Big data in sociologisch onderzoek[1]

Voorbeelden, kansen en uitdagingen

Rense Corten

SOC 15 (1): 45–68
DOI: 10.5117/SOC2019.1.003.CORT

Abstract

References to 'big data' have by now become commonplace in both the public debate and in scientific research, and also in sociological research new forms of large scale digital data are increasingly important. This article aims to provide an overview of the role and relevance of such new forms of data for sociological research. Using examples, I illustrate that these data allow for answering existing sociological questions in new ways, but also trigger new questions. Next, I sketch the opportunities that new digital data offer for the development of new methods and theory, as well as the most important drawbacks of big data. Finally, I argue that, so far, sociology as a discipline has failed to take full advantage of the opportunities that new digital data offer, and as a result risks losing relevance as compared to other field as well as private commercial parties. I conclude with some suggestions to alleviate this situation.

Keywords: big data, digitalisering, sociale netwerken, methoden

Inleiding: wat zijn 'big data' in sociologisch onderzoek?

'Big data' zijn al lang niet meer 'the next big thing'; zowel in het publieke debat als in wetenschappelijk onderzoek is de term inmiddels gemeengoed geworden, en is het besef doorgedrongen dat de ongeëvenaarde stortvloed aan (veelal digitale) gegevens die we 'big data' noemen een grote impact heeft op de maatschappij. Ook aan de (Nederlandse) sociologie gaat deze ontwikkeling niet voorbij, zoals geïllustreerd door het groeiende aantal

opleidingen dat aandacht besteed aan 'data science' of 'computational social science' of door het Actualiteitencollege dat de NSV in 2018 aan dit thema wijdde. In deze bijdrage probeer ik een beeld te schetsen van hoe big data momenteel in sociologisch onderzoek gebruikt worden, welke kansen er mijns inziens nog liggen en wat daarbij de grootste uitdagingen zijn. Hiervoor is het echter belangrijk eerst vast te stellen wat we eigenlijk precies bedoelen met 'big data'.

Hoewel de term 'big data' veel gebruikt wordt, is er momenteel geen eenduidige en algemeen geaccepteerde definitie voorhanden. Een zeer pragmatische definitie is dat data 'big' zijn wanneer ze niet in het werkgeheugen van een typische desktopcomputer passen, en er dus speciale krachtige hardware (en mogelijk speciale software) nodig is om de data te kunnen analyseren (Lazer en Radford 2017). Hoewel praktisch toepasbaar, is deze definitie duidelijk nogal tijdgebonden: gezien de snelle technologische ontwikkelingen zijn datasets die enkele jaren geleden nog als 'big' golden volgens deze definitie inmiddels te analyseren met een gewone huis-tuin-en-keuken-laptop.

Een populair alternatief is om big data te definiëren aan de hand van de zogenaamde 'drie V's': *Volume*, *Variety* en *Velocity* (Katal, Wazid en Goudar 2013; Khurshid et al. 2018)[2]. Ofwel: big data zijn groot in omvang, zijn gevarieerd in inhoud en komen in hoog tempo binnen. Het nadeel van deze definitie is dat zij circulair is (wat is 'groot'?).

Een veel inhoudelijker en meer sociologisch gemotiveerde definitie is die van McFarland, Lewis en Goldberg (2016), die big data definiëren als 'The in-context documentation of individual participant behavior, preferences, and attributes'. Hoewel deze definitie inhoudelijk goed aansluit op de praktijk van sociologisch big data-onderzoek (zoals we zullen zien), is deze definitie weer zo breed dat zij in principe ook van toepassing zou zijn op veel 'traditioneel' empirisch sociologisch onderzoek, dat we doorgaans toch niet als big data zouden classificeren. In plaats van te zoeken naar een scherpe definitie kunnen we ook eerst eens kijken naar enkele voorbeelden van typen data die doorgaans als big data beschouwd worden, en die mogelijk interessant zijn voor sociologisch onderzoek.

Netwerken en uitingen op sociale media. Een belangrijk deel van de Nederlandse bevolking (en van de wereldwijde bevolking; Cabañas, Cuevas en Cuevas 2018; Facebook 2019) maakt gebruik van sociale media als Facebook, Instagram en Twitter. Dit levert grootschalige data op over niet alleen allerlei soorten relaties tussen mensen, maar ook over inhoudelijke uitingen die gebruikers delen met hun contacten: politieke meningen,

media- en cultuurconsumptie, belangrijke gebeurtenissen in de levensloop, et cetera.

Registerdata in openbaar bestuur. Overheden en overheidsdiensten op allerlei niveaus verzamelen data over burgers en bedrijven om aan hun wettelijke taken te kunnen voldoen. Het gaat hierbij om gegevens uit de Gemeentelijke Basisadministratie, gegevens over inkomens en toeslagen verzameld door de Belastingdienst, uitkeringsgegevens bij UWV, gegevens over wetsovertredingen bij politie en justitie, et cetera.

Commerciële databases. Ook bedrijven verzamelen op grote schaal gegevens over hun klanten, en dat gebeurt in toenemende mate digitaal. Denk bijvoorbeeld aan gegevens over koopgedrag bij online winkels op basis van klantenkaarten en cultuurconsumptie bij online streamingdiensten als Netflix en Spotify, maar ook aan verzekeringsgegevens (inclusief zorggebruik) en gegevens over financiële transacties bij banken.

Hoge resolutie (sensor)data van smartphones. Smartphones verzamelen doorlopend data over hun gebruikers, niet alleen in termen van online activiteit, maar ook door middel van ingebouwde sensoren: locatiegegevens via GPS en GSM-masten en spraakgegevens bij het voeren van telefoongesprekken.

Corpora van tekst. Door de voortschrijdende digitalisering komen in toenemende mate grote corpora van geschreven tekst beschikbaar uit media, cultuur, wetenschap, et cetera. Voorbeelden zijn Google Books of digitale krantenarchieven.

Grootschalige veldexperimenten. Een minder in het oog springende maar niettemin vanuit sociologisch perspectief zeer relevante ontwikkeling is het gebruik van grootschalige veldexperimenten, die meestal online worden uitgevoerd met duizenden of zelfs miljoenen deelnemers (Bond et al. 2012; Centola 2010). Hoewel niet iedereen deze vorm van onderzoek direct als big data zal herkennen, is de schaal op zijn minst veel groter dan in voorgaand experimenteel onderzoek.

Deze voorbeelden hebben een aantal kenmerken gemeen die ze onderscheiden van sociologisch onderzoek zoals we dat gewend zijn. Ten eerste zijn ze *grootschalig*, in de zin dat het aantal observaties meestal veel groter is dan tot dan toe gebruikelijk bij onderzoek naar soortgelijke onderwerpen, en in sommige gevallen gaat het zelfs om datasets die een hele samenleving beslaan. Ten tweede gaat het tegelijkertijd vaak ook om zeer gedetailleerde data op het individuele niveau, waarbij vooral belangrijk is dat individuele gedragingen of kenmerken doorlopend in plaats van steekproefsgewijs worden gemeten ('always on'). Ten derde gaat het

in veel gevallen om *spontaan gedrag,* dat niet door de onderzoeker aangespoord is met bijvoorbeeld vragenlijsten. In die zin gaat het vaak ook om data die niet *reactief* zijn, in de zin dat onderzochte individuen hun gedrag niet veranderen als gevolg van het onderzoek (in deze context wordt ook wel gesproken van 'gevonden' of 'toevallige' data). Ten vierde is dergelijk onderzoek vaak opvallend *goedkoop* (hoewel soms verre van gratis), in ieder geval in vergelijking met meer 'traditionele' benaderingen van soortgelijke onderwerpen. Zo kan men met behulp van sociale media netwerken onderzoeken op een schaal die met vragenlijstonderzoek onbetaalbaar zou zijn, of met behulp van mobiele telefoons de welvaartsverdeling van een land in kaart brengen met een nauwkeurigheid die vergelijkbaar is met die van (veel duurder) vragenlijstonderzoek (Blumenstock, Cadamuro en On 2015). Tot slot gaat het om dataverzamelingen die relatief *snel* zijn, met name in vergelijking met conventioneel vragenlijstonderzoek.

