# Samenvatting

## Over Statistical Disclosure Control technieken
### Ter bescherming van persoonsgegevens in een open data context

**Achtergrond, reikwijdte en onderzoeksvragen**
De ontwikkelingen op het gebied van data - in termen van bijvoorbeeld hun volume, variëteit en snelheid - verhogen de risico's op onthulling van persoonsgegevens, ofwel dataonthulling. Enerzijds maakt de groei (in omvang) van een dataset het moeilijk om risico's bij het vrijgeven van data die verborgen zijn in de dataset (dat wil zeggen de *intrinsieke risicofactoren*) te detecteren en onder controle te krijgen. Anderzijds maakt de groei (in omvang of aantal) van andere datasets – ofwel de toename van achtergrondkennis die beschikbaar is voor andere partijen – het moeilijk om de risico's voor het vrijgeven van data te bepalen en onder controle te krijgen: hier betreft het risico's die zich kunnen voordoen bij het combineren van de data met andere datasets (dat wil zeggen de *extrinsieke risicofactoren*). Bijgevolg wordt het voor gegevensbeheerders moeilijker om hun data te openen, dat wil zeggen: hun data te delen met specifieke groepen, individuen of het publiek.
Het vrijgeven van gevoelige informatie over personen kan gebeuren wanneer persoonsgegevens worden overgedragen, opgeslagen of geanalyseerd. Mechanismen voor informatiebeveiliging, zoals dataencryptie en toegangscontrole, kunnen worden gebruikt om data tijdens transport, opslag of analyse te beschermen. Wanneer er al toegang is verkregen tot de data (zij het legitiem of onrechtmatig), is het nog steeds mogelijk om gevoelige informatie over personen onrechtmatig te onthullen (ongeoorloofd gegevensgebruik). Zelfs als direct-identificerende informatie (zoals namen) uit de data wordt verwijderd, kan iemand die (al dan niet op een legitieme of onrechtmatige wijze) toegang heeft verkregen tot die data, statistische onthullingsmethoden gebruiken om sommige data-items alsnog te identificeren, met name door andere informatiebronnen te gebruiken. De term 'burgemeester van Amsterdam' in een dataset kan bijvoorbeeld de identiteit van een persoon onthullen als men al weet wie die burgemeester is of als men dit kan achterhalen met een Google-zoekopdracht. Gegevensbeheerders kunnen op hun beurt de *Statistical Disclosure Control* (SDC-) tools gebruiken om de intrinsieke en extrinsieke risico's voor het vrijgeven van data te verkleinen.
SDC-tools zijn gericht op het elimineren van zowel direct als indirect identificeerbare informatie in een dataset, terwijl de datakwaliteit (dat wil zeggen de bruikbaarheid van de data) zo veel mogelijk wordt gehandhaafd. Direct identificerende informatie (zoals namen en burgerservicenummers) en indirect identificerende informatie (zoals de combinatie van geboortedatum, postcode en geslacht) in een dataset dragen respectievelijk bij aan de intrinsieke en extrinsieke risicofactoren. SDC-tools kunnen worden toegepast op zowel microdatasets als geaggregeerde datasets.
De reikwijdte van deze studie beperkt zich tot de SDC-tools die gericht zijn op het beschermen van microdatasets. Dit zijn datasets die informatie over individuen en individuele eenheden zoals huishoudens bevatten. Binnen deze studie houden we ons met name bezig met het beschermen van datasets uit het justitie domein voor open data doeleinden. Deze focus is gekozen omdat het Nederlandse ministerie van Justitie en Veiligheid van plan is haar open data initiatieven te intensiveren teneinde de transparantie en verantwoording te verbeteren. In deze context is het doel van de studie om SDC-tools te onderzoeken die gericht zijn op het beschermen van

microdatasets. Daartoe definiëren en behandelen we de volgende onderzoeksvragen:

1 Wat zijn de wettelijke beperkingen die relevant zijn voor op SDC-gebaseerde gegevensbescherming, in het bijzonder voor het openen van data uit het justitiedomein?
2 Wat zijn de belangrijkste functionaliteiten van beschikbare SDC-tools voor het beschermen van persoonsgegevens en het behoud van de bruikbaarheid van data?
3 Hoe kan achtergrondkennis worden verdisconteerd in de op SDC-gebaseerde bescherming van persoonsgegevens?
4 Wat zijn (andere) veelbelovende SDC-functionaliteiten of -methoden (voorgesteld in de literatuur)?

