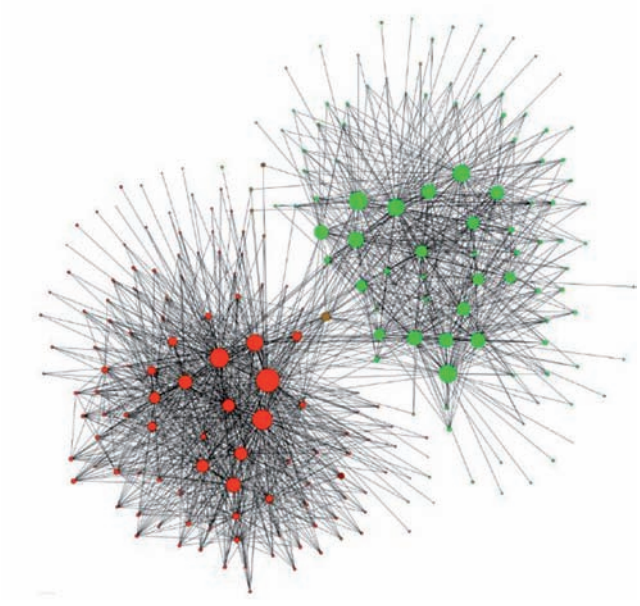
Clara Stegehuis
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
c.stegehuis@tue.nl

Maatschappij Faces of Science

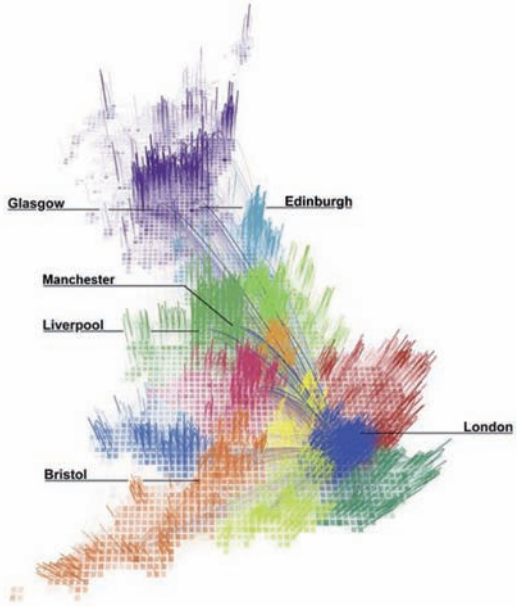
Netwerken met groepsstructuren

Op de website Facesofscience.nl bloggen jonge talentvolle wetenschappers over hun ervaringen en passies. Zo laten ze aan een breed publiek zien hoe de wetenschappelijke wereld eruitziet. Faces of Science is een initiatief van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), De Jonge Akademie en NEMO Kennislink. Sinds vorig jaar is een van de deelnemende bloggers Clara Stegehuis. Zij is promovenda aan de Technische Universiteit Eindhoven in de vakgroep Stochastiek, onder begeleiding van Remco van der Hofstad en Johan van Leeuwaarden. Ze doet hier onder andere onderzoek naar hoe groepsstructuren in netwerken de verspreiding van epidemieën beïnvloeden. In dit artikel vertelt zij over haar onderzoek.