Samenvattend kunnen we concluderen dat de term 'big data' op zich weinig behulpzaam is bij het beschrijven van het diffuse geheel aan nieuwe datavormen dat geassocieerd wordt met begrippen als big data, data science of computational social science: enerzijds is onduidelijk wanneer data precies 'groot' zijn, en anderzijds dekt alleen 'groot' ook nauwelijks de lading. Liever spreek ik dan ook, vrij naar Lazer en Radford (2017), van 'nieuwe vormen van digitale data'. In het vervolg van dit artikel zal ik beide begrippen echter door elkaar gebruiken.

In dit artikel zal ik pogen een beeld te geven van de kansen die deze nieuwe vormen van data bieden voor het beantwoorden van sociologische vragen, wat er op dit gebied al gebeurt in sociologisch onderzoek, welke invloed dit mogelijk heeft op toekomstig onderzoek en wat mogelijke problemen en uitdagingen zijn. Een voorbehoud is daarbij op zijn plaats: om pragmatische redenen zal ik in mijn overzicht veelvuldig putten uit eigen onderzoek, en het is onvermijdelijk dat het geschetste beeld daar enigszins door gekleurd zal zijn. Dit artikel pretendeert dan ook zeker niet het eerste of meest complete overzicht op dit gebied te zijn (zie bijvoorbeeld Lazer et al. 2009; Lazer en Radford 2017; Shah, Cappella en Neuman 2015). Het is bovendien, gezien de stormachtige ontwikkeling van dit onderzoeksveld, onmogelijk om in een kort artikel een uitputtend overzicht te bieden. Ik verwijs de geïnteresseerde lezer dan ook graag door naar een van de vele uitstekende boeken die de afgelopen jaren op die gebied verschenen zijn, zoals Ackland (2013), González-Baillón (2017) of Salganik (2018).

Oude en nieuwe vragen

Wat kunnen we nu met deze nieuwe vormen van data in sociologisch onderzoek? Om de discussie wat te structureren maak ik hier een onderscheid tussen twee manieren om dit soort data te benutten, hoewel ander indelingen uiteraard ook mogelijk zijn. Ten eerste kunnen we proberen 'oude', *bestaande vragen* op nieuwe manieren te beantwoorden (vgl. Molina en Garib 2019). Het gaat dan om al vaak gestelde en onderzochte sociologische vragen, waar we echter om uiteenlopende redenen nog geen bevredigend antwoord op hebben. Big data kunnen dan nieuwe, en soms zelfs baanbrekende antwoorden bieden. Ten tweede roepen nieuwe vormen van data ook *nieuwe vragen* op. Hierbij gaat het, gezien de aard van de data, vaak om vragen rondom de opkomst van het internet, digitalisering en meer in het algemeen de impact van technologie op sociale processen. Van beide typen vragen zal ik het vervolg van dit artikel voorbeelden schetsen, waarbij ik zal putten uit eigen onderzoek maar ook uit toonaangevend internationaal onderzoek van collega's.

Een nieuwe antwoord op een oude vraag: bestaat het 'small world effect' echt?

Het zogenoemde 'small world effect' (kleine-wereldeffect) en de bijbehorende 'six degrees of separation' is een van die relatief zeldzame sociaalwetenschappelijke ontdekkingen die zo invloedrijk waren dat ze deel zijn geworden van het dagelijks spraakgebruik en waarvan het grote publiek daardoor inmiddels – ironisch genoeg – de sociaalwetenschappelijke oorsprong vergeten is.[3] Dit kleine-wereldeffect verwijst naar de voor velen uit het dagelijks leven herkenbare ervaring dat sociale afstanden tussen mensen vaak verrassend kort zijn. Dit effect werd voor het eerst beschreven in het werk van Robert Milgram (Korte en Milgram 1970; Milgram 1967), die met behulp van een vindingrijke op kettingbrieven gebaseerde methode liet zien dat respondenten in zijn onderzoekspopulatie op zijn hoogst zes 'handdrukken' van elkaar verwijderd waren, een resultaat dat vervolgens zijn eigen leven is gaan leiden als de bewering dat iedereen op de wereld maximaal zes stappen van elkaar verwijderd zou zijn.

Nu is de bevinding dat afstanden in netwerken kort zijn op zich niet per se verrassend; basale grafentheorie laat zien dat in netwerken met puur willekeurig gekozen verbindingen gemiddelde afstanden naar verwachting ook kort zijn. *Sociale* netwerken zijn echter niet willekeurig maar kennen een bepaalde structuur, en zijn in het bijzonder *geclusterd*, de zin dat er

overlap in relaties bestaat (iemands vrienden zijn bijvoorbeeld vaak ook met elkaar bevriend). In een netwerk met veel overlap zijn sociale afstanden naar verwachting juist erg groot. Het bijzondere aan kleine-wereldnetwerken is dan ook dat ze deze twee eigenschappen, clustering én korte afstanden, *tegelijkertijd* bezitten. Pas in de jaren negentig kwamen Watts en Strogatz (1998) met een simulatiemodel dat het ontstaan van dergelijke structuren verklaart, in een baanbrekend artikel dat vrijwel eigenhandig een nieuw en zeer invloedrijk natuurwetenschappelijk onderzoeksveld veroorzaakte ('network science'). Watts en Strogatz lieten bovendien zien dat kleine-wereldstructuren belangrijke implicaties hebben: zo verspreiden informatie en ziektes zich in zulke netwerken veel sneller dan in netwerken die slechts geclusterd zijn. Dit illustreert het belang van Milgrams vroege ontdekking.

Bij Milgrams oorspronkelijke onderzoek zijn echter wel de nodige kanttekeningen te plaatsen. Zo had het onderzoek geen betrekking op de hele wereld, maar alleen op een paar locaties binnen de Verenigde Staten. Nog belangrijker wellicht is dat een substantieel deel van de verstuurde kettingbrieven überhaubt nooit aan kwam – het beroemde 'zes stappen'-resultaat was alleen gebaseerd op de kettingbrieven die wél aankwamen. Deze en een aantal andere problemen zijn reden genoeg om Milgrams fascinerende bevinding aan een verdere toets te willen onderwerpen.

Het probleem is echter dat dit met 'conventionele' methoden van dataverzameling voor sociale netwerken lastig is. Zulk onderzoek, doorgaans gebaseerd op vragenlijsten, valt traditioneel uiteen in twee categorieën: *sociometrisch* onderzoek en onderzoek naar *egonetwerken.* In het eerste geval worden, binnen een relatief kleine, goed afgebakende groep zoals een school of een bedrijf, alle relaties tussen de individuen in die groep (bijvoorbeeld vriendschappen) in kaart gebracht (Moody 2001). In het tweede geval worden respondenten uit een steekproef die representatief is voor een bepaalde populatie over hun directe relaties (het 'egonetwerk') bevraagd. Het voordeel van de sociometrische benadering is dat daarmee netwerkpaden tussen individuen (cruciaal voor het vaststellen van het kleine-wereldeffect) zichtbaar worden, maar het nadeel is dat dit, om logistieke, praktische en financiële redenen, alleen mogelijk is binnen relatief kleine groepen, terwijl de kleine-wereldhypothese juist over complete samenleving of in ieder geval grote netwerken gaat. Omdat de egonetwerkenbenadering gebruik maakt van representatieve steekproeven maakt deze het mogelijk uitspraken te doen over netwerken op samenlevingsniveau (zie bijvoorbeeld Mollenhorst, Völker en Flap 2008), maar juist omdat het gaat

om steekproeven van losse individuen is het niet mogelijk te kijken naar netwerkpaden tussen deze individuen.

