TIC na sociedade moderna tem sofrido uma evolução permanente desde que, nos anos 90, se tornou possível o acesso generalizado à Internet. Aquilo que inicialmente parecia um canal alternativo para o acesso a informação e entretenimento, tornou-se hoje numa verdadeira virtualização da sociedade onde um volume outrora impensável de dados é continuamente recolhido, transportado e armazenado.

No entanto, as capacidades tecnológicas de recolha e armazenamento de dados avançaram muito mais depressa do que a nossa capacidade de os processar. São muitas, hoje em dia, as entidades com acesso a fontes contínuas de informação, sejam elas leituras de consumos de água, gás e eletricidade, cliques de rato ou ativações de botões em controlos remotos. O sentimento de que a informação é valiosa determina que o seu armazenamento seja feito, sem que exista uma ideia clara de que utilização lhe pode ser dada.

Existem porém grandes benefícios para a informação recolhida como a publicação de estatísticas oficiais, encontrar associações de dados através de *data mining* em diversas áreas com especial ênfase em registos médicos, ou mesmo de forma a melhorar a experiência dos utilizadores aperfeiçoando por exemplo os resultados das pesquisas online. Estes benefícios esbarram por vezes em restrições impostas pela obrigatoriedade de proteger a privacidade de terceiros. Surge assim o problema de como publicar informação referente a pessoas, de forma a obter diversos benefícios protegendo ao mesmo tempo a sua privacidade.

1.2 MOTIVAÇÃO

O desafio tecnológico que se coloca nestes casos é o de fornecer uma solução que resolva o conflito de interesses (legítimos) entre aqueles que detêm informação e a pretendem valorizar, e

aqueles a quem a informação diz respeito e cuja privacidade pode ser de alguma forma comprometida.

Como solução deste conflito aparecem duas tecnologias promissoras, as técnicas baseadas na noção de privacidade diferencial e os algoritmos de anonimização de dados.

O modelo de privacidade diferencial implica que os resultados das computações obtidas pelos analistas sejam indistinguíveis quando efetuadas a dois conjuntos de dados que diferem em um elemento. Quase como se cada elemento decidisse não participar no conjunto de dados.

Os algoritmos de anonimização de dados consistem em manipular o conteúdo de um conjunto de dados de forma a que não seja possível a identificação de registos individuais protegendo assim a sua privacidade.

Em ambos os casos, o requisito funcional é o de preservar, tanto quanto possível, a utilidade dos resultados devolvidos pelos modelos.

O objetivo desta dissertação é o estudo destas tecnologias de forma teórica e experimental. Não só quando utilizadas de forma individual mas também estudar sua a interação, explorando a possibilidade da utilização das técnicas de privacidade diferencial e dos algoritmos de anonimização em conjunto. Em concreto é explorado o caso em que partindo de um conjunto de dados já anonimizado é de seguida aplicado o modelo de privacidade diferencial.

1.3 ESTRUTURA DO DOCUMENTO

Este documento encontra-se dividido em seis capítulos, em que neste primeiro capítulo é realizada uma introdução ao tema e apresentado o contexto em que este se insere e a motivação que levou à abordagem deste problema.

No segundo capítulo é detalhado o estado da arte relevante para esta dissertação. São apresentadas as duas principais soluções para o problema de privacidade em sistemas de dados, os algoritmos de anonimização de dados, derivados do mecanismo de *k-anonimity*[1] e o modelo de *differential privacy*[2] e seus variantes como a noção de privacidade denominada por *crowd blending privacy*[3], encontrando-se assim o capítulo sobre o estado da arte dividido em três secções.

No terceiro capítulo é explorada uma das principais temáticas desta dissertação, a composição dos modelos de *differential privacy* com o algoritmos de anonimização de *k-anonimity*. Neste capítulo é apresentado um caso de estudo em concreto que utiliza ambos os modelos em simultâneo. São também apresentadas dois tipos de *queries* de contagem para funcionarem neste

contexto de estudo e um ataque surpreendente que pode ocorrer com a utilização composta dos dois modelos neste contexto.

No quarto capítulo são detalhados os testes experimentais realizados no âmbito desta dissertação. É detalhado o código implementado para a realização destes testes, assim como os conjuntos de dados e bibliotecas auxiliares utilizadas para os testes.

No quinto capítulo é procurada uma solução para o ataque demonstrado no terceiro capítulo e são apresentadas três aproximações. Uma aproximação puramente teórica, uma aproximação híbrida e uma aproximação puramente experimental.

Por fim no sexto capítulo é realizada uma síntese do trabalho efetuado e resultados obtidos. É também feita uma reflexão sobre o trabalho futuro que poderia ser desenvolvido de forma a completar o trabalho realizado nesta dissertação.

2

ESTADO DA ARTE

Neste capítulo é apresentado o estado da arte relativo ao contexto desta dissertação. Encontra-se dividido em três secções. Na primeira, chamada critérios e mecanismos de anonimização, são apresentados os principais algoritmos de anonimização de dados nomeadamente o mecanismo de *k-anonimity* e os seus derivados. Na segunda secção é mostrado o modelo de *differential privacy*, e por fim na terceira secção são apresentadas variações deste modelo.

2.1 CRITÉRIOS E MECANISMOS DE ANONIMIZAÇÃO

O desenvolvimento do mecanismo de *k*-anonimity por Sweeney foi motivado pelo incidente de quebra de privacidade em Massachusetts [1]. Era prática normal entre organizações publicarem dados relativos a pessoas apenas removendo os identificadores mais óbvios, tais como o nome, morada e número de telefone com a ideia de que desta forma a sua privacidade era assegurada. Assim a Group Insurance Commission (GIC) de Massachusetts publicou um conjunto de dados supostamente anonimizado sobre registos médicos de pacientes que eram clientes desta companhia de seguros. Apesar dos identificadores óbvios terem sido removidos, estes dados apresentavam para cada indivíduo o seu código postal, data de nascimento e sexo. Sweeney mostrou em [4] que estes três atributos são suficientes para provavelmente identificar 87% (216 milhões em 248 milhões) da população dos Estados Unidos da América. Tendo estes dados, Sweeney cruzou-os com a lista pública de registo de votantes de Cambridge Massachusetts, que entre outros atributos possuía o nome, código postal, data de nascimento e sexo. Assim conseguiu ligar vários nomes presentes no registo de votantes aos correspondentes registos médicos publicados pela companhia de seguros. A título de exemplo, os registos médicos do governador de Massachusetts da altura, William Weld, estavam presentes nos dados do GIC e foram identificados. Tendo os dados exatos do governador, obtidos através da lista de votos de Cambridge, haviam apenas seis indivíduos nos dados do GIC com a mesma data de nascimento, três deles eram do

sexo masculino e desses três apenas um tinha o mesmo código postal. De notar que mesmo sem acesso à lista pública de registo de votos de Cambridge Massachusetts, este tipo de ataques continuava a ser possível desde que se conseguisse obter os atributos do código postal, sexo e data de nascimento a partir de outra fonte, o que hoje em dia não aparenta ser uma tarefa muito difícil principalmente com a emergência das redes sociais.

Podemos observar um exemplo deste tipo de ataque nas seguintes tabelas. Mesmo que a tabela com os registos médicos não mostre o nome dos pacientes, podemos facilmente ligar o nome à doença usando os restantes atributos. Conseguimos neste exemplo descobrir que o João teve pneumonia. A partir dos dados públicos presentes no registo de votantes temos acesso à informação detalhada sobre o João e verificamos assim que só existe um registo na segunda tabela coerentes com estes dados.

Registo de Votantes

Nome	Idade	Sexo	Código-Postal
João	21	M	4700
Maria	31	F	1017
Joana	27	F	4475
António	35	M	4815
Tiago	40	M	8200

Tabela 1.: Registo de Votantes

Registos Médicos

Idade	Sexo	Código-Postal	Doença
21	M	4700	Pneumonia
21	F	4740	Cancro
22	M	4810	Gripe
22	M	4730	HIV
23	F	4700	Gripe
24	M	4700	Tuberculose

Tabela 2.: Registos Médicos

Devido a este problema Sweeney necessitava de um mecanismo que limitasse a possibilidade de se cruzarem os dados publicados com dados externos. A noção de privacidade proposta por Sweeney em [1] foi a de k-anonimity em que para cada registo presente numa tabela publicada têm que existir pelo menos outros k-1 registos idênticos a ele em relação a um conjunto de Quasi-Identifiers definidos previamente. k-anonimity foi o primeiro mecanismo de privacidade baseado

na noção de *"blending in a crowd"* uma vez que os registos da tabela publicada têm que estar misturados no meio de outros registos semelhantes.

Neste contexto, os dados são normalmente referentes a pessoas e estão organizados em tabelas cujas linhas representam os tuplos e as colunas os atributos. Um tuplo contém uma relação entre uma pessoa e o conjunto de atributos que lhe correspondem. Os atributos podem ser divididos em três tipos: identificadores, Quasi-Identifiers e atributos sensíveis. Os atributos identificadores são por exemplo o nome ou número de telefone da pessoa, são atributos que identificam claramente uma pessoa e são geralmente omitidos. Os Quasi-Identifiers são um conjunto de atributos que podem ser conjugados com informação externa de forma a revelar a identidade de registos individuais. Os atributos sensíveis são aqueles que um adversário não consegue descobrir, ao contrário dos Quasi-Identifiers. É de esperar que a pessoa que possui os dados e os quer publicar consiga identificar nos seus dados os atributos que poderão também aparecer em informação externa e desta forma identificar de forma correta os Quasi-Identifiers. Caso contrário os dados publicados poderão ser menos anónimos do que o esperado e comprometer a privacidade dos indivíduos neles presentes.

Definição 1 (Quasi-Identifiers [1])**.** *Dada uma tabela* $T(A_1, A_2, ..., A_n)$*, o conjunto dos atributos Quasi-Identifiers* $Q_T = \{A_1, A_2, ..., A_m\} \subseteq \{A_1, A_2, ..., A_n\}$ *é o conjunto mínimo de atributos que podem ser conjugados com informação externa de forma a revelar a identidade pessoal de registos individuais.*

No exemplo apresentado anteriormente na tabela 1 em (citar tabela 1 de cima) para a tabela de registos médicos o conjunto dos Quasi-Identifiers $Q_T = \{\text{Idade, Sexo, Código-Postal}\}$ sendo neste caso o atributo sensível a doença.

Definição 2 (k-anonimity [1])**.** *Uma tabela* T *satisfaz a propriedade de k-anonimity em relação ao conjunto dos Quasi-Identifiers* Q_T *se e só se para cada registo* r *em* T *existam pelo menos outros (k-1) registos em* T *que sejam indistinguiveis de* r *em relação a* Q_T

Desta maneira, mesmo que um adversário saiba os valores dos atributos Quasi-Identifiers de um dado indivíduo e também saiba que o registo desse indivíduo se encontra numa tabela T que satisfaça a condição de k-anonimity, temos a garantia de que ele não consegue determinar qual o registo em T que corresponde ao indivíduo com uma probabilidade maior do que $1/k$.

A um grupo de registos de uma tabela que são indistinguíveis entre si em relação a um conjunto de Quasi-Identifiers dá-se frequentemente o nome de classe de equivalência.

Sweeney propôs em [5] duas formas de atingir k-anonimity: por generalização e por omissão. Com generalização os valores reais são substituídos por valores menos específicos mas seman-

ticamente consistentes. Dado um certo domínio, podem existir várias formas de o generalizar. Normalmente valores numéricos como a idade ou código-postal são generalizados para um intervalo. Por exemplo 21 pode-se transformar em [20-25]. Valores categóricos tais como sexo ou nacionalidade são frequentemente generalizados para um conjunto de valores ou um valor singular de domínio mais abrangente. Por exemplo Portugal poderia ser generalizado para {Portugal, Espanha} ou para Europa. Com a técnica de omissão simplesmente se remove um tuplo completo ou parte dele de forma a satisfazer a definição de *k*-anonimity. Neste caso existe uma maior perda de informação sendo o resultado final menos útil e como tal esta técnica é apenas usada em conjunto com a generalização em casos específicos. Um exemplo dum caso destes acontece quando numa dada tabela a anonimizar existe um número baixo de registos cujos valores diferem substancialmente dos restantes registos. Nesta situação obtemos um melhor resultado final em relação à utilidade dos dados se simplesmente removermos estes registos em vez de generalizarmos os valores com intervalos muito grandes.

A tabela 3 apresenta a versão anonimizada da tabela 2 que contém um exemplo de registos médicos com apenas o nome removido.

Idade	Sexo	Código-Postal	Doença
21	M-F	[4700-4800]	Pneumonia
21	M-F	[4700-4800]	Cancro
22	M	[4700-4900]	Gripe
22	M	[4700-4900]	HIV
[23-24]	M-F	4700	Gripe
[23-24]	M-F	4700	Tuberculose

Tabela 3.: Exemplo de *k*-anonimity, com *k*=2 e $Q_T = \{\text{Idade, Sexo, Código-Postal}\}$

Esta tabela satisfaz a definição de *k*-anonimity com *k*=2. Podemos considerar que se encontra dividida em três classes de equivalência, cada uma com pelo menos *k* elementos indistinguíveis entre si relativamente ao conjunto dos Quasi-Identifiers. Se esta tabela fosse publicada em vez da tabela 2 apresentada no exemplo anterior, o adversário, mesmo sabendo os valores dos Quasi-Identifiers do João obtidos através da tabela 1, já não conseguiria identificar a doença do João. Apenas conseguiria determinar que o seu registo seria um dos dois primeiros tuplos da tabela (na primeira classe de equivalência) e que a sua doença tanto poderia ser pneumonia como cancro.

No entanto a definição de *k*-anonimity tem alguns problemas inerentes e está sujeita a alguns ataques. Um dos mais óbvios é o ataque por homogeneidade dos atributos sensíveis que foi evidenciado em [6] por Machanavajjhala et al.. Geralmente nos problemas de anonimização um dos

principais objetivos é proteger os atributos sensíveis dos indivíduos. Porém, o problema passa pelo facto de a definição de k-anonimity não oferecer segurança neste aspeto pois a definição é apenas restritiva relativamente aos Quasi-Identifiers. Há assim em certos casos a possibilidade de se descobrirem os atributos sensíveis de um indivíduo mesmo sem o associar diretamente a um registo numa tabela. Por exemplo considerando uma tabela que satisfaça o principio de k-anonimity em que todos os atributos sensíveis de uma classe de equivalência são iguais. Apesar de não ser possível fazer corresponder um indivíduo a um dos registos dessa classe de equivalência, conseguimos determinar o valor do seu atributo sensível com probabilidade de 1 pois todos os registos dessa classe têm o mesmo atributo sensível. Para resolver esta questão, Machanavajjhala et al. [6] propôs a noção de *l-diversity*.

Definição 3 (ℓ-diversity)**.** *Uma tabela T satisfaz a propriedade de* ℓ-diversity *se para cada classe de equivalência de T os registos nela contidos possuem pelo menos* l *valores de atributos sensíveis diferentes.*

Uma vez que o requisito da definição de ℓ*-diversity* garante que cada classe de equivalência contenha sempre pelo menos ℓ valores de atributos sensíveis distintos, a probabilidade de um adversário descobrir um atributo sensível de um indivíduo é sempre inferior a $1/l$.

A definição de ℓ*-diversity* garante também ℓ*-anonimity* uma vez que cada classe de equivalência tem que ter pelo menos ℓ registos.

Naturalmente, nestas definições de privacidade, quanto maior for o k, ou o ℓ no caso da ℓ*-diversity*, maiores são as garantias de segurança oferecidas pelas definições. Quanto maiores forem estes valores, mais difícil se torna no entanto satisfazer estas propriedades.

Um outro problema a que estas definições estão sujeitas é que não suportam a republicação de dados. Elas apenas garantem a segurança de uma dada tabela garantindo que esta segue os princípios de privacidade estabelecidos (i.e. *k-anonimity*, ℓ*-diversity*). Isto é, cada uma de várias tabelas publicadas podem ser consideradas seguras individualmente mas no entanto estarem sujeitas a inferências quando são observadas em conjunto. Esta condição torna-se bastante limitante pois hoje em dia os dados estão continuamente a ser recolhidos e existe uma grande exigência para que se encontrem atualizados. Este caso torna-se evidente se pensarmos por exemplo num hospital que está frequentemente a receber pacientes novos e quer fornecer aos seus investigadores dados sempre atualizados mas anónimos.

Para exemplificar este problema suponhamos que um hospital publica numa primeira fase a tabela 4 com registos médicos e posteriormente publica a tabela 5 que contém mais dois registos. Ambas as tabelas satisfazem a condição de ℓ*-diversity* para $\ell = 2$ e por consequência satisfazem também *k-anonimity* para $k = 2$.

Idade	Sexo	Código-Postal	Doença
[20-25]	M-F	[4700-4800]	Pneumonia
[20-25]	M-F	[4700-4800]	Cancro
[25-30]	M	4600	Gripe
[25-30]	M	4600	HIV
[30-40]	M-F	[4800-4900]	Gripe
[30-40]	M-F	[4800-4900]	Tuberculose
[30-40]	M-F	[4800-4900]	Cancro

Tabela 4.: Primeira publicação

Idade	Sexo	Código-Postal	Doença
[20-25]	M-F	[4700-4800]	Pneumonia
[20-25]	M-F	[4700-4800]	Cancro
[20-25]	M-F	[4700-4800]	Asma
[25-30]	M	4600	Gripe
[25-30]	M	4600	HIV
[30-35]	M-F	[4800-4900]	Gripe
[30-35]	M-F	[4800-4900]	Tuberculose
[35-40]	M-F	[4800-4900]	Cancro
[35-40]	M-F	[4800-4900]	Pneumonia

Tabela 5.: Segunda publicação

Neste contexto tal como no exemplo anterior considera-se que os adversários têm conhecimento sobre os Quasi-Identifiers dos indivíduos que querem atacar e neste caso sabem também em que tabelas os registos de um indivíduo estão publicados. Isto porque não é de todo descabido pensar que um adversário consiga saber quando o indivíduo sobre o qual ele esta interessado é por exemplo admitido num hospital e a partir daí deduzir em que tabelas ele se encontra. São de seguida mostrados dois tipos de inferências que um atacante poderia fazer olhando para as duas publicações.

