Privacy Preserving Techniques

Op verantwoorde wijze analyseren van verspreide onderzoeksdata

Onze samenleving is in toenemende mate 'datagestuurd'. Gegevens uit verschillende bronnen worden gecombineerd om tot nieuwe inzichten te komen, of om het effect van beleid te meten. Het kunnen combineren en analyseren van gegevens uit verschillende bronnen is voor zowel de wetenschap als voor de overheid en natuurlijk ook het bedrijfsleven van grote waarde.

Echter, deze gegevens zijn vaak persoonlijk van aard, waardoor het combineren van dit soort bronnen niet zonder meer verantwoord of toegestaan is. Wet- en regelgeving (denk aan de Algemene Verordening Gegevensbescherming, de AVG) stellen eisen aan het verwerken van dergelijke persoonsgegevens. Daarnaast hebben de verschillende beheerders van deze databronnen verschillende en mogelijk tegenstrijdige belangen, en is het niet noodzakelijkerwijs zo dat ze elkaar volledig vertrouwen. Dit is zeker het geval als het gaat om gevoelige dan wel waardevolle gegevens die niet zomaar met willekeurige andere partijen gedeeld mogen worden. Vandaar dat het CBS, de Rijksuniversiteit Groningen, en de Universiteit Maastricht onderzoek doen naar een verantwoorde wijze van analyseren van verspreide onderzoeksdata.

Combineren datasets CBS, RUG en UM

Het CBS levert, als het statistische bureau van Nederland met betrouwbare statistische informatie en data, inzicht in maatschappelijke vraagstukken. Zij doet dit op basis van een groeiende hoeveelheid data die uit een steeds groter aantal verschillende bronnen komt en waar het CBS wettelijk gezien toegang toe heeft.
Wetenschappelijke onderzoeksinstellingen zoals de Rijksuniversiteit Groningen en de Universiteit Maastricht verzamelen ook steeds meer en steeds gedetailleerdere informatie bij de uitvoering van wetenschappelijke onderzoeksprojecten. Denk bijvoorbeeld aan longitudinale onderzoeksprojecten als Lifelines van de RUG en het Maastrichtse CARRIER-project voor het voorspellen van coronaire hartziekte, die de gezondheid van mensen over een lange periode meten op basis van een groot aantal medische indicatoren. Voor het beantwoorden van sommige vragen is het hierbij noodzakelijk om de verzamelde onderzoeksgegevens te combineren met andere datasets, bijvoorbeeld die van het CBS. Toegang tot data van het CBS is onder strikte voorwaarden mogelijk voor universitaire instellingen.

Dilemma's bij grootschalige data-analyses

Dit roept meteen een aantal dilemma's op. Zo ligt met het verder uitbreiden van de 'spanwijdte' van de data die door het CBS en/of de universiteiten verzameld dan wel geanalyseerd wordt een inbreuk op de privacy al snel op de loer. Ook neemt het risico op 'function creep' (het gebruik van data voor andere doeleinden dan waarvoor het oorspronkelijk verzameld is) toe.
Daarnaast is het de vraag of de mogelijke inbreuk opweegt tegen de potentiële maatschappelijke waarde van dergelijke grootschalige data-analyses. Zeker nu er meer oog is voor het risico van discriminatie en uitsluiting door niet transparante, niet controleerbare, 'slimme' algoritmen die gebruikmaken van bevooroordeelde (biased) data. De te beantwoorden vraag en de keuze van de databronnen is uitdrukkelijk relevant. En ook de kwaliteit van de data die door die bronnen geleverd worden.

Toename in data en databronnen

Tot nu toe heeft het CBS deze kwesties (deels) weten te adresseren door zelf zeer strikt regie te voeren over de verzamelde data en de analyses die daar vervolgens op los worden gelaten. Data wordt door het CBS zelf opgeslagen, en op dit moment maakt het CBS haar data beschikbaar via een zogenaamde Remote Access-omgeving.