Deze problemen illustreren hoe inventief Milgrams methode was, maar hebben er ook voor gezorgd dat een striktere toets van de kleine-wereld-hypothese lang op zich heeft laten wachten. De opkomst van sociale media heeft dit echter veranderd. Op sociale-mediaplatforms zoals Facebook houden gebruikers zelf lijsten van 'vrienden' bij en ondernemen zij allerlei vormen van sociale interactie met deze vrienden. Cruciaal is dat al deze interacties, juist door de online aard van het platform, automatisch op de servers van het platform worden opgeslagen en dus gebruikt kunnen worden voor onderzoek.[4] Als bron van informatie over sociale netwerken hebben sociale-mediadata een aantal belangrijke voordelen ten opzichte van de 'conventionele' databronnen. Ten eerste hebben sociale-mediadata met sociometrische datasets gemeen dat alle relaties (als gedefinieerd in de context van sociale media) tussen gebruikers in kaart gebracht worden, wat het mogelijk maakt lengtes van netwerkpaden te meten, maar anders dan met conventionele datasets kan dit voor hele grote groepen, soms zelfs voor complete samenlevingen. Ten tweede worden sociale relaties gemeten zonder tussenkomst van een onderzoeker: gebruikers leggen spontaan hun vriendschappen vast, en in die zin maken sociale media het mogelijk direct gedrag te observeren, zonder gebruik te hoeven maken van vragenlijsten, met alle nadelen van dien (bijv. Paik en Sanchagrin 2013). Ten derde hebben sociale-mediadata potentieel een zeer hoge resolutie; in principe (hoewel deze data niet altijd beschikbaar zijn) is het mogelijk alle veranderingen in sociale interacties in detail door de tijd te volgen. Tot slot is dataverzameling via sociale media, in vergelijking met vragenlijstonderzoek, bijzonder goedkoop en snel.

Als voorbeeld van een toets van de kleine-wereldhypothese kunnen we kijken naar data van het sociale-mediaplatform Hyves, dat tussen 2004 en 2013 in Nederland actief was als netwerkplatform en in die tijd een aanzienlijke populariteit verwierf: op het hoogtepunt in 2010 had Hyves ruim tien miljoen leden, meer dan driekwart van de Nederlandse bevolking. In dat jaar maakt Hyves ook data beschikbaar voor onderzoek: Corten (2012) analyseert een compleet (geanonimiseerd) 'snapshot' van het vriendschapsnetwerk tussen de leden (zie ook Takes en Kosters 2011). Uit die analyse blijkt dat Hyves destijds inderdaad een 'kleine wereld' was: niet alleen was het netwerk met een gemiddelde clustercoëfficient (Watts en Strogatz 1998) van .18 relatief geclusterd[5], maar daarnaast bleek negentig procent van alle paren in het netwerk maximaal zeven netwerkstappen van elkaar verwijderd te zijn. Hiermee voldoet het Hyves-netwerk aan beide kenmerken van

een 'kleine wereld'. Soortgelijke resultaten zijn gevonden voor Facebook (Ugander et al. 2011), MSN (Leskovec en Horvitz 2008), en andere sociale-medianetwerken (Chun et al. 2008). Dankzij dit soort onderzoek weten we nu dat Milgrams spraakmakende ontdekking, en Watts' en Strogatz' latere theoretische interpretatie daarvan, inderdaad generaliseerbaar zijn naar complete samenlevingen en zelfs grotere verbanden (in het geval van Facebook en MSN). Tegelijkertijd genereert dergelijk onderzoek naar zeer grote sociale netwerken ook weer nieuwe vragen: zo lijken bepaalde empirische regelmatigheden die gevonden worden voor grote (niet-sociale) netwerken systematisch niet of minder op te gaan voor sociale netwerken (Corten 2012; Jackson en Rogers 2007).

Zijn zwakke banden echt diverser dan sterke banden?

Een tweede voorbeeld van hoe 'nieuwe' data kunnen helpen om 'oude' vragen beter te beantwoorden gaat over de diversiteit van *zwakke banden*: relaties in netwerken die gekenmerkt worden door relatief lage interactiefrequenties (bijvoorbeeld kennissen of verre familie). In een van de meest geciteerde artikelen uit de hedendaagse sociologie stelt Granovetter (1973) dat, ietwat tegenintuïtief, dergelijke zwakke banden voor bepaalde doeleinden nuttiger zijn dan sterke banden (zoals hechte vriendschappen). In tegenstelling tot sterke banden hebben zwakke banden namelijk de potentie sociale contexten te overbruggen: waar goede vrienden vaak in allerlei opzichten op elkaar lijken (McPherson, Smith-Lovin en Cook 2001) en in dezelfde contexten verkeren, zijn zwakke banden relatief vaker met mensen die een ander sociale achtergrond hebben en in andere kringen verkeren. Een belangrijk gevolg hiervan, volgens Granovetter, is dat individuen nuttige informatie (zoals, in zijn invloedrijke artikel, over vacatures) relatief vaker via zwakke dan via sterke banden ontvangen.[6]

Hoewel er inmiddels veel onderzoek gedaan is naar die veronderstelde 'kracht van zwakke banden' (Mouw 2003), is er relatief weinig bekend over de kenmerken van zwakke banden zelf. De reden ligt enigszins voor het hand: juist door de kenmerkende lagere interactiefrequentie van zwakke banden is het lastig deze banden te meten met conventionele vragenlijstmethoden. Het is nu eenmaal makkelijker respondenten te vragen naar hun vijf beste vrienden (dat wil zetten, sterke banden) dan naar hun vage kennissen, waarvan het aantal makkelijk tot honderden kan oplopen (Hill en Dunbar 2003). Ook voor dit probleem bieden sociale-mediadata uitkomst. Platforms zoals Facebook zijn er immers juist op ontworpen om gebruikers hun sociale relaties in kaart te laten brengen, juist ook als het om zwakkere relaties gaat.

In Hofstra, Corten, Van Tubergen en Ellison (2017) maken we van dit feit gebruik om een belangrijke implicatie van Granovetters theorie te toetsen: dat zwakke banden diverser zijn dan sterke banden, in de zin dat ze in sterkere mate mensen uit van verschillende sociale achtergronden verbinden. We kijken daarbij specifiek naar diversiteit in etniciteit onder Nederlandse adolescenten. Hiervoor combineren we data uit vragenlijstonderzoek, waarmee we de sterke banden (in dit geval: beste vrienden) van deze adolescenten meten, met gegevens over hun Facebooknetwerken, waarbij we die Facebooknetwerken dus interpreteren als meting van zwakke banden.[7] Omdat we ook de etniciteit van de beste vrienden en de Facebookvrienden meten, kunnen we de Granovetters bewering toetsen dat zwakke banden relatief vaker voorkomen tussen leden van verschillende sociale groepen dan sterke banden.

De resultaten zijn verrassend: gemiddeld genomen blijken zwakke banden, als gemeten via Facebook, vrijwel precies zo divers te zijn als netwerken tussen vrienden. Voor adolescenten uit de meerderheidsgroepering (dat wil zeggen, met een Nederlandse achtergrond) zijn zwakke banden zelfs iets vaker met mensen uit hun eigen etnische groep dan sterke banden. Voor adolescenten uit minderheidsgroeperingen (dat wil zeggen, met een migratieachtergrond) vinden we wel dat zwakke banden iets diverser zijn. Op het eerste gezicht lijkt dit in tegenspraak met Granovetters zeer invloedrijke theorie. Kijken we echter meer in detail naar wat de samenstelling van persoonlijke netwerken van adolescenten bepaalt, dan blijken niet de mechanismen die Granovetter veronderstelde het probleem, maar eerder de condities waaronder die mechanismen moeten opereren. De sociale contexten waarin adolescenten hun vrienden leren kennen (scholen) zijn namelijk zodanig gesegregeerd dat, met name voor leden van de etnische meerderheid, zwakke banden niet 'de kans krijgen' als brug tussen sociale groepen te fungeren; in veel sociale contexten zijn simpelweg onvoldoende leden van minderheidsgroeperingen voorhanden. Hiermee stellen we nieuwe randvoorwaarden aan de geldigheid van een bestaande (en invloedrijke) theorie, op een manier die zonder gebruik van digitale data erg lastig zou zijn geweest.[8] Op vergelijkbare wijze zijn ook andere hypothesen die volgen uit de theorie, zoals de voorspelling dat relaties over langere geografische afstanden zwakker zijn, op nieuwe manieren te toetsen (Park, Blumenstock en Macy 2018).

Wat is het effect van netwerkstructuur op (economisch) succes van gemeenschappen?

Ook een laatste voorbeeld van 'oude vragen op nieuwe manieren beantwoorden', opnieuw (deels) uit eigen onderzoek, neemt Granovetters (1973)

'The Strength of Weak Ties' als uitgangspunt, maar richt zich daarbij op een in de literatuur wat onderbelicht gebleven hypothese uit die theorie. Granovetter beweerde namelijk niet alleen dat individuen met meer zwakke banden succesvoller zijn, maar ook dat iets soortgelijks geldt voor *gemeenschappen*. Omdat informatie zich, volgens de theorie, efficiënter verspreidt via zwakke banden dan via sterke banden, zijn gemeenschappen met meer zwakke banden relatief beter in staat zich te organiseren, waardoor deze gemeenschappen op allerlei vlakken succesvoller zouden zijn. In zijn oorspronkelijke artikel levert Granovetter enkele voorbeelden die deze stelling ondersteunen.

