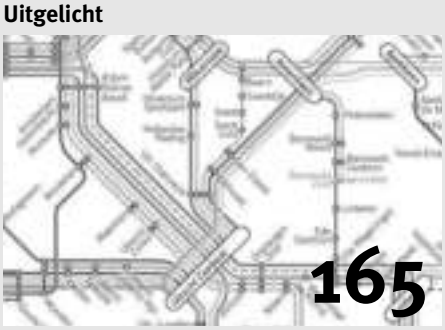
Inhoud | Contents

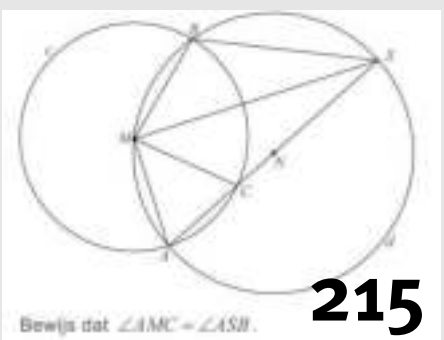
Artikelen | Features

Rubrieken | Departments

Uitgelicht



Een van de netwerken die in dit themanummer worden besproken is het spoornetwerk, in een artikel van Leo Kroon en Lex Schrijver.



Het examen vwo wiskunde B 2015 wordt besproken door Wim Caspers.

Colofon

Het Nieuw Archief voor Wiskunde is een uitgave van het Koninklijk Wiskundig Genootschap en verschijnt vier maal per jaar. Het tijdschrift richt zich op een ieder die zich beroepsmatig met wiskunde bezighoudt, als academisch of industrieel onderzoeker, student, leraar, journalist of beleidsmaker. Het stelt zich ten doel te berichten over ontwikkelingen in de wiskunde in het algemeen en in de Nederlandse wiskunde in het bijzonder.

ISSN 0028-9825

Adres
Nieuw Archief voor Wiskunde
Centrum Wiskunde & Informatica
Postbus 94079
1090 GB Amsterdam
tel. 020-5924288
redactie@nieuwarchief.nl
www.nieuwarchief.nl

Abonnementen/Subscriptions
Ledenadministratie KWG
Postbus 1064, 7940 KB Meppel
admin@wiskgenoot.nl, tel. 085-0160252
Abonnementsperiode/Subscription period
Een abonnement gaat in op de dag dat uw aanmelding ontvangen wordt. De opzegtermijn is twee maanden vóór het verstrijken van de abonnementsperiode.
The subscription starts on the day that your application is received. The term of notice is two months before the end of the current subscription period.
Exchange subscriptions
Koninklijk Wiskundig Genootschap Library
Centrum voor Wiskunde en Informatica
Postbus 94079, NL-1090 GB Amsterdam
KWG-exchange@cwi.nl, fax +31 20 5924026

Vormgeving
Kitty Molenaar (ontwerp)
Susanne Laws (omslag, begeleiding binnenwerk)
Programmatuur (CONTEXT)
Hans Hagen, PRAGMA-ADE, Hasselt (Overijssel)
Druk
Drukkerij Ten Brink, Meppel

Koninklijk Wiskundig Genootschap

Het Koninklijk Wiskundig Genootschap (KWG) is een landelijke vereniging van beoefenaren van de wiskunde. In 1778 opgericht onder de zinspreuk: 'Een onvermoeide arbeid komt alles te boven' is het 's werelds oudste nationale wiskundegenootschap. Leden ontvangen het Nieuw Archief voor Wiskunde als onderdeel van hun lidmaatschap.
De contributie voor gewone leden bedraagt in 2015 € 90 per jaar. Voor gepensioneerden en werklozen € 45. Voor studenten aio's en oio's van een Nederlandse universiteit of hbo-opleiding € 35. Studenten en pas afgestudeerden kunnen een jaar lang gratis lid worden. Voor leden van de wiskundige vakverenigingen VvS+OR, NVvW en NVORWO geldt het gereduceerd tarief van € 70. Leden die het Nieuw Archief voor Wiskunde niet willen ontvangen, betalen slechts € 50 contributie. Losse nummers zijn te bestellen voor € 20.
Members of *SIAM*, the *Société Mathématique de France*, the *Gesellschaft für Angewandte Mathematik und Mechanik*, the *Deutsche Mathematiker-Vereinigung*, and of the *American, Australian, Belgian, Indian, London, Spanish,* and *South-African Mathematical Societies* living outside of the Netherlands pay € 70 instead of the full membership fee of € 90 a year. Conversely, these foreign societies, as well as the VvS+OR and the NVORWO, have reduced membership fees for KWG-members. A postage charge of € 10 is added for members living abroad but in Europe, and a charge of € 12,50 for members living outside of Europe.
Instituten in Nederland en buitenland kunnen zich abonneren op het Nieuw Archief voor Wiskunde voor € 100 (Nederland), € 110 (Europa) of € 112,50 (buiten Europa).
Website: www.wiskgenoot.nl
Ledenadministratie
Ledenadministratie KWG
Postbus 1064, 7940 KB Meppel
admin@wiskgenoot.nl, tel. 085-0160252
Bestuur (wiskgenoot@wiskgenoot.nl)
Geurt Jongbloed (TUD, voorzitter), Fetsje Bijma (VU, penningmeester), Joke Blom (CWI, secretaris), Erik van den Ban (UU), Theo van den Bogaart (HU), Sonja Cox (UvA), André Ran (VU), Herman te Riele (CWI), Mark Veraar (TUD)

Informatie voor auteurs

Kopij
Het Nieuw Archief voor Wiskunde is een gerediger tijdschrift. Kopij dient elektronisch te worden aangeboden aan de hoofdredactie. Uitgebreide informatie en auteursinstructies kunt u vinden op www.nieuwarchief.nl .
Reprints
Auteurs ontvangen een elektronische reprint in pdf-formaat.
Volgende nummers

nummer	*verschijningsdatum*	*deadline kopij*
4	07–12–2015	verstreken
1	07–03–2016	01–11–2015
2	06–06–2016	01–02–2016
3	05–09–2016	01–05–2016

Instituutsleden

De publicaties van het Koninklijk Wiskundig Genootschap worden mede mogelijk gemaakt door de bijdragen van de volgende instituten.

Centrum Wiskunde & Informatica
Rijksuniversiteit Groningen
Universiteit van Amsterdam
Universiteit Leiden
Vrije Universiteit
Universiteit Utrecht
Technische Universiteit Eindhoven
Eurandom
Technische Universiteit Delft
Universiteit Twente
Radboud Universiteit Nijmegen
Freudenthal Instituut

Op het omslag

Dit nummer is grotendeels gewijd aan het thema 'netwerken'. De wereld om ons heen zit vol met netwerken van allerlei soort. Er wordt veel wetenschappelijk onderzoek naar gedaan, ook door wiskundigen. Zie het redactioneel hiernaast en de artikelen op pagina's 161–209.

Netwerken

Dit themanummer van het *Nieuw Archief voor Wiskunde* staat in het teken van *netwerken* en wat wiskundigen daar zoal over te zeggen hebben. Informeel gesproken kan men zeggen dat een netwerk een geheel van met elkaar verbonden punten beschrijft. Het is duidelijk dat met zo'n brede 'definitie' een veelheid van fenomenen uit de wereld om ons heen onder de noemer van netwerken geschaard kan worden. Denk bijvoorbeeld aan netwerken in de natuur, zoals allerlei chemische structuren, neurale verbindingen in het brein of planeten in een sterrenstelsel. Maar ook in de context van de mens zijn er voorbeelden te over: auto's in verkeersnetwerken, vrienden in sociale netwerken, de overdracht van virussen binnen populaties, mobiele apparaten in draadloze netwerken. De lijst van relevante netwerken lijkt eindeloos en het is dan ook niet vreemd dat ze het object zijn van intensief wetenschappelijk onderzoek. Ook in de wiskunde zien we een veelheid aan netwerken, in allerlei soorten en maten: ze kunnen deterministisch zijn of stochastisch, statisch of dynamisch, klein of groot, ruimtelijk of virtueel.

Het Zwaartekrachtproject NETWORKS, dat vorig jaar gestart is, heeft als doel netwerken te beschrijven en te analyseren, om te komen tot goede, werkbare algoritmes die de *performance* van netwerken kunnen verbeteren of zelfs optimaliseren. In het eerste artikel van dit nummer geven we een meer gedetailleerd beeld van het project, waarin veel inspiratie wordt geput uit de 'echte' netwerken zoals we die elke dag tegenkomen. De maatschappelijke uitdagingen waarmee zulke netwerken ons confronteren gaan vaak hand in hand met prachtige wiskunde en in deze editie zien we daar een aantal voorbeelden van.

Zo staat in ons land het onderzoek naar optimale dienstroosters voor treinen op een ongekend hoog niveau, getuige ook de internationale erkenning voor de ontwikkelde expertise. Leo Kroon en Lex Schrijver geven een beeld van de, geavanceerde maar direct toepasbare, wiskundige technieken die hierin een rol spelen. Leo van Iersel legt uit hoe we, gebruikmakend van concepten uit de grafentheorie, kunnen reconstrueren hoe organismen zijn ontstaan uit verre voorouders als gevolg van allerlei evolutionaire processen. Mark de Berg beschrijft hoe problemen uit de frequentietoekenning in draadloze communicatienetwerken aangepakt kunnen worden met combinatorische technieken. In een bijdrage die gecoördineerd werd door Karen Aardal komen toepassingen uit de gezondheidszorg aan bod, waarbij de focus ligt op het proactief plannen van ambulancediensten. Vergelijkbare technieken blijken ook gebruikt te kunnen worden bij het plaatsen van brandweercentrales.

Stochastische netwerken, waarin deeltjes zich volgens een bepaald probabilistisch mechanisme bewegen, zijn al meer dan honderd jaar een belangrijk onderzoeksthema. Onno Boxma, Stella Kapodistria en Michel Mandjes beschrijven een aantal klassen van zulke stochastische netwerken waarvoor nette gesloten oplossingen bestaan voor allerlei prestatiematen, en laten tevens zien dat de analyse uitermate gecompliceerd kan worden zodra men die klasse van standaardnetwerken verlaat. Deeltjesmodellen staan ook in de laatste twee bijdragen centraal. Sem Borst, Johan van Leeuwaarden en Peter van de Ven bespreken modellen voor draadloze communicatie. Draadloze netwerken zijn groot en druk, waardoor signalen verstoord kunnen raken. De uitdaging is om zonder centrale coördinatie de capaciteit te verdelen naar tevredenheid van de netwerkgebruikers en tevens de interferentie binnen de perken te houden. De laatste bijdrage betreft deeltjesmodellen vanuit een natuurkundig perspectief. Zoals wordt uiteengezet door Diego Garlaschelli, Frank den Hollander en Andrea Roccaverde, zijn er tal van fysica-geïnspireerde complexe netwerken, waarvan de onderliggende structuur de nodige wiskundige uitdagingen biedt.

Johan van Leeuwaarden en Michel Mandjes, gastredacteuren
Faculteit Wiskunde en Informatica, TU/e, resp. Korteweg-de Vries Instituut voor Wiskunde, UvA

Servicepagina

Service page

Koninklijk Wiskundig Genootschap

❑ **Secretariaat**
Joke Blom, CWI
Postbus 94079, 1090 GB Amsterdam
wiskgenoot@wiskgenoot.nl
www.wiskgenoot.nl

❑ **Wintersymposium KWG 2016**
Op zaterdag 9 januari 2016 vindt in het Academiegebouw te Utrecht het wintersymposium van het KWG plaats. Het symposium richt zich op wiskundedocenten in het voortgezet onderwijs en andere belangstellenden. Dit jaar is het thema 'Analytische meetkunde'. Dit is onderdeel van de curriculumvernieuwing wiskunde op havo en vwo. Doel van het symposium is contacten tussen wiskundedocenten en beroepsmatige beoefenaars van wiskunde te stimuleren. Sprekers zijn dit jaar onder anderen: Jeroen Spandaw (TU Delft), Frits Beukers (UU) en Jaap Top (RUG). Nadere informatie over het programma en de aanmelding wordt in de loop van het najaar bekend gemaakt.

❑ **BeNeLuxMC 2016 in Amsterdam**
Op 22–23 maart 2016 wordt onder auspiciën van het KWG, en in samenwerking met het Belgische en het Luxemburgse Wiskundig Genootschap, in Amsterdam het BeNeLux Mathematisch Congres georganiseerd. Hierin is bevat het 52ste Nederlands Mathematisch Congres. Een van de hoofdsprekers tijdens het congres is James Maynard die zal spreken over de laatste spectaculaire ontwikkelingen met betrekking tot het priemparenprobleem, waaraan hij een belangrijke bijdrage heeft geleverd. Aan het einde van de eerste dag worden in het Tropenmuseum twee publiekslezingen gegeven, door Marcus du Sautoy en Ionica Smeets. Daarna kan men tegen gereduceerde prijs deelnemen aan een dinerbuffet. Noteert u BeNeLuxMC 2016 alvast in uw agenda?

❑ **Oproep voor nominaties N.G. de Bruijnprijs**
Tot uiterlijk 1 oktober 2015 kunnen kandidaten nog worden voorgedragen, zie de oproep op wiskgenoot.nl. Voordrachten richten aan prof.dr. H.W. Broer, h.w.broer@rug.nl.

Platform Wiskunde Nederland

❑ **Bureau PWN**
Science Park 123, L013, 1098 XG Amsterdam
bureau@platformwiskunde.nl
pr@platformwiskunde.nl (publiciteit)
Tel: 020-5924006/06-51892525
www.platformwiskunde.nl

❑ **PWN Nieuwsbrief**
PWN publiceert eens in de drie maanden een PWN Nieuwsbrief, die op de website te vinden is. Actuele PWN-informatie wordt verder zowel via de website als via Twitter gegeven.

European Mathematical Society

❑ **EMS website**
www.euro-math-soc.eu

❑ **Oproep voor nominaties EMS-prijzen**
Eenieder die op 30 juni 2016 nog geen 35 jaar is en de prijs niet eerder ontvangen heeft, kan genomineerd worden. Sluitingsdatum is 1 november 2015. www.euro-math-soc.eu/ems-prizes

❑ **EMS Newsletter**
De elektronische versie van de meest recente uitgave van de EMS Newsletter is beschikbaar via www.ems-ph.org/publications.php.

International Mathematical Union

❑ **IMU website**
www.mathunion.org

❑ **IMU Newsletter**
De twee-maandelijkse IMU Newsletter is beschikbaar via www.mathunion.org/imu-net.

Recente publicaties van het KWG

❑ **Epsilon Uitgaven** (www.epsilon-uitgaven.nl)
81. Functionaalanalyse, Heinz Hanßmann, €28, 2015.
80. Christiaan Huygens: Het Slingeruurwerk, Jan Aarts, €32, 2015.
79. Galoistheorie, Frans Keune, €24, 2015.

❑ **Zebra-reeks** (Epsilon Uitgaven)
45. Inversie, Jacques Jansen , €10, 2015.
44. Labyrinten en doolhoven, Wim Kleijne, €10, 2015.
43. De Strepen van de Zebra, Geertje Hek, €10, 2015.
42. Sangaku's, Hans van Lint en Jeanne Breeman, €10, 2015.
41. Met passer, liniaal en neusislat, Ad Meskens en Paul Tytgat, €10, 2015.

Manuscripten voor Epsilon Uitgaven kunt u mailen naar redactie@epsilon-uitgaven.nl.

Agenda

Upcoming Events

september 2015

14–18 september 2015
❑ **Verification of Concurrent and Distributed Software**
Workshop georganiseerd door D. Gurov (Stockholm), M. Huisman (UT), J.J. Hunt (Karlsruhe) en A. Poetzsch-Heffter (Kaiserslautern).
plaats *Lorentz Center@Oort, Leiden*
info *lorentzcenter.nl*

18 september 2015
❑ **Finale Nederlandse Wiskunde Olympiade**
37 onderbouwleerlingen, 39 vierdeklassers en 47 vijfdeklassers zijn door naar de finale van de Nederlandse Wiskunde Olympiade 2015.
plaats *TU Eindhoven*
info *wiskundeolympiade.nl*

20 september 2015
❑ **Hoe snel zijn computers in de toekomst?**
Wakker Worden Kinderlezing waarin informaticus Harry Buhrman vertelt over nieuwe technologieën en kwantumcomputers.
plaats *NEMO, Amsterdam*
info *e-nemo.nl*

25 september 2015
❑ **Wiskundetoernooi 2015**
Leerlingen van scholen uit het hele land buigen zich in teams van 4 of 5 over verschillende wiskundeopgaven. Het thema is dit jaar 'Winkunde: geluk of strategie?'.
plaats *Radboud Universiteit Nijmegen*
info *ru.nl/wiskundetoernooi*

25 september 2015
❑ **Discovery Festival 2015**
Een avondje uit met stof tot nadenken: wetenschappelijke experimenten, muziek, kunst en inspirerende nieuwe ideeën.
plaats *Amsterdam*
info *discoveryfestival.nl*

oktober 2015

3 oktober 2015
❑ **Junior Wiskunde Olympiade**
Wiskundewedstrijd voor scholieren uit de onderbouw van havo en vwo.
plaats *Vrije Universiteit, Amsterdam*
info *few.vu.nl*

3 oktober 2015
❑ **Amsterdam Science Park Open Dag**
Jaarlijkse open dag van Amsterdam Science Park is dé gelegenheid om een kijkje te nemen achter de schermen van wetenschappelijk onderzoek.
plaats *Amsterdam Science Park*
info *amsterdamsciencepark.nl*

3 oktober 2015
❑ **European Ig Nobel Show 2015**
Terugblik op de tien onderzoeken waarvoor op 17 september aan de Havard Universiteit de Ig Nobelprijzen zijn uitgereikt.
plaats *NEMO, Amsterdam*
info *ignobelnight.nl*

5 oktober 2015
❑ **Dag van de Leraar**
Internationale dag van de leraar, waarop de Leraren van het Jaar worden gekozen in de categorieën basisonderwijs, speciaal onderwijs, voortgezet onderwijs en middelbaar beroepsonderwijs.
plaats *overal*
info *dagvandeleraar.nl*

5–9 oktober 2015
❑ **Shape Constrained Inference: Open Problems and New Directions**
Workshop georganiseerd door Dragi Anevski (Lund), Geurt Jongbloed (Delft) en Wolfgang Polonik (Davis, CA).
plaats *Lorentz Center@Snellius, Leiden*
info *lorentzcenter.nl*

7 oktober 2015
❑ **Mini-Workshop on Symplectic Geometry**
Workshop met sprekers Leonid Polterovich (Tel Aviv), Alexandru Oancea (Parijs), Barney Bramham (Bochum) en Sheila Sandon (Strasbourg).
plaats *VU, Amsterdam*
info *www.gqt.nl*

7–9 oktober 2015
❑ **Veertigste Woudschoten Conferentie**
Jaarlijkse conferentie georganiseerd door de Nederlands/Vlaamse Werkgemeenschap Scientific Computing (WSC).
plaats *Conferentiecentrum Woudschoten, Zeist*
info *wsc.project.cwi.nl/conferentieN.php*

Suggesties voor deze agenda zijn welkom.
agenda@nieuwarchief.nl

8 oktober 2015
❑ **Van Dantzig Seminar**
Landelijke serie seminars in de statistiek. Deze keer met spreker Martin Wainwright.
plaats *VU, Amsterdam*
info *few.vu.nl/~bkk320/vandantzig*

22 oktober 2015
❑ **Challenges for the next 25 years**
Symposium ter gelegenheid van het 25-jarig bestaan van de European Mathematical Society.
plaats *Institut Henri Poincaré, Parijs*
info *euro-math-soc.eu*

23–24 oktober 2015
❑ **Krommen en constructies**
Masterclass voor leerlingen van 5/6 vwo.
plaats *Utrecht*
info *www.uu.nl/u-talent*

26–30 oktober 2015
❑ **GQT Graduate School and Colloquium**
Bijeenkomst van het GQT-cluster georganiseerd door Raf Bocklandt (Newcastle), Gil Cavalcanti (UU) en Maarten Solleveld (RU).
plaats *Conferentiecentrum Woudschoten, Zeist*
info *www.gqt.nl*

28–30 oktober 2015
❑ **Workshop on Automatic Sequences, Number Theory, and Aperiodic Order**
Workshop met ander anderen Boris Adamczewski (Marseille), Michael Baake (Bielefeld), Henk Bruin (Wenen) en Hendrik Lenstra (UL).
plaats *TU Delft*
info *ewi.tudelft.nl*

november 2015

4 november 2015
❑ **ICAB Conferentie 2015**
Gratis toegankelijke conferentie voor iedereen die werkt in het academisch bètaonderwijs, met als thema 'Opleiden... Breed of Smal?'.
plaats *Het Kasteel, Groningen*
info *icab.nl*

7 november 2015
❑ **Jaarvergadering NVvW**
Jaarlijkse vergadering en studiedag van de Nederlandse Vereniging van Wiskundeleraren.
plaats *Ichthus College, Veenendaal*
info *nvvw.nl*

8 november 2015
❑ **Waarom zijn er geen vierkante slakken?**
Wakker Worden Kinderlezing waarin wiskundige Han Peters vertelt over wiskundige patronen en vormen in de natuur.
plaats *NEMO, Amsterdam*
info *e-nemo.nl*

9–11 november 2015
❑ **Stochastics Meeting Lunteren 2015**
44ste jaarlijkse bijeenkomst voor stochastici.
plaats *Congrescentrum De Werelt, Lunteren*
info *www.eurandom.nl/STAR*

9–13 november 2015
❑ **Moduli Spaces and Arithmetic Geometry**
Workshop georganisserd door Chai Ching-Li (Pennsylvania), Ben Moonen (RU) en Jaap Top (RUG).
plaats *Lorentz Center@Oort, Leiden*
info *lorentzcenter.nl*

13 november 2015
❑ **Voorronde A-lympiade**
Voorronde van de jaarlijkse wedstrijd waarin wiskunde A centraal staat, voor leerlingen van 5 havo en 5/6 vwo.
plaats *eigen school*
info *www.uu.nl/onderwijs/wiskunde-a-lympiade*

13 november 2015
❑ **Wiskunde B-dag**
Leerlingen van 5 havo en 5/6 vwo met wiskunde B in hun profiel werken in teams van 3 of 4 aan een open wiskundig probleem.
plaats *deelnemende scholen*
info *www.uu.nl/onderwijs/wiskunde-b-dag*

26–27 november 2015
❑ **Diamant symposium**
Tweejaarlijks symposium van het DIAMANT-cluster.
plaats *Congrescentrum De Werelt, Lunteren*
info *websites.math.leidenuniv.nl/diamant*

december 2015

7–11 december 2015
❑ **Quantum Random Walks and Quantum Algorithms**
Workshop georganiseerd door Harry Buhrman (CWI) en Frank den Hollander (LU).
plaats *Lorentz Center@Snellius, Leiden*
info *lorentzcenter.nl*

januari/februari 2016

9 januari 2016
❑ **Wintersymposium KWG 2016**
Jaarlijks symposium van het KWG met dit jaar als thema 'Analytische meetkunde', hetgeen onderdeel is van de curriculumvernieuwing wiskunde op havo en vwo.
plaats *Academiegebouw Utrecht*
info *wiskgenoot.nl*

25–29 januari 2016
❑ **Studiegroep Wiskunde met de Industrie**
Jaarlijkse studiegroep waarin wiskunde en industrie samen problemen oplossen.
plaats *IMAPP, Nijmegen*
info *www.ru.nl/imapp*

25–27 januari 2016
❑ **15th Winter School Mathematical Finance**
Winterschool over financiële wiskunde met onder andere minicursussen van Dirk Becherer (Berlijn) en Fred Espen Benth (Oslo).
plaats *De Werelt, Lunteren*
info *staff.science.uva.nl/~spreij/winterschool/winterschool.html*

5–6 februari 2016
❑ **Nationale Wiskunde Dagen**
22ste editie van de tweedaagse conferentie voor wiskundeleraren.
plaats *Noordwijkerhout*
info *www.uu.nl/nationale-wiskunde-dagen*

Nieuws | News

Deze rubriek is een kroniek van wiskundige activiteiten in Nederland. Toekomstige activiteiten worden aangekondigd en van voorbije activiteiten wordt verslag gedaan. Wilt u uw aankondiging of verslag in deze rubriek geplaatst zien? Stuur ons dan uw bijdrage van ± 350 woorden, zo mogelijk met illustratie. De redactie behoudt zich het recht voor berichten te weigeren of in te korten.

nieuws@nieuwarchief.nl

Spinozapremie voor Aad van der Vaart

Voor zijn baanbrekend onderzoek in de statistiek ontvangt de Leidse hoogleraar Stochastiek Aad van der Vaart de NWO-Spinozapremie.

