Visualisatie of interpretatie?

Een kritische reflectie op visualisaties van online netwerken

Martijn Weghorst
3240800
Universiteit Utrecht
November 2010

Het begint een ware hype te worden; elke politicus lijkt tegenwoordig te moeten *twitteren*. Dit gaat zo ver dat het al tot de nodige politieke opstootjes heeft geleid. Maxime Verhagen werd in 2008 een keer door toenmalig premier Balkenende op de vingers getikt omdat hij een foto van een kabinetsraad via Twitter had verspreid. En tijdens de kabinetsformatie in 2010 vroeg informateur Opstelten zelfs om een 'twitterstilte', omdat de formatie volgens hem gebaat was bij geduld en rust. Politici maken gretig gebruik van Twitter; Meer dan de helft van de verkozen politici tijdens de Tweede Kamer verkiezingen van 2010 gebruikt Twitter of heeft dat in het verleden gedaan.

Natuurlijk zijn politici niet de enigen die gebruik maken van online sociale netwerken. Facebook is immers niet zomaar de 500 miljoen leden gepasseerd. Zulke netwerken bezitten dan ook enorm veel data, en dat biedt interessante mogelijkheden voor een nieuw soort onderzoek.[1] Steeds vaker zie ik de meest prachtige visualisaties van databases verschijnen met meer of minder opmerkelijke bevindingen.[2]

Deze visualisaties zijn echter niet puur het gevolg van harde feiten, maar ook van keuzes van de onderzoeker. Het begint bijvoorbeeld al bij de keuze van het onderzoeksmateriaal: welke data gebruik je wel voor je onderzoek en welke niet. Daarom wil ik me in dit paper gaan richten op de vraag hoe waardevol zulke visualisaties nou eigenlijk zijn en er aan de hand van een voorbeeldonderzoek eventueel een kritische reflectie op geven.

Voorbeeldonderzoek

Dat voorbeeldonderzoek zal gaan over het Nederlandse politieke landschap. Ik wil in ieder geval van alle huidige Tweede Kamerleden hun Twitteraccount (mits ze dat hebben) meenemen. Van elke gebruiker zal ik onderzoeken met wie hij bevriend is. Daarna wil ik bekijken welke accounts door meerdere Tweede Kamerleden worden gevolgd en ook die accounts onderzoeken. Ik hoop dan een database te krijgen bestaande uit enkele honderden politiek relevante Twitteraccounts, inclusief alle onderliggende relaties. Daarnaast zal ik van alle accounts uit die database de berichten van een bepaalde periode downloaden en kort analyseren en op die manier proberen uit te vinden welke issues vaak besproken werden.

Om het paper overzichtelijk te houden, zal ik het onderzoek verdelen in twee fasen:

- **Fase 1:** Het construeren en analyseren van de Haagse twittersfeer. Wie nemen deel, welke groepen zijn vertegenwoordigd, wie volgt wie, wie praat met wie, wie is populair?
- **Fase 2:** Bekijken welke issues er in een bepaalde periode speelden. Welke issues waren hot en op welk moment?

Aanvankelijk was ik ook nog van plan om een derde fase uit te voeren. Daarin zou ik de communicatie onder twitteraars onder de loep nemen, in combinatie met een gevonden issue, en zou ik onderzoeken of er op Twitter sprake is van discussie. De tijd en ruimte om ook deze fase af te ronden ontbraken echter.

[1] Een mooi overzicht van zulke nieuwe onderzoeksmethoden voor digitale data is te vinden op de website van Richard Rogers, digitalmethods.net

[2] Zie bijvoorbeeld deze grafiek over het humeur van de Amerikaanse bevolking: http://twittermood.s3.amazonaws.com/images/poster-large.png

Fase 1 – Wie is het (*Constructie en analyse van het netwerk*)

Zoals gezegd ben ik begonnen met het construeren van de *Haagse Twittersfeer*. Het woord construeren dient hier niet te letterlijk genomen te worden; Ik begon immers met een bestaande (offline) community, de Tweede Kamer. De eerste stap van het onderzoek was dan ook het downloaden van een lijst van alle Tweede Kamerleden van Wikipedia en het omzetten van die lijst naar een database. Vervolgens heb ik van alle Kamerleden uitgezocht of zij een Twitter account hebben. Als dat het geval was, heb ik hun *screen name* (Twitternaam) in de database gezet.[3] Nadat ik van alle politici had bekeken of ze twitteren, heb ik degenen die dat niet doen uit de database verwijderd (hier komt dus het 'construeren' om de hoek kijken).

De volgende stap was om met behulp van de Twitter Application Programming Interface[4] van elke politicus in de database uit te zoeken wie hij of zij via Twitter volgt (de *friends*) en deze relaties toe te voegen aan de database. Natuurlijk volgen Tweede Kamerleden ook mensen van buiten de Tweede Kamer, zoals journalisten, webloggers en lokale politici. Al die accounts samen noem ik de 'tweede laag'. Na alle relaties gedownload te hebben, heb ik meteen een deel van de tweede laag weer verwijderd. Als iemand door minder dan drie Tweede Kamerleden wordt gevolgd maakte hij wat mij betreft geen onderdeel uit van de 'Politieke Twittersfeer'. Later zal ik nog meer accounts verwijderen.

Na deze stap stonden er enkele honderden accounts in de database. Ik had echter alleen de friends van Tweede Kamerleden opgehaald, en daardoor misten de onderlinge relaties tussen mensen uit de tweede laag, en wist ik ook niet welke Tweede Kamerleden iemand uit de tweede laag volgt. Daarom heb ik ook van alle mensen uit de tweede laag met vijf of meer volgers in de Tweede Kamer de *friends* gedownload.

Bij deze stap ontstond wel een probleem: namelijk dat er mensen zijn die door heel erg veel mensen uit de tweede laag worden gevolgd, maar door bijna niemand uit de Tweede Kamer (een voorbeeld hiervan is de Amsterdamse politicus Alexander Scholtes, hij heeft slechts vier volgers in de Tweede Kamer, maar maar liefst 81 volgers in de tweede laag). Daarom stelde ik opnieuw criteria: als iemand door mininaal 10% van de accounts (dit komt neer op 46 accounts) in laag 2 wordt gevolgd, behoort hij tot het netwerk. Een probleem daarbij was dat laag 2 veel journalisten bevatte, en dat deze journalisten elkaar nogal intensief volgen. Daarom wilde ik niet alle accounts uit laag 2 die door meer dan 10% gevolgd worden in de index houden, en moeten ze alsnog minimaal 3 volgers uit laag 1 hebben. Kortom, mensen die:

- ofwel 5 of meer volgers in de Tweede Kamer
- ofwel 3 of meer volgers in de Tweede Kamer EN 46 of meer volgers in de tweede laag

hebben, behoren tot het netwerk. Daar komt nog bij dat ze minimaal 10 tweets moeten hebben geplaatst.

Criteria

Hieruit blijkt dus de hoeveelheid beslissingen die ik moest nemen om tot een dataset te komen. Daarbij moest ik ook nog rekening houden met externe factoren, zoals de limieten van de Twitter API. De dataset mocht niet te groot worden, omdat het dan niet meer mogelijk zo zijn om van iedereen een aantal tweets te downloaden. Dat is dan ook een van de redenen om niet gewoon alle accounts te behouden.