Een meer grootschalige toets van deze hypothese loopt echter, met conventionele vragenlijstmethoden, snel tegen beperkingen van de data aan. Het probleem in dit geval is dat je, om deze hypothese heel precies te toetsen, niet alleen gedetailleerde informatie over netwerkenstructuren *binnen* gemeenschappen nodig hebt, maar ook vergelijkingen wilt kunnen maken *tussen* gemeenschappen, en, om statistisch betrouwbare uitspraken te kunnen doen op het niveau van gemeenschappen, liefst tussen een aanzienlijk aantal gemeenschappen. Juist deze combinatie van detail op het microniveau en variantie op het macroniveau is met vragenlijstonderzoek vaak lastig te realiseren, hoewel – volgens sommigen (bijv. Coleman 1990) – dit juist het soort vragen is waar sociologie idealiter over zou moeten gaan.

Nieuwe vormen van sociale big data bieden echter nieuwe kansen om dit type hypothesen te toetsen. In een eerste en originele poging om de relatie tussen sociaal kapitaal en succes op gemeenschapsniveau te onderzoeken maken Eagle, Macy, en Claxton (2010) gebruik van zeer grootschalige Britse data over telefoongesprekken om netwerken te reconstrueren. Ze laten daarbij zien dat de diversiteit van netwerken in lokale gemeenschappen, een concept dat dicht in de buurt komt van Granovetters (1973) ideeën over zwakke banden, inderdaad sterk samenhangt met economisch succes van deze gemeenschappen. Voor Nederland toetsten wij soortgelijke hypothesen met behulp van de eerder genoemde sociale-mediadata van het platform *Hyves*, waarbij we keken naar het verband tussen netwerkstructuur en economisch succes op gemeenteniveau (Norbutas en Corten 2017). De resultaten zijn in grote lijnen consistent met die van Eagle, Macy en Claxton (2010) en met Granovetters (1973) theorie: gemeenschappen met meer 'bruggen' in het lokale netwerk zijn economisch welvarender.

De bovenstaande voorbeelden uit eigen onderzoek maken gebruik van sociale-mediadata, maar er zijn tal van voorbeelden van onderzoek dat 'oude' vragen onderzoekt met andere vormen van 'nieuwe' data, zoals naar de sociale stratificatie van roem in gedrukte media op basis van digitale

krantenarchieven (Van de Rijt et al. 2013), naar de rol van rituelen bij flirten aan de hand spraakopnames bij speeddating (McFarland, Jurafsky en Rawlings 2013), naar cumulatief voordeel in culturele markten met grootschalig online experiment (Salganik, Dodds en Watts 2006), of naar sociale segregatie met gebruik van sensordata (GPS) uit smartphones (Silm en Ahas 2014; Toomet et al. 2015).

Nieuwe vragen

In het voorgaande hebben we voorbeelden besproken van onderzoek dat bestaande vragen op nieuwe (en hopelijk betere) manieren beantwoordt met gebruik van nieuwe vormen van data. Deze nieuwe data maken het echter soms ook mogelijk *nieuwe* vragen te stellen, die vóór de beschikbaarheid van deze datavormen niet aan de orde waren. Door de aard van de data gaat het hierbij vaak om vragen gerelateerd aan de opkomst van het internet of onlinegedrag. Een vraag die (op het moment van schrijven) bijvoorbeeld erg in de aandacht staat bij zowel sociale wetenschappers als beleidsmakers en media gaat over de rol van sociale media in het publieke debat, en dat met name de vraag of sociale media fungeren als 'echokamers' waarin gebruikers vooral in aanraking komen met uitingen die hun eigen mening reflecteren. Op zijn beurt zou dit verschijnsel politieke polarisatie kunnen vergroten en uiteindelijk sociale cohesie negatief kunnen beïnvloeden. Hoewel dit verband in het publieke debat vaak al als gegeven aangenomen wordt (Tempelman 2019), zijn de resultaten uit sociaalwetenschappelijk onderzoek nog minder eenduidig (Bakshy, Messing en Adamic 2015; Del Vicario et al. 2016, 2017).

Een gerelateerde vraag gaat over de rol van sociale media in massamobilisatie. Gebeurtenissen als de Arabische Lente tussen 2010 en 2012, de 'Project X'-rellen in Haren in 2012, of, meer recent, de protesten door 'Gele Hesjes' in met name Frankrijk, doen vermoeden dat de opkomst van sociale media het eenvoudiger hebben gemaakt om grote groepen mensen te mobiliseren voor protesten (González-Bailón et al. 2011), wat mogelijk weer impact zou kunnen hebben op de stabiliteit van samenlevingen. Een hiermee verweven vraag die in het publieke debat veel stof doet opwaaien gaat over verspreiding van 'nepnieuws' via sociale media, en de invloed daarvan op de publieke opinie en politieke processen (Allcot en Gentzkow 2017).

Een meer fundamentele[9] vraag die we in eigen onderzoek hebben bekeken gaat over het ontstaan van orde in anonieme online gemeenschappen. De opkomst van het internet maakt niet alleen economisch verkeer zonder fysiek contact veel makkelijker (ook voor particulieren onderling, zoals op eBay), maar maakt het ook mogelijk dat zulk verkeer plaatsvindt

tussen volslagen vreemden. Op het Dark Web, dat deel van het internet dat alleen toegankelijk is door middel van software dat internetverkeer geheel anonimiseert,[10] zijn bijvoorbeeld online marktplaatsen ontstaan waar illegale producten, vooral drugs, verhandeld worden door individuen wiens identiteit niet of nauwelijks te achterhalen is: niet alleen is hun internetverkeer versleuteld, maar ook betalingen verlopen buiten het reguliere bancaire systeem om door het gebruik van BitCoin of andere cryptovaluta. Het is duidelijk dat hier een vertrouwensprobleem bestaat tussen koper en verkoper: de koper moet er op vertrouwen dat de verkoper na betaling de middelen met de beloofde kwaliteit levert, maar heeft geen mogelijkheden om persoonlijk verhaal te halen mocht dat niet gebeuren. Daarbij kan de koper ook niet, zoals in de reguliere 'bovengrondse' economie, terugvallen op de wet en de overheid om zijn belangen te beschermen.

Desondanks bloeien deze marktplaatsen (Martin 2014). Antwoord op de vraag waarom leert ons mogelijk niet alleen iets over deze specifieke drugsmarktplaatsen, maar ook over de meer algemene vraag hoe samenwerking en vertrouwen ontstaan in contexten waarin men niet kan terugvallen op traditionele sociale structuren of instituties en waarin sociale orde vrijwel geheel berust op zelforganiserende verbanden. Deze bredere vraag wordt, in een tijd waarin steeds meer sociale en economische activiteiten zich verplaatsen naar het internet, steeds relevanter.

Een mechanisme dat vaak verantwoordelijk wordt gehouden voor het creëren van vertrouwen in legale online marktplaatsen zoals eBay is het gebruik van *reputatiesystemen*, die de gebruikers in staat stellen hun interactiepartners in het openbaar te beoordelen via gestandaardiseerde scores (bijvoorbeeld de alomtegenwoordige 'vijf sterren') of geschreven recensies (bijv. Diekmann et al. 2014). Een open vraag over dit soort systemen is of zij even effectief zijn in de anonieme context van het Dark Web, waar geen wet of overheid bestaat om in het uiterste geval goed gedrag af te dwingen. In Przepiorka, Corten en Norbutas (2017) onderzoeken we deze vraag aan de hand van duizenden transacties van Silk Road, een van de eerste grote online drugsmarktplaatsen. We vinden inderdaad dat verkopers met betere reputatiescores niet alleen sneller, maar ook voor hogere prijzen verkopen, wat impliceert dat het vertrouwensprobleem bij hen inderdaad kleiner is. Hiermee laten we zien dat mechanismen die voor sociale orde zorgen in door wetten en instituties gereguleerde maatschappijen óók functioneren in de minimalistische, wetteloze en daarmee bijna hobbesiaanse wereld van het Dark Web.