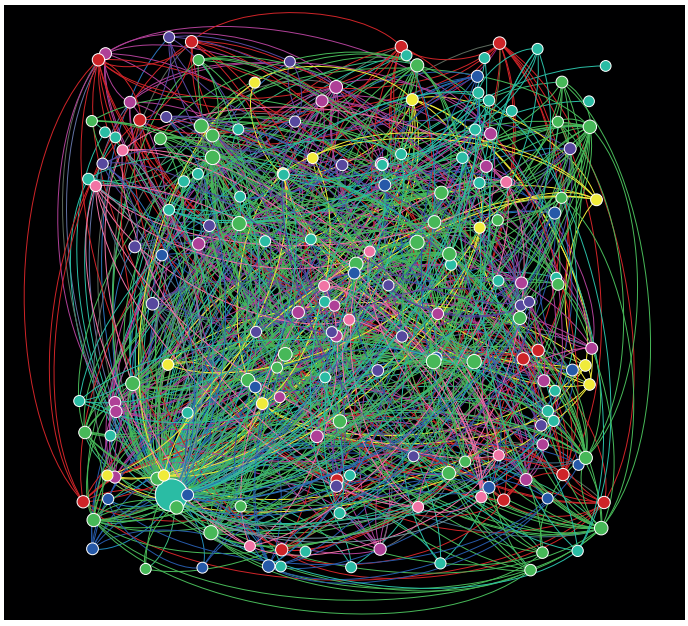
Complexe netwerken kun je overal terugvinden. Je kunt bijvoorbeeld denken aan netwerken in je brein, netwerken op het internet, netwerken op sociale media of communicatienetwerken. Een voorbeeld van een communicatienetwerk is te zien in Figuur 1, die het telefoonnetwerk van België laat zien. Punten in dit netwerk zijn Belgen, en er is een lijn tussen twee Belgen als ze elkaar ooit een keer gebeld hebben. Het is meteen duidelijk dat België verdeeld is in twee groepen. Een groep van Franstaligen, en een groep van Vlamingen, die onderling vaak bellen. Tussen de twee groepen zijn maar weinig lijnen, wat niet zo gek is gezien de taalbarrière. Maar ook landen zonder taalbarrière kunnen verdeeld zijn in groepen. In het plaatje van het telefoonnetwerk van Groot-Brittannië in Figuur 2 zijn vergelijkbare groepsstructuren te zien. Mensen uit dezelfde provincie bellen elkaar veel vaker dan mensen



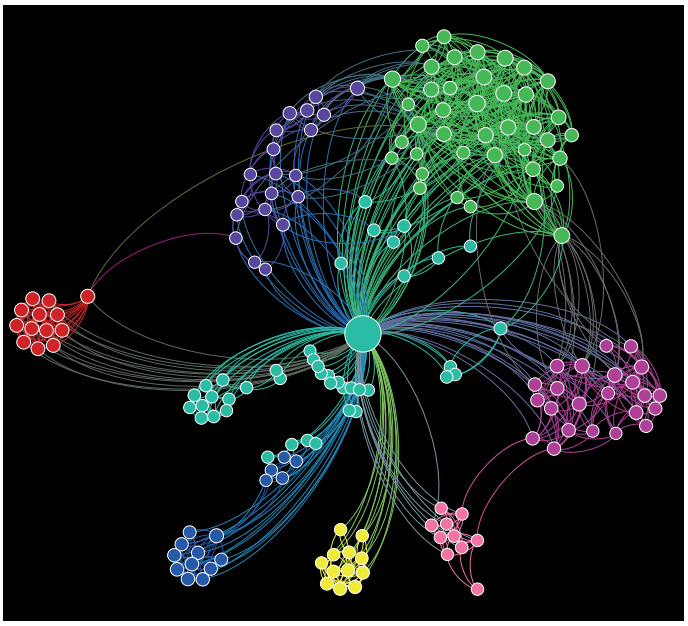
Figuur 1 Belgisch telefoonnetwerk [1].



Figuur 2 Engels telefoon netwerk [4].



Figuur 3 Mijn LinkedIn-netwerk.



Figuur 4 Mijn LinkedIn-netwerk, wat logischer geordend.

uit andere provincies. Londen verbindt alle verschillende groepen in het telefoonnetwerk met elkaar.

Het interessante is dat heel andere soorten netwerken vergelijkbare groepsstructuren zien. In Figuur 3 zie je bijvoorbeeld een plaatje van mijn LinkedIn-netwerk. Punten in dit netwerk zijn mijn LinkedIn-connecties, en een lijn betekent dat twee van mijn vrienden ook met elkaar bevriend zijn. Op het eerste gezicht lijkt dit netwerk een grote chaos. Toch zien we in Figuur 4 dat het netwerk verborgen groepsstructuren bevat. Maar hoe kun je deze groepsstructuren wiskundig uit de chaos van het vorige plaatje halen? Dit is een moeilijke vraag waar onder andere wiskundigen, natuurkundigen en informatici zich het afgelopen decennium over gebogen hebben. Het vinden van deze verborgen groepsstructuren heeft ook veel praktische toepassingen, doordat groepsstructuren in allerlei verschillende soorten netwerken voorkomen. Je kunt bijvoorbeeld denken aan adverteerders in sociale netwerken die zich willen richten op bepaalde groepen van mensen met dezelfde interesses of woonplaats.