Ao primeiro ataque dá-se o nome de ataque por diferença no qual um atacante que tem conhecimento que um indivíduo se encontra numa de duas tabelas publicadas calcula a diferença entre as classes de equivalência a que o indivíduo pertenceria em ambas as publicações. Desta forma um adversário pode obter um resultado que contém um número de registos inferior a k ou ℓ quebrando assim o principio de *k-anonimity* e/ou ℓ*-diversity*. Por exemplo imaginemos que um

adversário sabe que o André, que é seu vizinho e como tal o seu Código-Postal é 4750 foi admitido recentemente no hospital. Desta forma sabe que o registo do André se encontra na segunda publicação e não na primeira. Olhando para a segunda publicação (tabela 4) ele só sabe que o André poderá ter uma de três doenças $\{$Peumonia, Cancro, Asma$\}$. Mas fazendo a diferença com a classe de equivalência a que o André pertenceria na primeira publicação (tabela 5) ele conclui que o André foi admitido no hospital com asma. De notar que este ataque só é possível em bases de dados incrementais em que apenas são permitidas fazer inserções.

O segundo ataque é conhecido como ataque por intersecção. Neste caso um adversário sabendo que o indivíduo que quer atacar está presente em duas publicações diferentes, calcula a interceção das classes de equivalência em que o indivíduo se encontra dentro das tabelas publicadas de forma a tentar obter mais informação sobre o atributo sensível do seu alvo. Tal como no ataque anterior o resultado deste cálculo poderá reduzir as possibilidades de valores para o atributo sensível do indivíduo ou até mesmo revelar o valor do atributo. Para exemplificar este ataque suponhamos que o adversário sabe que o Mário tem 37 anos de idade e já se encontra em tratamento no hospital há bastante tempo e como tal tem a certeza que o seu registo está presente em ambas as tabelas. Observando a primeira publicação (tabela 4) o adversário sabe que o Mário tem uma das seguintes doenças $\{$Gripe, Tuberculose, Cancro$\}$. Olhando agora para a segunda publicação (tabela 5) tem a informação de que o Mário tem uma das duas doenças seguintes $\{$Cancro, Pneumonia$\}$. Desta forma fazendo a intersecção entre estes dois conjuntos o adversário descobre que o Mário tem cancro. Este segundo ataque funciona na presença de inserções e remoções pois estando o alvo presente em ambas as publicações este não vai ser afetado pelo aparecimento ou desaparecimento de novos registos uma vez que estes vão desaparecer com a interceção dos dois conjuntos.

Para solucionar estes ataques foi proposto em [7] por Barbosa et al. um algoritmo simples, genérico e eficiente. O objetivo deste algoritmo é analisar cada nova publicação e modificá-la de forma a evitar os ataques acima mencionados. Este algoritmo chamando BPG deve ser usado como uma camada superior de forma a se poder realizar republicação de dados de forma segura. Desta forma ele recebe uma tabela já anonimizada de acordo com certos princípios de proteção tais como *k-anonimity* e *l-diversity*. De seguida o algoritmo vai verificar todas as publicações anteriores que foram sendo guardadas ao longo do tempo e averiguar se com a presença da nova publicação os ataques por diferença ou interceção conseguem violar os princípios de segurança. Caso os princípios já não se satisfaçam na presença de diferenças ou interseções com a nova tabela, o algoritmo tenta modificar as classes de equivalência da nova publicação de forma a evitar estes ataques. Se existirem classes de equivalência que necessitem ser alteradas e isto não

seja possível estas poderão ser omitidas da publicação de modo a preservar a privacidade dos dados.

As noções de privacidade como *k-anonimity* e ℓ*-diversity* tiveram inicialmente algum sucesso na literatura devido a serem de fácil implementação e serem as primeiras soluções para o problema de privacidade em sistemas de dados. Possuem alguns problemas inerentes como os apresentados anteriormente e a principal crítica a estes modelos reside no facto das suas definições apenas restringirem a forma de como o output final tem que ficar e não restringem o algoritmo usado para atingir a tabela final. Outra critica existente é que estes modelos dependem do conhecimento que os adversários já possuem sobre os seus alvos (i.e. identificação correta dos Quasi-Indentifiers) e por causa disto uma incorreta identificação dos Quasi-Identifiers ou a existência de informação externa não prevista, pode sempre levar a ataques.

2.2 DIFFERENTIAL PRIVACY

Surgiu em [2] por Cynthia Dwork uma nova noção de privacidade chamada *differential privacy*. Os modelos utilizados até esse momento como *k-anonimity* e ℓ*-diversity* dependiam fortemente no conhecimento externo que os adversários possuíam sobre os seus alvos. Quem pretendia proteger os dados tinha que saber que atributos os adversários poderiam ter acesso de forma externa que pudessem cruzar com as tabelas publicadas de forma a quebrar a privacidade dos indivíduos nelas presentes. Isto na prática é difícil de realizar. Ao contrário dos modelos anteriores, a *differential privacy* não faz praticamente nenhum pressuposto sobre o conhecimento que um atacante pode ter. A noção de *differential privacy* oferece uma definição mais rigorosa que tenta garantir que a adição ou remoção de um registo numa base de dados não altera significativamente o resultado de uma análise sobre a mesma. Dwork com esta definição quer capturar a ideia de que o risco de um indivíduo ter a sua privacidade violada não deve aumentar substancialmente por este fazer parte de uma base de dados estatística.

No contexto da análise de dados privados surgem normalmente dois modelos, o interativo e o não interativo. No modelo interativo a base de dados é mantida por um coletor de dados confiável, que pode ser uma pessoa ou uma organização, que altera as verdadeiras respostas às perguntas feitas pelos utilizadores de modo a proteger a privacidade dos indivíduos contidos na base de dados. No modelo não interativo o coletor dos dados ou publica de uma só vez uma versão anonimizada dos dados ou calcula e publica uma série de estatísticas sobre os dados normalmente sobre a forma de histogramas ou tabelas de contingência. O modelo de *differential privacy* permite tratar ambos estes casos. Foi mostrado em [8] no entanto que o modelo não interativo é

mais difícil de implementar principalmente devido à distância temporal entre a anonimização dos dados e a sua utilização pois torna-se difícil fornecer utilidade a algo que ainda não se encontra especificado na altura da publicação.

De forma geral, a *differential privacy* requer que a resposta a uma qualquer *query* seja probabilisticamente indistinguível quer o registo de um indivíduo esteja ou não presente numa base de dados. Note-se que pode haver na mesma divulgação de informação sobre o indivíduo, mas esta garantia assegura que esta divulgação não é devido à presença do indivíduo nos dados.

Definição 4 (ϵ-*differential privacy* [2])**.** *Um algoritmo aleatório* $K : D \rightarrow R$ *fornece* ϵ-differential privacy *se para duas bases de dados* D_1 *e* D_2 *que diferem no máximo em um elemento e para qualquer* $O \subseteq R$*,*

$$Pr[K(D_1) \subseteq O] \leq e^{\epsilon} \cdot Pr[K(D_2) \subseteq O]$$

Olhando para a definição podemos concluir que ϵ*-differential privacy* oferece uma definição forte de segurança. Se K satisfaz ϵ*-differential privacy* pode-se concluir que publicar $K(D)$ não constitui uma violação de privacidade de qualquer tuplo t pertencente a D. Isto porque mesmo que t não esteja incluído em D, caso em que consideramos que a privacidade de t está assegurada, com probabilidade semelhante as mesmas conclusões podem ser retiradas sobre os dados.

Esta definição na prática pode ser difícil de satisfazer. Desta forma é por vezes utilizada uma definição ligeiramente mais fraca proposta em [9] que permite uma probabilidade de erro δ.

Definição 5 ((δ,ϵ)-*differential privacy*) [9])**.** *Um algoritmo aleatório* $K : D \rightarrow R$ *fornece* (δ,ϵ)-differential privacy *se para duas bases de dados* D_1 *e* D_2 *que diferem no máximo em um elemento e para qualquer* $O \subseteq R$*,*

$$Pr[K(D_1) \subseteq O] \leq e^{\epsilon} \cdot Pr[K(D_2) \subseteq O] + \delta$$

De forma a satisfazer a definição de *differential privacy* Dwork sugere em [2], [10] a adição de ruído aleatório obtido através de uma distribuição probabilística de Laplace às verdadeiras respostas de uma *query*. A quantidade de ruído adicionado a cada resposta depende do grau de sensibilidade de cada *query* e do parâmetro de segurança *epsilon*. O grau de sensibilidade global de uma *query* é a diferença máxima que a resposta a essa *query* poderia ter caso fosse feita em duas bases de dados que diferem no máximo em um elemento.

Definição 6 (Grau de Sensibilidade [2])**.** *Seja* $q(D)$ *uma* query *a uma base de dados* D*, a sensibilidade de* q*, dada por* Δ_q *é a diferença máxima que* q *pode tomar quando se adiciona ou remove um tuplo de* D*:*

$$\Delta_q = max \, \|q(D_1) - q(D_2)\|$$

para todos os D_1, D_2 que diferem no máximo em um elemento.

Por exemplo o grau de sensibilidade para a *query*: "Quantas pessoas têm mais de 25 anos de idade?"é 1 pois adicionando ou removendo um registo da base de dados a resposta a esta *query* vai ser alterada no máximo em 1.

Assim um mecanismo K que queira responder a uma *query* q sobre uma base de dados D satisfazendo a definição de *ϵ-differential privacy*, tem que publicar $K(D) = q(D) + X$ onde X é o ruído aleatório retirado da distribuição probabilística de Laplace com valor médio de 0 e escala Δ_q/ϵ.

A distribuição de Laplace tem a seguinte função densidade de probabilidade:

$$f(x, \mu, b) = \frac{1}{2 \cdot b} \cdot e^{-\frac{|x-\mu|}{b}}$$

onde para um dado valor x, o parâmetro do valor médio μ e o parâmetro escala b, a função retorna a probabilidade de, retirando um valor aleatório da distribuição probabilística de Laplace, esse valor ser x.

Para o nosso contexto, o valor médio é sempre centrado a 0 e o valor escala b é definido a Δ_q/ϵ.

Assim, o ruído acrescentado a cada *query* por este modelo depende do parâmetro sensibilidade, Δ_q e do parâmetro de segurança *epsilon*, ϵ, em concreto na forma Δ_q/ϵ. A figura 1 mostra a distribuição de probabilidades de Laplace para alguns valores fixos.

Naturalmente este mecanismo produz resultados com mais utilidade quando o valor de sensibilidade de uma *query* é baixo, pois será necessário adicionar menos ruído.

O mecanismo de adição de ruído proveniente de uma distribuição de Laplace apenas funciona no entanto para *queries* de valor numérico uma vez que não faz sentido adicionar ruído a valores não numéricos. De forma a se conseguir responder a *queries* não numéricas preservando *differential privacy* foi proposto em [11] um mecanismo exponencial. No entanto este modelo é não interativo o que significa que todas as *queries* têm que ser feitas antecipadamente ao contrário do mecanismo de Laplace. Outro inconveniente é que o mecanismo exponencial é altamente ineficiente.

A mais conhecida implementação deste modelo de privacidade é o PINQ [12] de *Privacy Integrated Queries*. O sistema foi criado por um investigador da Microsoft, Frank McSherry, é implementado em C# e utiliza o sistema de *queries* LINQ. O PINQ age como uma camada

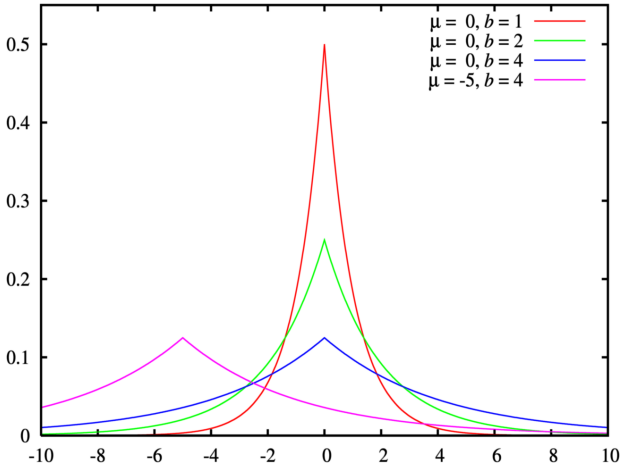


Figura 1.: Distribuição de probabilidades de Laplace

intermédia entre uma base de dados e o analista de dados deixando-o efetuar *queries* à base de dados de forma segura, aplicando o modelo de *differential privacy*.

Outra conhecida implementação do modelo de *differential privacy* é o sistema FUZZ[13] que surge dois anos mais tarde tentando colmatar alguns dos problemas do PINQ.

2.3 VARIANTES DE DIFFERENTIAL PRIVACY

Surgiu recentemente em 2012 a noção de *Crowd Blending Privacy* proposta em [3]. Os autores deste modelo queriam chegar a um modelo de privacidade que não sofresse das falhas que os modelos anteriores possuíam. O modelo de *k-anonimity* apesar de tentar captar a ideia de "*blending in a crowd*", pois obriga os registos de uma tabela publicada a estarem misturados no meio de outros registos com atributos semelhantes, não o consegue fazer de forma a permitir uma análise teórica análoga à existente para *differential privacy* pois este modelo apenas restringe o output das tabelas publicadas e não o mecanismo para as obter.

Precisamente uma das ideias chave do modelo mais consensual na literatura de atingir privacidade, *differential privacy*, é a ideia de que a privacidade deve ser uma propriedade do mecanismo que a obtém e não apenas do resultado final. Porém as principal criticas dos autores ao modelo de *differential privacy* são que para além de algumas técnicas usadas para atingir *differential privacy* serem pouco eficientes estas necessitam obrigatoriamente da adição de ruído aos seus resultados de modo a preservar a privacidade dos indivíduos presentes numa base de dados. Esta adição de ruído por sua vez faz com que os resultados publicados percam utilidade.

De forma a dar a volta a este problema, os autores deste novo modelo exploraram a possibilidade de em muitos casos os dados que são recolhidos para fins estatísticos podem ser obtidos através de amostras aleatórias da população. Intuitivamente o facto de a amostragem da população ser aleatória já fornece algumas garantias básicas de privacidade e isto pode ajudar a reduzir a quantidade de ruído necessária à privatização dos dados [14].

O objetivo dos autores passava então por encontrar uma nova definição de privacidade que capture corretamente a ideia de "*blending in a crowd*", que seja mais fraca que *differential privacy* de forma a ser mais eficiente e a conseguir fornecer maior utilidade e que em situações onde os dados são recolhidos através de uma amostragem aleatória à população a definição consegue atingir ambas *differential privacy* e *zero knowledge privacy*.

A definição de *zero knowledge privacy* foi proposta em [15] de forma a endereçar a privacidade em contextos de redes sociais. È uma definição estritamente mais forte que a de *differential privacy*. De forma geral, podemos interpretar a definição de *differential privacy* dizendo que um atacante não consegue aprender mais sobre um dado indivíduo do que aquilo que consegue aprender olhando para todos os restantes registos de uma base de dados. No entanto no contexto das redes sociais, pode existir um grupo de registos cuja informação esteja mais relacionada com o indivíduo (i.e. os seus amigos). Nestes casos a definição de *differential privacy* pode não fornecer garantias de segurança suficientes. Surge assim o modelo de *zero knowledge privacy* que diz que tudo o que um atacante pode saber sobre um dado indivíduo, também o pode saber analisando a informação de apenas k indivíduos presentes na base de dados. Se k for igual ao número total de indivíduos a definição de *zero knowledge privacy* torna-se igual à definição de *differential privacy*.

Adicionalmente mesmo que a amostragem dos dados não seja completamente aleatória e o atacante possa inclusive saber que certos indivíduos se encontram ou não na base de dados, a definição deve ser na mesma válida e fornecer garantias de segurança mesmo de forma independente.

A definição de privacidade a que os autores chegaram foi a de *crowd blending privacy*, mas para a definirem precisavam primeiro de introduzir a noção de indistinguibilidade. Esta definição diz que dois tuplos são indistinguíveis em relação a um dado mecanismo se para qualquer base de dados, substituindo um tuplo pelo outro o mecanismo dá sensivelmente o mesmo resultado.

Definição 7 (ϵ-*indistinguishable* [15])**.** *Dados dois tuplos t_1 e t_2 e um mecanismo San dizemos que t_1 e t_2 são* ϵ-indistinguishable *por San se:*

$$San(D, t_1) \approx_\epsilon San(D, t_2), \forall D$$

Normalmente t_1 e t_2 são registos pertencentes a dois indivíduos. Se t_1 e t_2 são indistinguíveis por San, então o mecanismo não consegue distinguir estes dois indivíduos independentemente de quem mais se possa encontrar na base de dados. Estando a noção de indistinguibilidade definida, os autores propõem então a definição de *crowd blending privacy*.

Definição 8 ((k,ϵ)-*Crowd Blending Privacy*) [15]**.** *Um mecanismo San satisfaz a definição de* (k,ϵ)-Crowd Blending Privacy *se para qualquer base dados D e qualquer indivíduo t contido em D, ou*

- *t é* ϵ-indistinguishable *de pelo menos k indivíduos em D, ou*
- *t é essencialmente ignorado:* $San(D) \approx_\epsilon San(D \backslash \{t\})$

Esta definição de privacidade requer através da primeira cláusula que cada indivíduo presente numa base de dados tem que ser indistinguível de pelo menos outros k indivíduos da base de dados. A segunda cláusula é de certa forma uma relaxação à primeira permitindo a existência na base de dados de indivíduos que não são indistinguíveis de k outros indivíduos desde que o mecanismo San os ignore garantindo que $San(D) \approx_\epsilon San(D \backslash \{t\})$.