Uiteraard is er naast de bovenstaande voorbeelden nog veel meer mooi onderzoek naar 'nieuwe' vragen aan de hand van nieuwe digitale data.

Noemenswaardig zijn bijvoorbeeld studies naar de impact van de opkomst van online datingplatforms op homogamie (Rosenfeld en Thomas 2012), naar discriminatie in de deeleconomie (Edelman, Luca en Svirsky 2017) of naar samenwerkingsprocessen in de totstandkoming van Wikipedia (Tsvetkova et al. 2017).

Implicaties

Nieuwe methoden

Naast de ontwikkeling van nieuwe methoden van dataverzameling, waarvan hierboven al enkele aan de orde gekomen zijn, leidt de beschikbaarheid van nieuwe datavormen ook tot de ontwikkeling van nieuwe methoden voor dataverwerking en -analyse. In de eerste plaats zijn deze data, voor zover het om big data in de traditionele betekenis gaat, vaak zo grootschalig dat deze niet meer in het werkgeheugen van conventionele computers ingelezen kunnen worden. Om dergelijke grote datasets toch te kunnen verwerken zijn speciale parelelle rekenmethodes ontwikkeld zoals MapReduce (Dean en Ghemawat 2008), geïmplementeerd in populaire applicaties als Apache Hadoop (bijv. White 2012) en meer recent Apache Spark (Zaharia et al. 2016).

Behalve de verwerking van big data vraagt ook de *analyse* van grote datasets soms om nieuwe methoden. Zo zijn veel 'klassieke' methoden van sociale netwerkanalyse (Wasserman en Faust 1994) in de praktijk niet toepasbaar op de zeer grote netwerkdatasets die tegenwoordig beschikbaar zijn uit sociale media en andere digitale bronnen – de berekeningen zouden onwerkbaar lang duren, zelfs op heel snelle computers. Dit heeft geleid tot de ontwikkeling van een reeks van methoden voor de analyse van zeer grote netwerken die speciaal toegesneden zijn op een efficiënt gebruik van rekenkracht (Newman 2018).

Naast dergelijke uitdagingen biedt de rijkheid van nieuwe digitale data ook kansen. Men kan daarbij, in de context van netwerkanalyse, denken aan methoden die vooral zinvol zijn bij hele grote netwerken (zoals het detecteren van hechte gemeenschappen binnen netwerken; Blondel et al. 2008), maar daarbuiten ook van methoden voor kwantitatieve tekstanalyse en 'natural language processing', zoals bijvoorbeeld het meten van 'sentiment' in geschreven tekst (Ceron et al. 2014), of in het verlengde daarvan, de kwantitatieve analyse van *gesproken* tekst (Cioffi-Revilla 2017; McFarland et al. 2013). Een interessant aspect aan deze ontwikkeling is dat we hierin een convergentie zien van kwalitatieve en kwantitatieve onderzoeksmethoden:

dankzij de grootschaligheid én rijkheid van digitale data is het nu mogelijk kwalitatieve informatie uit zulke data (zoals geschreven of gesproken tekst) op een zodanig systematische manier te verwerken dat deze ook in kwantitatieve vervolganalyses gebruikt kan worden. Een belangrijke ontwikkeling daarbij is het toenemende gebruik van *machine learning* voor (semi-) automatische classificatie en patroonherkenning in tekstdata en andere kwalitatieve data (Grimmer 2015; Hofstra en De Schipper 2018; Molina en Garib 2019).

Een heel andersoortige methodologische innovatie die mogelijk wordt gemaakt door online interactie is de ontwikkeling van grootschalige online (veld)experimenten, die de sterke causale interpretatie van gecontroleerde, gerandomiseerde experimenten combineren met de externe validiteit van een natuurlijke setting (Salganik 2018). Mooie voorbeelden van experimenten waarin onderzoekers gebruik maakten van het internet om natuurlijke situaties op een gecontroleerde maar grootschalige manier na te bootsen zijn onderzoek door Centola (2010) naar de impact van netwerkstructuur op de verspreiding van gedrag en door Salganik et al. (2006) naar succes in culturele markten. Varianten hierop zijn het gebruik maken van bestaande contexten zoals sociale-medianetwerken of online markten voor het uitvoeren van veldexperimenten (Bond et al. 2012; Van de Rijt et al. 2014) of natuurlijke experimenten (Phan en Airoldi 2015).

Nieuwe theorie

Aanvankelijk werd de opkomst van big data door sommigen verwelkomd als het 'einde van theorie' in de wetenschap (Anderson 2008). Hoewel deze verregaande stelling vooralsnog niet bewaarheid lijkt (zie bijv. boyd en Crawford 2012; González-Bailón 2013), is er wel reden aan te nemen dat de beschikbaarheid van nieuwe digitale data, naast methodologische, ook theoretische implicaties zal hebben. Ik noem hier drie voorbeelden ter illustratie.

Ten eerste roepen nieuwe vormen van geografische data de vraag op wat redelijkerwijs als de geografische context voor sociale beïnvloeding en andere sociale processen beschouwd moet worden. Hoewel sociologen en sociaal geografen zich lang vooral gericht hebben op (woon)wijken als de belangrijkste geografische sociale context (Sampson, Morenoff, en Gannon-Rowley 2002), laat recent onderzoek met behulp van GPS-tracking zien waar mensen daadwerkelijk hun tijd doorbrengen, en met wie (Lazer en Radford 2017; Toomet et al. 2015).

Ten tweede maakt de koppeling van grootschalige databestanden uit verschillende sociale domeinen het mogelijk sociale systemen *als geheel* te

onderzoeken. Waar sociologisch onderzoek zich nu nog vaak beperkt tot afzonderlijke sociale domeinen zoals het onderwijs, de arbeidsmarkt of zorg, biedt dit de mogelijkheid theorie te ontwikkelen over hoe individuen zich gedragen in deze verschillende domeinen en hoe sociale processen in deze domeinen interacteren (McFarland et al. 2016).

Ten derde stellen meer en meer gedetailleerde data over sociale interacties het nut van het theoretisch paradigma van *sociale netwerken*, dat juist in de afgelopen decennia een sterke opmars heeft gekend, ter discussie. Onderzoek dat sociale relaties analyseert in termen van sociale netwerken gaat over het algemeen uit van *discrete* relaties met helder gedefinieerde *rollen* (bijvoorbeeld 'vrienden' of 'collega's') die redelijk *stabiel* zijn. Nieuwe data in hoge resolutie (bijvoorbeeld sensordata) stellen deze drie steunpilaren van sociale netwerkanalyse ter discussie (Kitts 2014). Netwerkonderzoekers construeren sociale netwerken doorgaans als aggregaties over een vastgestelde tijdsperiode (met bijvoorbeeld een enquêtevraag als: 'met wie hebt u in het afgelopen kwartaal het vaakst gesproken?'), en het is relatief eenvoudig voor te stellen dat aggregatie over een langere tijdsperiode leidt tot hechtere en meer stabiele sociale structuren, die zich relatief makkelijk als sociale netwerken laten analyseren. Nieuwe vormen van data zoals sensordata maken het echter mogelijk sociale interacties in verregaande mate te de-aggregeren, soms zelfs tot op de minuut nauwkeurig. Enerzijds biedt dit de mogelijkheid metingen van sociale rollen ('vrienden') in detail te vergelijken met het corresponderende gedrag, maar anderzijds neemt met het toenemende detailniveau van observatie natuurlijkerwijs het aantal tegelijk geobserveerde sociale interacties en daarmee de mate van sociale structuur af. Op zijn beurt roept dit de vraag op of het sociale netwerk als metafoor nog wel zo nuttig is, en of een theorie van 'wederzijds afhankelijke dynamiek' (Kitts 2014) niet vruchtbaarder zou zijn.

Uitdagingen

Beperkingen van big data

Uiteraard zijn big data of nieuwe digitale data, zoals elke vorm van data, niet zonder problemen. Ook hier is al het nodige over gezegd en geschreven (zie bijv. Hampton 2017; Lewis 2015); ik zal me hier beperken tot de belangrijkste bezwaren. Ten eerste zijn dit soort data, omdat ze typisch niet verzameld zijn met het oogmerk van wetenschappelijk onderzoek, vaak *incompleet*, in de zin dat ze niet altijd de informatie bevatten waarmee de

theoretische constructen uit de onderzoeksvraag op een valide manier geoperationaliseerd kunnen worden (Lazer 2015).