De criteria die ik stelde voor accounts om wel behouden te blijven roepen ook vragen op. Het lijkt bijvoorbeeld raar dat ik iemand met 8 volgers in laag 2 en 5 volgers in laag 1 wel verwijderde, terwijl ik dat niet deed met iemand met 4 volgers in laag 1 en 40 volgers in laag 2. Echter, 5 volgers in laag 1 is in principe net zo knap (en zeker politiek gezien relevanter) dan 45 volgers in laag 2: beiden komen neer op ongeveer 10%. De reden dat mensen met weinig volgers in laag 1 en veel volgers in laag 2 vaak geslachtofferd zijn is dat de kans groot is dat ze geen politicus zijn, maar journalist of weblogger.

In bovenstaande wordt dus al heel goed duidelijk dat er bij het samenstellen van een dataset altijd sprake is van keuzes en criteria. Ik hoop geïllustreerd te hebben dat dat niet erg hoeft te zijn, als er maar duidelijk

[3] Ik kan niet garanderen dat er niet enkele fake accounts tussen zitten. Ik heb wel mijn best gedaan om die te filteren.

[4] Met een *application programming* interface (API) bedoel ik software waarmee je toegang hebt tot *raw data* van een externe bron, in dit geval Twitter.

vermeld wordt welke criteria zijn toegepast, en als die criteria maar goed uitgelegd en onderbouwd worden.

Overigens, zelfs als ik had besloten om geen accounts te verwijderen (om het toepassen van criteria te omzeilen), dan nog had ik moeten kiezen hoeveel lagen ik toe zou voegen aan de dataset (alleen friends, of ook friends van friends, of ook friends van friends van friends, etcetera).

Groepen

Terug naar het voorbeeldonderzoek. De dataset die ontstond na het toepassen van alle criteria besloeg 606 accounts, inclusief al hun onderlinge relaties.[5] Van al deze 606 accounts ben ik handmatig gaan bepalen tot welke 'groep' zij behoren: variërend van journalist tot Europarlementariër. Na afloop daarvan ben ik de groepen gaan reduceren; zowel Europarlementariërs als (kandidaat-)Tweede Kamerleden als raadsleden behoren bijvoorbeeld tot de groep 'Politici'. De groepen die ontstonden zijn:

Groep	**Aantal accounts**
Politici Tweede Kamer, Europees parlement, Lokale politici, Oud politici	286
Journalisten TV, kranten, columnisten	135
Overig Headlines, Partijaccounts, Twittermeta, Organisaties	103
Weblogger	29
BN'ers	28
PR Campagnemedewerkers, communicatiedeskundigen	25

Tabel 1 – Aantal accounts per groep

De politici zijn als volgt over de partijen verdeeld (het zijn er in totaal geen 286, omdat van sommige politici de partij niet wordt meegenomen of niet relevant is, zoals Rita Verdonk):

Partij	**TK zetels**	**TK accounts**	**Totaal accounts**
VVD	31	19	92
PvdA	30	13	50
PVV	24	4	5
CDA	21	15	22
SP	15	6	10
D'66	10	7	37
GroenLinks	10	8	50
Christenunie	5	2	7
SGP	2	2	4
Partij voor de Dieren	2	2	2

Tabel 2 – Aantal accounts per politieke partij

Deze tabel geldt vooral ter ondersteuning van het verdere onderzoek. In de grafieken die volgen zullen de kleine partijen waarschijnlijk niet duidelijk naar voren komen, en mogelijk zelfs gecombineerd worden met een grotere partij die dichtbij ligt op het politieke spectrum.

[5] Het opvragen van data kon niet allemaal tegelijk, ik kon slechts 150 requests per uur doen (Twitter policy). Er kunnen daarom veranderingen zijn opgetreden in de tijdspanne dat ik mijn data verzamelde. Er kunnen extra follows zijn ontstaan in de tijd tussen het opvragen van friends uit de eerste laag en die uit de tweede laag. Ik houd hier verder geen rekening mee, omdat dit voor het geheel zeer waarschijnlijk te verwaarlozen is.

We weten nu dus welke groepen vertegenwoordigd worden in de opgebouwde dataset. Er zijn meerdere dingen die opvallen, zoals het hoge aantal journalisten. We kunnen hier alvast uit concluderen dat er veel verbindingen bestaan tussen journalisten en politici. De relatie tussen politiek en pers lijkt dus veranderd. Vroeger werd een persconferentie belegd waar veel journalisten op afkwamen die allemaal enkele vragen konden stellen, of was er sprake van een langer interview tussen journalist en politicus. Nu is het anders: journalist en politicus hebben de mogelijkheid om elkaar direct en zonder enige obstakels iets te vragen of te melden. Of ze dat ook daadwerkelijk doen, dus of de groepen elkaar berichten sturen, zal ik later nog deels proberen te onderzoeken.

Iets anders dat niet ik niet helemaal verwachtte is het aantal BN'ers dat zich in het netwerk bevindt. Voor elk account in de index geldt dat het een minimum aantal volgers moet hebben onder politici, dus ook voor BN'ers. Waarom politici precies BN'ers volgen, en of zij ook met elkaar communiceren, kan een interessant startpunt zijn voor vervolgonderzoek.

Tot slot iets dat niet uit de tabel blijkt: ook de Amerikaanse politici Barack Obama en Al Gore bevinden zich in het netwerk. Het is niet logisch dat zij veelvuldig communiceren met Nederlandse politici en/of journalisten. Twitter wordt dus niet alleen gebruikt om te communiceren, maar ook om een bepaald iemand uitsluitend te volgen. Dit kan ook een mogelijke verklaring zijn voor de BN'ers in het netwerk.

Verbindingen in kaart brengen

Inmiddels is een redelijk hermetisch netwerk ontstaan: er zijn 606 accounts en elk account heeft minimaal 5 eenzijdige verbindingen met een ander account binnen het netwerk. Om beter inzicht te krijgen in de sterkte van de relaties tussen accounts zal ik van elk account een aantal tweets downloaden. (Nog) niet om de inhoud te analyseren, wel om te kunnen kijken wie elkaar 'persoonlijke' berichten sturen.

Duidelijk is dat er binnen het netwerk zowel sprake is van horizontale als verticale verbindingen. Tweede Kamerleden volgen elkaar onderling (horizontaal), maar volgen ook veel mensen uit de tweede laag (verticaal, en anders zou er überhaupt geen sprake zijn van een tweede laag).

Vanuit de database zal ik een soort lijsten maken die kunnen worden ingelezen in het programma Gephi.[6] Met behulp van Gephi kun je sociale netwerken analyseren en bijvoorbeeld bekijken wie er samen communities (*clusters*) vormen. Zo hoop ik met Gephi uit te vinden of partijgenoten communities vormen, en of dat vooral geldt voor volgen of ook voor @replies[7]. Hierbij dient aangemerkt te worden dat Gephi volgens mij vooral informatie geeft over een netwerk in z'n geheel en minder over afzonderlijke *nodes* (nodes zijn in dit geval Twitter accounts). Alle grafieken uit dit paper zijn in groter formaat te downloaden op http://www.martijnweghorst.nl/Scriptie.

