GROEP GEGEVENSBESCHERMING ARTIKEL 29

0829/14/NL
WP 216

Advies 5/2014 over anonimiseringstechnieken

Goedgekeurd op 10 april 2014

DE GROEP VOOR DE BESCHERMING VAN PERSONEN IN VERBAND MET DE VERWERKING VAN PERSOONSGEGEVENS,

opgericht bij Richtlijn 95/46/EG van het Europees Parlement en de Raad van 24 oktober 1995,

gezien artikel 29 en artikel 30 van die richtlijn,

gezien het reglement van orde van de Groep,

HEEFT HET VOLGENDE ADVIES GOEDGEKEURD:

SAMENVATTING

In dit advies analyseert de Groep gegevensbescherming artikel 29 (hierna „de Groep" genoemd) de doeltreffendheid en beperkingen van bestaande anonimiseringstechnieken door ze te toetsen aan het EU-rechtskader inzake gegevensbescherming. Voorts geeft de Groep aanbevelingen om ervoor te zorgen dat bij het hanteren van deze technieken altijd afdoende rekening wordt gehouden met het daaraan inherente restrisico van identificatie.

De Groep is zich bewust van het potentieel dat anonimisering biedt, in het bijzonder als strategie om niet alleen individuele personen, maar ook de samenleving in het algemeen te laten profiteren van „open gegevens" en tegelijk de daaraan verbonden risico's te beperken voor de betrokkenen. Niettemin is uit casestudy's en onderzoekspublicaties gebleken dat het allesbehalve een eenvoudige opgave is een echt anonieme dataset aan te leggen met behoud van zo veel mogelijk onderliggende informatie als nodig is om het beoogde doel te verwezenlijken.

In de zin van Richtlijn 95/46/EG en andere ter zake dienende EU-rechtsinstrumenten wordt anonimisering bewerkstelligd door persoonsgegevens zodanig te verwerken dat elke mogelijkheid tot identificatie van betrokkenen onherroepelijk wordt uitgesloten. Daarbij moeten de voor de verwerking verantwoordelijken rekening houden met diverse factoren. Tevens dient te worden gekeken naar alle middelen „waarvan mag worden aangenomen" dat zij „redelijkerwijs" door degene die voor de verwerking verantwoordelijk is dan wel door enige andere derde in te zetten zijn om een persoon te identificeren.

Anonimisering strekt tot verdere verwerking van persoonsgegevens en moet als zodanig beantwoorden aan het verenigbaarheidsvereiste door recht te doen aan de wettelijke grondslagen en omstandigheden van de verdere verwerking. Daar komt nog bij dat geanonimiseerde gegevens weliswaar buiten de werkingssfeer van de gegevensbeschermingswetgeving vallen, maar dat de betrokkenen niettemin nog steeds bescherming genieten op grond van andere bepalingen (zoals die welke het vertrouwelijke karakter van communicatie waarborgen).

De belangrijkste anonimiseringstechnieken, meer bepaald randomisatie en generalisatie, worden nader toegelicht in dit advies. Bijzondere aandacht wordt besteed aan ruistoevoeging, permutatie, differentiële privacy, aggregatie (samenvoeging), *k*-anonimiteit, *l*-diversiteit en *t*-gelijkenis. In dit advies worden de beginselen van die technieken uiteengezet, worden de sterke en zwakke punten ervan belicht en wordt verduidelijkt welke fouten en vergissingen vaak worden gemaakt bij het gebruik van elke techniek.

Voorts wordt dieper ingegaan op de deugdelijkheid van elke techniek door toetsing aan drie criteria:

(i) de herleidbaarheid, dat wil zeggen de mogelijkheid om een persoon te individualiseren,

(ii) de koppelbaarheid, dat wil zeggen de mogelijkheid om records in verband te brengen met een persoon, en

(iii) de deduceerbaarheid, dat wil zeggen de mogelijkheid om persoonsgebonden informatie af te leiden.

Een goed begrip van de belangrijkste sterke en zwakke punten van elke techniek bevordert de opzet van een geschikt anonimiseringsproces in een bepaalde situatie.

Ook het begrip pseudonimisering komt hier ter sprake om misverstanden en misvattingen uit de weg te ruimen. Pseudonimisering mag niet worden gezien als synoniem van anonimisering. Pseudonimisering beperkt louter de koppelbaarheid van een dataset aan de oorspronkelijke identiteit van een betrokkene, en is bijgevolg een nuttige maatregel om gegevens te beveiligen.

De conclusie van dit advies is dat anonimiseringstechnieken niet alleen privacywaarborgen kunnen bieden, maar ook kunnen bijdragen aan de opzet van efficiënte anonimiseringsprocessen, zij het alleen wanneer bij de invulling daarvan voor technische onderbouwing wordt gezorgd. Met andere woorden, de randvoorwaarden (contextuele situatie) en de doelstelling(en) van het anonimiseringsproces moeten duidelijk worden uiteengezet om de daarmee beoogde doeleinden te verwezenlijken en tegelijk bruikbare gegevens te produceren. De optimale oplossing dient per geval te worden vastgesteld, zo nodig door verschillende technieken te combineren. De in dit advies aangereikte aanbevelingen voor de praktijk moeten daarbij houvast bieden.

Tot slot dienen de voor de verwerking verantwoordelijken zich rekenschap te geven van het feit dat een geanonimiseerde dataset nog steeds restrisico's kan inhouden voor de betrokkenen. Anonimisering en re-identificatie zijn immers actuele thema's waarnaar volop onderzoek wordt verricht en waarover regelmatig nieuwe ontdekkingen worden gepubliceerd. Bovendien zijn zelfs anoniem gemaakte gegevens, zoals statistieken, bruikbaar om bestaande persoonsprofielen te verrijken, wat nieuwe kwesties en problemen op het gebied van gegevensbescherming kan doen rijzen. Een en ander maakt duidelijk dat anonimisering geen eenmalige exercitie is en dat de voor de verwerking verantwoordelijken de daarmee samenhangende risico's periodiek opnieuw moeten bekijken.

1 Inleiding

Apparaten, sensoren en netwerken genereren een grote verscheidenheid aan gegevens in grote hoeveelheden terwijl de kosten voor gegevensopslag verwaarloosbaar klein worden. Daarom bestaat er bij het grote publiek een groeiende belangstelling in en vraag naar het hergebruik van die gegevens. „Open gegevens" bieden de samenleving, personen en organisaties onmiskenbare voordelen, maar alleen als ieders rechten op bescherming van persoonsgegevens en de persoonlijke levenssfeer worden geëerbiedigd.

Anonimisering kan een goede strategie zijn om de voordelen te behouden en de risico's te verminderen. Zodra een dataset echt anoniem gemaakt is en personen niet langer identificeerbaar zijn, is de Europese gegevensbeschermingswetgeving niet langer van toepassing. Casestudy's en onderzoekspublicaties hebben echter aangetoond dat het allesbehalve een sinecure is om op basis van een uitgebreide verzameling van persoonsgegevens een echt anonieme dataset aan te leggen met behoud van zo veel mogelijk onderliggende informatie als nodig is om het beoogde doel te verwezenlijken. Zo kan een als anoniem aangemerkte dataset bijvoorbeeld op zodanige wijze worden gecombineerd met een andere dataset dat het mogelijk wordt de identiteit van een of meer personen te achterhalen.

In dit advies analyseert de Groep de doeltreffendheid en beperkingen van bestaande anonimiseringstechnieken door die te toetsen aan het EU-rechtskader inzake gegevensbescherming. Voorts formuleert de Groep aanbevelingen om deze technieken op behoedzame en verantwoorde wijze toe te passen bij het opzetten van een anonimiseringsproces.

2 Definities en juridische analyse

2.1. Definities in het EU-rechtskader

In overweging 26 van Richtlijn 95/46/EG komt anonimisering ter sprake door te stellen dat anoniem gemaakte gegevens buiten de werkingssfeer van de gegevensbeschermingswetgeving vallen:

> *„Overwegende dat de beschermingsbeginselen moeten gelden voor elk gegeven betreffende een geïdentificeerde of identificeerbare persoon; dat, om te bepalen of een persoon identificeerbaar is, moet worden gekeken naar alle middelen waarvan mag worden aangenomen dat zij redelijkerwijs door degene die voor de verwerking verantwoordelijk is dan wel door enig ander persoon in te zetten zijn om genoemde persoon te identificeren; dat de beschermingsbeginselen niet van toepassing zijn op gegevens die op zodanige wijze anoniem zijn gemaakt dat de persoon waarop ze betrekking hebben niet meer identificeerbaar is; dat de gedragscodes in de zin van artikel 27 een nuttig instrument kunnen zijn om een indicatie te geven omtrent de middelen waarmee de gegevens anoniem kunnen worden gemaakt en kunnen worden bewaard in een vorm die identificatie van de betrokkene niet langer mogelijk maakt."*[1]

[1]Voorts zij opgemerkt dat deze zienswijze ook wordt gehanteerd in overweging 23 van het voorstel voor een Verordening van het Europees Parlement en de Raad betreffende de bescherming van natuurlijke personen in

Bij nadere lezing bevat overweging 26 een conceptuele definitie van het begrip anonimisering. Om gegevens anoniem te maken, is het volgens overweging 26 noodzakelijk voldoende elementen uit die gegevens te verwijderen zodat de betrokkene niet langer identificeerbaar is. De gegevens moeten meer in het bijzonder op zodanige wijze worden verwerkt dat ze niet langer bruikbaar zijn om een natuurlijke persoon te identificeren door „alle middelen waarvan mag worden aangenomen dat zij redelijkerwijs [...] in te zetten zijn" door degene die voor de verwerking verantwoordelijk is dan wel door enige andere derde. Een belangrijke factor daarbij is dat de verwerking onomkeerbaar moet zijn. In de richtlijn wordt niet verduidelijkt hoe een dergelijk proces om gegevens niet-identificeerbaar te maken (de-identificatie) uitgevoerd kan of dient te worden[2]. De nadruk ligt op het resultaat, namelijk dat de gegevens het niet mogelijk mogen maken de betrokkene te identificeren met „alle" middelen „waarvan mag worden aangenomen" dat zij „redelijkerwijs" in te zetten zijn. Voorts wordt verwezen naar gedragscodes als nuttig instrument om mogelijke anonimiseringsmechanismen uit te werken, en naar de bewaring in een vorm die identificatie van de betrokkene „niet langer mogelijk maakt". De richtlijn stelt bijgevolg een bijzonder hoge norm.

Ook in de e-privacyrichtlijn (Richtlijn 2002/58/EG) wordt in sterk vergelijkbare zin verwezen naar „anonimisering" en „anonieme gegevens". Overweging 26 luidt:

> *„Verkeersgegevens die worden gebruikt voor de marketing van communicatiediensten of voor de levering van diensten met toegevoegde waarde moeten ook worden gewist of anoniem gemaakt na de levering van de dienst".*

Dienovereenkomstig bepaalt artikel 6, lid 1, het volgende:

> *„Verkeersgegevens met betrekking tot abonnees en gebruikers die worden verwerkt en opgeslagen door de aanbieder van een openbaar elektronischecommunicatienetwerk of -dienst, moeten, wanneer ze niet langer nodig zijn voor het doel van de transmissie van communicatie, worden gewist of anoniem gemaakt, onverminderd de leden 2, 3 en 5, alsmede artikel 15, lid 1."*

Bovendien bepaalt artikel 9, lid 1:

> *„Wanneer andere locatiegegevens dan verkeersgegevens die betrekking hebben op gebruikers of abonnees van elektronischecommunicatienetwerken of -diensten verwerkt kunnen worden, mogen deze gegevens slechts worden verwerkt wanneer zij anoniem zijn gemaakt of wanneer de gebruikers of abonnees daarvoor hun toestemming hebben gegeven, voorzover en voor zolang zulks nodig is voor de levering van de dienst met toegevoegde waarde."*

De onderliggende redenering is dat het resultaat van anonimisering als op persoonsgegevens toegepaste techniek volgens de huidige stand van de techniek even permanent moet zijn als uitwissing. De verwerking van persoonsgegevens moet met andere woorden voorgoed onmogelijk worden gemaakt.[3]

---

verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens (algemene verordening gegevensbescherming): „[o]m te bepalen of een persoon identificeerbaar is, dienen alle middelen in aanmerking te worden genomen waarvan redelijkerwijs te verwachten valt dat zij door de voor de verwerking verantwoordelijke, of door ieder ander worden gebruikt om de persoon te identificeren".

[2]Op dit begrip wordt dieper ingegaan op blz. 8 van dit advies.

[3]Dienaangaande zij herinnerd aan de definitie van anonimisering in internationale normen, zoals ISO 29100, zijnde het proces waarbij persoonlijk identificeerbare informatie (PII) onomkeerbaar wordt gewijzigd, en wel op zodanige wijze dat de belangrijkste betrokkene daarvan niet langer direct of indirect kan worden geïdentificeerd door degene die voor de verwerking van de PII verantwoordelijk is, noch alleen, noch in samenwerking met

2.2. Juridische analyse

Bij nader inzien komen uit de wetteksten over anonimisering in de richtinggevende EU-rechtsinstrumenten inzake gegevensbescherming duidelijk vier belangrijke kenmerken naar voren:

- Anonimisering kan voortvloeien uit de verwerking van persoonsgegevens met het doel elke mogelijkheid tot identificatie van de betrokkene onherroepelijk uit te sluiten.

- Aangezien er in de EU-wetgeving geen bindende norm is voorgeschreven, zijn diverse anonimiseringstechnieken denkbaar.

- Bijzondere aandacht moet worden besteed aan contextuele factoren: er moet worden gekeken naar „alle" middelen „waarvan mag worden aangenomen" dat zij „redelijkerwijs" door degene die voor de verwerking verantwoordelijk is dan wel door derden in te zetten zijn om de persoon te identificeren; daarbij moet met name worden gelet op de middelen die momenteel, volgens de huidige stand van de techniek, redelijkerwijs in te zetten zijn (gezien de toenemende rekenkracht van computers en het groeiende aantal beschikbare hulpmiddelen).

- Er zijn inherente risico's verbonden aan anonimisering: het is zaak die risicofactor te laten meewegen bij de beoordeling van de deugdelijkheid van elke anonimiseringstechniek en tevens rekening te houden met de gebruiksmogelijkheden van gegevens die door middel van die techniek „anoniem werden gemaakt"; ook de ernst en waarschijnlijkheid van het risico moeten worden ingeschat.

In dit advies wordt veeleer dan „anonimiteit" of „anonieme gegevens" de aanduiding „anonimiseringstechniek" gehanteerd om in verband met een technisch-organisatorische maatregel die beoogt gegevens „anoniem" te maken nadrukkelijk te wijzen op het daaraan inherente restrisico van re-identificatie.

2.2.1. Rechtmatigheid van het anonimiseringsproces

Allereerst zij opgemerkt dat anonimisering een techniek is die op persoonsgegevens wordt toegepast om de betrokkene onherroepelijk niet-identificeerbaar te maken. Het uitgangspunt is derhalve dat de persoonsgegevens moeten zijn verzameld en verwerkt met inachtneming van de toepasselijke wetgeving inzake de bewaring van gegevens in een identificeerbare vorm.

In dit verband is het anonimiseringsproces, zijnde de verwerking van zulke persoonsgegevens om ze anoniem te maken, een voorbeeld van „verdere verwerking". Als zodanig moet die

---

enige andere derde (ISO 29100:2011). Het onomkeerbare karakter van de in de persoonsgegevens aangebrachte wijziging staat ook in deze ISO-norm centraal. Vanuit dit oogpunt sluit deze definitie grotendeels aan bij de beginselen en begrippen die ten grondslag liggen aan Richtlijn 95/46/EG. Dat geldt ook voor de definities in bepaalde nationale wetgevingen (bijvoorbeeld in Italië, Duitsland en Slovenië). Daarin wordt de nadruk gelegd op de niet-identificeerbaarheid en wordt gesteld dat re-identificatie „onevenredig veel moeite" moet kosten (DE, SI). De Franse gegevensbeschermingswetgeving bepaalt echter dat zelfs wanneer de re-identificatie van de betrokkene uiterst veel moeite kost en hoogst onwaarschijnlijk is, de gegevens verder als persoonsgegevens dienen te worden aangemerkt. Hier wordt met andere woorden niet verwezen naar de „redelijkheidstoets".

verwerking voldoen aan de verenigbaarheidstoets overeenkomstig de richtsnoeren die door de Groep worden aangereikt in haar advies 3/2013 over doelbinding[4].