A definição de *crowd blending privacy* garante a privacidade de um indivíduo t pois se este é indistinguível de um grupo de k outros indivíduos, o mecanismo não vai publicar informação sobre t para além das características gerais do grupo sobre o qual ele é indistinguível. Isto porque o indivíduo t poderia ser substituído por qualquer outro indivíduo daquele grupo que os resultados publicados pelo mecanismo seriam praticamente os mesmos. Desta forma é garantido que um mecanismo que satisfaça *crowd blending privacy* não publica informação privada de um indivíduo t.

Assim é mostrado pelos autores em [15] que esta definição é mais fraca que *differential privacy* conseguindo assim uma definição menos restritiva capaz de produzir resultados com mais utilidade. Apesar desta noção de privacidade oferecer garantias de segurança de forma independente, o objetivo é ser utilizada em dados que tenham sido retirados de uma amostragem aleatória à população. O principal teorema apresentado pelos autores mostra que se a definição de *crowd blending privacy* caso seja utilizada e em dados provenientes de uma amostra aleatória da população consegue atingir níveis de privacidade mais fortes como *zero knowledge privacy* que por sua vez implica *differential privacy*.

3

COMPOSIÇÃO DE MECANISMOS DE ANONIMIZAÇÃO

O objetivo desta dissertação e principal contribuição passa pela exploração e pelo estudo das técnicas de *k-anonimity* e *differential privacy*. Não apenas quando são utilizadas de forma independente, mas especialmente estudar o caso em que estas duas técnicas que surgem de contextos diferentes e que raramente se tocam na literatura são utilizadas em simultâneo.

Mais concretamente, o objetivo é a análise da composição destas duas técnicas sobre um caso real. Isto é, o estudo do caso em que partindo de um dado *dataset* lhe aplicamos primeiro um algoritmo de anonimização de *k-anonimity* e em seguida ao resultante *dataset* já anonimizado fazemos agora *queries* segundo o modelo de *differential privacy*.

Este caso pode ocorrer num contexto real da seguinte forma. Considerando que existe um coletor de dados que trata do seu armazenamento e responde a *queries* feitas por utilizadores aos dados segundo o modelo de *differential privacy*. Por sua vez, este coletor de dados pode já ter recebido os dados pré-processados sobre a forma de um algoritmo de anonimização *k-anonimity* por exemplo. Isto pode ocorrer devido aos dados recebidos serem por exemplo provenientes de uma instituição médica que se encontra sobre obrigação de proteger a privacidade dos seus pacientes e tenha decidido apenas publicar dados sensíveis referentes aos seus pacientes de forma anonimizada com um algoritmo de *k-anonimity*.

Formalmente podemos definir este caso de estudo como:

$$DP_{\epsilon,\Delta_q} \circ q \circ A_k \circ D$$

onde D é um conjunto de dados, A_k um algoritmo de anonimização de *k-anonimity* com parâmetro de segurança k, q uma *query* e DP_{ϵ,Δ_q} o ruído adicionado à *query* proveniente do método de *differential privacy* onde ϵ é o parâmetro de segurança *epsilon* e Δ_q a sensibilidade da *query*.

3.1 TIPOS DE QUERIES NO CONTEXTO DE DADOS GENERALIZADOS

Para a realização de alguns dos testes aos conjuntos de dados surgiu a necessidade de se efetuarem *queries* de contagem quer aos conjuntos de dados originais, quer aos conjuntos de dados após anonimização. A título de exemplo, uma *query* deste tipo pode ser: "Qual o número de pessoas cuja idade está entre os 30 e 35 anos ?".

Ora, efetuar uma *query* deste tipo ao conjunto de dados normal, onde os valores exatos dos atributos estão todos especificados, é um procedimento direto e trivial.

No entanto isto já não acontece quando queremos aplicar este tipo de *queries* ao conjunto de dados já anonimizado uma vez que não temos agora acesso ao valor exato dos atributos como tínhamos no caso anterior. Isto porque após a anonimização dos dados, os valores dos atributos encontram-se substituídos por um intervalo de valores. Mais concretamente onde inicialmente tínhamos a indicação de que a idade de um certo indivíduo era de 25 anos, podemos ter após anonimização a indicação de que a idade desse mesmo indivíduo é por exemplo [23-28]. De notar que nem todos os valores necessitam de generalização, podendo estes ser representados por [25-25]. Este caso contudo não é frequente e no geral os intervalos têm mais do que um valor possível.

Surge assim o problema de como transportar as *queries* de contagem ditas tradicionais para os conjuntos de dados anonimizados, ou seja como efetuar *queries* a um conjunto de dados cujos valores dos atributos aparecem generalizados.

Para solucionar este problema foram definidos dois tipos de *queries* de contagem para funcionarem neste contexto em que os valores dos atributos se encontram generalizados.

3.1.1 *Queries de Inclusão*

Ao primeiro tipo de *queries* demos o nome de *queries* de inclusão. Estando os valores dos atributos dos indivíduos generalizados em intervalos, quando efetuamos *queries* deste tipo consideramos para a contagem apenas os indivíduos cujo intervalo do atributo em questão está completamente contido no intervalo definido pela *query*.

Definição 9 (*Query* de Inclusão)**.** *Uma* query *de contagem feita a um conjunto de dados generalizado é inclusiva quando apenas consideramos para a contagem os indivíduos cujo intervalo do atributo ou atributos questionados pela* query *estão completamente contidos no intervalo definido pela mesma.*

Por exemplo se uma *query* quiser contar todos os indivíduos cuja idade está entre o intervalo [20-30], um dado indivíduo cujo atributo idade esteja generalizado para:

- [23-26], é naturalmente incluído na contagem pois todo o intervalo está contido no intervalo da *query*.
- [35-40], é ignorado pela contagem devido a todo o intervalo se encontrar fora dos limites da *query*.
- [28-32], é também ignorado pela contagem porque apesar de parte da generalização estar incluída no intervalo da *query*, esta não está totalmente incluída, o que é necessário numa *query* de inclusão.

Deste modo, com *queries* deste género estamos possivelmente a omitir da contagem indivíduos que seriam seguramente contabilizados caso a mesma *query* fosse realizada ao conjunto de dados original, antes da anonimização.

Isto pode acontecer no terceiro caso do exemplo anterior, imaginando que um indivíduo tem 28 anos de idade no conjunto de dados original e após anonimização a sua idade é generalizada para [28-32]. Ora a *query* [20-30] do exemplo sendo realizada ao conjunto de dados original iria seguramente considerar este indivíduo de 28 anos na contagem, mas sendo realizada ao mesmo conjunto de dados mas anonimizado já não consideraria a idade [28-32] para a contagem.

O valor retornado por uma qualquer *query* de contagem de inclusão feita a um conjunto de dados anonimizado vai assim ser sempre menor ou igual do que o valor retornado pela mesma *query* feita ao conjunto de dados original antes de efetuada a anonimização.

Isto ocorre pois de todos os indivíduos que são apanhados pela *query* antes da anonimização, alguns deles já não vão ser apanhados pela mesma *query* após a anonimização. O que significa que o valor da contagem pode diminuir. Por outro lado, todos os indivíduos que não são considerados pela *query* ao conjunto de dados original, não têm hipótese nenhuma de serem considerados pela *query* agora inclusiva ao conjunto de dados após anonimização. Pois como uma generalização tem de incluir sempre no seu intervalo o valor original, o intervalo da generalização iria no máximo sobrepor-se a um dos limites da *query* tal como no ponto 3 do exemplo anterior onde já foi visto que este caso não é considerado pelas *queries* de inclusão. Assim o valor da contagem não pode aumentar, apenas diminuir.

3.1.2 *Queries de Sobreposição*

Foi dado o nome de *queries* de sobreposição ao segundo tipo de *queries* de contagem definidas para trabalhar neste contexto.

Quando uma *query* deste tipo é efetuada a um conjunto de dados anonimizado são considerados para a contagem todos os indivíduos cuja generalização do ou dos atributos relevantes para a *query* se encontre abrangida pelo intervalo da *query*, mesmo que o intervalo da *query* abranja apenas uma parte da generalização e não todo o intervalo como era requisito para as *queries* de inclusão.

Definimos mais formalmente esta *query* da seguinte forma.

Definição 10 (*Query* de Sobreposição)**.** *Uma* query *de contagem feita a um conjunto de dados generalizado é de sobreposição quando consideramos para a contagem os indivíduos cujo intervalo do atributo ou atributos questionados pela* query *estão incluídos de alguma forma no intervalo definido pela mesma.*

Exemplificando com valores reais tal como foi feito para as *queries* de inclusão, se uma *query* quiser contar todos os indivíduos cuja idade está entre o intervalo [20-30], um dado indivíduo cujo atributo idade esteja generalizado para:

- [23-26], é incluído na contagem pois todo o intervalo se encontra contido no intervalo da *query*.
- [35-40], não é considerado para a contagem pois todo o intervalo do atributo está fora dos limites definidos pela *query*.
- [28-32], é incluído na contagem de uma *query* de sobreposição porque apesar de todo o intervalo da generalização do atributo não estar completamente contido nos limites da *query*, isso não é necessário. É suficiente para este caso que apenas parte da generalização esteja dentro do intervalo da *query*.

A diferença entre os dois tipos de *queries* reside só no terceiro ponto, no qual as *queries* de inclusão ignoram os registos cuja generalização dos atributos não se encontrem totalmente incluídos no intervalo definido pela *query* e as *queries* de sobreposição acrescentam estes mesmos registos à contagem.

Usando *queries* de sobreposição estamos assim a possivelmente incluir na contagem registos que não seriam contabilizados caso a mesma *query* fosse efetuada ao conjunto de dados pré-anonimização, antes dos atributos serem generalizados.

Novamente uma instância desta ocorrência pode ser baseada no terceiro ponto do exemplo. Considerando um indivíduo com 32 anos de idade presente no conjunto de dados original e que após anonimização o seu atributo idade é generalizado para [28-32]. A *query* [20-30] apresentada no exemplo sendo efetuada ao conjunto de dados original não incluiria este indivíduo de 32 anos no resultado dessa contagem, mas no entanto se a mesma *query* agora de sobreposição for aplicada ao conjunto de dados já anonimizado esta já vai incluir o registo cuja idade é [28-32] na contagem.

Também paralelamente ao que foi demonstrado para as *queries* de inclusão, pode-se mostrar que o valor retornado por uma qualquer *query* de sobreposição feita ao conjunto de dados anonimizado vai ter sempre um valor maior ou igual do que o valor retornado pela mesma *query* efetuada ao conjunto de dados original antes da anonimização, nunca menor.

A mesma ideia pode ser usada, mas de forma inversa. Isto é, já foi visto que de todos os indivíduos que não são considerados pela *query* antes da anonimização, alguns deles vão ser incluídos pela mesma *query* após anonimização, ou seja o valor da contagem pode subir. Por outro lado, todos os indivíduos que são considerados pela *query* de contagem ao conjunto de dados original antes da anonimização vão também ser sempre considerados pela mesma *query* quando efetuada ao conjunto de dados já anonimizado. Como a generalização de um atributo inclui sempre no seu intervalo o valor original, se o valor original do atributo estava contido no intervalo da *query*, o intervalo da sua generalização vai no máximo sobrepor-se a um dos limites da *query*, caso demonstrado no terceiro ponto do exemplo anterior que é na mesma contabilizado pelas *queries* de sobreposição. Ou seja o valor da contagem não pode descer, apenas subir.

3.2 PROBLEMA COM A COMPOSIÇÃO DAS TÉCNICAS

Com o estudo e composição das duas principais técnicas de obtenção de privacidade em sistemas de dados, *differential privacy* e *k-anonimity*, um dos objetivos da dissertação passaria por utilizando os dois modelos em simultâneo tentar reduzir os valores dos parâmetros de segurança mas mantendo os mesmos níveis de privacidade. Isto seria um resultado natural e intuitivo uma vez que a fase de anonimização retira de facto informação e qualidade ao *dataset* o que à partida significa que com a composição das duas técnicas o nível de segurança obtido não poderia baixar. No entanto o que mostramos em seguida é precisamente o contrário, que esta intuição não é correta. Na verdade a utilização dos dois modelos em conjunto pode levar a uma perda do nível de segurança e quebrar a garantia de privacidade do modelo de *differential privacy*.

No contexto desta dissertação, o nosso caso de estudo é aquele em que partimos de um *dataset* que é anonimizado por um algoritmo de anonimização de *k-anonimity* onde sobre isto são de seguida efetuadas *queries* de acordo com o modelo de *differential privacy*.

Ao longo da dissertação são utilizadas principalmente *queries* de contagem. Segundo o modelo de *differential privacy* as *queries* de contagem têm sensibilidade de 1. A sensibilidade de uma *query* é um parâmetro de calibração do modelo de *differential privacy* cujo valor é dado pela variação máxima que a reposta a essa mesma *query* pode ter quando executada a dois *datasets* que diferem no máximo em um elemento. Ora dada uma qualquer *query* de contagem, adicionando ou removendo um elemento a um conjunto de dados, a resposta a essa *query* vai sofrer no máximo uma alteração de 1 devido ao elemento que foi acrescentado ou removido.

No entanto no nosso cenário de estudo, calibrar o parâmetro da sensibilidade a 1 mesmo para *queries* de contagem pode levar a uma perda do nível de segurança podendo até a garantia de privacidade do modelo de *differential privacy* ser destruída. Isto ocorre porque a adição ou remoção de um indivíduo ao *dataset* neste caso pode causar alterações significativas nas classes de equivalência geradas pelo algoritmo de anonimização. Estas alterações podem incluir o aparecimento ou desaparecimento de classes de equivalência inteiras. Isto faz com que a variação máxima que a resposta a uma *query* neste contexto pode ter, possa tomar valores muito maiores do que 1 quando se acrescenta ou remove um indivíduo ao *dataset*. Desta maneira a premissa do modelo de *differential privacy* pode assim ser quebrada. Relembrando a definição de privacidade do modelo de *differential privacy* já apresentada no estado da arte no capítulo 2.2 a sua definição é a seguinte:

$$Pr[K(D_1) \subseteq O] \leq e^{\epsilon} \cdot Pr[K(D_2) \subseteq O]$$

Em concreto, um algoritmo K está de acordo com o modelo de *differential privacy* se quando aplicado a dois conjuntos de dados D_1 e D_2 que diferem no máximo em um elemento, as probabilidades do algoritmo dar como output um certo resultado em ambos os casos não podem diferir por um fator maior que do que e^{ϵ}. Se isto acontecer, um adversário não consegue determinar se um indivíduo pertence ao *dataset* ou não, visto que em ambos os casos a probabilidade do mesmo resultado ser retornado é sensivelmente a mesma, menor que e^{ϵ}.

O objetivo é assim o de mostrar que no nosso caso de estudo, onde efetuamos *queries* segundo o modelo de *differential privacy* em cima de um *dataset* já anonimizado por um algoritmo de anonimização, dados dois *datasets* que diferem apenas em um elemento, o fator das probabilidades de duas *queries* de contagem efetuadas a cada um dos *datasets* segundo o modelo de *differential privacy* com o parâmetro de sensibilidade calibrado a 1 retornarem um certo resul-

tado pode ser maior que e^{ϵ}. Se isto ocorrer, a garantia dada pelo modelo de *differential privacy* é quebrada.

Para mostrar isto foram realizados os seguintes passos.

Em primeiro lugar foi observado o resultado de uma anonimização ao conjunto de dados com o parâmetro de segurança k selecionado a um valor comum na literatura, 10. Com isto é pretendido encontrar no resultado da anonimização classes de equivalência com exatamente dez elementos. A intuição é a de que removendo um dos indivíduos pertencentes a uma destas classes do *data-set* original, obteremos uma anonimização substancialmente diferente do que com o indivíduo presente. Pois com a remoção deste elemento, a classe de equivalência onde este se encontrava não pode continuar a existir após a anonimização uma vez que esta já continha o número mínimo de elementos permitidos pela restrição de *k-anonimity* de 10. Garantimos assim que removendo um elemento ocorrem diferenças significativas entre duas anonimizações que normalmente vão passar pelo desaparecimento da classe de equivalência, formando os restantes elementos uma nova classe ou juntando-se a uma classe mais abrangente.

De seguida é necessário calcular as probabilidades de duas *queries* neste contexto retornarem um certo valor de forma a se poder mostrar que de facto a fórmula pode ser quebrada. Para isto após efetuadas as *queries* são registados os seus verdadeiros valores, que se tudo tiver corrido bem até aqui, serão distintos com uma diferença superior a 1. De seguida é necessário obter a probabilidade de, após a adição do ruído do modelo de *differential privacy*, que origina de uma distribuição aleatória de Laplace de parâmetro Δ_q/ϵ, o valor final ser por exemplo X. Ora, sendo os verdadeiros valores das respostas antes da adição de ruído dados por A e B, o que queremos na verdade são as probabilidades da distribuição de Laplace retornar os valores de Y e Z, onde $X = A + Y$ e $X = B + Z$. Desta forma, as probabilidades de no nosso contexto a composição dos modelos retornar um certo valor, são de facto as probabilidades da distribuição de Laplace retornar os valores de Y e Z.

Olhando para a fórmula de segurança do modelo de *differential privacy* da seguinte maneira:

$$\frac{Pr[K(D_1) \subseteq O]}{Pr[K(D_2) \subseteq O]} \leq e^{\epsilon}$$

caso $\frac{Pr[Y]}{Pr[Z]} \geq e^{\epsilon}$, a garantia de privacidade deste modelo é quebrada.

Para se obterem as probabilidades da distribuição de Laplace retornar Y e Z é utilizada a função densidade de probabilidade já apresentada na secção 2.2 que é a seguinte:

$$f(x, \mu, b) = \frac{1}{2 \cdot b} \cdot e^{-\frac{|x-\mu|}{b}}$$

onde x é o valor para o qual queremos saber a probabilidade, que será assim substituído por Y ou Z, μ é o centro da distribuição que é 0 e b é o parâmetro escala que será substituído por Δ_q / ϵ.