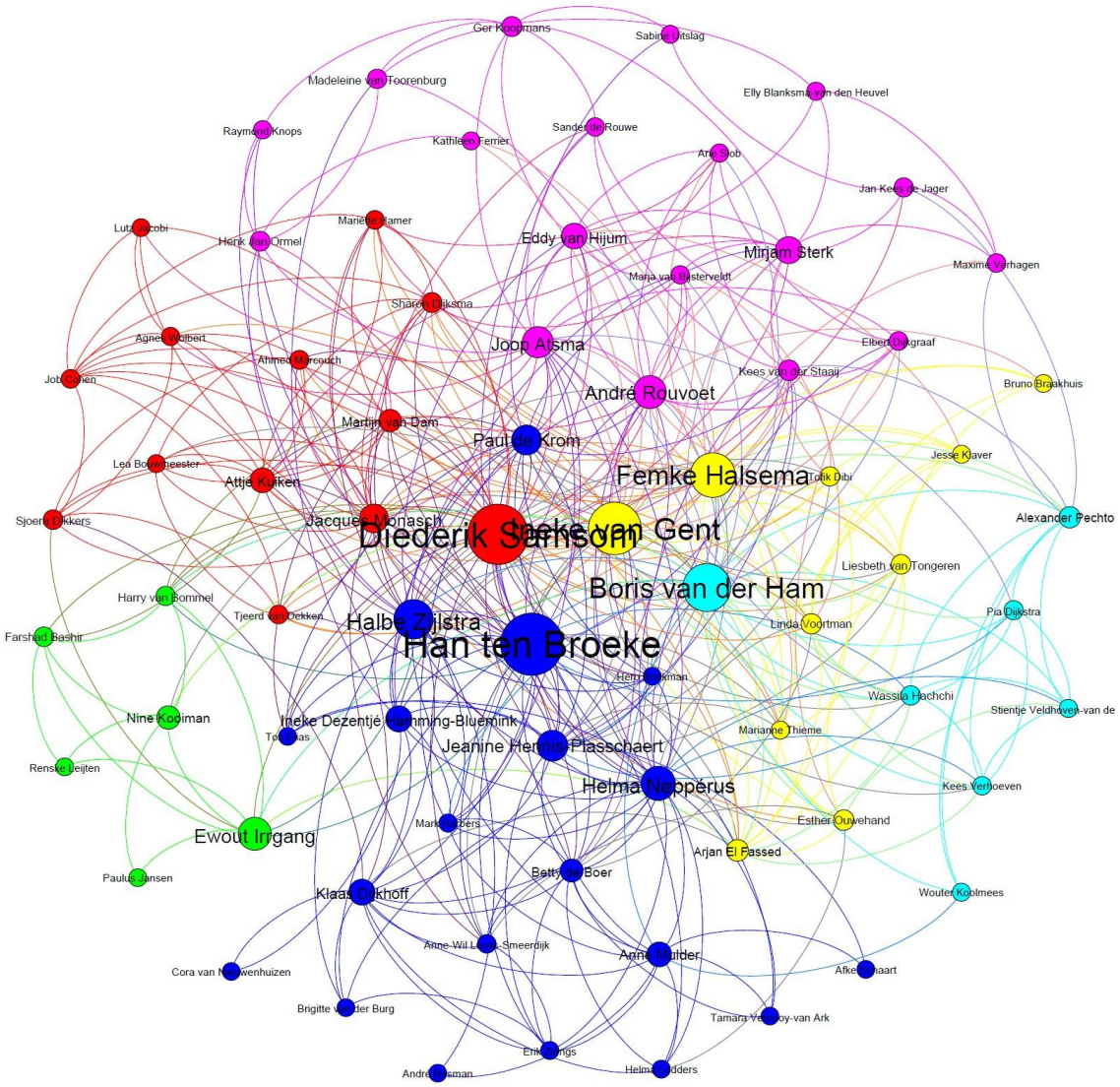
Tweede Kamerleden

Ik ben begonnen met het maken van een dataset van alle Tweede Kamerleden en hun onderlinge *follows.* Bovendien tellen alleen de follows die wederzijds zijn. Een voorbeeld zijn Femke Halsema en Alexander Pechtold: zij zijn beiden Tweede Kamerlid én ze volgen elkaar. Voor deze dataset geldt dat elk account een bepaalde kleur heeft, afhankelijk van de partij waarvoor hij in de Tweede Kamer zit (de kleuren zijn willekeurig gekozen en de partijen zijn gegroepeerd naar hun plek in het politieke spectrum):

VVD, PVV	GroenLinks, PvdD	SP
PvdA	CDA, ChristenUnie, SGP	D'66

[6] Zie voor meer informatie over het maken van zulke lijsten (gml-bestanden) de website van Gephi, gephi.org

[7] Een @reply is een tweet die direct aan iemand anders is gericht.



Grafiek 1 – Tweezijdig volgen in de eerste laag (Fruchterman Reingold)

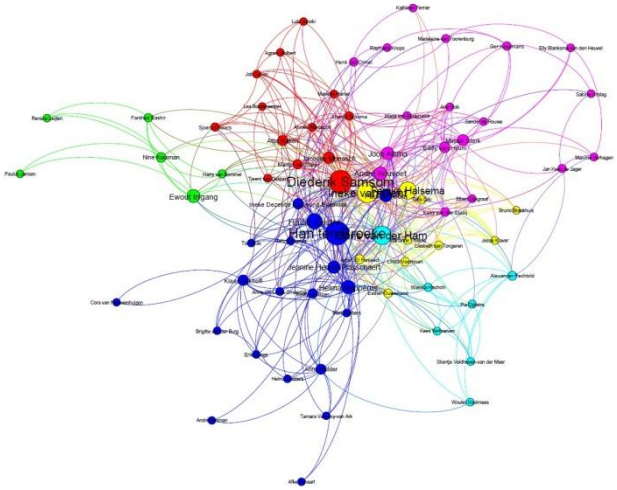
Allereerst iets over de grootte van de afzonderlijke nodes. De grootte is gebaseerd op het aantal verbindingen dat een node heeft en is dus in principe niet afhankelijk van een algoritme. Toch bepaal je tot op zekere hoogte zelf hoe groot een node wordt; Je stelt namelijk de schaal in van groottes. Is de schaal groter, dan is ook de grootste node groter. Ook dit kan dus bedrieglijk zijn.

Uit grafiek 1 blijkt dat Tweede Kamerleden binnen de Tweede Kamer over het algemeen vooral partijgenoten volgen. De vorming van clusters[8] komt bijna een op een overeen met de werkelijke partijen. Hiermee is overigens niet gezegd dat alle politici alleen maar partijgenoten volgen. Het gaat hier immers om wederzijdse, tweezijdige verbindingen: het kan dus best dat iemand veel accounts van buiten de partij volgt, maar dat dat andersom niet het geval is (dat hij/zij niet door diegenen 'teruggevolgd' wordt).