Wat is een groep?

Om te ontdekken waar de groepen zich in een netwerk bevinden, moet eerst duidelijk zijn wat een groep precies is. Als we naar de bovenstaande plaatjes kijken, is het intuïtief duidelijk dat een groep aan twee eigenschappen voldoet:

- Er zijn relatief veel verbindingen tussen punten in dezelfde groep.
- Er zijn relatief weinig verbindingen tussen punten in verschillende groepen.

Eén van de eerste methoden om groepsstructuren in netwerken te vinden gebruikt precies deze twee eigenschappen van groepen. Deze methode maximaliseert de *modulariteit* [3]. Stel dat we een netwerk op een bepaalde manier opdelen in K groepen. De *modulariteit* M van deze specifieke opdeling is dan

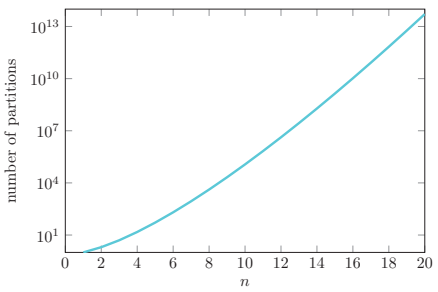
$$M = \sum_{c=1}^{K} \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2.$$

Hier is L het totale aantal lijnen in het netwerk en L_c het aantal lijnen binnen groep c. k_c is de som van de graden van alle punten in groep c, waar de graad van een punt het aantal lijnen is dat met dat punt verbonden is. Als we de modulariteit willen maximaliseren, zegt de eerste term in deze vergelijking dus dat we zo veel mogelijk lijnen binnen de groepen willen hebben. Maar dit alleen is niet genoeg. Als we alleen het aantal lijnen binnen de groepen maximaliseren, dan is het altijd het beste om het netwerk in één grote groep met alle punten erin te verdelen. Dan zitten alle lijnen uit het netwerk namelijk binnen een groep. Daarom heeft de vergelijking ook nog een tweede term. Deze tweede term beschrijft het verwachte aantal lijnen binnen groep c als we het netwerk opnieuw zouden tekenen met hetzelfde aantal lijnen en dezelfde graden, maar de lijnen compleet willekeurig met elkaar verbinden. Modulariteit meet dus hoe veel meer lijnen er binnen de groepen zijn dan dat je eigenlijk verwacht had.

Veel manieren om een netwerk op te delen

De vergelijking voor modulariteit kan voor iedere verdeling van een netwerk in groepen berekenen hoe goed die verdeling is. Maar helaas lost dit het probleem van het vinden van de groepsstructuren nog niet op. Een naïeve manier om de beste opdeling in groepen te vinden is om voor alle mogelijke opdelingen te berekenen wat de modulariteit is, en daaruit de opdeling met de hoogste modulariteit te kiezen. Figuur 5 laat zien hoe veel manieren er zijn om een netwerk met n punten op te delen in groepen.

Dit maakt meteen wel duidelijk dat we nooit alle mogelijke manieren om een netwerk in groepen te verdelen af kunnen gaan: bij een netwerk met maar 20 punten



Figuur 5 Het aantal manieren om een netwerk van n punten in groepen op te delen.

moet je al voor 10^{13} verschillende opdelingen de modulariteit berekenen! Het is voor grote netwerken dus onmogelijk om de opdeling met de hoogste modulariteit te vinden. Gelukkig zijn er slimmere methoden gevonden om toch een goede opdeling van het netwerk te krijgen.