Van der Vaart verdient deze meest prestigieuze wetenschapsprijs van Nederland omdat hij met zijn onderzoek naar wiskundige statistiek tot de internationale top van zijn vakgebied behoort, zo meldt NWO in het laudatio. Zijn boeken over schattingstheorieën zijn heel invloedrijke werken. Met zijn toegepaste statistiek levert Van der Vaart waardevolle bijdragen aan onder andere medische beeldvorming en statistische genetica. Vijftien jaar geleden zette hij in Nederland de Bayesiaanse niet-parametrische statistiek op de kaart. Mede dankzij Van der Vaarts verdiensten is deze richting uitgegroeid tot een toonaangevend onderzoeksonderwerp, aldus het laudatio. *leidenuniv.nl*

Nederlands succes op Internationale Wiskunde Olympiade

Bij de Internationale Wiskunde Olympiade in Thailand heeft het Nederlandse team drie bronzen medailles behaald. Deze medailles werden behaald door Bob Zwetsloot (17, Teylingen College, Noordwijkerhout), Eva van Ammers (17, Spinoza Lyceum, Amsterdam) en Yuhui Cheng (19, Wolfert Tweetalig, Rotterdam), die alle drie twee van de zes pittige wiskundeopgaven op wisten te lossen. Teamlid Tim Brouwer (18, Da Vinci College, Leiden) loste één van de zes opgaven op en behaalde hiermee een eervolle vermelding.

De Internationale Wiskunde Olympiade vond van 8 tot 16 juli plaats in Chiang Mai, Thailand. Op elk van de twee wedstrijddagen kregen de leerlingen drie opgaven van hoog wiskundig niveau voor hun kiezen, waar ze vier en een half uur aan konden werken. *wiskundeolympiade.nl*

Wiskundigen waarderen hun studie

Volgens een onderzoek van Vacatures.nl zijn wiskundigen van alle pas afgestudeerden van de Nederlandse universiteiten het meest enthousiast over hun studie. Universitair wiskundigen geven hun opleiding gemiddeld een $8{,}1$. De gemiddelde score van alle universitaire studies is $7{,}4$.

Het enthousiasme voor hun studie kan deels worden verklaard door de snelheid waarmee wiskundigen een baan op niveau vinden. Gemiddeld heeft een afgestudeerde wiskundige daar iets meer dan drie maanden voor nodig. De succeskansen bij sollicitaties zijn ongekend groot: de gemiddelde wiskundige starter solliciteert vier keer, waarbij hij vrijwel altijd op gesprek mag komen (gemiddeld $3{,}5$ sollicitatiegesprekken op $4{,}2$ brieven). Daarnaast speelt ook het salaris een rol: anderhalf jaar na afstuderen verdient een wiskundige gemiddeld 2.975 euro per maand, bijna 400 euro meer dan de gemiddelde academicus.

Het onderzoek is gebaseerd op het Elsevier-rapport 'Beste Banen 2015'. Voor het onderzoek zijn bijna 6000 hbo'ers en academici ondervraagd, anderhalf jaar na hun afstuderen. Zij rondden hun opleiding af in 2012 en 2013. *vacatures.nl en onderzoek.elsevier.nl*

Wifi in trein onveilig

Het wifi-netwerk in treinen van de NS blijkt volstrekt onveilig. Hannes Mühleisen, onderzoeker aan het CWI, verzamelde vanaf zijn woonboot nabij Amsterdam CS met twee simpele antennes vijf maanden lang data van tienduizenden treinreizigers en kon zo ongemerkt privacygevoelige gegevens achterhalen. De NS doet, ondanks herhaalde waarschuwingen, niets aan het probleem.

Volgens Mühleisen hoef je geen expert te zijn om het internetverkeer van treinreizigers af te tappen: “De antennes die ik heb gebruikt, kosten nog geen 30 euro per stuk en als je een beetje zoekt op internet kun je precies vinden hoe je informatie kunt onderscheppen en misbruiken.” Hij kon zien welke websites treinreizigers bezoeken en welke applicaties zij gebruiken, maar heeft er bewust voor gekozen om geen privacygevoelige gegevens op te slaan.

De NS zegt dat ze haar reizigers zo laagdrempelig mogelijk toegang wil bieden tot internet, maar reizigers wel waarschuwt voor de risico’s, zoals bijvoorbeeld het invoeren van creditcardgegevens via het wifi-netwerk. Mühleisen vindt dat onvoldoende: “De NS is een van de grootste internetproviders van ons land. Dan heb je ook een verantwoordelijkheid om je klanten te beschermen tegen cybercriminelen.” *decorrespondent.nl*

Foto: Marco Raaphorst

Stieltjesprijs voor Ziyang Gao en Bram Gorissen

Op voordracht van de selectiecommissie heeft WONDER net als vorig jaar besloten om — bij hoge uitzondering en gezien de uitzonderlijke kwaliteit — ook dit jaar twee Stieltjesprijzen uit te reiken, voor twee proefschriften op twee heel verschillende vakgebieden.

Ziyang Gao won de prijs voor zijn proefschrift *The mixed Ax-Lindemann theorem and its applications to the Zilber-Pink conjecture*, waarop hij op 24 november 2014 gepromoveerd is in Leiden bij prof.dr. S.J. Edixhoven en prof.dr. E. Ullmo (IHÉS, Université Paris-Sud).

Bram Gorissen won de prijs voor zijn proefschrift *Practical robust optimization techniques and improved inverse planning of HDR brachytherapy*, waarop hij op 19 september 2014 cum laude gepromoveerd is in Tilburg bij prof.dr.ir. D. den Hertog en copromotor dr.ir. A.L. Hoffmann.

Sinds 1996 wordt jaarlijks een Stieltjesprijs uitgereikt. Vanaf 2010 wordt deze prijs toegekend onder auspiciën van onderzoekschool WONDER voor het beste proefschrift in de wiskunde, dat in dat betreffende jaar is verdedigd aan een Nederlandse universiteit. De prijs bestaat uit een oorkonde en een bedrag van 2.500 euro, beschikbaar gesteld door de Stichting Compositio Mathematica. *Fred Bakker*

Bayesiaanse statistiek niet zo robuust als gedacht

De veelgebruikte Bayesiaanse statistiek is niet zo robuust als vaak wordt gedacht. Onderzoeker Thijs van Ommen van het CWI heeft ontdekt dat voor een bepaald type problemen de Bayesiaanse statistiek niet-bestaande patronen in data vindt. Van Ommen is op 10 juni op dit onderwerp gepromoveerd aan de Universiteit Leiden.

Bayesiaanse statistiek wordt vaak gebruikt om vast te stellen of een hypothese juist of onjuist is op basis van de bewijslast, en geeft een maat voor de zekerheid van deze conclusie. Deze vorm van statistiek wordt onder andere in de machine learning gebruikt. Van Ommen heeft ontdekt dat Bayesiaanse statistiek niet robuust is als bepaalde aannames in het model een klein beetje verkeerd zijn. Hij ontwierp verschillende datasets waarin de Bayesiaanse statistiek niet-bestaande patronen vond, gebaseerd op willekeurige ruis in de data. De datasets hadden allemaal realistische eigenschappen en zouden prima als echte experimentele data kunnen voorkomen.

De fouten treden op bij zogenaamde regressieanalyse. In deze vorm van data-analyse zoekt een onderzoeker naar de relatie tussen twee of meer variabelen, de ene bekend en de andere onbekend. Als hierbij modellen worden gebruikt die niet helemaal correct zijn, zoals bij een aanname dat de ruis een specifieke kansverdeling volgt, is er een risico dat onzinnige conclusies worden getrokken. Van Ommen stelt het probleem niet alleen vast, maar levert ook direct de oplossing in de vorm van een toevoeging aan de Bayesiaanse statistiek. Deze toevoeging, SafeBayes, voorkomt de genoemde problemen in regressieanalyse. Naar verwachting wordt dit binnenkort toegevoegd aan statistische software als R en SPSS. *cwi.nl*

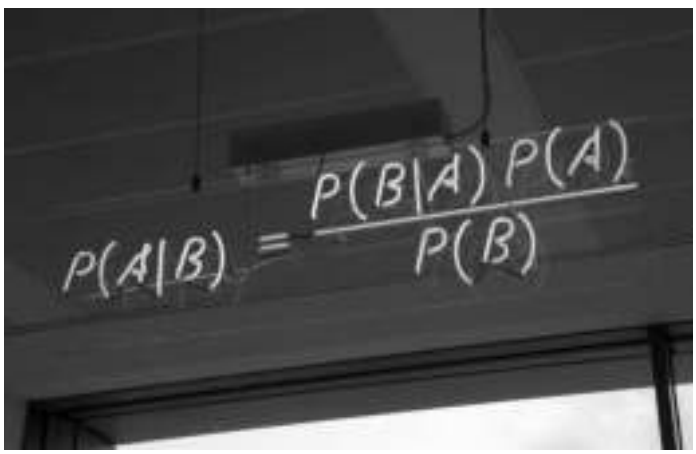


Foto: Matt Buck

Eduard de Jager overleden

Op 17 juli is Eduard de Jager, emeritus hoogleraar mathematische fysica aan de UvA, overleden. Hij is 88 jaar geworden. Van 1965–1968 was De Jager hoogleraar toegepaste wiskunde aan de TH Twente, en van 1968 tot aan zijn emeritaat in 1992 was hij verbonden aan de UvA als hoogleraar mathematische fysica. Van 1970–1975 was hij Managing Editor van het *Nieuw Archief voor Wiskunde*. *uva.nl*

Indagationes Mathematicae gedigitaliseerd

Het wiskundetijdschrift *Indagationes Mathematicae* is eigendom van het KWG en wordt uitgegeven door Elsevier. Het tijdschrift komt voort uit de *Proceedings van de Koninklijke Nederlandse Akademie van Wetenschappen Ser. A*. Recent heeft Elsevier de jaargangen 1951–1968 gedigitaliseerd. Hiermee zijn alle artikelen die ooit verschenen zijn bij *Indagationes Mathematicae* en zijn voorganger digitaal beschikbaar: 1895–1950 bij KNAW Digital Library en 1951–heden bij ScienceDirect. *Tom Koornwinder*

Johan van Leeuwaarden
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
j.s.h.v.leeuwaarden@tue.nl

Michel Mandjes
Korteweg-de Vries Instituut voor Wiskunde
Universiteit van Amsterdam
m.r.h.mandjes@uva.nl

NETWORKS

NETWORKS is een tien jaar lopend (2014–2023) Zwaartekrachtproject waarin, met de steun van het Ministerie van Onderwijs, Cultuur en Wetenschap, fundamenteel onderzoek gedaan wordt naar de eigenschappen van netwerken. Het project zal bruggen slaan tussen wiskunde en informatica, theorie en praktijk, en wetenschap en maatschappij. En bovenal zal NETWORKS veel jonge talenten opleiden die deze bruggen gaan bewandelen. Johan van Leeuwaarden en Michel Mandjes zijn lid van het projectteam.

In de wereld om ons heen hebben we te maken met zeer omvangrijke netwerken. Technologische ontwikkelingen hebben geleid tot grootschalige infrastructuren voor vervoer van mensen, producten, informatie en meer. In Nederland zijn miljoenen mensen verbonden door verkeers- en energienetwerken, in de wereld communiceren miljarden mensen via Facebook of het internet. Deze massale netwerken bieden zowel kansen als bedreigingen en het is daarom van cruciaal belang dat we ze begrijpen, zodat we in staat zijn ze te verbeteren en in bedwang te houden.

De taal die in het project NETWORKS wordt gebruikt is die van de wiskunde. Netwerken kunnen daarmee op een abstract niveau worden beschreven, met als doel universele eigenschappen of beginselen te ontdekken en nader te onderzoeken. Als we het brein beter begrijpen kunnen we ook iets leren over hoe we het internet zouden moeten inrichten of over hoe een epidemie aan banden gelegd kan worden. Wiskunde kan ook helpen bij het omgaan met de grootschaligheid, complexiteit en onvoorspelbaarheid die we in netwerken tegenkomen. Hierbij wordt met name een beroep gedaan op twee deelgebieden van de wiskunde: de *stochastiek* en de *algoritmiek*. Op deze manier raken deze twee onderzoeksvelden, die zich tot dusver vooral naast elkaar ontwikkelden, steeds meer met elkaar verweven. Tegelijkertijd is er sprake van een toenemende interactie met allerlei andere wetenschapsgebieden.

Stochastiek is de wiskunde van de grilligheid en onvoorspelbaarheid. Het is de uitdaging om, in weerwil van die intrinsieke onzekerheid, tot wiskundig betrouwbare uitspraken te komen. De meeste netwerken die we kennen zijn inherent stochastisch; tot op bepaalde hoogte onvoorspelbaar door gedrag van mensen, apparaten of veranderende omstandigheden (het weer, de economie of 'toevalligheden' van allerlei aard). De invloed van die toevallige factoren wordt versterkt door de karakteristieke hoge connectiviteit van netwerken (denk aan de beroemde theorie van 'six degrees of separation') en de doorgaans hoge belasting waaronder netwerken opereren. Onder deze 'kritieke' condities kan een kleine stochastische verstoring een desastreuze invloed hebben, zoals ondermaatse prestaties of in extreme gevallen zelfs het bezwijken van het hele netwerk.

Het is dus van belang om te begrijpen onder welke omstandigheden netwerken instabiel worden of ander ongewenst gedrag gaan vertonen. Waar ligt de grens, of het kookpunt, vanaf waar het netwerk niet meer naar behoren functioneert? Welke transmissiesnelheden kunnen via draadloze netwerken gerealiseerd worden? Hoeveel auto's kan het wegennet faciliteren terwijl de kans op massale filevorming gering blijft? Wat is het minimale percentage van de bevolking dat ingeënt moet zijn om een eventuele epidemie te voorkomen? Deze vragen illustreren dat er een grote behoefte bestaat aan fundamenteel nieuwe denkwijzen om grote stochastische netwerken te doorgronden, niet alleen om ze efficiënt te kunnen benutten, maar ook om eventuele catastrofale ontwikkelingen af te wenden.

Algoritmes spelen hierin een cruciale rol. Populair gezegd is een algoritme een recept dat voorschrijft hoe een netwerk bestuurd wordt. Zo zijn er algoritmes die de kortste weg van A naar B bepalen, of de opslag en transport in een energienetwerk coördineren. Ook aan Google ligt een algoritme ten grondslag: het beroemde *PageRank*. Doordat de onderliggende netwerken steeds

Illustratie: Ryu Tajiri

groter en complexer worden, ontstaan er forse algoritmische uitdagingen. Algoritmes moeten snel en accuraat zijn, en bovendien kunnen omgaan met veranderende omstandigheden. Denk aan een aangepaste dienstregeling voor treinen wanneer delen van het netwerk wegvallen. Of een internetprovider die beslist dat het dataverkeer langs andere routes gestuurd moet worden als bepaalde knooppunten gebukt gaan onder overbelasting. Of aan energievoorraden die aangesproken zouden kunnen worden op windstille dagen. Dit zijn voorbeelden van uitdagingen die, zij het verpakt in een abstracte formulering, onderzocht worden in NETWORKS.

Voorbeelden van samenwerking

We bespreken nu voorbeelden van samenwerking binnen NETWORKS waarbij de stochastiek en de algoritmiek nader tot elkaar komen. Een tweede gemeenschappelijk punt is dat de samenwerking telkens wordt gemotiveerd door grote praktische thema's en aanleiding geeft tot fundamentele onderzoeksvragen.

Het eerste voorbeeld betreft de *besliskunde*. De mathematisch besliskundigen in Nederland ontmoeten elkaar in januari, tijdens de jaarlijkse conferentie, traditiegetrouw gehouden in congrescentrum De Werelt te Lunteren. Er wordt door veel besliskunigen aan netwerken gewerkt, maar een ieder doet dat vauit een eigen perspectief. Op hoog niveau zijn er twee stromingen te onderscheiden, die slechts in beperkte mate met elkaar communiceren: de deterministische en de stochastische besliskunde. En hoewel vaak naar dezelfde toepassingen wordt gekeken, verkiezen de deterministici toch geheel andere wiskundige methoden dan de stochastici. Neem als voorbeeld een communicatienetwerk, zoals het internet. Met een beschrijving van de infrastructuur van dit netwerk, in termen van een graaf met een verzameling knopen verbonden door lijnen, heeft men een wiskundig object waaraan op veel manieren gerekend kan worden. Zo zou je met een deterministisch algoritme het kortste pad kunnen berekenen tussen twee knopen, of een route die anderszins optimaal is. Voor het internet alleen al zijn er talloze algoritmes die kunnen worden uitgevoerd om bepaalde eigenschappen van de internetgraaf te analyseren en aan de hand daarvan voorspellingen te doen of beslissingen te nemen over de optimale route voor een datapakket; binnen de informatica is dit dan ook een belangrijk onderwerp.

Stochastici zijn geïnteresseerd in dezelfde vraagstukken, maar voegen dan vaak een bron van onzekerheid toe, bijvoorbeeld door het gedrag van de gebruikers mee te nemen in het wiskundig model. In dat geval blijft de internetgraaf onveranderd, maar is er sprake van een onzeker proces van datapakketjes dat over de graaf stroomt, waarbij de individuele knooppunten vaak worden beschreven door *wachtrijen* waarin de datapakketjes tijdelijk kunnen worden opgeslagen. In andere modellen kan ook de graaf zelf onzeker zijn, bijvoorbeeld als er verbindingen (tijdelijk) wegvallen en bijkomen. Zelfs het algoritme dat de route kiest kan stochastisch zijn, als er bij wijze van spreken een muntje wordt gegooid dat bepaalt of een datapakket langs de ene of de andere route wordt gestuurd. Deze 'willekeurige' algoritmes worden steeds belangrijker, omdat ze schaalbaar zijn, in tegenstelling tot algoritmes die 'alle' informatie gebruiken (preciezer gezegd: algoritmes die gebaseerd zijn op een gedetailleerde beschrijving van de toestand van het gehele netwerk van dat moment).

Communicatienetwerken spelen een belangrijke rol in NETWORKS. Het onderzoek richt zich op allerlei uitdagende problemen op het gebied van vaste en draadloze netwerken, waarbij er in het bijzonder ook aandacht besteed wordt aan geavanceerde optische netwerken.

NETWORKS houdt zich naast communicatienetwerken ook bezig met verkeersnetwerken. Bij het ontwerpen van een verkeersnetwerk bestaat de eerste fase uit ruwe berekeningen waarbij het leidende principe is dat de capaciteit van de wegen voldoende groot moet zijn, met een planningshorizon van wellicht enkele decennia. De benodigde berekeningen zijn, in essentie, combinatorisch van aard: vanuit de huidige infrastructuur wil men tegen minimale kosten tot een netwerk komen dat aan de gestelde eisen voldoet. Op kleinere tijdschaal speelt het echter een rol dat verkeersstromen niet deterministisch zijn, maar juist inherent stochastisch. Er zijn allerlei grillige effecten die van invloed zijn; denk aan de verschillen in rijgedrag, maar ook aan de invloed van het weer. Om te komen tot een realistische analyse van de capaciteit van wegen, moeten deze stochastische factoren worden meegenomen. Bovendien moet men ook kunnen inspelen op ongelukken en werkzaamheden, bijvoorbeeld door verkeer langs een alternatieve route te sturen. Hiermee komen we dan op het gebied van de algoritmiek terecht: gegeven de huidige condities, moet het juiste advies aan de automobilisten worden gegeven. Het onderliggende algoritme moet snel ('real-time') en adaptief zijn; er is geen tijd voor tijdrovend rekenwerk. NETWORKS brengt de disciplines bij elkaar die nodig zijn bij het oplossen van dergelijke algoritmische vraagstukken. Op een vergelijkbare manier wordt binnen NETWORKS ook naar energienetwerken gekeken.

Een derde voorbeeld van samenwerking betreft *toevallige grafen* die worden gebruikt om complexe netwerken te beschrijven. Waar bij een verkeersnetwerk de infrastructuur doorgaans als gegeven wordt beschouwd, is dat bij veel andere complexe netwerken niet langer het geval. Miljarden mensen over de hele wereld wisselen informatie uit over Facebook en dat levert grafen met knopen en verbindingen op die eenvoudigweg te complex zijn om op detailniveau in kaart te brengen. Een manier om de complexiteit te lijf te gaan is door gebruik te maken van het concept van toevallige grafen (random graphs). De bekendste toevallige graaf is vernoemd naar Paul Erdős en Alfréd Rényi: de zogenaamde *Erdős–Rényi-graaf* bestaat uit n knopen, waarbij ieder paar knopen met een vaste kans p verbonden is met een lijn.

Toevallige grafen liggen op het snijvlak van de kansrekening (stochastiek) en grafentheorie (algoritmiek). Veel eigenschappen van toevallige grafen kan men bestuderen door alleen lokale informatie van de infrastructuur te kennen. Zo kan de structuur van het netwerk worden beschreven door vanuit een knoop, steeds de aangrenzende knopen in kaart te brengen. Op dit principe gebaseerde algoritmes zijn recentelijk erg krachtig gebleken: ze zijn schaalbaar (in de zin dat ze toegepast kunnen worden in grote netwerken), maar tegelijkertijd vertonen ze onder bepaalde voorwaarden bijna-optimaal gedrag.

Om de infrastructuur van een online sociaal netwerk op schaalbare wijze te benutten, kan men een algoritme zo ontwerpen dat iedereen de hem of haar beschikbare informatie aan vrienden doorgeeft, die het vervolgens weer doorgeven aan hun vrienden. Ondanks het feit dat het onderliggende netwerk zeer omvangrijk is, zal de informatie zich als een lopend vuurtje verspreiden. Op wiskundig niveau gebeurt bij het ontstaan van een epidemie (computervirussen op communicatienetwerken, of ziektes op populaties) echter in feite hetzelfde. Ook hier geldt: greep krijgen op deze netwerkfenomenen vraagt om de bundeling van de algoritmische en stochastische krachten.

Diversiteit, gemeenschappelijke uitdagingen

NETWORKS is een groot onderzoeksproject dat bestaat uit een aanzienlijk aan-

tal uiteenlopende thema's: ruimtelijke netwerken, dynamische netwerken, kwantumnetwerken, transportnetwerken, communicatienetwerken en energienetwerken. Op voorheen zelfstandig opererende gebieden en onderwerpen zijn inmiddels zes universitair docenten aangesteld en meer dan twintig promovendi. Daarnaast zijn sinds de start van NETWORKS steeds meer onderzoekers betrokken geraakt. Binnen het project spelen outreach en kennisuitwisseling via workshops een belangrijke rol en is er bovendien een trainingsprogramma opgezet dat bedoeld is voor de verdere inhoudelijke scholing van de staf en promovendi.

Ondanks de schaal en heterogeniteit van het NETWORKS-programma zijn de uitgangspunten herkenbaar. We schetsten al eerder de gedachten achter het bundelen van stochastiek en algoritmiek, maar daarnaast wordt een prominente rol gespeeld door twee overkoepelende uitdagingen. De eerste uitdaging is het introduceren van *zelforganisatie* in echte netwerken. Het achterliggende idee is om algoritmes te ontwerpen waarmee het netwerk in feite zichzelf aanstuurt. Het netwerk neemt dan met lokale niet-volledige informatie op gedistribueerde wijze beslissingen, maar vertoont desondanks op globaal niveau nagenoeg optimaal gedrag. De tweede uitdaging betreft het bouwen van een universele theorie voor intelligente netwerken. Die intelligente netwerken worden aangestuurd door de hierboven besproken real-time algoritmes. Het doel is om op abstract niveau te formaliseren wat intelligentie betekent in de context van netwerken.