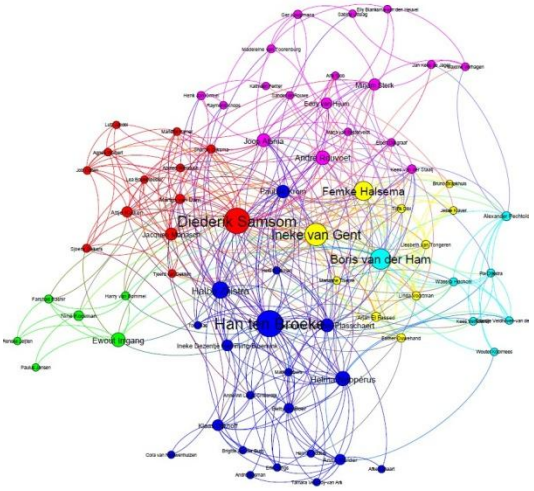
Op het eerste gezicht zijn er enkele uitzonderingen, zoals Boris van der Ham en Paul de Krom. Het is verleidelijk om iets te stellen in de trant van "Van der Ham zit als D66'er tussen de Groenlinks'ers". Dit hoeft echter niet het geval te zijn: Van der Ham kan ook banden hebben met SP'ers en PvdA'ers en op die manier 'naar links getrokken worden'. De posities van individuele nodes in een Gephi grafiek zijn sowieso twijfelachtig: elke keer dat je een grafiek maakt van een bepaalde dataset kunnen er nodes op een andere plek staan. Om daadwerkelijk iets te kunnen zeggen over individuele gevallen zou dan ook verder onderzoek noodzakelijk zijn (als het überhaupt mogelijk is).

[8] Ik heb ook een grafiek gemaakt zonder kleuren vooraf en Gephi clusters laten vormen en de grafiek in laten kleuren. Deze grafiek is te vinden op martijnweghorst.nl als grafiek A.

Iets anders dat grafiek 1 duidelijk maakt is dat er een soort kern bestaat van politici die (veel) meer volgers hebben dan gemiddeld. Hierbij dient echter rekening gehouden te worden met het feit dat grafiek 1 gebaseerd is op een bepaald algoritme, namelijk het *Fruchterman Reingold* algoritme. Van deze grafiek heb ik in navolging van Bernhard Rieder in zijn blogpost *One network and four algorithms* ook nog andere versies gemaakt: één met het *Force Atlas* algoritme en één met het *Yifan Hu* algoritme. De drie grafieken lijken behoorlijk veel op elkaar, maar er is ook een duidelijk verschil: de mate waarin er een kern aanwezig is. Bij het Yifan Hu algoritme was die kern heel erg geconcentreerd, bij het Force Atlas algoritme veel minder. Het lijkt misschien een detail, maar het kan in bepaalde gevallen misschien behoorlijk van invloed zijn op de interpretatie van een grafiek:



Grafiek 2 – Tweezijdig volgen in de eerste laag (Yifan Hu)



Grafiek 3 – Tweezijdig volgen in de eerste laag (Force Atlas)

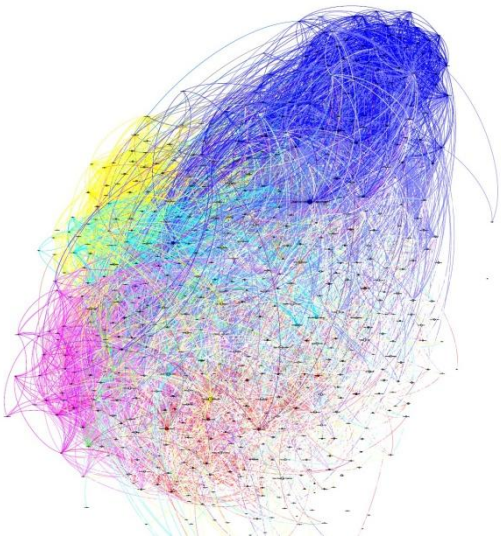
In het vervolg van dit onderzoek zal ik uitsluitend nog gebruik maken van het Force Atlas algoritme. Ik veronderstel het vanaf hier duidelijk dat zo'n algoritme van invloed kan zijn op de resulterende grafiek.

Ik heb ook geprobeerd een grafiek te maken van het aantal @replies dat men aan elkaar stuurde, maar daarvoor had ik simpelweg te weinig data. Er kwamen slechts enkele gevallen naar voren waarbij twee Tweede Kamerleden elkaar een bericht hadden gestuurd in de onderzochte tijdspanne.

Het volledige netwerk

Natuurlijk heb ik ook een grafiek gemaakt van het gehele netwerk, dus inclusief de tweede laag. Het zou kunnen dat de clusters niet meer zo duidelijk zijn, simpelweg omdat in de tweede laag een groot deel van de accounts onafhankelijk is; ze maken geen deel uit van een van de politieke partijen. Het zou echter ook kunnen dat de tweede laag clusters juist wel versterkt, omdat mensen uit dezelfde partij dezelfde externe mensen (bijvoorbeeld bloggers) volgen en omdat in de tweede laag ook nog politici zitten (bijvoorbeeld lokale politici).

Ik heb geprobeerd om deze vraag door middel van een grafiek te beantwoorden. Dat heb ik gedaan door alle politici (in dit geval dus ook lokale politici, Europarlementariërs, etcetera) opnieuw een kleurtje te geven. Alle overige accounts zijn in dit geval wit. Vervolgens heb ik de dataset in Gephi geladen en gekeken wie wie volgt.



Grafiek 4 – Clusterverandering door laag

De grafiek die hier uit voortkwam verschafte enigszins duidelijkheid; De 'tweede laag' bleek het netwerk in ieder geval niet volledig op z'n kop te gooien. Er is immers nog steeds een redelijk onderscheid tussen

kleuren. Echter: geel en turquoise, Groenlinks en D'66 dus, lijken zich behoorlijk (nog meer dan in de vorige grafiek) te versmelten. En hoewel voor individuele gevallen geldt dat er aan een positie niet te veel conclusies verbonden mogen worden, is dat voor hele groepen meer het geval. Lokale politici, Europarlementariërs en andere actieve leden van deze twee partijen volgen elkaar dus relatief intensief en/of volgen veelal dezelfde andere mensen.h

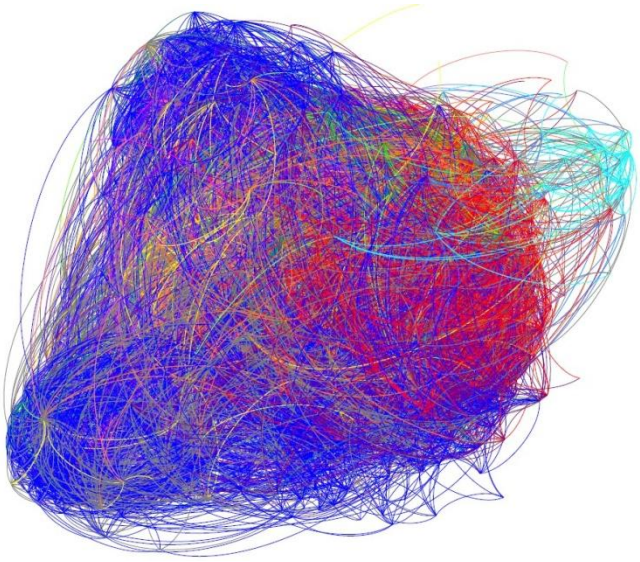
De rode groep, de PvdA, komt in deze grafiek niet zo sterk naar voren, terwijl toch al is gebleken dat er redelijk veel twitterende PvdA'ers zijn. Echter, als er heel erg sterk wordt ingezoomd op de grafiek (zie hiervoor het PDF-bestand op martijnweghorst.nl), blijkt dat de PvdA'ers wel degelijk sterk geconcentreerd zijn. Het lijkt alleen zo te zijn dat de witte lijnen (van de niet-politici) over de rode lijnen heen getekend zijn. Hier blijkt dus weer dat zulke grafieken ontzettend voorzichtig geïnterpreteerd dienen te worden en dat er veel zaken zijn die de grafiek ongewenst kunnen beïnvloeden.

