Pim van der Hoorn
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
w.l.f.v.d.hoorn@tue.nl

Onderzoek

Wel of geen schaalvrije netwerken? – Een verhaal van extremen

Onze wereld wordt gedomineerd door een grote hoeveelheid complexe systemen. Veel van deze systemen kunnen beschreven worden als een netwerk (een graaf in puur wiskundige termen). Netwerken bestaan uit elementen (knopen) en verbindingen tussen deze elementen. Voorbeelden zijn overal te vinden. Zo zijn er sociale netwerken, waar de knopen vaak personen zijn en de verbindingen sociale relaties vertegenwoordigen. Maar denk ook aan het Internet dat ons in staat stelt om met computers aan de andere kant van de aarde te communiceren, of aan onze hersenen. Pim van der Hoorn gaat verder in op zogenaamde schaalvrije netwerken. Dat zijn netwerken waarvoor er geen typische schaal is voor het aantal verbindingen die de knopen van een netwerk hebben. Het aantal volgers op Twitter is hier een goed voorbeeld van.

Een belangrijke eerste-orde-eigenschap van een netwerk is het aantal verbindingen die de knopen hebben (de graden van de knopen). In de beginjaren van de netwerkwetenschap was dit dan ook de eerste eigenschap van een netwerk die bestudeerd werd. Dit leidde al snel tot een bijzondere ontdekking. Tot grote verbazing van verschillende onderzoekers bleek dat de verdeling van de graden in verschillende netwerken een wonderbaarlijk universeel gedrag vertoonde. Als men een plot maakte van de empirische verdeling van de graden op een log-log-schaal, dan leek deze erg op een rechte lijn met een negatieve coëfficiënt. Dit gedrag was nog duidelijker aanwezig als men naar de inverse kansverdelingsfunctie keek.

Laten we de fractie van het aantal knopen met graad k in een netwerk aanduiden met $P(k)$. Als deze functie een neerwaartse rechte lijn is op een log-log-schaal dan moet $P(k)$ een machtswet hebben;

$$\log(P(k)) = -\gamma \log(K) + C' \iff P(k) = Ck^{-\gamma}.$$

In de artikelen in de netwerkwetenschap werd dit aangeduid met de volgende formule:

$$P(k) \sim k^{-\gamma}. \tag{1}$$

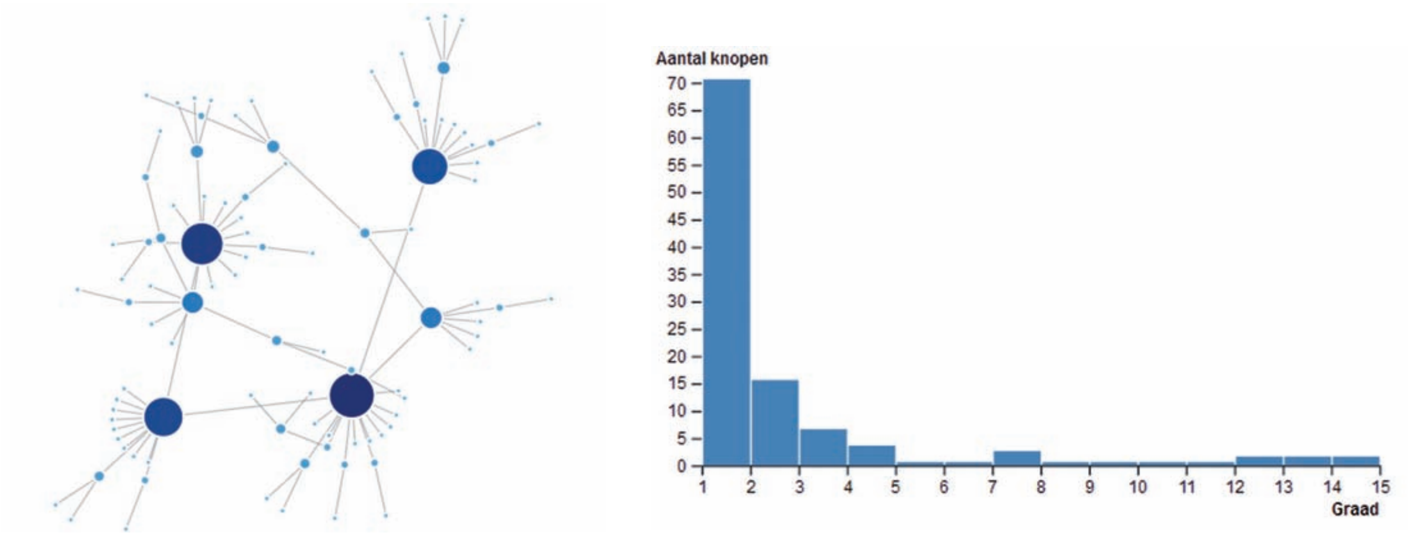
In dit geval zeggen wetenschappers dat het netwerk een machtswet heeft.