Dit betekent dat anonimisering in beginsel haar rechtsgrondslag vindt in alle in artikel 7 genoemde redenen (inclusief de behartiging van het gerechtvaardigde belang van de voor de verwerking verantwoordelijke) mits tevens wordt voldaan aan de in artikel 6 van de richtlijn beschreven eisen inzake gegevenskwaliteit, en er naar behoren rekening wordt gehouden met de specifieke omstandigheden en alle factoren die ter sprake komen in het advies van de Groep over doelbinding[5].

Voorts dient nadrukkelijk te worden gewezen op de bepalingen in artikel 6, lid 1, onder e), van Richtlijn 95/46/EG (alsook in artikel 6, lid 1, en in artikel 9, lid 1, van de e-privacyrichtlijn), op grond waarvan persoonsgegevens „in een vorm die het mogelijk maakt de betrokkenen te identificeren" niet langer mogen worden bewaard dan voor de verwezenlijking van de doeleinden waarvoor zij worden verzameld of vervolgens worden verwerkt, noodzakelijk is.

Op zich komt uit deze bepaling duidelijk naar voren dat persoonsgegevens ten minste anoniem moeten worden gemaakt „door standaardinstellingen" (met inachtneming van verschillende juridische eisen, zoals die welke in de e-privacyrichtlijn worden genoemd met betrekking tot verkeersgegevens). Wil de voor de verwerking verantwoordelijke deze persoonsgegevens bewaren zodra de doeleinden van de oorspronkelijke of verdere verwerking zijn bereikt, dan dient er gebruik te worden gemaakt van anonimiseringstechnieken om identificatie onherroepelijk te voorkomen.

Naar het oordeel van de Groep is anonimisering, als voorbeeld van verdere verwerking van persoonsgegevens, aan te merken als verenigbaar met de oorspronkelijke doeleinden van de verwerking, zij het alleen op voorwaarde dat het anonimiseringsproces ertoe strekt op betrouwbare wijze anoniem gemaakte informatie te produceren in de hier weergegeven zin.

Voorts zij beklemtoond dat anonimisering dient plaats te vinden met inachtneming van de juridische beperkingen die het Europees Hof van Justitie (HvJ) in herinnering heeft gebracht in zijn arrest in zaak C-553/07 (*College van burgemeester en wethouders van Rotterdam v M.E.E. Rijkeboer*), met betrekking tot de noodzaak om de gegevens in identificeerbare vorm te bewaren, bijvoorbeeld teneinde betrokkenen in staat te stellen hun rechten van toegang uit te oefenen. In dit arrest verklaart het HvJ: „*Artikel 12, sub a, van richtlijn [95/46] verlangt van de lidstaten dat zij niet alleen voor het heden, maar ook voor het verleden voorzien in een recht op toegang tot informatie over de ontvangers of de categorieën ontvangers van de gegevens en tot de inhoud van de verstrekte informatie. Het staat aan de lidstaten, een termijn voor de bewaring van deze informatie en een daarop afgestemde toegang daartoe vast te stellen die een juist evenwicht vormen tussen, enerzijds, het belang van de betrokkene om zijn*

[4]Advies 3/2013 van de Groep gegevensbescherming artikel 29 is te vinden op: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[5]Zo moet er met name een beoordeling ten gronde worden uitgevoerd in het licht van alle relevante omstandigheden en met bijzondere aandacht voor de volgende essentiële factoren:
a) het verband tussen de doeleinden waarvoor de persoonsgegevens zijn verzameld en de doeleinden waarvoor ze verder worden verwerkt;
b) de context waarin de persoonsgegevens zijn verzameld en de redelijke verwachtingen van de betrokkenen ten aanzien de verdere verwerking daarvan;
c) de aard van de persoonsgegevens en het effect van de verdere verwerking op de betrokkenen;
d) de garanties die de voor de verwerking verantwoordelijke biedt met het oog op een eerlijke verwerking alsook om ongewenste gevolgen voor de betrokkenen te vermijden.

persoonlijke levenssfeer te beschermen, met name door middel van de bij richtlijn 95/46 voorziene mogelijkheden om de voor de verwerking verantwoordelijke in te schakelen en zich tot de rechter te wenden, en, anderzijds, de last die de verplichting tot bewaring van die informatie inhoudt voor de voor de verwerking verantwoordelijke."

Dat is met name relevant wanneer een voor de verwerking verantwoordelijke zich wat anonimisering betreft beroept op artikel 7, onder f), van Richtlijn 95/46/EG: het gerechtvaardigde belang van de voor de verwerking verantwoordelijke dient steeds te worden afgewogen tegen de rechten en fundamentele vrijheden van de betrokkenen.

Zo heeft de Nederlandse gegevensbeschermingsautoriteit bijvoorbeeld in 2012-2013 onderzoek uitgevoerd naar het gebruik van technologieën voor deep packet inspection door vier exploitanten van mobiele netten. Daaruit bleek dat rechtsgrondslag werd ontleend aan artikel 7, onder f), van Richtlijn 95/46/EG om verkeersgegevens inhoudelijk anoniem te maken zo snel mogelijk nadat die gegevens werden verzameld. Artikel 6 van de e-privacyrichtlijn bepaalt namelijk dat verkeersgegevens met betrekking tot abonnees en gebruikers die worden verwerkt en opgeslagen door de aanbieder van een openbaar elektronischecommunicatienetwerk of -dienst, zo snel mogelijk moeten worden gewist of anoniem gemaakt. Aangezien zulks is toegestaan krachtens artikel 6 van de e-privacyrichtlijn, bestaat in casu een overeenkomstige rechtsgrondslag in artikel 7 van de richtlijn gegevensbescherming. Deze redenering kan ook worden omgekeerd: is een specifieke vorm van gegevensverwerking niet toegestaan op grond van artikel 6 van de e-privacyrichtlijn, dan kan geen rechtsgrondslag worden ontleend aan artikel 7 van de richtlijn gegevensbescherming.

2.2.2. Identificeerbaarheid van geanonimiseerde gegevens

De Groep heeft zich in haar advies 4/2007 gebogen over het begrip persoonsgegeven en meer in het bijzonder aandacht besteed aan de bouwstenen die kunnen worden onderscheiden in de definitie van dit begrip in artikel 2, onder a), van Richtlijn 95/46/EG, waaronder het element „geïdentificeerd of identificeerbaar" van die definitie. Tevens heeft de Groep dienaangaande geconcludeerd: „[g]eanonimiseerde gegevens zijn dan anonieme gegevens betreffende een persoon die eerder identificeerbaar was, maar nu niet meer kan worden geïdentificeerd".

Bijgevolg heeft de Groep al verduidelijkt dat de in de richtlijn genoemde toets om te bepalen of er sprake is van „middelen waarvan mag worden aangenomen dat zij redelijkerwijs [...] in te zetten zijn" als criterium wordt voorgesteld om na te gaan of het anonimiseringsproces voldoende privacyveilig is en met andere woorden de identificatie van de betrokkenen „redelijkerwijs" onmogelijk maakt. De bijzondere context en omstandigheden van een specifiek geval zijn rechtstreeks van invloed op de identificeerbaarheid. In de technische bijlage bij dit advies wordt onderzocht welke gevolgen de keuze van de meest geschikte techniek heeft.

Zoals hierboven beklemtoond, zijn onderzoek en instrumentarium in volle ontwikkeling. Bovendien neemt de rekenkracht van computers gestaag toe. Het is bijgevolg niet doenbaar, noch zinvol om een uitputtende opsomming te geven van de omstandigheden waarin identificatie niet langer mogelijk is. Niettemin dient rekening te worden gehouden met sommige essentiële factoren die hieronder worden geïllustreerd.

In de eerste plaats kan worden gesteld dat de voor de verwerking verantwoordelijken zich moeten toespitsen op de concrete middelen die noodzakelijk zijn om de anonimiseringstechniek terug te draaien. Daarbij moeten ze met name aandacht besteden aan

de benodigde kosten en knowhow om die middelen toe te passen, en tevens de ernst en waarschijnlijkheid daarvan inschatten. Ze dienen bijvoorbeeld de met de anonimisering gepaard gaande kosten en moeite (zowel qua de benodigde tijd als wat de vereiste middelen betreft) af te zetten tegen de toenemende beschikbaarheid, niet alleen van goedkope technische middelen om personen in datasets te identificeren, maar ook van andere datasets die publiekelijk toegankelijk zijn (bijvoorbeeld in het kader van een opengegevensbeleid). Daarbij dienen ze aandacht te schenken aan de talloze praktijkgevallen waarbij een onvolledige anonimisering in aanzienlijke mate afbreuk heeft gedaan aan de belangen van de betrokkenen of bij gelegenheid onherstelbare schade daaraan heeft toegebracht.[6] Opgemerkt zij dat het risico van identificatie mettertijd kan toenemen en ook afhangt van de ontwikkeling van informatie- en communicatietechnologie. Eventuele wettelijke bepalingen moeten derhalve op technologisch neutrale wijze worden geformuleerd en bij uitstek rekening houden met veranderingen in de ontwikkeling van het IT-potentieel.[7]

In de tweede plaats zijn „de middelen waarvan mag worden aangenomen dat zij redelijkerwijs in te zetten zijn [...] om genoemde persoon te identificeren" die welke worden gebruikt „door degene die voor de verwerking verantwoordelijk is dan wel door enig ander persoon". Het is daarom van wezenlijk belang in te zien dat wanneer een voor de verwerking verantwoordelijke de originele (identificeerbare) gegevens niet verwijdert op gebeurtenisniveau, en een deel van die dataset doorgeeft (bijvoorbeeld na het verwijderen of maskeren/afschermen van identificeerbare gegevens), de resulterende dataset nog steeds valt onder de noemer van persoonsgegevens. Uitsluitend wanneer de voor de verwerking verantwoordelijke de gegevens dermate samenvoegt (aggregeert) dat de individuele gebeurtenissen niet langer identificeerbaar zijn, kan de resulterende dataset als anoniem worden aangemerkt. Wanneer een organisatie bijvoorbeeld gegevens over reizigersbewegingen verzamelt, worden de individuele reispatronen op gebeurtenisniveau nog steeds met persoonsgegevens gelijkgesteld voor elke partij, en wel zolang de voor de verwerking verantwoordelijke (of enige andere partij) toegang heeft tot de oorspronkelijke onbewerkte gegevens, ook al werden de direct identificerende gegevens („identificatoren") verwijderd uit de aan derden doorgegeven dataset. Wanneer echter de voor de verwerking verantwoordelijke de onbewerkte gegevens verwijdert, en alleen geaggregeerde statistieken doorgeeft aan derden op hoog niveau, zoals „voor reisroute X zijn er op maandag 160 % meer reizigers dan op dinsdag", moeten die als anonieme gegevens worden beschouwd.

Een doeltreffende anonimiseringsoplossing verhindert dat een persoon in een dataset wordt geïndividualiseerd, dat twee records in een dataset (of in twee afzonderlijke datasets) met elkaar in verband worden gebracht en dat uit die dataset informatie wordt afgeleid. In het algemeen is het verwijderen van direct identificerende gegevens op zich dus niet voldoende om zeker te stellen dat de betrokkene niet langer identificeerbaar is. Doorgaans moeten verdergaande maatregelen worden genomen om identificatie te voorkomen. Een en ander hangt opnieuw af van de omstandigheden en de doeleinden van de verwerking waarvoor de anonieme gegevens zijn bestemd.

[6] Het is belangwekkend dat in overweging 23 van de recentelijk (21 oktober 2013) door het Europees Parlement ingediende amendementen op de ontwerptekst van de algemene verordening gegevensbescherming uitdrukkelijk wordt bepaald: „Om uit te maken of van middelen redelijkerwijs te verwachten valt dat zij zullen worden gebruikt om de persoon te identificeren, dienen alle objectieve factoren, zoals de kosten van en de tijd benodigd voor identificatie, in aanmerking te worden genomen, rekening houdend met de beschikbare technologie op het tijdstip van verwerking en de technologische ontwikkeling".

[7] Zie Advies 4/2007 van de Groep gegevensbescherming artikel 29, blz. 15.

PRAKTIJKGEVAL
DNA-profielen zijn een voorbeeld van persoonsgegevens waarvoor, gezien het unieke karakter van sommige profielen, een risico van identificatie bestaat wanneer de toegepaste techniek er uitsluitend toe strekt de identiteit van de donor te verwijderen. Uit de literatuur[8] is bekend dat zelfs wanneer DNA-materiaal „anoniem" wordt gedoneerd, de identiteit van bepaalde personen toch kan worden achterhaald door publiekelijk beschikbare genetische informatiebronnen (bijvoorbeeld geslachts- en overlijdensregisters, resultaten van gegevensopvragingen of query's met zoekmachines) te combineren met de metagegevens over DNA-donoren (tijdstip van donatie, leeftijd, woonplaats).

Beide anonimiseringstechnieken[9] – randomisatie en generalisatie – vertonen tekortkomingen. Dit neemt niet weg dat elke techniek in de gegeven omstandigheden en context geschikt kan zijn om het beoogde doel te verwezenlijken zonder de privacy van de betrokkenen in gevaar te brengen. Het moet duidelijk zijn dat „identificatie" niet alleen te verstaan is als de mogelijkheid om iemands naam en/of adres te achterhalen, maar ook verwijst naar de identificeerbaarheid door gegevens te herleiden tot de persoon, met elkaar in verband te brengen en af te leiden. Bovendien doen de voornemens van de voor de verwerking verantwoordelijke of van de ontvanger van de gegevens niet ter zake bij de beoordeling van de vraag of de gegevensbeschermingswetgeving van toepassing is: de gegevensbeschermingsvoorschriften behouden hun gelding zolang de gegevens identificeerbaar zijn.

Een dataset waarop een anonimiseringstechniek is toegepast (anoniem gemaakt en vrijgegeven door degene die oorspronkelijk voor de verwerking verantwoordelijk was) mag rechtmatig door een derde worden verwerkt zonder rekening te houden met de vereisten inzake gegevensbescherming, mits de betrokkenen direct noch indirect kunnen worden geïdentificeerd in de oorspronkelijke dataset. Niettemin zijn derden verplicht rekening te houden met de hierboven genoemde contextuele factoren en aanwijzingen (waaronder de specifieke kenmerken die werden toegepast door degene die oorspronkelijk voor de verwerking verantwoordelijk was) om te bepalen hoe ze deze geanonimiseerde gegevens voor eigen doeleinden zullen gebruiken en meer bepaald combineren: derden kunnen inderdaad op verschillende manieren aansprakelijk worden gesteld voor de daaruit voortvloeiende gevolgen. Indien deze factoren en kenmerken een onaanvaardbaar risico doen ontstaan dat betrokkenen worden geïdentificeerd, valt de verwerking opnieuw binnen de werkingssfeer van de gegevensbeschermingswetgeving.

In de bovenstaande uiteenzetting wordt geen aanspraak op volledigheid gemaakt. Veeleer wordt beoogd algemene richtsnoeren aan te reiken die houvast bieden bij het beoordelen van de mate van identificeerbaarheid van een bepaalde dataset die werd geanonimiseerd door middel van de verschillende beschikbare technieken. Alle bovengenoemde factoren vormen evenzovele risicofactoren die moeten worden afgewogen, enerzijds door de voor de verwerking verantwoordelijken wanneer ze datasets anonimiseren en anderzijds door derden wanneer ze „anoniem gemaakte" datasets voor eigen doeleinden gebruiken.

2.2.3. Risico's verbonden aan het gebruik van geanonimiseerde gegevens

Wanneer de voor de verwerking verantwoordelijken overwegen anonimiseringstechnieken toe te passen, moeten zij zich rekenschap geven van de volgende risico's.

[8] Zie John Bohannon, „Genealogy Databases Enable Naming of Anonymous DNA Donors" in *Science*, vol. 339, nr. 6117, 18 januari 2013, blz. 262.

[9] De belangrijkste kenmerken van en verschilpunten tussen beide anonimiseringstechnieken worden hieronder uiteengezet in deel 3 getiteld „Technische analyse".