Beide uitdagingen zijn aanzienlijk: de eerste vanwege het feit dat de toepasbaarheid centraal staat, wat de onderzoekers dwingt om de resultaten te toetsen aan de praktische uitvoerbaarheid, en de tweede als gevolg van het ontbreken van alomvattende theorievorming op het meest globale, abstracte niveau. Met de combinatie van deze twee uitdagingen, waarbij praktische toetsing en theorievorming dus hand in hand moeten gaan, komen we terecht bij de kern van NETWORKS. Intelligente netwerken worden bestuurd door algoritmes die in real-time werken, die kunnen leren, die kunnen omgaan met grillige omstandigheden en die uiteindelijk een netwerk zelfstandig kunnen laten opereren. We hoeven dan ook de infrastructuur van het brein, het internet of het verkeersnetwerk niet volledig te kennen om het toch te kunnen besturen. De wiskunde zal de komende jaren algoritmes tot stand brengen om de steeds toenemende omvang en complexiteit van netwerken het hoofd te bieden. Kunnen we die weerbarstige netwerken, waarvan wij allen afhankelijk zijn, beheersen door intelligente, zelfstandig opererende, maar niettemin betrouwbare algoritmes te ontwikkelen?

Meer informatie over NETWORKS is te vinden op de website van het project: www.thenetworkcenter.nl.

Leo Kroon
Rotterdam School of Management
Erasmus Universiteit Rotterdam, en
NS Reizigers, Proceskwaliteit en Innovatie
lkroon@rsm.nl

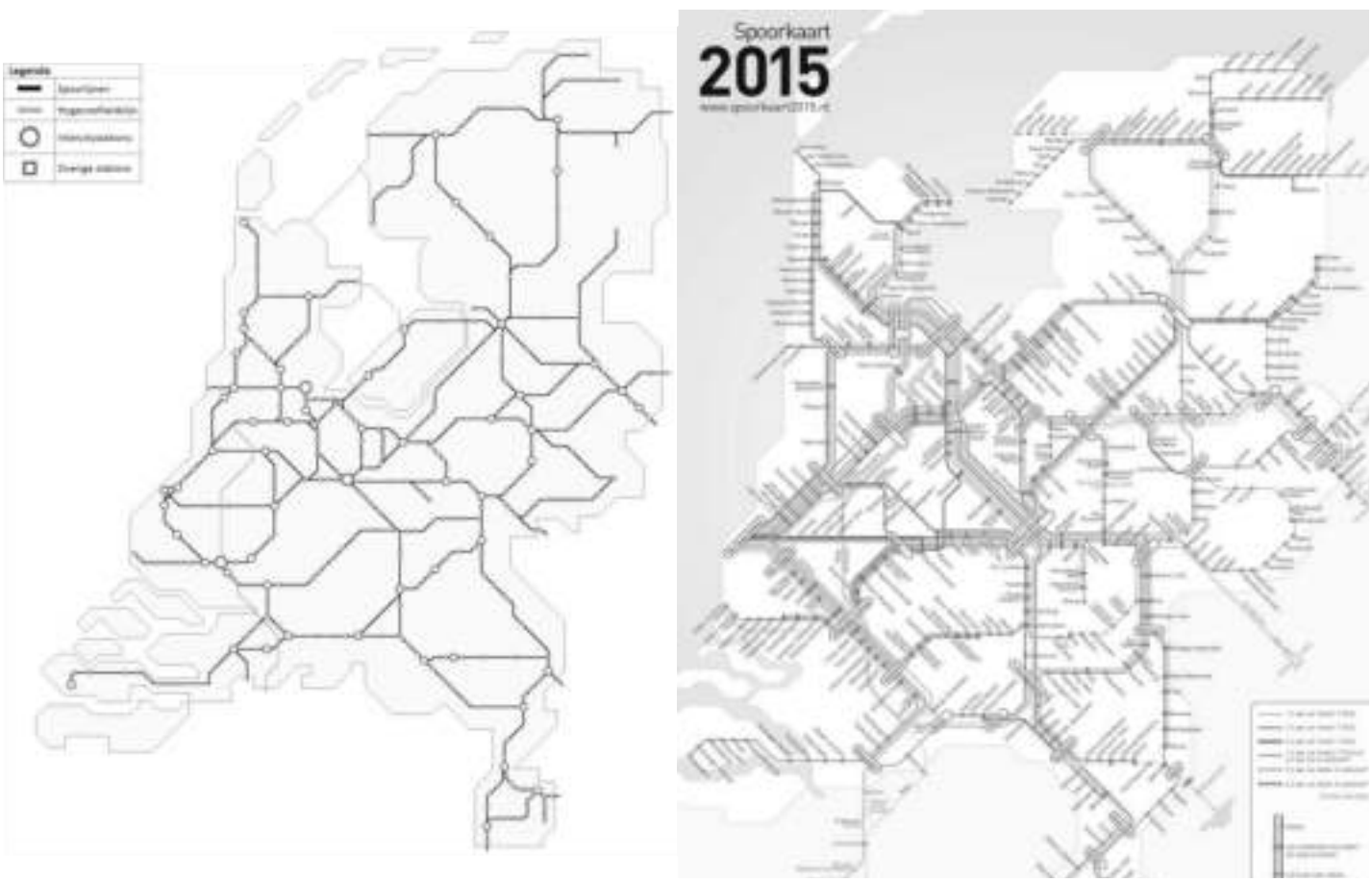
Lex Schrijver
Korteweg-de Vries Instituut voor Wiskunde
Universiteit van Amsterdam, en
Centrum Wiskunde & Informatica
lex@cwi.nl

Spoornetwerken

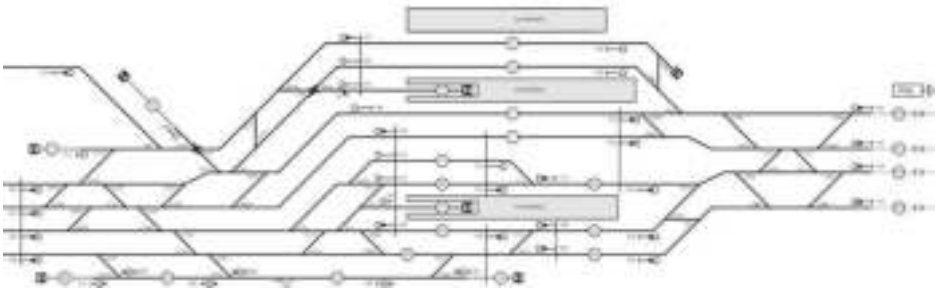
Leo Kroon en Lex Schrijver hebben samen met teams bij NS, EUR en CWI algoritmiek ontwikkeld waarmee de dienstregeling van NS gepland kan worden, alsmede de bijbehorende routering van treinen door stations, de materieelomloop, en de personeelsdiensten. Deze algoritmen worden regelmatig door NS gebruikt, sinds hun eerste integrale toepassing ten behoeve van het 'Spoorboekje 2007'. Hierbij ging de dienstregeling van NS eind 2006 grondig op de schop. Binnen deze algoritmen spelen netwerken, in verschillende vormen, een belangrijke rol.

Wie aan spoorwegen denkt, denkt al snel aan netwerken. Daarbij springt het netwerk van spoortrajecten als eerste in het oog, zie Figuur 1 (links): een netwerk van stations die door middel van sporen met elkaar verbonden zijn. Een meer gedetailleerd netwerk, dat over het netwerk van spoortrajecten heen ligt, is het netwerk van spoorlijnen, zie Figuur 1 (rechts): de rechtstreekse treinverbindingen, ieder met een herkomst, een bestemming, een frequentie, en een lijst van stations waar gestopt wordt. En als je verder inzoomt, bijvoorbeeld op station Gouda in het netwerk van spoortrajecten, dan zie je het veel gedetailleerdere spoornetwerk in Figuur 2.

Minder zichtbaar zijn de virtuele wiskundige netwerken die een onmisbare rol spelen



Figuur 1 Het netwerk van spoortrajecten (links) en het netwerk van spoorlijnen (rechts).



Figuur 2 Infrastructuur van station Gouda.

Zwolle	9 56	10 26	10 56	11 26	11 56	12 26	12 56	13 26	13 56	14 26
Dalfsen	10 05	10 34	11 05	11 34	12 05	12 34	13 05	13 34	14 05	14 34
Emmen	10 15	10 44	11 15	11 44	12 15	12 44	13 15	13 44	14 15	14 44
Marienberg	|	10 51	|	11 51	|	12 51	|	13 51	|	14 51
Handenberg	10 26	10 57	11 26	11 57	12 26	12 57	13 26	13 57	14 26	14 57
Grameergen	|	11 01	|	12 01	|	13 01	|	14 01	|	15 01
Coevorden	10 36	11 07	11 36	12 07	12 36	13 07	13 36	14 07	14 36	15 07
Dalen	|	11 11	|	12 11	|	13 11	|	14 11	|	15 11
Nieuw Amsterdam	10 44	11 18	11 44	12 18	12 44	13 18	13 44	14 18	14 44	15 18
Emmen Bargeres	|	11 23	|	12 23	|	13 23	|	14 23	|	15 23
Emmen A	10 51	11 28	11 51	12 28	12 51	13 28	13 51	14 28	14 51	15 28

Figuur 3 Dienstregeling Zwolle – Emmen.

bij de operationele planning van het spoorsysteem, zoals de planning van de dienstregeling, de materieelomloop en de personeelsinzet, en de routering van treinen over emplacementen.

De klassieke netwerkbegrippen *potentiaal*, *spanning* en *stroom*, bekend uit de elektriciteitsleer, blijken ook van belang in de logistiek van de spoorwegen. Deze begrippen zijn uiteraard belangrijk in de energievoorziening aan het elektrische spoorwegmaterieel. Maar, zoals we zullen zien, potentialen, spanningen en stromen in netwerken vormen ook een belangrijk gereedschap bij het *plannen* van verschillende spoorprocessen. De inzet op de volgende bladzijde geeft enkele relevante definities en eigenschappen.

We richten ons in het bijzonder op twee fasen in het planningsproces:

1. het bepalen van de dienstregeling,
2. het bepalen van de materieelomloop.

In geval 1 worden de *eisen* waaraan de dienstregeling moet voldoen voorgesteld als een netwerk, en de dienstregeling als een potentiaal op dit netwerk. De punten van dit netwerk representeren de vast te stellen vertrek- en aankomsttijden. Deze modellering gaat terug op het werk van Van der Aalst, Van Hee en Voorhoeve [12]. In geval 2 wordt de *dienstregeling zelf* als netwerk voorgesteld, en de materieelomloop als stroom in dit netwerk. Deze onderwerpen komen aan de orde in de volgende twee paragrafen. De laatste paragraaf geeft een beknopte beschrijving van nog enkele andere onderwerpen.

Figuur 4 Treinen van series 46 en 15 op het traject Amsterdam – Amersfoort.

De dienstregeling en het netwerk van eisen

Het Basis Uur Patroon

De basis van de Nederlandse dienstregeling is het *Basis Uur Patroon* (BUP): de dienstregeling herhaalt zich elke 60 minuten. Dus als een bepaalde trein op tijdstip t van station s vertrekt, dan vertrekt er ook op de tijdstippen $\ldots, t-120, t-60$ en $t+60, t+120, \ldots$ een vergelijkbare trein van station s, met dezelfde bestemmingen. Hieraan worden spitstreinen toegevoegd, en in de avond en nacht en in het weekend zijn er andere afwijkingen, maar het BUP is het onderliggende patroon. Veel treinseries hebben overigens een 30-minutenfrequentie, maar dat kan eenvoudig in het model worden meegenomen.

Het betekent bijvoorbeeld dat elk uur, om 26 minuten over het hele uur, een stoptrein van Zwolle naar Emmen vertrekt, van 's morgens vroeg tot 's avonds laat. Zie Figuur 3. (De voorbeelden in dit artikel zijn gebaseerd op de situatie op het Nederlandse spoorwegnet, maar komen niet steeds overeen met de huidige (2015) dienstregeling en overige omstandigheden.)

Elke uurlijkse vertrek- en aankomsttijd kan worden gegeven door de minuut in het uur, dus door een element uit de verzameling $\mathbb{Z}_{60} = \mathbb{Z}/60\mathbb{Z}$ van getallen modulo 60. Het vaststellen van de tijden in het BUP komt dan neer op het toekennen van elementen uit $\mathbb{Z}_{60}$ aan variabelen. Wat zijn deze variabelen en waaraan moeten ze voldoen?

De uurlijkse treinen

Voor elke uurlijkse treinserie en elke af te leggen etappe (rit) definiëren we twee punten, namelijk een punt voor de vertrektijd en een punt voor de aankomsttijd. Bekijk, als voorbeeld, de treinserie 15, een uurlijkse intercitytrein van Amsterdam naar Amersfoort. Zie Figuur 4. Deze stopt onderweg in Hilversum. Dit geeft vier punten in het netwerk:

$$v_{15,\mathrm{Asd,V}},\ v_{15,\mathrm{Hvs,A}},\ v_{15,\mathrm{Hvs,V}} \text{ en } v_{15,\mathrm{Amf,A}}. \quad (1)$$

Hierin zijn Asd, Hvs en Amf de afkortingen voor Amsterdam, Hilversum en Amersfoort, en staan V en A voor vertrektijd en aankomsttijd. Vervolgens voeren we ook pijlen in, overeenkomend met de verschillende etappes en stationnementen van treinserie 15:

$$(v_{15,\mathrm{Asd,V}}, v_{15,\mathrm{Hvs,A}}), (v_{15,\mathrm{Hvs,A}}, v_{15,\mathrm{Hvs,V}}) \text{ en } (v_{15,\mathrm{Hvs,V}}, v_{15,\mathrm{Amf,A}}). \quad (2)$$

Figuur 5 Eisen aan tijden in treinserie 15 Amsterdam – Hilversum – Amersfoort.

Stroom en spanning

In de context van dit artikel is een *netwerk* (meestal) hetzelfde als een *gerichte graaf*: een geordend paar $N = (V, E)$, waarbij V een eindige verzameling is en E een deelverzameling van $V \times V$ met $(v, v) \notin E$ voor elke $v \in V$. De elementen van V en E heten de *punten* respectievelijk de *pijlen* van het netwerk. Een pijl (u, v) *vertrekt uit* u en *komt aan in* v.
Als R een additieve abelse groep is, dan is een *stroom* een functie $f : E \to R$ die voldoet aan de *wet van stroombehoud*:

$$\sum_{e \in \delta^{\text{in}}(v)} f(e) = \sum_{e \in \delta^{\text{uit}}(v)} f(e) \quad \text{voor elke } v \in V.$$

Hierin is $\delta^{\text{in}}(v)$ de verzameling in v aankomende pijlen, en $\delta^{\text{uit}}(v)$ de verzameling uit v vertrekkende pijlen.
Een functie $g : E \to R$ heet een *spanning* als er een functie $p : V \to R$ (een *potentiaal*) bestaat zo dat

$$g(u, v) = p(v) - p(u) \quad \text{voor elke } (u, v) \in E.$$

De verzameling S van stromen en de verzameling T van spanningen vormen elk een deelmoduul van het R-moduul R^E.
Als R een commutatieve ring met eenheid is, dan voldoen S en T aan de volgende dualiteit. Het *inwendig product* $\langle x, y \rangle$ van elementen $x, y \in R^E$ is als gebruikelijk

$$\langle x, y \rangle := \sum_{e \in E} x(e) y(e).$$

Voor $X \subseteq R^E$ kan men dan een orthogonaal complement $X^\perp$ definiëren:

$$X^\perp := \{y \in R^E \mid \langle x, y \rangle = 0 \text{ voor elke } x \in X\}.$$

Dan is het niet moeilijk om te bewijzen:

$$T = S^\perp \quad \text{en} \quad S = T^\perp.$$

Ook geldt dat als B een opspannende boom in N is, dan is elke functie op de pijlen *in* B uniek uit te breiden op de pijlen *buiten* B zodat een spanning ontstaat. Evenzo kan elke functie op de pijlen *buiten* B uniek worden uitgebreid op de pijlen *in* B zodat een stroom ontstaat.
Een basisalgoritme uit de besliskunde, geïnitieerd door Ford en Fulkerson [3], en *sterk polynomiaal* gemaakt door Tardos [10], geeft het volgende. Als R de ring $\mathbb{Z}$ van gehele getallen is, en er voor elke $(u, v) \in E$ een convexe deelverzameling $I(u, v)$ van $\mathbb{Z}$ gegeven is, dan kan snel (dat wil zeggen, binnen polynomiaal-begrensde tijd) een stroom f gevonden worden zo dat

$$f(u, v) \in I(u, v) \quad \text{voor elke } (u, v) \in E$$

(mits zo'n stroom bestaat). Sterker: als ook nog een functie $c : E \to \mathbb{Z}$ (de *kostenfunctie*) gegeven is, dan kan snel zo'n stroom f gevonden worden waarvoor $\langle c, f \rangle$ minimaal is. Hetzelfde geldt voor spanningen.

Voor elke etappe zijn onder- en bovengrenzen voorgeschreven voor de rijtijden, en voor elke tussenstop zijn onder- en bovengrenzen voorgeschreven voor de halteertijd.

Dus voor elke pijl (u, v) is er een interval $I(u, v)$ voorgeschreven waarin het verschil van opvolgende vertrek en aankomsttijden dient te liggen. Zoals in het Amsterdam – Amersfoort voorbeeld:

$$\begin{aligned} &I(v_{15,\text{Asd,V}}, v_{15,\text{Hvs,A}}) = [20, 22], \\ &I(v_{15,\text{Hvs,A}}, v_{15,\text{Hvs,V}}) = [1, 2], \\ &I(v_{15,\text{Hvs,V}}, v_{15,\text{Amf,A}}) = [12, 13]. \end{aligned} \tag{3}$$

Hier is $[a, b]$ het interval van gehele getallen x met $a \leq x \leq b$ modulo 60. Voorgeschreven is dus dat de rijtijd van Amsterdam naar Hilversum minimaal 20 minuten bedraagt, en maximaal 22 minuten. De twee andere eisen hebben een vergelijkbare betekenis. We kunnen deze eisen aan de treintijden grafisch weergeven zoals in Figuur 5. Een aan deze eisen voldoende dienstregeling correspondeert dus met een $\mathbb{Z}_{60}$-waardige potentiaal p zo dat $p(v) - p(u) \in I(u, v)$ voor elke pijl (u, v), oftewel met een $\mathbb{Z}_{60}$-waardige spanning g zo dat $g(u, v) \in I(u, v)$ voor elke pijl (u, v).

Voor de dienstregeling van treinserie 15 die weergegeven wordt in Figuur 4 geeft dit

$$\begin{aligned} &p(v_{15,\text{Asd,V}}) = 27, \\ &p(v_{15,\text{Hvs,A}}) = 48, \\ &p(v_{15,\text{Hvs,V}}) = 49, \\ &p(v_{15,\text{Amf,A}}) = 2. \end{aligned} \tag{4}$$

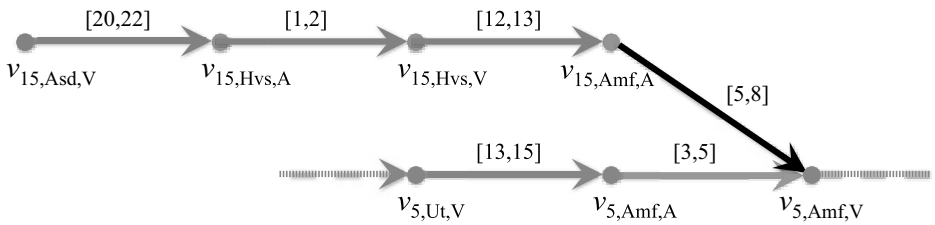
Deze waarden voldoen inderdaad aan de eisen. Immers:

$$\begin{aligned} &g(v_{15,\text{Asd,V}}, v_{15,\text{Hvs,A}}) \\ &\quad = p(v_{15,\text{Hvs,A}}) - p(v_{15,\text{Asd,V}}) \\ &\quad = 48 - 27 \equiv 21 \in [20, 22], \\ &g(v_{15,\text{Hvs,A}}, v_{15,\text{Hvs,V}}) \\ &\quad = p(v_{15,\text{Hvs,V}}) - p(v_{15,\text{Hvs,A}}) \\ &\quad = 49 - 48 \equiv 1 \in [1, 2], \\ &g(v_{15,\text{Hvs,V}}, v_{15,\text{Amf,A}}) \\ &\quad = p(v_{15,\text{Amf,A}}) - p(v_{15,\text{Hvs,V}}) \\ &\quad = 2 - 49 \equiv 13 \in [12, 13]. \end{aligned} \tag{5}$$

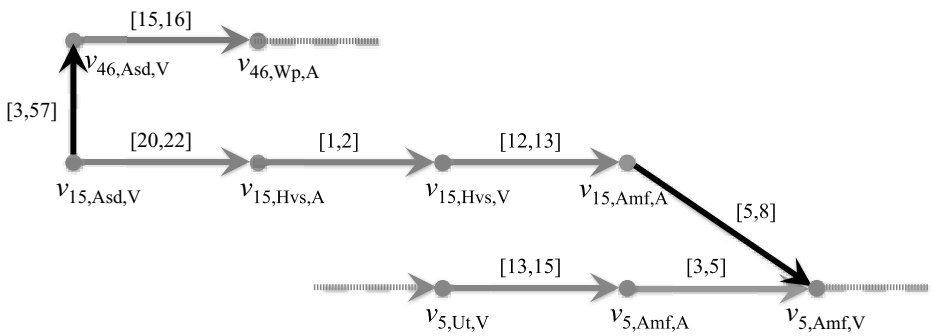
Dit zijn enkele eisen voor één treinserie. Alle treinseries samen geven een groot aantal los van elkaar liggende, gerichte paden, elk corresponderend met een treinserie, heen of terug.

Aansluitingen

Samenhang *tussen* de treinseries ontstaat door de gewenste overstapmogelijkheden. Zo is er een uurlijkse intercity treinserie 5 van Rotterdam via Amersfoort naar Groningen, waarop de treinserie 15 uit Amsterdam in Amersfoort dient aan te sluiten. Daartoe moet de trein naar Groningen minimaal 5 en maximaal 8 minuten na aankomst van de trein uit Amsterdam van Amersfoort vertrekken. Hierdoor ontstaat een nieuwe pijl, tussen punten van twee verschillende treinseries: $(v_{15,\text{Amf,A}}, v_{5,\text{Amf,V}})$, met als interval



Figuur 6 Aansluiting van treinserie 15 op treinserie 5 in Amersfoort.



Figuur 7 Voorkomen van conflict tussen treinseries 15 en 46 op Amsterdam – Weesp.

$$I(v_{15,\mathrm{Amf,A}}, v_{5,\mathrm{Amf,V}}) = [5, 8]. \quad (6)$$

Dit is grafisch weergegeven in Figuur 6. Alle overstapvoorwaarden tezamen geven al veel meer samenhang aan het netwerk.

Voorkomen van conflicten

Daarnaast moeten zogenaamde *conflicten* worden vermeden: twee treinen mogen niet gelijktijdig van hetzelfde spoor gebruik maken, bijvoorbeeld op een kruising of op enkelspoor, maar ook langs een traject: om te zorgen dat er voldoende ruimte tussen twee treinen wordt gepland, wordt een minimale *opvolgtijd* van drie minuten geëist, dit ook om kleine vertragingen op te kunnen vangen.

Als voorbeeld de treinserie 46 van Amsterdam naar Lelystad, zie Figuur 4. Deze gebruikt op het traject Amsterdam – Weesp hetzelfde spoor als de bovengenoemde treinserie 15 van Amsterdam naar Amersfoort. Op dit traject moeten de treinen een minimale afstand van drie minuten in tijd hebben. Dit geeft dan de pijl $(v_{15,\mathrm{Asd,V}}, v_{46,\mathrm{Asd,V}})$ met

$$I(v_{15,\mathrm{Asd,V}}, v_{46,\mathrm{Asd,V}}) = [3, 57]. \quad (7)$$

Hierdoor wordt dus voorkomen dat het verschil in de vertrektijden vanuit Amsterdam gelijk is aan 58, 59, 0, 1 of 2 (modulo 60) – dan zouden de treinen te dicht op elkaar zitten. Zie Figuur 7. Deze constraints kunnen sterker gemaakt worden door ook de minimale rijtijdverschillen tussen de treinen mee te nemen.

Op deze manier kunnen ook de beperkingen op enkelsporige trajecten en kruisingen worden beschreven. Deze eisen zorgen voor de grootste toevoeging van pijlen aan het netwerk, en, ondanks de ruime intervallen, ook voor de grootste toevoeging van complexiteit.