Het feit dat er zoveel netwerken met machtswetten werden gevonden kan inderdaad verrassend genoemd worden. Immers, deze netwerken beschrijven een ongekend breed spectrum aan complexe systemen, die niets met elkaar te maken hebben. Maar er was nog meer wonderlijks aan de hand.

Geen typische schaal voor de graden

Toen wetenschappers de exponent van de machtswet gingen schatten, bleek dat deze in het merendeel van de netwerken tussen de twee en drie lag. Dergelijke machtswetten werden als *schaalvrij* aangeduid en vaak werd dan ook het netwerk schaalvrij genoemd.

Een manier om de term schaalvrij te begrijpen is als volgt. Als je een kansverdeling hebt van de vorm $Ck^{-\gamma}$ met $2 < \gamma < 3$, dan heeft deze verdeling een eindig gemiddelde maar een oneindige variantie. Dit betekent dat er geen *typische* schaal is voor de graden. Er is een gemiddelde graad, maar vanwege de oneindige variantie komen we ook knopen tegen met graden die vele orden groter zijn.



Illustratie: D3.js, Herb Susmann

Figuur 1 Voorbeeld van een netwerk met 100 knopen uit het Barabási-Albert model.

Een mooi voorbeeld van dit schaalvrij gedrag is het netwerk van volgers op Twitter. Hier zijn de knopen gebruikers en er is een gerichte link van gebruiker a naar gebruiker b $(a \to b)$ als b gevolgd wordt door a. Het aantal volgers dat je hebt is dus gelijk aan de inkomende graad en dit wordt vaak gezien als een maat voor je invloed of populariteit. Een meting uit 2016 liet zien dat het gemiddeld aantal volgers in Twitter 717 is. Dit getal verbleekt echter bij het aantal volgers van de top 100 populairste gebruikers, die allemaal miljoenen volgers hebben. Voormalig president van de Verenigde Staten, Barack Obama en zangeres Katie Pery hebben zelfs een duizelingwekkend aantal van meer dan honderdmiljoen volgers.

Schaalvrije netwerken zijn overal

In de beginjaren van de netwerkwetenschap werd artikel op artikel gepubliceerd waarin de auteurs nieuwe netwerken vonden waarin de gradenverdeling een machtswet vertoonde. In veel van deze gevallen wisten de wetenschappers de exponent van de machtswet tussen de twee en drie te schatten. Het leek er dus op dat we omringt zijn door schaalvrije netwerken. Maar niemand begreep echter waar deze vandaan kwamen. Hoe kon het toch dat zoveel verschillende netwerken een schaalvrije gradenverdeling hadden?

In 1999 gaven de nu zeer bekende netwerkwetenschappers Réka Albert en László Barabási een antwoord op dit vraagstuk [1]. Zij introduceerde een dynamisch model voor netwerken, waar elke keer een nieuwe knoop aan het netwerk wordt toegevoegd die een vast aantal verbindingen met bestaande knopen maakt. De crux was dat de nieuwe knoop zijn buren koos met een kans die proportioneel was met de huidige graad van de knopen. Dit betekent dat als een knoop een hoge graad heeft, deze ook veel vaker nieuwe buren zal krijgen en dus een als maar hogere graad.

Albert en Barabási lieten zien dat dit model op een natuurlijke manier leidt tot een gradenverdeling die schaalvrij is. Een realisatie van een netwerk uit het Albert–Barabási-model is te zien in Figuur 1, met ernaast het histogram van de graden.