- Een bekend misverstand is dat gepseudonimiseerde gegevens worden gelijkgesteld met geanonimiseerde gegevens. In het deel getiteld „Technische analyse" wordt uitgelegd dat gepseudonimiseerde gegevens niet gelijk te stellen zijn met geanonimiseerde informatie omdat de versleutelde gegevens (pseudoniemen) nog steeds herleidbaar zijn tot de individuele betrokkene en koppelbaar tussen verschillende datasets. Pseudonimiteit laat de deur open voor identificatie en valt bijgevolg binnen de werkingssfeer van het rechtskader inzake gegevensbescherming. Dat is met name van belang in de context van wetenschappelijk, statistisch of historisch onderzoek.[10]

PRAKTIJKGEVAL
Het bekende „AOL(America On Line)-incident" illustreert goed de algemeen verspreide misvattingen over pseudonimisering. In 2006 werd een database met 20 miljoen zoektermen van ruim 650 000 gebruikers over een periode van drie maanden voor het publiek toegankelijk gemaakt. De enige maatregel ter bescherming van de privacy bestond erin de AOL-gebruikersidentificatie te vervangen door een numeriek attribuut. Het gevolg daarvan was dat sommige gebruikers publiekelijk werden geïdentificeerd en gelokaliseerd. Gepseudonimiseerde tekenreeksen (strings) van gegevensopvragingen of query's met zoekmachines, met name wanneer die gekoppeld zijn aan andere attributen, zoals IP-adressen of andere configuratieparameters van clientcomputers, hebben een zeer hoog identificerend gehalte.

- Een tweede vergissing is ervan uit te gaan dat naar behoren geanonimiseerde gegevens (die voldoen aan alle hierboven genoemde voorwaarden en criteria, en die per definitie buiten de werkingssfeer van de richtlijn gegevensbescherming vallen) individuele personen eender welke garanties ontzeggen, op de allereerste plaats omdat het gebruik van die gegevens onder andere wetgevende instrumenten kan vallen. Zo verbiedt artikel 5, lid 3, van de e-privacyrichtlijn bijvoorbeeld de opslag van en toegang tot elk soort „informatie" (met inbegrip van niet-persoonlijke informatie) op eindapparatuur zonder toestemming van de abonnee/gebruiker omdat dit valt onder het ruimere beginsel van de vertrouwelijkheid van communicatie.

- Van een derde vergissing is sprake wanneer wordt voorbijgegaan aan het effect dat naar behoren geanonimiseerde gegevens in bepaalde omstandigheden hebben op personen, meer in het bijzonder bij profilering. Het recht op bescherming van de persoonlijke levenssfeer is verankerd in artikel 8 van het EVRM en in artikel 7 van het Handvest van de grondrechten van de EU. Ook al is de gegevensbeschermingswetgeving mogelijk niet langer van toepassing op dit soort gegevens, het omgaan met geanonimiseerde datasets die voor gebruik door derden zijn vrijgegeven, kan in wezen afbreuk doen aan het privacyrecht. De grootste zorgvuldigheid is geboden bij het omgaan met geanonimiseerde informatie, zeker wanneer die (vaak in combinatie met andere gegevens) wordt gebruikt om beslissingen te nemen die (weliswaar indirecte) gevolgen kunnen hebben voor personen. Zoals de Groep beklemtoont in dit advies en meer in detail verduidelijkt in haar advies over het begrip doelbinding (Advies 3/2013)[11], moet de gewettigde verwachting van de betrokkenen ten aanzien van de verdere verwerking van hun gegevens worden getoetst aan de relevante contextuele factoren, zoals de relatie tussen de betrokkenen en de voor de verwerking verantwoordelijken, de toepasselijke wettelijke verplichtingen en de transparantie van verwerkingen.

[10] Zie ook Advies 4/2007 van de Groep gegevensbescherming artikel 29, blz. 18-20.
[11] Dit advies is te vinden op: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

3 Technische analyse, deugdelijkheid van technologieën en klassieke fouten

Aan anonimisering wordt op verschillende manieren en met een wisselende mate van doeltreffendheid invulling gegeven. In dit deel staat een overzicht van de aandachtspunten waarmee de voor de verwerking verantwoordelijken rekening moeten houden wanneer ze die technieken toepassen. Daarbij moeten zij meer in het bijzonder letten op de garanties die volgens de huidige stand van de techniek worden geboden, en zich rekenschap geven van drie risico's die van essentieel belang zijn voor het anonimiseringsproces:

- *herleidbaarheid*, zijnde de mogelijkheid om een persoon in de dataset te individualiseren door sommige of alle records uit te lichten;
- *koppelbaarheid*, zijnde de mogelijkheid om ten minste twee records over dezelfde betrokkene of groep betrokkenen met elkaar in verband te brengen (in dezelfde database of in twee verschillende databases). Wanneer een aanvaller (bijvoorbeeld door de correlatie te analyseren) kan vaststellen dat twee records aan een en dezelfde groep personen zijn gerelateerd, zonder personen binnen deze groep te kunnen individualiseren, dan doorstaat de techniek de „herleidbaarheidstoets", maar niet de koppelbaarheidstoets;
- *deduceerbaarheid*, zijnde de mogelijkheid om de waarde van een persoonskenmerk („attribuut") met grote waarschijnlijkheid af te leiden uit de waarden van een reeks andere attributen.

Een anonimiseringsoplossing die deze drie risico's uitsluit, is in voldoende mate bestand tegen re-identificatie op basis van de meest waarschijnlijke en redelijke middelen die worden gebruikt door de voor de verwerking verantwoordelijke en enige derde. De Groep benadrukt dienaangaande dat momenteel onderzoek wordt verricht naar technieken om gegevens niet-identificeerbaar en anoniem te maken. Daaruit is op consistente wijze gebleken dat geen enkele techniek in wezen vrij is van tekortkomingen. In grote lijnen kan anonimisering op twee manieren worden benaderd: enerzijds door te ***randomiseren*** en anderzijds door te ***generaliseren***. In dit advies komen ook andere begrippen ter sprake, zoals *pseudonimisering, differentiële privacy*, *l-diversiteit* en *t-gelijkenis*.

De in dit deel van het advies gebruikte termen hebben de volgende betekenis: een dataset is een geheel of verzameling van gegevensrecords betreffende personen (de betrokkenen). Elke record is gerelateerd aan één betrokkene en bestaat uit een reeks waarden (ook „informatie-elementen" of „ingangen" genoemd, bijvoorbeeld 2013) voor elk attribuut (bijvoorbeeld het jaar). Een dataset kan ook de vorm aannemen van een tabel (of reeks tabellen) of, zoals tegenwoordig steeds meer het geval is, een grafiek met verklarende aantekeningen (geannoteerde grafiek) of met gewogen waarden (gewogen grafiek). De voorbeelden in dit advies hebben betrekking op tabellen, maar gaan ook op voor de andere grafische voorstellingswijzen van records. Combinaties van attributen die verband houden met een betrokkene of een groep betrokkenen worden quasi-identificatoren genoemd. In sommige gevallen kan een dataset meerdere records over dezelfde persoon bevatten. Een „aanvaller" is een derde (niet zijnde de voor de verwerking verantwoordelijke, noch de gegevensverwerker) die per ongeluk of opzettelijk toegang krijgt tot de oorspronkelijke records.

3.1. Randomisatie

Randomisatie wordt gerekend tot de groep technieken waarmee de waarheidsgetrouwheid van gegevens wordt gewijzigd met het doel die gegevens los te koppelen van de persoon. Als de gegevens voldoende *at random* zijn (dat wil zeggen willekeurig of onbepaald), is het niet langer mogelijk om ze te herleiden tot een specifieke persoon. Door randomisatie op zich wordt de uniciteit of eenduidigheid van elke record niet verminderd: er bestaat nog steeds een één-op-éénrelatie tussen de record en de betrokkene. Wel kan randomisatie bescherming bieden tegen deductieve aanvallen en ook risico's van deduceerbaarheid verminderen. Door randomisatie te combineren met generalisatietechnieken worden betere privacywaarborgen geboden. Waar nodig moeten aanvullende technieken worden toegepast om zeker te stellen dat een record niet herleidbaar is tot een enkele persoon.

3.1.1. Ruistoevoeging

Ruistoevoeging is een techniek die vooral nuttig is wanneer attributen belangrijke negatieve gevolgen kunnen hebben voor personen. De attributen in de dataset worden daarbij gewijzigd om ze minder nauwkeurig te maken, zij het met behoud van de algemene verdeling. Wie werkt met een dataset waaraan ruis is toegevoegd, gaat ervan uit dat de waarden accuraat zijn, terwijl dat slechts tot op zekere hoogte het geval is. Stel dat de lichaamslengte van een persoon oorspronkelijk tot op één centimeter precies werd gemeten. Na anonimisering door ruistoevoeging bevat de dataset lengtematen die slechts tot op ± 10 cm nauwkeurig zijn. Wordt deze techniek op doeltreffende wijze toegepast, dan is het voor een derde niet mogelijk om een persoon te identificeren, noch om de oorspronkelijke gegevens terug te rekenen of te achterhalen hoe de gegevens werden gewijzigd.

Doorgaans moet ruistoevoeging worden gecombineerd met andere anonimiseringstechnieken, zoals het verwijderen van doorzichtige attributen en quasi-identificatoren. De omvang van de toegevoegde ruis dient te worden bepaald volgens het vereiste informatiegehalte en het effect van de bekendmaking van beveiligde attributen op de individuele privacy.

3.1.1.1. Privacywaarborgen

- *Herleidbaarheid:* ook al zijn de records minder betrouwbaar, het blijft mogelijk om ze te herleiden tot een persoon (mogelijk zonder de identiteit te kunnen vaststellen).
- *Koppelbaarheid:* het blijft mogelijk om de records van dezelfde persoon met elkaar in verband te brengen, maar de records zijn minder betrouwbaar. Dit betekent dat een echte record kan worden gekoppeld aan een op kunstmatige wijze toegevoegde record (dat wil zeggen aan ruis). In sommige gevallen kan een betrokkene bij een onjuiste attributie aanzienlijk meer risico lopen dan wanneer de records juist worden geattribueerd.
- *Deduceerbaarheid:* deductieve aanvallen zijn mogelijk, maar de slaagkans is kleiner en er kunnen foutpositieven (en foutnegatieven) ontstaan.

3.1.1.2. Vaak gemaakte fouten

- Inconsistente ruis toevoegen: is ruis uit semantisch oogpunt niet neutraal (dat wil zeggen „buitenproportioneel" en in strijd met de logica tussen attributen in een dataset), dan kan de aanvaller die toegang heeft tot de database de ruis uitfilteren en in sommige gevallen de ontbrekende informatie-elementen opnieuw genereren. Wanneer

de dataset bovendien te sterk gespreid[12] is, blijven de als ruis toegevoegde informatie-elementen koppelbaar via een externe bron.

- Ervan uitgaan dat kan worden volstaan met het toevoegen van ruis: ruistoevoeging is een aanvullende maatregel die het voor aanvallers moeilijker maakt om persoonsgegevens te achterhalen. Tenzij de mate van ruis in de dataset groter is dan het informatiegehalte, mag er niet van worden uitgegaan dat ruistoevoeging op zich toereikend is om gegevens anoniem te maken.

3.1.1.3. Tekortkomingen van ruistoevoeging

Een welbekend re-identificatie-experiment werd uitgevoerd op de klantendatabase van de aanbieder van videostreamingdiensten Netflix. In deze database met meer dan 100 miljoen filmrecensies werd door bijna 500 000 gebruikers een waarderingscijfer op een schaal van 1 tot 5 toegekend aan ruim 18 000 films. Overeenkomstig het interne privacybeleid van Netflix werden de gegevens „anoniem gemaakt" door alle identificerende klantengegevens te verwijderen, behalve de filmrecensies en datums. Vervolgens gaf Netflix de database vrij voor het publiek en werden de geometrische eigenschappen ervan geanalyseerd door onderzoekers. Ruis werd toegevoegd naarmate iets hogere of lagere waarderingscijfers werden toegekend.

Desondanks bleek 99 % van de gebruikersrecords eenduidig identificeerbaar te zijn in de dataset door als selectiecriteria 8 waarderingscijfers en datums met een foutenmarge van 14 dagen te gebruiken. Zelfs met minder strenge selectiecriteria (2 waarderingscijfers en een foutenmarge van 3 dagen) bleef 68 % van de gebruikers identificeerbaar.[13]

3.1.2. Permutatie

Deze techniek bestaat erin de attribuutwaarden in een tabel in willekeurige volgorde van plaats te verwisselen zodat bepaalde waarden op kunstmatige wijze worden gekoppeld aan andere betrokkenen. Permutatie is nuttig wanneer de exacte verdeling van elk attribuut binnen de dataset behouden moet blijven.

Permutatie kan ook als een bijzondere vorm van ruistoevoeging worden beschouwd. Met een klassieke techniek voor ruistoevoeging worden de attributen gewijzigd met aselecte (gerandomiseerde) waarden zonder voorspelbaar patroon. Het kan een hele opgave zijn op consistente wijze ruis te genereren. Bovendien is dit geen privacyveilige oplossing wanneer alleen de attribuutwaarden in geringe mate worden veranderd. Permutatie is een alternatieve techniek om de waarden in de dataset te veranderen door ze gewoon anders te ordenen tussen twee records. Deze omwisseling zorgt ervoor dat het bereik en de verdeling van de waarden identiek blijven, maar verandert wel de correlaties tussen waarden en personen. Gesteld dat tussen twee of meer attributen een logische relatie of statistische correlatie bestaat, dan gaat deze relatie verloren wanneer de attributen onafhankelijk van elkaar van plaats worden verwisseld (gepermuteerd). Vandaar het belang om een verzameling van aan elkaar gerelateerde attributen anders te ordenen zonder de logische relatie ertussen te verbreken. Zo niet, kan een aanvaller de anders geordende attributen identificeren en de permutatie ongedaan maken.

[12] Dit begrip wordt nader toegelicht op blz. 34 van de bijlage.

[13] Narayanan, A., & Shmatikov, V. (mei 2008). „Robust de-anonymization of large sparse datasets" in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, blz. 111-125, IEEE.

Stel een deelverzameling van attributen in een medische dataset, bijvoorbeeld „redenen voor ziekenhuisopname/symptomen/behandelende afdeling". Doorgaans bestaat er een sterke logische relatie die de waarden aan elkaar koppelt. Wordt slechts één van die waarden van plaats verwisseld, dan kan een aanvaller die permutatie opsporen en zelfs terugdraaien.

Net als bij ruistoevoeging is permutatie op zich niet voldoende om gegevens anoniem te maken. Het is ook van belang dat doorzichtige attributen/quasi-identificatoren worden verwijderd.

3.1.2.1. Privacywaarborgen

- *Herleidbaarheid:* net als bij ruistoevoeging blijft het mogelijk om de records te individualiseren, dat wil zeggen te herleiden tot een persoon, ook al zijn de records minder betrouwbaar.
- *Koppelbaarheid:* door attributen en quasi-identificatoren van plaats te verwisselen, kan worden vermeden dat attributen intern of extern in het „juiste" verband worden gebracht met een dataset; niettemin blijft een risico bestaan dat een „onjuist" verband wordt gelegd wanneer een echt informatie-element wordt gerelateerd aan een andere betrokkene.
- *Deduceerbaarheid:* de dataset blijft vatbaar voor deducties, met name wanneer attributen aan elkaar gecorreleerd zijn of sterke logische relaties bezitten. De aanvaller weet echter niet op welke attributen de permutatie werd toegepast en moet er dan ook van uitgaan dat zijn deductie op een onjuiste hypothese kan berusten. Bijgevolg bestaat er alleen nog een risico van probabilistische deduceerbaarheid.

3.1.2.2. Vaak gemaakte fouten

- Het verkeerde attribuut selecteren: indien toegepast op attributen die niet privacygevoelig zijn of geen privacyrisico inhouden, levert de permutatie geen wezenlijke voordelen op voor de bescherming van persoonsgegevens. Immers, zolang de gevoelige/risicohoudende attributen nog aan het oorspronkelijke attribuut gekoppeld zijn, kan de aanvaller gevoelige informatie over personen extraheren.
- Aselecte permutatie van attributen: wanneer twee attributen nauw aan elkaar gecorreleerd zijn, levert een aselecte (willekeurige) permutatie van attributen geen wezenlijke privacywaarborgen op. Deze vaak gemaakte fout wordt geïllustreerd in tabel 1.
- Ervan uitgaan dat kan worden volstaan met permutatie: permutatie op zich is, net als ruistoevoeging, geen garantie voor anonimiteit en moet worden gecombineerd met andere technieken, zoals het verwijderen van doorzichtige attributen.