Clustering in het volledige netwerk

We hebben nu gezien dat Tweede Kamerleden via Twitter vooral partijgenoten volgen. De vraag is of dat ook op het volgende 'niveau' geldt: volgen politici vooral andere politici, journalisten vooral journalisten, etcetera? Ik heb dus voor het gehele netwerk gekeken of er sprake is van een soort clustering, dit keer in het geval van groepen (politici, journalisten, webloggers, etc.). Opnieuw gaf ik elk account een kleurtje:

Politici	Overig	PR
Journalisten	Webloggers	BN'ers

Grafiek 5 lijkt misschien nogal chaotisch, maar hij geeft wel degelijk enkele dingen duidelijk weer. Zo vormen de BN'ers een beetje een apart groepje aan de buitenkant van het netwerk. Het zou kunnen dat die BN'ers met relatief weinig anderen een 'vriendschap' (een tweezijdige relatie) hebben, behalve dan met andere BN'ers. In dat geval zou er sprake kunnen zijn van een *two-step flow of communication*; er moeten immers bepaalde personen zijn die een soort tussenstap vormen tussen BN'ers en politici en journalisten. Wie die personen precies zijn zou onderwerp kunnen zijn van vervolgonderzoek.



Grafiek 5 – Wederzijds volgen door groepen

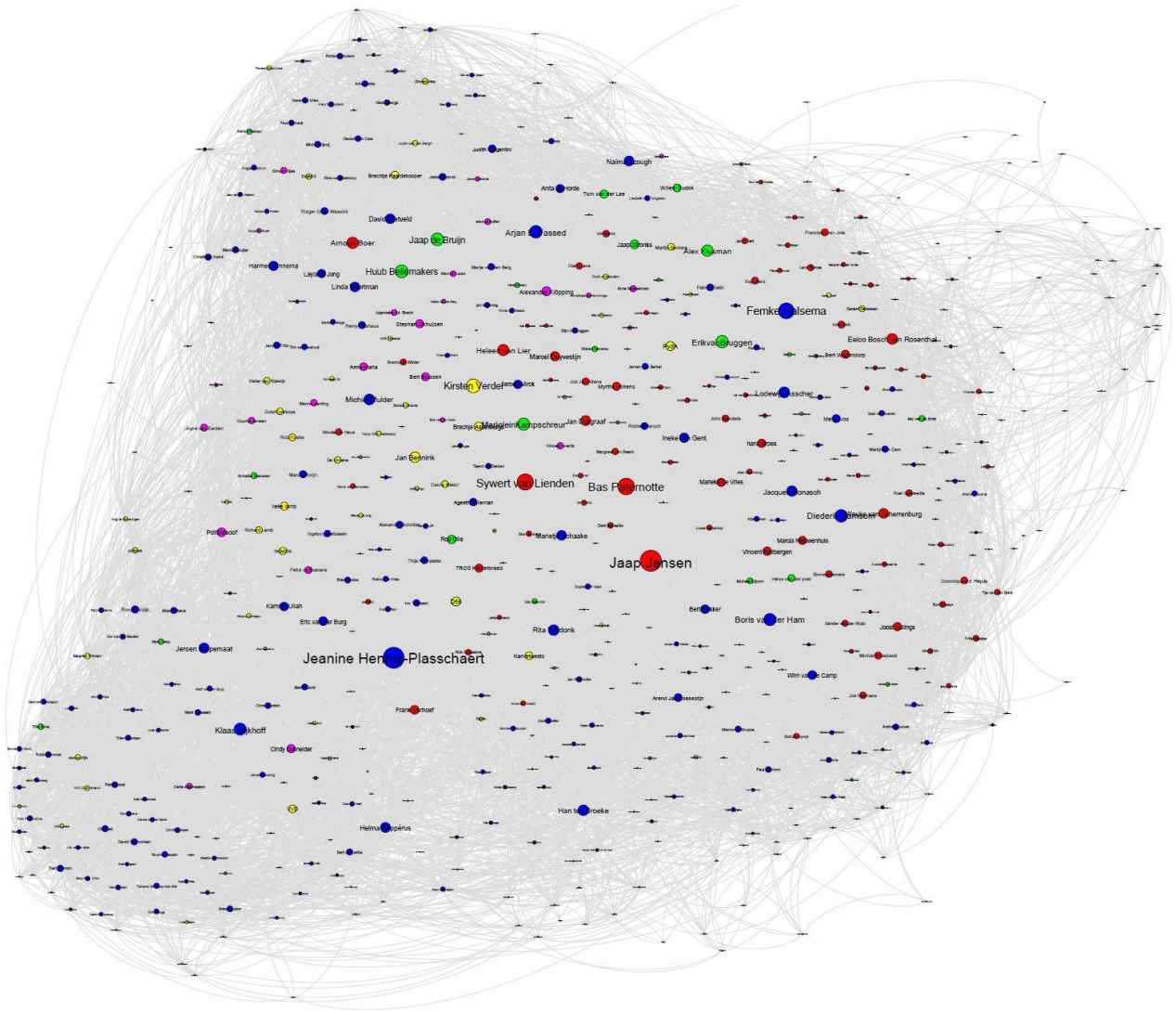
Ook hier geldt echter dat de Gephi grafiek niet zonder meer uitsluitsel geeft over de reden van de positie van de BN'er groep. Er is namelijk ook een tweede mogelijke verklaring: namelijk dat BN'ers in elke groep ongeveer evenveel vrienden hebben en dus niet extra goed in een van de bestaande groepen passen.

Iets anders dat opvalt in grafiek 5 is de rode vlek van journalisten. Hoewel het natuurlijk duidelijk is dat journalisten ook in contact staan met politici, lijkt de rode vlek aan te geven dat journalisten relatief vaak elkaar volgen. We kunnen hier voorzichtig uit concluderen dat accounts toch weer redelijk gegroepeerd blijven volgens de offline wereld. Dit zou kunnen bevestigen dat mensen blijven communiceren met anderen uit hun eigen groep, zoals dat ook al bleek bij partijgenoten. Een uitzondering hierop vormen de gele accounts, maar dat klopt omdat dat alle 'overige' accounts zijn en die dus ook niet samen een gesloten groep vormen. Dit is dus meer een bevestiging dan een ontkrachting.