Vanaf dit moment ontstond er een groei in studies naar schaalvrije netwerken. In het bijzonder naar het gedrag van verschillende processen op deze netwerken, zoals de verspreiding van epidemieën. De algemene tendens onder wetenschappers was dat onze wereld wordt gedomineerd door schaalvrije netwerken en dus moesten we zoveel mogelijk over hun structuur en gedrag te weten komen. Een mooie conclusie van de eerste tien jaar onderzoek naar netwerken. Niets meer aan doen.

Wat is eigenlijk de definitie?

Er schuilt echter een klein addertje onder het groene gras van schaalvrije netwerken: er is niet echt een consensus onder wetenschappers over wat formule (1) nu precies betekent.

Op het eerste oog lijkt er niets aan de hand met (1). Immers, het teken $\sim$ wordt over het algemeen geassocieerd met asymptotisch equivalent en dus moet het wel zo zijn dat netwerkwetenschappers bedoelen dat

$$\lim_{k \to \infty} \frac{P(k)}{k^{-\gamma}} = 1.$$

Maar er is bijna geen enkel artikel in netwerkwetenschap te vinden waarin (1) ook daadwerkelijk zo geïnterpreteerd wordt. Vaker nog lijkt men te bedoelen dat $\frac{P(k)}{k^{-\gamma}} \to c > 0$ als $k \to \infty$. Maar in veel gevallen schrijft men slechts (1) zonder enige uitleg wat hiermee bedoeld wordt.

Waar men het wel over eens lijkt te zijn is de achterliggende intuïtie. Namelijk dat de staart van $P(k)$ op log-log-schaal lijkt op een neerwaartse rechte lijn. Oftewel, dat de staart van de verdeling op een machtswet lijkt.

Dit gebrek aan consensus over de interpretatie van een van de meest gebruikte formules in netwerkwetenschap is problematisch.

Als we willen vaststellen of een netwerk schaalvrij is, dan moeten we eerst controleren of de gradenverdeling aan (1) voldoet. Vervolgens moeten we de exponent γ schatten. Maar als we niet weten wat (1) betekent, hoe kunnen we dan vaststellen of de gradenverdeling hieraan voldoet, of de exponent schatten?

Schaalvrije netwerken zijn zeldzaam

Eind januari 2018 verschijnt een preprint online getiteld 'Scale-free networks are rare' (schaalvrije netwerken zijn zeldzaam). De auteurs van dit artikel hadden ook opgemerkt dat de term schaalvrij door de jaren heen van allerlei vormen had aangenomen en dat er niet echt een eenduidige definitie meer was. Daarnaast constateerde zij, terecht, dat de methodes die in het verleden gebruikt waren om de gradenverdeling van netwerken te analyseren statistisch gezien niet heel stevig in hun schoenen stonden. Ze vonden het dus tijd om de lang bestaande claim dat schaalvrije netwerken overal waren eens goed tegen het licht te houden.

Hun interpretatie van (1) was dat de empirische gradenverdeling van een netwerk, vanaf een gegeven graad $k_{\min}$, een exacte machtswet moest hebben:

$$P(k) = Ck^{-\gamma} \quad \text{voor alle } k \geq k_{\min}. \tag{2}$$

Dit is zeker een verdeling die voldoet aan de eis dat de staart op een machtswet lijkt. Het is er namelijk een.

De echte mooie eigenschap van deze verdeling is dat hij door twee parameters, γ en $k_{\min}$, wordt gedefinieerd. Als je vervolgens schatters hebt voor deze parameters kan je een hypothesetest voor deze aanname ontwikkelen. Dit is precies wat de auteurs hebben gedaan, plus nog wat andere toevoegingen die voor dit verhaal niet relevant zijn. Vervolgens pasten ze deze test toe op een groot aantal verschillende netwerken. Wat bleek, slechts een schamele vier procent van de bestudeerde netwerken slaagde voor hun schaalvrijtest. Hun conclusie was dan ook dat schaalvrije netwerken helemaal niet zo alomvertegenwoordigd zijn als werd aangenomen. Sterker nog, ze blijken erg zeldzaam te zijn. Hun werk werd begin 2019 in *Nature Communication* gepubliceerd [2].