Assim, estando o ataque apresentado de forma teórica é de seguida mostrado um caso concreto com valores reais que confirmam esta hipótese.

Após a anonimização do *dataset* census-10k, *dataset* sintético detalhado na secção 4.1, com o algoritmo de Mondrian calibrado para k igual a dez, foram procuradas no *dataset* resultante classes de equivalência com exatamente dez elementos. Entre outras, foi encontrada a seguinte classe de equivalência: [[65:67], [1:1], [35000:36000]]. Isto significa que todos os dez indivíduos cujos atributos Quasi-Identifiers foram generalizados para esta classe têm a sua idade entre 65 e 67 anos, são do sexo masculino e o seu código-postal encontra-se entre os valores de 35000 e 36000. De seguida é removido do conjunto de dados original um dos indivíduos que faria parte desta classe. Neste caso foi removido o elemento da linha 45 do *dataset* com os seguintes dados: 67, 1, 35000, 120328. De seguida o conjunto de dados original é anonimizado novamente mas agora sem este elemento. Observando agora o resultado desta anonimização podemos reparar que de facto a classe de equivalência a que este elemento pertencia, [[65:67], [1:1], [35000:36000]], desapareceu. Em concreto as diferenças entre estas duas anonimizações em que uma delas contém menos um elemento que a outra são mostradas na seguinte tabela:

Dataset 1	Nº	Dataset 2	Nº
[65:67], [1:1], [33000:34000]	15	[65:67], [1:1], [33000:36000]	24
[65:67], [1:1], [35000:36000]	10		

Tabela 6.: Diferenças entre as duas anonimizações

A coluna representada pelo *Dataset* 1 mostra as classes de equivalência que existem após a primeira anonimização e não existem na segunda. A segunda coluna mostra o caso contrário. As três classes de equivalência apresentadas na tabela são as únicas afetadas pela remoção de um elemento do conjunto de dados original, todo o restante resultado da anonimização é igual. Neste caso com a remoção do conjunto de dados original de um dos elementos que após anonimização fazia parte da classe [[65:67], [1:1], [35000:36000]], após nova anonimização os restantes nove elementos desta classe juntaram-se a uma outra formando uma classe maior e mais abrangente com vinte e quatro elementos como se pode ver na tabela 6.

De forma a se quebrar a fórmula apresentada anteriormente é agora necessário realizar uma *query* com ruído adicionado às suas respostas segundo o modelo de *differential privacy* a ambos os *datasets* e comparar as probabilidades de obtenção de um certo resultado. No entanto com

uma qualquer *query* aleatória não conseguiremos quebrar a fórmula. A *query* efetuada terá de abranger a diferença encontrada entre as duas anonimizações de forma a existir ataque. Se a *query* utilizada for exatamente a *query* que pergunta quantos indivíduos existem na classe de equivalência que desapareceu de uma anonimização para a outra, [[65:67], [1:1], [35000:36000]], isto é quantos elementos no conjunto de dados têm entre 65 e 67 anos de idade, são do sexo masculino e o seu código-postal encontra-se entre 35000 e 36000, então irá ocorrer a quebra da fórmula. Considerando a interpretação inclusiva, esta *query* retorna o valor de 10 para o primeiro *dataset* e o valor de 0 para o segundo uma vez que o intervalo do atributo código-postal na nova classe que apareceu não esta totalmente incluído no intervalo perguntado pela *query*. Se a *query* for por outro lado considerada de sobreposição, a contagem retornada mantém-se a 10 para o primeiro caso, mas para o segundo é agora de 24 pois já inclui a nova classe na contagem.

Ora aplicando agora o modelo de *differential privacy*, que consiste na adição de ruído aleatório proveniente de uma distribuição de Laplace usando como parâmetros os valores da sensibilidade e do *epsilon* às respostas da *query*, a calibração do parâmetro da sensibilidade a 1, o usual para *queries* de contagem, não vai ser suficiente para cobrir a diferença entre as respostas desta *query* aos dois conjuntos de dados que variam apenas em um elemento e a garantia de privacidade será assim quebrada.

Em concreto, dos valores possíveis de serem retornados pela *query* após a adição de ruído vamos verificar as probabilidades de o valor retornado ser de 10. Para a interpretação inclusiva, sendo os verdadeiros valores retornados antes da adição de ruído 10 e 0 para cada um dos *datasets*, para se obter o valor final de 10, a distribuição de Laplace tem que retornar para o primeiro caso o ruído de 0 e para o segundo o ruído de 10. Sendo o modelo de *differential privacy* calibrado com sensibilidade de 1 e *epsilon* de 0.1, a distribuição de Laplace tem assim valor médio de 0 e escala de $\frac{1}{0.1}$. Utilizando a função densidade de probabilidade, a tabela 7 indica a probabilidade desta distribuição retornar alguns valores exemplo, incluindo aqueles necessários.

Com a observação da tabela, as probabilidades de após a adição de ruído se obter o valor de 10, são de 0.0500 e 0.0183 respetivamente. Calculando o rácio entre estas probabilidades, $\frac{0.0500}{0.0183} = 2.7322$, o valor obtido é superior ao valor de $e^{\epsilon} = 1.1051$. Isto significa que neste contexto, a garantia fornecida pelo modelo de *differential privacy*, de que a privacidade de um indivíduo não é quebrada quando participa num *dataset* visto que com a mesma probabilidade os mesmos resultados podem ser obtidos mesmo que ele não esteja presente no mesmo, é destruída. Um adversário conseguiria assim detetar a presença de um indivíduo num dos *datasets*, visto que as probabilidades dos mesmos resultados serem obtidos a partir de uma *query* a ambos os *datasets* são substancialmente diferentes (maiores que e^{ϵ}).

X	Pr[X]
-10	0.0183
-9	0.0203
-8	0.0224
-7	0.0248
-6	0.0274
-5	0.0303
-4	0.0335
-3	0.0370
-2	0.0409
-1	0.0452
0	0.0500
1	0.0452
2	0.0409
3	0.0370
4	0.0335
5	0.0303
6	0.0274
7	0.0248
8	0.0224
9	0.0203
10	0.0183

Tabela 7.: Função Densidade da Distribuição de Laplace de escala $\frac{1}{0.1}$

Para a mesma *query* mas sendo agora considerada de sobreposição, o ataque continua a ser válido. Neste caso, sendo os verdadeiros valores retornados antes da adição de ruído 10 e 24 para cada um dos *datasets*, para se obter o valor final de 10, a distribuição de Laplace tem que retornar para o primeiro caso o ruído de 0 e para o segundo o ruído de -14. As probabilidades de isto acontecer são agora, 0.0500 e 0.0123 respetivamente e o valor do seu rácio é agora $\frac{0.0500}{0.0123} = 4.0650$. Este valor é ainda maior que o anterior devido à diferença entre os verdadeiros valores antes da adição de ruído ser também maior.

Se forem utilizados valores mais baixos para o parâmetro de segurança *epsilon*, ocorre na mesma a possibilidade de ataque. As probabilidades obtidas pela função densidade da distribuição de Laplace calibrada agora com o parâmetro escala a $\frac{1}{0.01}$ estão apresentadas na tabela 8.

X	Pr[X]
-10	0.00452
-9	0.00456
-8	0.00461
-7	0.00466
-6	0.00470
-5	0.00475
-4	0.00480
-3	0.00485
-2	0.00490
-1	0.00495
0	0.00500
1	0.00495
2	0.00490
3	0.00485
4	0.00480
5	0.00475
6	0.00470
7	0.00466
8	0.00461
9	0.00456
10	0.00452

Tabela 8.: Função Densidade da Distribuição de Laplace de escala $\frac{1}{0.01}$

Por exemplo para *epsilon* de 0.01 e para a mesma *query* por inclusão, as probabilidades seriam agora de $\frac{0.00500}{0.00452} = 1.10619$, valor também maior e^{ϵ}, sendo o *epsilon* agora mais baixo, $e^{\epsilon} = 1.01005$

Com esta demonstração conseguimos assim mostrar que no nosso caso de estudo com a utilização conjunta dos modelos de *k-anonimity* e *differential privacy* podemos estar a obter um nível de segurança inferior ao que obteríamos se utilizássemos apenas um dos métodos. Esta conjunção dos modelos pode inclusive quebrar a garantia de privacidade oferecida pelo modelo de *differential privacy* que assegura que dados dois *datasets* que diferem apenas em um elemento e dada uma qualquer *query* efetuada a ambos os *datasets*, a probabilidade dessa *query* retornar um certo resultado num dos *datasets* não pode diferir da probabilidade de retornar o mesmo resultado no outro *dataset* por mais do que um fator multiplicativo de e^{ϵ}, sendo ϵ o parâmetro de segurança do modelo. Um adversário consegue deste modo diferenciar os resultados obtidos por

uma *query* quando é adicionado um elemento ao *dataset*. De acordo com o modelo de *differential privacy*, isto não pode acontecer.

Este ataque resulta da exploração do um dos parâmetros do modelo de *differential privacy*, a sensibilidade de uma *query*. Em concreto da ausência de calibração deste parâmetro para a correta utilização dos métodos no nosso contexto de estudo.

Esta necessidade de calibração do valor da sensibilidade não nos era inicialmente óbvia devido às *queries* de contagem terem naturalmente sensibilidade de 1 quando o modelo é utilizado sozinho. Contudo com a utilização simultânea dos dois métodos principais de obtenção de privacidade em sistemas de dados este já não é o caso e a calibração do parâmetro de sensibilidade com o valor de 1 neste contexto pode levar a um ataque surpreendente como foi visto nesta secção.

Surge assim o problema de identificar qual o valor correto para o parâmetro da sensibilidade das *queries* de contagem no nosso contexto de estudo. Este problema é debatido e explorado no capítulo 5.

4

ANÁLISE EXPERIMENTAL

Uma vez que no capítulo 5, onde serão exploradas respostas para o problema levantado no capítulo anterior, serão utilizados alguns resultados experimentais deste mesmo capítulo, este é apresentado primeiro.

De forma a melhor compreender os métodos de anonimização apresentados no estado da arte, em particular os modelos de *k-anonimity* e *differential privacy*, foi implementada uma aplicação em JAVA para testar e verificar os resultados da aplicação destes métodos em conjuntos de dados.

Em concreto com estes testes experimentais é pretendido verificar os efeitos que estes mecanismos têm num conjunto de dados, quer quanto utilizados individualmente quer quanto utilizados no nosso contexto de estudo, simultaneamente. Desta forma poderemos também melhor avaliar as aproximações efetuadas no capítulo 5.

Neste capítulo são primeiro detalhados na secção 4.1 os *datasets* utilizados para a realização dos testes. Na secção 4.2 são mostrados os detalhes das implementações dos modelos de *k-anonimity* e *differential privacy*. Por fim a secção 4.3 contém os testes propriamente ditos realizados a estes modelos.

4.1 CONJUNTOS DE DADOS

Para a realização dos testes experimentais foram utilizados dois conjuntos de dados diferentes.

O primeiro *dataset* é o mesmo que o utilizado pelos autores do sistema FUZZ [13], uma implementação do modelo de *differential privacy*. Este conjunto de dados é sintético e baseado nos dados de um censo realizado à população dos Estados Unidos da América retirados de [16].

Por motivos de simplicidade o *dataset* foi de seguida formatado sendo-lhe retirados alguns atributos ficando o *dataset* com 10000 registos tendo cada um os quatro atributos seguintes: idade, sexo, código-postal e salário. O atributo sensível deste *dataset* é o atributo salário. A tabela 9 mostra o intervalo de valores possíveis para cada atributo.

Para referência futura chamámos a este *dataset* de census-10k.

Atributo	Idade	Sexo	Código-Postal	Salário
Valor Mínimo	20	1	20000	10009
Valor Máximo	69	2	69000	300057

Tabela 9.: Intervalo de valores possíveis para cada atributo

Para contrastar com o *dataset* sintético apresentado acima foi também utilizado um *dataset* real originalmente obtido do *Adult dataset* proveniente de um censo à população norte-americana que pode ser obtido em http://ipums.org [17]. Foram utilizados em vários outros trabalhos na literatura, nomeadamente em [7], [18], [19], [20], [21] duas derivações deste dataset denominadas por OCC e SAL. Estes *datasets* possuem 600000 tuplos e cinco atributos onde o atributo sensível do dataset OCC é *occupation* enquanto que o atributo sensível do *dataset* SAL é *salary*.

No âmbito deste trabalho foram utilizados os primeiros 50000 tuplos do *dataset* SAL e todos os cinco atributos: idade, sexo, nível de educação, local de nascimento e salário. Os valores de cada atributo encontram-se discretizados em categorias sendo o número de valores diferentes possíveis para cada atributo muito menor do que no *dataset* anterior. O intervalo de valores possíveis para cada atributo é apresentado na tabela 10.

Para referência futura este *dataset* ficou com o nome de SAL-50k.

Atributo	Idade	Sexo	Nível de Educação	Local de Nascimento	Salário
Valor Mínimo	16	1	1	1	1
Valor Máximo	94	2	17	57	50

Tabela 10.: Intervalo de valores possíveis para cada atributo

4.2 k-ANONIMITY E DIFFERENTIAL PRIVACY

Nesta secção são detalhados em maior pormenor os modelos utilizados na prática para a realização dos testes experimentais assim como as bibliotecas externas utilizadas.

Apesar de existirem diversos outros algoritmos de anonimização de *k-anonimity*, sendo os mais conhecidos o Datafly [22], Incognito [23] e Anatomy [20], onde o ultimo é um algoritmo para atingir l-diversity o que implica *k-anonimity*, o algoritmo de anonimização escolhido foi o Mondrian [24] pois é o mais consensual na literatura sendo relativamente eficaz.

Para anonimizar os *datasets* com o algoritmo de Mondrian foi decidido utilizar a UTD Anonimization ToolBox [25] disponibilizada ao público pelo laboratório de privacidade e segurança de dados da universidade do Texas em Dallas, uma vez que não faria sentido efetuar a nossa própria implementação do algoritmo de Mondrian devido a já existirem várias e comprovadas implementações deste complexo algoritmo licenciadas para uso público e a implementação do mesmo se encontrar fora do contexto desta tese.

Esta *toolbox* é implementada em JAVA e funciona internamente com uma base de dados embebida, SQLite [26], de forma a facilitar a manipulação de *datasets* de tamanho elevado. Recebe como input dois ficheiros: o *dataset* a anonimizar e um ficheiro de configuração.

Os dados a anonimizar têm que se encontrar num ficheiro de texto onde cada linha corresponde a um registo em que os seus atributos estão separados por um ou mais caracteres delimitadores que podem ser definidos no ficheiro de configuração. A título de exemplo o delimitador por defeito é a virgula.

O ficheiro de configuração está estruturado em formato XML e pode ser editado para se alterarem os parâmetros da anonimização, o nome e caminho para os ficheiros que contêm os *datasets*, indicar os seus delimitadores, o número de atributos e quais deles são atributos sensíveis ou Quasi-Identifiers. Todos estes parâmetros podem também ser passados por argumento correndo a aplicação por linha de comandos. Em especial o nome e caminho para o ficheiro de configuração apenas pode ser indicado desta forma.

Após a anonimização, a *toolbox* escreve para um ficheiro de texto o resultado com a mesma formatação do *dataset* que recebeu como input sendo a única diferença naturalmente nos atributos Quasi-Identifiers que têm os seus valores específicos substituídos por valores generalizados sob a forma de intervalos.

Desta forma, a nossa aplicação de testes tem de seguida que ler este ficheiro resultante da anonimização dos dados através da *toolbox* e carregar os dados para as estruturas de dados correspondentes para posteriormente serem realizados os testes necessários.

Já a implementação do método de *differential privacy* foi efetuado de raiz, uma vez que apenas consiste em adicionar ao verdadeiro valor da resposta a uma *query*, uma quantidade de ruído aleatório obtido através de uma distribuição probabilística de Laplace. Esta quantidade de ruído depende do valor do *epsilon* definido para o modelo de segurança e da sensibilidade de cada *query* como já foi visto anteriormente. Neste contexto, temos assim de retirar um valor aleatório de uma distribuição de Laplace centrada em 0 e com escala Δ_q / ϵ, sendo Δ_q a sensibilidade da *query* em questão e ϵ o valor do *epsilon* definido para o modelo de segurança.

Para isto foi utilizada a seguinte fórmula.

Seja U uma variável aleatória retirada de uma distribuição uniforme de intervalo $]-\frac{1}{2}, \frac{1}{2}]$, a variável aleatória

$$X = \mu - b \cdot sgn(U) \cdot \ln(1 - 2 \cdot |U|)$$

tem uma distribuição de Laplace com os parâmetros μ e b, onde μ é o parâmetro da localização e b o parâmetro escala e onde *sgn* é a função sinal que retorna o sinal de um número real. Ou seja, sendo x um número real:

$$sng(x) = \begin{cases} -1 & \text{s } x < 0 \\ 0 & \text{se } x = 0 \\ 1 & \text{se } x > 0 \end{cases}$$

Com esta fórmula conseguimos assim retirar valores aleatórios de uma distribuição de Laplace com os parâmetros à nossa escolha e deste modo realizar *queries* aos *datasets* de acordo com o método de *differential privacy*.

4.3 UTILIDADE DOS DADOS

Este teste tem como objetivo medir a utilidade dos dados resultantes da aplicação dos métodos de *k-anonimity* e *differential privacy* quer quando usados sozinhos, quer quando usados em simultâneo. Tentaremos com isto obter uma aproximação à quantidade de ruído acrescentada aos conjuntos de dados por estes métodos.

Para isso são efetuadas séries de *queries* aos conjuntos de dados antes e depois da aplicação dos métodos e comparados os resultados para desta forma calcular o erro médio introduzido pelos métodos. Este teste à utilidade dos dados produzidos é baseado no teste utilizado pelos autores do método de m-invariance em [18].