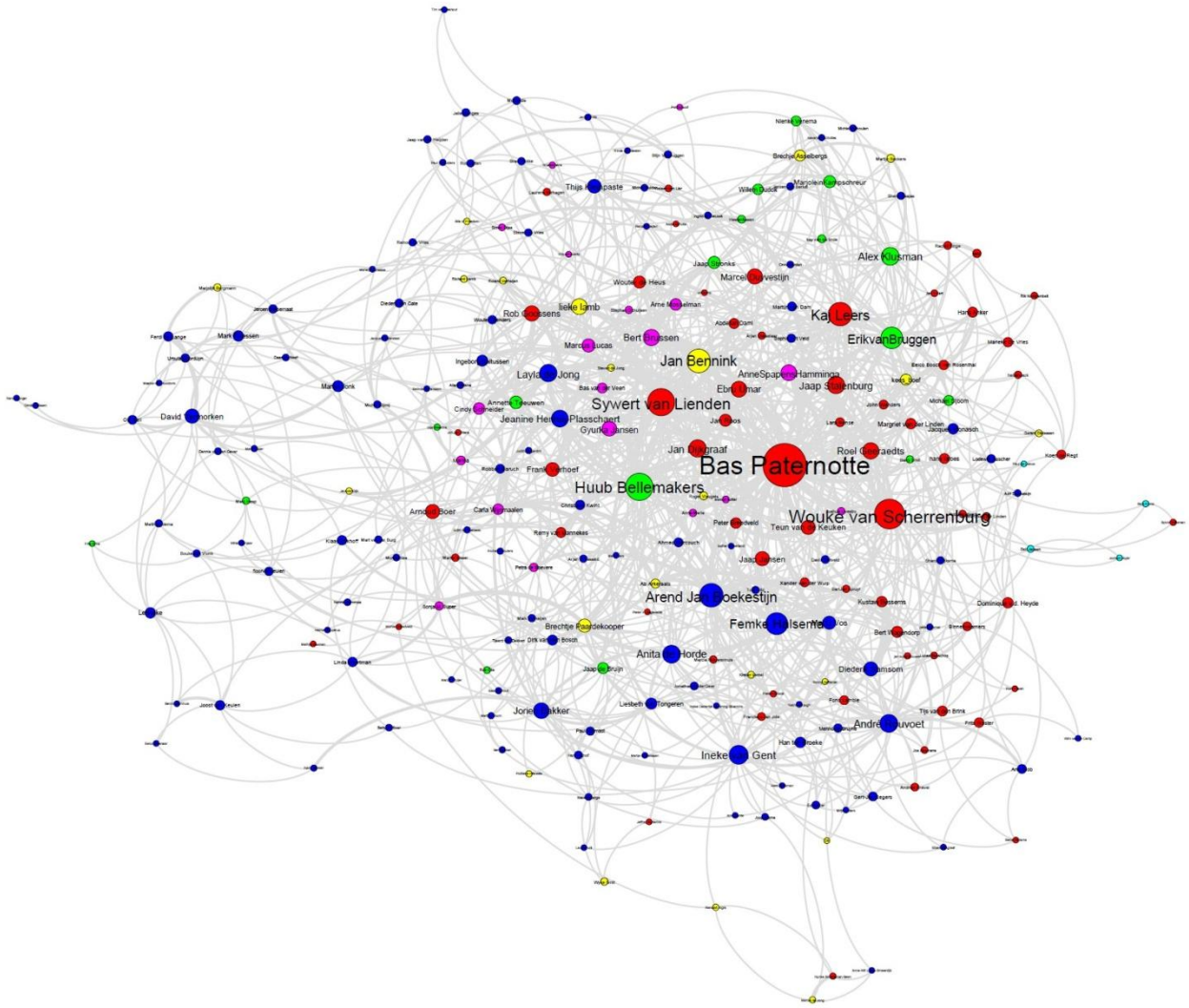
Tweet en volggedrag in het volledig netwerk

Tot nu toe heb ik me in het voorbeeldonderzoek alleen geconcentreerd op *friends* en *follows*, dus op wie wie volgt. Dat iemand een ander volgt wil echter nog helemaal niet zeggen dat die personen ook daadwerkelijk met elkaar communiceren via Twitter. Daarom heb ik twee vergelijkende grafieken gemaakt van het hele netwerk: een die gebaseerd is op wederzijds volgen, en een die gebaseerd is op wederzijds tweeten. Als blijkt dat er verschillen zijn tussen die grafieken is er dus een verschil tussen wie wie volgt en wie wie berichten stuurt.

Grafiek 6 is dezelfde als grafiek 5, maar dan uitvergroot, met de namen van accounts erbij en met grijze verbindingslijnen in plaats van gekleurde. Grafiek 7 heb ik gemaakt met behulp van een dataset waarin een relatie bestaat wanneer twee accounts een @reply naar elkaar hebben gestuurd. Beide grafieken hebben dezelfde kleuren als grafiek 5.



Grafiek 6 – Wederzijds volgen binnen het hele netwerk



Grafiek 7 – Wederzijds tweeten binnen het hele netwerk

De eerste observatie is dat er veel meer nodes zijn in grafiek 6. Dat is logisch, niet iedereen die elkaar volgt heeft elkaar in de afgelopen twee weken een bericht gestuurd. Wat wel vreemd is, is dat er zich in de kern van deze grafiek amper tot geen politici bevinden. Ik heb hier geen goede verklaring voor, verder onderzoek zou hier eventueel meer duidelijkheid over kunnen verschaffen.

Daarnaast lijken de journalisten wat beter gerepresenteerd te worden in grafiek 7 dan in grafiek 6. Dit duidt er op dat journalisten relatief veel contact hebben, en niet slechts passief volgen. Van deze journalisten, en ook van het gehele netwerk, is Bas Paternotte in grafiek 7 veruit de grootste node. Bovendien is hij in grafiek 7 relatief[9] veel groter dan in grafiek 6. Dit duidt er op dat deze journalist van onder andere HP/De tijd ontzettend veel twittert en in hoge mate contact zoekt met andere twitteraars. Hetzelfde geldt voor andere nodes die in grafiek 7 relatief groter zijn dan in grafiek 6, zoals Wouke van Scherrenburg en Arend Jan Boekestijn.

Deze 'ontdekking' van Bas Paternotte is puur te danken aan de opzet van dit onderzoek; Zoiets was ik vrijwel onmogelijk te weten gekomen bij ‘traditioneel’ onderzoek. Het account van Paternotte kwam automatisch bovendrijven door het downloaden van follows/friends en tweets en door daar grafieken van te maken, en niet omdat er al een bepaalde verwachting over hem was. Dit is dus een voordeel van de data-miningachtige methode die ik voor dit voorbeeldonderzoek heb gebruikt.

[9] Het woord relatief is van belang: zoals eerder uitgelegd bepaal je tot op zekere hoogte zelf hoe groot een node wordt.

Fase 2 – Waarover gaat het (*Issue-analyse*)

In fase 1 heb ik gekeken naar visualisaties van sociale netwerken. Dat is echter niet het enige dat gedaan wordt met Twitter data. Er wordt bijvoorbeeld ook gekeken naar de inhoud van tweets. Een voorbeeld zijn zogenaamde Twitter *StreamGraphs*[10]. Voor zo'n grafiek worden de laaste 1000 tweets met een bepaald woord erin gedownload. Vervolgens wordt geanalyseerd op welk moment de meeste van die tweets werden geschreven en met welke woorden het opgegeven het woord het meest samen voorkomt.

In deze fase van het voorbeeldonderzoek wil ik iets soortgelijks maken. Ik wil echter niet een grafiek maken op basis van één woord, maar op basis van alle hashtags[11] die door personen uit het netwerk van fase 1 gebruikt werden. Ik zal een grafiek maken waarop de aandacht voor een issue wordt uitgezet tegen de tijd. Helaas moet ik me door de limieten van de Twitter API wel beperken tot een momentopname.