Een voorbeeld hiervan is om te beginnen met het opdelen van het netwerk in n groepen, een groep voor ieder punt. Dan kijkt het algoritme voor alle paren van twee groepen of de modulariteit groter wordt als de punten uit die twee groepen samen in één groep worden gestopt. Zo ja, dan worden de punten uit deze twee groepen inderdaad samengevoegd tot één groep. Hierna is het netwerk opgedeeld in grotere groepen. Daarna kijkt het algoritme weer of het samenvoegen van twee groepen soms een grotere modulariteit oplevert. Dit gaat zo door tot het samenvoegen van twee groepen niet meer leidt tot een hogere modulariteit.

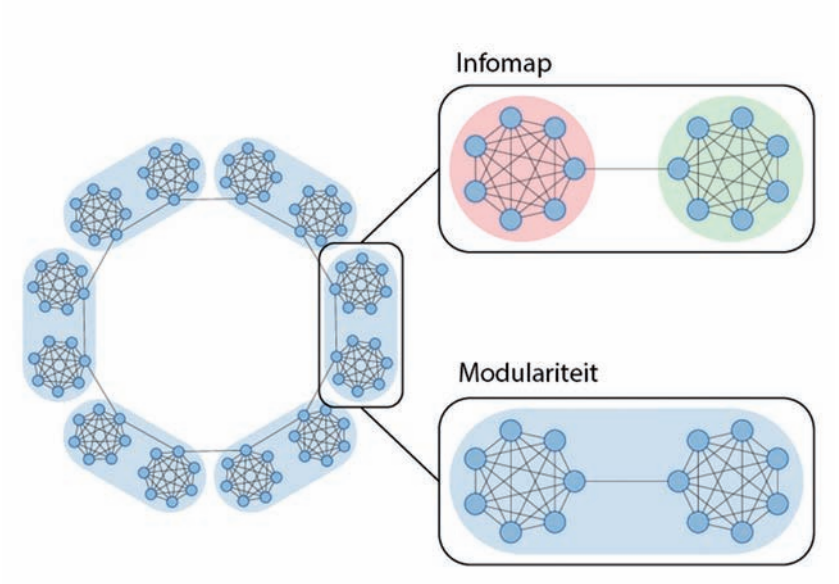
Resolutielimiet

Ondanks deze slimme methode om groepen met hoge modulariteit te vinden, is er nog een probleem met modulariteit. Het blijkt namelijk dat de modulariteit in een netwerk met n punten het hoogst is als er alleen maar groepen zijn met minstens $\sqrt{n}$ punten [2]. Deze beperking wordt ook wel de resolutielimiet genoemd. Stel dat we in het netwerk van alle 2 miljard Facebookgebruikers zoeken naar groepen, dan moeten deze groepen dus minimaal vijftigduizend personen bevatten! Dit is veel groter dan iedere vriendengroep op Facebook die je eigenlijk in het netwerk zou willen terugvinden. Figuur 6 laat een ander voorbeeld zien waar deze resolutielimiet ervoor zorgt dat de juiste groepsstructuur niet teruggevonden wordt met modulariteit. Als we de grote ring groot genoeg maken, dan bevat het netwerk te weinig punten om de kleine groepen nog te kunnen ontdekken, omdat die kleiner zijn dan $\sqrt{n}$.

Een kaart van het netwerk

Een meer recente methode om groepen in netwerken te vinden heeft deze resolutielimiet niet. De *Infomap*-methode [5] bekijkt een random wandeling over het netwerk: een persoon die van punt naar punt over de lijnen van het netwerk loopt. Iedere keer kiest hij met gelijke kans een van de buren van het punt waar hij nu is om naartoe te gaan. Als een netwerk groepsstructuren bevat, dan verwacht je dat deze *random walker* relatief veel tijd doorbrengt binnen een groep voordat hij deze verlaat, omdat groepen relatief veel lijnen bevatten. Tussen de verschillende groepen zijn veel minder lijnen, de random walker zal dus niet zo vaak van groep naar groep bewegen. Maar hoe gebruik je deze intuïtie om de groepsstructuren te ontdekken?