3.1.2.3. Tekortkomingen van permutatie

Het volgende voorbeeld toont aan dat de aselecte permutatie van attributen onvoldoende privacywaarborgen oplevert zolang er logische relaties tussen verschillende attributen bestaan. Ook na anonimisering valt het inkomen van elke persoon eenvoudig af te leiden uit de functie (en het geboortejaar). Zo valt in een oogopslag uit de gegevens op te maken dat de CEO in de tabel naar alle waarschijnlijkheid geboren is in 1957 en het hoogste salaris heeft, terwijl de werkloze geboren is in 1964 en het laagste inkomen heeft.

Jaar	Geslacht	Functie	Inkomen (na permutatie)
1957	m.	Ingenieur	70 000
1957	m.	CEO	5 000
1957	m.	Werkloos	43 000
1964	m.	Ingenieur	100 000
1964	m.	Manager	45 000

Tabel 1. Voorbeeld van een ondoeltreffende anonimisering door permutatie van gecorreleerde attributen

3.1.3. Differentiële privacy

Differentiële privacy[14] wordt gerekend tot de groep randomisatietechnieken, maar volgt een andere benaderingswijze. Daar waar ruis van tevoren wordt toegevoegd wanneer de dataset moet worden vrijgegeven, kan differentiële privacy worden toegepast wanneer de voor de verwerking verantwoordelijke geanonimiseerde beelden (*views*) van een dataset genereert en tegelijk een exemplaar van de oorspronkelijke gegevens behoudt. Deze anoniem gemaakte beelden worden doorgaans gegenereerd via een deelverzameling van gegevensopvragingen (query's) voor een specifieke derde. Aan de deelverzameling is achteraf bewust aselecte (willekeurige) ruis toegevoegd. Dankzij differentiële privacy kan de voor de verwerking verantwoordelijke bepalen hoeveel ruis hij moet toevoegen en in welke vorm dat moet gebeuren om de nodige privacywaarborgen te bieden.[15] In dit verband is het van het allergrootste belang voortdurend erop toe te zien (ten minste voor elke nieuwe gegevensopvraging) of de mogelijkheid bestaat dat een persoon wordt geïdentificeerd in de queryresultaten. Niettemin zij erop gewezen dat technieken van differentiële privacy de oorspronkelijke gegevens niet veranderen. Zolang de oorspronkelijke gegevens bestaan, kan de voor de verwerking verantwoordelijke in de queryresultaten personen identificeren door rekening te houden met alle middelen waarvan redelijkerwijs te verwachten valt dat zij worden gebruikt. Deze resultaten zijn ook als persoonsgegevens te beschouwen.

De op differentiële privacy gebaseerde benadering heeft als voordeel dat de datasets aan daartoe gemachtigde derden worden verstrekt in antwoord op een specifieke gegevensopvraging (query), en niet zozeer door een enkele dataset vrij te geven. Als hulpmiddel bij een privacyaudit kan de voor de verwerking verantwoordelijke een lijst met alle query's en gegevensopvragingen bijhouden om zeker te stellen dat derden uitsluitend toegang krijgen tot de gegevens waarvoor zij gemachtigd zijn. Ook op een query kunnen anonimiseringstechnieken, zoals ruistoevoeging of substitutie, worden toegepast om de privacy verder te beschermen. Er is nog nader onderzoek nodig om een vraag-en-antwoordmechanisme te ontwikkelen dat degelijke interactiemogelijkheden biedt en tegelijk in staat is gegevensopvragingen relatief accuraat te beantwoorden (dat wil zeggen met zo min mogelijk ruis) met vrijwaring van de privacy.

Om deductieve en koppelingsaanvallen tegen te gaan, moeten de van een entiteit afkomstige gegevensopvragingen worden bijgehouden en moet de over betrokkenen ingewonnen informatie worden gemonitord. Databases die berusten op „differentiële privacy" mogen met

[14]Dwork, C. (2006), „Differential privacy" in *Automata, languages and programming*, blz. 1-12, Springer Berlin Heidelberg.

[15]Cf. Ed Felten (2012) „Protecting privacy by adding noise". URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

andere woorden niet toegankelijk worden gemaakt voor open zoekmachines die geen functies bieden voor de traceerbaarheid van de entiteiten die gegevens opvragen.

3.1.3.1. Privacywaarborgen

- *Herleidbaarheid:* wanneer de output alleen uit statistieken bestaat en indien zorgvuldig afgewogen regels op de dataset worden toegepast, kunnen de antwoorden niet worden gebruikt voor de herleiding tot een persoon.
- *Koppelbaarheid:* bij gebruik van meervoudige gegevensopvragingen kan het mogelijk zijn persoonsgebonden informatie-elementen in twee antwoorden met elkaar in verband te brengen.
- *Deduceerbaarheid:* meervoudige gegevensopvragingen maken het mogelijk informatie over personen of groepen af te leiden.

3.1.3.2. Vaak gemaakte fouten

- Onvoldoende ruis toevoegen: om te vermijden dat een verband wordt gelegd met achtergrondkennis, is het zaak zo weinig mogelijk informatie te verstrekken over het feit of een specifieke betrokkene of groep betrokkenen al dan niet heeft bijgedragen aan de dataset. Het belangrijkste probleem vanuit gegevensbeschermingsoogpunt is het vermogen om de juiste hoeveelheid ruis te genereren en aan de echte antwoorden toe te voegen met het doel de individuele privacy te beschermen, en tegelijk ervoor te zorgen dat de vrijgegeven antwoorden bruikbaar blijven.

3.1.3.3 Tekortkomingen van differentiële privacy

Elke gegevensopvraging (query) op zichzelf staand behandelen: door queryresultaten te combineren, kan ongewild geheime informatie openbaar worden gemaakt. Wordt geen querygeschiedenis bijgehouden, dan kan een aanvaller meerdere gegevensopvragingen verzenden naar een database die berust op „differentiële privacy" om zo de uitgevoerde steekproef in amplitude te beperken totdat een specifiek persoonskenmerk van een enkele betrokkene of groep betrokkenen op deterministische wijze of met een zeer hoge waarschijnlijkheidsgraad naar voren komt. Bovendien is het een misvatting te denken dat de gegevens anoniem zijn voor de derde, terwijl de voor de verwerking verantwoordelijke de betrokkene nog steeds kan identificeren in de oorspronkelijke database door rekening te houden met alle middelen waarvan redelijkerwijs te verwachten valt dat zij worden gebruikt.

3.2. Generalisatie

Generalisatie wordt gerekend tot de tweede groep anonimiseringstechnieken. Deze benaderingswijze bestaat erin de attributen van de betrokkenen te generaliseren (veralgemenen) of af te zwakken (dilueren) door de schaalgrootte of omvang te wijzigen (dat wil zeggen een regio in plaats van een stad, een maand in plaats van een week). Generalisatie kan een efficiënte manier zijn om herleiding tot de persoon uit te sluiten, maar is niet in alle gevallen geschikt om gegevens op doeltreffende wijze anoniem te maken. Er zijn specifieke en geavanceerde kwantitatieve benaderingen nodig om koppelbaarheid en deduceerbaarheid tegen te gaan.

3.2.1. Aggregatie en *k*-anonimiteit

Technieken voor aggregatie en *k*-anonimiteit beogen te voorkomen dat een betrokkene wordt geïndividualiseerd door die samen te voegen met ten minste *k* andere personen. Daartoe worden de attribuutwaarden op zodanige wijze gegeneraliseerd dat alle personen dezelfde waarde gemeen hebben. Door de mate van gedetailleerdheid van een locatie te verminderen (= grovere „granulariteit"), bijvoorbeeld een land in plaats van een stad, wordt een groter aantal betrokkenen bestreken. Individuele geboortedatums kunnen worden gegeneraliseerd tot een datuminterval of gegroepeerd op maand of op jaar. Andere numerieke attributen (bijvoorbeeld salaris, gewicht, lichaamslengte of geneesmiddeldosering) kunnen worden gegeneraliseerd door middel van intervalwaarden (bijvoorbeeld een salarisbedrag tussen 20 000 EUR en 30 000 EUR). Deze methoden zijn ook bruikbaar wanneer quasi-identificatoren kunnen ontstaan door puntenwaarden van attributen aan elkaar te correleren.

3.2.1.1. Privacywaarborgen

- *Herleidbaarheid:* aangezien *k* gebruikers nu dezelfde attributen gemeen hebben, is het niet langer mogelijk een persoon te individualiseren in een groep van *k* gebruikers.
- *Koppelbaarheid:* ook al is de koppelbaarheid minder groot, het blijft mogelijk records met elkaar in verband te brengen in groepen van *k* gebruikers. Binnen deze groep is de kans dat twee records overeenstemmen met dezelfde pseudo-identificatoren gelijk aan 1 op *k* (wat veel groter kan zijn dan de kans dat deze informatie-elementen niet koppelbaar zijn).
- *Deduceerbaarheid:* de belangrijkste zwakke plek in het *k*-anonimiteitsmodel is dat deductieve aanvallen niet worden voorkomen. Immers, gesteld dat alle *k* personen tot dezelfde groep behoren en tevens bekend is van welke groep een persoon deel uitmaakt, dan is de waarde van deze eigenschap eenvoudig te achterhalen.

3.2.1.2. Vaak gemaakte fouten

- Sommige quasi-identificatoren over het hoofd zien: de drempelwaarde van *k* is van kritiek belang voor de *k*-anonimiteit. Hoe groter de waarde van *k*, des te sterker de privacywaarborgen. Een vaak gemaakte fout bestaat erin de waarde *k* op kunstmatige wijze te verhogen door minder quasi-identificatoren in aanmerking te nemen. Wanneer er minder quasi-identificatoren zijn, kunnen eenvoudiger clusters van *k* gebruikers worden samengesteld in verband met het identificerende gehalte dat eigen is aan de andere attributen (met name wanneer bepaalde attributen gevoelig zijn of een zeer hoge entropie hebben, zoals bij uiterst zeldzame attributen). Het niet in aanmerking nemen van alle quasi-identificatoren bij de keuze van het te generaliseren attribuut kan een fatale vergissing blijken te zijn. In dit geval worden sommige personen niet beschermd door de generalisatie in zoverre bepaalde attributen bruikbaar zijn om een persoon te individualiseren in een cluster van *k* personen (zie voorbeeld in tabel 2).
- Kleine waarde van *k*: een soortgelijk probleem doet zich voor wanneer er wordt getracht een kleine waarde van *k* te bereiken. Is *k* te klein, dan is de zwaarte of het gewicht van elke persoon in een cluster te groot, waardoor de slaagkans van deductieve aanvallen toeneemt. Indien *k* bijvoorbeeld gelijk is aan 2, bestaat meer kans dat twee personen dezelfde eigenschap gemeen hebben, dan wanneer *k* groter is dan 10.
- Personen met dezelfde zwaarte niet groeperen: ook het groeperen van een reeks personen met ongelijk verdeelde attributen kan problemen opleveren. De record van

een persoon heeft een wisselend effect op een dataset: sommige personen stellen een belangrijk deel van de informatie-elementen voor, terwijl anderen van te verwaarlozen betekenis zijn. Daarom is het van belang ervoor te zorgen dat *k* groot genoeg is zodat niemand een overwicht heeft van de informatie-elementen in een cluster.

3.1.3.3. Tekortkomingen van *k*-anonimiteit

Het belangrijkste probleem van *k*-anonimiteit is dat deductieve aanvallen niet worden voorkomen. Wanneer de aanvaller in het volgende voorbeeld weet dat in de dataset een specifieke persoon is opgenomen met 1964 als geboortejaar, weet hij ook dat die persoon een hartaanval heeft gekregen. Is het bovendien bekend dat deze dataset afkomstig is van een Franse organisatie, dan ligt het ook voor de hand dat elke persoon in Parijs woont aangezien de postcodes van Parijs altijd beginnen met de drie cijfers 750*.

Jaar	Geslacht	Postcode	Diagnose
1957	m.	750*	Hartaanval
1957	m.	750*	Cholesterol
1957	m.	750*	Cholesterol
1964	m.	750*	Hartaanval
1964	m.	750*	Hartaanval

Tabel 2. Voorbeeld van een gebrekkige toepassing van *k*-anonimiteit

3.2.2. *L*-diversiteit en *t*-gelijkenis

L-diversiteit breidt *k*-anonimiteit uit om deterministische deductieve aanvallen uit te sluiten door ervoor te zorgen dat elk attribuut in elke equivalentieklasse ten minste *l* verschillende waarden heeft.

Het hoofddoel bestaat erin het aantal equivalentieklassen met een geringe variabiliteit in attributen te beperken. Op die manier krijgt een aanvaller die achtergrondkennis over een specifieke betrokkene bezit altijd met een hoge onzekerheidsfactor te maken.

L-diversiteit is nuttig als bescherming tegen deductieve aanvallen wanneer de attribuutwaarden goed verdeeld zijn. Niettemin zij beklemtoond dat deze techniek het lekken van informatie niet verhindert wanneer de attributen binnen een partitie ongelijk verdeeld zijn dan wel deel uitmaken van een beperkt waardebereik of scala van semantische betekenissen. Probabilistische deductieve aanvallen vormen in fine een bedreiging voor *l*-diversiteit.

T-gelijkenis is een verfijnde vorm van *l*-diversiteit waarbij equivalentieklassen worden gecreëerd die gelijkenis vertonen met de aanvankelijke verdeling van attributen in de tabel. Deze techniek is nuttig wanneer het van belang is dat de gegevens zo nauw mogelijk aansluiten bij de oorspronkelijke gegevens. Daartoe wordt als aanvullende voorwaarde opgelegd dat er niet alleen binnen elke equivalentieklasse ten minste *l* verschillende waarden moeten bestaan, maar ook dat elke waarde zo vaak als nodig is wordt voorgesteld om een spiegelbeeld te krijgen van de oorspronkelijke verdeling van elk attribuut.

3.2.2.1. Privacywaarborgen

- *Herleidbaarheid:* net als *k*-anonimiteit kunnen *l*-diversiteit en *t*-gelijkenis ertoe bijdragen dat databaserecords niet worden herleid tot een persoon.

- *Koppelbaarheid:* wat dat betreft, bieden *l*-diversiteit en *t*-gelijkenis geen betere garanties dan *k*-anonimiteit. Hier doet zich hetzelfde probleem voor als met elke cluster: de kans dat dezelfde informatie-elementen aan een en dezelfde betrokkene toebehoren, is groter dan 1 op *n* (waarbij *n* het aantal betrokkenen in de database is).
- *Deduceerbaarheid:* de belangrijkste verbetering van *l*-diversiteit en *t*-gelijkenis ten opzichte van *k*-anonimiteit is dat het niet langer mogelijk is deductieve aanvallen op te zetten tegen een op *l*-diversiteit of *t*-gelijkenis gebaseerde database met een betrouwbaarheidsniveau van 100 %.

3.2.2.2. Vaak gemaakte fouten

- Gevoelige attribuutwaarden beschermen door ze te mengen met andere gevoelige attributen: de aanwezigheid van twee attribuutwaarden in een cluster is niet voldoende om de privacy te waarborgen. De verdeling van gevoelige waarden in elke cluster moet eigenlijk gelijkenis vertonen met de verdeling van die waarden in de volledige populatie, of op zijn minst uniform zijn in de gehele cluster.

3.2.2.3. Tekortkomingen van *l*-diversiteit

In de onderstaande tabel wordt *l*-diversiteit bewerkstelligd voor het attribuut „Diagnose". Wetende dat in de tabel een persoon staat die in 1964 geboren is, valt niettemin met een zeer hoge waarschijnlijkheid daaruit op te maken dat die persoon een hartaanval heeft gekregen.

Jaar	Geslacht	Postcode	Diagnose
1957	m.	750*	**Hartaanval**
1957	m.	750*	**Cholesterol**
1957	m.	750*	**Cholesterol**
1957	m.	750*	**Cholesterol**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Cholesterol**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**
1964	m.	750*	**Hartaanval**

Tabel 3. Op *l*-diversiteit gebaseerde tabel met ongelijk verdeelde waarden voor het attribuut „Diagnose"

Naam	Geboortedatum	Geslacht
Smith	1964	m.
Rossi	1964	m.
Dupont	1964	m.
Jansen	1964	m.
Garcia	1964	m.