As *queries* usadas para os testes são *queries* de agregação devido a serem as mais utilizadas para tarefas de *datamining* e análises a grandes quantidades de dados. Mais concretamente são utilizadas *queries* de contagem.

Cada *query* impõe restrições sobre todos os atributos do *dataset*. Isto inclui todos os atributos Quasi-Identifiers e o atributo sensível. No caso do conjunto de dados census-10k isto significa três atributos Quasi-Identifiers e um atributo sensível e no caso do conjunto de dados SAL-50k, quatro atributos Quasi-Identifiers e um atributo sensível. A *query* retorna a contagem de todos os indivíduos que obedeçam às restrições da *query*.

Mais formalmente cada *query* é da forma:

SELECT COUNT (*) FROM D WHERE
$Cond(A_1^{qi})$ AND ... AND $Cond(A_n^{qi})$ AND $Cond(A^s)$

Onde D é um dos dois *datasets* de teste, A_1^{qi}, ..., A_n^{qi} são os atributos Quasi-Identifiers do respetivo *dataset* e A^s o atributo sensível. Para cada atributo A, $cond(A)$ impõe uma restrição sobre esse atributo na forma de um intervalo I. I é gerado aleatoriamente e está contido no domínio de A. O tamanho do intervalo I não é aleatório mas sim dependente do parâmetro RR. O tamanho de I é assim dado por $|A| * RR$, onde $|A|$ é o domínio do atributo A e RR o parâmetro *RangeRatio*. Este parâmetro é dado em percentagem e é igual para todos os atributos da mesma *query*. A utilização deste parâmetro é necessária pois sem a sua inclusão os resultados variariam muito. A não utilização deste parâmetro iria também frequentemente originar intervalos muito curtos fazendo com que o resultado da contagem retornada pela *query* fosse zero ou perto disso, resultado este que não oferece grande utilidade à posterior análise e comparação de resultados. Quanto maior for o valor do parâmetro RR mais abrangente será a *query* em relação ao conjunto de dados e de forma geral maior será o resultado da contagem retornada pela *query*.

Desta forma, uma *query* deste tipo é por exemplo: Quantos indivíduos no *dataset* têm idade contida no intervalo [36,66] e o seu código-postal está entre [27000,57000] e o seu sexo esta entre [2,2] e o seu salário entre [54000,150000].

Assim, são efetuados testes a quatro cenários diferentes para comparação dos resultados obtidos.

No primeiro cenário são feitas as *queries* aos conjuntos de dados originais antes de qualquer anonimização. Este cenário servirá de ponto base sobre o qual os restantes casos serão comparados.

O segundo caso consiste em realizar as *queries* aos conjuntos de dados já anonimizados com o algoritmo de anonimização de Mondrian. Para isto e como já foi detalhado na secção anterior é utilizada a UTD Anonimization ToolBox. Este caso será utilizado para estimar a quantidade de erro introduzida nos dados pelo algoritmo de anonimização.

O terceiro cenário tem como objetivo ser utilizado para tentar medir a quantidade de erro obtido através da utilização do método de *differential privacy*. Assim, neste cenário as *queries* são realizadas aos conjuntos de dados originais mas de forma a estarem de acordo com o método de *differential privacy*. Para isto, aos verdadeiros valores das respostas às *queries* é adicionado ruído aleatório proveniente de uma distribuição probabilística de Laplace.

Por fim, o quarto cenário pretende estimar a utilidade dos dados resultantes da aplicação dos métodos de *k-anonimity* e *differential privacy* em simultâneo. Este caso na prática reflete a

conjunção dos dois casos anteriores. São efetuadas *queries* seguindo o modelo de *differential privacy* diretamente sobre o *dataset* já anonimizado com o algoritmo de anonimização de Mondrian.

Para cada um dos três últimos cenários, o cálculo do erro de uma *query* depende do valor da contagem obtido nesse cenário e do valor da contagem obtido no primeiro caso apresentado, o ponto base de comparação que apresenta o verdadeiro valor das respostas sem qualquer erro.

O erro de uma *query* em qualquer um dos 3 últimos cenários é assim dado pela seguinte fórmula:

$$\frac{|countReal - countEstimado|}{countReal}$$

onde *countReal* é o valor da contagem obtido pela *query* no primeiro cenário, e *countEstimado* é o valor da contagem obtido pela *query* num dos três cenários seguintes dependendo qual dos casos está a ser testado.

Desta forma o erro é medido em percentagem. Se por exemplo o valor do *countReal* for de 1000 e o valor do *countEstimado* for de 800, o resultado da fórmula dá o valor de 0,2. Neste caso dizemos que o valor retornado por esta *query* em concreto apresenta um erro de 20% em relação ao *countReal*. As *queries* em que o valor do *countReal* é igual a zero são ignoradas. Esta ocorrência é rara e apenas acontece esporadicamente para os valores de RR mais baixos.

Assim para a realização dos testes, a aplicação JAVA lê de ficheiro duas versões do mesmo conjunto de dados. Uma versão original, tal como as apresentadas na secção 4.1, e uma versão anonimizada com o algoritmo de Mondrian obtida através da UTD Anonimization Toolbox. A aplicação carrega as duas versões para duas estruturas de dados diferentes: uma referente ao *dataset* original com todos os atributos bem definidos e explícitos, e outra referente ao *dataset* anonimizado onde todos os atributos encontram-se generalizados para intervalos.

Para a obtenção dos resultados finais foram efetuadas 5000 *queries* e calculado o erro médio de forma a testar extensivamente os métodos e obter uma melhor visão do verdadeiro erro evitando possíveis valores *outliers*.

Em cada uma das 5000 iterações é gerada uma *query* aleatória de acordo com os parâmetros definidos. Esta *query* é aplicada em todos os cenários apresentados anteriormente e o seu resultado é registado. De seguida o erro desta *query* aleatória é calculado para cada um dos cenários usando a fórmula anteriormente mostrada. Estes erros são guardados e no final de todas as iterações é calculado o erro médio de cada cenário.

Resultados

De seguida são apresentados os resultados obtidos por estes testes. Os resultados são apresentados em gráficos e estão divididos em duas partes principais. Na primeira os testes são realizados sobre o *dataset* census-10k apresentado na secção 4.1. Encontra-se subdividida em três partes referentes aos três cenários de estudo. De forma a contrastar com o conjunto de dados census-10k, são apresentados na segunda parte os resultados dos testes agora realizados sobre o *dataset* SAl-50k também apresentado na secção 4.1.

Conjunto de dados Census-10k

Nas três subsecções seguintes são mostrados os testes realizados aos três cenários de teste, *k-anonimity*, *differential privacy* e os dois modelos em simultâneo, efetuados sobre o *dataset* sintético census-10k.

K-anonimity

Nesta primeira parte mostram-se os resultados correspondentes aos testes efetuados ao algoritmo de anonimização de Mondrian. Estes resultados provêem da comparação dos valores de contagem obtidos através do primeiro e segundo cenários detalhados anteriormente. São apresentados em dois gráficos. No primeiro gráfico o valor de RR é fixo a 0.6 e faz-se variar o parâmetro k. O segundo gráfico mostra o caso contrário, o valor do parâmetro k é fixo a 10 e faz-se variar o valor do parâmetro *RangeRatio*.

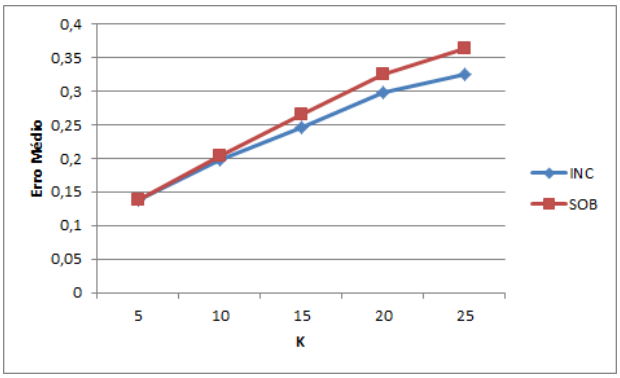


Figura 2.: Erro médio relativo k-anonimity vs. k (RR=0,6)

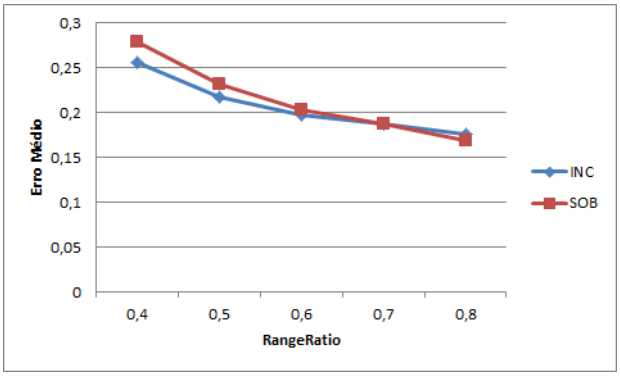


Figura 3.: Erro médio relativo k-anonimity vs. RR (k=10)

Da observação destes dados podemos verificar o impacto que a variação dos parâmetros de k e *RangeRatio* têm no erro médio relativo no algoritmo de anonimização de Mondrian. Os resultados obtidos confirmam aquilo que de certa forma já se esperava.

Como se pode ver na figura 2, o aumento do k leva também a um aumento do erro médio, variando entre cerca de 15% e 35% para k de 5 e 25 respetivamente. Isto porque um valor de k mais elevado obriga a que as classes de equivalência sejam também maiores, no mínimo com k elementos, o que requer mais generalização dos dados fazendo com que as *queries* tenham no geral um erro maior.

A figura 3 mostra o erro médio relativo do algoritmo de anonimização de Mondrian em função do parâmetro RR com o valor de k fixo a 10. Podemos observar que o aumento do valor do parâmetro *RangeRatio* origina uma diminuição do erro. Este resultado era também expectável pois valores altos do parâmetro RR tornam as *queries* mais abrangentes retornando valores de contagem maiores onde no geral este tipo de análise agregada é mais eficaz.

Em ambos estes casos a diferença no erro médio relativo entre as *queries* de inclusão e sobreposição não é significativa. A maior disparidade de valores entre os dois tipos de *query* ocorre para k igual a 25 na primeira imagem e para RR igual a 0.4 na segunda, casos onde o valor do erro médio é o mais alto.

Differential Privacy

De seguida são apresentados os resultados dos testes realizados ao método de *differential privacy*. Neste caso de estudo os resultados provêem agora da comparação dos valores de contagem obtidos através do primeiro e terceiro cenários referidos atrás. De forma similar aos testes sobre o modelo de *k-anonimity* os resultados são apresentados em dois gráficos. O primeiro gráfico

mostra a variação do parâmetro de segurança *epsilon* dado pelo símbolo ϵ com o valor de RR fixo a 0.6. No segundo gráfico o valor de ϵ é fixo a 0.01 e faz-se variar o valor do parâmetro *RangeRatio*.

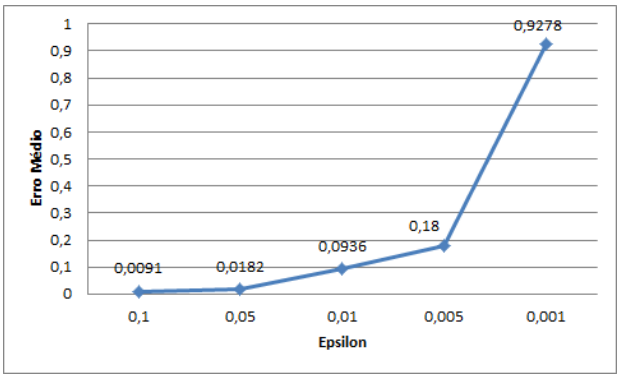


Figura 4.: Erro médio relativo differential privacy vs. ϵ (RR=0,6)

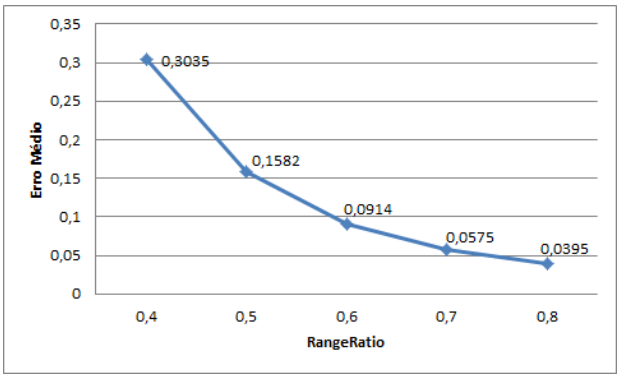


Figura 5.: Erro médio relativo differential privacy vs. RR (ϵ=0,01)

Como já seria previsto a partir do funcionamento do modelo de *differential privacy* pode-se ver na figura 4 que a diminuição do valor do parâmetro de segurança ϵ gera um aumento do erro médio relativo das *queries*.

De forma particular é possível notar que o fator de aumento do erro é praticamente constante com a diminuição do valor de ϵ. Isto é, reduzindo por exemplo o valor de ϵ dez vezes leva também a um aumento de dez vezes no erro médio obtido pelas *queries*. Isto acontece para todos os valores do gráfico. Por exemplo, pode-se notar que a alteração do valor de ϵ de 0.1 para 0.01 corresponde a um aumento no erro médio de 0.0091 para 0.0936. Isto acontece porque neste modelo de *differential privacy*, o ruído adicionado às verdadeiras respostas é obtido de forma

aleatória através de uma distribuição probabilística de Laplace com valor médio de 0 e escala Δ_q / ϵ. Como a sensibilidade das *queries* de contagem utilizadas nestes testes é sempre 1, o ruído adicionado varia apenas com o valor de *epsilon*.

Por esta mesma razão podemos dizer que o erro acrescentado pelo modelo de *differential privacy* é de certa forma estático, ou seja, é independente do *dataset* a que é aplicado, das *queries* efetuadas (desde que não alterem a sensibilidade) e do resultado das mesmas. Depende apenas dos parâmetros sensibilidade e *epsilon*. Uma vez que na figura 5 o parâmetro de segurança ϵ é fixo a 0.01, nota-se um decréscimo do erro médio com o aumento do parâmetro *RangeRatio*. Por um lado, o aumento do valor de RR faz com que o valor das contagens retornadas pelas *queries* seja mais alto. Por outro lado, o ruído absoluto acrescentado não varia, pois os valores de *epsilon* e da sensibilidade não sofrem alterações neste teste. Assim, o erro médio relativo vai decrescer em percentagem como se pode observar no gráfico.

K-anonimity + Differential Privacy

Por fim são apresentados os resultados referentes aos testes realizados à utilização simultânea dos modelos de *k-anonimity*, em concreto o algoritmo de anonimização de Mondrian, e *differential privacy*. A cada uma das 5000 *queries* feitas sobre o conjunto de dados já anonimizado é ainda acrescentado ruído de acordo com o modelo de *differential privacy*. Os dados obtidos estão também compilados em dois gráficos. Em ambos os gráficos o parâmetro *RangeRatio* tem o seu valor fixo a 0.6. No primeiro gráfico o valor do parâmetro de segurança do modelo de *differential privacy*, ϵ, é fixo a 0.01 sendo mostrada a variação do parâmetro do modelo de *k-anonimity*, k. A segunda imagem mostra a inversão dos parâmetros, k é fixo com um valor de 10 e ϵ varia entre 0.1 e 0.001.

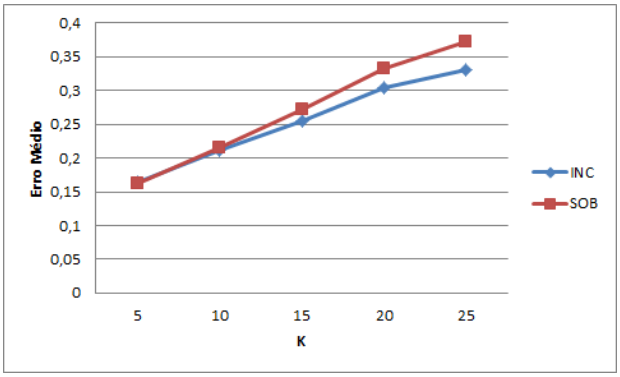


Figura 6.: Erro médio relativo K-Anon + DP vs. k (RR=0,6, ϵ =0,01)

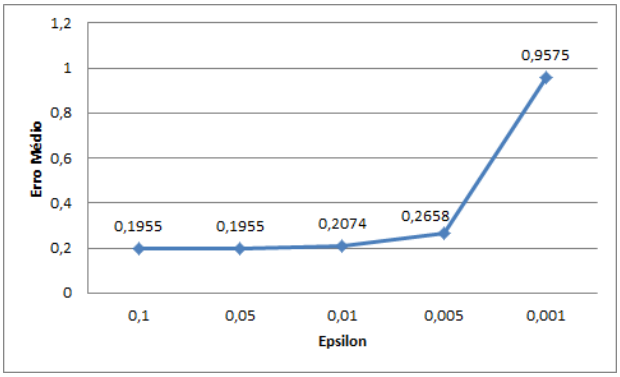


Figura 7.: Erro médio relativo K-Anon + DP vs. ϵ (RR=0,6, k =10)

Apesar de se mostrar agora o resultado obtido da aplicação dos dois métodos em conjunto, os traços gerais do comportamento de ambos os algoritmos que já tinham sido vistos nas imagens anteriores são também evidentes nestes gráficos. Observa-se que o aumento do parâmetro k e a diminuição do parâmetro ϵ levam também a um aumento do erro médio relativo.

Na figura 6 a diferença entre *queries* de inclusão e sobreposição só começa a ser notável para os valores de k mais altos. Uma vez que neste caso o valor de ϵ se encontra fixo, o gráfico é principalmente influenciado pela variação de k, acabando a curva por ser relativamente semelhante à do gráfico 2 onde é mostrado o erro produzido pelo algoritmo de Mondrian sozinho. Diferem no facto desta curva ter os seus valores ligeiramente superiores devido à presença também do modelo de *differential privacy* com valor de *epsilon* de 0.01.