Deze resultaten ontketenden een flinke discussie onder netwerkwetenschappers. Wetenschappers die heilig overtuigd waren van het bestaan van schaalvrije netwerken, waaronder Albert en Barabási, wezen erop dat definitie (2) te restrictief was en trokken de conclusie dus in twijfel. Een belangrijke onderbouwing van hun bezwaar is dat bijna geen enkel bekend model voor schaalvrije netwerken een gradenverdeling heeft die aan (2) voldoet.

Er waren echter ook veel wetenschappers die blij waren dat er eindelijk een duidelijke rigoureuze definitie en test waren, ook al was (2) wellicht iets te streng. Daarnaast hadden veel van hen al langer het vermoeden dat zeker niet het merendeel van netwerken schaalvrij was, wat nu ondersteund werd door de claim van de auteurs.

Maar wie heeft er gelijk? Hoe zit het echt met de gradenverdeling van netwerken? Is er misschien iets dat wiskunde kan doen om dit geschil te beslechten?

Machtswetten en reguliere variatie

Het probleem met machtswetten in netwerken schreeuwt gewoon om een wiskundige aanpak. Zeker omdat een grote kracht van de wiskunde het vermogen is om intuïties te abstraheren, om te zetten in rigoureuze concepten en deze vervolgens te gebruiken om diepgaande vraagstukken op te lossen. En de situatie van schaalvrije netwerken is gelukkig geen uitzondering.

Laten we beginnen met het idee van *een rechte lijn op een log-log-schaal*. Dit brengt ons bij de theorie van regulier variërende (*regularly-varying*) functies. Een functie $f: \mathbb{R} \to \mathbb{R}$ is regulier variërend met exponent α, $f \in \mathcal{R}_{-\alpha}$, als het van de volgende vorm is:

$$f(x) = \ell(x)x^{-\alpha} \quad \text{met } \alpha > 0,$$

en waarbij $\ell(x)$ een langzaam variërende functie is,

$$\forall t > 0 \quad \lim_{x \to \infty} \frac{\ell(tx)}{\ell(x)} = 1.$$

Een kansverdeling heet regulier variërend als $\bar{F} \in \mathcal{R}_{-\alpha}$, met $\bar{F}(x) = 1 - F(x)$ de inverse kansverdelingsfunctie. Als we aannemen dat de kansfunctie voor graden van de vorm $P(k) = Ck^{-\gamma}$ is, dan geldt dat $\bar{F}(k) = C'k^{-(\gamma-1)}$. Dus de exponent α en γ zijn aan elkaar gerelateerd door de formule $\gamma = \alpha + 1$.

De klasse $\mathcal{R}_{-\alpha}$ is ongelofelijk divers. Zij bevat de pure machtswetten en die van de vorm (2). Maar ook functies zoals $\log(x)x^{-\alpha}$ zijn toegestaan. Het belangrijkste is dat een langzaam variërende functie altijd langzamer groeit dan elke macht van x. Dit betekent dat als we een functie $f \in \mathcal{R}_{-\alpha}$ op een log-log-schaal bekijken, dan zal de staart inderdaad bijna op een neerwaartse rechte lijn met coëfficiënt $-\alpha$ lijken. De regulier variërende kansverdelingen bieden dus de perfecte abstractie van de intuïtie achter het concept van machtswetten in netwerken, en in het bijzonder van schaalvrije netwerken.

Er is echter een probleem met deze mooie abstractie. De klasse $\mathcal{R}_{-\alpha}$ heeft geen parametrische beschrijving. Sterker nog, het heeft een niet-aftelbare oneindige dimensie, in de vorm van de langzaam variërende functies. Een mooie hypothesetest voor lidmaatschap van $\mathcal{R}_{-\alpha}$ lijkt dus uitgesloten, laat staan voor het schatten van de zo belangrijke exponent α.

