Zoeken in een extreem grote database lijkt op googlen. Hoe haal je zeldzame patronen uit de resultaten?

Door HERBERT BLANKESTEIJN

De Amerikaanse inlichtingendienst NSA onderschept het telefoon- en internetverkeer – niet alleen in Amerika, maar ook in Europa, bleek afgelopen week. Om hoeveel data gaat het eigenlijk? En hoe haalt de NSA daar bruikbare informatie uit voor het opsporen van terroristen?

Het is niet aannemelijk dat de NSA elke dag een kopietje van alle internetverkeer draait. De website Gigaom.com heeft uitgerekend dat alleen al het opslaan van, bijvoorbeeld, de jaarproductie aan data van Facebook 10 miljoen dollar kost. Zulke budgetten heeft de NSA niet: het hele PRISM-programma, dat de inlichtingendienst gebruikt om de eindeloze stroom aan data snel te kunnen scannen, kost volgens de presentatie die klokkenluider Edward Snowden naar buiten heeft gebracht jaarlijks 20 miljoen dollar.

Het ligt dus voor de hand dat PRISM een geautomatiseerde procedure is voor het opvragen van specifieke gegevens, en geen stortplaats waar alles blindelings wordt gedumpt. Ook op een andere manier bezien is het niet waarschijnlijk dat de NSA 'alles' zelf opslaat: internetbedrijven als Yahoo, Google, Facebook, Amazon en Microsoft houden er vele datacenters op na. De opslagfaciliteit van de NSA zou een even grote capaciteit moeten hebben als al deze centra samen. Zo'n faciliteit is er niet.

Dat zou wel kunnen veranderen: op dit moment zijn er twee datacentra van de NSA in aanbouw, een in de staat Utah dat dit jaar wordt voltooid en een in Maryland dat in 2016 klaar moet zijn. Samen kosten ze 2 à 3 miljard dollar. Gissingen over de capaciteit gaan in de richting van een yottabyte. Dat is een miljoen maal een miljoen terabyte – oftewel 1 byte met 24 nullen erachter. Ter vergelijking: als je nu een laptop koopt, is de harde schijf meestal 0,5 à 1 terabyte groot. Een yottabyte is ook het duizendvoudige van wat het hele internet in 2015 jaarlijks zal produceren.

Hoe zoekt een geheime dienst in een extreem grote database? Op zich is dat bekend terrein. „Eén mogelijkheid is: je bent geïnteresseerd in een patroon en je gaat zoeken of dat er is", zegt Wil van der Aalst, hoogleraar informatiesystemen aan de TU Eindhoven. Zoeken naar concrete kenmerken is niet wezenlijk anders dan googlen. Je kunt je zoekvraag bijvoorbeeld beperken tot data afkomstig van een aantal IP-adressen. Een IP-adres geeft aan uit welke regio of van welke internetaanbieder gegevens afkomstig zijn. Je kunt bijvoorbeeld alleen zoeken binnen e-mails uit één land.

Zulke zoekvragen zijn voor een database dagelijkse kost. Hoe beter de data zijn ingeperkt, hoe sneller het antwoord komt. Daarom is het aannemelijk dat de NSA niet alle mails 'leest', zelfs niet op een geautomatiseerde manier. Veel doelmatiger is het om uit te gaan van verkeersgegevens, zoals president Obama heeft gezegd. Verkeersgegevens – wie communiceert met wie, wanneer en waarvandaan – vormen een veel kleinere voorraad data.

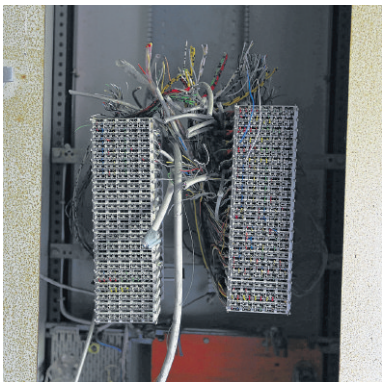
Door onderzoek daarvan valt te bepalen welke gesprekken beluisterd en welke mails gelezen moeten worden.

De NSA gebruikt hiervoor vergelijkbare technieken als grote internetbedrijven als Google, Yahoo en Facebook, zoals de databasesoftware Hadoop, waarmee je datacollecties kunt beheren en doorzoeken. Hadoop is *open source*, wat inhoudt dat de programmacode vrij beschikbaar is, vrij mag worden gebruikt en ook door iedereen te bewerken is. De NSA heeft een eigen uitbreiding van Hadoop geschreven, genaamd Accumulo. Accumulo maakt de zoekresultaten afhankelijk van de autorisatie van degene die zoekt – en bewijst daarmee opmerkelijk genoeg de privacy een dienst.