Hierbij is een grote toename in gebruik waar te nemen. En zoals hierboven al geschetst neemt de omvang van de data exponentieel toe, en worden de databronnen steeds diverser. Meer bestanden en steeds groter wordende datasets naar het CBS halen, is niet per definitie wenselijk: naast de hierboven genoemde issues eist dat steeds meer wat betreft datatransport en datamanagement, en is het noodzakelijk vertrouwen in het CBS hoog.

Degelijke technische data-infrastructuur

Een manier om hier aan tegemoet te komen, is een werkwijze waarbij de data van verschillende organisaties (afkomstige van publieke en private organisaties) bij de bron blijft, maar beschikbaar komt voor het uitvoeren van statistisch onderzoek in het kader van beleid of wetenschap met een uiterst scherp oog voor privacybescherming.
De ontsluiting van deze data vraagt om een degelijke technische data-infrastructuur, die werkt voor onderzoekers en voldoet aan de eisen van privacybescherming, informatiebeveiliging, de CBS-wet en een verwerking die voldoet aan de eisen die de AVG hieraan stelt. De samenwerking van het CBS met de Rijksuniversiteit Groningen (RUG) en de Universiteit Maastricht (UM) beoogt een dergelijke infrastructuur te ontwikkelen.

Data ontsluiten, koppelen en aggregeren

Doel van de samenwerking is om innovatieve benaderingen (technisch, organisatorisch, juridisch en anderszins) te ontwikkelen om data die zich bij verschillende bronnen bevindt, te ontsluiten, te koppelen en te aggregeren tot informatie, die niet op een andere wijze verkregen had kunnen worden. Uitgangspunt hierbij is dat de brondata bij de leverende partij blijft, en het gebruik van gegevens enerzijds antwoord geeft op de onderzoeksvraag en anderzijds plaatsvindt binnen het kader van de privacywetgeving (o.a. AVG en CBS-wet).
Daarbij geldt dat, hoewel een samenwerking in de regel gebouwd is op vertrouwen, het voor de te ontwikkelen werkwijze *niet* noodzakelijk is dat de betrokken partijen elkaars werkwijze en technologie hoeven te kennen en daarop moeten vertrouwen. De te ontwikkelen methodologie en de technische data-infrastructuur moeten de privacy en de veiligheid van het gebruik van de data tot een zo hoog mogelijk niveau ondersteunen.

Innovatie op het gebied van veilig data uitwisselen

Vanuit de techniek spelen zogenaamde Privacy Preserving Analytics hierbij een belangrijke rol: zij bieden de mogelijkheid de privacy bij de analyse van zeer gevoelige data geautomatiseerd te waarborgen. Deze methoden kunnen de juiste informatie boven tafel krijgen, waarbij het verplaatsen van de data waar mogelijk wordt beperkt en alleen die gevallen overblijven waarbij dat voor de analyse essentieel is. Nieuwe technologie maakt zo innovatie op het gebied van veilig data uitwisselen mogelijk.

Scope

We wijzen er hierbij nadrukkelijk op dat daarmee niet alle dilemma's (waarvan een aantal zijn geschetst in de inleiding) van tafel zijn: het risico van bias in de data en daarmee de uitkomsten blijft aanwezig, net zoals de vraag of de uiteindelijke analyses wel antwoord geven op de daadwerkelijke maatschappelijke vraag: “not everything that counts can be measured, and not everything that can be measured counts”.
Ook helpt het toepassen van privacy preserving analytics niet om te voldoen aan de eis van doelbinding. Doelbinding vereist dat bij het verzamelen van gegevens het doel waarvoor deze gegevens verzameld worden duidelijk en limitatief beschreven moet zijn, en dat deze gegevens niet naderhand voor andere doeleinden gebruikt mogen worden. Ook niet als deze verwerking op een verder privacy vriendelijke manier plaatsvindt.

Privacy Preserving Analytics: vijf technieken

Welke mogelijkheden zijn er, technisch gezien, voor het inrichten van een platform voor Privacy Preserving Analytics, waarbij de te analyseren (micro)data bij de verschillende bronnen zelf

opgeslagen blijft, (en waarbij de verschillende partijen elkaar maar beperkt hoeven te vertrouwen)? In dit stuk bespreken we kort de volgende benaderingen:

I. Met hulp van een trusted third party (TTP)
II. Secure multiparty computation (SMC)
III. Trusted execution environments (TEEs)
IV. Rekenen met versleutelde gegevens (homomorphic encryption)
V. Gepseudonimiseerd datadelen

Naast deze vijf benaderingen zijn er nog meer technieken in beeld zoals het toepassen van (gegeneerde) synthetische data en het inzetten van beveiligde virtuele research workspaces.