De theorie van extreme waarden

Maar ook hier biedt de wiskunde een uitkomst. Deze keer komt die vanuit extreme-waardentheorie. Een vakgebied van de wiskunde waar de extreme waarden van kansverdelingen worden bestudeerd. Een klassiek vraagstuk is het volgende: laat $(X_n)_{n \geq 1}$ een rijtje onafhankelijke stochastische variabelen zijn met dezelfde kansverdeling F en definieer de lokale maxima $M_n = \max\{X_1, \ldots, X_n\}$. Wat is dan de asymptotische verdeling van M_n?

Het antwoord op deze vraag komt in de vorm van de fundamentele Fisher–Tippett-stelling, die de mogelijke distributielimieten van gecentreerde en genormaliseerde maxima beschrijft.

Fisher–Tippett-stelling. *Stel dat er rijen d_n en $c_n > 0$ bestaan zodat*

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H \quad \textit{als } n \to \infty,$$

waar de stochastische variabel H niet met kans 1 een constante is. Dan behoort de kansverdeling van H tot een van de drie extreme-waardenkansverdelingen: 1. Frechèt, 2. Gumbel, of 3. Weibull.

De stelling draagt de naam van de auteurs die als eerste vaststelde wat de mogelijk verschillende limietverdelingen konden zijn. Bewijzen van de stelling zijn tot stand gekomen door werk van Gnedenko [4], de Haan [5] en Weissman [12]. Een volledig bewijs kan tegenwoordig gevonden worden in elk standaardwerk zoals bijvoorbeeld [10].

Maar het verhaal wordt nog mooier. Deze drie klassen van kansverdelingen kunnen namelijk samengevoegd worden in de zogeheten *algemene extreme-waardekansverdeling*

$$H_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right) & \text{als } \xi \neq 0, \\ \exp(-\exp(-x)) & \text{als } \xi = 0. \end{cases}$$

De Frechèt-distributies corresponderen met $\xi > 0$, Weibul met $\xi < 0$ en de Gumbel-verdeling is wanneer $\xi = 0$.

Het cruciale van deze algemene beschrijving van de distributielimieten van de maxima is de parameter ξ, die de *extreme-waarden-index* wordt genoemd. Het is deze parameter waarmee we de exponent α kunnen schatten, zonder ons druk te maken over the exacte vorm van de langzaam variërende functie.

Dit wordt mogelijk gemaakt door een ander fundamenteel resultaat uit de extreme-waardentheorie. De klasse van kansverdelingen waarvan de verdeling van de maxima $\frac{M_n - d_n}{c_n}$ naar de Frechèt-kansverdeling met $\xi > 0$ convergeert, zijn precies de regulier variërende kansverdelingen met $\alpha = \frac{1}{\xi}$ (voor een bewijs zie bijvoorbeeld [3]). De moraal is dat als we in staat zijn om de extreme-waarden-index ξ te schatten, we ook kunnen concluderen of de verdeling tot de klasse $\mathbb{R}_{-1/\xi}$ behoort en dus of zij een machtswet beschrijft, zoals netwerkwetenschappers daarover denken. In het bijzonder correspondeert een waarde van $\frac{1}{2} < \xi < 1$ met een exponent $1 \leq \alpha < 2$ $(2 < \gamma < 3)$ en dus met een verdeling met eindig gemiddelde en oneindige variantie. Een schaalvrije kansverdeling dus.

Nu krijgen parameters in wiskunde zelden een eigen naam als ze niet van fundamenteel belang zijn. Het zal dan ook niet als een verrassing komen dat een groot deel van het onderzoek in extreme-waardentheorie zich bezighoudt met het efficiënt en correct schatten van ξ, op basis van een eindig aantal observaties. Jaarlijks worden meerdere artikelen gepubliceerd waarin een nieuwe of verbeterde versie van een bestaande schatter voor ξ wordt geïntroduceerd. Er is dus zeker geen gebrek aan robuuste hamers om op de spijker "Wanneer heeft mijn netwerk een machtswet?" te slaan.

Waar we nu staan

We hebben zojuist een mooie rondrit gemaakt. Vanuit het probleem van het analyseren van machtswetten in netwerken en het bestaan van schaalvrije netwerken, via regulier-variërende functies door naar de extreme-waardentheorie en weer terug naar machtswetten in netwerken. De conclusie is dat de meest brede en rigoureuze definitie voor machtswetten in netwerken regulier-variërende functies zijn en dat we hun exponent $\alpha = \gamma - 1$ kunnen schatten via de extreme-waarden-index ξ.