Já na figura 7 estando o valor de k fixo a 10, a curva demonstra maioritariamente as consequências da evolução do valor de ϵ. Por esta razão a figura apresenta apenas os valores das *queries* de inclusão, uma vez que a diferença entre *queries* de inclusão e sobreposição para k igual a 10 é praticamente nula como se pode ver na figura 2.

De forma geral o erro médio obtido pela utilização combinada dos modelos de *k-anonimity* e *differential privacy* é, como seria expectável, sempre maior do que o erro médio produzido por apenas um dos modelos, para os mesmos parâmetros de segurança. Nota-se também que o erro obtido pela utilização dos dois modelos em conjunto é inferior à soma dos erros obtidos pela utilização dos modelos individualmente. Por exemplo na figura 7 o erro obtido devido ao uso das duas técnicas em conjunto para k = 10 e ϵ = 0.005 é de 0.2658 enquanto que na figura 2 podemos verificar que o erro obtido apenas do modelo de *k-anonimity* também com k = 10 é de 0.2 e na figura 4 pode ser visto que o erro da proveniente da utilização individual do modelo de *differential privacy* para ϵ = 0.005 é de 0.18. Ora a soma dos dois erros obtidos a partir dos modelos quando usados individualmente, 0.20 e 0.18, dá claramente um valor superior ao erro

obtido da utilização dos modelos em simultâneo, 0.2658. Isto acontece também para os restantes valores.

Uma possível razão para isto ocorrer é a de que os dois erros relativos a cada um dos modelos têm de certa forma origens diferentes. No caso do modelo de *k-anonimity* o valor da contagem retornado pelas *queries* ou é sempre maior ou sempre menor que o verdadeiro valor dependendo se a *query* for de inclusão ou de sobreposição tal como já foi mostrado na secção 3.1.

No caso do modelo de *differential privacy* o erro provém da adição de ruído obtido através de uma distribuição probabilística de Laplace às verdadeiras respostas de uma *query* e como tal o valor da contagem pode agora ser ou maior ou menor que o verdadeiro valor.

Assim, na utilização dos dois modelos em conjunto os erros proveniente de ambos os modelos podem por vezes anularem-se um ao outro acabando o erro final por ser desta maneira menor que a soma dos erros provenientes da utilização individual dos modelos.

Conjunto de dados SAL-50k

São agora apresentados os resultados obtidos da realização dos testes sobre o *dataset* real SAL-50k apresentado na secção 4.1. Os testes efetuados são exatamente os mesmos dos realizados ao *dataset* census-10k.

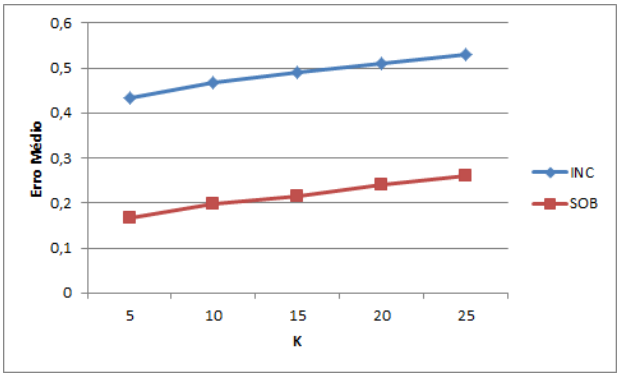


Figura 8.: Erro médio relativo *k*-anonimity vs. *k* (*RR*=0,6)

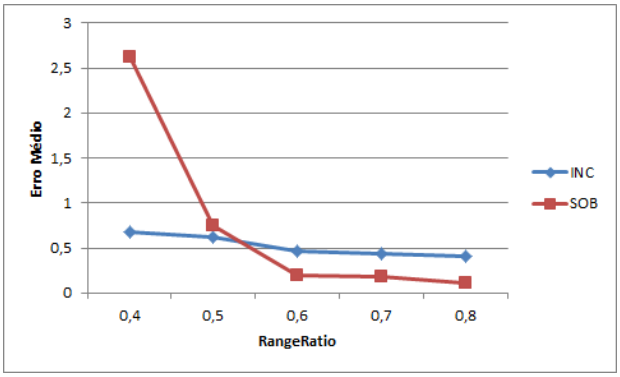


Figura 9.: Erro médio relativo k-anonimity vs. RR (k=10)

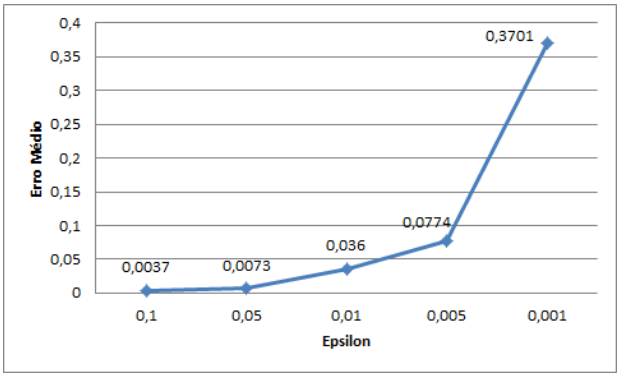


Figura 10.: Erro médio relativo differential privacy vs. ϵ (RR=0,6)

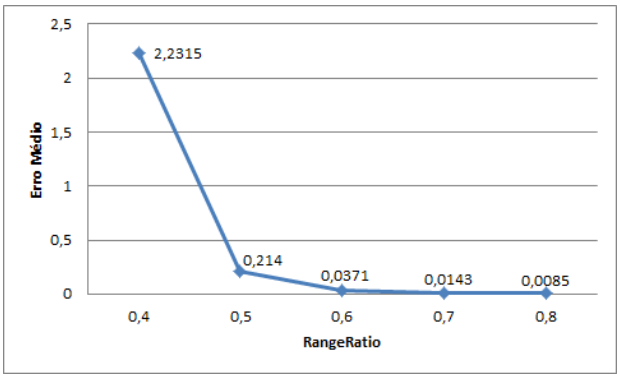


Figura 11.: Erro médio relativo differential privacy vs. RR (ϵ=0,01)

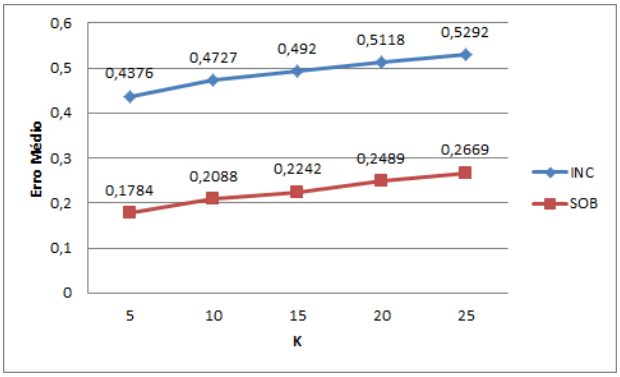


Figura 12.: Erro médio relativo K-Anon + DP vs. k (RR=0,6, ϵ =0,01)

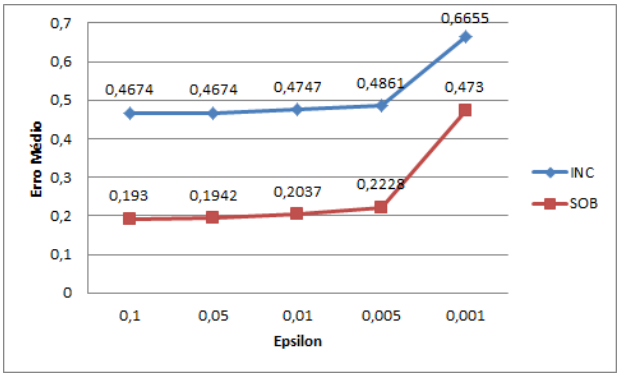


Figura 13.: Erro médio relativo K-Anon + DP vs. ϵ (RR=0,6, k =10)

De forma geral, a variação dos gráficos utilizando este conjunto de dados segue a linha coerente de resultados obtidos da utilização do *dataset* anterior. O aumento do valor do parâmetro k e a diminuição do valor do parâmetro ϵ origina um aumento do erro médio relativo, assim como valores mais altos do parâmetro RR levam a um valor de erro médio mais baixo.

Por outro lado, a diferença do erro médio entre as *queries* de inclusão e sobreposição é bastante significativa neste *dataset* como pode ser observado nas figuras 8 e 9.

Uma vez que este conjunto de dados SAL-50k é cinco vezes maior que o census-10k, os valores dos erros dos gráficos relativos ao modelo de *differential privacy*, figuras 10 e 11, são mais baixos que os do conjunto de dados anterior. Como os parâmetros utilizados nos testes são os mesmos, a quantidade de ruído adicionada pelo modelo é a mesma, mas como os valores retornados pelas contagens são maiores devido ao *dataset* ser bastante maior, os valores dos erros, que são apresentados em percentagem, diminuem. Devido ao *dataset* ser maior, a variação do

parâmetro RR tem também um efeito mais acentuado nos resultados, principalmente no modelo de *differential privacy*.

Os resultados da utilização dos dois modelos em conjunto, figuras 12 e 13, são em tudo semelhantes aos do conjunto de dados census-10k com a exceção da disparidade existente entre os erros das *queries* de inclusão e sobreposição.

Da realização destes testes podemos concluir que os resultados da aplicação dos métodos de *k-anonimity* e *differential privacy* usados de forma individual ou conjunta dependem largamente dos parâmetros de segurança escolhidos para a sua aplicação. Desta forma, a correta calibração destes modelos é fundamental para a obtenção de resultados satisfatórios. Que por um lado se atinjam níveis aceitáveis de segurança e que por outro se mantenha a utilidade dos dados garantindo que estes não sofrem grande deterioração.

5

CALIBRAÇÃO DA SENSIBILIDADE

Este capítulo surge no seguimento do capítulo 3 onde é apresentado o caso de estudo no contexto desta dissertação e o problema encontrado que pode dar origem a um ataque surpreendente que leva à quebra da garantia de segurança oferecida pelo modelo de *differential privacy*.

Ficou mostrado que utilizar o parâmetro da sensibilidade com o valor de 1 como seria normal para *queries* de contagem é altamente no contexto da utilização simultânea dos dois mecanismos. Resta assim a pergunta de qual será de facto o valor apropriado para a correta calibração do parâmetro sensibilidade do método de *differential privacy*.

A sensibilidade de uma *query*, dada por Δ_q, é a alteração máxima que a reposta a essa mesma *query* pode ter quando efetuada a dois *datasets* que diferem em apenas um elemento. O parâmetro é utilizado pelo modelo de *differential privacy* e tem influencia direta na quantidade de ruído adicionada ás verdadeiras respostas de uma *query*. Juntamente com o parâmetro de segurança *epsilon* são usados para formar uma distribuição probabilística de Laplace com valor médio de 0 e escala Δ_q / ϵ de onde é retirado o ruído aleatório a adicionar às respostas. Assim, quanto maior for o valor da sensibilidade e menor o valor de *epsilon*, maior será o ruído acrescentado e menor utilidade terão os resultados obtidos.

No nosso contexto o valor da sensibilidade irá depender essencialmente do pré-processamento efetuado pelo algoritmo de anonimização. Ao longo deste capítulo são assim efetuados testes experimentais de forma a melhor compreender que efeitos este pré-processamento por via do algoritmo de anonimização de Mondrian pode ter no parâmetro da sensibilidade.

No entanto, a imprevisibilidade dos resultados obtidos por uma anonimização com o algoritmo de Mondrian, nomeadamente das diferenças que ocorrem entre duas anonimizações onde antes de uma delas é removido ou adicionado um elemento ao conjunto de dados, coloca uma questao interessante, à qual tentamos responder com três abordagens distintas. Uma aproximação puramente teórica, uma aproximação puramente experimental e uma aproximação híbrida.

5.1 APROXIMAÇÃO TEÓRICA

Nesta primeira secção deste capítulo o problema levantado nas secções anteriores é explorado de forma puramente teórica.

Uma vez que existe grande incerteza relativamente às diferenças que podem ocorrer entre duas anonimizações sucessivas em que uma delas é efetuada após a adição ou remoção de um indivíduo do *dataset* original, as ideias estudadas nesta secção são hipotéticas. Isto é, são baseadas nos dados teóricos concretos sobre o funcionamento dos modelos no contexto do estudo do problema abordado nesta dissertação mas são assumidas condições hipotéticas que um algoritmo de anonimização teria de ter de forma a se poderem retirar conclusões aceitáveis sobre um correto valor do parâmetro da sensibilidade.

Supomos assim que existe um algoritmo de anonimização de *k-anonimity* que garante certas condições. Nomeadamente garante que dados dois *datasets* que diferem apenas num elemento, o resultado de anonimizar cada um destes *datasets* apenas pode diferir em uma classe de equivalência. Isto é, entre os resultados destas duas anonimizações, todas as classes de equivalência são iguais exceto uma. Este caso é na verdade o melhor que se pode considerar. Pois a adição ou remoção de um elemento pode causar sempre no mínimo a alteração de uma classe.

Por exemplo, se for adicionado ao conjunto de dados um indivíduo *outlier*, ou seja um indivíduo cujos valores dos atributos se encontrem distantes dos restantes elementos do *dataset*, não existirá nenhuma classe de equivalência onde este elemento se possa simplesmente acrescentar. Teria neste caso que ser criada uma nova classe de equivalência onde este novo elemento se poderia inserir.

O caso contrário ocorre quando é removido um elemento que formaria uma classe de equivalência com o número de elementos mínimo permitido pela restrição de *k-anonimity*. Este foi o caso retratado no capítulo anterior e obriga também a que a classe de equivalência onde o elemento se encontrava desapareça.

Assim, supondo que existe um algoritmo de anonimização de *k-anonimity* que assegure esta condição, o parâmetro da sensibilidade teria que ser calibrado a um valor igual ao tamanho da classe alterada.

Desta forma o ataque demonstrado no capítulo anterior já não teria efeito. Mesmo que as duas *queries* efetuadas perguntem exatamente quantos indivíduos se encontram numa dada classe de equivalência e num dos casos a classe de equivalência desapareça, a diferença de resultados entre as duas *queries* vai agora estar protegida pelo valor de sensibilidade atualizado.

No entanto, a definição de sensibilidade de uma *query* apresentada pela autora do modelo de *differential privacy*, Cynthia Dwork, indica que o valor da sensibilidade é a alteração *máxima* que a resposta a uma *query* pode ter quando aplicada a dois conjuntos de dados que diferem apenas em um elemento. Temos desta maneira que ter em conta o pior caso, isto é, o tamanho da classe de equivalência que aparece ou desaparece com cujo valor estamos a calibrar o parâmetro sensibilidade ter o maior número de elementos possível.

De forma geral os algoritmos de anonimização não impõem restrições ao tamanho máximo que uma classe de equivalência pode tomar, apenas restringem o tamanho mínimo que é dado pelo parâmetro k. O seu funcionamento é no entanto o de procurar que as classes de equivalência tenham uma quantidade de elementos o mais baixa possível de forma a maximizar a utilidade dos dados, reduzindo a generalização necessária minimizando assim o ruído adicionado.

O algoritmo de Mondrian apresenta no entanto no seu artigo original um limite teórico para a dimensão das classes de equivalência.Este limite é apresentado na seguinte fórmula:

$$2 \cdot d \cdot (k-1) + m$$

O parâmetro d representa a dimensão do *dataset* que é o número total de atributos a anonimizar, k representa naturalmente o parâmetro de segurança algoritmo e m indica o número de elementos repetidos, com exatamente os mesmos valores dos atributos contidos nesta classe de equivalência.

No entanto esta fórmula acaba por não impor um limite superior fixo para o tamanho das classes estando este sempre dependente da quantidade de elementos repetidos no conjunto de dados. No pior caso todos os elementos poderiam ser repetidos sendo o tamanho máximo de uma classe de equivalência igual ao tamanho do *dataset*. Este caso à partida não seria muito útil de forma a se retirarem conclusões sobre a calibração da sensibilidade.

Desta forma acrescentamos agora outra condição ao algoritmo de anonimização hipotético que temos vindo a discutir nesta secção. Assumimos assim que este algoritmo de anonimização para além de garantir que dados dois *datasets* que diferem apenas num elemento, o resultado de anonimizar cada um destes *datasets* apenas pode diferir numa classe de equivalência, garante também que um conjunto de dados não pode ter elementos repetidos, assunção relativamente fácil de efetuar imaginando um qualquer *dataset* real em que existe apenas um registo para cada indivíduo. Assim, o valor do parâmetro sensibilidade a ser utilizado no nosso caso de estudo, que implica o uso dos modelos de *differential privacy* e *k-anonimity* em conjunto, seria calibrado a $2 \cdot d \cdot (k-1)$.

Obteríamos assim um valor razoável para o parâmetro da sensibilidade que por um lado seria seguro contra o ataque mencionado no capítulo anterior e que por outro produziria resultados

com utilidade aceitável não sendo a quantidade de ruído acrescentado exagerado. Por exemplo, para o *dataset* census-10k,com 3 atributos Quasi-Identifiers, e para uma utilização usual do parâmetro de k, 10, teríamos as classes de equivalência limitadas a 54 elementos. O parâmetro da sensibilidade seria desta forma calibrado com o valor de 54.

Já tendo sido verificado este caso, pode ainda ser relaxado ligeiramente. Assumimos assim que apesar de o tamanho máximo de cada classe de equivalência ser neste caso 54, as classes raramente atingem esse valor, tendo por norma cerca de metade do tamanho máximo. Na verdade como será visto na próxima secção, esta assunção não esta longe do que realmente acontece para este *dataset*. Desta forma o parâmetro da sensibilidade seria agora de 27.

De forma a perceber que impacto a calibração da sensibilidade a estes valores teria de fato nos resultados obtidos, o erro médio que se obtém para *queries* de contagem neste contexto é mostrado na seguinte tabela. O valor do erro médio é apresentado em percentagem tal como no capítulo de análise experimental.