Periodic Event Scheduling

Het bepalen van een dienstregeling die aan de gestelde eisen voldoet komt nu dus neer op het vinden van een potentiaal $p : V \to \mathbb{Z}_{60}$ die voldoet aan

$$p(v) - p(u) \in I(u, v) \quad \text{voor elke pijl } (u, v). \quad (8)$$

Dat wil zeggen, een spanning $g : E \to \mathbb{Z}_{60}$ die voldoet aan

$$g(u, v) \in I(u, v) \text{ voor elke pijl } (u, v). \quad (9)$$

Dit *Periodic Event Scheduling Problem* (PESP) is in zijn algemeenheid een NP-volledig probleem: graafkleuring kan er eenvoudig toe worden gereduceerd. Dit in tegenstelling (als NP≠P) tot het geval waarin we $\mathbb{Z}_{60}$ zouden vervangen door $\mathbb{Z}$, waarmee, zoals we in de inzet meldden, het probleem snel oplosbaar zou zijn.

Je zou natuurlijk kunnen stellen dat elke $\mathbb{Z}$-oplossing ook een $\mathbb{Z}_{60}$-oplossing is, maar er kan een $\mathbb{Z}_{60}$-oplossing bestaan zonder dat er een $\mathbb{Z}$-oplossing bestaat.

NP-volledigheid is echter niet het ultieme antwoord: NP-volledigheid heeft betrekking op de asymptotische moeilijkheid van het probleem, terwijl hier één concreet, eindig probleem voorligt. We beschrijven twee aanpakken om een potentiaal te vinden, als die bestaat.

Constraint propagation

We kunnen de eisen beschrijven in een $V \times V$-matrix M waarvan de elementen deelverzamelingen van $\mathbb{Z}_{60}$ zijn, dus als matrix in $\mathcal{P}(\mathbb{Z}_{60})^{V \times V}$. Als (u, v) een pijl is, definiëren we $M_{u,v} := I(u, v)$. Als (u, v) geen pijl is en $u \neq v$, dan $M_{u,v} := \mathbb{Z}_{60}$, en als $u = v$, dan $M_{u,v} := \{0\}$.

Dergelijke matrices kunnen vermenigvuldigd worden in de $(\cap, +)$-algebra, als variant van de gebruikelijke $(+, \times)$-algebra. Als A en B matrices zijn met elementen in $\mathcal{P}(\mathbb{Z}_{60})$, vervangen we $+$ door $\cap$ en $\times$ door $+$, en krijgen we het $(\cap, +)$-product $A * B$ als:

$$(A * B)_{i,j} = (A_{i,1} + B_{1,j}) \cap \cdots \cap (A_{i,n} + B_{n,j}), \quad (10)$$

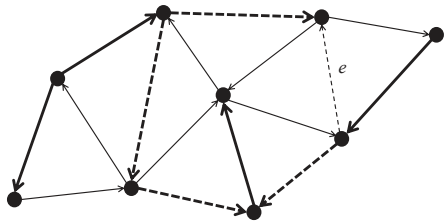
waarbij $X + Y := \{x + y \mid x \in X, y \in Y\}$ voor $X, Y \subseteq \mathbb{Z}_{60}$.

Het is niet moeilijk in te zien dat als $p(v) - p(u) \in A_{u,v}$ voor elke $u, v \in V$, dan ook $p(v) - p(u) \in (A^{*2})_{u,v}$ voor elke $u, v \in V$, waarbij A^{*2} het $(\cap, +)$-kwadraat is van A. Bovendien geldt $(A^{*2})_{u,v} \subseteq A_{u,v}$ voor elke $u, v \in V$, doordat $A_{v,v} = \{0\}$ voor elke $v \in V$.

Daarom kunnen we, beginnend met $A := M$, en A iteratief vervangend door A^{*2}, de eisen beschreven in de initiële matrix M verscherpen, totdat geen van de elementen van de matrix A strikt kleiner wordt. Dit is een voorbeeld van *constraint propagation*.

Mocht na een aantal iteraties blijken dat een element in A gelijk is aan $\emptyset$, dan weten we dat er geen potentiaal (dat wil zeggen dienstregeling) bestaat die aan alle eisen voldoet. Dit is de terugmelding van het algoritme, mogelijk met een beschrijving van de stappen die tot $\emptyset$ hebben geleid. Hierdoor wordt een minimale verzameling eisen geïdentificeerd waarvan er tenminste één verruimd moet worden om een dienstregeling mogelijk te maken.

Als alle elementen in de uiteindelijke matrix A niet-leeg zijn, geeft dit een instrument om gericht te zoeken naar een oplossing met een vertakkend zoekalgoritme. Overigens is



Figuur 8 De dikke (doorlopende of gestreepte) pijlen geven een opspannende boom weer. De gestreepte (dikke of dunne) pijlen geven de met pijl e gevormde *fundamentele cykel* C_e weer.

het bestaan van een toegelaten oplossing nog steeds niet gegarandeerd. Dit is de basis van het op het Centrum Wiskunde & Informatica door Schrijver en Steenbeek [9] voor NS en ProRail ontwikkelde programma CADANS, waarmee NS toekomstige dienstregelingen opstelt en ProRail toekomstige uitbreidingen van de infrastructuur doorrekent.

Zoeken in de duale ruimte

Een alternatieve manier voor het vinden van een toegelaten dienstregeling is de volgende. Daartoe verlaten we het *modulo* 60 *model* en zien we de intervallen $I(u,v)$ als intervallen van gehele getallen. Als $g : E \to \mathbb{Z}$ een toegelaten spanning is (dat wil zeggen, voldoende aan (9) mod 60), dan bestaat er een functie $k : E \to \mathbb{Z}$ zo dat

$$g(u,v) \in I(u,v) + 60k(u,v) \quad \text{voor elke pijl } (u,v). \tag{11}$$

De modulo 60-voorwaarde wordt nu omzeild door er een geheel (nog vast te stellen) veelvoud van 60 bij op te tellen. Zoals we in de inzet al vermeldden, kan voor vaste k snel worden beslist of er een spanning $g : E \to \mathbb{Z}$ bestaat die aan (11) voldoet. We kunnen de zoekprocedure dan omdraaien: we zoeken naar een functie $k : E \to \mathbb{Z}$ zodanig dat er een spanning $g : E \to \mathbb{Z}$ bestaat die voldoet aan (11).

Nu geldt dat als $k, k' \in \mathbb{Z}^E$ en $k - k'$ tot T behoort (dat wil zeggen een spanning is), dan is voor elke oplossing g die hoort bij k, $g' := g + 60(k' - k)$ een oplossing die hoort bij k' (want g' is een spanning en g' voldoet aan (11) met k vervangen door k'). Dus uit iedere klasse in de qoutiëntruimte $\mathbb{Z}^E/T$ hoeven we slechts één k te onderzoeken. Dit betekent dat we dan in feite zoeken in de duale ruime S^* van $\mathbb{Z}$-lineaire functies $S \to \mathbb{Z}$ (doordat de functie $\mathbb{Z}^E \to S^*$ met $k \mapsto (x \mapsto \langle k, x \rangle)$ kern $S^\perp = T$ heeft, en dus $S^* \cong \mathbb{Z}^E/T$).

Nog steeds zijn er oneindig veel mogelijkheden om k te kiezen, maar de keuze kan beperkt worden tot een eindig aantal door een opspannende boom B te kiezen. We nemen daarbij aan dat het netwerk samenhangend is — dat is in Nederland inderdaad het geval.

Dan blijven er nog maar eindig veel mogelijkheden over voor de waarden van $k(e)$ als e niet in B zit. Dit kunnen we zien door de cykel C_e te beschouwen die e met B maakt (de fundamentele cykel). De intervallen langs de pijlen in C_e impliceren een onder- en een bovengrens voor $k(e)$. Zie Figuur 8.

Dan bevat elke klasse van $\mathbb{Z}^E/T$ precies één $k : E \to \mathbb{Z}$ zo dat $k(e) = 0$ op elke e in B. Dit volgt uit het feit dat voor elke $k \in \mathbb{Z}^E$ precies één spanning g bestaat zodat $g(e) = k(e)$ voor elke e in B.

Het aantal mogelijkheden voor $k(e)$ is afhankelijk van de *ruimte* die deze cykel toelaat. Soms is $k(e)$ hierdoor volledig vastgelegd, soms is er keuze uit slechts twee mogelijkheden, soms ook is het aantal mogelijkheden groter. Maar in ieder geval wordt het zoekproces flink gereduceerd.

Een optimale dienstregeling

Uiteraard wil men niet alleen een *toegelaten* dienstregeling hebben, maar zelfs een *optimale*. Maar de vraag daarbij is natuurlijk ook: wat is het criterium voor een optimale dienstregeling? Er zijn verschillende, soms tegenstrijdige, criteria voor de kwaliteit van een dienstregeling.

Gegeven een toegelaten dienstregeling, kunnen de vertrek- en aankomsttijden met lineaire programmering geschoven worden, mits de volgordes der treinen op de trajecten vast zijn. In wiskundige termen: als we k in (11) vasthouden, kunnen we in polynomiale tijd voor een willekeurige kostenfunctie op de spanning g (de verschillen van de tijden) een spanning van minimale totale kosten vinden. Dit wordt echter een stuk lastiger als we ook k vrij laten en willen optimaliseren — dan staan we weer voor een NP-volledig probleem.

Er kan bijvoorbeeld gestreefd worden naar een dienstregeling waarin de reistijden voor de reizigers zo kort mogelijk zijn. Dat wil zeggen: korte rijtijden, korte halteertijden, en korte overstaptijden. Maar een dergelijke dienstregeling is waarschijnlijk weinig robuust onder verstoringen: het dempend vermogen is beperkt. Daarom is een ander doel om een geschikt gekozen rijtijdspeling in de dienstregeling in te bouwen. Deze kan gebruikt worden om een eventuele vertraging er weer uit te rijden. Kroon e.a. [5] beschrijven een *stochastisch optimalisatiemodel* om de rijtijdspeling optimaal te verdelen over de dienstregeling. Daarnaast hebben Heidergott, Olsder en Van der Woude [4] modellen ontwikkeld op basis van $(\max, +)$-algebra waarmee de robuustheid van een dienstregeling doorgerekend kan worden.

Bovendien kan in een onverstoorde situatie de rijtijdspeling gebruikt worden om zo min mogelijk energie te gebruiken door zo véél mogelijk zonder tractie uit te rollen vóór iedere haltering. Met behulp van Pontryagins Maximum Principe [7] kan een snelheidsprofiel berekend worden waarbij onderweg zo min mogelijk energie gebruikt wordt. Het totale energieverbruik per rit is daarbij dalend in de geplande rijtijd: hoe langer de rijtijd, hoe meer er uitgerold kan worden. Het vinden van een geschikte rijtijd voor een rit is daardoor een trade-off tussen snelheid, robuustheid en energieverbruik.

Materieelomloop en dienstregelingsnetwerk

Als de dienstregeling is vastgesteld, zijn de bepaling van de materieelomloop en de personeelsinzet de volgende relevante logistieke problemen. We beperken ons hier in eerste instantie tot het eerste. We gaan hierbij uit van de volgende (inmiddels gewijzigde)

Figuur 9 Een koploper van type III.

Figuur 10 Een koploper van type IV.

treinnummer	2123	2127	2131	2135	2139	2143	2147	2151	2155	2159	2163	2167	2171	2175	2179	2183	2187	2191
Amsterdam V		6.48	7.55	8.56	9.56	10.56	11.56	12.56	13.56	14.56	15.56	16.56	17.56	18.56	19.56	20.56	21.56	22.56
Rotterdam A		7.55	8.58	9.58	10.58	11.58	12.58	13.58	14.58	15.58	16.58	17.58	18.58	19.58	20.58	21.58	22.58	23.58
Rotterdam V	7.00	8.01	9.02	10.03	11.02	12.03	13.02	14.02	15.02	16.00	17.01	18.01	19.02	20.02	21.02	22.02	23.02	
Roosendaal A	7.40	8.41	9.41	10.43	11.41	12.41	13.41	14.41	15.41	16.43	17.43	18.42	19.41	20.41	21.41	22.41	23.54	
Roosendaal V	7.43	8.43	9.43	10.45	11.43	12.43	13.43	14.43	15.43	16.45	17.45	18.44	19.43	20.43	21.43			
Vlissingen A	8.38	9.38	10.38	11.38	12.38	13.38	14.38	15.38	16.38	17.40	18.40	19.39	20.38	21.38	22.38			

treinnummer	2108	2112	2116	2120	2124	2128	2132	2136	2140	2144	2148	2152	2156	2160	2164	2168	2172	2176
Vlissingen V			5.30	6.54	7.56	8.56	9.56	10.56	11.56	12.56	13.56	14.56	15.56	16.56	17.56	18.56	19.55	
Roosendaal A			6.35	7.48	8.50	9.50	10.50	11.50	12.50	13.50	14.50	15.50	16.50	17.50	18.50	19.50	20.49	
Roosendaal V		5.29	6.43	7.52	8.53	9.53	10.53	11.53	12.53	13.53	14.53	15.53	16.53	17.53	18.53	19.53	20.52	21.53
Rotterdam A		6.28	7.26	8.32	9.32	10.32	11.32	12.32	13.32	14.32	15.32	16.32	17.33	18.32	19.32	20.32	21.30	22.32
Rotterdam V	5.31	6.29	7.32	8.35	9.34	10.34	11.34	12.34	13.35	14.35	15.34	16.34	17.35	18.34	19.34	20.35	21.32	22.34
Amsterdam A	6.39	7.38	8.38	9.40	10.38	11.38	12.38	13.38	14.38	15.38	16.40	17.38	18.38	19.38	20.38	21.38	22.38	23.38

Tabel 1 Dienstregeling Amsterdam – Vlissingen v.v.

dienstregeling van de treinen Amsterdam – Vlissingen v.v. Zie Tabel 1.

De treinen stoppen bij meer stations, maar Rotterdam en Roosendaal zijn de enige plaatsen waar onderweg materieel kan worden *afgetrapt* of *bijgeplaatst*. Dit kan ook aan de eindpunten Amsterdam en Vlissingen.

Ook gaan we, in eerste instantie, uit van één type twee-richtingtreinstellen, de zogenaamde koplopers type III (zie Figuur 9), die onderling gekoppeld kunnen worden tot langere treinen. Een treinstel dat afgetrapt is van een trein op een station, kan daarna weer bijgeplaatst worden aan een andere trein op dit station.

Per trein heeft NS voor elk van de deeltrajecten een schatting gemaakt van het aantal reizigers en dat vertaald in het minimale aantal gekoppelde treinstellen dat op dat deeltraject dient te rijden. Zie Tabel 2.

Met deze gegevens en condities is de vraag wat het minimale aantal koplopers is dat voor de dienst Amsterdam – Vlissingen moet worden ingezet.

Het dienstregelingsnetwerk

Een optimale materieelomloop kan worden bepaald door het probleem te modelleren als een stroomprobleem in het volgende virtuele netwerk, het *dienstregelingsnetwerk* [2, 13]. Figuur 11 geeft het dienstregelingsnetwerk weer voor de in Tabel 1 gegeven dienstregeling.

In het dienstregelingsnetwerk $N = (V, E)$ corresponderen de schuine pijlen met treinritten op een van de trajecten tussen Amsterdam, Rotterdam, Roosendaal en Vlissingen, terwijl de verticale pijlen corresponderen met een tijdsinterval op een van deze stations tussen twee opvolgende vertrek- of aankomsten. De pijlen lopen steeds met de tijd mee.

De punten van het netwerk zijn de eindpunten van deze treinritten — kruispunten van twee schuine pijlen 'onderweg' geven geen punten van het netwerk. De punten van het netwerk komen dan overeen met de 198 tijdstippen die in Tabel 1 vermeld staan. Daarnaast zijn twee extra punten toegevoegd, een beginpunt s en een eindpunt t.

Een materieelomloop is nu een stroom $f : E \to \mathbb{Z}_+$, waarbij $f(e)$ de volgende betekenis heeft. Als e correspondeert met een treinrit, dan geeft $f(e)$ het aantal gekoppelde koplopers aan dat voor deze treinrit wordt ingezet. Als e een tijdsinterval tussen twee opvolgende gebeurtenissen op station X voorstelt, dan is $f(e)$ het aantal koplopers dat op station X aanwezig is tijdens dit interval. In s en t hoeft de stroom niet aan de wet van stroombehoud te voldoen — dit kan desgewenst alsnog bewerkstelligd worden door s en t te identificeren.

Bij elke pijl e is de *vraagfunctie* $d(e)$ gezet: het minimum aantal koplopers zoals gegeven in Tabel 2. Als e niet met een treinrit correspondeert, dan is $d(e) = 0$.

Het aantal in te zetten koplopers is dan gelijk aan de minimale hoeveelheid stroom die uit s vertrekt en die op elke pijl aan de ondergrens voldoet. Klassieke netwerk- of lineaire programmeringsmethoden leveren binnen een fractie van een seconde een optimale materieelomloop op die aan alle ondergrenzen voldoet. Voor het netwerk in Figuur 11 blijkt 22 het minimale aantal in te zetten koplopers te zijn.

Deze methode kan worden aangepast en uitgebreid naar verschillende varianten van het probleem. Door de vier pijlen aankomend in t per station te koppelen aan de vier pijlen vertrekkend uit s kan bijvoorbeeld een volledig periodieke omloop worden bepaald. In plaats van het aantal koplopers te minimaliseren, kan men ook het aantal gemaakte *bakkilometers* minimaliseren, of een lineaire combinatie hiervan met het aantal koplopers. Bovendien kunnen bovengrenzen worden gesteld aan de stroom, als bovengrens aan de lengte van de trein of aan de opstelcapaciteit op een station. De methode blijft ook snel

treinnummer	2123	2127	2131	2135	2139	2143	2147	2151	2155	2159	2163	2167	2171	2175	2179	2183	2187	2191
Amsterdam – Rotterdam		3	4	3	3	2	2	2	2	3	4	4	4	2	2	1	2	1
Rotterdam – Roosendaal	1	2	3	3	2	2	2	2	2	4	5	4	3	2	2	1	1	
Roosendaal – Vlissingen	3	2	2	2	2	2	2	2	2	3	4	3	2	2	1			

treinnummer	2108	2112	2116	2120	2124	2128	2132	2136	2140	2144	2148	2152	2156	2160	2164	2168	2172	2176
Vlissingen – Roosendaal			1	3	3	3	2	2	2	2	2	2	3	2	2	1	1	
Roosendaal – Rotterdam		2	3	4	3	4	2	2	2	2	2	2	3	2	1	1	1	1
Rotterdam – Amsterdam	1	2	4	4	3	4	3	2	2	3	3	4	4	3	2	1	1	1

Tabel 2 Ondergrenzen aan het aantal koplopers per deeltraject.

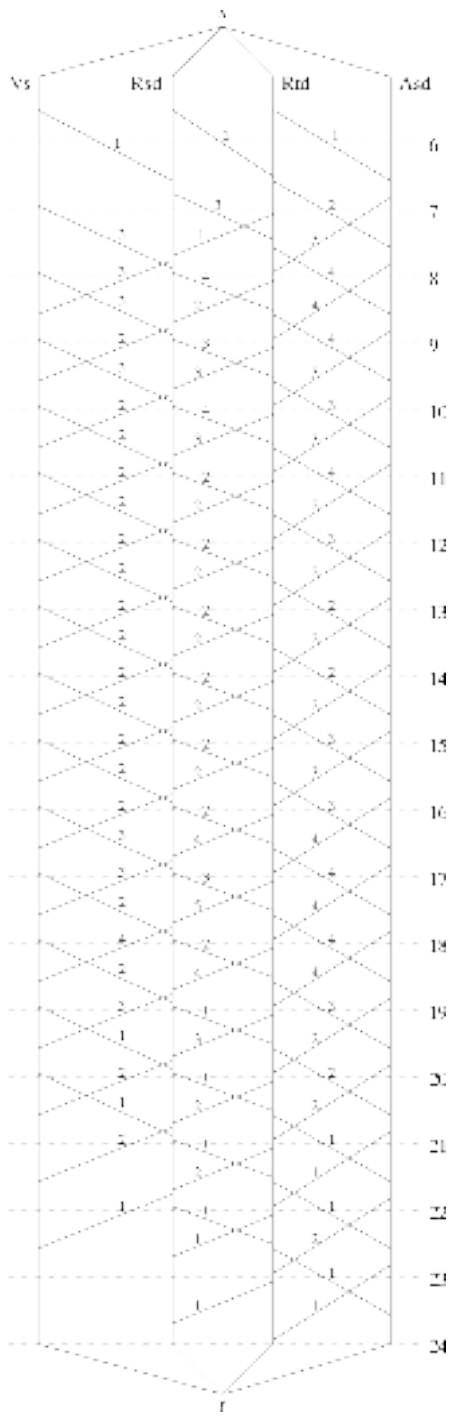
als de materieelomloop het hele Nederlandse intercity-netwerk omvat.

Koppelen van verschillende materieeltypes
Het probleem wordt echter een stuk lastiger als er treinstellen van meer dan één type materieel kunnen worden gekoppeld tot een trein. Als voorbeeld geven we weer de koplopers, waar naast type III in Figuur 9, elk bestaande uit 3 *bakken* (rijtuigen), ook type IV bestaat, elk bestaande uit 4 bakken, zie Figuur 10.

De *drietjes* en *viertjes* kunnen onderling gekoppeld worden. Met combinaties van drietjes en viertjes kan meer *op maat* worden gereden, maar het optimaliseren van de materieelomloop, en het herstellen van verstoringen in de materieelomloop wordt lastiger.

Het omloopprobleem vertaalt zich nu in een *multi-commodity* stroomprobleem op het dienstregelingsnetwerk, waarin ondergrenzen worden gesteld aan de *combinatie* van drietjes en viertjes: de drietjes vormen op zichzelf een stroom, evenals de viertjes, maar het totale aantal aangeboden zitplaatsen per trein wordt bepaald door een lineaire combinatie van de aantallen drietjes en viertjes in de trein. Dit maakt het probleem op zichzelf al lastiger — in feite NP-volledig. De constraint matrix is niet meer totaal unimodulair, en lineaire programmering levert daardoor niet meer direct een geheeltallige oplossing op.

En daar komt nog iets bij: de volgorde van de drietjes en viertjes in een trein is nu ook van belang. Als bijvoorbeeld een drietje in een station moet worden afgetrapt, dan dient dit zich aan het eind van de trein te bevinden. Bijplaatsen gebeurt steeds aan de kop van de trein. Het toegelaten zijn van een stroom wordt dus niet alleen bepaald door de aantallen drietjes en viertjes per trein, maar ook door het *woord* gevormd door de volgorde van drietjes en viertjes, zoals 343 of 334. Door toevoeging van binaire *compositievariabelen* die de toegelaten treincomposities beschrijven, tezamen met *transitievariabelen* voor de toegelaten overgangen, kunnen dergelijke problemen met een geheeltallig optimaliseringspakket zoals CPLEX binnen korte tijd worden opgelost (zie [8]). Hierbij helpt het dat de transitievariabelen vanzelf geheeltallig zijn als de compositievariabelen dat zijn.



Figuur 11 Het dienstregelingsnetwerk Amsterdam – Vlissingen (Vs = Vlissingen, Rsd = Roosendaal, Rtd = Rotterdam, Asd = Amsterdam).

En er is nog meer

Reisadvies
Het dienstregelingsnetwerk kan ook gebruikt worden voor het geven van reisadviezen aan reizigers, zoals door middel van de NS Reisplanner. Een reis van een vertrekstation (inclusief een vertrektijdstip) naar een bestemming kan gerepresenteerd worden als een gericht pad in een netwerk van dezelfde structuur als het eerder aangegeven dienstregelingsnetwerk, maar waarin nu *alle* halteringen van *alle* treinen zijn meegenomen. Een zo kort mogelijke reis is een kortste pad van vertrekstation en -tijd naar de bestemming.