Het zal niet onverwacht zijn dat deze link met de extreme-waardentheorie geheel onbekend was bij netwerkwetenschappers. Voor veel hen lijkt dit dan ook waarschijnlijk meer op een ondoorgrondelijke goocheltruc. Maar dit is precies de kracht van het abstracte denken in de wiskunde. Alleen door een probleem tot zijn kern te reduceren en in een algemene vorm te beschrijven, is het mogelijk om bruggen naar andere gebieden te slaan en zo de kennis uit die gebieden te gebruiken om het originele probleem op te lossen.

In een recent artikel van mijzelf en drie co-auteurs hebben wij de extreme-waarden-index ingezet om gradenverdelingen in netwerken te bestuderen [11]. Onze resultaten laten zien dat het aannemelijk is dat een redelijk aantal netwerken een machtswet heeft en dat een flink deel hiervan schaalvrij kan zijn. Maar het is zeker niet zo dat dit voor bijna alle netwerken opgaat. De conclusie is dus dat de waarheid omtrent het bestaan van schaalvrije netwerken, hoe kon het ook anders, ergens in het midden ligt.

Aan de ene kant verschaft de link met de extreme-waardentheorie netwerkwetenschappers een grote gereedschapskist om de gradenverdelingen van netwerken te analyseren. Maar er is ook goed nieuws voor de wiskundigen. Want, hoewel veel theorie beschikbaar is, is niet alles direct toepasbaar op netwerken. Allereerst zijn de hier genoemden resultaten voor continue kansverdelingen terwijl graden in netwerken discreet zijn. Daarnaast wordt voor consistentie van de schatters voor ξ aangenomen dat de data onafhankelijk is, wat niet waar is als deze data uit de graden in een netwerk bestaat. Desondanks lijken verschillende simulaties en experimenten aan te geven dat de methodes ook werken in deze, minder schone, situatie. Er is dus een echte noodzaak om de huidige theorie uit te breiden en zijn toepasbaarheid op gradenverdelingen van netwerken met solide bewijzen te ondersteunen. Dit biedt kansen voor het optuigen van nieuwe en mooie wiskunde. Iets wat wiskundigen maar al te graag doen.

Om verder te lezen over dit onderwerp in een bredere context verwijzen we de geïnteresseerde lezer naar een reeks artikelen op de Network Pages (www.networkpages.nl, [6–9]), het online outreach platform van het onderzoeksprogramma NETWORKS.

Referenties

1 Albert-László Barabási en Réka Albert, Emergence of scaling in random networks, *Science* 286(5439) (1999), 509–512.

2 Anna D. Broido en Aaron Clauset, Scale-free networks are rare, *Nature Communications* 10(1) (2019), 1017.

3 Paul Embrechts, Claudia Klüppelberg en Thomas Mikosch, *Modelling Extremal Events: for Insurance and Finance,* Vol 33, Springer, 2013.

4 Boris Gnedenko, Sur la distribution limite du terme maximum d'une serie aleatoire, *Ann. Math.* 44(3) (1943), 423–453.

5 L. de Haan, *On Regular Variation and its Application to the Weak Convergence of Sample Extremes*, Mathematisch Centrum, 1970.

6 Remco van der Hofstad, How the popular become even more popular, www.networkpages.nl

7 Remco van der Hofstad en Julia Komjathy, Degrees in graphs IV: degrees in large real-world networks, www.networkpages.nl.

8 Pim van der Hoorn, Scale-free networks, a controversial topic solved by extreme mathematics, www.networkpages.nl.

9 Julia Komjathy, Why the whole world has seen Gangnam style, www.networkpages.nl.

10 Sidney I. Resnick, *Extreme Values, Regular Variation and Point Processes*, Springer, 2013.

11 Ivan Voitalov, Pim van der Hoorn, Remco van der Hofstad en Dmitri Krioukov, Scale-free networks well done, *Physical Review Research* 1(3) (2019), 033034.

12 Ishay Weissman, On location and scale functions of a class of limiting processes with application to extreme value theory, *The Annals of Probability* 3(1) (1975), 178–181.