**Methodologie en resultaten**

Om de onderzoeksvragen te beantwoorden, hebben we een uitgebreide literatuurstudie uitgevoerd over de relevante onderwerpen, zoals privacy bevorderende technologieën, SDC-methoden, procedures voor gegevensbeschermingseffectbeoordeling, (nieuwe) wet- en regelgeving en open data initiatieven. Verder hebben we onze tussentijdse resultaten gepresenteerd aan verschillende (expert)groepen (zoals data-analisten, privacy-experts, trainees en hogeschoolstudenten) om de grenzen van de reikwijdte te verfijnen, relevante onderwerpen te selecteren en de resultaten en aanpak te controleren.

Voor het beantwoorden van de eerste onderzoeksvraag hebben we bovendien semigestructureerde interviews afgenomen met drie experts op het gebied van gegevensbescherming die ervaring hebben met privacywetten en -voorschriften. Verder hebben we, om de tweede onderzoeksvraag te beantwoorden, een aantal experimenten uitgevoerd om een voorlopige indicatie te krijgen van de bruikbaarheid en schaalbaarheid van de SDC-tools.

Hieronder beschrijven we in het kort de belangrijkste resultaten van het onderzoek per onderzoeksvraag.

*Over wettelijke beperkingen*

In het licht van de Algemene Verordening Gegevensbeveiliging (AVG, 2016), kunnen SDC-tools worden gebruikt om de gegevensminimalisatie, doelbinding en proportionaliteitsprincipes te realiseren. SDC-technologieën kunnen met name inzicht verschaffen in en mechanismen bieden voor (a) het transformeren van onbewerkte data, (b) het beoordelen van het nut van de onbewerkte en getransformeerde data, (c) het schatten van de onthullingsrisico's van de onbewerkte en getransformeerde data, en (d) het maken van afwegingen tussen de bruikbaarheid van de data en risico’s verbonden aan het vrijgeven van data. We concluderen dat deze op SDC-gebaseerde inzichten en SDC-mechanismen, noodzakelijk zijn voor gegevensbeheerders om aan de AVG te voldoen bij het delen en openen van hun data.

Pseudonimisering en anonimisering zijn twee belangrijke termen binnen het domein van SDC-tools. Deze termen zijn niet uniform gedefinieerd en worden op verschillende manieren gebruikt in het juridische en technologische domein. We stellen vast dat bijvoorbeeld de meeste data-anonimiseringsmechanismen in de technologische zin kunnen worden beschouwd als data-pseudonimiseringsmechanismen in de AVG-zin. Als onderdeel van de context van onze studie, gaan we in op deze terminologische verschillen.

Data uit het justitiedomein betreffen voornamelijk gevoelige persoonsgegevens (bijvoorbeeld data over strafrechtspleging en wetshandhaving). Het niet opnemen van persoonlijke (identificerende) informatie speelt een belangrijke rol – zo niet een noodzakelijke rol – bij het openen van data uit het justitie domein. Daarom onder-

zoeken we ook wanneer een dataset kan worden beschouwd als zijnde zonder persoonlijke informatie (of anoniem) conform de AVG. Hiertoe stellen we het idee van een drempel voor om de grens van data-anonimiteit te markeren. Deze drempel is in principe afhankelijk van de context (en tijd). Dat wil zeggen dat deze drempel afhankelijk is van bijvoorbeeld beschikbare technologieën en hun vooruitgang, andere beschikbare gegevensbronnen en de motivatie voor en kosten van heridentificering. Daarom kunnen risico's voor het vrijgeven van gegevens in de toekomst toenemen – gegevens die op dit moment anoniem zijn, kunnen niet-anonieme persoonsgegevens worden, omdat de drempelwaarde voor anonimiteit met de tijd toeneemt. Soms kan het drempelwaarde echter afnemen, bijvoorbeeld als de achtergrondkennis verdwijnt.