Tabel 4. Een aanvaller die weet dat deze personen in tabel 3 staan, kan daaruit opmaken dat ze een hartaanval hebben gekregen

4. Pseudonimisering

Bij pseudonimisering wordt één attribuut (dat doorgaans uniek is) in een record vervangen door een ander attribuut. De natuurlijke persoon is dus nog steeds indirect identificeerbaar. Bijgevolg is pseudonimisering op zich niet voldoende om een dataset volledig anoniem te maken. Niettemin wordt deze techniek hier besproken vanwege de talloze misvattingen en vergissingen die gepaard gaan met het gebruik ervan.

Pseudonimisering vermindert de koppelbaarheid tussen een dataset en de oorspronkelijke identiteit van een betrokkene, en is als zodanig een nuttige beveiligingsmaatregel, maar geen anonimiseringsmethode.

Het resultaat van pseudonimisering kan losstaan van de beginwaarde (bijvoorbeeld een door de voor de verwerking verantwoordelijke gegenereerd aselect getal of een door de betrokkene gekozen familienaam) of kan worden afgeleid uit de oorspronkelijke waarden van een attribuut of verzameling van attributen, bijvoorbeeld een hashfunctie of een encryptiesysteem.

Hier volgen de meest gebruikelijke pseudonimiseringstechnieken:

- Encryptie met een geheime sleutel: in dit geval kan degene die de sleutel bezit elke betrokkene eenvoudig opnieuw identificeren door de dataset te decoderen. De persoonsgegevens zijn immers nog steeds in de dataset opgenomen, zij het in gecodeerde vorm. Gesteld dat een geavanceerd encryptiesysteem werd toegepast, dan is decodering uitsluitend mogelijk wanneer de sleutel bekend is.
- Hashfunctie: deze functie retourneert voor een invoer van willekeurige omvang (één enkel attribuut of een verzameling van attributen) een uitvoer met vaste grootte, en kan niet worden teruggedraaid. Met andere woorden, het bij encryptie bestaande risico dat het proces wordt teruggedraaid, is hier niet aanwezig. Is het bereik van invoerwaarden van de hashfunctie echter bekend, dan bestaat de mogelijkheid de hashfunctie opnieuw daarop toe te passen (*replay*-aanval) om de juiste waarde voor een specifieke record af te leiden. Werd een dataset bijvoorbeeld gepseudonimiseerd door de hashfunctie op een nationaal identificatienummer toe te passen, dan is dit nummer eenvoudig af te leiden door alle mogelijke invoerwaarden te *hashen* en het resultaat te vergelijken met de waarden in de dataset. Hashfuncties zijn doorgaans ontworpen om zo snel mogelijk berekend te kunnen worden, en staan bloot aan brutekrachtaanvallen[16]. Er kunnen ook vooraf berekende tabellen worden gemaakt om een grote verzameling van hashwaarden massaal terug te rekenen.

  Door gebruik te maken van een *salted*-hashfunctie (waarbij aan het gehashte attribuut een willekeurige waarde, „salt" genoemd, wordt toegevoegd), verkleint de kans dat de invoerwaarde wordt herleid. Toch blijft het met redelijke middelen mogelijk de oorspronkelijke attribuutwaarde terug te rekenen die verborgen zit in het resultaat van de *salted*-hashfunctie[17].
- *Keyed*-hashfunctie met opgeslagen sleutel: dit is een bijzondere hashfunctie waarbij een geheime sleutel als aanvullende invoer wordt gebruikt (in tegenstelling tot een *salted*-hashfunctie waar de *salt* doorgaans niet geheim is). Een voor de verwerking

[16] Bij dit soort aanvallen worden alle mogelijke invoerwaarden uitgeprobeerd met het doel correspondentietabellen op te bouwen.

[17] Dat is met name het geval wanneer het attribuuttype (naam, sofinummer, geboortedatum enz.) bekend is. Als aanvullend rekentechnisch vereiste kan een hashfunctie voor sleutelafleiding worden gebruikt, waarbij de berekende waarde meermaals wordt gehasht met een korte *salt*.

verantwoordelijke kan de functie opnieuw toepassen op het attribuut met gebruikmaking van de geheime sleutel. Voor een aanvaller is het veel moeilijker om de functie te herhalen zonder de sleutel te kennen: het aantal uit te proberen mogelijkheden is namelijk zo groot dat dit vrijwel onbegonnen werk zou zijn.

- Deterministische encryptie of *keyed*-hashfunctie met verwijdering van de sleutel: deze techniek komt erop neer voor elk attribuut in de database een willekeurig (aselect) getal te kiezen als pseudoniem en vervolgens de correspondentietabel te verwijderen. Deze oplossing vermindert[18] de koppelbaarheid tussen de persoonsgegevens in de dataset en de gegevens betreffende diezelfde persoon in een andere dataset waar een verschillend pseudoniem wordt gebruikt. Zelfs met een geavanceerd algoritme zou het voor een aanvaller rekentechnisch moeilijk zijn om de functie te decoderen of te herhalen (*replay*-aanval), want dat zou betekenen dat elke mogelijke sleutel moet worden uitgeprobeerd aangezien de sleutel niet bekend is.
- Tokeniseren: deze techniek wordt doorgaans (maar niet uitsluitend) toegepast in de financiële sector om kaartidentificatienummers te vervangen door waarden die minder nuttig zijn voor een aanvaller. Deze techniek vloeit voort uit de vorige en komt erop neer mechanismen voor eenrichtingsencryptie toe te passen of via een indexfunctie een volgnummer of willekeurig (aselect) getal toe te wijzen dat rekenkundig niet af te leiden valt uit de oorspronkelijke gegevens.

4.1. Privacywaarborgen

- *Herleidbaarheid:* de records blijven herleidbaar tot de persoon aangezien deze laatste nog steeds wordt geïdentificeerd door een uniek attribuut dat voortvloeit uit de pseudonimiseringsfunctie (= het gepseudonimiseerde attribuut).
- *Koppelbaarheid:* de records zijn eenvoudig met elkaar in verband te brengen wanneer hetzelfde gepseudonimiseerde attribuut wordt gebruikt om naar dezelfde persoon te verwijzen. Zelfs wanneer voor dezelfde betrokkene verschillende gepseudonimiseerde attributen worden gebruikt, kunnen de records via andere attributen met elkaar in verband worden gebracht. Alleen wanneer geen enkel ander attribuut in de dataset het mogelijk maakt de betrokkene te identificeren en wanneer het oorspronkelijke attribuut volledig werd losgekoppeld van het gepseudonimiseerde attribuut (onder meer door de oorspronkelijke gegevens te verwijderen), kunnen twee datasets waarin verschillende gepseudonimiseerde attributen worden gebruikt niet met elkaar in verband worden gebracht.
- *Deduceerbaarheid:* deductieve aanvallen om de werkelijke identiteit van een betrokkene te achterhalen zijn mogelijk in een dataset of tussen verschillende databases die hetzelfde gepseudonimiseerde attribuut voor een persoon gebruiken, of wanneer de pseudoniemen doorzichtig zijn en de oorspronkelijke identiteit van de betrokkene niet naar behoren verhullen.

4.2. Vaak gemaakte fouten

- Aannemen dat een gepseudonimiseerde dataset anoniem is: voor de verwerking verantwoordelijken gaan er vaak van uit dat het verwijderen of vervangen van een of meer attributen voldoende is om de dataset anoniem te maken. Talloze praktijkgevallen bewijzen het tegendeel: wanneer alleen de ID wordt gewijzigd, blijft

[18] Dit hangt af van de andere attributen in de dataset en van het feit of de oorspronkelijke gegevens werden verwijderd.

de betrokkene identificeerbaar zolang de dataset nog quasi-identificatoren bevat dan wel andere waarden of attributen waarmee een persoon kan worden geïdentificeerd. Vaak kan een persoon in een gepseudonimiseerde dataset even gemakkelijk worden geïdentificeerd als met de oorspronkelijke gegevens. De dataset kan alleen als anoniem worden beschouwd wanneer extra stappen worden ondernomen, bijvoorbeeld attributen wegnemen en generaliseren, de oorspronkelijke gegevens verwijderen of op zijn minst samenvoegen tot op een hoog aggregatieniveau.

- Vaak gemaakte fouten bij gebruik van pseudonimisering als techniek om koppelbaarheid te verminderen:

    - Dezelfde sleutel gebruiken in verschillende databases: de mogelijkheid om verschillende datasets van elkaar los te koppelen hangt in sterke mate af van het gebruik van een coderingsalgoritme alsook van het feit of verschillende gepseudonimiseerde attributen in diverse contexten herleidbaar zijn tot een enkele persoon. Het is dus van belang in verschillende databases niet dezelfde sleutel te gebruiken om de koppelbaarheid te verminderen.
    - Verschillende sleutels („roulerende sleutels") gebruiken voor verschillende gebruikers: het kan aanlokkelijk zijn verschillende sleutels te gebruiken voor verschillende groepen gebruikers en de sleutel op gebruiksbasis te veranderen (bijvoorbeeld dezelfde sleutel gebruiken om 10 informatie-elementen over dezelfde gebruiker vast te leggen). Wanneer deze bewerking echter onzorgvuldig wordt uitgevoerd, kunnen patronen ontstaan, waardoor de beoogde voordelen deels tenietgaan. Wanneer de sleutel bijvoorbeeld op roulerende basis wordt gebruikt overeenkomstig bijzondere voorschriften voor specifieke personen, kan gemakkelijker een verband worden gelegd tussen persoonsgebonden informatie-elementen. Ook het feit dat een terugkerend gepseudonimiseerd gegeven in de database verdwijnt wanneer een nieuw gegeven verschijnt, kan erop wijzen dat beide records naar dezelfde natuurlijke persoon verwijzen.
    - De sleutel bewaren: wanneer de geheime sleutel samen met de gepseudonimiseerde gegevens wordt bewaard en de gegevens raken gecompromitteerd, dan kan de aanvaller de gepseudonimiseerde gegevens eenvoudig in verband brengen met het oorspronkelijke attribuut. Dat is ook het geval wanneer de sleutel afgezonderd van de gegevens wordt bewaard, maar niet op een veilige manier.

4.3. Tekortkomingen van pseudonimisering

- *Gezondheidszorg*

1. Naam, adres en geboortedatum	2. Tijdvak van bijzondere bijstandsuitkering	3. Body Mass Index (BMI)	6. Referentienummer cohortonderzoek
[redacted]	< 2 jaar	15	QA5FRD4
	> 5 jaar	14	2B48HFG
	< 2 jaar	16	RC3URPQ
	> 5 jaar	18	SD289K9
	< 2 jaar	20	5E1FL7Q

Tabel 5. Voorbeeld van pseudonimisering door een eenvoudig terug te rekenen hashfunctie (naam, adres en geboortedatum)

Een dataset werd aangelegd om het verband te onderzoeken tussen het gewicht van een persoon en een bijzondere bijstandsuitkering. In de oorspronkelijke dataset stonden de naam, het adres en de geboortedatum van de betrokkene, maar die gegevens werden verwijderd. Het referentienummer van het cohortonderzoek werd met behulp van een hashfunctie gegenereerd op basis van de verwijderde gegevens. Ook al zijn de naam, het adres en de geboortedatum uit de tabel verwijderd, de referentienummers van het cohortonderzoek zijn eenvoudig terug te rekenen wanneer de naam, het adres en de geboortedatum van de betrokkene bekend zijn en daarnaast ook geweten is welke hashfunctie werd gebruikt.

- *Sociale netwerken*

Aangetoond werd[19] dat het op sociale netwerken mogelijk is gevoelige informatie over specifieke personen te extraheren uit de relatiegrafiek (*social graph*), zelfs wanneer op die gegevens „pseudonimiseringstechnieken" werden toegepast. De aanbieder van een sociaal netwerk ging er ten onrechte van uit dat pseudonimisering voldoende privacyveilig was om identificatie te voorkomen, en besloot daarop gegevens aan andere ondernemingen te verkopen voor marketing- en reclamedoeleinden. In plaats van echte namen gebruikte de aanbieder bijnamen (aliassen). Dit bleek echter onvoldoende om de gebruikersprofielen anoniem te maken in de mate dat de unieke interpersoonlijke relaties kunnen dienen als identificator.

- *Locatiegegevens*

Onderzoekers van MIT[20] analyseerden recentelijk een gepseudonimiseerde dataset bestaande uit de over een periode van 15 maanden verzamelde ruimtelijke en temporele mobiliteitscoördinaten van 1,5 miljoen mensen op een grondgebied met een straal van 100 km. Daaruit bleek dat 95 % van de populatie individualiseerbaar was op basis van vier locatiepunten, en dat slechts twee locatiepunten voldoende waren om ruim 50 % van de betrokkenen te individualiseren (waarbij een van beide punten bekend verondersteld wordt, zijnde naar alle waarschijnlijkheid „thuis" of „kantoor"). Er was dus zeer weinig ruimte

[19] A. Narayanan en V. Shmatikov, „De-anonymizing social networks" in *30th IEEE Symposium on Security and Privacy*, 2009.
[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen en V. Blondel, „Unique in the Crowd: The privacy bounds of human mobility" in *Nature* nr. 1376, 2013.

voor privacybescherming, ook al werden de identiteitsgegevens van de personen gepseudonimiseerd door hun echte attributen te vervangen door andere labels.

5. Conclusies en aanbevelingen

5.1. Conclusies

Er wordt volop onderzoek verricht naar technieken om gegevens niet-identificeerbaar of anoniem te maken. In voorliggend advies werd op consistente wijze aangetoond dat aan elke techniek voor- en nadelen verbonden zijn. In de meeste gevallen is het niet mogelijk minimale aanbevelingen op te stellen voor de te gebruiken parameters. Elke dataset moet immers per geval worden bekeken.

Vaak is er voor betrokkenen nog een restrisico verbonden aan een geanonimiseerde dataset. Immers, ook al kan de record van een persoon niet meer als zodanig worden opgehaald, toch blijft de mogelijkheid bestaan om informatie over die persoon te verzamelen met gebruikmaking van andere (al dan niet publiekelijk) beschikbare informatiebronnen. Benadrukt zij dat een gebrekkig anonimiseringsproces niet alleen directe gevolgen heeft voor de betrokkenen (ergernis, tijdverlies en het gevoel de controle te verliezen door onbewust of zonder voorafgaande toestemming in een cluster te worden opgenomen), maar ook indirecte bijeffecten kan veroorzaken wanneer een betrokkene als gevolg van de verwerking van geanonimiseerde gegevens onbedoeld het doelwit van een aanval wordt. Dat is des te meer het geval wanneer de aanvaller kwaadwillige bedoelingen heeft. Daarom benadrukt de Groep dat anonimiseringstechnieken uitsluitend privacywaarborgen kunnen bieden wanneer ze naar behoren worden toegepast. Het is derhalve zaak om de randvoorwaarden (contextuele situatie) en de doelstelling(en) van het anonimiseringsproces duidelijk vast te stellen teneinde de beoogde mate van anonimiteit te bewerkstelligen.

5.2. Aanbevelingen

- Sommige anonimiseringstechnieken hebben inherente beperkingen. De voor de verwerking verantwoordelijken moeten terdege rekening daarmee houden voordat ze deze of gene techniek toepassen om een anonimiseringsproces op te zetten. Zij moeten zich rekenschap geven van de met de anonimisering nagestreefde doeleinden, zoals het beschermen van de privacy van personen wanneer een dataset openbaar wordt gemaakt of het vrijgeven van een informatie-element uit een dataset.
- Geen van de in dit advies uiteengezette technieken beantwoordt met zekerheid aan de drie criteria voor een doeltreffende anonimisering, namelijk dat het niet mogelijk mag zijn een persoon te individualiseren (herleidbaarheid), persoonsgebonden records met elkaar in verband te brengen (koppelbaarheid) en persoonsgegevens af te leiden (deduceerbaarheid). Niettemin kan deze of gene techniek sommige van die risico's geheel of ten dele ondervangen. Het is derhalve zaak om zorgvuldig af te wegen hoe een op zichzelf staande techniek kan worden toegepast in de specifieke situatie die aan de orde is. Voorts moet worden bekeken of een combinatie van die technieken ertoe kan bijdragen het resultaat beter bestand te maken tegen privacyschendingen.