I. Met behulp van een Trusted Third Party (TTP)

De klassieke oplossing voor dit soort problemen waar verschillende samenwerkende partijen elkaar niet of maar beperkt vertrouwen, is om de hulp van een vertrouwde derde partij (in het Engels een Trusted Third Party, oftewel TTP) in te schakelen. Notarissen en banken zijn bekende voorbeelden van dergelijke partijen die succesvol hun rol hebben gevonden in het handelsverkeer. In de context van het combineren van verschillende gegevensverzamelingen voor het uitvoeren van een statistische analyse zou het betekenen dat elk van de partijen haar gegevens tijdelijk onder beheer van de TTP stelt (hetzij door een kopie te sturen, dan wel de TTP de toegang tot de lokale dataset te verlenen). De TTP kan vervolgens eenvoudig de analyse uitvoeren (omdat alle data bij één partij beschikbaar dan wel bereikbaar is) en de resultaten aan de deelnemende partijen teruggeven.
Een andere optie is dat de TTP de verschillende gegevensverzamelingen koppelt en (gefilterd) voor analyse ter beschikking stelt aan de samenwerkende partijen. Ter afsluiting verwijdert de TTP de onderliggende gegevens (of wordt de toegang tot de lokale kopie afgesloten).

Voor- en nadeel

Nadeel van deze benadering is dat privacy (alsmede de betrouwbaarheid van de resultaten overigens) staat of valt in dit geval volledig met de vraag of het vertrouwen in de TTP wel terecht is.
Voordeel is dat dit (computationeel gezien) de meest efficiënte en eenvoudig te realiseren vorm van het uitvoeren van dergelijke analyses is.

Deelnemende partij als vertrouwde partij

Een variant van deze oplossing is om één van de deelnemende partijen als vertrouwde partij aan te wijzen. Dit kan bijvoorbeeld de partij met de grootste of gevoeligste dataverzameling zijn. Deze variant wordt in de praktijk vaak toegepast, en is bijvoorbeeld de methode die nu vaak toegepast wordt in samenwerkingen met het CBS. In dat geval wordt de data door CBS gecombineerd en geanalyseerd en worden de resultaten via een veilige Remote Access-omgeving gedeeld met de andere deelnemers.
Vanwege de al eerder genoemde nadelen, en het feit dat het volledig versturen van grote datasets problematisch is, is dit niet langer een gewenste oplossing.

II. Secure Multiparty Computation (SMC)

Tamelijk recente ontwikkelingen in de cryptografie maken het mogelijk het idee van een op een TTP-gebaseerde oplossing te realiseren zonder daadwerkelijk gebruik te maken van een centrale TTP. En dus zonder daadwerkelijk één bepaalde partij te moeten vertrouwen. Dat klinkt natuurlijk tegenstrijdig, maar kan daadwerkelijk gedaan worden door gebruik te maken van zogenaamde Secure Multiparty Computation (SMC) technieken.

Gezamenlijke rol van TTP

SMC beschrijft de stappen die een willekeurige groep deelnemers moeten doorlopen (door het doen van lokale berekeningen en het uitwisselen van cryptografisch versleutelde berichten) om gezamenlijk het resultaat van een bepaalde functie te berekenen. Elk van de deelnemers heeft een deel van de invoer voor die functie. En elk van de deelnemers krijgt de garantie dat zijn of haar eigen invoer volledig privé blijft: geen van de andere deelnemers krijgt iets over deze invoer te weten (tenzij die

informatie uit het resultaat van de functie af te leiden is).
Omdat de deelnemers actief meedoen met het SMC-protocol, en omdat het protocol zelf publiek is, weten de deelnemers ook zeker dat deze privacy-garantie daadwerkelijk gewaarborgd is. Het is alsof alle deelnemers gezamenlijk de rol van de centrale TTP in virtuele zin realiseren, zonder dat deze centrale TTP daadwerkelijk bestaat, en zonder dat deze TTP (of een willekeurige andere entiteit) inzage krijgt in de privé invoer van elk van de deelnemers.