*Over de belangrijkste functionaliteiten van SDC-tools*
In deze studie hebben we drie niet-commerciële *open source software* SDC-tools onderzocht, namelijk: μ-ARGUS, ARX en sdcMicro. Enerzijds heeft het onderzoek van de tools ons in staat gesteld om (a) een inzicht te krijgen in de belangrijkste SDC-functionaliteiten, (b) hands-on ervaring op te doen met SDC-tools (door te experimenteren met deze tools), en (c) te leren van de ervaringen van de onderzoekgemeenschap en academische wereld. Anderzijds leidde het onderzoek van de SDC-tools (samen met onze literatuurstudie) ertoe dat we de SDC-tools konden karakteriseren op basis van een generiek functioneel model dat uit vier componenten bestaat:

- *datatransformatie* waarin een originele microdataset getransformeerd wordt naar een microdataset met behulp van SDC-methoden en -modellen;
- *dataonthullingsrisicometing* waarmee de onthullingsrisico's in de getransformeerde microdataset gekwantificeerd kunnen worden door middel van het in overweging te nemen van verschillende onthullingsscenario’s en mogelijke koppelingen;
- *bruikbaarheidsmeting* waarin de gegevenskwaliteit van de getransformeerde microdataset in termen van bruikbaarheid gekwantificeerd wordt; en
- *privacy-utility-evaluatie* waarmee afwegingen gemaakt kunnen worden tussen de onthullingsrisico's en bruikbaarheid van de getransformeerde microdataset.

Met behulp van het functionele model bieden we inzicht in de belangrijkste functionaliteiten van de SDC-tools, per component van het functionele model. De datatransformatie component omvat SDC-methoden (zoals verwijdering, onderdrukking, pseudonimisering, generalisatie, permutatie, perturbatie en anatomisatie) en SDC-modellen (zoals k-anonimity, l-diversity, t-closeness, k-map en δ-presence). Over het algemeen wordt een combinatie van SDC-methoden gebruikt om een SDC-model te realiseren en een combinatie van SDC-modellen wordt binnen een SDC-tool gerealiseerd. De data-onthullingsrisicometing neemt de onthullingsscenario's en aspecten van de mate van uniekheid van data-items in beschouwing. Deze data-onthullingsrisicometing omvat twee categorieën risicometingen: elementaire metingen (zoals de waarden van k en l in k-anonimity en l-diversity) en geavanceerde metingen (die op hun beurt weer steunen op het definiëren van data onthullingsscenario's, zoals het scenario van de openbare aanklager, journalist en marketeer aanvaller). De bruikbaarheidsmeting omvat algemene metingen (zoals de onderscheidingsmaatstaf) en maatregelen voor speciale doeleinden (zoals classificatiemaatstaven en maatregelen voor classificatieprestaties). De privacy-utility-evaluatiecomponent vertrouwt voornamelijk op menselijke expertise om een afweging te maken tussen de onthullingsrisico's en de bruikbaarheid van de getransformeerde microdataset op basis van de hierboven genoemde metingen.

Daarnaast stellen we een raamwerk voor om de niet-functionele aspecten van SDC-tools te onderzoeken, op basis van een bruikbaarheidsperspectief dat relevant is voor onze studie (d.w.z. voor datamanagers die meer willen weten over SDC-tools).