Omdat Twitter maar toegang geeft tot de laatste 3200 tweets van een gebruiker kon ik maar tweets downloaden van een periode van twee weken. Anders zou ik van bepaalde accounts (bijvoorbeeld Bas Paternotte) niet alle tweets hebben (omdat hij meer dan 3200 berichten plaatste in een ruimere tijdspanne).[12] Daardoor heb ik van elk account in de database alle tweets die hij/zij tussen 16 september en 1 oktober 2010 schreef gedownload.

Aanvankelijk was ik van plan om issues te destilleren aan de hand van woorden. Ik wilde elke tweet opdelen in de gebruikte woorden, en vervolgens een index maken van alle gevonden woorden en het aantal keer dat ze gebruikt werden. Het bleek echter makkelijker om uit te gaan van hashtags. Niet iedereen in het netwerk maakt gebruik van hashtags, maar het bleek voldoende om issues te vinden. Het gevolg is wel dat tweets die wel over een issue gingen, maar waar geen hashtag in gebruikt werd, niet meegenomen zijn in de analyse.

Hashtags

Zoals gezegd zou ik issues proberen te vinden aan de hand van hashtags. Het was dus zaak om van elke tweet die ik had gedownload de gebruikte hashtag(s) te vinden. Daardoor kreeg ik een index van alle hashtags die gebruikt werden tussen 15 september en 1 oktober. Van al deze hashtags verwijderde ik degenen die ofwel minder dan 25 keer gebruikt waren, ofwel door minder dan 5 verschillende gebruikers gebruikt waren. Dit leverde een index op van 93 verschillende hashtags. Van elke overgebleven hashtag heb ik vervolgens bepaald hoe vaak deze in totaal per dag was gebruikt, en door hoeveel verschillende gebruikers, om eventuele pieken te kunnen onderscheiden.

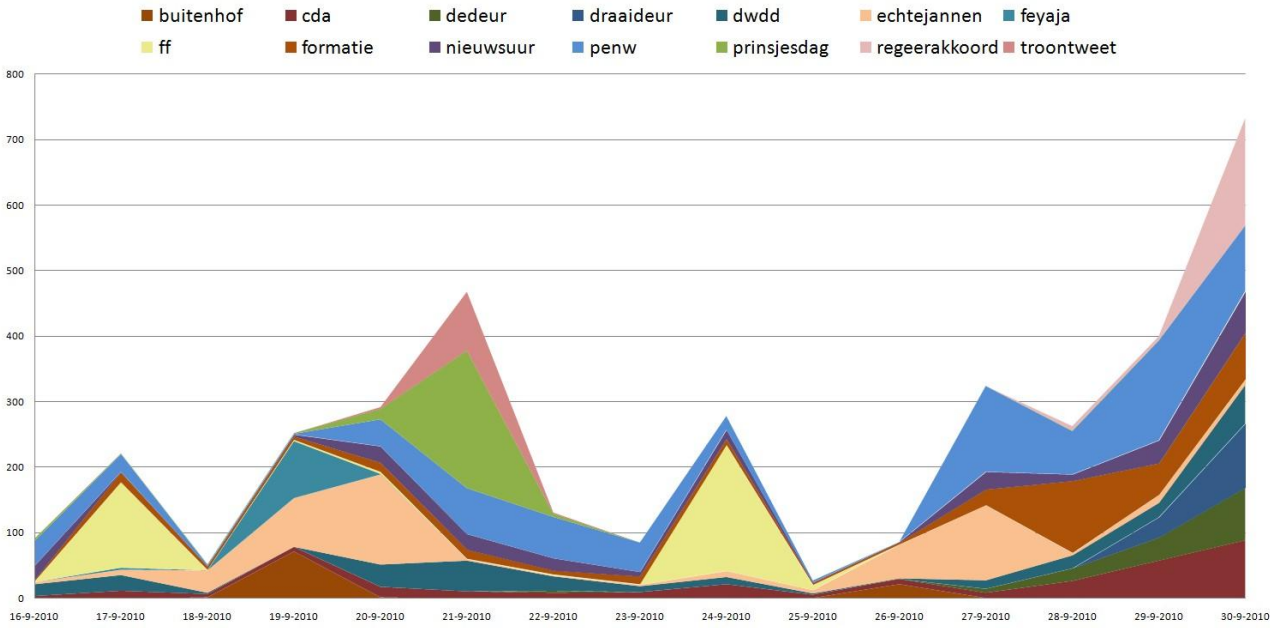
Deze hele index van het gebruik van hashtags per dag heb ik vervolgens geïmporteerd in een Microsoft Excel 2007 sheet.[13] In die sheet maakte ik twee tabbladen aan: eentje waarbij van elke hashtag per dag werd aangegeven hoe vaak 'ie in totaal gebruikt werd, en eentje waarbij per dag werd aangegeven door hoeveel verschillende gebruikers 'ie gebruikt werd. Van beide lijsten maakten ik een grafiek. Voor de grafiek van het totale gebruik werden alleen hashtags meegenomen die op een dag meer dan 50 keer waren gebruikt. Voor de grafiek van het gebruik per gebruiker werden alleen hashtags meegenomen die op een bepaalde dag door meer dan 20 verschillende gebruikers waren gebruikt.

[10] Zie voor meer informatie http://www.neoformix.com/Projects/TwitterStreamGraphs/view.php

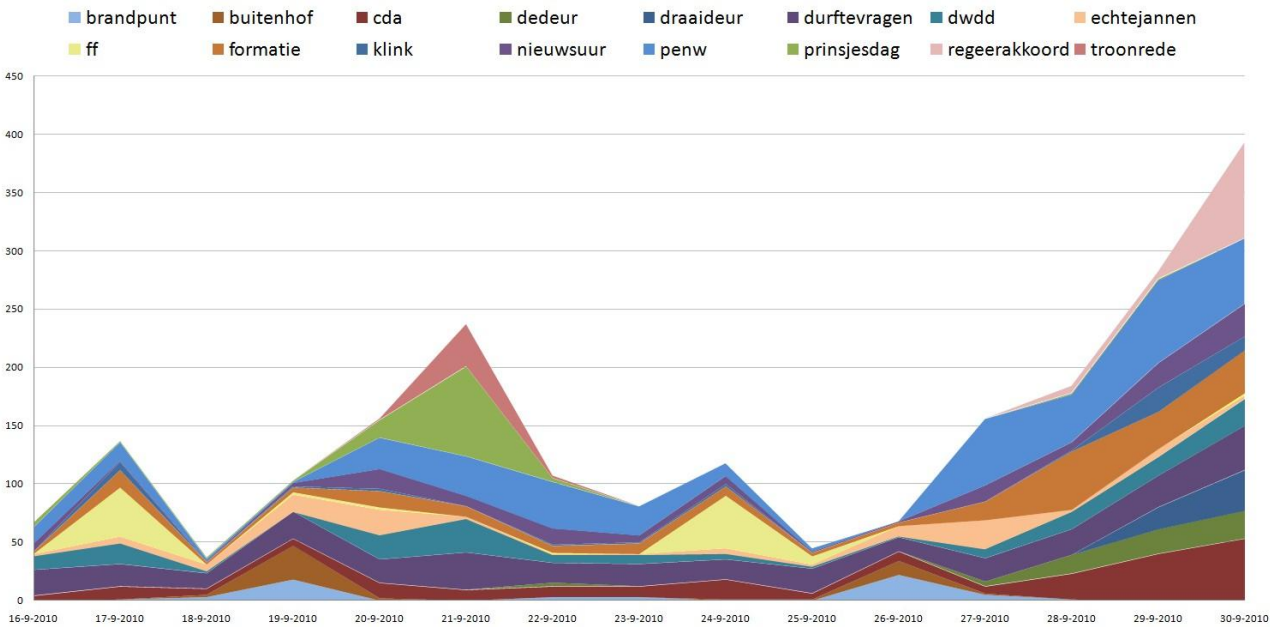
[11] Hashtags worden op Twitter gebruikt om het onderwerp van een tweet aan te geven (bijvoorbeeld dwdd voor het TV-programma De Wereld Draait Door)

[12] Aanvankelijk was het idee was om in deze fase een vergelijkbaar onderzoek te doen als Rogers in hoofdstuk 5 van *Information Politics on the Web*, en dus groepen/accounts aan issues te koppelen in een Gephi(-achtige) grafiek. Dat bleek door de API-restricties en tijdgebrek echter onmogelijk.