Reisadvies

Gilze-Rijen → Buitenpost
Vertrek dinsdag 23 juni 2015 rond 09:00

Heenreis: Vertrek 09:32 Aankomst 13:14

Dinsdag 23 juni 2015

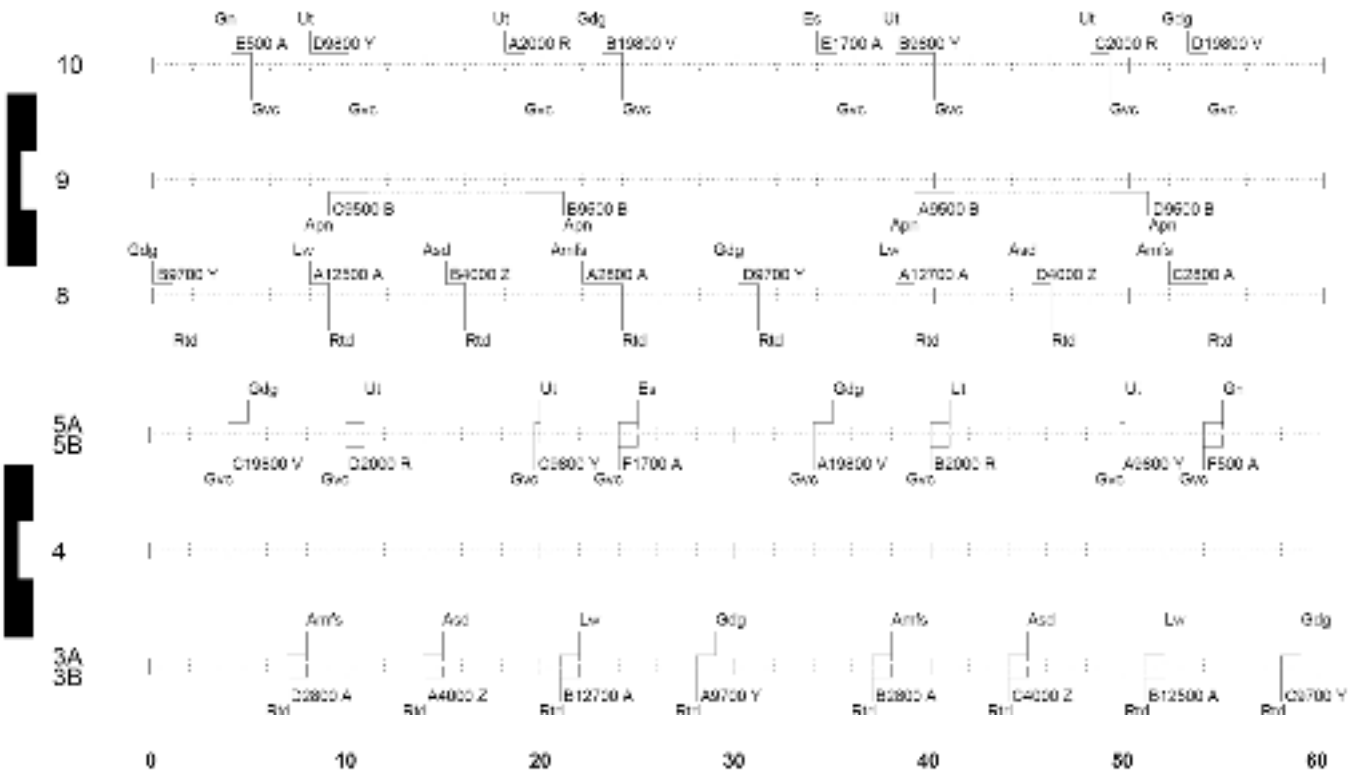
Dit reisadvies houdt geen rekening met wijzigingen in de treindienst op de dag zelf. Plan voor een actueel reisadvies uw reis vlak voor vertrek.

Tijd	Station / Halte	Spoor	Richting	Reisdetails
09:32	Gilze-Rijen	4	Utrecht Centraal	Sprinter (NS)
10:00	's-Hertogenbosch	3b		
10:09	's-Hertogenbosch	4	Schiphol	Intercity (NS)
10:37	Utrecht Centraal	7		
10:50	Utrecht Centraal	11	Groningen	Intercity (NS)
12:44	Groningen	4b		
12:51	Groningen	6a	Leeuwarden	Stoptrein (Arriva)

Reisdetails: Geen AVR-NS

13:14	Buitenpost	2		

Figuur 12 Reisadvies Gilze-Rijen – Buitenpost.



Figuur 13 Perrontoewijzing voor een deel van de sporen van station Gouda.

Wanneer de reiziger zo min mogelijk wil overstappen, dan moet het dienstregelingsnetwerk echter uitgebreid worden met extra punten en pijlen: de punten worden onderverdeeld in treinpunten en stationspunten. Treinpijlen verbinden achtereenvolgende treinpunten behorend bij dezelfde trein. Stationspijlen verbinden in de tijd opeenvolgende stationspunten. Treinpijlen representeren de verplaatsing van een reiziger in een trein van het beginpunt van de pijl (station en tijd) naar het eindpunt van de pijl (station en tijd). Stationspijlen representeren het wachten van een reiziger op een station tijdens een overstap. In- en uitstappijlen verbinden treinpunten en stationspunten met elkaar. Nu is een reis met zo weinig mogelijk overstaps te vinden als een pad van herkomst naar bestemming dat een minimaal aantal in- en uitstappijlen gebruikt.

De Reisplanner van NS is gebaseerd op het werk van Tulp [11]. De Reisplanner geeft meestal meerdere mogelijkheden om van herkomst naar bestemming te komen: de snelste mogelijkheid, en ook eventuele langere reismogelijkheden maar met minder overstaps.

Diensten voor machinisten en conducteurs

Net zoals een reis van een reiziger een pad is in het (uitgebreide) dienstregelingsnetwerk, is een dienst van een machinist of conducteur dat ook. Daarbij dient een dergelijk pad wel aan een aantal extra randvoorwaarden te voldoen: zo mag het pad niet te lang zijn (in tijd), het moet een pauze op een geschikt punt (in tijd en plaats) bevatten, en het moet weer terugkeren naar zijn beginpunt (in plaats). Daarnaast moet ook nog aan constraints op standplaatsniveau voldaan worden: zo mogen per standplaats het aantal diensten en de gemiddelde dienstlengte niet te groot zijn.

Het complete planningsprobleem voor de diensten van de machinisten op één dag kan dan beschreven worden als het vinden van een zo klein mogelijke set van toegelaten paden in het dienstregelingsnetwerk, die gezamenlijk alle treinpijlen in het dienstregelingsnetwerk overdekken en gezamenlijk aan de standplaatsconstraints voldoen. Het resulterende probleem is een *Set Covering*-probleem met extra randvoorwaarden. Aangezien het aantal toegelaten paden exponentieel is in het aantal ritten in de dienstregeling, worden dergelijke modellen meestal met dynamische kolomgeneratie opgelost, zie [1].

Routering in stations

Bij het zoeken naar een dienstregeling is tot nu toe buiten beschouwing gelaten of het mogelijk is de diverse treinen op een stationsemplacement zo te routeren dat er geen conflicten ontstaan.

Gegeven de vertrek- en aankomsttijden van de treinen op een station, kan het routeren van de treinen door dat station gebaseerd worden op lijsten per trein met alle mogelijke routes die een trein door dat station kan volgen. Hierbij is een route een combinatie van een fysieke route (een reeks sporen en wissels) en een tijdstip. Het tijdstip is zo gekozen, dat de aankomst- en vertrektijd van een trein langs het perron gelijk is aan de vastgestelde aankomst- en vertrektijd. Een route kan gezien worden als een gericht pad in een *time-expanded* versie van een emplacement zoals in Figuur 2 wordt weergegeven.

Voor het oplossen van dit routeringsprobleem kan gebruik gemaakt worden van het ongerichte *conflictennetwerk* $K = (V, E)$: iedere combinatie van een trein en een mogelijke route voor die trein wordt in het conflictennetwerk weergegeven door een enkel punt. Twee punten worden met elkaar verbonden als de bijbehorende trein-route combinaties met elkaar conflicteren. Hierbij conflicteren twee routes met elkaar als ze één of meerdere infrastructuur-elementen (sporen, wissels, et cetera) tegelijk willen bezetten of reserveren.

Het probleem is nu het selecteren van precies één punt per trein, zodanig dat géén van de geselecteerde punten met elkaar verbonden zijn, en zodanig dat het totale gewicht van de geselecteerde punten maximaal is. Hierbij geeft het gewicht van een punt aan hoe gewenst het is dat de bijbehorende route aan de bijbehorende trein wordt toegewezen. Dit is een zogenaamd *Node Packing*-probleem.

Als $z_{t,r}$ een binaire beslissingsvariabele is die weergeeft of route r wel of niet aan trein t wordt toegewezen, dan worden de constraints als volgt weergegeven:

$$z_{t,r} + z_{t',r'} \leq 1 \quad \forall \{(t,r),(t',r')\} \in E. \tag{12}$$

Voor het oplossen van dit probleem helpt het als niet slechts *paren* van verbonden punten verboden worden, maar als de *maximale cliques* in het conflictennetwerk gebruikt worden. Daarmee worden de constraints als volgt:

$$\sum_{(t,r)\in c} z_{t,r} \leq 1 \quad \forall c \subseteq K : c \text{ is een maximale clique.} \tag{13}$$

Als het doel is om het totale gewicht van de geselecteerde punten te maximaliseren, dan kan CPLEX instanties met een periode van één uur zelfs voor de grotere stations in korte tijd oplossen. Complexer wordt het als het doel is om de robuustheid van de routering te maximaliseren. Dan is het zaak om kruisende treinbewegingen zoveel mogelijk te vermijden, en de overblijvende kruisende treinbewegingen zoveel mogelijk in tijd te spreiden. Dit leidt tot een doelfunctie met kwadratische termen (die wel weer te lineariseren is). Figuur 13 geeft een perrontoewijzing voor een deel van de sporen van station Gouda.

Tot slot

In dit artikel hebben we slechts een topje van de ijsberg kunnen laten zien van de toepassingen van wiskundige netwerken ten behoeve van de optimalisatie van spoorwegsystemen. Een deel van de in het voorgaande beschreven modellen wordt inmiddels succesvol toegepast in de *planningsprocessen* bij NS en ProRail, zie [6].

De volgende uitdaging is om deze modellen ook geschikt te maken voor de *real-time bijsturing* bij vertragingen of verstoringen. De grootste uitdaging daarbij is de beschikbare rekentijd: een model dat een uur nodig heeft om uit te rekenen wat er over één minuut moet gebeuren is duidelijk niet bruikbaar in de praktijk. Daarnaast is er in de bijsturing inherent sprake van een grotere onzekerheid dan in de planning. Op dit gebied zijn inmiddels significante stappen vooruit gezet, maar er zijn ook nog veel problemen op te lossen.

Referenties

1 E.J.W. Abbink, M. Fischetti, L.G. Kroon, G. Timmer en M.J.C.M. Vromans, Reinventing Crew Scheduling at Netherlands Railways, *Interfaces* 35 (2005), 393–401.

2 T.E. Bartlett, An algorithm for the minimum number of transport units to maintain a fixed schedule, *Naval Research Logistics Quarterly* 4 (1957), 139–149.

3 L.R. Ford en D.R. Fulkerson, Maximal flow through a network, *Canadian Journal of Mathematics* 8 (1956), 399–404.

4 B. Heidergott, G.J. Olsder en J. van der Woude, *Modeling and Analysis of Synchronized Systems: A Course on Max-Plus Algebra and Its Applications*, Princeton University Press, Princeton, NJ, 2005.

5 L.G. Kroon, G. Maróti, M. Retel Helmrich, M.J.C.M. Vromans en R. Dekker, Stochastic Improvement of Cyclic Railway Timetables, *Transportation Research Part B* 42 (2008), 553–570.

6 L.G. Kroon, D. Huisman, E.J.W. Abbink, P.J. Fioole, M. Fischetti, G. Maróti, A. Schrijver, A. Steenbeek en R.J. Ybema, The New Dutch Timetable: The OR Revolution, *Interfaces* 39(1) (2009), 6–17.

7 L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze en E.F. Misiichenko, *The Mathematical Theory of Optimal Processes*, Interscience Publishers, New York, 1962.

8 A. Schrijver, Planning van materieelomlopen, Technical Report, Centrum Wiskunde & Informatica, Amsterdam, 2003, www.cwi.nl/~lex pvm.pdf.

9 A. Schrijver en A. Steenbeek, Spoorwegdienstregelingontwikkeling, Technical Report, Centrum Wiskunde & Informatica, Amsterdam, 1993, www.cwi.nl/~lex/sdo.pdf.

10 É. Tardos, A strongly polynomial minimum cost circulation algorithm, *Combinatorica* 5 (1985), 247–255.

11 E. Tulp, *Searching timetable networks*, proefschrift, Vrije Universiteit Amsterdam, 1991.

12 W. van der Aalst, K.M. van Hee en M. Voorhoeve, The DONS rail scheduling system, *ATPN Case Studies Tutorial*, 1994, pp. 1–12.

13 J.W.H.M.T.S.J. van Rees, Een studie omtrent de circulatie van materieel, *Spoor- en Tramwegen* 38 (1965), 363–367.

Leo van Iersel
Faculteit Elektrotechniek, Wiskunde en Informatica
TU Delft
l.j.j.v.iersel@gmail.com

Hoe zijn ze verwant?

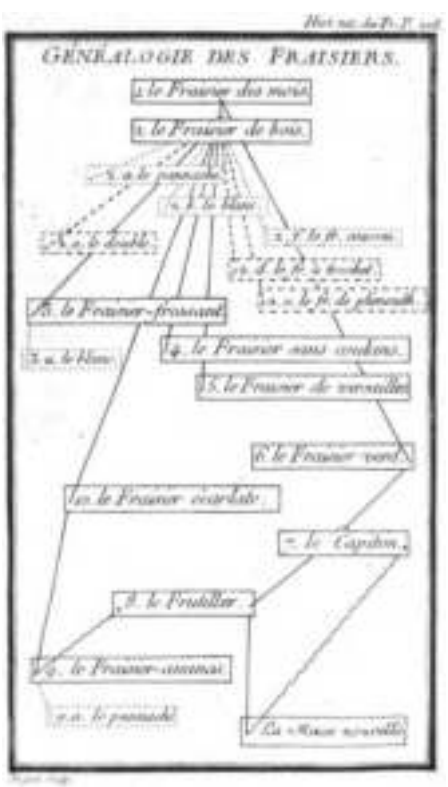
Hoe kunnen we reconstrueren hoe hedendaagse organismen zijn ontstaan uit verre voorouders door verschillende evolutionaire processen? Dit is het doel van onderzoek op het gebied van *fylogenetische netwerken*: grafen die evolutionaire verwantschappen beschrijven. Leo van Iersel ontving in 2011 een Veni-beurs van NWO voor onderzoek op dit gebied. Eind 2014 is hij begonnen als tenure track universitair docent aan de TU Delft. In dit artikel legt hij uit wat fylogenetische netwerken precies zijn en welke wiskundige problemen in dit vakgebied naar voren komen.

Darwins beroemde boek *The Origin of Species* (1859) bevatte precies één illustratie: een *fylogenetische boom*, oftewel een diagram dat schematisch weergeeft hoe soorten splitsen en zo nieuwe soorten vormen. Dit was niet de eerste keer dat zo'n diagram getekend werd: al voor de publicatie van *The Origin of Species* gaven biologen evolutionaire relaties weer in diagrammen. Interessant is dat die diagrammen lang niet altijd bomen waren.

In 1755 publiceerde Buffon al een diagram dat de evolutionaire relaties tussen verschillende hondenrassen weergaf [12]. Dit diagram was echter geen boom maar een netwerk. Hondenrassen splitsen namelijk niet alleen, maar nieuwe rassen kunnen ook gevormd worden uit combinaties van andere rassen. Een dergelijk diagram noemen we nu een *fylogenetisch netwerk*. De term 'fylogenetisch' bestaat uit de oud Griekse woorden *phulè* (volksstam) en *genesis* (wording). Een fylogenetisch netwerk beschrijft dus hoe groepen organismen (bijvoorbeeld stammen) zijn ontstaan uit andere groepen.

Deze netwerken zijn echter niet alleen relevant voor het bestuderen van verschillende stammen of rassen binnen één soort. Ook nieuwe soorten kunnen gevormd worden uit combinaties van andere soorten. Een mooi voorbeeld daarvan is de vorming van hybrides bij planten, zoals in het fylogenetische netwerk voor aardbeien in Figuur 1.

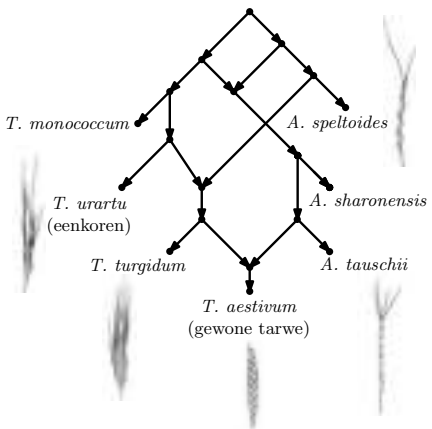
Toen Buffon honden bestudeerde, Duchenne aardbeien en Darwin vogels op de Galapagoseilanden, stelden ze allemaal dezelfde vraag: "Hoe zijn ze verwant?" Om dit uit te vinden moeten we achterhalen wat er miljoenen jaren geleden gebeurd is. Geen makkelijke taak. Nu, 260 jaar nadat Buffon zijn eerste fylogenetische netwerk publiceerde, zijn wiskundigen en informatici bezig om methodes te ontwikkelen om deze vraag systematisch te beantwoorden met behulp van DNA-data. Er is al veel bekend over het geval dat de verwantschappen beschreven kunnen worden door een fylogenetische boom. Het reconstrueren van fylogenetische netwerken is echter veel uitdagender. Niet alleen zijn er veel meer netwerken dan bomen, zelfs voor een gegeven netwerk is het vaak moeilijk om te bepalen hoe goed de data op dit netwerk passen. Pas recent zijn de eerste fylogenetische netwerken gepubliceerd die met behulp van computerprogramma's gegenereerd zijn, zoals het netwerk voor tarwe in Figuur 2.



Figuur 1 Een fylogenetisch netwerk getekend in 1766 door Antoine Nicolas Duchesne, dat de evolutionaire relaties beschrijft tussen verschillende aardbeiensoorten (in de ononderbroken rechthoeken) en -rassen [3, 14]. Het ras dat 'La Race nouvelle' wordt genoemd is door Duchesne op 17-jarige leeftijd ontdekt.

Wat is een fylogenetisch netwerk precies?

Er zijn verschillende definities van fylogenetische netwerken in omloop. Het belangrijkste onderscheid is dat tussen gerichte en ongerichte netwerken. Waar eerdere studies zich vooral concentreerden op ongerichte netwer-



Figuur 2 Een fylogenetisch netwerk dat laat zien hoe moderne tarwe is ontstaan uit een combinatie van verschillende oude tarwesoorten [13, 15]. De punten met twee inkomende pijlen beschrijven hybridisaties.

ken, wordt er steeds meer onderzoek gedaan over gerichte netwerken. De richtingen van de pijlen in zo'n netwerk geven de richting van evolutie aan. Ze geven dus een expliciete hypothese over de evolutionaire geschiedenis van de bestudeerde objecten. In dit artikel zal ik me beperken tot gerichte netwerken, die als volgt gedefiniëerd kunnen worden.

Definitie. Een *fylogenetisch netwerk* voor een verzameling X is een gerichte graaf met de volgende eigenschappen:

1. er zijn geen gerichte circuits;
2. er zijn geen punten met één inkomende en één uitgaande pijl (*overbodige* punten);
3. er is één punt zonder inkomende pijlen (de *wortel*);
4. de punten zonder uitgaande pijlen (de *bladeren*) zijn elk gelabeld met een element van X;
5. elk element van X is het label van één blad.

Een fylogenetisch netwerk is *binair* als elk punt hoogstens twee inkomende en hoogstens twee uitgaande pijlen heeft en elk punt met twee inkomende pijlen precies één uitgaande pijl heeft. Het netwerk voor tarwe uit Figuur 2 is een voorbeeld van een binair fylogenetisch netwerk.

In toepassingen in de biologie bevat de verzameling X een aantal namen van hedendaagse soorten, rassen of stammen. Voor het gemak gaan we uit van soorten. Elk blad stelt dus een hedendaagse soort voor. Een punt met meerdere uitgaande pijlen stelt een splitsing van een soort in twee of meer nieuwe soorten voor. Een punt met meer dan één inkomende pijl stelt voor dat een nieuwe soort ontstaan is uit een combinatie van eerdere soorten. Veelvoorkomende voorbeelden hiervan zijn de vorming van hybrides bij planten, reassortment bij virussen en genoverdracht bij bacteriën. Zo'n punt met minstens twee inkomende pijlen wordt een *reticulatie* genoemd.

Fylogenetische netwerken zijn een generalisatie van *fylogenetische bomen*, die we nu eenvoudig kunnen definiëren als fylogenetische netwerken zonder reticulaties. Fylogenetische bomen beschrijven net als fylogenetische netwerken evolutionaire relaties. Bomen zijn daarin echter veel beperkter dan netwerken. In een boom stelt elk intern punt een splitsing van een soort in twee of meer soorten voor. Het ontstaan van een soort uit een combinatie van eerdere soorten kan dus niet door een boom beschreven worden.

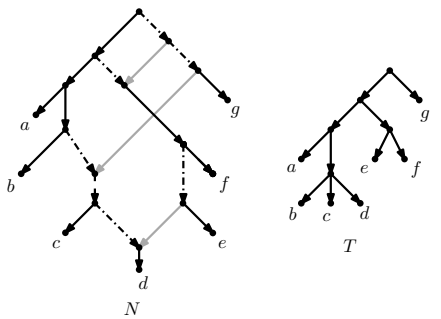
Bomen combineren tot een netwerk

Stel dat de evolutie van een verzameling X van soorten het best beschreven kan worden door een fylogenetisch netwerk N. Het DNA van deze soorten bestaat uit verschillende genen. Voor het gemak ga ik ervan uit dat genen in hun geheel worden overgeërfd. Als dat niet het geval is dan kunnen kleinere stukjes DNA beschouwd worden die wel in hun geheel worden overgeërfd.

In een punt van N met twee inkomende pijlen wordt dus een deel van de genen via de ene pijl geërfd en de rest van de genen via de andere pijl. De evolutie van een enkel gen kan daardoor beschreven worden door een boom, sterker nog, een boom die in het netwerk zit. Wat ik bedoel met 'in het netwerk' wordt geformaliseerd door de onderstaande definitie en geïllustreeerd in Figuur 3.

Definitie. Laat T een fylogenetische boom voor X zijn en N een fylogenetisch netwerk voor X. Dan is T *bevat* in N als we T kunnen verkrijgen uit N door het verwijderen van punten en pijlen en het samentrekken van pijlen.

Het *samentrekken* van een pijl van u naar v betekent dat je de pijlen die uit v vertrekken nu uit punt u laat vertrekken, de punten die in v aankomen nu in punt u laat aankomen, en vervolgens het punt v verwijdert. Waarom kan het nodig zijn om pijlen samen te trekken? Ten eerste, bij het verwijderen van punten en pijlen ontstaan er overbodige punten met één inkomende en één uitgaande pijl. Het *onderdrukken* van een overbodig punt betekent dat de inkomende pijl samengetrokken wordt. Alle overbodige punten moeten onderdrukt worden omdat deze niet toegestaan zijn in een fylogenetisch netwerk, en dus ook niet in een fylogenetische boom.



Figuur 3 Het fylogenetische netwerk N voor tarwe, met de bladeren voor het gemak herlabeld tot a,b,c,d,e,f en g en een fylogenetische boom T die bevat is in N. Boom T is bevat in N omdat je T uit N kunt verkrijgen door de grijze pijlen te verwijderen en alle onderbroken pijlen samen te trekken.

Er is echter nog een belangrijke reden om pijlen samen te trekken, wanneer de boom T niet binair is. Niet-binaire bomen worden in de praktijk gebruikt om onzekerheid uit te drukken. Bijvoorbeeld boom T in Figuur 3 geeft aan dat het onduidelijk is in welke volgorde b, c en d zijn afgesplitst van hun gemeenschappelijke voorouder. Dus in een netwerk dat deze boom bevat kunnen a, b en c in een willekeurige volgorde afsplitsen. Om T dan uit een deelgraaf van N te verkrijgen moeten pijlen samengetrokken worden.