Estes valores foram calculados de acordo com o modelo de testes utilizado no capítulo anterior. Em concreto é utilizado um valor de k de 10 para o algoritmo de *k-anonimity*, e um valor de *epsilon* de 0.1 para o modelo de *differential privacy*.

Sensibilidade	Erro Médio
54	0.528
27	0.309

Tabela 11.: Erro médio para aproximação teórica da sensibilidade

O resultado obtido é considerável calibrando o valor da sensibilidade a 54, mas pode ser considerado aceitável em alguns contextos, principalmente se utilizado para um *datasets* de tamanho superior a este onde seria expectável o valor do erro médio diminuir. Utilizando o valor de 27 o ruído acrescentado ao *dataset* é bastante bom, não resultando num valor de erro médio muito elevado, em linha com os resultados do capítulo anterior de analise experimental.

Esta aproximação puramente teórica é bastante otimista nos pressupostos que supõe ao algoritmo de anonimização. Assume por um lado que entre duas anonimizações a dois *dataset* que diferem apenas em um elemento só existe uma classe de equivalência de diferença. Esta diferença na verdade pode ser bastante maior como será visto na próxima secção. Por outro lado assume também que um *dataset* não pode ter elementos repetidos, assunção esta mais fácil de aceitar. É também verificado o caso em que se assume que apesar de ser apresentado um limite para o tamanho de cada classe, este limite não é atingido sendo considerado um valor metade deste. Com

estas suposições obteríamos resultados bons relativamente às próximas aproximações e iríamos neste caso ter também uma utilização dos modelos tal como o nosso contexto de estudo indica evitando por completo o ataque demonstrado no capítulo 3.

5.2 APROXIMAÇÃO HÍBRIDA

Nesta secção é apresentada uma aproximação híbrida ao valor do parâmetro da sensibilidade. As ideias exploradas nesta secção têm por base conceitos teóricos sobre o comportamento dos algoritmos que são testados e validados experimentalmente. As principais assunções realizadas na secção anterior foram ao tamanho máximo de cada classe de equivalência e ao número de classes alteradas entre duas anonimizações em que os seus *datasets* originais variam apenas em um elemento. Ambos estes parâmetros são testados experimentalmente e feita assim uma aproximação ao valor da sensibilidade baseado nestas mesmas experiências.

A ideia explorada é a de que de forma a garantir que o ataque apresentado no capítulo anterior não possa ocorrer, o parâmetro da sensibilidade tem que ser calibrado ao valor resultante da multiplicação do tamanho máximo que uma classe pode tomar com o número máximo de classes alteradas, isto é que podem desaparecer ou aparecer, entre duas anonimizações a dois *datasets* que diferem em um elemento. Com a calibração do parâmetro da sensibilidade a este valor a garantia dada pelo modelo de *differential privacy* é mantida válida pois a diferença que poderia ocorrer entre duas *queries* será cobrida por este valor.

Todos os pormenores relativos aos testes experimentais são detalhados no capítulo 4. Os testes realizados para determinar o tamanho máximo de uma classe de equivalência foram aplicados ao conjunto de dados sintético, census-10k, e ao conjunto de dados real, SAL-50k. Cada *dataset* possui 3 e 4 atributos Quasi-Indentifiers respetivamente. Ambos os *datasets* foram anonimizados pelo algoritmo de *k-anonimity* de Mondrian com um valor de k igual a 10.

Para o *dataset* census-10k foram encontradas 655 classes de equivalência com 24 tamanhos distintos. O tamanho máximo encontrado foi de 35 elementos. A distribuição completa das classes pode ser vista na seguinte figura.

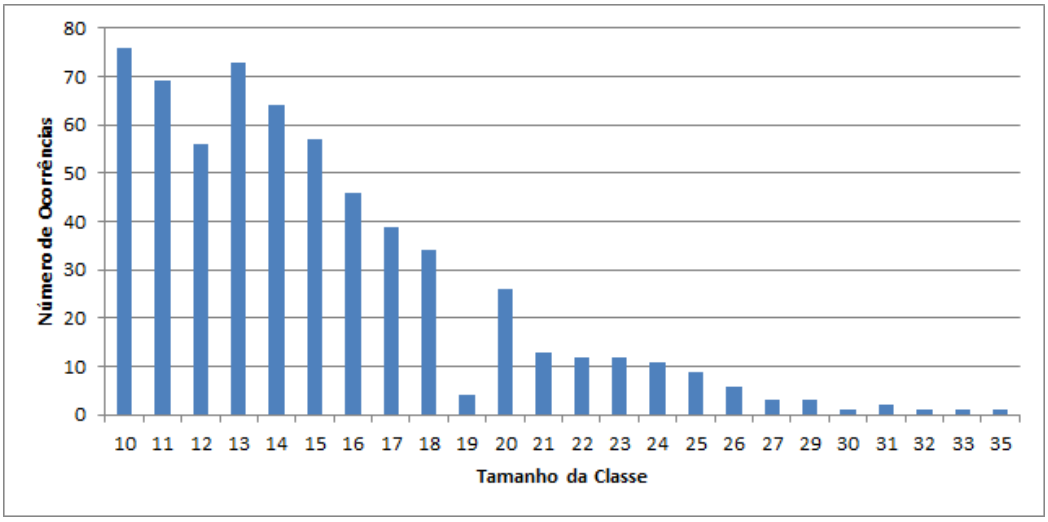


Figura 14.: Distribuição de classes da anonimização de census-10k

Este resultado vem em linha com o funcionamento dos algoritmos de anonimização que tentam que o tamanho das classes seja o mais próximo do mínimo possível de forma a minimizar o ruído acrescentado. Nota-se assim no gráfico que a maior densidade de classes ocorre para tamanhos mais baixos entre 10 e 20. As classes com tamanhos maiores não implicam ainda a existência de elementos repetidos sendo 35 menor que $2 \cdot d \cdot (k-1) + m$, com d igual a 3, k igual a 10 e m igual a 0.

Para o conjunto de dados real SAL-50k que contém 50000 indivíduos em vez dos 10000 do *dataset* anterior os resultados são apresentados no seguinte gráfico.

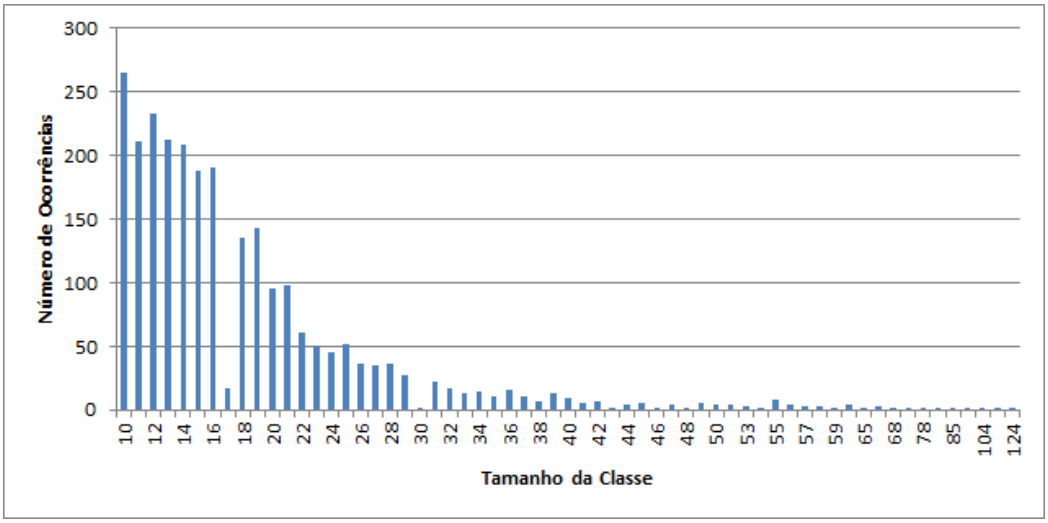


Figura 15.: Distribuição de classes da anonimização de SAL-50k

A anonimização do *dataset* SAL-50k resulta em 2735 classes de equivalência com 61 tamanhos de classe diferentes. A classe de equivalência com maior número de elementos encontrada tem 124 elementos. Neste caso as classes de equivalência encontram-se muito mais dispersas estando a maior concentração nos tamanhos mais baixos. As maiores classes obtidas por esta anonimização contêm agora elementos repetidos. Apesar deste *dataset* ser real, a existência de elementos repetidos deve-se ao fato de o *dataset* se encontrar discretizado. É assim possível a existência de elementos com a mesma idade, sexo, nível de educação e local de nascimento. O conteúdo das duas maiores classes obtidas por esta anonimização com a indicação dos elementos repetidos, classe com 124 elementos e classe com 116, são mostradas em anexo nas tabelas 16 e 17. Devido à quantidade de elementos repetidos presentes nestas classes e comparando esse valor com a fórmula para o tamanho máximo de uma classe de equivalência do algoritmo de Mondrian estas classes ainda poderiam albergar mais elementos não estando o seu tamanho máximo ainda atingido.

De seguida foi realizado um teste para verificar quantas classes são alteradas entre duas anonimizações de *k-anonimity* de Mondrian efetuadas a dois conjuntos de dados que diferem em um elemento. Para isto é primeiro realizada uma anonimização do *dataset* para servir de comparação. De seguida é removido ou adicionado um elemento ao *dataset*. O elemento a ser removido é escolhido aleatoriamente. O elemento a ser acrescentado ao *dataset* é criado também com atributos aleatórios cujos valores para cada atributo estão entre os intervalo de valores possíveis para esse conjunto de dados. Ambos estes testes são feitos individualmente para poderem ser comparados. É utilizado o valor de 10 para o parâmetro k e são efetuadas para cada caso 1000 iterações. Após

cada iteração são verificadas as diferenças entre o *dataset* resultante da anonimização e o *dataset* anonimizado anteriormente. Uma vez que o conjunto de dados census-10k demora cerca de 30 segundos a ser anonimizado enquanto que o conjunto de dados SAL-50k demora cerca de 17 minutos este teste foi realizado apenas para o *dataset* census-10k. Os resultados obtidos para os dois casos, adição e remoção de um elemento aleatoriamente são apresentados nas seguintes tabelas.

Classes Alteradas	Ocorrências
0	651
3	61
4	46
5	14
6	25
7	13
8	11
10	11
14	8
16	11
22	8
23	4
25	13
27	16
28	2
30	5
33	6
34	1
36	5
50	28
51	4
109	2
110	51
111	2
112	1
113	1

Tabela 12.: Alterações com adição de um elemento aleatório

Analisando os resultados obtidos, não são encontradas diferenças significativas entre os dois casos. Para ambas as situações em cerca de 65% das iterações não há alterações de classes inteiras. Nestes casos o elemento é simplesmente adicionado ou removido a uma das classes de

Classes Alteradas	Ocorrências
0	648
3	59
4	61
5	25
6	14
7	11
8	3
9	12
10	10
12	1
14	3
16	10
18	2
25	13
26	2
27	15
33	14
49	1
50	37
109	8
110	41
111	1
112	9

Tabela 13.: Alterações com remoção de um elemento aleatório

equivalência já existentes, sofrendo estas uma alteração de cardinalidade de 1. Para as restantes cerca de 35% das vezes que se adiciona ou remove um elemento do *dataset* aleatoriamente há alterações de classes completas entre duas anonimizações cujos conjuntos de dados originais variam em um elemento. Este valor é surpreendentemente alto e significa que em qualquer um destes casos o ataque demonstrado no capítulo anterior era possível caso a *query* efetuada incluísse estas diferenças. As classes que sofrem alterações encontram-se sempre próximas do elemento removido ou adicionado em relação aos valores dos atributos. Sempre que ocorrem diferenças entre os dois *datasets* o que costuma acontecer é que sensivelmente metade das classes existem num dos resultados da anonimização e não no outro, e vice-versa. Isto é cerca de metade do valor das classes de equivalência alteradas apresentado nas tabelas é exclusivo a cada um dos

resultados, o resultado da anonimização realizado antes dos testes para efeitos comparativos e o resultado de uma anonimização quando ocorre a remoção ou adição de um elemento.

Tendo agora os valores máximos de ambos os parâmetros testados experimentalmente, 113 para o tamanho máximo de classes alteradas e 35 para o tamanho máximo de uma classe, repara-se que a multiplicação destes dois parâmetros resulta em valores bastante elevados rondando os 3000 para o conjunto de dados census-10k.

Este valor é no entanto muito alto, pode até ser um valor extremo em demasia uma vez não existirem na anonimização do *dataset* census-10k cerca de 100 classes de equivalência com tamanho de sensivelmente 30 elementos pois as classes de tamanhos maiores como se pode ver no gráfico 14 representam apenas uma pequena parte do conjunto de dados anonimizado.

Foi desta forma realizado outro teste experimental que contabiliza o número total de elementos pertencentes às classes de equivalência que apareceram ou desapareceram entre as anonimizações. Não existindo diferenças significativas nos resultados quer se adicione ou remova um elemento de forma aleatória ao *dataset*, os resultados deste teste são apresentados apenas em uma tabela. Devido ao elevado tamanho da mesma esta é colocada em anexo em 18.

Olhando para estes resultados, os valores mais altos indicam agora um total de cerca de 1600 tuplos alterados. Valor que é cerca de metade do obtido através da simples multiplicação dos dois parâmetros estimados experimentalmente. Observa-se também que apenas cerca de 10% destes valores encontram-se acima de 500, valor ainda muito elevado.

Na tabela 14 apresentam-se os valores do ruído médio que se obteria calibrando o valor da sensibilidade utilizando esta aproximação. É utilizado um valor de k de 10, e um valor de *epsilon* de 0.1.

Sensibilidade	Erro Médio
3000	27.172
500	4.568

Tabela 14.: Erro médio para aproximação híbrida da sensibilidade

Os dados desta tabela foram obtidos utilizando os testes à utilidade dos *datasets* realizados no capítulo anterior. Os erros são apresentados em percentagem.

Como seria de esperar pelos valores estimados para a sensibilidade, o erro médio obtido por *queries* de contagem nestas condições é extremamente elevado. Os valores retornados neste contexto diferem das verdadeiras respostas em cerca de 27 vezes, valor inutilizável num contexto real.

Esta aproximação do valor da sensibilidade seria assim bastante alta e sendo o modelo calibrado com um valor desta magnitude obteríamos resultados de muito baixa utilidade. Este é de fato um valor limite, pessimista que mostra o pior caso possível não sendo possível assim o ataque mostrado no capítulo 3 ser efetuado de maneira nenhuma. Mesmo com um relaxamento do valor máximo, em que o modelo continuaria seguro em 90% dos casos, o valor de ruído obtido mesmo sendo relativamente mais baixo é ainda muito elevado e continuaria a ser extremamente penalizador para a utilidade dos dados obtidos.

5.3 APROXIMAÇÃO EXPERIMENTAL

Nesta secção é realizada uma aproximação puramente experimental para o valor do parâmetro da sensibilidade. O objetivo é tentar realmente estimar o efeito que uma anonimização pode ter num *dataset* de forma a calcular o valor da sensibilidade. A ideia explorada nesta secção é que sendo o modelo de *differential privacy* seguro quando utilizado sozinho para *queries* de contagem calibradas com sensibilidade de 1, estimar experimentalmente o efeito que uma anonimização pode ter em *queries* de contagem para calibrarmos assim o valor da sensibilidade de forma a cobrir este efeito estimado.

O teste realizado é igual ao efetuado na secção 4.3 que mede a utilidade dos dados após a aplicação dos modelos. No entanto os resultados são aqui apresentados de forma diferente. Em vez de serem apresentado os valores dos erros médios das *queries* efetuadas, é apresentada a distribuição desses mesmos erros. É assim efetuada uma anonimização com o algoritmo de Mondrian para k igual a 10. São efetuadas *queries* de contagem ao conjunto de dados antes e após a anonimização e verificado o erro em percentagem. O teste foi agora realizado apenas para o *dataset* census-10k de forma aos resultados obtidos serem comparáveis com os das restantes aproximações efetuadas nas secções anteriores.

Os resultados obtidos encontram-se nos seguintes dois gráficos em que no primeiro as *queries* utilizadas são de inclusão e no segundo de sobreposição.

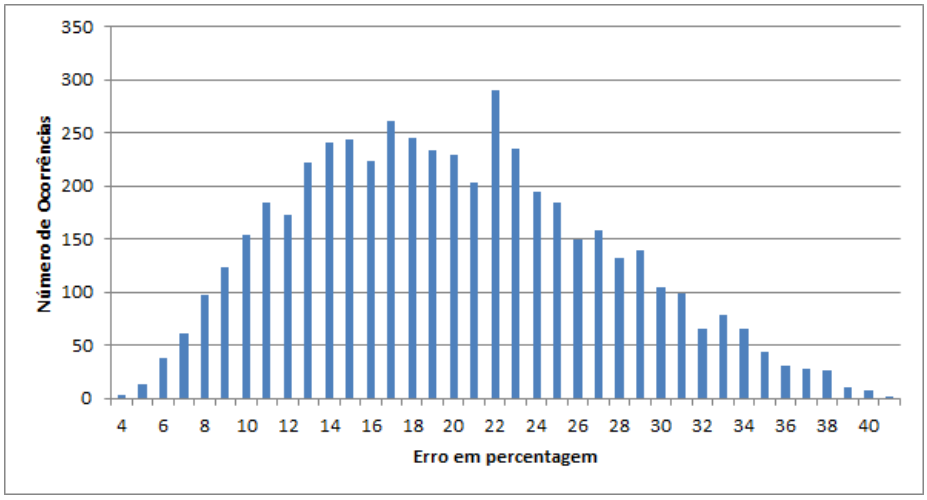


Figura 16.: Distribuição de erros para queries de Inclusão

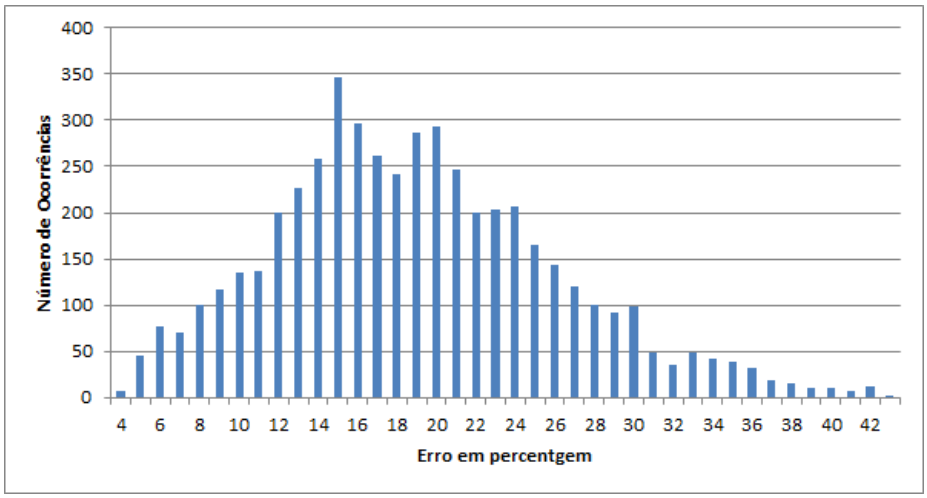


Figura 17.: Distribuição de erros para queries de sobreposição

Cada coluna dos gráficos representa o número de *queries* que tiveram um erro em percentagem arredondado para baixo, ou seja a coluna 10 por exemplo contém todos os valores entre 10 e 11, correspondente a essa coluna. As diferenças encontradas entre os dois gráficos correspondentes aos dois tipos de *queries* não são significantes.