Ten tweede zijn digitale data, zoals sociale-mediadata, vaak *niet representatief* voor de doelpopulatie. Een goed voorbeeld hiervan is Twitter, dat vanwege de relatieve toegankelijkheid van de data veel gebruikt wordt voor onderzoek maar duidelijk geen afspiegeling vormt voor de volwassen populatie van westerse samenlevingen (Hargittai 2015; Schober et al. 2016).

Een gerelateerd probleem is dat databronnen vaak *tijdgebonden* zijn, in de zin dat de populatie die een systeem (zoals een sociale-mediaplatform) gebruikt of de wijze waarop deze een systeem gebruikt kan veranderen over de tijd. Zo kan een verandering in de publieke opinie als gemeten via sociale media veroorzaakt worden door een daadwerkelijke verandering van de publieke opinie onder de bevolking, maar ook door een verschuiving in het deel van de bevolking dat de betreffende dienst gebruikt (Salganik 2018).

Ten vierde zijn data afkomstig uit digitale systemen soms '*algoritmisch vervuild*' (Salganik 2018), in de zin dat er buiten het zicht van de onderzoeker processen in het systeem ingebouwd zijn die het gedrag van gebruikers beïnvloeden. Een bekend voorbeeld hiervan zijn de algoritmes die gebruikers potentiële 'vrienden' suggereren op platformen als Facebook en LinkedIn.

Ten vijfde kunnen big data *ruis* bevatten die door de omvang van de data niet eenvoudig te detecteren is maar wel tot misleidende conclusies kan leiden. Een berucht voorbeeld is de sentimentanalyse van semafoonberichten na de aanslagen van 11 september 2001. Waar de oorspronkelijke analyse een sterke toename in 'woede' liet zien, bleek uit een heranalyse dat deze toename geheel voor rekening kwam van één apparaat dat berichten verstuurde die niets met de aanslagen te maken hadden (Pury 2011).

Tot slot zijn big data, hoewel die doorlopend op grote schaal verzameld worden, vaak *niet beschikbaar* voor wetenschappelijk onderzoek. In veel gevallen worden deze data verzameld door private partijen zoals sociale-mediaplatformen (Facebook) wiens verdienmodel gebaseerd is op het bezit van data over gebruikers, of die om andere redenen zeer terughoudend zijn met het delen van data. Dit geldt in veel gevallen ook voor data verzameld voor publieke doeleinden. Om toch toegang te krijgen tot zulke data zijn onderzoekers vaak genoodzaakt (geheimhoudings)overeenkomsten aan te gaan met eigenaren van de data (zoals sociale-mediaplatformen), wat in sommige gevallen de vrijheid van het onderzoek begrenst, de schijn van afhankelijk creëert en bovendien de mogelijkheden voor verificatie en replicatie door andere onderzoekers beperkt. Hampton (2017) trekt wat dat

betreft een parallel met etnografisch onderzoek, waar het verkrijgen van toegang tot de onderzoekspopulatie een belangrijke rol speelt in het onderzoeksproces en waar replicatie ook vaak problematisch is.

Kapers op de kust?

De bovenstaande problemen suggereren dat hoewel omgang met nieuwe digitale data vaak om nieuwe technieken en vaardigheden vraagt, klassieke sociaalwetenschappelijke onderzoeksvaardigheden verre van irrelevant zijn geworden: sociale wetenschappers zijn vaak juist bedreven in het inschatten van representativiteit, het opschonen van data, omgang met meetfouten, et cetera. Daarnaast hebben zij vaak ook de inhoudelijke veldkennis in huis om de validiteit van nieuwe digitale data op waarde te kunnen schatten. Toch wordt een belangrijk deel van het hedendaagse big data-onderzoek, óók waar het sociaalwetenschappelijke onderwerpen betreft, gedaan door onderzoekers met een achtergrond in bètawetenschappen, zoals natuurkunde of informatica (of het meer recentelijk modieuze 'data science'). Deze onderzoekers profiteren daarbij van hun sterkere technische vaardigheden bij het verzamelen en analyseren van big data, zoals kennis van programmeren, omgang met grote databases, machine learning en het gebruik (en de beschikbaarheid) van de juiste hardware (zoals HPC-clusters) en mogelijk ook van ruimere financiering en meer prestige bij het grote publiek.

Hoewel de bijdragen vanuit deze vakgebieden zeer waardevol en soms zelfs baanbrekend zijn, is de dominantie van bètawetenschappers – of eigenlijk: het gebrek aan betrokkenheid van sociale wetenschappers – in een aantal opzichten ook problematisch. Ten eerste kennen, zoals al opgemerkt, big data tal van problemen waarvoor kennis van sociaalwetenschappelijke onderzoeksmethoden zeer bruikbaar is. Ten tweede ontberen onderzoekers met een bèta-achtergrond doorgaans kennis van sociologisch onderzoek en vooral sociologische theorie, wat er toe leidt dat velen opnieuw het wiel uitvinden.[11] Uiteraard leidt dit soms tot vernieuwende inzichten – het baanbrekende onderzoek van Watts en Strogatz (1998) naar 'small worlds' is een treffend voorbeeld – maar meer in het algemeen is het gebrek aan verbinding met de sociologische literatuur niet bevorderlijk voor de groei van kennis, onder meer doordat het leidt tot een wildgroei aan ad-hocmodellen die moeilijk generaliseerbaar zijn naar andere contexten (González-Bailón 2013). Ten derde lopen de sociale wetenschappen als discipline het risico aan relevantie te verliezen als zij het aan andere disciplines overlaten te profiteren van de mogelijkheden die nieuwe digitale data bieden. Dit geldt

misschien in het bijzonder voor de sociologie, aangezien big data zich bij uitstek lenen voor het beantwoorden van sociologische vragen.

Ook uit een andere hoek kunnen sociale wetenschappers 'concurrentie' verwachten als het gaat om het bestuderen van sociale processen met big data, namelijk vanuit de ICT-sector zelf, die een belangrijk deel van de relevante data produceert. Een bedrijf als Facebook heeft bijvoorbeeld een eigen data science-afdeling, die niet alleen vrijuit de beschikking heeft over wat waarschijnlijk de grootste sociologische dataset ter wereld is, maar ook over de bijbehorende hardwarefaciliteiten, en daarbij ook de middelen heeft om experts aan te trekken tegen salarissen die universiteiten nooit zouden kunnen bieden.[12] Hoewel bijvoorbeeld het Facebook-team regelmatig publiceert, ook in wetenschappelijke tijdschriften (bijv. Bakshy et al. 2015) is het onderzoek dat deze bedrijven doen is niet gebonden aan dezelfde regels omtrent transparantie, onderzoeksethiek en het delen van data die gelden voor academische onderzoekers, zoals pijnlijk geïllustreerd door de controverse rondom het onderzoek van Kramer et al. (2014), waarin Facebook zonder voorafgaande toestemming emoties van gebruikers manipuleerde. Deze controverse liep hoog op mede omdat het in een prominent wetenschappelijk tijdschrift gepubliceerd werd, maar er is alle reden om aan te nemen dat platformbedrijven als Facebook regelmatig dergelijke experimenten op hun gebruikers uitvoeren (bijvoorbeeld om nieuwe functies van hun product te testen), zonder daar de buitenwereld over te informeren. Dit leidt tot wat ik de 'big-dataparadox' in de sociale wetenschappen zou willen noemen: hoewel meer sociaalwetenschappelijk relevante data worden verzameld dan ooit, hebben commerciële partijen vandaag de dag een betere uitgangspositie voor het bestuderen van de samenleving dan (academische) sociale wetenschappers. Omdat deze partijen niet de verplichting of de noodzaak kennen de daarbij geproduceerde kennis te publiceren, is de verwachting reëel dat veel van deze kennis niet in het publieke domein terecht zal komen maar in het bezit zal blijven van enkele grote private partijen, wat op zijn beurt de mogelijkheden beperkt om via democratische wegen maatschappelijke problemen aan te pakken.