In de onderstaande tabel staat een overzicht van de sterke en zwakke punten van de desbetreffende technieken door toetsing aan de drie basisvereisten:

	Bestaat er nog risico van herleidbaarheid?	Bestaat er nog risico van koppelbaarheid?	Bestaat er nog risico van deduceerbaarheid?
Pseudonimisering	Ja	Ja	Ja
Ruistoevoeging	Ja	Vermoedelijk niet	Vermoedelijk niet
Substitutie	Ja	Ja	Vermoedelijk niet
Aggregatie of *k*-anonimiteit	Nee	Ja	Ja
L-diversiteit	Nee	Ja	Vermoedelijk niet
Differentiële privacy	Vermoedelijk niet	Vermoedelijk niet	Vermoedelijk niet
Hashen/tokeniseren	Ja	Ja	Vermoedelijk niet

Tabel 6. Sterke en zwakke punten van de anonimiseringstechnieken

- De optimale oplossing moet per geval worden gekozen. Een oplossing (dat wil zeggen een alomvattend anonimiseringsproces) die beantwoordt aan de drie criteria, is in voldoende mate bestand tegen identificatie op basis van de meest waarschijnlijke en redelijke middelen die inzetbaar zijn door de voor de verwerking verantwoordelijke of enige derde.
- Wanneer een voorstel niet aan een van deze drie criteria voldoet, moeten de risico's van identificatie zorgvuldig worden geëvalueerd. De uitkomst van deze evaluatie moet kenbaar worden gemaakt aan de bevoegde instantie indien de nationale wetgeving voorschrijft dat die het anonimiseringsproces moet toetsen of goedkeuren.

Om de risico's van identificatie terug te dringen, moeten de volgende goede praktijken worden toegepast:

<u>Goede praktijken inzake anonimisering</u>

Algemeen:

- Elke benadering die neerkomt op „vrijgeven en vergeten” is af te raden. Gelet op het restrisico van identificatie moeten de voor de verwerking verantwoordelijken:
    - 1. Nieuwe risico's in kaart brengen en het (de) restrisico('s) periodiek opnieuw bekijken;
    - 2. Beoordelen of de controlemiddelen voor de onderkende risico's toereikend zijn, en die waar nodig aanpassen; EN
    - 3. De risico's monitoren en beheersen.
- In het kader van dergelijke restrisico's moet aandacht worden besteed aan het identificerende gehalte van het eventuele niet-geanonimiseerde segment van een dataset, met name in combinatie met het geanonimiseerde segment, alsook aan mogelijke correlaties tussen attributen (bijvoorbeeld gegevens over de geografische locatie en het welvaartspeil).

Contextuele factoren

- De doeleinden die met de geanonimiseerde dataset worden nagestreefd, moeten duidelijk worden vastgesteld omdat ze van het allergrootste belang zijn om het risico van identificatie in te schatten.
- Tegelijk moeten ook alle relevante contextuele factoren in overweging worden genomen, bijvoorbeeld de aard van de oorspronkelijke gegevens, de opgezette controlemechanismen (met inbegrip van beveiligingsmaatregelen ter beperking van de toegang tot de datasets), steekproefgrootte (kwantitatieve kenmerken), beschikbaarheid van openbare informatiebronnen (voor de ontvangers van de gegevens), geplande vrijgave van gegevens aan derden (al dan niet met beperkingen, bijvoorbeeld op het internet enzovoort).
- De nodige aandacht moet worden besteed aan eventuele aanvallers door na te gaan in hoeverre de gegevens doelgerichte aanvallen kunnen uitlokken (wat dat betreft, zijn de gevoeligheid van de informatie en de aard van de gegevens ook hier van fundamenteel belang).

Technische factoren

- De voor de verwerking verantwoordelijken moeten bekendmaken welke anonimiseringstechniek dan wel combinatie van technieken wordt toegepast, met name wanneer zij van plan zijn de geanonimiseerde dataset vrij te geven.
- Doorzichtige (bijvoorbeeld zeldzame) attributen/quasi-identificatoren moeten uit de dataset worden verwijderd.
- Wanneer bij randomisatie technieken voor ruistoevoeging worden toegepast, moet het aan de records toegevoegde ruisniveau worden bepaald als functie van een attribuutwaarde (er mag bijgevolg geen buitenproportionele ruis worden toegevoegd), het effect voor de betrokkenen van de te beveiligen attributen en/of de spreiding van de dataset.
- Wanneer bij randomisatie differentiële privacy wordt toegepast, moet rekening worden gehouden met de noodzaak om gegevensopvragingen (query's) bij te houden. Query's die de privacy schenden, kunnen dankzij deze traceerbaarheid worden opgespoord. Dat is van belang omdat het inbreukmakende karakter van query's cumulatief is.
- Worden generalisatietechnieken toegepast, dan heeft de voor de verwerking verantwoordelijke er alle belang bij meerdere generalisatiecriteria toe te passen, zelfs voor een en hetzelfde attribuut. Er dient met andere woorden meer dan één detailniveau („granulariteit”) voor locaties of meer dan één tijdsinterval te worden geselecteerd. De verdeling van de attribuutwaarden in de populatie moet bepalend zijn voor de keuze van het toe te passen criterium. Niet alle verdelingen zijn generaliseerbaar. Anders gezegd, bij generalisatie kan geen algemeen geldende benadering („one-size-fits-all”) worden gevolgd. De variabiliteit binnen equivalentieklassen moet worden gewaarborgd, bijvoorbeeld door een specifieke drempelwaarde te selecteren op basis van de hierboven genoemde „contextuele factoren” (steekproefgrootte enzovoort). Wordt de vastgestelde drempelwaarde niet bereikt, dan moet de steekproef in kwestie buiten beschouwing worden gelaten (of moet een ander generalisatiecriterium worden gehanteerd).

BIJLAGE
Handreiking anonimiseringstechnieken

A.1. Inleiding

EU-breed wordt het begrip anonimiteit anders geïnterpreteerd: in sommige lidstaten wordt daaronder verstaan computationele anonimiteit (het moet reken- of computertechnisch moeilijk zijn, zelfs voor de voor de verwerking verantwoordelijke met de hulp van een andere partij, een van de betrokkenen direct of indirect te identificeren); in andere lidstaten wordt dit begrip opgevat als volkomen anonimiteit (het moet onmogelijk zijn, zelfs voor de voor de verwerking verantwoordelijke met de hulp van een andere partij, om een van de betrokkenen direct of indirect te identificeren). Niettemin verwijst „anonimisering" in beide betekenissen naar het proces waarmee gegevens anoniem worden gemaakt. Het verschil heeft te maken met hetgeen aanvaardbaar wordt geacht in termen van het risico dat de identiteit opnieuw wordt vastgesteld (= re-identificatie).

Diverse praktijkgevallen zijn denkbaar waarin gebruik wordt gemaakt van geanonimiseerde gegevens, zoals sociale enquêtes/opinieonderzoeken, statistische analyses en de ontwikkeling van nieuwe diensten/producten. Soms kunnen dergelijke activiteiten van algemene aard zelfs gevolgen hebben voor specifieke betrokkenen, waardoor het schijnbaar anonieme karakter van de verwerkte gegevens tenietgaat. Daar zijn talloze voorbeelden van, gaande van de opzet van doelgerichte marketinginitiatieven tot de tenuitvoerlegging van publiekrechtelijke maatregelen op basis van profilering, gedragingen of mobiliteitspatronen van gebruikers[21].

Helaas bestaan er afgezien van algemene verklaringen geen beproefde maatstaven om op voorhand in te schatten hoeveel tijd of moeite nodig is om de identiteit opnieuw vast te stellen na de verwerking of, omgekeerd, om de meest aangewezen procedure te selecteren met het doel de kans te verkleinen dat in een vrijgegeven database verwijzingen naar een geïdentificeerde groep betrokkenen voorkomen.

De „kunst van het anonimiseren", zoals deze praktijken soms worden genoemd in wetenschappelijke publicaties[22], is een nieuwe tak van de wetenschap die nog in de kinderschoenen staat. Er bestaan diverse praktijken om het identificerende gehalte van datasets te verminderen. Niettemin moet het duidelijk zijn dat het merendeel van die praktijken niet belet dat de verwerkte gegevens in verband wordt gebracht met de betrokkenen. In bepaalde gevallen werden personen met succes geïdentificeerd in anoniem geachte datasets, in andere situaties was sprake van foutpositieven.

Algemeen beschouwd bestaan er twee verschillende benaderingen: de eerste berust op het generaliseren van attributen, de tweede op het randomiseren. Door de details en eigenheden van die praktijken van naderbij te bekijken, kunnen we nieuwe inzichten verwerven in het identificerende gehalte van gegevens, en het eigenlijke begrip persoonsgegeven in een ander licht stellen.

A.2. Anonimisering door randomisatie

Een mogelijkheid om gegevens anoniem te maken, bestaat erin de werkelijke waarden te wijzigen teneinde de geanonimiseerde gegevens los te koppelen van de oorspronkelijke

[21] Bijvoorbeeld de zaak TomTom in Nederland (zie de toelichting in punt 2.2.3).

[22] Jun Gu, Yuexian Chen, Junning Fu, HuanchunPeng, Xiaojun Ye, „Synthesizing: Art of Anonymization, Database and Expert Systems Applications Lecture Notes" in *Computer Science*, Springer, vol. 6261, 2010, blz. 385-399.

waarden. Er bestaat een breed scala aan methoden om dat doel te bereiken, gaande van ruistoevoeging tot permutatie van gegevens. Het zij benadrukt dat het verwijderen van een attribuut een extreme vorm van randomisatie is, waarbij het attribuut in kwestie volledig wordt gemaskeerd of afgeschermd door ruis.

In bepaalde omstandigheden wordt met de algehele verwerking niet zozeer de vrijgave van een gerandomiseerde dataset beoogd, maar veeleer het toegankelijk maken of ontsluiten van gegevens door middel van query's. In dit geval bestaat het risico voor de betrokkene erin dat een aanvaller informatie kan extraheren uit een reeks verschillende query's zonder dat de voor de verwerking verantwoordelijke daar weet van heeft. Om te waarborgen dat de in de dataset opgenomen personen anoniem blijven, mag het niet mogelijk zijn daaruit op te maken dat een betrokkene heeft bijgedragen aan de dataset. Op die manier kan geen verband worden gelegd met welke achtergrondkennis dan ook waarover een aanvaller beschikt.

Het risico van re-identificatie kan verder worden teruggedrongen door waar nodig ruis toe te voegen aan het antwoord op de query. Deze benadering, in vakliteratuur ook differentiële privacy[23] genoemd, wijkt af van de hierboven toegelichte benaderingen doordat de uitgevers van informatie meer controle over de gegevenstoegang hebben dan het geval is bij openbaarmaking. Met het toevoegen van ruis worden twee belangrijke doelstellingen nagestreefd: enerzijds de privacy van betrokkenen in de dataset beschermen en anderzijds ervoor zorgen dat de vrijgegeven informatie bruikbaar blijft. Zo moet de mate van ruis meer bepaald in verhouding staan tot de intensiteit van gegevensopvragingen (wanneer te veel op de persoon gerichte query's accuraat worden beantwoord, verhoogt de kans dat personen worden geïdentificeerd). Er bestaat momenteel geen enkele techniek die een volkomen privacyveilige methode waarborgt. Daarom moet de geslaagde toepassing van randomisatie per geval worden beoordeeld. Er zijn gevallen bekend waarin informatie over de persoonskenmerken van een (al dan niet in de dataset opgenomen) betrokkene is gelekt, terwijl de voor de verwerking verantwoordelijke ervan uitging dat de dataset naar behoren gerandomiseerd was.

Het kan nuttig zijn specifieke voorbeelden te bespreken om te verduidelijken welke tekortkomingen verbonden zijn aan randomisatie als anonimiseringstechniek. In het kader van interactieve toegang bijvoorbeeld, kunnen als privacyvriendelijk beschouwde query's een risico inhouden voor de betrokkenen. Stel bijvoorbeeld dat de aanvaller weet dat een subgroep S van personen is opgenomen in de dataset met informatie over het voorkomen van attribuut A binnen populatie P. In dat geval is het mogelijk door een gewone query bestaande uit twee vragen, namelijk „hoeveel personen in populatie P bezitten attribuut A?" en „hoeveel personen in populatie P, behalve de personen die behoren tot subgroep S, bezitten attribuut A?", (differentiërend) te bepalen hoeveel personen in subgroep S daadwerkelijk attribuut A bezitten, zowel op deterministische wijze als door deductieve kansrekening. De privacy van de personen in subgroep S kan hoe dan ook ernstig in het gedrang komen. De ernst van de privacybedreiging hangt meer in het bijzonder af van de aard van attribuut A.

Ook wanneer een betrokkene niet in de dataset is opgenomen, maar zijn relatie met gegevens in de dataset bekend is, kan de vrijgave van de dataset een ernstige bedreiging vormen voor zijn privacy. Stel bijvoorbeeld dat het geweten is dat „de attribuutwaarde A van het doelwit met hoeveelheid X verschilt van het gemiddelde van de populatie". In dit geval kan de aanvaller door de databasebeheerder gewoon te vragen een privacyvriendelijke bewerking uit

[23]Cynthia Dwork, „Differential Privacy", *International Colloquium on Automata, Languages and Programming* (ICALP), 2006, blz. 1-12.

te voeren, zoals het gemiddelde berekenen van attribuut A, nauwkeurig een persoonsgegeven van de betrokkene in kwestie afleiden.

Wanneer onnauwkeurigheden in de werkelijke waarden van een database worden ingevoegd, komt het erop aan bijzonder zorgvuldig te werk te gaan. Er moet voldoende ruis worden toegevoegd om de privacy te beschermen, maar niet zo veel dat de gegevens onbruikbaar worden. Zijn er bijvoorbeeld zeer weinig betrokkenen die een bijzonder persoonskenmerk (attribuut) bezitten of gaat het om een bijzonder gevoelig attribuut, dan verdient het aanbeveling om in plaats van het werkelijke aantal een bereik op te geven of een algemene zin te formuleren, zoals „een klein aantal gevallen, mogelijk zelfs nihil". De onzekerheid die daardoor ontstaat, zorgt ervoor dat de privacy van de betrokkene gevrijwaard blijft, zelfs wanneer de aanvaller op voorhand weet dat er ruis aan de gepubliceerde gegevens werd toegevoegd. Wat de bruikbaarheid van de gegevens betreft, is het van belang de onnauwkeurigheid naar behoren in te voeven om ervoor te zorgen dat de gegevens nog kunnen dienen voor statistische doeleinden of voor de besluitvorming.

Er moet verder worden nagedacht over randomisatie van databases en gegevenstoegang via differentiële privacy. Eerst en vooral kan de juiste mate van vervorming aanzienlijk variëren afhankelijk van de context (soort gegevensopvragingen/query's, populatiedichtheid in de database, aard van het attribuut en identificerend gehalte dat daaraan inherent is). Bijgevolg bestaat er geen algemeen geldende oplossing. Bovendien kan de context mettertijd veranderen, zodat het interactieve mechanisme dienovereenkomstig moet worden aangepast. Om ruis te kalibreren, moet worden bijgehouden aan welke cumulatieve privacyrisico's het interactieve mechanisme de betrokkenen blootstelt. Wanneer de „privacydrempel" is bereikt en de betrokkenen specifieke risico's lopen als gevolg van een nieuwe query, moet in het mechanisme voor gegevenstoegang een waarschuwingssysteem worden opgenomen dat de voor de verwerking verantwoordelijke helpt bepalen in welke mate de werkelijke persoonsgegevens telkens moeten worden vervormd.

Ook de verwijdering (of wijziging) van attribuutwaarden verdient aandacht. Een veelgebruikte oplossing om sommige atypische attribuutwaarden te behandelen, bestaat erin de gegevensverzameling betreffende de atypische personen of de atypische waarden zelf te verwijderen. In dit laatste geval is het van belang ervoor te zorgen dat het ontbreken van de waarde in se geen mogelijkheid biedt om een betrokkene te identificeren.