Voor een gegeven netwerk en een gegeven boom is het trouwens al NP-moeilijk om te beslissen of het netwerk de boom bevat.

Een veel bestudeerde aanpak voor het construeren van fylogenetische netwerken uit DNA is de volgende tweestapsmethode. In de eerste stap wordt voor elk gen een fylogenetische boom gegenereerd. Voor deze stap zijn goede en snelle methodes beschikbaar. De tweede stap is om de verkregen fylogenetische bomen te combineren tot een fylogenetisch netwerk. Het doel is om een zo simpel mogelijk netwerk te vinden dat alle bomen bevat. Dit kunnen we als volgt formaliseren. Voor het gemak beperken we ons tot binaire netwerken.

Probleem: Minimum Reticulatie (MinRet).
Gegeven: een verzameling $\mathcal{T}$ van fylogenetische bomen, elk voor dezelfde verzameling X van labels.
Vind: een binair fylogenetisch netwerk N voor X dat elke boom in $\mathcal{T}$ bevat en een minimum aantal reticulaties heeft.

MinRet is een NP-moeilijk probleem, zelfs als $\mathcal{T}$ slechts twee binaire bomen bevat. Voor dit speciale geval is er echter een elegante karakterisering van het probleem, die gebruikt kan worden om het probleem relatief snel op te lossen.

Bos van overeenstemming

Stel $\mathcal{T}$ bestaat uit twee binaire bomen T_1 en T_2. Het idee is om pijlen uit T_1 en T_2 te verwijderen zodanig dat beide bomen in hetzelfde bos veranderen. Zo'n bos heet een 'bos van overeenstemming' omdat elke component van het bos in zekere zin consistent is met beide bomen. Beide bomen zijn het dus 'eens' over de evolutie van de soorten in één component van het bos.

Als T_1 en T_2 twee binaire bomen voor X zijn, dan is een *bos van overeenstemming* voor T_1 en T_2 een bos dat uit elk van T_1 en T_2 verkregen kan worden door een deel van de pijlen en ongelabelde punten te verwijderen en overbodige punten te onderdrukken.

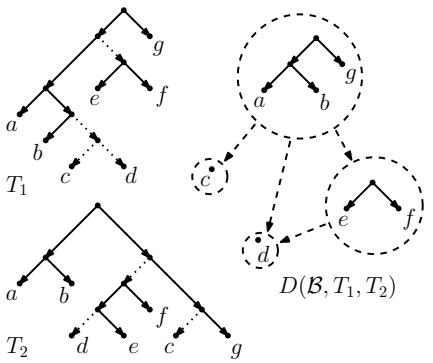
Stel nu dat N een netwerk is dat T_1 en T_2 bevat. Stel dat we voor elk punt in N met twee inkomende pijlen beide pijlen verwijderen, en vervolgens alle ongelabelde bladeren verwijderen en overbodige punten onderdrukken totdat er geen ongelabelde bladeren of overbodige punten meer zijn. Dan wordt N veranderd in een bos. Sterker nog, we verkrijgen een bos van overeenstemming voor T_1 en T_2.

Het is nu verleidelijk om te denken dat we uit elk bos van overeenstemming voor T_1 en T_2 ook een netwerk kunnen maken dat beide bomen bevat. Dit is echter alleen mogelijk als het bos aan de volgende acycliciteitsvoorwaarde voldoet. We definiëren een gerichte graaf $D(\mathcal{B}, T_1, T_2)$ die beschrijft hoe de componenten van $\mathcal{B}$ zich verhouden in de bomen T_1 en T_2. Deze gerichte graaf $D(\mathcal{B}, T_1, T_2)$ heeft een punt voor elke component van $\mathcal{B}$ en een pijl van (het punt voor) een component C_1 naar (het punt voor) een component C_2 als er een gericht pad is van de wortel van C_1 naar de wortel van C_2 in tenminste één van T_1 en T_2.

Definitie. Een bos van overeenstemming $\mathcal{B}$ voor T_1 en T_2 is *acyclisch* als de gerichte graaf $D(\mathcal{B}, T_1, T_2)$ acyclisch is (dat wil zeggen geen gerichte circuits bevat).

Stelling (Baroni e.a. [2]). *Als T_1 en T_2 twee binaire bomen voor X zijn, dan bestaat er een binair netwerk met k reticulaties dat T_1 en T_2 bevat dan en slechts dan als T_1 en T_2 een acyclisch bos van overeenstemming hebben met $k+1$ componenten.*

Het oplossen van MinRet is dus equivalent aan het vinden van een acyclisch bos van overeenstemming met zo min mogelijk componenten. Zo'n bos noemen we een *acyclisch bos van maximum overeenstemming*.



Figuur 4 Deze twee fylogenetische bomen T_1 en T_2 hebben een bos van overeenstemming $\mathcal{B}$ met vier componenten. Dit zijn de componenten die je krijgt als je de gestippelde lijnen uit T_1 en T_2 verwijdert en vervolgens overbodige punten onderdrukt. De gerichte graaf $D(\mathcal{B},T_1,T_2)$ geeft aan hoe de componenten van $\mathcal{B}$ zich verhouden in T_1 en T_2. Omdat $D(\mathcal{B},T_1,T_2)$ geen gerichte circuits bevat is $\mathcal{B}$ een *acyclisch* bos van overeenstemming. Volgens de stelling van Baroni e.a. is er dus een netwerk met drie reticulaties dat T_1 en T_2 bevat (namelijk het netwerk N uit Figuur 3).

De acycliciteit blijkt het belangrijkste obstakel te zijn voor het vinden van een efficiënt benaderingsalgoritme voor MinRet. Dit probleem blijkt namelijk net zo moeilijk te benaderen als het probleem Directed Feedback Vertex Set (DFVS): maak een gerichte graaf acyclisch door zo min mogelijk punten te verwijderen.

Stelling [11]. *Er bestaat een constante-factor-benaderingsalgoritme voor* MinRet *beperkt tot twee binaire bomen dan en slechts dan als een dergelijk algoritme bestaat voor* DFVS.

Een algoritme is een *constante-factor benaderingsalgoritme* als er een constante c bestaat zodanig dat het algoritme in polynomiale tijd een oplossing vindt die maximaal c maal slechter is dan een optimale oplossing. Of voor MinRet en DFVS een dergelijk algoritme bestaat is een belangrijk open probleem. DFVS was één van de eerste 21 problemen waarvan is bewezen dat ze NP-volledig zijn, door Richard Karp in een beroemd artikel uit 1971 [10], maar nog steeds is het niet bekend of er een constante-factor-benaderingsalgoritme voor bestaat.

Gelukkig is DFVS in de praktijk goed op te lossen met behulp van geheeltallig programmeren. In combinatie met een benaderingsalgoritme voor het vinden van een bos van maximum overeenstemming geeft dit een praktisch benaderingsalgoritme voor MinRet beperkt tot twee binaire bomen, wat zelfs uitgebreid kan worden voor niet-binaire bomen [7].

Door de bomen het bos niet meer zien

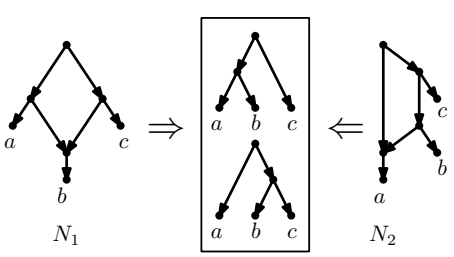
We hebben gezien dat voor het oplossen van MinRet de relatie met acyclische bossen van overeenstemming van groot belang is. Helaas bestaat deze relatie, beschreven in de stelling van Baroni e.a., alleen voor instanties van twee bomen. Voor instanties met drie of meer bomen geldt de relatie nog maar één kant uit. Het aantal componenten in een acyclisch bos van maximum overeenstemming geeft alleen een ondergrens op het aantal reticulaties in een optimaal netwerk. Zelfs voor instanties bestaande uit drie binaire bomen wordt MinRet erg uitdagend. Het theoretisch snelste algoritme voor dit geval heeft looptijd $1609891840^k p(n)$, met k het aantal reticulaties in een optimaal netwerk en $p(n)$ een polynoom in het aantal bladeren n [8]. Voor algemene instanties, waarin een willekeurig aantal niet-binaire bomen is toegestaan, is het niet bekend of er een algoritme met looptijd $f(k)\cdot p(n)$ bestaat met f een functie van k en p een polynoom in n.

Welke informatie is nodig?

Als we willen reconstrueren hoe de evolutie van een groep soorten er precies uit heeft gezien, dan maken we een fylogenetisch netwerk. Maar hoe weten we zeker dat we het juiste netwerk hebben? In welke gevallen wordt een netwerk uniek bepaald door de data? Als de data, zoals hierboven, bestaan uit fylogenetische bomen, dan kan het zijn dat het netwerk uniek bepaald is maar in het simpele voorbeeld in Figuur 5 is dat bijvoorbeeld niet zo.

Voor fylogenetische bomen is er veel onderzoek gedaan naar dit soort vraagstukken. Een fylogenetische boom wordt bijvoorbeeld uniek bepaald door de verzameling *triplets* die het bevat. *Triplets* zijn fylogenetische bomen met elk drie bladeren. Bovendien is er een polynomiale-tijd-algoritme (Aho e.a. [1]) om, gegeven een willekeurige verzameling triplets, te bepalen of er een fylogenetische boom bestaat die deze triplets bevat.

Dit wordt gebruikt voor zogenaamde 'superboom'-methodes. Stel dat voor een aantal verschillende deelverzamelingen van X een fylogenetische boom bekend is. Kunnen de-



Figuur 5 Twee fylogenetische netwerken N_1 en N_2 voor $\{a,b,c\}$ die allebei precies dezelfde verzameling bomen bevatten.

ze fylogenetische bomen dan samengevoegd worden tot een fylogenetische 'superboom' voor X die elke gegeven boom bevat? Deze vraag kunnen we nu gemakkelijk beantwoorden. Eerst vinden we voor elke invoerboom de verzameling triplets die het bevat. Daarna bepalen we of er een fylogenetische boom bestaat die de vereniging van de verkregen triplet verzamelingen bevat, met het algoritme van Aho e.a. Voor het geval dat er geen boom bestaat die *alle* triplets bevat zijn er tal van heuristieken ontwikkeld die toch een redelijke superboom genereren.

Maar waardoor wordt een fylogenetisch netwerk uniek bepaald? En kunnen we 'supernetwerk'-methodes ontwikkelen?

Elk fylogenetisch netwerk voor X induceert, voor elke deelverzameling X' van X, een fylogenetisch netwerk voor X', volgens de volgende definitie. Laat $LSV(X')$ (Laatste Stabiele Voorouder) het laatste punt zijn dat ligt op alle gerichte paden van de wortel van het netwerk naar een blad met label in X'.

Definitie. Gegeven een fylogenetisch netwerk N voor X en een deelverzameling $X' \subsetneq X$, wordt het *deelnet* $N|X'$ verkregen door

1. alle punten en pijlen te nemen die op een gericht pad liggen van $LSV(X')$ naar een blad met label in X';
2. alle overbodige punten te onderdrukken en parallelle pijlen te vervangen door enkele pijlen totdat er geen overbodige punten of parallelle pijlen meer zijn.

De verzameling deelnetten geïnduceerd door een netwerk N voor X is nu gedefinieerd als

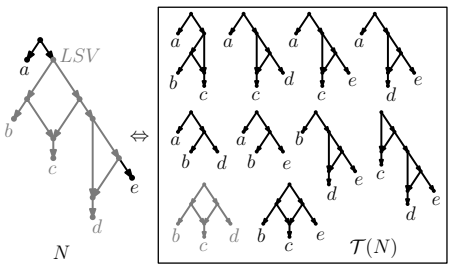
$$\mathcal{S}(N) = \{N|X' : X' \subsetneq X\}.$$

Net zo als triplets fylogenetische bomen zijn met drie bladeren, kunnen we *trinetten* definiëren als fylogenetische netwerken met drie bladeren. De verzameling trinetten geïnduceerd door een fylogenetisch netwerk N wordt dan gedefinieerd als

$$\mathcal{T}(N) = \{N|X' : X' \subsetneq X, |X'| = 3\}.$$

Helaas blijkt dat fylogenetische netwerken in het algemeen niet uniek bepaald worden door de verzameling trinetten die ze induceren, en ook niet door de hele verzameling geïnduceerde deelnetten [6].

De tegenvoorbeelden zijn echter erg complex: het aantal reticulaties groeit exponentieel in het aantal bladeren. Redelijk simpele netwerken worden wel uniek bepaald



Figuur 6 Een fylogenetisch netwerk N en de verzameling $\mathcal{T}(N)$ van alle trinetten die het bevat. Laat bijvoorbeeld $X'=\{b,c,d\}$. Dan is het punt LSV het laatste punt dat ligt op alle gerichte paden van de wortel van N naar een blad met label in $\{b,c,d\}$. Alle punten en pijlen die op een gericht pad liggen van LSV naar een blad met label in $\{b,c,d\}$ zijn grijs gekleurd. Als we in deze grijze deelgraaf nu alle overbodige punten onderdrukken en parallelle pijlen vervangen door enkele pijlen totdat er geen overbodige punten of parallelle pijlen meer zijn, dan verkrijgen we het grijze deelnet $N|\{b,c,d\} \in \mathcal{T}(N)$. Het blijkt dat in dit geval N uniek bepaald wordt door $\mathcal{T}(N)$.

door $\mathcal{T}(N)$ [9]. Dit is bijvoorbeeld het geval voor netwerken met maximaal twee reticulaties per 2-samenhangende deelgraaf (*2-samenhangend* betekent dat de deelgraaf samenhangend blijft als je een willekeurige pijl verwijdert). Het geldt ook voor netwerken waarin elk punt dat geen blad is tenminste één uitgaande pijl heeft naar een punt dat geen reticulatie is. Bijvoorbeeld de verzameling $\mathcal{T}(N)$ van trinetten in Figuur 6 bepalen het netwerk N in de figuur, want N voldoet aan beide voorwaarden. Het is echter onbekend waar precies de grens ligt tussen netwerken die wel en niet uniek bepaald worden door hun verzameling geïnduceerde trinetten.

Dit soort vraagstukken zijn van groot belang voor toepassingen in de biologie. Een belangrijke eis aan een methode voor het maken van fylogenetische netwerken (of bomen) is dat de methode *consistent* is, dat wil zeggen dat de methode het juiste netwerk produceert indien het volledige en foutloze data krijgt aangeboden. In deze zin kan een methode die fylogenetische bomen combineert tot een fylogenetisch netwerk nooit consistent zijn. Hetzelfde geldt voor een 'supernetwerk' methode die trinetten of deelnetwerken combineert tot een volledig netwerk. Maar als we ons beperken tot de genoemde klassen van netwerken die wel uniek bepaald worden door hun trinetten, dan is het wel mogelijk om consistente supernetwerk-methodes te ontwikkelen.

Kunnen we het DNA rechtstreeks gebruiken?

Voor het maken van fylogenetische bomen zijn tal van methodes ontwikkeld. Tegenwoordig is de belangrijkste bron van data natuurlijk DNA. Er zijn methodes die in twee stappen met het DNA werken. In een eerste stap kan bijvoorbeeld een afstand tussen elk tweetal soorten berekend worden, of een triplet voor elk drietal soorten, op basis van het DNA. Een tweede stap is dan om een boom te vinden die zo goed mogelijk voldoet aan de afstanden of triplets. De meest nauwkeurige methodes werken echter rechtstreeks met het DNA. Ze zoeken een boom die een zekere score maximaliseert. De score is een functie van de boom en de DNA-sequenties, en kan bijvoorbeeld gebaseerd zijn op een waarschijnlijkheidsberekening of op het *parsimony*-principe.

Het idee van *parsimony* (gierigheid) is om een fylogenetische boom te vinden waarop het gegeven DNA met zo min mogelijk mutaties geëvolueerd zou kunnen zijn. Het grote voordeel van parsimony is dat de parsimonyscore van een fylogenetische boom en gegeven DNA-sequenties in polynomiale tijd berekend kan worden. Merk eerst op dat de posities in de DNA-sequenties onderling onafhankelijk zijn en dat je dus elke positie apart kunt bekijken. Voor elke positie moet je dan het volgende probleem oplossen. Hierin is $\mathcal{P}$ de verzameling mogelijke letters in de sequenties. Dus bijvoorbeeld voor DNA is $\mathcal{P} = \{A, C, G, T\}$.

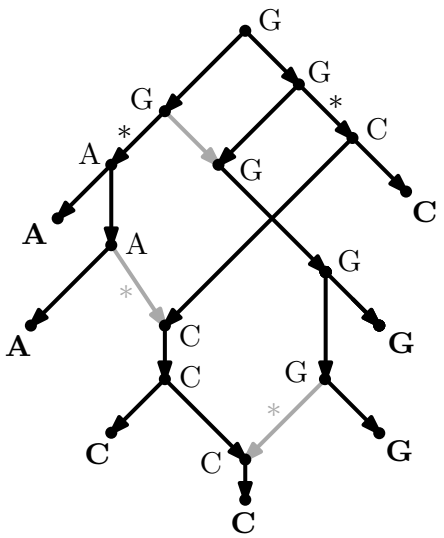
Probleem: Parsimony voor Bomen.
Gegeven: een fylogenetische boom en een label $\ell(x) \in \mathcal{P}$ voor elk blad x.
Vind: een label $\ell(v) \in \mathcal{P}$ voor elk intern punt v zodanig dat het aantal pijlen (u, w) met $\ell(u) \neq \ell(w)$ minimaal is.

Dit minimum aantal wordt de *parsimonyscore* genoemd. Dit probleem kan in polynomiale tijd opgelost worden door middel van dynamisch programmeren [5].

Voor fylogenetische netwerken zijn er twee interessante generalisaties van deze score. De eerste generalisatie is wiskundig gezien misschien de meest logische.

Probleem: Netwerk Parsimony.
Gegeven: een fylogenetisch netwerk en een label $\ell(x) \in \mathcal{P}$ voor elk blad x.
Vind: een label $\ell(v) \in \mathcal{P}$ voor elk intern punt v zodanig dat het aantal pijlen (u, w) met $\ell(u) \neq \ell(w)$ minimaal is.

Als $|\mathcal{P}| = 2$ dan is dit probleem direct gerelateerd aan het bekende MinCut-probleem, waarin een ongerichte graaf gegeven is met twee speciale punten s en t en gevraagd wordt om zo min mogelijk lijnen uit de graaf te verwijderen zodat er geen pad tussen s en t meer bestaat. Hoe is dit gerelateerd aan Netwerk Parsimony? Stel dat een netwerk is ge-



Figuur 7 Stel het fylogenetische netwerk voor tarwe is gegeven samen met de (vetgedrukte) labels van de bladeren. De figuur geeft één mogelijke labelling van de interne punten. Voor deze labelling zien we dat er in totaal vier pijlen zijn waar het label verandert, gemarkeerd met een $*$. De optimale waarde van Netwerk Parsimony is dus maximaal 4. Als we echter naar de boom kijken die bestaat uit de zwarte pijlen, dan zijn er maar twee pijlen waar het label verandert (de grijze $*$'en). De optimale waarde van Boom-in-netwerk Parsimony is dus maximaal 2.

geven met een label uit $\mathcal{P}$ voor elk blad. Neem voor het gemak aan dat $\mathcal{P} = \{S, T\}$. Stel nu dat we alle bladeren met label S samenvoegen tot een enkel punt dat we s noemen, en alle bladeren met label T samenvoegen tot een enkel punt dat we t noemen. Dan komt het oplossen van MinCut in de verkregen graaf op hetzelfde neer als Netwerk Parsimony oplossen in het oorspronkelijke netwerk [4].

Voor $|\mathcal{P}| = 2$ is dit probleem dus in polynomiale tijd op te lossen door middel van het bekende Edmonds–Karp-algoritme voor MaxFlow. Voor $|\mathcal{P}| > 2$ wordt het probleem NP-moeilijk maar is het nog wel goed benaderbaar.

Vanuit biologisch perspectief is de volgende generalisatie echter logischer.

Probleem: Boom-in-netwerk Parsimony.
Gegeven: een fylogenetisch netwerk N voor X en een label $\ell(x) \in \mathcal{P}$ voor elk blad x.
Vind: een fylogenetische boom voor X die bevat is in N en een zo klein mogelijke parsimony score heeft.

Waarom is Boom-in-netwerk Parsimony biologisch gezien een logischere generalisatie? Kijk wat er bij de reticulaties gebeurd met de sequenties. Hier worden de sequenties van twee 'voorouders' gecombineerd tot de sequentie van een 'hybride'. Kijk je echter naar een enkele positie van het DNA, dan wordt deze van een van de twee voorouders geërfd. De evolutie van een enkele DNApositie kan dus altijd beschreven worden door een boom — een boom die bevat is in het netwerk. Helaas is dit Boom-in-netwerk Parsimony probleem veel moeilijker op te lossen of te benaderen dan Netwerk Parsimony [4]. Alleen als we ons beperken tot netwerken met weinig reticulaties per 2-samenhangende deelgraaf dan kunnen we deze score snel berekenen.

Tot slot

Het reconstrueren van fylogenetische netwerken blijkt veel moeilijker te zijn dan het maken van fylogenetische bomen. Netwerken zijn tenslotte veel complexer dan bomen. Toch kunnen we in veel gevallen ver komen zolang we ons beperken tot relatief simpele netwerken. Dit is nuttig in de biologie omdat veel praktische fylogenetische netwerken vrij simpel zijn, zoals bijvoorbeeld de netwerken voor aardbeien en tarwe uit het begin van dit artikel. Toch komen er ook veel complexere fylogenetische netwerken voor in de biologie, bijvoorbeeld voor bacteriën. Zulke netwerken zullen we waarschijnlijk nooit tot in alle details kunnen reconstrueren. Dit is echter ook niet altijd noodzakelijk voor biologen. Belangrijker is het om het netwerk in hoofdlijnen te kunnen schetsen, zodat biologen er belangrijke conclusies uit kunnen trekken over de relaties tussen verschillende soorten.

Referenties

1 A.V. Aho, Y. Sagiv, T.G. Szymanski en J.D. Ullman, Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions, *SIAM Journal on Computing* 100(3) (1981), 405–421.

2 M. Baroni, S. Grünewald, V. Moulton en C. Semple, Bounding the number of hybridisation events for a consistent evolutionary history, *Journal of Mathematical Biology* 51(2) (2005), 171–182.

3 A.N. Duchesne, *Histoire naturelle des fraisiers, contenant Les vues d'Économie réunies à la Botanique; et suivie de Remarques Particulières sur plusieurs points qui ont rapport à l'Histoire naturelle générale*, Didot le jeune & C.J. Panckoucke, Parijs, 1766.

4 M. Fischer, L.J.J. van Iersel, S.M. Kelk en C. Scornavacca, On Computing the Maximum Parsimony Score of a Phylogenetic Network, *SIAM Journal on Discrete Mathematics* 29(1) (2015), 559–585.

5 W. Fitch, Toward defining the course of evolution: Minimum change for a specific tree topology, *Systematic Zoology* 20 (1971), 406–416.

6 K.T. Huber, L.J.J. van Iersel, V. Moulton en T. Wu, How Much Information is Needed to Infer Reticulate Evolutionary Histories? *Systematic Biology* 64(1) (2015), 102–111.

7 L.J.J. van Iersel, S.M. Kelk, N. Lekić en C. Scornavacca, A practical approximation algorithm for solving massive instances of hybridization number for binary and nonbinary trees, *BMC Bioinformatics* 15 (2014), 127.

8 L.J.J. van Iersel, S.M. Kelk, N. Lekić, C. Whidden en N. Zeh, Hybridization Number on Three Trees (2014), arXiv:1402.2136 [cs.DS].

9 L.J.J. van Iersel en V. Moulton, Trinets encode tree-child and level-2 phylogenetic networks, *Journal of Mathematical Biology* 68(7) (2014), 1707–1729.

10 R.M. Karp, Reducibility Among Combinatorial Problems, in R.E. Miller en J.W. Thatcher (eds.), *Complexity of Computer Computations*, Plenum, New York, 1972, pp. 85–103.

11 S.M. Kelk, L.J.J. van Iersel, N. Lekić, S. Linz, C. Scornavacca en L. Stougie, Cycle killer... qu'est-ce que c'est? On the comparative approximability of hybridization number and directed feedback vertex set, *SIAM Journal on Discrete Mathematics* 26(4) (2012), 1635–1656.