Stel dat je in woorden wil beschrijven hoe een random walker zich door het netwerk heen beweegt. Om dit te doen, moet je ieder punt in het netwerk een naam geven. Je beschrijft bij elke stap de naam van het punt waar de walker zich naartoe beweegt. In een groot netwerk heb je dan veel verschillende namen nodig. Een slimmere manier om de route te beschrijven is zoals een route op een kaart. Niet alle straten in Nederland hebben een unieke naam, er zijn verschillende Stationsstraten in Nederland. Maar binnen een bepaalde stad zijn straatnamen natuurlijk wel uniek. Als je een route door Nederland beschrijft, dan zul je waarschijnlijk de straatnamen gebruiken van de stad waar je nu bent. Alleen als je naar een andere stad toe gaat, gebruik je de naam van die andere stad. In een netwerkperspectief zijn we dus eigenlijk op zoek naar de steden en dorpen in het netwerk: gebieden met veel straten.



Figuur 6 De methode met modulariteit vindt de kleine groepen in dit netwerk niet terug, terwijl de Infomap-methode dit wel doet.

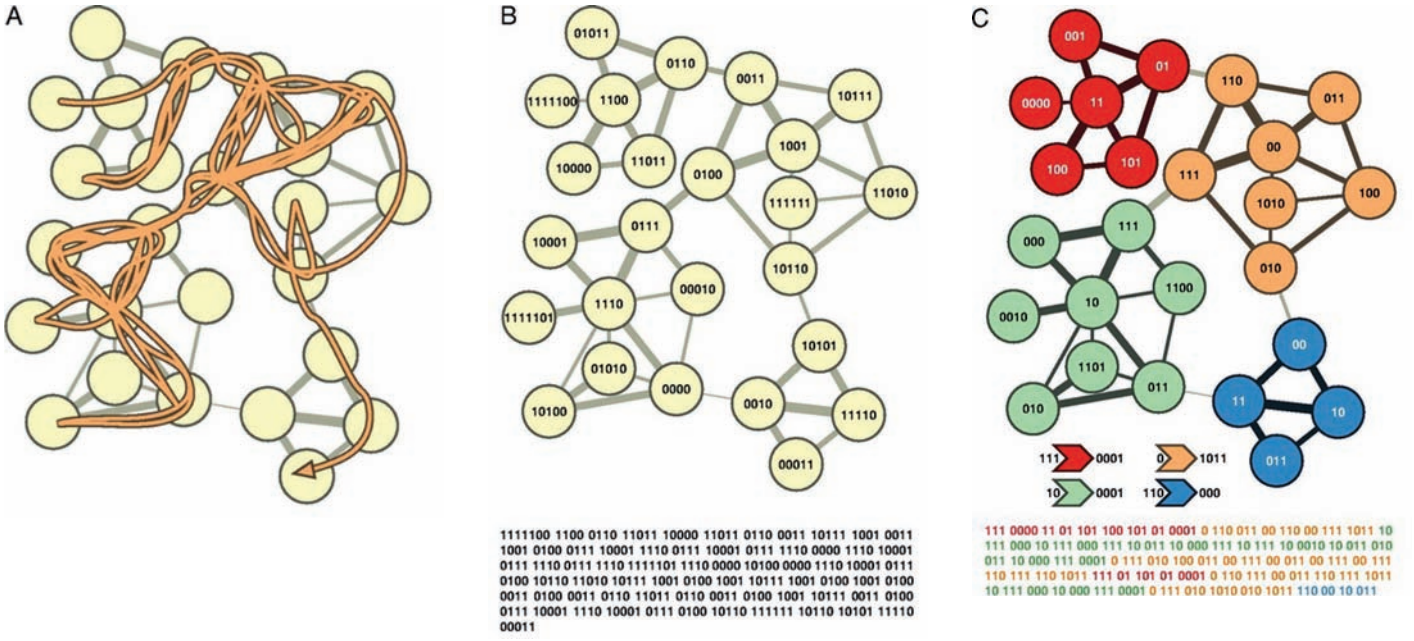
De beschrijving van een wandeling

Stel dat we binaire namen geven aan punten in een netwerk dat is opgedeeld in groepen. Namen binnen een groep moeten uniek zijn. Als een groep veel punten bevat, worden de binaire namen van punten in die groep dus erg lang. Alle groepen hebben ook een unieke binaire naam (dit zijn de 'steden'). We beschrijven de route van de random walker door steeds de naam van het punt te noemen waar hij naartoe gaat. Als de random walker van groep wisselt, moeten we daarnaast de naam van de nieuwe groep opschrijven. De beschrijving wordt dus erg lang als de random walker vaak van groep verandert, omdat we steeds zowel de naam van het punt waar de walker naartoe gaat, als de bijbehorende groepsnaam moeten opschrijven. De beschrijving wordt juist kort als we groepen hebben waar de random walker lang in blijft rondlopen. Als we de groepen te groot maken, wordt de beschrijving juist weer erg lang, omdat de unieke binaire namen van de punten dan erg lang worden. De vraag is dan: wat is de opdeling in groepen die de beschrijving van de random walk zo kort mogelijk maakt?

Voor een opdeling van het netwerk in groepen berekent de *Map equation* hoe veel binaire getallen gemiddeld nodig zijn om een stap van de random walker te beschrijven:

$$L = q_{\text{uit}} H(\text{uit}) + \sum_{c=1}^{K} q_{\text{in}}^{c} H_c(\text{in}).$$

Hier is q_{uit} de kans dat een willekeurige stap van de random walker een stap tussen twee verschillende groepen is, en q_{in}^{c} de kans dat een willekeurige stap van de