De geheime dienst heeft Accumulo vorig jaar zelfs beschikbaar gemaakt als open source software. Het internetbedrijf Sqrrl, mede opgericht door ex-NSA'er Adam Fuchs, helpt nu op zijn beurt bedrijven deze software nuttig te gebruiken. Met Sqrrl kunnen telecombedrijven bijvoorbeeld storingen opsporen door in hun communicatiekanalen te zoeken naar woorden als 'kapot'. Online winkels kunnen ermee voorspellen wat klanten die artikel A hebben gekocht, nog meer zouden willen hebben.

Het zoeken van een concreet patroon is relatief makkelijk. Maar zo'n duidelijke vraagstelling is niet altijd ideaal. Wie alleen kijkt naar communicatie met Arabische landen, mist binnenlands terrorisme. Patronen die je niet zoekt, zul je waarschijnlijk niet vinden. Een voorbeeld. Edward Snowden heeft volgens *The New York Times* in Hongkong een bijeenkomst gehad met juristen, waar hij zijn gasten vroeg hun mobieltjes in de koelkast te leggen om ze te isoleren van het radionetwerk. Achteraf kun je vaststellen dat het een interessante gebeurtenis is om naar te zoeken: een handvol mobieltjes die op dezelfde locatie tegelijkertijd van het net verdwijnen. Maar dan moet je wel weten dat je ernaar moet zoeken.

Hoe ontdek je zoiets? „Als je zeldzame patronen zoekt, en niet precies weet wát, helpt het soms als je visuele voorstellingen maakt", zegt de Eindhovense hoogleraar Van der Aalst. „Als mens zie je daar makkelijk patronen in. Je ziet bijvoorbeeld veel activiteit op een bepaald tijdstip." Het bestuderen van grafische voorstellingen van Big Data, zogeheten Big Graphs, heeft de aandacht van de NSA, zoals blijkt uit een bijdrage van twee medewerkers van de NSA eind mei op een conferentie aan de Carnegie Mellon-universiteit.



Bliksem kost net zo veel levens als terrorisme

Big Graphs zoeken de grenzen op van wat op dit moment mogelijk is qua omvang van bestanden. Bestanden van vele exabytes (miljoenen terabytes) zijn geen uitzondering, en die zijn zelfs voor supercomputers zware kost, zoals de NSA-delegatie op de conferentie voorrekende.

Meer data maken de resultaten betrouwbaarder – als je tenminste in staat bent de data goed te analyseren. Boekt de NSA de gewenste resultaten? Critici wijzen erop dat de geheime diensten gebeurtenissen als 9/11 en de aanslag op de marathon van Boston niet hebben kunnen voorkomen. En dat terwijl in beide gevallen vooraf aandacht was geweest voor de latere daders. Maar de NSA kan simpelweg niet alle data analyseren, doordat er meer wordt verzameld dan de dienst aankan.

Er zijn dus prominente gevallen die ten onrechte niet zijn gesignaleerd (in jargon: de vals-negatieven). En dan heb je ook nog de vals-positieven: mensen die ten onrechte als verdachte worden aangemerkt. Juist bij het zoeken in grote aantallen naar zeldzame eigenschappen is dat een moeilijk te bestrijden probleem. Epidemiologen weten dat. Stel: je onderzoekt een miljoen individuen van wie er naar verwachting 100 het gezochte kenmerk hebben: ze hebben een zeldzame ziekte of ze zijn terrorist. En stel dat de diagnose in maar 1 procent van de gevallen onjuist is. Dan levert de test naast de 99 'echte' gevallen ook zo'n 10.000 vals-positieven op: 1 procent van een miljoen. Ondanks de bijna perfecte test heb je honderd keer zoveel onterecht verdachten als echte terroristen.

En, hoe pijnlijk het ook is voor de NSA: terrorisme is vooral relevant als schrikbeeld, niet als werkelijke dreiging. John Mueller, hoogleraar nationale veiligheid aan Ohio State University, zette al in 2004 feiten en cijfers op een rijtje in het blad *Regulation*. Amerika had jaarlijks 40.000 verkeersdoden te betreuren. Het aantal doden door terrorisme is in geen enkel jaar meer dan een paar honderd geweest, behalve in 2001.

Mueller somde verschijnselen op die grofweg evenveel levens eisen als terrorisme: blikseminslagen, botsingen met herten en pinda-allergie. Zelfs als je de aanslag op het WTC in 2001 meerekent, is de kans dat een Amerikaan sterft door terrorisme ongeveer 1 op 75.000 in een periode van 80 jaar. De kans dat dezelfde Amerikaan omkomt in het verkeer is duizend keer zo groot: 1 op 80, aldus Mueller.

Zijn de effecten van terrorisme zo beperkt dankzij effectieve preventie, dus mede door afluisteren? Dat is niet waarschijnlijk: om vliegen in de VS per passagierskilometer even onveilig te maken als autorijden, zou elke maand een 9/11 moeten plaatsvinden, stelt Mueller. Zelfs de NSA claimt niet dat ze zoveel dood en verderf weet te voorkomen. Als het erom gaat levens te redden, zijn de NSA-dollars elders waarschijnlijk beter besteed.