12 G.-L. Leclerc, graaf de Buffon, *Histoire naturelle générale et particulière*, Vol. 5, Imprimerie Royale, Parijs, 1755, pp. 228–229.

13 T. Marcussen, S.R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, International Wheat Genome Sequencing Consortium, K.S. Jakobsen, B.B. Wulff, B. Steuernagel, K.F. Mayer en O.A. Olsen, Ancient hybridizations among the ancestral genomes of bread wheat, *Science* 345 (2014), 1250092.

14 D. Morrison, The second phylogenetic network (1766), phylonetworks.blogspot.nl/2012/04/second-phylogenetic-network-1766.html.

15 D. Morrison, Complex hybridizations in wheat, phylonetworks.blogspot.nl/2015/01/complex-hybridizations-in-wheat.html.

Mark de Berg
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
mdberg@win.tue.nl

Conflictvrije kleuringen voor frequentietoekenning in draadloze netwerken

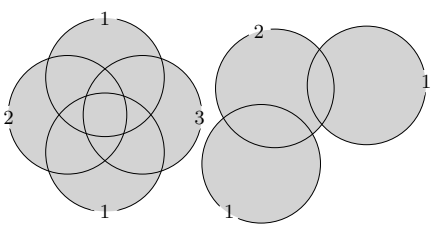
Een kleuring van een verzameling van n schijven in het platte vlak wordt *conflictvrij* genoemd als het volgende geldt voor elk punt q dat in een of meer schijven ligt: ten minste één van de schijven die q bevatten heeft een kleur die niet voorkomt onder de andere schijven die q bevatten. Conflictvrije kleuringen zijn gerelateerd een graafkleuringen, en modelleren toekenningen van frequenties aan zendmasten waarbij interferentieproblemen voorkomen worden. In dit artikel zal Mark de Berg onder andere laten zien dat er altijd een conflictvrije kleuring bestaat die maar $O(\log n)$ kleuren gebruikt.

Draadloze netwerken, bijvoorbeeld voor mobiele telefonie, bestaan uit basisstations (zendmasten) die de communicatie voor gebruikers van het netwerk verzorgen. Als een gebruiker binnen bereik is van meerdere zendmasten, dan heeft hij in principe meerdere mogelijkheden ter beschikking om de communicatie te laten verlopen. Als verschillende zendmasten echter dezelfde frequentie gebruiken, dan kan de communicatie met die zendmasten verstoord worden door interferentie. Dit kan natuurlijk voorkomen worden door elke zendmast een eigen, unieke frequentie toe te kennen, maar het uitgeven van veel verschillende frequenties is ongewenst. De vraag is dus: is het echt nodig om alle zendmasten een eigen frequentie te geven, of kunnen we ook met minder frequenties toe? In dit artikel zullen we zogenaamde conflictvrije kleuringen bekijken, die het bovenstaande probleem modelleren.

In onze modellering gaan we ervan uit dat een zendmast alle gebruikers kan bereiken binnen een gegeven vaste afstand. (Dit is in de praktijk niet precies het geval: het bereik van een zendmast wordt onder andere beïnvloed door gebouwen en atmosferische omstandigheden, en sommige zendmasten kunnen krachtiger dan andere zijn.) In ons model is het bereik van een zendmast dus een schijf met de zendmast als middelpunt en met een vaste straal. Het toekennen van een frequentie aan een zendmast wordt nu gemodelleerd als het toekennen van een kleur aan de bijbehorende schijf, waarbij we voor het gemak de i-de kleur identificeren met de integer i. Dit leidt tot het volgende abstracte probleem.

Laat $\mathcal{S} := \{S_1, \ldots, S_n\}$ een verzameling van n schijven in het vlak zijn, elk met dezelfde straal. Voor een punt $q \in \mathbb{R}^2$ definiëren we $\mathcal{S}(q) := \{S_i \in \mathcal{S} : q \in S_i\}$ als de verzameling schijven die q bevatten.

Een *kleuring* van $\mathcal{S}$ met c kleuren is een afbeelding κ die aan elke schijf $S_i \in \mathcal{S}$ een kleur $\kappa(S_i) \in \{1, \ldots, c\}$ toekent. Een kleuring κ wordt *conflictvrij* genoemd als het volgende geldt voor elk punt $q \in \mathbb{R}^2$: als $\mathcal{S}(q) \neq \emptyset$ dan is er een schijf in $\mathcal{S}(q)$ met een unieke kleur, dat wil zeggen een schijf $S_i \in \mathcal{S}(q)$ zodanig dat $\kappa(S_j) \neq \kappa(S_i)$ voor alle $S_j \in \mathcal{S}(q) \setminus \{S_i\}$. In de toepassing die we in gedachten hebben zou een gebruiker op locatie q dus kunnen communiceren via de zendmast behorend bij



Figuur 1 Voorbeeld van een conflictvrije kleuring van zeven schijven: elk punt in de vereniging van de schijven is bevat in ten minste een schijf met een unieke kleur.

S_i, want alle andere zendmasten die binnen bereik zijn gebruiken een andere frequentie. Figuur 1 illustreert het concept van conflictvrije kleuringen.

De vraag die we nu willen beantwoorden is: wat is het kleinste aantal kleuren, als functie van n, zodat we voor elke verzameling $\mathcal{S}$ van n schijven een conflictvrije kleuring kunnen vinden? Wat formeler: als $\chi_{\text{cf}}(\mathcal{S})$ het minimale aantal kleuren is dat nodig is om $\mathcal{S}$ conflictvrij te kleuren, dan willen we

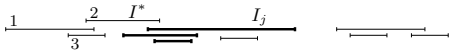
$$\chi_{\text{cf}}(n) := \max_{|\mathcal{S}|=n} \chi_{\text{cf}}(\mathcal{S})$$

bepalen.

Conflictvrije kleuringen zijn gerelateerd aan graafkleuringen. In een gewone graafkleuring moeten de knopen van een graaf op een zodanige manier gekleurd worden dat de eindpunten van elke kant in de graaf verschillend gekleurd zijn. De beroemde *vierkleurenstelling* zegt dat elke planaire graaf met vier kleuren gekleurd kan worden. Laat $\mathcal{G}_\mathcal{S}$ nu de intersectiegraaf van $\mathcal{S}$ zijn, dat wil zeggen de graaf met een knoop voor elke schijf $S_i \in \mathcal{S}$ en een kant tussen twee schijven S_i en S_j als $S_i \cap S_j \neq \emptyset$. Dan geeft een kleuring van de graaf $\mathcal{G}_\mathcal{S}$ een conflictvrije kleuring van de schijven. Immers, voor elk punt $q \in \mathbb{R}$ snijden alle schijven in $\mathcal{S}(q)$ elkaar, dus alle schijven in $\mathcal{S}(q)$ moeten verschillend gekleurd zijn. Helaas is de intersectiegraaf van een verzameling schijven niet noodzakelijk planair. Sterker, als alle schrijven elkaar snijden dan is de intersectiegraaf $\mathcal{G}$ een volledige graaf en zijn er dus n kleuren nodig voor een gewone kleuring van $\mathcal{G}_\mathcal{S}$. Dit betekent niet dat een conflictvrije kleuring van $\mathcal{S}$ in dit geval ook n kleuren nodig heeft. Zo zijn de linker vier schijven in Figuur 1 conflictvrij gekleurd met drie kleuren, terwijl de intersectiegraaf van deze vier schijven volledig is. We zullen zien dat, hoe de schijven ook liggen, er altijd een conflictvrije kleuring bestaat met maar $O(\log n)$ kleuren.

Het eendimensionale geval

Om meer inzicht in het probleem te krijgen bestuderen we eerst de eendimensionale versie, waarbij een verzameling $\mathcal{I} := \{I_1, \ldots, I_n\}$ van n intervallen in $\mathbb{R}^1$ conflictvrij gekleurd moet worden. Een conflictvrije kleuring is hier gedefinieerd analoog aan het tweedimensionale geval: voor elk punt $q \in \mathbb{R}^1$ moet gelden dat $\mathcal{I}(q)$, de verzameling van intervallen die q bevatten, ten minste één interval heeft met een kleur die uniek is binnen $\mathcal{I}(q)$. Het blijkt dat er in deze eendimensionale versie maar drie kleuren nodig zijn voor een conflictvrije kleuring. Het volgende algoritme genereert zo'n kleuring. We gaan er in de beschrijving voor het gemak van uit dat de er geen eindpunten van intervallen samenvallen; het is niet moeilijk om het algoritme aan te passen als dit wel zo is.



Figuur 2 Het kleuringsalgoritme voor intervallen. Op het moment dat I^* behandeld wordt zijn er al drie intervallen gekleurd. De intervallen in $\mathcal{I}_{\text{in}}$ zijn dikgedrukt. Het algoritme is in geval (i), en zal I_j kleur 1 geven; de twee andere intervallen in $\mathcal{I}_{\text{in}}$ krijgen kleur 3. In de volgende stap zal I^* veranderd worden in I_j ,en zal geval (ii) van toepassing zijn.

Laat I_1 het interval zijn met het meest linkse linker eindpunt. We geven interval I_1 kleur 1 en kleuren daarna de overige intervallen als volgt.

Laat $\mathcal{I}^*$ de verzameling intervallen zijn die nog gekleurd moeten worden, en laat I^* het meest rechtse al gekleurde interval zijn. In het begin geldt dus $\mathcal{I}^* = \mathcal{I} \setminus \{I^*\}$ en $I^* = I_1$, maar in de loop van het algoritme zullen intervallen uit $\mathcal{I}^*$ verwijderd worden en zal I^* veranderen. We zullen er echter voor zorgen dat I^* altijd kleur 1 of 2 heeft. Bekijk nu alle intervallen in $\mathcal{I}^*$ die hun linker eindpunt in I^* hebben. Laat $\mathcal{I}_{\text{in}}$ deze verzameling intervallen zijn. Er zijn twee gevallen, geïllustreerd in Figuur 2.

i. Als $\mathcal{I}_{\text{in}}$ ten minste één interval bevat dat niet helemaal binnen I^* ligt, laat dan $I_j \in \mathcal{I}_{\text{in}}$ het interval zijn dat het verst naar rechts uitsteekt. We kleuren I_j als volgt: $\kappa(I_j) := 2$ als $\kappa(I^*) = 1$ en $\kappa(I_j) := 1$ als $\kappa(I^*) = 2$. Alle andere intervallen in $\mathcal{I}_{\text{in}}$ krijgen kleur 3. Tenslotte verwijderen we de zojuist gekleurde intervallen uit $\mathcal{I}^*$, en veranderen we I^* in I_j.
ii. Als alle intervallen in $\mathcal{I}_{\text{in}}$ bevat zijn in het interval I^* — het geval $\mathcal{I}_{\text{in}} = \emptyset$ valt hier onder —, geef dan deze intervallen kleur 3 en verwijder ze uit I^*. Neem daarna van alle overblijvende intervallen in $\mathcal{I}^*$ het interval I_j met het meest linkse linker eindpunt, geef I_j kleur 1, verwijder I_j uit $\mathcal{I}^*$, en verander I^* in I_j.

Nadat we het relevante geval hebben afgehandeld wordt het proces herhaald met de nieuwe $\mathcal{I}^*$ en I^*, net zolang tot $\mathcal{I}^* = \emptyset$ en dus alle intervallen gekleurd zijn. Dit leidt tot de volgende stelling.

Stelling 1. *Elke verzameling van n intervallen in $\mathbb{R}^1$ kan conflictvrij gekleurd worden met drie kleuren.*

Bewijs. Het bovenstaande algoritme gebruikt drie kleuren. Dat de kleuring conflictvrij is, volgt uit de volgende twee observaties.

Ten eerste zal een interval met kleur 1 nooit een ander interval met kleur 1 overlappen. Immers, als een interval I_j kleur 1 krijgt, dan krijgen in de stap daarna alle intervallen die I_j overlappen kleur 2 of 3. Eenzelfde argument laat zien dat intervallen met kleur 2 elkaar nooit overlappen.

Ten tweede ligt elk punt q dat bevat is in een interval I_k van kleur 3 ook in een interval van kleur 1 of 2. (De intervallen van kleur 3 hoeven dus door geen enkel punt q 'gebruikt' te worden.) Om dit in te zien, bekijk het moment waarop I_k gekleurd wordt. Het linker eindpunt van I_k ligt in het interval dat op dat moment I^* is, en I_k steekt minder ver uit dan het interval I_j dat op dat moment kleur 1 of 2 krijgt. Dus $I_k \subset I^* \cup I_j$, hetgeen de claim impliceert. □

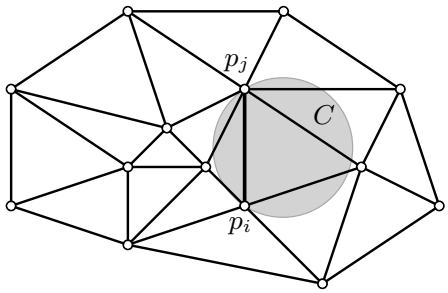
Het tweedimensionale geval

Het tweedimensionale geval, waarin we schijven in het vlak willen kleuren, is een stuk lastiger dan het eendimensionale geval. Zoals eerder aangegeven zullen we aannemen dat alle schijven in $\mathcal{S}$ dezelfde straal hebben. Zelfs dan blijkt het niet altijd mogelijk om een conflictvrije kleuring te geven met een constant aantal kleuren.

Stelling 2. *Voor elke n is er een verzameling van n schijven in $\mathbb{R}^2$ waarvoor elke conflictvrije kleuring $\Omega(\log n)$ kleuren gebruikt.*

Bewijs. Bekijk de verzameling $\mathcal{S} := \{S_1, \ldots, S_n\}$ waarin S_i een schijf is met straal 1 en middelpunt $(i/n, 0)$. Merk op dat er voor elke paar indices i, j met $1 \leqslant i < j \leqslant n$ een punt q in het vlak bestaat waarvoor geldt dat $\mathcal{S}(q) = \{S_i, S_{i+1}, \ldots, S_j\}$. Omdat er een punt q is met $\mathcal{S}(q) = \mathcal{S}$, moet er een schijf $S_k \in \mathcal{S}$ zijn met een unieke kleur. Stel dat $k \leqslant n/2$ en bekijk de verzameling $\mathcal{S}' := \{S_{k+1}, \ldots, S_n\}$. Merk op dat $|\mathcal{S}'| \geqslant \lfloor n/2 \rfloor$. (Als $k > n/2$, dan bekijken we $\{S_1, \ldots, S_{k-1}\}$.) Ook in $\mathcal{S}'$ moet er weer een schijf zijn met een kleur die uniek is binnen $\mathcal{S}'$, aangezien er een punt q is met $\mathcal{S}(q) = \mathcal{S}'$, enzovoorts. (Binnen $\mathcal{S} \setminus (\mathcal{S}' \cup \{S_k\}$ kan de kleur wel hergebruikt worden.) We concluderen dat het aantal benodigde kleuren, $K(n)$, voor deze configuratie van n schijven voldoet aan de recurrente betrekking $K(n) \geqslant 1 + K(\lfloor n/2 \rfloor)$, waaruit de stelling volgt. □

Gelukkig wordt het niet veel erger dan in bovenstaande stelling: elke verzameling van n even grote schijven kan conflictvrij gekleurd worden met $O(\log n)$ kleuren. Het algoritme hiervoor gebruikt de Delaunay-triangulatie,



Figuur 3 Voorbeeld van een Delaunay triangulatie. De grijze cirkel C illustreert de lege-cirkel eigenschap van het paar p_i,p_j.

die we kort introduceren voordat we het algoritme beschrijven.

De Delaunay-triangulatie

Laat $\mathcal{P}$ een verzameling van n punten in het vlak zijn. Een *triangulatie* van $\mathcal{P}$ is een verdeling van het convexe omhulsel van $\mathcal{P}$ in driehoeken, waarbij de verzameling hoekpunten van de driehoeken gelijk is aan $\mathcal{P}$. Een bijzondere triangulatie is de *Delaunay-triangulatie*, genoemd naar de Russische wiskundige Boris Delaunay (1890–1980). De Delaunay-triangulatie DT(P) wordt verkregen door een lijnstuk te trekken tussen elk paar punten $p_i, p_j \in P$ dat de *lege-cirkeleigenschap* heeft: er is een cirkel C met p_i en p_j op de rand die verder leeg is, dat wil zeggen, die geen andere punten van $\mathcal{P}$ bevat in het inwendige of op de rand; zie Figuur 3. (Als er vier of meer punten van $\mathcal{P}$ precies op een cirkel liggen, dan hoeft het verbinden van alle paren met de lege-cirkeleigenschap geen volledige triangulatie op te leveren. Indien gewenst kan een volledige triangulatie verkregen worden door het toevoegen van een aantal extra verbindingen, maar voor onze toepassing is dat niet nodig.)

De Delaunay-triangulatie wordt voor allerlei toepassingen gebruikt en heeft prachtige eigenschappen. Zo is de Delaunay-triangulatie de duale van het bekende Voronoi-diagram: twee punten p_i, p_j zijn verbonden in DT(P) dan en slechts dan als de Voronoi-cellen van p_i en p_j buren zijn in Vor(P). (Het Voronoi-diagram Vor(P) is de opdeling van het platte vlak in Voronoi-cellen, één per punt $p_i \in P$, zodanig dat de cel van p_i precies die punten $q \in \mathbb{R}^2$ bevat waarvoor p_i het dichtstbijzijnde punt in $\mathcal{P}$ is.) Voor ons zijn maar twee eigenschappen van de Delaunay-triangulatie van belang: de lege-cirkeleigenschap en het feit dat DT(P) een planaire graaf is (met $\mathcal{P}$ als de verzameling knopen, en de verbindingen tussen de punten als kanten).

Conflictvrije kleuring

We keren nu terug naar het kleuringsprobleem voor een verzameling $\mathcal{S}$ van even grote schijven in het vlak. We zullen dit probleem eerst transformeren naar een kleuringsprobleem op punten.

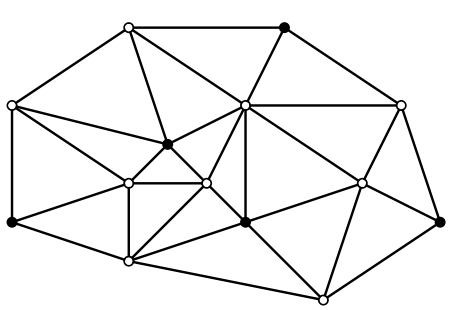
Laat p_i het middelpunt van de schijf $S_i \in \mathcal{S}$ zijn, en laat $\mathcal{P} := \{p_i : S_i \in \mathcal{S}\}$ de verzameling van alle middelpunten zijn. Laat r de gemeenschappelijk straal zijn van de schijven in $\mathcal{S}$. Merk op dat voor elk punt $q \in \mathbb{R}^2$ en elke schijf $S_i \in \mathcal{S}$ geldt dat $q \in S_i$ dan en slechts dan als $p_i \in D_q$, waarbij D_q de schijf is met middelpunt q en straal r. Definieer $\mathcal{P}(D_q)$ als de verzameling van punten $p_i \in \mathcal{P}$ die in D_q liggen. Dan is het vinden van een conflictvrije kleuring voor $\mathcal{S}$ equivalent aan het vinden van een conflictvrije kleuring voor $\mathcal{P}$, waarbij een kleuring voor $\mathcal{P}$ conflictvrij is als het volgende geldt voor elk punt $q \in \mathbb{R}^2$: als $\mathcal{P}(D_q) \neq \emptyset$ dan is er een punt $p_i \in \mathcal{P}(D_q)$ met een unieke kleur. De Delaunay-triangulatie geeft een elegant algoritme om een conflictvrije kleuring voor $\mathcal{P}$ te vinden, zoals hieronder beschreven.

Het algoritme werkt in $O(\log n)$ fasen. In de eerste fase geven we een aantal punten kleur 1, in de tweede fase geven we een aantal punten kleur 2, enzovoorts, totdat alle punten een kleur hebben gekregen. Het selecteren van de punten die in de k-de fase gekleurd worden, gaat als volgt.

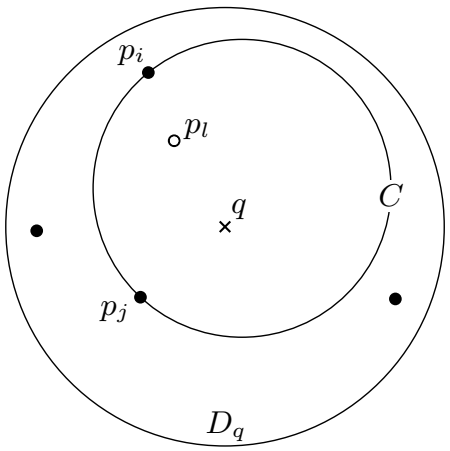
Laat $\mathcal{P}_k \subseteq \mathcal{P}$ de deelverzameling van de punten zijn die nog geen kleur hebben gekregen, en laat DT($\mathcal{P}_k$) de Delaunay-triangulatie van $\mathcal{P}_k$ zijn. Een *independent set* in DT($\mathcal{P}_k$) is een deelverzameling van punten uit $\mathcal{P}_k$ die niet met elkaar verbonden zijn in DT($\mathcal{P}_k$); zie Figuur 4.

Omdat DT($\mathcal{P}_k$) een planaire graaf is, heeft DT($\mathcal{P}_k$) een independent set $\mathcal{I}_k$ met ten minste $n_k/4$ punten. We selecteren de punten in $\mathcal{I}_k$ als de punten die kleur k krijgen. De verzameling $\mathcal{P}_{k+1}$ die voor volgende fase overblijft, is dus $\mathcal{P}_k \setminus \mathcal{I}_k$.

Nadat we alle punten uit $\mathcal{P}$ gekleurd hebben volgens het bovenstaande algoritme, geven we elke schijf $S_i \in \mathcal{S}$ de kleur van het bijbehorende middelpunt. Dit leidt tot het volgende resultaat.



Figuur 4 Een independent set (de zwarte punten) in de Delaunay triangulatie.



Figuur 5 Illustratie bij het bewijs van Stelling 3. De zwarte punten zijn de punten uit $\mathcal{P}(D_q) \cap \mathcal{I}_k$.

Stelling 3. *Elke verzameling $\mathcal{S}$ van n schijven in $\mathbb{R}^2$ kan conflictvrij gekleurd worden met $O(\log n)$ kleuren.*

Bewijs. Bekijk het bovenstaande algoritme om de verzameling $\mathcal{P}$ van middelpunten van de schijven te kleuren. Het algoritme begint met n punten, en kleurt ten minste een kwart van de overgebleven punten in elke fase. Als n_k het aantal overgebleven punten aan het begin van de k-de fase is, dan geldt dus $n_{k+1} \leqslant 3n_k/4$. Het aantal fasen, en daarmee het aantal kleuren, is daarom $O(\log n)$.

Blijft over te bewijzen dat de kleuring conflictvrij is. Laat $q \in \mathbb{R}^2$ een punt in het vlak zijn, en laat $\mathcal{S}(q)$ de verzameling schijven zijn die q bevatten. Neem aan dat $\mathcal{S}(q) \neq \emptyset$. We willen bewijzen dat er een schijf $S_i \in \mathcal{S}(q)$ is met een unieke kleur. Zoals al eerder opgemerkt komt dit overeen met te bewijzen dat $\mathcal{P}(D_q)$, de verzameling middelpunten die in D_q liggen, een punt met een unieke kleur heeft. Om dit te bewijzen zullen we beargumenteren dat de hoogst voorkomende kleur in $\mathcal{P}(D_q)$ uniek is.