Dit kader omvat de volgende criteria:
1 toegankelijkheid of eenvoudige beschikbaarheid, bijvoorbeeld *open source*, gratis en platformonafhankelijk;
2 gebruiksgemak, bijvoorbeeld eenvoudige import, verwerking van gegevens, en export van gegevens, en een heldere gebruikersinterface;
3 leergemak, bijvoorbeeld beschikbaarheid en kwaliteit van documentatie, community-ondersteuning en intuïtiviteit van de tool;
4 uitbreidbaarheid, bijvoorbeeld integratiemogelijkheid met andere software, aantal actieve ontwikkelaars, recente onderhoudsactiviteiten en ondersteuning door ontwikkelaars.

Ten slotte beschrijven we een experiment voor het testen van een specifiek aspect van de prestaties - de uitvoeringstijd - van de drie onderzochte SDC- tools. Daartoe hebben we de verschillen in de functionaliteiten van de drie SDC-tools meegenomen om zo een uniforme manier te vinden om deze tools te testen. Het experiment heeft tot doel (a) praktisch uitvoerbaar te zijn en (b) zo veel mogelijk vergelijkbare tests voor deze tools te leveren. Het experiment is als volgt opgezet:
- gebruik ARX om een aantal generalisatie-instellingen te vinden, gerangschikt volgens hun datafunctionaliteit, zoals berekend door ARX;
- neem de eerste generalisatie-instelling op uit de bovenstaande lijst;

Voer ARX, µ-ARGUS en sdcMicro uit voor de gekozen generalisatie-instelling, meet hun uitvoeringstijden.

Ons onderzoek naar de functionele aspecten van de SDC-tools laat zien dat ARX relatief meer toegankelijk lijkt voor nieuwkomers en *early adopters*. Maar µ-ARGUS en sdcMicro zijn daarentegen relatief beter geschikt voor meer ervaren experts.

*Over achtergrondkennis*

Achtergrondkennis – steeds meer beschikbaar voor indringers – is een belangrijke extrinsieke risicofactor. Achtergrondkennis omvat de informatie in voor het publiek beschikbare databanken of directory's (zoals kiesregisters, telefoongidsen, handelsgidsen, registers van beroepsverenigingen), in persoonlijke en informele contacten (vanwege of via bijvoorbeeld lokale nabijheid), in sociale media; of in organisatiedatabases (beschikbaar voor, bijvoorbeeld, overheidsinstanties en commerciële bedrijven). Tijdens het SDC-proces gericht op het in kaart brengen van de attributen worden sommige kenmerken van microdatasets aangeduid als Quasi-ID's (QID's). QID's zijn attributen die indringers kunnen gebruiken om de identiteit van sommige betrokkenen, beschikbaar in externe informatiebronnen, te koppelen aan de gegevensitems in de getransformeerde microdataset. Bij het beschermen van microdatasets via SDC-tools wordt daarom de achtergrondinformatie die beschikbaar is voor indringers vastgelegd door de QID's op de juiste manier te definiëren. We merken op dat er geen universele manier is om attributen in kaart te brengen, bijvoorbeeld om QID's te definiëren. Daarom moeten gegevensbeheerders deze attribuuttoewijzing zorgvuldig uitvoeren binnen een SDC-proces om de risico's te beperken en de onthullingsniveaus op acceptabele niveaus te houden.

*Over veelbelovende SDC-functionaliteiten*

Onderzoek naar het bereik van SDC-functionaliteiten, dat gebaseerd is op het bestuderen van de drie SDC-tools en de literatuur, heeft ons in staat gesteld een visie te ontwikkelen voor het bundelen van de krachten van deze tools en voor het uitbreiden van deze tools in de toekomst. We identificeren een aantal SDC-functionali-

teiten die nuttig zijn om te worden opgenomen in (toekomstige) SDC-tools, in het bijzonder voor het beschermen van data uit het justitiedomein:

- risicobeoordeling bepaald op basis van werkelijke populatie data (bijvoorbeeld het aantal inwoners van een bepaalde leeftijdscategorie met een specifieke opleiding);
- semiautomatische datatransformatie, maar samen met de bij het proces betrokken gebruikers;
- data-anonimisering op basis van de kenmerken van data uit het justitie domein (omgaan met, bijvoorbeeld, doorlopende publicatie en locatie-afhankelijkheid).