Randomisatie kan ook ertoe strekken attributen te vervangen (substitutie). Een bekend misverstand is het gelijkstellen van anonimisering met encryptie (codering) of versleuteling. Aan deze misvatting liggen twee aannamen ten grondslag. De eerste aanname is dat de record als „geanonimiseerd" wordt beschouwd zodra sommige attributen van een databaserecord zijn gecodeerd (bijvoorbeeld naam, adres of geboortedatum) of zodra deze attributen na een versleutelingsbewerking, zoals een *keyed*-hashfunctie, zijn vervangen door een schijnbaar gerandomiseerde tekenreeks. De tweede aanname is dat anonimisering efficiënter is wanneer de sleutel voldoende lang en het encryptiealgoritme voldoende geavanceerd is. Met name onder voor de verwerking verantwoordelijken is dit een wijdverbreid misverstand dat verduidelijking verdient, temeer omdat dit ook opgaat voor pseudonimisering en de schijnbaar lagere privacyrisico's daarvan.

In de eerste plaats beogen deze technieken totaal andere doeleinden: als beveiligingsmaatregel is encryptie erop gericht de vertrouwelijkheid te waarborgen van een communicatiekanaal tussen geïdentificeerde partijen (mensen, apparaten of hardware/software) teneinde te voorkomen dat er wordt afgeluisterd of dat informatie onbedoeld openbaar wordt gemaakt. Bij versleuteling worden de gegevens semantisch vertaald volgens een geheime sleutel.

Anonimisering daarentegen beoogt te beletten dat personen worden geïdentificeerd door te vermijden dat attributen heimelijk in verband worden gebracht met een betrokkene.

In wezen zijn codering en versleuteling echter niet geschikt om een betrokkene niet-identificeerbaar te maken: de oorspronkelijke gegevens blijven namelijk beschikbaar of herleidbaar zolang de voor de verwerking verantwoordelijke die in handen heeft. Wanneer persoonsgegevens alleen semantisch worden vertaald, zoals bij versleuteling, blijft het mogelijk de oorspronkelijke gegevensstructuur te herstellen door het algoritme terug te draaien of via brutekrachtaanvallen naargelang van de aard van de encryptie- of coderingssystemen, of nog door gegevenslekken. Ook al kunnen geavanceerde coderingstechnieken gegevens beter beschermen door ze onbegrijpelijk te maken voor entiteiten die de decoderingssleutel niet kennen, ze resulteren niet noodzakelijkerwijs in anonimisering. Zolang de sleutel of de oorspronkelijke gegevens beschikbaar zijn (zelfs bij een vertrouwde derde partij (TTP - *Trusted Third Party*) die contractueel verplicht is op te treden als bewaarnemer (*escrow agent*) van geheime sleutels), blijven betrokkenen identificeerbaar.

Wordt de sterkte van het encryptiemechanisme als enige maatstaf gehanteerd om te bepalen in hoeverre een dataset „anoniem gemaakt" is, dan kan een verkeerd beeld ontstaan: de betrouwbaarheid van een coderingsmechanisme of hashfunctie hangt immers ook af van tal van andere technische of organisatorische factoren. In de literatuur zijn heel wat geslaagde aanvallen bekend waarbij het algoritme volledig werd omzeild door zwakke plekken in de sleutelbewaring (bijvoorbeeld het bestaan van een minder veilige methode voor gegevenstoegang) of andere menselijke factoren (zoals zwakke wachtwoorden voor sleutelafleiding) uit te buiten. Tot slot zij opgemerkt dat een encryptiesysteem met een bepaalde sleutelgrootte ontworpen is om de vertrouwelijkheid slechts gedurende een bepaalde periode te waarborgen (de grootte van de meeste huidige sleutels zal omstreeks 2020 moeten worden aangepast), terwijl een anonimiseringsproces niet in de tijd mag worden beperkt.

Het is de moeite waard dieper in te gaan op de beperkingen die eigen zijn aan randomisatie (of substitutie en verwijdering) van attributen. Nuttige aanwijzingen kunnen worden verkregen door te kijken naar diverse slechte voorbeelden uit een recent verleden, waarbij randomisatie als anonimiseringstechniek werd gebruikt, en door na te gaan waarom dat mislukt is.

Een welbekend geval waarin een slecht geanonimiseerde dataset werd vrijgegeven, betreft de Netflix-prijswedstrijd[24]. Elke generieke record in een database waarvan een aantal attributen van een betrokkene werd gerandomiseerd, kan nog steeds in de volgende twee subrecords worden opgesplitst: {gerandomiseerde attributen, niet-gerandomiseerde attributen}, waarbij niet-gerandomiseerde attributen elke combinatie van schijnbaar niet-persoonsgebonden gegevens zijn. Uit de dataset van de Netflix-prijswedstrijd komt duidelijk naar voren dat elke record kan worden vertegenwoordigd door een punt in een meerdimensionale ruimte, waarbij het niet-gerandomiseerde attribuut een coördinaat is. Volgens deze techniek is elke dataset te beschouwen als een puntenverzameling in die meerdimensionale ruimte met een hoge spreidingsgraad, wat betekent dat de punten van elkaar verwijderd kunnen zijn. Die afstand kan dermate groot zijn dat na het opdelen van de ruimte in grote gebieden, elk gebied slechts één record bevat. Zelfs na het toevoegen van ruis worden de records niet dicht genoeg bij elkaar gebracht om datzelfde meerdimensionale gebied te delen. In het Netflix-experiment bijvoorbeeld waren de records voldoende eenduidig voor slechts 8 filmrecensies binnen een

[24] Arvind Narayanan, Vitaly Shmatikov, „Robust De-anonymization of Large Sparse Datasets", *IEEE Symposium on Security and Privacy*, 2008, 111-125.

interval van 14 dagen. Nadat ruis werd toegevoegd aan de filmrecensies (waarderingscijfers) en datums, waren er geen overlappende gebieden meer. Met andere woorden, dezelfde selectie van slechts 8 gerecenseerde films leverde een vingerafdruk op van de toegekende waarderingscijfers die niet werden gedeeld door twee betrokkenen in de database. Op basis van deze geometrische observatie vergeleken onderzoekers de schijnbaar anonieme Netflix-dataset met een andere openbare database met filmrecensies (IMDb - *Internet Movie Database*), en slaagden ze erin gebruikers te vinden die dezelfde films binnen hetzelfde tijdsinterval hadden gerecenseerd. Aangezien voor de meeste gebruikers een één-op-éénovereenkomst bestond, kon de uit de IMDb-database opgehaalde aanvullende informatie worden geïmporteerd in de vrijgegeven Netflix-dataset om zo de schijnbaar geanonimiseerde records identificeerbaar te maken.

Het moet worden beklemtoond dat dit een algemene eigenschap is: het overige segment van een „gerandomiseerde" database heeft nog een zeer hoog identificerend gehalte naarmate de combinatiemogelijkheden tussen de andere (niet-gerandomiseerde) attributen zeldzamer zijn. De voor de verwerking verantwoordelijken moeten zich rekenschap daarvan geven wanneer zij randomisatie kiezen als anonimiseringstechniek.

Bij talloze re-identificatie-experimenten werd een soortgelijke benadering gevolgd waarbij twee databases in dezelfde subruimte worden geprojecteerd. In een recent verleden werd deze uiterst krachtige re-identificatiemethode veelvuldig toegepast op uiteenlopende gebieden. Zo werd bijvoorbeeld bij een identificatie-experiment in een sociaal netwerk[25] gebruikgemaakt van de relatiegrafiek (*social graph*) van de gebruikers, gepseudonimiseerd met behulp van labels. In dit geval bestonden de identificerende attributen uit een lijst met contacten/vriendschapsrelaties van elke gebruiker omdat de kans uiterst klein is dat twee personen een identieke lijst met contacten/vriendschapsrelaties hebben. Uit deze intuïtieve veronderstelling bleek dat een subgrafiek bestaande uit interne verbindingen tussen een zeer klein aantal knooppunten een topologische vingerafdruk vormt die in het netwerk verborgen is, en dat na het identificeren van dit subnetwerk een groot deel van het volledige sociale netwerk kan worden blootgelegd. De resultaten van een soortgelijke aanval kunnen als volgt in cijfers worden uitgedrukt: met minder dan 10 knooppunten (die een miljoen verschillende subnetwerkconfiguraties opleveren, waarbij elk subnetwerk een topologische vingerafdruk vormt) kan een sociaal netwerk met ruim 4 miljoen gepseudonimiseerde knooppunten en 70 miljoen verbindingen het doelwit vormen van re-identificatieaanvallen. Dit betekent dat de privacy van een groot aantal verbindingen in gevaar kan komen. Benadrukt zij dat deze re-identificatiebenadering niet alleen geldt in de specifieke context van sociale netwerken, maar in het algemeen kan worden aangepast voor andere databases waar relaties tussen gebruikers worden vastgelegd (bijvoorbeeld telefooncontacten, e-mailverkeer, datingsites enzovoort).

Een andere manier om een schijnbaar anonieme record te identificeren bestaat erin de schrijfstijl (stylometrie)[26] te analyseren. Er zijn al algoritmen ontwikkeld die metrieke gegevens extraheren uit syntactisch ontlede tekst, waaronder de gebruiksfrequentie van specifieke woorden, het voorkomen van specifieke grammaticale patronen en het soort leestekens (interpunctie). Deze eigenschappen kunnen allemaal worden gebruikt om een schijnbaar anonieme tekst in verband te brengen met de schrijfstijl van een geïdentificeerde auteur. Onderzoekers hebben de schrijfstijl van ruim 100 000 blogs uitgelicht en zijn vandaag in staat de auteur van een gepubliceerd artikel automatisch te identificeren met een

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

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/

nauwkeurigheid van bijna 80 %. De verwachting is dat deze techniek nog accurater zal worden wanneer andere signalen worden benut, zoals locatie- of andere metagegevens in de tekst.

Het identificerende gehalte dat eigen is aan de semantiek van een record (dat wil zeggen het niet-gerandomiseerde recordsegment) moet meer aandacht krijgen in de onderzoekswereld en de industrie. In een recent verleden (2013)[27] werden de identiteiten van DNA-donoren teruggerekend. Daaruit blijkt dat weinig vooruitgang is geboekt sinds het welbekende AOL-incident (2006): toen werd een database met 20 miljoen zoektermen van ruim 650 000 gebruikers over een periode van 3 maanden openbaar gemaakt, met als gevolg dat een aantal AOL-gebruikers werd geïdentificeerd en gelokaliseerd.

Locatiegegevens vormen een andere categorie van gegevens die zelden anoniem kunnen worden gemaakt door alleen de identiteiten van de betrokkenen te verwijderen of door sommige attributen gedeeltelijk te coderen. Aangezien de mobiliteitspatronen van mensen in voldoende mate uniek zijn, kan het semantische segment van locatiegegevens (de plaatsen waar de betrokkene zich op een specifiek tijdstip bevond), zelfs zonder andere attributen, uiteenlopende kenmerken van een betrokkene onthullen[28]. Dat werd bij herhaling aangetoond in representatief academisch onderzoek[29].

Wat dat betreft, is voorzichtigheid geboden bij het gebruik van pseudoniemen om betrokkenen afdoende te beschermen tegen het lekken van identiteitsgegevens en attributen. Wanneer pseudonimisering erin bestaat een identiteit te vervangen door een andere unieke code, is het naïef te veronderstellen dat de informatie daarmee op afdoende wijze niet-identificeerbaar is gemaakt. Daardoor wordt immers voorbijgegaan aan de complexiteit van identificatiemethoden en de veelsoortige contexten waarin die toepassing vinden.

A.3. Anonimisering door generalisatie

De op het generaliseren van attributen gebaseerde benadering kan aan de hand van een eenvoudig voorbeeld worden verduidelijkt.

Stel dat een voor de verwerking verantwoordelijke beslist een eenvoudige tabel vrij te geven waarin drie informatie-elementen of attributen staan: een identificatienummer dat uniek is voor elke record, een locatie-ID die de betrokkene koppelt aan zijn woonplaats, en een eigenschapsidentificatie die een kenmerk van de betrokkene aangeeft. Gesteld verder dat die eigenschap een van twee verschillende waarden kan aannemen, algemeen aangeduid met {E1, E2}:

---

[27] Genetische gegevens zijn het voorbeeld bij uitstek van gevoelige gegevens die re-identificatie van donoren mogelijk maken wanneer de „anonimisering" er uitsluitend in bestaat de identiteiten van donoren te verwijderen. Zie het hierboven in punt 2.2.2 aangehaalde voorbeeld. Zie ook John Bohannon, „Genealogy Databases Enable Naming of Anonymous DNA Donors" in *Science*, vol. 339, nr. 6117, 18 januari 2013, blz. 262.

[28] Sommige nationale wetgevers hebben dit probleem aangepakt. In Frankrijk bijvoorbeeld worden gepubliceerde locatiestatistieken anoniem gemaakt door generalisatie en permutatie. Zo publiceert het Franse bureau voor de statistiek (INSEE) statistieken die worden gegeneraliseerd door alle gegevens samen te voegen in een gebied dat 40 000 m² bestrijkt. Toch is het detailniveau van de dataset voldoende groot om de gegevens bruikbaar te houden. Permutaties voorkomen de-anonimiseringsaanvallen in gespreide gebieden. Algemener bieden de aggregatie en permutatie van deze categorie van gegevens sterke waarborgen tegen deductieve en de-anonimiseringsaanvallen (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. en Blondel, V.D., „Unique in the Crowd: The privacy bounds of human mobility" in *Nature Scientific Reports* 3, nr. 1376, 2013.

Reeks-ID	Locatie-ID	Eigenschap
Nr. 1	Rome	E1
Nr. 2	Madrid	E1
Nr. 3	Londen	E2
Nr. 4	Parijs	E1
Nr. 5	Barcelona	E1
Nr. 6	Milaan	E2
Nr. 7	New York	E2
Nr. 8	Berlijn	E1

Tabel A1. Steekproef van betrokkenen volgens locatie en eigenschappen E1 en E2

Wanneer iemand (de aanvaller) van tevoren weet dat in de tabel een specifieke betrokkene (het doelwit) is opgenomen die in Milaan woont, dan kan hij daaruit afleiden dat de enige betrokkene met die locatie-ID, zijnde persoon nr. 6, ook eigenschap E2 bezit.

Dit zeer elementaire voorbeeld illustreert de belangrijkste onderdelen van elke identificatieprocedure die wordt toegepast op een schijnbaar geanonimiseerde dataset. Er is namelijk een aanvaller die (per ongeluk of opzettelijk) beschikt over achtergrondkennis in verband met sommige of alle betrokkenen in een dataset. De aanvaller probeert die achtergrondkennis in verband te brengen met de gegevens in de vrijgegeven dataset om zich een beter beeld te vormen van de kenmerken van die betrokkenen.

Om ervoor te zorgen dat de gegevens minder doeltreffend of minder direct in verband kunnen worden gebracht met achtergrondkennis, kan de voor de verwerking verantwoordelijke zich toespitsen op de locatie-ID en de stad waar de betrokkene woont vervangen door een ruimer gebied, bijvoorbeeld het land. De tabel ziet er dan als volgt uit:

Reeks-ID	Locatie-ID	Eigenschap
Nr. 1	Italië	E1
Nr. 2	Spanje	E1
Nr. 3	Verenigd Koninkrijk	E2
Nr. 4	Frankrijk	E1
Nr. 5	Spanje	E1
Nr. 6	Italië	E2
Nr. 7	Verenigde Staten	E2
Nr. 8	Duitsland	E1

Tabel A2. Tabel A1 na generalisatie volgens nationaliteit

Met deze opnieuw geaggregeerde gegevens kan de aanvaller aan de hand van zijn achtergrondkennis over een geïdentificeerde betrokkene (bijvoorbeeld „het doelwit woont in Rome en staat in de tabel") geen duidelijke conclusie trekken over de eigenschap van die betrokkene. Dat komt doordat de twee Italianen in de tabel verschillende eigenschappen bezitten, respectievelijk E1 en E2. Voor de aanvaller betekent dit dat de eigenschap van het doelwit voor 50 % onvoorspelbaar is. Dit eenvoudige voorbeeld illustreert welke gevolgen generalisatie heeft op de anonimiseringspraktijk. Deze generalisatieslag halveert weliswaar de kans dat een Italiaans doelwit wordt geïdentificeerd, maar blijft zonder uitwerking voor een doelwit op andere locaties (bijvoorbeeld in de Verenigde Staten).