Figuur 7 (A) Een voorbeeld van een random walk over een netwerk; (B) De binaire beschrijving van de random walk zonder groepsstructuren mee te nemen; (C) De binaire beschrijving van de random walk opgedeeld in groepen is een stuk korter [5].

random walker een stap binnen groep c is. De term $H(\text{uit})$ beschrijft het gemiddelde aantal bits van de naam van een groep waar de walker naartoe beweegt, en $H_c(\text{in})$ het gemiddelde aantal bits van de naam van een punt in groep c waar de walker naartoe beweegt. Door de opdeling te vinden die deze vergelijking minimaliseert, vinden we dus de opdeling die de beschrijving van de wandeling het kortst maakt. Dit is dan de opdeling die het netwerk het meest efficiënt in 'steden' opdeelt.

Met dezelfde slimme algoritmes die een opdeling met hoge modulariteit vinden, kunnen we nu juist de opdelingen vinden die een lage waarde van de map equation hebben. Opdelingen die door Infomap gevonden worden kunnen wel echt anders zijn dan de opdelingen die gebaseerd zijn op modulariteit. In Figuur 6 zien we bijvoorbeeld dat de Infomap methode er wel in slaagt om de kleine groepsstructuren in het netwerk te ontdekken.

Overlappende groepen

Maar er zijn nog een hele hoop problemen die deze twee methoden om groepen te vinden niet oplossen. Op sociale netwerken horen de meeste mensen bijvoorbeeld thuis in verschillende groepen: familie, sportvrienden en collega's bijvoorbeeld. In dit soort netwerken is het dus helemaal niet mogelijk om iedereen aan één bepaalde groep toe te wijzen omdat de groepsstructuren in sociale netwerken elkaar vaak overlappen. Om bijvoorbeeld groepsstructuren te vinden die mensen op sociale media indelen op basis van hun interesses, heb je dus een methode nodig die overlappende groepsstructuren vindt. De twee methoden die hierboven beschreven staan kunnen deze overlappende groepen niet vinden. Kortom, het probleem van groepen vinden in netwerken is met deze twee methoden nog lang niet opgelost.

Referenties

1 V.D. Blondel, J.D. Guillaume, R. Lambiotte en E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.* (2008).

2 S. Fortunato en M. Barthélemy, Resolution limit in community detection, *PNAS* 204 (2007), 36–41.

3 M. Girvan en M.E.J. Newman, Finding and evaluating community structure in networks, *Phys. Rev. E* 69(2) (2004), 026113.

5 C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades en S.H. Strogatz, Redrawing the map of Great Britain from a network of human interactions, *PLOS ONE* 5(12) (2010), e14248.

6 M. Rosvall en C.T. Bergstrom, Maps of information flow reveal community structure in complex networks, *PNAS* 105 (2008), 1118.