**Discussie en vervolgonderzoek**

Gegevensbeschermingstechnologieën, in het algemeen, en SDC- tools in het bijzonder, kunnen geen 100% bescherming bieden tegen data-onthullingsrisico's. Dit kan met name worden toegeschreven aan de extrinsieke risicofactoren in de (data-) omgeving. Daarom moet men realistisch zijn over de mogelijkheden van databeschermingstechnologieën en het toepassen ervan mag geen vals gevoel van privacy geven. Aangezien er over het algemeen geen enkele oplossing is om gegarandeerde privacy te bieden, pleiten veel professionals ervoor om een op risico gebaseerde benadering voor databescherming aan te nemen, in plaats van een strikt gegarandeerde gegevensbeschermingsmethode. Dit vereist dat databescherming wordt beschouwd als een continu risicobeheerproces en niet als een eenmalige bewerking met een binaire uitkomst (resulterend in voor altijd anoniem zijn of voor altijd niet-anoniem zijn). Wij denken dat SDC-tools een essentieel onderdeel zijn van een dergelijk risicobeheerproces. Om ervoor te zorgen dat gegevensbeheerders AVG-compliant worden bij het delen en openen van hun data, dienen SDC-tools te worden opgenomen in het proces voor gegevensbeschermingseffectbeoordeling (DPIA). Zij kunnen daarmee de risico's identificeren en controleren middels dataminimalisatie, terwijl de datakwaliteit voor het doel aanvaardbaar blijft. Daarom pleiten wij er ook voor dat de SDC-tools worden gebruikt om domeinexperts te ondersteunen en dus niet te vervangen. *Samenvattend zien we het toepassen van SDC-tools als een noodzakelijke stap voor het realiseren van het zorgvuldigheidsprincipe dat vraagt om voldoende inspanningen om persoonsgegevens in een bepaalde context te beschermen*.

SDC-tools bieden een breed scala aan functionaliteiten, opties en configuratiemogelijkheden voor gegevensbeheerders. In de praktijk is het echter niet eenvoudig om deze tools te gebruiken en te configureren, juist wanneer er zo veel opties zijn om uit te kiezen. Het gebruik en de configuratie van deze tools worden nog omslachtiger en complexer als ook wordt gekeken naar de verscheidenheid van de gegevens die moeten worden beschermd en de diversiteit van de dataomgeving waarin de gegevensbescherming moet worden uitgevoerd. Verder moet men ook de parameters van SDC-tools en -methoden kunnen interpreteren en afstemmen om het besluitvormingsproces van dataminimalisatie adequaat te ondersteunen. *Daarom adviseren wij verder onderzoek te doen naar de toepassing van SDC-tools op justitiële gegevens, met name door een aantal concrete studies uit te voeren met operationele gegevens uit het justitiedomein*.

Ten slotte zien we op basis van de inzichten die in dit onderzoek de volgende mogelijkheden voor toekomstig onderzoek:

- Onderzoek naar de noodzaak en gevolgen van anonimiteit in de AVG-zin, ook bij oe verantwoordelijke voor de gegevensverwerking en voor open-data-initiatieven;
- een workflow ontwikkelen voor het in de praktijk gebruiken van een SDC-tool;
- een leidraad bieden voor de configuratie en interpretatie van SDC-parameters en -resultaten;

- een methodologie ontwikkelen voor effectieve samenwerking tussen verschillende belanghebbenden in het data-anonimiseringsproces, zodat SDC-tools effectief in de praktijk kunnen worden gebruikt;
- een aantal studies uitvoeren om de SDC-vereisten van datasets voor het justitie domein voor het delen van gegevens (inclusief het openen van gegevens) in kaart te brengen;
- het ontwikkelen van aanvullende (wettelijke) maatregelen die nodig zijn voor, tijdens en na het beschermen van gegevens met SDC-tools.