Weerbarstige praktijk

Het berekenen van een statistiek over verschillende databronnen is niets anders dan het berekenen van een, weliswaar complexe, functie met elk van deze databronnen als private invoer. Vandaar dat SMC-technieken in theorie de perfecte oplossing voor dit probleem zouden kunnen vormen. De praktijk is weerbarstiger, helaas. Ten eerste is het middels SMC berekenen van een functie zeer complex. Om ervoor te zorgen dat geen van deelnemende partijen iets te weten komt over de invoer van één van de andere deelnemers, moet iedere invoer versleuteld worden, maar wel op een zodanige manier dat er wel willekeurige berekeningen op uitgevoerd kunnen worden.
Ten tweede zijn de uit te voeren stappen in hoge mate afhankelijk van de te berekenen functie: grofweg komt het erop neer dat je voor iedere te berekenen statistiek een volledig nieuw SMC-programma moet schrijven (en aan de deelnemende partijen moet distribueren). Dat maakt dat SMC alleen in een zeer beperkt aantal situaties praktisch toepasbaar is.

III. Trusted Execution Environments (TEEs)

Trusted Execution Environments (TEEs) kun je grofweg zien als verzegelde computers die één specifieke taak uitvoeren, waar zelfs hun eigenaar of beheerder geen invloed op uit kan oefenen, en waarbij deze ook niet bij de tussenresultaten kan. Door ieder gebruik te maken van een TEE kunnen de deelnemers gezamenlijk de analyse uitvoeren, zonder gebruik te maken van een TTP, en zonder gebruik te maken van complexe SMC-methoden, terwijl toch de individuele databestanden van de deelnemers beschermd worden tegen misbruik door anderen.
Daar waar in een SMC-protocol de deelnemers zelf hun *eigen* gegevens beschermen door ze eerst te versleutelen voor ze gedeeld worden met de andere deelnemers (dit is overigens een grove simplificatie van hoe SMC echt werkt), zorgt een TEE ervoor dat deelnemers de gegevens van *andere* deelnemers niet kunnen misbruiken, simpelweg door de mogelijke verwerkingen die een deelnemer kan uitvoeren op de gegevens van anderen hardwarematig te beperken.

Toepassingen

TEEs worden veel gebruikt in beveiligde omgevingen, bijvoorbeeld om ervoor te zorgen dat de privésleutel van een Certification Authority alleen gebruikt kan worden om certificaten te ondertekenen. Ook worden ze gebruikt in moderne smartphones om ervoor te zorgen dat de cryptografische sleutels die gebruikt worden om gevoelige data op de telefoon te versleutelen echt alleen beschikbaar zijn als de gebruiker de telefoon ontgrendeld heeft met zijn vingerafdruk, gelaatsscan of toegangscode.

Aparte computeromgeving

De TEE creëert in feite een aparte, met behulp van hardware beveiligde, computeromgeving die enkel en alleen een op voorhand bekende en goedgekeurde berekening kan uitvoeren. Door deze berekening de gewenste statistische analyse te laten uitvoeren, en daarna de data te laten verwijderen, krijgen de andere deelnemers voldoende zekerheid dat hun gevoelige gegevens niet voor andere doeleinden misbruikt worden.

IV. Rekenen met versleutelde gegevens: Homomorphic Encryption

Een laatste techniek die we kort willen bespreken is homomorphe encryptie. Deze vorm van versleuteling laat het toe om te ‘rekenen met versleutelde gegevens’, maar in een beperktere setting dan de SMC-technieken die we hierboven besproken hebben. Met andere woorden: met homomorphe encryptie kan maar een beperkte klasse van analyses op een privacyvriendelijke manier gedaan worden, vergeleken met de algemeen toepasbare SMC-technieken.