Al met al is de situatie dus zorgelijk: niet alleen laten sociologen grote kansen liggen voor het doen van vernieuwend onderzoek met nieuwe digitale data, ze dreigen daarin ook te worden voorbijgestreefd door zowel collega-wetenschappers uit andere disciplines als door private partijen. Wat moet er gebeuren om in deze situatie verandering aan te brengen? Ik besluit met een paar suggesties.

Ten eerste moeten sociologen de aansluiting vinden met onderzoekers uit andere disciplines die zich bezighouden met nieuwe digitale data,

waarbij complementariteit van kennis en methoden voorop moet staan. Hoewel er het nodige aan te merken is op het concept 'data science', bieden de initiatieven die momenteel onder deze noemer aan veel universiteiten zijn of worden opgezet goede kansen om dergelijke samenwerkingen op te zetten, en sociale wetenschappers zouden er goed aan doen zich nadrukkelijk met deze initiatieven te bemoeien.

Ten tweede zou ook onder sociale wetenschappers enige 'geletterdheid' op het gebied van big data gewenst zijn, en idealiter begint dit in de curricula van sociaalwetenschappelijke universitaire opleidingen. Hierbij kan, op het bachelorniveau, gedacht worden aan inhoudelijke kennis van de mogelijkheden én beperkingen van nieuwe digitale data en op het (research) masterniveau aan zaken als basale programmeervaardigheden (omdat veel software voor analyse van big data, gezien de snelle ontwikkelingen, niet binnen de gangbare statistiekpakketten beschikbaar is), kennis van relationele databases en tekstanalyse.

Tot slot is meer aandacht nodig voor onderzoeksdesigns die niet afhankelijk zijn van de medewerking van platformbedrijven als Facebook voor het verzamelen van data. Enerzijds helpt dit wetenschappers de problemen te omzeilen die ontstaan bij samenwerking met dergelijke partijen (zie boven), en anderzijds kan zulk onafhankelijk onderzoek tegenwicht bieden aan onderzoek dat – soms met ondoorzichtige belangen – gedaan wordt door private partijen. Goede voorbeelden van onafhankelijke onderzoeksstrategieën zijn de grootschalige online experimenten door Centola (2010, 2011) en Salganik et al. (2006), de online veldexperimenten door Aral en Walker (2011) en Van de Rijt et al. (2014), of het onafhankelijke platform Inside Airbnb[13] dat met eigen 'scraping' verzamelde data van het bekende verhuurplatform beschikbaar maakt voor onderzoek. Bovenal vraagt de beschikbaarheid van nieuwe vormen van digitale data om creativiteit van sociale wetenschappers in het stellen van nieuwe vragen en het onderzoeken van oude vragen op nieuwe manieren, om de mogelijkheden die deze nieuwe data bieden ten volle te benutten.

Noten

1 Dit artikel is gebaseerd op de gelijknamige lezing uitgesproken bij gelegenheid van het Actualiteitencollege 2018, georganiseerd door de Nederlandse Sociologische Vereniging in Den Haag op 29 november 2018. Mijn dank gaat uit naar Marissa Bultman voor de onderzoeksondersteuning bij het schrijven van dit artikel.

2 Varianten op deze definitie gaan uit van 5, 7 of zelfs 9 'V's'.

3 Mertons *self-fulfilling prophecy* (1968: 19) is een ander voorbeeld.

4 Onder de aanname dat de data door het platform beschikbaar gemaakt worden voor onderzoek; hier komen we later nog op terug.
5 Een willekeurig gevormd netwerk met dezelfde dichtheid als het Hyves-netwerk zou een gemiddelde clustercoëfficiënt van vrijwel nul hebben.
6 Hier is een duidelijke parallel met de kleine-wereldtheorie van Watts en Strogatz (1998).
7 Uiteraard bevatten Facebooknetwerken ook sterke banden, met bijvoorbeeld vrienden en familie. Deze kunnen we tot op zeker hoogte identificeren in de data, maar dit blijkt geen invloed te hebben op de resultaten.
8 Een alternatieve interpretatie van deze resultaten is de vergelijking tussen *online* sociale relaties en *offline* relaties. In principe zou dit een voorbeeld zijn van 'een nieuwe vraag met nieuwe data beantwoorden'. Uit ander onderzoek (Dunbar et al. 2015) weten we echter al dat online- en offlinenetwerken niet fundamenteel van elkaar verschillen.
9 'Fundamenteel' als in tegenstelling tot 'toegepast'; ik wil hiermee geen waardeoordeel geven over het belang van dit onderzoek in vergelijking met het voorgaande.
10 Zie www.torproject.org.
11 De titel van Pentlands (2014) invloedrijke boek *Social Physics* is wat dat betreft veelzeggend.
12 Zo wordt het 'computational social science team' van Facebook geleid door Lada Adamic, die daarvoor een glanzende academische carrière kende als prominent netwerkonderzoeker. Duncan Watts werkte jarenlang voor Yahoo en Microsoft.
13 http://insideairbnb.com/

Literatuur

Ackland, Robert (2013) *Web social science: Concepts, data and tools for social scientists in the digital age*. London: Sage.

Allcot, H. en M. Gentzkow (2017) Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2): 211-236.

Anderson, Chris (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, juni 23.

Aral, Sinan en Dylan Walker (2011) Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks. *Management Science*, 57(9): 1623-1639.

Bakshy, E., S. Messing en L.A. Adamic (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130-1132.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte en Etienne Lefebvre (2008) Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Blumenstock, Joshua, Gabriel Cadamuro en Robert On (2015) Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264): 1073-1076.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle en James H. Fowler (2012) A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415): 295-298.

boyd, danah en Kate Crawford (2012) Critical Questions for Big Data. *Information, Communication & Society*, 15(5): 662-679.

Cabañas, Jose González, Angel Cuevas en Rubén Cuevas (2018) Facebook Use of Sensitive Data for Advertising in Europe. arXiv preprint, arXiv: 1802.05030.

Centola, Damon (2010) The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996): 1194-1197.

Centola, Damon (2011) An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060): 1269-1272.

Ceron, Andrea, Luigi Curini, Stefano M. Lacus en Giuseppe Porro (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, 16(2): 340-358.

Chun, Hyunwoo, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon en Hawoong Jeong (2008) Comparison of Online Social Relations in Terms of Volume vs. Interaction: A Case Study of Cyworld. In: *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, 57-70.

Cioffi-Revilla, Claudio (2017) *Introduction to computational social science*. New York: Springer.

Coleman, James S (1990) *Foundations of Social Theory*. Cambridge: Belknap.

Corten, Rense (2012) Composition and structure of a large online social network in the Netherlands. *PLOS one*, 7(4): e34760.

Dean, Jeffrey en Sanjay Ghemawat (2008) MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1): 107-113.

Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley en Walter Quattrociocchi (2016) The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 201517441.

Del Vicario, Michela, Fabiana Zollo, Guido Caldarelli, Antonio Scala en Walter Quattrociocchi (2017) Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*, 50:6-16.

Diekmann, Andreas, Ben Jann, Wojtek Przepiorka en Stefan Wehrli (2014) Reputation formation and the evolution of cooperation in anonymous online markets. *American Sociological Review*, 79(1): 65-85.

Dunbar, Robin I.M., Valerio Arnaboldi, Marco Conti en Andrea Passarella (2015) The structure of online social networks mirrors those in the offline world. *Social Networks*, 43: 39-47.

Eagle, Nathan, Michael Macy en Rob Claxton (2010) Network diversity and economic development. *Science*, 328(5981): 1029-1031.

Edelman, Benjamin, Michael Luca en Dan Svirsky (2017) Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, 9(2): 1-22.

Facebook (2019) Company Info. Verkregen op 1 april 2019, https://newsroom.fb.com/company-info/.

González-Bailón, Sandra (2013) Social Science in the Era of Big Data. *Policy & Internet*, 5(2): 147-160.

González-Bailón, Sandra (2017) *Decoding the social world: Data science and the unintended consequences of communication*. Cambridge: MIT Press.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero en Yamir Moreno (2011) The dynamics of protest recruitment through an online network. *Scientific Reports*, 1: 197.

Granovetter, Mark S. (1973) The Strength of Weak Ties. *American Journal of Sociology*, 78(6): 1360-1380.

Grimmer, Justin (2015) We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, 48(1): 80-83.

Hampton, Keith N. (2017) Studying the Digital: Directions and Challenges for Digital Methods. *Annual Review of Sociology*, 43(1): 167-188.