A primeira distribuição apresenta uma média de 19 e desvio padrão de 7.3 enquanto que para as *queries* de sobreposição a distribuição tem uma média de 18 e desvio padrão de 7.1. Novamente no entanto para corretamente estimar o valor da sensibilidade é necessário ter em conta o pior

caso, os valores mais altos dos gráficos que rondam os 40%, sendo o valor máximo encontrado o de 43% com uma ocorrência.

Com esta aproximação puramente experimental o valor da sensibilidade poderia ser assim calibrado a um valor correspondente a 43% do resultado retornado por uma *query* de contagem efetuada ao conjunto de dados anonimizado. Sendo este o erro máximo que uma anonimização pode ter no resultado de uma *query* em comparação com o *dataset* antes de ser anonimizado, então com esta calibração do parâmetro da sensibilidade o modelo de *differential privacy* já terá em conta este possível desvio e o ataque referido no capítulo anterior devido à falta de correta calibração da sensibilidade com a utilização dos dois modelos de *k-anonimity* e *differential privacy* não ocorrerá.

Tal como foi feito para as aproximações anteriores, este valor máximo é relaxado de forma a se tentar obter uma aproximação melhor.

Assim, assumindo informalmente que as distribuições de erros apresentadas nas figuras anteriores se assemelham a uma distribuição normal, representada na imagem 18, 95% dos valores encontram-se a dois desvios padrão de distancia do valor médio. Neste caso poderíamos assim afirmar que em 95% das vezes, calibrando o valor da sensibilidade a de cerca de 33% do valor obtido pela contagem era suficiente para que a utilização simultânea dos dois mecanismos de proteção de privacidade fosse segura. Este valor acaba por não ser muito inferior ao valor máximo.

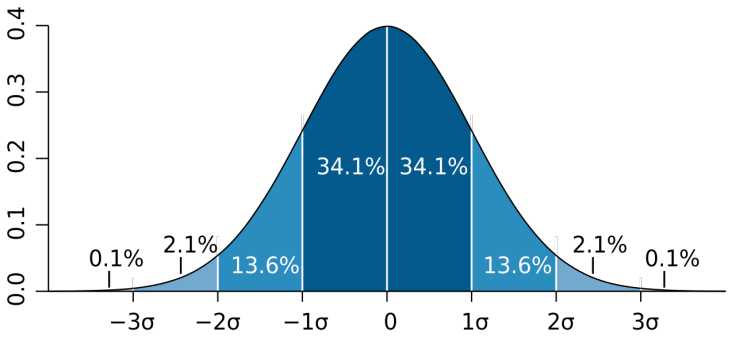


Figura 18.: Distribuição Normal

No entanto, mesmo com o relaxamento do valor máximo o ruído acrescentado é bastante elevado. Em concreto o valor de contagem retornado pelas *queries* apresentadas nos gráficos tem um valor médio de cerca de 1000. Ora 43% disto significaria em média um valor de 430 para o parâmetro da sensibilidade no primeiro caso e um valor de 330 para o segundo. Estes

valores são ainda bastante elevados. São apresentados na tabela seguinte os valores do ruído médio que se obteria calibrando o valor da sensibilidade segundo esta aproximação. Tal como nas aproximações anteriores estes valores são obtidos utilizando os testes do capítulo anterior com o parâmetro de k a 10 e o parâmetro de *epsilon* a 0.1.

Sensibilidade	Erro Médio
430	3.998
330	2.963

Tabela 15.: Erro médio para aproximação experimental da sensibilidade

Pela observação da tabela é retirado que os resultados obtidos por esta aproximação possuem utilidade também extremamente baixa, onde em média os valores retornados pelas *queries* neste contexto diferem das verdadeiras respostas em 4 vezes.

Mesmo para *datasets* de dimensões maiores que os utilizados nos testes, onde o ruído acrescentado pelo modelo de *differential privacy* é menos penalizador, continua neste caso a ser elevado sendo o valor da sensibilidade nesta aproximação calculado a partir do valor retornado pelas *queries*.

Na aproximação puramente teórica são mostradas as condições ideais sobre o qual se poderia estimar o valor da sensibilidade. Na segundo secção, na aproximação híbrida, é apresentado precisamente o caso oposto onde são mostrados os parâmetros estimados experimentalmente que formam o caso limite para a calibração da sensibilidade. Nesta secção puramente experimental é efetuada uma aproximação com base no que de facto acontece quando são realizadas *queries* de contagem ao *datasets*, no impacto que o pré-processamento por via do algoritmo de anonimização de Mondrian tem nos dados.

Assim, para condições ideais que são mostradas na primeira aproximação a composição das técnicas poderia ser utilizada de forma segura e com um nível de ruído aceitável. No entanto para os casos experimentados e utilizando o algoritmo de anonimização de Mondrian, o mais consensual na literatura, a correta calibração do parâmetro de sensibilidade de forma a evitar o ataque demonstrado no capítulo 3 obrigaria a um imenso acréscimo de ruído que na nossa opinião tornaria os resultados obtidos muito pouco úteis.

6

CONCLUSÃO

Este último capítulo tem como objetivo realizar uma síntese de todo o trabalho realizado nesta dissertação. São também apresentadas as contribuições fornecidas por este trabalho e fornecidas outras ideias de estudo que poderão ser exploradas futuramente de forma ao trabalho ser continuado e produzido um resultado ainda mais completo.

6.1 SÍNTESE DO TRABALHO

Após a realização do trabalho é nesta secção feita uma síntese do trabalho realizado nesta dissertação.

De inicio os objetivos da dissertação eram o estudo das duas principais técnicas para obtenção de privacidade em sistemas de dados, os modelos de *k-anonimity* e de *differential privacy* e da forma como poderiam interagir quando utilizadas em conjunto. Para dar inicio ao trabalho e ao documento escrito e de forma a adquirir maior conhecimento sobre estas áreas foi realizado o levantamento do estado da arte no capítulo 2. Este estado da arte contempla os modelos de *k-anonimity* e de *differential privacy* e alguns dos seus variantes como por exemplo o modelo de ℓ*-diversity* e as noções de *crowd blending privacy* e de *zero knowledge privacy*.

Após a realização do estado da arte foi começada a ser desenvolvida a aplicação JAVA que serviria de base para os testes efetuados aos modelos de *k-anonimity* e de *differential privacy* que constituem o capítulo 4. Tendo esta dissertação uma forte componente experimental, com os testes realizados foi obtida uma melhor compreensão sobre os modelos e o seu funcionamento. Ao longo da dissertação a aplicação foi sendo atualizada conforme as necessidades de novos testes ficando no final com mais de 2000 linhas de código.

Tiradas conclusões sobre os testes realizados e após algum trabalho teórico sobre a utilização conjunta dos modelos de *differential privacy* e *k-anonimity*, foi atingido no capítulo 3 o resultado que demonstra o ataque inesperado que pode ocorrer neste contexto. A normal calibração do parâmetro de sensibilidade para *queries* de contagem com o valor de 1, pode na verdade reduzir

o nível de privacidade obtido pelos modelos de forma individual quebrando até as garantias de segurança fornecidas pelo modelo de *differential privacy*.

A pergunta natural que surgiu de seguida foi a de qual será então o valor correto para o parâmetro de sensibilidade de forma a que este ataque não seja possível. Para responder a esta pergunta foram realizadas três aproximações distintas no capítulo 5. Teórica, híbrida e experimental. Com estas aproximações foram obtidos valores para o parâmetro da sensibilidade com o quais seria possível utilizar os modelos de *differential privacy* e *k-anonimity* em conjunto de acordo com o caso de estudo apresentado 3, de forma segura evitando o ataque por completo. No entanto esta segura calibração do parâmetro de sensibilidade traria um custo elevado. Foi assim concluído que a menos que o algoritmo de anonimização oferecesse as condições perfeitas descritas com a aproximação teórica, esta utilização composta dos modelos para ser efetuada de forma segura implicaria resultados com um nível de ruído muito elevado que inviabilizaria por completo a sua utilização.

6.2 CONTRIBUIÇÕES

Da realização desta dissertação resultaram diversas contribuições relevantes.

As nossas primeiras contribuições incluem o estudo geral e testes experimentais realizados aos modelos de *k-anonimity* e *differential privacy* não só quando utilizados de forma individual mas também quando utilizados em conjunto.

De forma a trabalhar com conjuntos de dados anonimizados surgiu o problema de como efetuar *queries* de contagem a dados que se encontram generalizados. Para resolver este problema foram definidas dois tipos de *queries* de contagem, as *queries* de inclusão e sobreposição. Com este tipo de *queries* conseguimos assim trabalhar de forma mais concreta e calculada sobre conjuntos de dados generalizados e efetuar da mesma forma os testes experimentais necessários.

A próxima e importante contribuição foi mostrar que para responder a *queries* de contagem, a composição de algoritmos de anonimização com o modelo de *differential privacy* pode efetivamente destruir a garantia de segurança oferecida pelo modelo. Este resultado não nos era à partida óbvio pois o que seria de esperar era que com a utilização dos dois modelos em conjunto iríamos atingir um nível de segurança maior, nunca menor. No entanto esta intuição mostrou-se estar perigosamente incorreta.

Por fim com o objetivo de descobrir quais as condições necessárias para possibilitar o uso dos dois modelos em conjunto foram realizadas três aproximações ao parâmetro da sensibilidade.

Foram encontrados valores que permitiriam a segura utilização da composição dos modelos mas que iriam acabar por produzir resultados de muito baixa utilidade excetuando condições ideais.

6.3 TRABALHO FUTURO

Como é natural, foram surgindo ao longo da realização da dissertação outros caminhos de estudo que, não podendo ser todos explorados, ficam assim referidos nesta secção.

Como trabalho futuro seria interessante verificar como é que outros tipos algoritmos baseados no pré-processamento de dados poderiam afetar a utilização do modelo de *differential privacy*. Será que mesmo algoritmos não relacionados com a proteção da privacidade de dados como por exemplo algoritmos ligados a tarefas de *datamining*, que normalmente aplicam um pré-processamento aos dados com o intuito de remover possível ruído e aumentar a qualidade dos dados para análise, teriam também um efeito destrutivo? Ou iriam desta vez permitir o relaxamento dos parâmetros de segurança do modelo de *differential privacy* de modo a produzir resultados com maior utilidade mas mantendo o mesmo nível de privacidade?

Por outro lado e com o objetivo de tornar este trabalho ainda mais completo, apesar do algoritmo de anonimização de Mondrian ser o mais consensual e utilizado na literatura, seria interessante explorar de como outros algoritmos de anonimização como o Incognito, Anatomy ou Datafly se comportariam quando utilizados também em conjunto com o modelo de *differential privacy*. Mesmo utilizados individualmente seria também interessante comparar valores relativos a utilidade dos dados para os diversos algoritmos de anonimização.

Todas as ideias teóricas e testes experimentais realizados nesta dissertação utilizaram por base *queries* agregadas do tipo de contagem. Como trabalho futuro teríamos também como objetivo testar outros tipos de *queries* e testar para quais o mesmo ataque ocorreria e com que implicações.

BIBLIOGRAFIA

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pp. 557–570, 2002.

[2] C. Dwork, "Differential privacy," in *ICALP*, pp. 1–12, Springer, 2006.

[3] J. Gehrke, M. Hay, E. Lui, and R. Pass, "Crowd-blending privacy," in *CRYPTO*, pp. 479–496, 2012.

[4] L. Sweeney, "Uniqueness of simple demographics in the u.s. population," 2000.

[5] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*L*-diversity: Privacy beyond *k*-anonymity," *TKDD*, vol. 1, no. 1, 2007.

[7] M. Barbosa, A. Pinto, and B. Gomes, "Generically extending anonymization algorithms to deal with successive queries," in *CIKM*, pp. 1362–1371, 2012.

[8] A. V. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *PODS*, pp. 211–222, 2003.

[9] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *EUROCRYPT*, pp. 486–503, 2006.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, pp. 265–284, 2006.

[11] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *FOCS*, pp. 94–103, 2007.

[12] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, pp. 19–30–, 2009.

[13] A. Haeberlen, B. C. Pierce, and A. Narayan, "Differential privacy under fire," in *Proceedings of the 20th USENIX Conference on Security*, pp. 33–33, 2011.

[14] N. Li, W. H. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, *k*-anonymization meets differential privacy," in *ASIACCS*, pp. 32–33, 2012.

[15] J. Gehrke, E. Lui, and R. Pass, "Towards privacy for social networks: A zero-knowledge based definition of privacy," in *TCC*, pp. 432–449, 2011.

[16] S. Hettich and S. D. Bay, "The uci kdd archive [http://kdd.ics.uci.edu]. university of california irvine, department of information and computer science."

[17] S. Ruggles, J. T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, and M. Sobek, "Integrated public use microdata series [https://www.ipums.org/]."

[18] X. Xiao and Y. Tao, "M-invariance: Towards privacy preserving re-publication of dynamic datasets," in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, pp. 689–700, 2007.

[19] G. Ghinita, P. Karras, P. Kalnis, and N. mamoulis, "Fast data anonymization with low information loss," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp. 758–769, 2007.

[20] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proceedings of the 32Nd International Conference on Very Large Data Bases*, pp. 139–150, 2006.

[21] X. Xiao, K. Yi, and Y. Tao, "The hardness and approximation algorithms for l-diversity," in *Proceedings of the 13th International Conference on Extending Database Technology*, pp. 135–146, 2010.

[22] L. Sweeney, "Datafly: A system for providing anonymity in medical data," in *Proceedings of the IFIP TC11 WG11.3 Eleventh International Conference on Database Securty XI: Status and Prospects*, pp. 356–381, 1998.

[23] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pp. 49–60, 2005.

[24] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proceedings of the 22Nd International Conference on Data Engineering*, pp. 25–, 2006.

[25] UT Dallas Security and Privacy Lab, "Utd anonymization toolbox [http://cs.utdallas.edu/dspl/cgi-bin/toolbox/]."

[26] D. Richard Hipp, "Sqlite [http://www.sqlite.org/]."

[27] C. Dwork, "Differential privacy: A survey of results," in *TAMC*, pp. 1–19, 2008.

[28] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *IEEE Symposium on Security and Privacy*, pp. 111–125, 2008.

[29] A. Hájek, "Interpretations of probability," in *The Stanford Encyclopedia of Philosophy*, 2012.

A

ANEXOS

Idade	Sexo	Nível de Educação	Local de Nascimento	Número de Ocorrências
36	1	10	30	1
36	1	10	31	1
36	1	10	32	1
36	1	10	34	10
36	1	10	35	1
36	1	10	36	24
36	1	6	34	1
36	1	6	36	5
36	1	8	34	3
36	1	8	36	2
36	1	9	36	5
37	1	10	30	2
37	1	10	31	4
37	1	10	33	2
37	1	10	34	11
37	1	10	36	40
37	1	4	36	1
37	1	5	34	1
37	1	7	36	3
37	1	8	34	1
37	1	8	36	1
37	1	9	30	1
37	1	9	34	1
37	1	9	36	2

Tabela 16.: Elementos da classe de equivalência com 124 elementos

Idade	Sexo	Nível de Educação	Local de Nascimento	Número de Ocorrências
34	1	1	36	1
34	1	10	30	2
34	1	10	32	1
34	1	10	33	2
34	1	10	34	5
34	1	10	36	28
34	1	5	36	1
34	1	7	36	3
34	1	8	33	1
34	1	8	34	1
34	1	8	36	2
34	1	9	34	1
34	1	9	36	4
35	1	10	30	2
35	1	10	31	2
35	1	10	33	1
35	1	10	34	9
35	1	10	35	2
35	1	10	36	36
35	1	6	36	1
35	1	7	36	1
35	1	8	35	1
35	1	8	36	4
35	1	9	35	1
35	1	9	36	4

Tabela 17.: Elementos da classe de equivalência com 116 elementos

Elementos alterados	Ocorrências	Elementos alterados	Ocorrências
0	648	93	6
39	5	95	5
41	14	97	4
43	6	99	3
45	7	101	1
47	2	103	2
49	12	135	7
51	9	137	5
53	6	139	5
55	8	153	1
57	4	155	7
59	11	161	2
60	1	162	3
61	1	180	3
63	2	259	1
67	2	295	2
68	2	403	8
69	4	404	7
71	1	449	11
73	5	450	2
75	2	469	2
79	3	491	14
81	1	809	27
85	1	810	11
87	5	1623	38
89	3	1624	12
91	1	1673	9

Tabela 18.: Total de elementos alterados