[13] Deze sheet kan ook worden gedownload op http://www.martijnweghorst.nl/Scriptie/Hashtags.xlsx



Grafiek 8 – Gebruik van hashtags



Grafiek 9 – Gebruik van hashtags gegroeperd per gebruiker

Beide grafieken kennen dezelfde twee grootste pieken: op 21 september en op 30 september. Deze data komen overeen met belangrijke gebeurtenissen. Dit jaar viel op 21 september Prinsjesdag, en 30 september was vlak voor het belangrijke CDA-congres over regeringsdeelname. Dat blijkt ook wel, want de hashtags die op 21 september werden gebruikt waren vooral *prinsjesdag* en *troonrede* (de bovenste twee vlakken), en op 30 september vooral *regeerakkoord*, *formatie* en *cda* (bovenste, onderste en oranje vlak).

Ook *dedeur* en *draaideur* werden richting 30 september veel gebruikt, en deze woorden hadden ook iets met de formatie te maken.[14]

De twee kleine piekjes zijn te wijten aan het fenomeen Follow Friday (hashtag ff), waarbij twitteraars op een vrijdag andere twitteraars aanbevelen aan hun vrienden. De overige hashtags die vaak en door veel verschillende twitteraars gebruikt werden zijn vooral TV-programma's: Brandpunt en Buitenhof op zondag en De Wereld Draait Door, Nieuwsuur en Pauw en Witteman op werkdagen.

Vergelijking

Bij een vergelijking van de grafieken 8 en 9 valt op dat er in grafiek 8 op 19 september een soort piek is die in grafiek 9 ontbreekt. Dit komt met name door de hashtag *feyaja* (daarmee werd de voetbalwedstrijd Feyenoord – Ajax die op die dag werd gespeeld aangeduid). Er werd dus op die dag wel veel over die wedstrijd getwittert, maar niet door veel verschillende mensen. Ook de Follow Friday pieken zijn in grafiek 8 een stuk groter. Dat kan komen omdat er een paar mensen zijn die het heel veel gebruiken, of omdat iedereen die het gebruikt dat een paar keer doet.

We hebben dus kunnen zien dat de twee grafieken enkele verschillen kennen. En hoewel het in dit geval niet echt andere conclusies zou opleveren, had dat wel gekund. Ook hier dient dus van te voren heel goed nagedacht te worden over de samenstelling en het gebruik van de dataset. Van de twee voorbeelden is de laatste zeer waarschijnlijk betrouwbaarder, omdat het dan niet zo kan zijn dat een issue *hot* is simpelweg omdat één iemand er veel over twittert.

Tot slot: zoals gezegd stelde ik criteria op voor de hashtags. Ze moesten minimaal één dag een bepaald aantal keer gebruikt zijn. Het zou kunnen dat die criteria ook nog van invloed zijn geweest en dat zonder die criteria er nog een piek was geweest. Dat zou gekund hebben als er een thema was waarvoor veel verschillende hashtags een paar keer gebruikt werden, in plaats van één hashtag heel vaak.

Conclusie

Ik heb in dit paper geprobeerd om aan de hand van een voorbeeldonderzoek een kritische benadering ten opzichte van nieuwe onderzoeksmethoden te geven. In fase 1 beschreef ik enkele mogelijke problemen bij het visualiseren van een sociaal netwerk. Ten eerste is zo'n visualisatie in veel gevallen gebaseerd op een dataset die een onderzoeker van te voren kiest en/of vormgeeft. De resultaten zijn dus afhankelijk van criteria die aan de dataset gesteld zijn en het is dus zaak om deze criteria duidelijk te beschrijven en te onderbouwen. Een mogelijke invloed van de keuze van de dataset bleek in de grafieken 6 (gebaseerd op follows) en 7 (gebaseerd op @replies).

Daarnaast is het in Gephi noodzakelijk om een algoritme te kiezen. Dat algoritme bepaalt uiteindelijk mede de positie van afzonderlijke nodes in de grafiek. Ook zo'n algoritme is dus van invloed op de uiteindelijke visualisatie. Een voorbeeld daarvan gaf ik in de grafieken 1, 2 en 3. Er was duidelijk zichtbaar dat het ene algoritme een grafiek met een duidelijkere kern opleverde dan het ander.

In fase 2 ging ik wat dieper in op het identificeren van veelbesproken issues binnen een netwerk. Ook daar was sprake van een criterium: namelijk hoe vaak een *hashtag* gebruikt moest zijn om in de grafiek te komen. En ook hier bleek dat de constructie van de dataset van belang is: het 'totale' gebruik van hashtags verschilde van het gebruik dat gegroepeerd was per account.

Kortom: de conclusies die uit zulke grafieken getrokken kunnen worden zijn lang niet altijd puur feitelijk, maar worden aangevuld met de eigen ideeën, kennis en interpretaties van de onderzoeker. En niet alleen in de beginfase, maar gedurende het hele onderzoek vinden beslissende momenten plaats: hoe wordt een dataset geïnitieerd, aan welke criteria wordt data vervolgens getoetst, welke algoritmen worden gebruikt

[14] Dedeur en draaideur waren hashtags die gebruikt werden om over het CDA-fractieoverleg over eventuele regeringsdeelname (dat achter gesloten deuren plaatsvond, vandaar) te twitteren.

om een grafiek te genereren, etcetera. Al deze momenten kunnen van bepalende invloed zijn op de resulterende grafieken en dus op conclusies die getrokken worden.

Tot slot wil ik nog opmerken dat dit paper dan wel een kritische reflectie is, maar geenszins een poging om onderzoek zoals beschreven als volledig nutteloos weg te zetten. Want hoewel visualisaties met de nodige voorzichtigheid gebruikt dienen te worden, kunnen ze denk ik wel degelijk een ondersteuning van argumenten vormen. Bovendien bleek op verschillende momenten dat er onverwachte dingen aan het licht kwamen als gevolg van de gekozen methoden. Twee voorbeelden daarvan zijn de groep BN'ers die blijkbaar een bijzonder soort relaties hebben met politici en de 'ontdekking' van Bas Paternotte.

Bronnen

Rieder, Bernhard. 2010, 10 6. *One network and four algorithms*. Opgeroepen op 9 November 2010, van The Politics of Systems: http://thepoliticsofsystems.net/2010/10/06/one-network-and-four-algorithms/

Rogers, Richard. 2004. *Information politics on the web.* Cambridge: MIT Press.