Hargittai, Eszter (2015) Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1): 63-76.

Hill, Russell A. en Robin I. Dunbar (2003) Social Network Size in Humans. *Human Nature*, 14(1): 53-72.

Hofstra, B., R. Corten, F. van Tubergen en N. B. Ellison (2017) Sources of Segregation in Social Networks: A Novel Approach Using Facebook. *American Sociological Review*, 82(3).

Hofstra, Bas en Niek C. de Schipper (2018) Predicting Ethnicity with First Names in Online Social Media Networks. *Big Data & Society*, 5(1): 2053951718761141.

Jackson, Matthew O. en Brian W. Rogers (2007) Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3): 890-915.

Katal, Avita, Mohammad Wazid en R. H. Goudar (2013) Big Data: Issues, Challenges, Tools and Good Practices. In: *Sixth international conference on contemporary computing*, vol. 2013: 404-409.

Khurshid, K., A. Khan, H. Siddique en I. Rashid (2018) Big Data-9Vs, Challenges and Solutions. *Technical Journal*, 23(03): 28-34.

Kitts, James A (2014) Beyond networks in structural theories of exchange: Promises from computational social science. In: *Advances in group processes*. Bingley: Emerald, 263-298.

Korte, Charles en Stanley Milgram (1970) Acquaintance networks between racial groups: Application of the small world method. *Journal of Personality and Social Psychology*, 15(2): 101.

Kramer, Adam D.I., Jamie E. Guillory en Jeffrey T. Hancock (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788-8790.

Lazer, David (2015) Issues of construct validity and reliability in massive, passive data collections. *The City Papers: An Essay Collection from The Decent City Initiative*. Verkregen op 25 april 2019, http://citiespapers.ssrc.org/issues-of-construct-validity-and-reliability-in-massive-passive-data-collections/.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy en Marshall Van Alstyne (2009) Computational Social Science. *Science*, 323(5915): 721-723.

Lazer, David en Jason Radford (2017) Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, 43(1): 19-39.

Leskovec, Jure en Eric Horvitz (2008) Planetary-scale views on a large instant-messaging network. In: *Proceeding of the 17th international conference on World Wide Web*. New York: ACM, 915-924.

Lewis, Kevin (2015) Three Fallacies of Digital Footprints. *Big Data & Society*, 2(2): 2053951715602496.

Martin, James (2014) *Drugs on the dark net: How cryptomarkets are transforming the global trade in illicit drugs*. New York: Springer.

McFarland, Daniel A., Dan Jurafsky en Craig Rawlings (2013) Making the connection: Social bonding in courtship situations. *American Journal of Sociology*, 118(6): 1596-1649.

McFarland, Daniel A., Kevin Lewis en Amir Goldberg (2016) Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist*, 47(1): 12-35.

McPherson, Miller, Lynn Smith-Lovin en James M. Cook (2001) Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1): 415-444.

Merton, Robert K. (1968) *Social Theory and Social Structure*. New York: Free Press.

Milgram, Stanley (1967) The Small World Problem. *Psychology Today*, 2: 60-67.

Molina, Mario en Filiz Garib (2019) Machine Learning for Sociology. *Annual Review of Sociology*, 45: in druk.

Mollenhorst, Gerald, Beate Völker en Henk Flap (2008) Social contexts and personal relationships: The effect of meeting opportunities on similarity for relationships of different strength. *Social Networks*, 30(1): 60-68.

Moody, James (2001) Race, School Integration, and Friendship Segregation in America. *American Journal of Sociology*, 107(3): 6790-716.

Mouw, Ted (2003) Social Capital and Finding a Job: Do Contacts Matter? *American Sociological Review*, 68(6): 868-898.

Newman, Mark E.J. (2018) *Networks*. Second Edition. Oxford: Oxford University Press.

Norbutas, L. en R. Corten (2017) Network structure and economic prosperity in municipalities: A large-scale test of social capital theory using social media data. *Social Networks*, 52: 120-134.

Paik, Anthony en Kenneth Sanchagrin (2013) Social Isolation in America: An Artifact. *American Sociological Review*, 78(3): 339-360.

Park, Patrick S., Joshua E. Blumenstock en Michael W. Macy (2018) The Strength of Long-Range Ties in Population-Scale Social Networks. *Science*, 362(6421): 1410-1413.

Pentland, Alex (2014) *Social physics: How social networks can make us smarter*. New York: Penguin.

Phan, Tuan Q. en Edoardo M. Airoldi (2015) A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences*, 112(21): 6595-6600.

Przepiorka, Wojtek, Lukas Norbutas en Rense Corten (2017) Order without Law: Reputation Promotes Cooperation in a Cryptomarket for Illegal Drugs. *European Sociological Review*, 33(6).

Pury, Cynthia L.S. (2011) Automation can lead to confounds in text analysis: Back, Küfner, and Egloff (2010) and the not-so-angry Americans. *Psychological science*, 22(6): 835.

Rijt, Arnout van de, Soong Moon Kang, Michael Restivo en Akshay Patil (2014) Field Experiments of Success-Breeds-Success Dynamics. *Proceedings of the National Academy of Sciences*, 111(19): 6934-6939.

Rijt, Arnout van de, Eran Shor, Charles Ward en Steven Skiena (2013) Only 15 Minutes? The Social Stratification of Fame in Printed Media. *American Sociological Review*, 78(2): 266-289.

Rosenfeld, M. J. en R. J. Thomas (2012) Searching for a Mate: The Rise of the Internet as a Social Intermediary. *American Sociological Review*, 77(4): 523-547.

Salganik, Matthew J (2018) *Bit by Bit*. Princeton: Princeton University Press.

Salganik, Matthew J., Peter Sheridan Dodds en Duncan J. Watts (2006) Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762): 854-856.

Sampson, Robert J., Jeffrey D. Morenoff en Thomas Gannon-Rowley (2002) Assessing 'Neighborhood Effects': Social Processes and New Directions in Research. *Annual Review of Sociology*, 28(1): 443-478.

Schober, Michael F., Josh Pasek, Lauren Guggenheim, Cliff Lampe en Frederick G. Conrad (2016) Social media analyses for social measurement. *Public Opinion Quarterly* 80(1): 180-211.

Shah, Dhavan V., Joseph N. Cappella en W. Russell Neuman (2015) Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. *The ANNALS of the American Academy of Political and Social Science*, 659(1): 6-13.

Silm, Siiri en Rein Ahas (2014) Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers*, 104(3): 542-559.

Takes, Frank W. en Walter A. Kosters (2011) Identifying prominent actors in online social networks using biased random walks. *BNAIC*, 23: 215-222.

Tempelman, Olaf (2019) Zuckerbergs ideaal van een verbonden wereld leidt juist tot een versplintering. *De Volkskrant*, 15 februari.

Toomet, Ott, Siiri Silm, Erki Saluveer, Rein Ahas en Tiit Tammaru (2015) Where Do Ethno-Linguistic Groups Meet? How Copresence during Free-Time Is Related to Copresence at Home and at Work. *PLOS ONE*, 10(5): e0126093.

Tsvetkova, Milena, Ruth García-Gavilanes, Luciano Floridi en Taha Yasseri (2017) Even good bots fight: The case of Wikipedia. *PloS one*, 12(2).

Ugander, J., B. Karrer, L. Backstrom en C. Marlow (2011) The Anatomy of the Facebook Social Graph. *Arxiv preprint arXiv*: 1111.4503.

Wasserman, Stanley en Katherine Faust (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Watts, Duncan J. en Steven H. Strogatz (1998) Collective Dynamics of 'Small World' Networks. *Nature*, 393: 440-442.

White, Tom (2012) *Hadoop: The definitive guide*. Sebastopol: O'Reilly Media.

Zaharia, Matei, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman en Michael J. Franklin (2016) Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11): 56-65.

Over de auteur

Rense Corten is universitair hoofddocent bij de afdeling Sociologie. Hij doet onderzoek naar samenwerking en vertrouwen in relatie tot (de dynamiek van) sociale netwerken, met empirische toepassingen zoals netwerken van adolescenten, sociale-medianetwerken, de deeleconomie, online criminele netwerken en labexperimenten. In 2016 ontving hij een NWO Vidi-subsidie voor een onderzoeksproject naar oorzaken en gevolgen van vertrouwen in de deeleconomie.
E-mail: r.corten@uu.nl