Juiste sleutel

Stel dat we versleuteling zien als een functie die een bericht (denk in deze context even aan een getal, bijvoorbeeld het energieverbruik van een huishouden) omzet in een versleutelde variant. We kunnen dat schrijven als $E_k(m)$. Met behulp van de juiste sleutel kan dit bericht weer ontsleuteld worden tot m. Normaal gesproken is versleuteling totaal willekeurig en is er geen verband tussen $E_k(m)$, en m. Bij homomorphe encryptie is dat *wel* het geval, en kun je, gegeven $E_k(a)$ en $E_k(b)$ de versleuteling van de som $E_k(a+b)$ uitrekenen als $E_k(a) + E_k(b)$.
Als de versleutelde waarden dus het energiegebruik van een huishouden voorstellen, dan kunnen we homomorphe encrypte gebruiken om op een privacyvriendelijke manier het totale energiegebruik van alle huizen in een wijk te berekenen en door te geven aan de netwerkbeheerder of de energieleverancier, zonder dat het individueel gebruik m_i van een huishouden bekend wordt. De slimme meter in ieder huis in de wijk berekent $E_k(m_i)$. Het koppelstation in de wijk berekent $E_k(m_1) + \cdots + E_k(m_n) = E_k(m_1 + \cdots + m_n)$ en stuurt dit door aan de netwerkbeheerder of de energieleverancier. Deze heeft de sleutel om aldus het berekende totale verbruik te ontsleutelen.

Wie krijgt de sleutel?

Een vergelijkbare methode kan gebruikt worden om berekeningen uit te voeren op metingen in verschillende databestanden (bijvoorbeeld om ze bij elkaar op te tellen, maar complexere berekeningen zijn ook mogelijk) zonder dat de individuele metingen bij andere partijen bekend raken. Wel moet hierbij goed nagedacht worden over de vraag welke partij of partijen de sleutel hebben om het eindresultaat te ontsleutelen. Dit probleem speelt niet bij SMC.

V. Gepseudonimiseerd data delen

Een laatste vorm van data delen die in deze context ook relevant is, is door gebruik te maken van pseudonimisering. De achterliggende gedachte is dat in veel gevallen het grootste probleem met het analyseren van data uit verschillende bronnen is dat deze privacygevoelig zijn. Met andere woorden, de individuele data-items in deze bronnen zijn te herleiden tot individuen, en daarmee een privacy-risico. Door de koppeling tussen het data-item en de persoon te verbreken, door middel van pseudonimisering, voordat het data item met een anderen gedeeld wordt, kan het privacy-risico beperkt worden.
Daartoe moet dan wel een veilige vorm van pseudonimisering gebruikt worden, en moet ook zeker zijn dat na het koppelen van verschillende bronnen de uiteindelijk dataverzameling niet toch eenvoudig te deanonymiseren is. Dit is vaak veel lastiger dan op het eerste gezicht lijkt: veel gepseudonimiseerde databestanden blijken later toch identificeerbare gegevens te bevatten.

Conclusie

We hebben een aantal mogelijke technieken besproken die kunnen helpen om data die zich bij verschillende bronnen bevindt, te ontsluiten, te koppelen en te aggregeren tot informatie, die niet op een andere wijze verkregen had kunnen worden. Maar zoals hierboven ook al is aangegeven worden daarbij niet alle problemen die met een dergelijke koppeling van databestanden kunnen ontstaan opgelost. Zaken als het bewaken van doelbinding kunnen niet alleen door middel van technische maatregelen worden bewerkstelligd. Daar zijn ook, mogelijk nieuwe, juridische en data governance benaderingen voor nodig. In de oplossingen die het samenwerkingsverband onderzoekt, worden ook deze juridische en procedurele aspecten integraal meegenomen.