Bovendien kan een aanvaller nog steeds informatie herleiden over een Spaans doelwit. Is de achtergrondkennis van het type „het doelwit woont in Madrid en staat in de tabel" of „het doelwit woont in Barcelona en staat in de tabel", dan kan de aanvaller met 100 % zekerheid afleiden dat het doelwit eigenschap E1 bezit. De generalisatie levert derhalve niet dezelfde mate van privacybescherming of weerstand tegen deductieve aanvallen op voor de volledige populatie in de dataset.

Wordt de hierboven gevolgde redenering doorgetrokken, dan zou men tot de conclusie kunnen komen dat koppelbaarheid te voorkomen is door de gegevens verdergaand te generaliseren, bijvoorbeeld volgens continent. In dit geval ziet de tabel er als volgt uit:

Reeks-ID	Locatie-ID	Eigenschap
Nr. 1	Europa	E1
Nr. 2	Europa	E1
Nr. 3	Europa	E2
Nr. 4	Europa	E1
Nr. 5	Europa	E1
Nr. 6	Europa	E2
Nr. 7	Noord-Amerika	E2
Nr. 8	Europa	E1

Tabel A3. Tabel A1 na generalisatie volgens continent

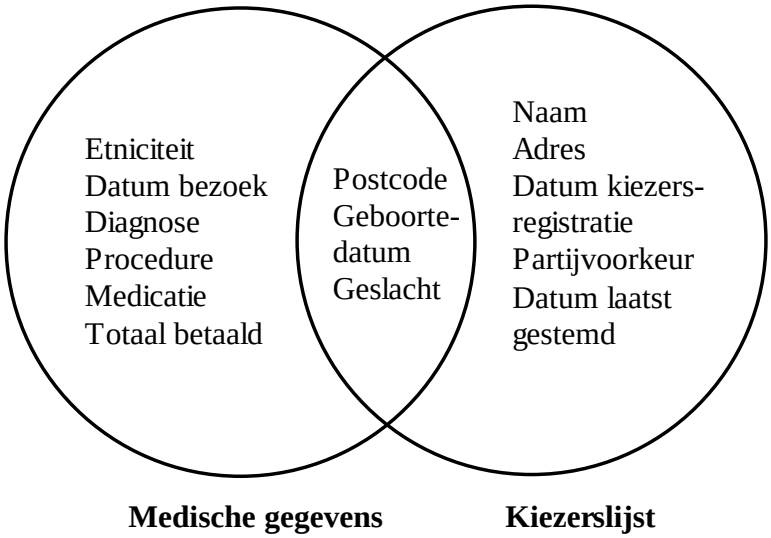
Na deze aggregatie zijn alle in de tabel opgenomen betrokkenen, uitgezonderd de betrokkene die in de Verenigde Staten woont, beschermd tegen koppelings- en identificatieaanvallen. Uit achtergrondkennis van het type „het doelwit woont in Madrid en staat in de tabel" of „het doelwit woont in Milaan en staat in de tabel" kan de aanvaller met een zekere mate van waarschijnlijkheid opmaken welke eigenschap van toepassing is op de betrokkene (E1 met een waarschijnlijkheid van 71,4 % en E2 met een waarschijnlijkheid van 28,6 %) zonder evenwel een direct verband te kunnen leggen. Bovendien gaat deze verdere generalisatie ten koste van een duidelijk en wezenlijk verlies aan informatie: uit de tabel komen geen correlaties naar voren tussen eigenschappen en locatie. Er valt met andere woorden niet uit te maken of een van beide eigenschappen met een grotere waarschijnlijkheid samenhangt met een specifieke locatie. De tabel maakt alleen de zogenaamde „marginale" verdelingen zichtbaar, zijnde de absolute kans dat eigenschappen E1 en E2 voorkomen in de gehele populatie (respectievelijk 62,5 % en 37,5 % in dit voorbeeld) en op het gehele continent (zoals gezegd, respectievelijk 71,4 % en 28,6 % in Europa en 100 % en 0 % in Noord-Amerika).

Uit dit voorbeeld blijkt tevens dat de generalisatiepraktijk invloed heeft op de praktische bruikbaarheid van gegevens. Vandaag zijn er technische hulpmiddelen voorhanden om van tevoren (dat wil zeggen alvorens een dataset vrij te geven) vast te stellen in hoeverre attributen moeten worden gegeneraliseerd om het risico te verminderen dat de in de tabel opgenomen betrokkenen worden geïdentificeerd, zonder in verregaande mate afbreuk te doen aan de bruikbaarheid van de vrijgegeven gegevens.

K-anonimiteit

Met *k*-anonimiteit wordt beoogd koppelingsaanvallen op basis van gegeneraliseerde attributen te voorkomen. Deze praktijk is ontstaan uit een re-identificatie-experiment eind de jaren negentig, toen een particulier Amerikaans bedrijf in de zorgsector een schijnbaar anoniem gemaakte dataset toegankelijk maakte voor het publiek. De anonimisering bestond erin de

namen van de betrokkenen te verwijderen. Toch bevatte de dataset nog gezondheidsgegevens en andere attributen, zoals postcode (locatie-ID van hun woonplaats), geslacht en de volledige geboortedatum. Aangezien hetzelfde triplet {postcode, geslacht, volledige geboortedatum} ook was opgenomen in andere publiekelijk toegankelijke registers (bijvoorbeeld de kiezerslijst), kon een universiteitsonderzoeker de identiteit van specifieke betrokkenen in verband brengen met de attributen in de vrijgegeven dataset. De aanvaller (in casu de onderzoeker) beschikte bijvoorbeeld over de volgende achtergrondkennis: „Ik weet dat er in de kiezerslijst één enkele betrokkene staat voor het specifieke triplet {postcode, geslacht, volledige geboortedatum}. De vrijgegeven dataset bevat een record met dit triplet". Zo werd empirisch vastgesteld[30] dat het merendeel (ruim 80 %) van de betrokkenen in het voor dit onderzoeksexperiment gebruikte publieke register eenduidig gekoppeld was aan een specifiek triplet, waardoor het mogelijk werd de betrokkenen te identificeren. Derhalve waren de gegevens in casu niet naar behoren anoniem gemaakt.



Figuur A1. Re-identificatie door gegevens met elkaar in verband te brengen

Om de slaagkans van soortgelijke koppelingsaanvallen te verminderen, werd geargumenteerd dat de voor de verwerking verantwoordelijken de dataset eerst dienen te onderzoeken. Daarbij moeten ze de attributen groeperen waarvan redelijkerwijs te verwachten valt dat zij door een aanvaller worden gebruikt om de vrijgegeven tabel in verband te brengen met andere aanvullende informatie. Elke groep moet ten minste *k* identieke combinaties van gegeneraliseerde attributen bevatten (dat wil zeggen een equivalentieklasse van attributen voorstellen). De datasets mogen vervolgens alleen worden vrijgegeven nadat ze in zulke homogene groepen zijn opgedeeld. De voor generalisatie geselecteerde attributen worden in de literatuur quasi-identificatoren genoemd, aangezien de kennis ervan als zodanig de betrokkenen onmiddellijk identificeerbaar maakt.

Tal van identificatie-experimenten hebben zwakke plekken aan het licht gebracht in tabellen die slecht werden geanonimiseerd op basis van *k*-anonimiteit. Mogelijke redenen daarvoor zijn bijvoorbeeld dat de andere attributen in een equivalentieklasse identiek zijn (wat het geval is voor de equivalentieklasse van Spaanse betrokkenen in voorbeeldtabel A2) of in zeer ongelijke mate verdeeld zijn, waarbij een specifiek attribuut het overwicht heeft; dat een

[30] L. Sweeney, „Weaving Technology and Policy Together to Maintain Confidentiality" in *Journal of Law, Medicine & Ethics*, vol. 25, nr. 2 en 3, 1997, blz. 98-110.

equivalentieklasse zeer weinig records bevat, waardoor in beide gevallen de waarschijnlijkheid afleidbaar is; of nog dat er geen wezenlijk „semantisch" verschil bestaat tussen de niet-gerandomiseerde attributen van de equivalentieklassen. Zo kan de kwantitatieve meting van zulke attributen in werkelijkheid verschillen, maar cijfermatig sterk gelijklopend zijn, of kan die gelden voor een reeks semantisch gelijksoortige attributen, bijvoorbeeld dezelfde mate van kredietrisico of dezelfde groep ziekten. Het gevolg daarvan is dat uit de dataset nog heel wat informatie kan worden afgeleid door koppelingsaanvallen[31]. Dienaangaande zij erop gewezen dat wanneer gegevens dun gespreid zijn (bijvoorbeeld wanneer een specifieke eigenschap zelden voorkomt in een geografisch gebied), en een eerste aggregatie het niet mogelijk maakt gegevens dermate te groeperen dat verschillende eigenschappen een voldoende aantal keren voorkomen (bijvoorbeeld een geografisch gebied waarin sommige eigenschappen nog steeds zelden voorkomen), de gegevens verder moeten worden geaggregeerd om de nagestreefde anonimisering te bewerkstelligen.

L-diversiteit

Voortbouwend op deze waarnemingen werden mettertijd varianten op *k*-anonimiteit voorgesteld. Voorts werden ook technische criteria vastgesteld om anonimisering door generalisatie te verbeteren en het risico van koppelingsaanvallen tegen te gaan. Deze berusten op de probabilistische eigenschappen van datasets. Zo wordt meer in het bijzonder een aanvullende randvoorwaarde geïntroduceerd, te weten dat elk attribuut in een equivalentieklasse ten minste *l* keren moet voorkomen. Op die manier krijgt een aanvaller te maken met een aanzienlijke onzekerheidsfactor wat de attributen betreft, zelfs wanneer hij achtergrondkennis over een specifieke betrokkene bezit. Dit komt erop neer te stellen dat een geselecteerde eigenschap ten minste een aantal keren moet voorkomen in een dataset (of partitie) om het re-identificatierisico te verkleinen. Dit doel wordt nagestreefd door de anonimisering op basis van *l*-diversiteit. Een voorbeeld daarvan is te vinden in tabel A4 (de oorspronkelijke gegevens) en tabel A5 (het resultaat van de verwerking). Daaruit wordt duidelijk dat wanneer de locatie-ID en leeftijd van de personen in tabel A4 naar behoren worden vastgelegd, de generalisatie van attributen leidt tot een forse toename van de onzekerheidsfactor met betrekking tot de werkelijke attributen van elke betrokkene in de enquête. Zelfs wanneer de aanvaller weet dat een betrokkene deel uitmaakt van de eerste equivalentieklasse, kan hij daaruit niet opmaken of een persoon eigenschap X, Y of Z bezit , aangezien er in die klasse (en in elke andere equivalentieklasse) ten minste één record bestaat die dergelijke eigenschappen bezit.

[31]Hierbij dient te worden beklemtoond dat er ook correlaties kunnen worden gemaakt nadat de gegevensrecords op attributen werden gegroepeerd. Wanneer de voor de verwerking verantwoordelijke weet welke correlaties hij wil controleren, kan hij de meest relevante attributen selecteren. Zo zijn enquêteresultaten van Pew Research Center niet blootgesteld aan zeer gedetailleerde deductieve aanvallen en kunnen die nog goed van pas komen om correlaties te vinden tussen demografie en interesses (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx)

Reeksnummer	Locatie-ID	Leeftijd	Eigenschap
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Tabel A4. Personen gegroepeerd op locatie, leeftijd en de drie eigenschappen X, Y en Z

Reeksnummer	Locatie-ID	Leeftijd	Eigenschap
1	11*	< 50	X
4	11*	< 50	Y
9	11*	< 50	Z
10	11*	< 50	Z
5	23*	> 50	Z
6	23*	> 50	X
7	23*	> 50	Y
8	23*	> 50	Y
2	12*	< 50	X
3	12*	< 50	Y
11	12*	< 50	Z
12	12*	< 50	Z

Tabel A5. Voorbeeld van tabel A4 met *l*-diversiteit

T-gelijkenis

In het specifieke geval dat attributen in een partitie ongelijk verdeeld zijn of deel uitmaken van een kleine reeks waarden of semantische betekenissen, wordt een benadering toegepast die *t*-gelijkenis wordt genoemd. Dit is een verdere verbetering van anonimisering door generalisatie, waarbij de gegevens zodanig worden geordend dat de equivalentieklassen een optimale afspiegeling zijn van de aanvankelijke verdeling van attributen in de oorspronkelijke dataset. Dat gaat via een procedure in twee stappen die als volgt verloopt. Tabel A6 is de oorspronkelijke database met niet-gerandomiseerde records van betrokkenen, gegroepeerd op locatie, leeftijd, salaris en twee reeksen semantisch gelijksoortige eigenschappen, respectievelijk (X1, X2, X3) en (Y1, Y2, Y3), bijvoorbeeld gelijksoortige kredietrisicoklassen, gelijksoortige ziekten. Eerst wordt de tabel verwerkt op basis van de *l*-diversiteit, waarbij $l = 1$ (tabel A7), door records te groeperen in semantisch gelijksoortige equivalentieklassen met een beperkte mate van anonimisering. Vervolgens wordt de tabel verwerkt om de *t*-gelijkenis te bewerkstelligen (tabel A8) met een hogere variabiliteit binnen elke partitie. Na de tweede stap bevat elke equivalentieklasse records uit beide categorieën eigenschappen. Op te merken valt dat de locatie-ID en leeftijd een verschillende mate van gedetailleerdheid („granulariteit") hebben in de diverse processtappen. Bijgevolg kunnen voor elk attribuut andere generalisatiecriteria nodig zijn om de nagestreefde anonimisering te

bewerkstelligen. Omgekeerd houdt dit ook specifieke technische maatregelen en extra rekenwerk in voor de voor de verwerking verantwoordelijken.

Reeksnummer	Locatie-ID	Leeftijd	Salaris	Eigenschap
1	1127	29	30 000	X1
2	1112	22	32 000	X2
3	1128	27	35 000	X3
4	1215	43	50 000	X2
5	1219	52	120 000	Y1
6	1216	47	60 000	Y2
7	1115	30	55 000	Y2
8	1123	36	100 000	Y3
9	1117	32	110 000	X3

Tabel A6. Personen gegroepeerd op locatie, leeftijd, salaris en twee categorieën eigenschappen

Reeksnummer	Locatie-ID	Leeftijd	Salaris	Eigenschap
1	11**	2*	30 000	X1
2	11**	2*	32 000	X2
3	11**	2*	35 000	X3
4	121*	> 40	50 000	X2
5	121*	> 40	120 000	Y1
6	121*	> 40	60 000	Y2
7	11**	3*	55 000	Y2
8	11**	3*	100 000	Y3
9	11**	3*	110 000	X3

Tabel A7. Op *l*-diversiteit gebaseerde versie van tabel A6

Reeksnummer	Locatie-ID	Leeftijd	Salaris	Eigenschap
1	112*	< 40	30 000	X1
3	112*	< 40	35 000	X3
8	112*	< 40	100 000	Y3
4	121*	> 40	50 000	X2
5	121*	> 40	120 000	Y1
6	121*	> 40	60 000	Y2
2	111*	< 40	32 000	X2
7	111*	< 40	55 000	Y2
9	111*	< 40	110 000	X3

Tabel A8. Op *t*-gelijkenis gebaseerde versie van tabel A6

Het zij duidelijk dat de doelstelling om de attributen van betrokkenen op gefundeerde wijze te generaliseren soms niet haalbaar is voor alle records, maar slechts voor een klein aantal daarvan. Goede praktijken moeten ervoor zorgen dat in elke equivalentieklasse meerdere personen voorkomen en dat elke deductieve aanval wordt uitgesloten. Om deze benadering toe te passen, moeten de voor de verwerking verantwoordelijken hoe dan ook de beschikbare gegevens grondig onderzoeken en door combinatie nagaan welke alternatieven mogelijk zijn (bijvoorbeeld een verschillende intervalgrootte, een ander detailniveau voor locatie of leeftijd enzovoort). Anders gezegd, anonimisering door generalisatie mag niet voortvloeien uit een ruwe opzet waarbij de voor de verwerking verantwoordelijken analytische attribuutwaarden in

een record vervangen door bereiks- of intervalwaarden. Er zijn inderdaad specifiekere kwantitatieve benaderingen vereist, zoals het onderzoeken van de entropie van attributen binnen elke partitie, of het meten van de deviatie tussen de oorspronkelijke verdeling van de attributen en de verdeling in elke equivalentieklasse.