Statistical disclosure control

Daarnaast zal ook de nodige aandacht besteed worden aan zogenaamde ‘output controle’-technieken. Deze statistical disclosure control-technieken worden traditioneel toegepast om te voorkomen dat de publicatie van geaggregeerde statistieken berekend over een verzameling gevoelige microdata ongewild toch details over bepaalde data-items bekend maken. Bijvoorbeeld omdat de gepubliceerde statistiek (denk bijvoorbeeld aan ‘gemiddeld inkomen’) berekend wordt over een dusdanig kleine groep personen dat toch met enige zekerheid een goede inschatting van het inkomen van één persoon gemaakt kan worden.
Het probleem is dat normaal gesproken statistical disclosure control toegang tot of kennis over de onderliggende microdata vereist om een betrouwbare inschatting te maken van de risico’s

die verbonden zijn met het publiceren van bepaalde statistische gegevens, of om maatregelen te nemen (bijvoorbeeld door bepaalde waarden weg te laten of af te ronden) om deze risico's te beperken. Nader onderzoek naar methoden om dit mogelijk maken in een setting waarbij de microdata zich bij verschillende bronnen bevindt, en de toegang tot deze data noodzakelijkerwijs beperkt is, is nodig.

Intersectiebepaling

Tenslotte is er het probleem van intersectiebepaling. De bovenstaande oplossingen gaan er namelijk van uit dat elke partij weet over welke datasubjecten zij data moet delen, oftewel wat de intersectie is tussen de partijen. Maar die informatie an sich kan al privacy-aspecten hebben. Als bijvoorbeeld data van kankerkliniek Maastro Clinic wordt gecombineerd met CBS-data dan leert CBS dat zij kanker hebben. Al met al is deze samenwerking tussen het CBS, Universiteit Maastricht, en de Rijksuniversiteit Groningen daarom van wezenlijk belang.

Contact

Wilt u meer weten over de samenwerking, neem dan contact op met

- RUG: Hans van Gestel – **j.p.w.van.gestel@rug.nl**
- UM: Marc Dolman – **m.dolman@maastrichtuniversity.nl**
- CBS: Paul Grooten – **pg.grooten@cbs.nl**

De auteur van dit artikel dr. J.H. (Jaap-Henk) Hoepman is als universitair hoofddocent verbonden aan de vakgroep IT Recht van de Rijksuniversiteit Groningen.

Betrokken instellingen

Het is voor de RUG essentieel dat het delen van data op een ethische en verantwoorde wijze plaatsvindt. Bijvoorbeeld met het UMCG, een van de grootste medische onderzoeksinstituten van Europa dat nauw gelieerd is aan de RUG. Dankzij de ontwikkeling van privacy-vriendelijke vormen van datadelen en data-analyse wordt het mogelijk om medische data beter te benutten en tegelijkertijd de privacy te waarborgen. Het Centrum voor Informatie Technologie speelt hierbij belangrijke rol. Er zijn al diverse oplossingen ontwikkeld op het gebied van pseudonimisatie, beveiligde werkomgevingen, datavirtualisatie, dataportalen en dergelijke.

Het CBS doet continu onderzoek naar nieuwe technologieën om data te benutten voor statistische doeleinden. Door de sterke toename van databronnen en databronhouders en de eisen aan privacy en security, lukt het niet altijd om alle data op een centrale plek te bundelen.
Vanuit de wettelijke taak van het CBS om statistische informatievoorziening van overheidswege te bevorderen, kan het CBS opgedane kennis en kunde inzetten om een data-infrastructuur te helpen realiseren die bestaat uit een stelsel van afspraken en technische voorzieningen.
De samenwerking met de RUG en UM draagt bij aan de totstandkoming van een verbeterde data-infrastructuur in Nederland voor statistisch gebruik.

Voor de UM is een veilig dataplatform van belang voor het analyseren van data van zowel het CBS als andere bronnen waaraan deze data is gelinkt, zoals de data die wordt verzameld binnen het Maastricht Universiteit Medisch Centrum (MUMC+). De omvang van de populatie, alsook het brede scala van aspecten waarover het CBS data verzamelt, maken het mogelijk om complexere modellen te leren en te corrigeren voor de inherente bias in datasets. Aangezien research data zoals medische gegevens om privacy-redenen niet zomaar gedeeld kunnen worden, zijn methodes nodig die het FAIR maken van data vergemakkelijken. Daarnaast kan de impact van onderzoek worden vergroot door het gebruik van aanvullende databronnen.