Laat k de hoogste kleur in $\mathcal{P}(D_q)$ zijn, laat $\mathcal{P}_k$ de verzameling overgebleven punten aan het begin van de k-de fase zijn, en laat $\mathcal{I}_k$ de independent set in DT($\mathcal{P}_k$) zijn die de kleur k krijgt. Merk op dat $\mathcal{P}(D_q) \cap \mathcal{I}_k \neq \emptyset$. We beweren dat $|\mathcal{P}(D_q) \cap \mathcal{I}_k| = 1$, hetgeen betekent dat de kleur k inderdaad uniek is in $\mathcal{P}(D_q)$. Om deze bewering te bewijzen zullen we laten zien dat $|\mathcal{P}(D_q) \cap \mathcal{I}_k| \geqslant 2$ tot een contradictie leidt. Het is niet moeilijk in te zien dat $|\mathcal{P}(D_q) \cap \mathcal{I}_k| \geqslant 2$ het volgende impliceert: er is een cirkel $C \subset D_q$ die twee punten $p_i, p_j \in \mathcal{P}(D_q) \cap \mathcal{I}_k$ op de rand heeft en verder

geen punten van $\mathcal{P}(D_q)$ bevat; zie Figuur 5. (Zo'n cirkel kan verkregen worden D_q op de juiste manier te krimpen tot aan de voorwaarden voldaan wordt.) De cirkel C moet echter wel een punt $p_l \in \mathcal{P}_k$ bevatten. Immers, als C verder leeg zou zijn dan zou het paar p_i, p_j de lege-cirkel eigenschap hebben, en dit zou betekenen dat p_i en p_j niet beide in de independent set $\mathcal{I}_k$ kunnen zitten. Omdat $p_l \in \mathcal{P}_k \setminus \mathcal{I}_k$, krijgt p_l een hogere kleur dan k, een contradictie met de definitie van k. □

Slotopmerkingen

Conflictvrije kleuringen werden in 2003 geïntroduceerd door Even e.a. [1], die onder andere bewezen dat elke verzameling van n schijven met $O(\log n)$ kleuren conflictvrij gekleurd kan worden. Hun bewijs is een stuk algemener dan de versie die we hier besproken hebben. Het werkt bijvoorbeeld ook voor verzamelingen *pseudo-disks*, dat wil zeggen verzamelingen van samenhangende gebieden zodanig dat de randen van elk paar gebieden elkaar maar in ten hoogste twee punten snijden. Sinds 2003 zijn er veel varianten van conflictvrije kleuringen bestudeerd; Smorodinski [2] geeft een overzicht van het werk op dit gebied. Toch zijn nog lang niet alle vragen beantwoord. Zo is het ondergrensvoorbeeld van Stelling 2 niet erg realistisch: in de praktijk zullen de de zendmasten nooit zo dicht bij elkaar geplaatst worden. Dat leidt bijvoorbeeld tot de vraag: met hoeveel kleuren kunnen we toe als geen geen enkel punt in het vlak binnen bereik is van meer dan k van de n zendmasten, voor $k \ll n$? Het voorbeeld van Figuur 2 laat zien dat er in dit geval soms $\Omega(\log k)$ kleuren nodig zijn, maar er een goede bovengrens is niet bekend. Zelfs voor het geval dat elke schijf maar k andere schijven snijdt is de best bekende bovengrens $O(\log^2 k)$ [2]. Ook over andere meer realistische modellen, bijvoorbeeld waarin er rekening mee wordt gehouden dat obstakels het bereik van de zendmasten kunnen beïnvloeden, is weinig bekend. Kortom: nog veel uitdagingen voor onderzoek!

Referenties

1 G. Even, Z. Lotker, D. Ron en S. Smorodinsky, Conflict-free colorings of simple geometric regions with applications to frequency assignment in cellular networks, *SIAM J. Comput.* 33 (2003), 94–136.

2 S. Smorodinski, Conflict-Free Coloring and its Applications, in I. Bárány, K.J. Böröczky, G. Fejes Tóth en J.Pach (red.) *Geometry – Intuitive, Discrete, and Convex*, Bolyai Society Mathematical Studies 24, Springer, 2013, pp. 331–389.

Karen Aardal
Delft Institute of Applied Mathematics
TU Delft
k.i.aardal@tudelft.nl

Thije van Barneveld
Centrum Wiskunde & Informatica
Amsterdam
t.c.van.barneveld@cwi.nl

Pieter van den Berg
Delft Institute of Applied Mathematics
TU Delft
p.l.vandenberg@tudelft.nl

Sandjai Bhulai
Afdeling Wiskunde
Vrije Universiteit Amsterdam
s.bhulai@vu.nl

Martin van Buuren
Centrum Wiskunde & Informatica
Amsterdam
m.van.buuren@cwi.nl

Theresia van Essen
Centrum Wiskunde & Informatica, Amsterdam, en
Delft Institute of Applied Mathematics, TU Delft
j.t.vanessen@tudelft.nl

Caroline Jagtenberg
Centrum Wiskunde & Informatica
Amsterdam
c.j.jagtenberg@cwi.nl

Geert Jan Kommer
Rijksinstituut voor Volksgezondheid en Milieu
Bilthoven
geertjan.kommer@rivm.nl

Guido Legemaate
Brandweer Regionale Eenheid Amsterdam, en
Centrum Wiskunde & Informatica, Amsterdam
g.a.g.legemaate@cwi.nl

Rob van der Mei
Centrum Wiskunde & Informatica
Amsterdam
r.d.van.der.mei@cwi.nl

Van reactieve naar proactieve planning van ambulancediensten

In dit artikel berichten Karen Aardal e.a. over het REPRO-project, een onderzoeksproject dat zich richt op het ontwikkelen van slimme voorspellings- en planningsmethoden voor ambulancediensten. In Nederland is de norm dat een ambulance in 95 procent van de spoedeisende gevallen binnen vijftien minuten ter plaatse moet zijn. Het nieuw ontwikkelde systeem kan hiervoor uitrekenen wat de optimale locaties van de standplaatsen zijn en de spreiding van ambulances over de standplaatsen.

In levensbedreigende situaties waarin elke seconde telt, kan het al dan niet op tijd ter plaatse zijn van een ambulance het verschil maken tussen leven en dood. In deze gevallen is het streven om zo snel mogelijk en bij voorkeur binnen vijftien minuten ter plaatse van de patiënt te zijn. Om zulke korte responstijden te realiseren is een uitgekiende planning van ambulanceritten noodzakelijk. Bij het realiseren van een efficiënte aansturing van ambulanceritten spelen allerlei vragen een rol, zowel op strategisch, tactisch als operationeel niveau. Typische vragen zijn: "Hoe kunnen we op de juiste wijze anticiperen en reageren op pieken in de vraag naar ambulanceritten?", "Hoeveel standplaatsen hebben we nodig en wat zijn de goede locaties voor die standplaatsen?", "Hoeveel teams moeten we minimaal inzetten om de gewenste servicegraad te realiseren, en waar en wanneer?" en "Hoe kunnen we op elk moment van de dag een goede dekking van ambulances over een regio realiseren door slimme dynamische en proactieve real-time herpositionering van ambulances?"

In het STW-project 'Van Reactieve naar Proactieve Planning van Ambulancediensten', kortweg REPRO, staan bovengenoemde vragen centraal. Binnen het project worden nieuwe modellen en oplossingsmethoden niet alleen ontwikkeld en geëvalueerd, maar ook daadwerkelijk toegepast in de praktijk. Het project is een samenwerking tussen het Centrum Wiskunde & Informatica (CWI), de Technische Universiteit Delft, de Vrije Universiteit Amsterdam, het Rijksinstituut voor Volksgezondheid en Milieu (RIVM), Ambulance Amsterdam, Witte Kruis Ambulancezorg en de regionale ambulancevoorzieningen Flevoland, Gooi- en Vechtstreek en Utrecht. Daarnaast zijn ook Brandweer Regionale Eenheid Amsterdam, hulpdiensten-softwarespecialist CityGIS, Politie Regionale Eenheid Amsterdam, Ambulancezorg Nederland, het Universitair Medisch Centrum Groningen, het Oranje Kruis en de ANWB bij het project betrokken. Het is belangrijk om op te merken dat de vraagstukken die voorkomen in de ambulancesector vergelijkbaar zijn met vraagstukken die voorkomen bij de planning van dienstverlening in andere sectoren, zoals de brandweer, de politie, de wegenwacht, maar ook

bij het managen van *spare parts* waarbij (dure) reserveonderdelen over een regio moeten worden verspreid zodanig dat ze snel kunnen worden ingezet als onderdelen van apparatuur en machines uitvallen. De wiskundige modellen die wij ontwikkelen en gebruiken hebben daardoor een breder nut dan alleen voor ambulanceplanning.

Wij modelleren de planningsproblemen rondom de ambulancezorg als *optimaliseringsproblemen in netwerken*. De netwerkstructuur is meer of minder expliciet aanwezig afhankelijk van het onderliggende vraagstuk. Om de algoritmes en oplossingen voor de vraagstukken te analyseren, testen en evalueren, hebben we een simulatietool ontwikkeld.

Een sterk complicerende, maar wetenschappelijk zeer uitdagende, factor bij het oplossen van dit soort planningsvraagstukken is de grote mate van *onzekerheid* die inherent is aan het ambulance-serviceproces. De bestaande planningsmethodieken gaan er doorgaans van uit dat de vraag naar en beschikbaarheid van ambulances van tevoren bekend zijn, terwijl veel van deze methoden zeer gevoelig zijn voor veranderingen van de invoerparameters.

In dit artikel zullen verschillende aspecten van ambulanceplanning worden belicht. In de volgende paragraaf worden verschillende modellen besproken voor het bepalen van de goede locaties voor standplaatsen, en de bijbehorende bezetting van ambulances. Daarna wordt ingegaan op locatiemodellen specifiek voor de dienstverlening van de brandweer. Vervolgens richten wij ons op modellen en algoritmen door dynamisch ambulance management, waarbij de dekking in realtime wordt verbeterd door middel van het uitvoeren van zogenaamde *proactieve relocaties* op basis van de actuele en toekomstige locaties van de ambulancevoertuigen en de incidenten. Dan volgt nog een paragraaf waarin wij verder ingaan op het optimaal plannen van besteld ambulancevervoer en een paragraaf waarin zal worden ingegaan op de simulatietool *Testing Interface for Ambulance Research* (TIFAR) waarin de ontwikkelde plannings- en relocatiealgoritmen zijn geïmplementeerd. Tot slot wordt in de laatste paragraaf een simulatiemodel voor ambulancemeldkamers gepresenteerd.

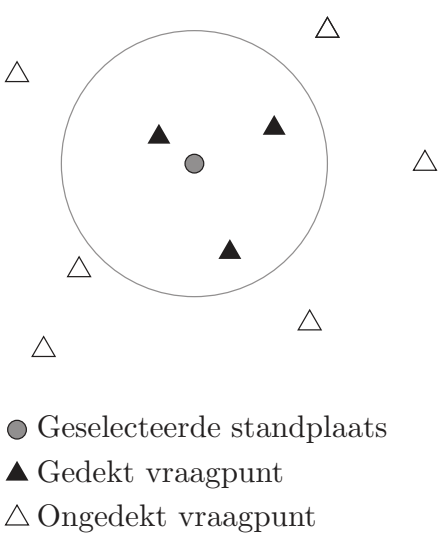
Voordat we op de verscheidene aspecten van de ambulanceplanning ingaan, geven we een korte beschrijving van hoe de ambulancesector in Nederland is georganiseerd. De organisatie van de ambulancezorg is verdeeld over 25 regio's. In elke regio is er een Regionale Ambulancevoorziening (RAV) verantwoordelijk voor het verlenen van de ambulancezorg, zowel voor het meldkamer- als het rittendomein. De ambulancesector in Nederland is de laatste jaren sterk in beweging. Door de invoering van de Tijdelijke Wet op Ambulancezorg worden de prestaties van de verschillende RAV's gemeten en beoordeeld aan de hand van een aantal prestatie-indicatoren. De door een RAV geleverde prestatie kan gevolgen hebben voor de financiering, waardoor het leveren van topkwaliteit ambulancezorg voor RAV's cruciaal is.

Binnenkomende meldingen worden via een standaarduitvraagprotocol door de meldkamercentralist geclassificeerd in drie urgentieklassen: A1, A2 en B. Een melding van een ernstige levensbedreigende situatie wordt geclassificeerd als een A1-urgentie, en in dat geval is de richtlijn dat er binnen vijftien minuten hulp ter plaatse moet zijn. Is de situatie minder ernstig, dan wordt de melding als A2 aangemerkt en geldt een responstijd van dertig minuten. Tot slot zijn er B-ritten. Dat zijn planbare ritten waarbij een patiënt bijvoorbeeld vanaf een ziekenhuis naar huis of naar een andere zorginstelling wordt vervoerd. In Nederland is het aantal binnenkomende meldingen ongeveer $1{,}1$ miljoen per jaar, waarvan ongeveer 49 procent als A1 wordt aangemerkt, 24 procent A2 en 27 procent B.

In het vervolg wordt deze Nederlandse situatie als uitgangspunt gebruikt.

Locatieproblemen voor ambulancevervoer

Een van de eerste problemen in de ambulancelogistiek waarop de wiskunde is toegepast is het bepalen van locaties voor standplaatsen. Het doel van de hiervoor gebruikte modellen is het verlagen van de aanrijtijden naar patiënten. Als prestatiemaat wordt hier vaak de dekking binnen een gegeven tijdslimiet gebruikt. Deze dekking is de fractie van het aantal oproepen dat binnen de dekkingsnorm wordt bereikt. In Nederland is deze norm vijftien minuten voor inzetten met A1-urgentie.



Figuur 1

Bij het definiëren van deze locatiemodellen wordt gebruik gemaakt van een verzameling met vraagpunten V en een verzameling met potentiële standplaatslocaties W. De rijtijd van een standplaats $i \in W$ naar een vraagpunt $j \in V$ wordt bekend verondersteld en wordt genoteerd als t_{ij}. In de rijtijd zit een zogenaamd 'pre-trip delay' ingebakken. Dat is de tijd tussen het binnenkomen van een melding en het moment dat de ambulance het basisstation verlaat. Op basis van deze rijtijden kan voor elk vraagpunt j een verzameling $W_j \subset W$ worden afgeleid die alle potentiële standplaatslocaties bevat die dekking kunnen verlenen aan vraagpunt j. Dus $W_j = \{i \in V | t_{ij} \leq r\}$, waarbij r de gestelde dekkingsnorm is. Om de dekking van de verschillende vraagpunten te kunnen wegen, wordt aan elk vraagpunt j een weging d_j toegekend. Voor deze weging zou bijvoorbeeld het aantal inwoners of het verwachte aantal oproepen in een gebied gebruikt kunnen worden. Zie Figuur 1.

De eerste en eenvoudigste modellen die gebruikt worden in de literatuur proberen de enkelvoudige dekking te maximaliseren. Hierbij wordt een gebied als gedekt beschouwd als tenminste één ambulance gestationeerd is binnen de gestelde norm. Dit model staat bekend als het *Maximum Coverage Location Problem* (MCLP) [7]. In dit model hebben we een binaire variabele x_i die waarde 1 aanneemt als er een ambulance is gestationeerd op locatie $i \in W$. Verder geeft de binaire variabele y_j aan of vraagpunt $j \in V$ is gedekt door tenminste één ambulance. Het model maximaliseert de dekking die behaald kan worden met maximaal p ambulances. Het model is als volgt:

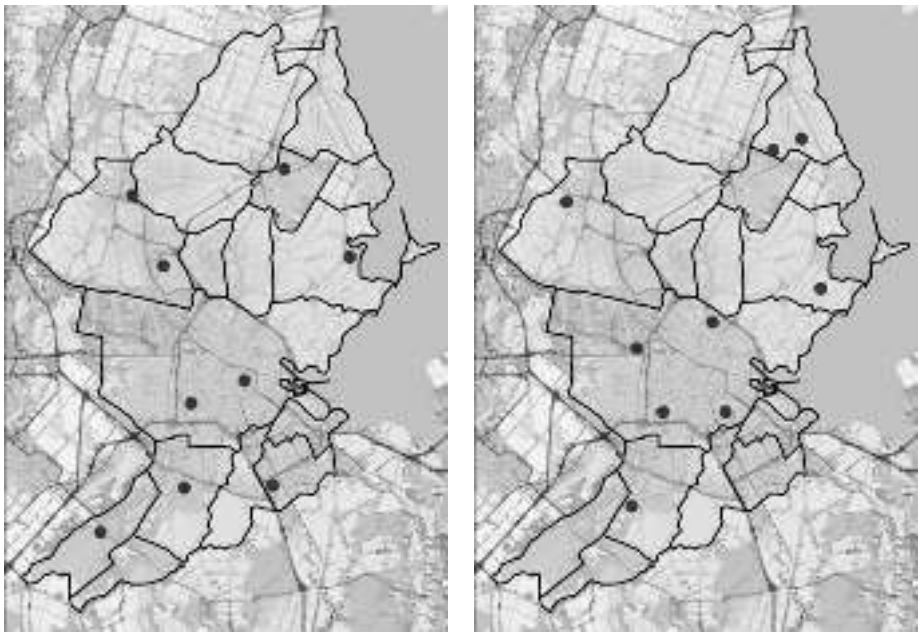
$$
\begin{aligned}
\max \quad & \sum_{j \in V} d_j y_j \\
\text{s.t.} \quad & \sum_{i \in W_j} x_i \geq y_j \quad (\forall j \in V), \\
& \sum_{i \in W} x_i = p, \\
& x_i \in \{0,1\} \quad (\forall i \in W), \\
& y_j \in \{0,1\} \quad (\forall j \in V).
\end{aligned}
$$

De eerste voorwaarde dwingt af dat de variabele y_j enkel waarde 1 krijgt als er daadwerkelijk een ambulance in staat is dit gebied te dekken. De tweede voorwaarde beperkt het aantal gebruikte standplaatsen. Dit relatief simpele model is NP-moeilijk en is een speciaal geval van het maximaliseren van een submodulaire functie. Ondanks de complexiteitsstatus blijkt dit model voor grote instanties nog goed oplosbaar. Het model is voor meer dan 10.000 gebieden oplosbaar binnen redelijke rekentijd. Dit terwijl grote RAV's in Nederland meestal niet meer dan 500 gebieden bevatten. Een gebied is typisch een postcodegebied als we de vier cijfers van de postcode beschouwen en is dan zowel een vraagpunt als een potentiële standplaatslocatie.

Hoewel MCLP vaak gebruikt is voor het bepalen van standplaatslocaties, is de toepasbaarheid beperkt als naast de standplaatsen ook de ambulanceverdeling bepaald dient te worden. In dat geval gaat MCLP ervan uit dat een ambulance altijd op zijn post aanwezig is, zodat een enkele ambulance voldoende is voor volledige dekking. In de praktijk is dit niet het geval en, met name in stedelijk gebied, is het van groot belang back-updekking te garanderen. We presenteren twee modellen die als maat voor de verwachte beschikbaarheid van de ambulances gebruik maken van het concept van een bezettingsgraad. Hier is de bezettingsgraad gedefinieerd als de verwachte totale werklast gedeeld door de totale beschikbare capaciteit.

Het eerste model dat deze bezettingsgraad q gebruikt is het *Maximum Availability Location Problem* (MALP) [21], waarbij met behulp van q wordt berekend hoeveel ambulances nodig zijn om met betrouwbaarheid α een ambulance beschikbaar te hebben. Door aan te nemen dat ambulances onafhankelijk van elkaar beschikbaar zijn, geeft dit dat $b = \lceil \frac{\log(1-\alpha)}{\log q} \rceil$ ambulances nodig zijn. Vervolgens wordt de dekking met tenminste b ambulances gemaximaliseerd.

Een andere aanpak is om de verwachte dekking, op basis van de bezettingsgraad q, mee te nemen in de doelfunctie. Dit wordt gedaan in het *Maximum Expected Coverage Location Problem* (MEXCLP) [8], waarbij de kans dat tenminste één ambulance beschikbaar is binnen de norm wordt gemaximaliseerd. Gegeven een dekking met k ambulances is de verwachte dekking $E_k = 1 - q^k$. Voor modelleerdoeleinden wordt vaak gebruik gemaakt van de marginale dekking van de k-de ambulance, gegeven door $E_k - E_{k-1} = q^{k-1}(1-q)$. Dit geeft de volgende doelstellingsfunctie:



Figuur 2 In de linkerplattegrond worden de huidige standplaatsen weergegeven. Rechts zien we de standplaatsen die het resultaat zijn van het oplossen van MEXCLP met tijdsafhankelijke input [3] met de extra eis dat elk gebied door tenminste één ambulance gedekt moet zijn.

$$\max \sum_{j \in V} \sum_{k=1}^{p} d_j (1-q) q^{k-1} y_{jk}.$$

Hierbij is y_{jk} een binaire variabele die aangeeft of vraagpunt j met tenminste k ambulances is gedekt, en x_i representeert nu het aantal ambulances op standplaats i. Deze variabele is niet langer binair, omdat meer dan één ambulance op een standplaats gestationeerd kan worden. Om de variabelen y_{jk} de juiste waarde te geven wordt de volgende voorwaarde toegevoegd:

$$\sum_{i \in W_j} x_i \geq \sum_{k=1}^{p} y_{jk} (\forall j \in V).$$

Er zijn vele uitbreidingen mogelijk van de genoemde modellen. Een mogelijke uitbreiding is om rekening te houden met gedifferentieerd vervoer. Veel ambulancediensten werken niet alleen met auto's die patiënten kunnen vervoeren, maar ook met zogenaamde 'rapid responders' die snel naar de patiënt kunnen komen om hem te stabiliseren in afwachting op vervoer naar het ziekenhuis. Het is makkelijk om het model uit te breiden met verscheidene voertuigtypes met erbij behorende responstijden. Voor het bepalen van ambulancestandplaatsen heeft dit tot nu toe geen hoge prioriteit gehad, maar deze uitbreiding wordt wel besproken in de volgende paragraaf over uitbreidingen naar de brandweren en verderop in de paragraaf over besteld vervoer. Een belangrijke uitbreiding die wel is onderzocht binnen REPRO is om rekening te houden met tijdsafhankelijkheid in de input. In [3] zijn tijdsafhankelijke aspecten toegevoegd aan het model MEXCLP. Wij geven in Figuur 2 resultaten van het oplossen van het model ontwikkeld in [3] met input van de regio Amsterdam.

Het bepalen van standplaatsen is een strategische vraag binnen de ambulanceplanning, en rekentijden mogen dan lang zijn zonder dat het gebruik van de modellen wordt belemmerd. Alle besproken modellen zijn NP-moeilijk en we kunnen daardoor verwachten dat rekentijden behoorlijk oplopen als de instanties groter worden. Onze ervaring is echter dat dit meevalt, in ieder geval als we problemen voor een regio oplossen en niet voor heel Nederland. Om een voorbeeld te noemen: als we het model MEXCLP nemen met input van de grootste regio in Nederland, regio Groningen, krijgen we een instantie met 456 geheeltallige variabelen, 8208 binaire variabelen en 457 constraints. Op een laptop kan je die instantie oplossen in ongeveer één minuut rekentijd met behulp van het pakket IBM CPLEX, dat gebruik maakt van een combinatie van *branch-and-bound* en *cutting planes*.

Voor meer referenties verwijzen we naar het overzichtsartikel [16].

Locatieproblemen voor brandweerkazernes

De problematiek van het bepalen van de optimale locaties van de standplaatsen speelt ook bij de brandweer. Typische vragen zijn:

"Wat zijn de optimale locaties van de brandweerkazernes?" en "Hoeveel brandweerwagens van de verschillende types moeten op elk van de kazernes gestationeerd zijn, zodanig dat de dekking zo groot mogelijk is?" Er zijn echter een paar belangrijke verschillen tussen de ambulance en de brandweer. Ten eerste zijn er veel minder incidenten waarbij de brandweer moet uitrukken, waardoor het zelden voorkomt dat meer dan één brandweerincident tegelijkertijd plaatsvindt. Ten tweede heeft de brandweer te maken met verschillende typen voertuigen, die afhankelijk van het type incident ingezet kunnen worden. Ten derde zijn de maximale aanrijtijden bij de brandweer veel kleiner dan ingeval van ambulances.

Net als in de vorige sectie is V de verzameling vraagpunten van mogelijke incidentlocaties en W de verzameling van potentiële basislocaties. De verzameling beschikbare voertuigtypes noemen we K. In totaal zijn er c_k voertuigen van type $k \in K$ beschikbaar. Aan elk vraagpunt $j \in V$ en voor elk voertuigtype $k \in K$ wordt een gewicht d_{jk} toegekend dat het relatieve belang aangeeft van het dekken van vraagpunt j door een voertuig van type k.

Voor gegeven vraagpunt $j \in V$ en voertuigtype $k \in K$ definiëren we $W_{jk} \subset W$ als de verzameling van alle locaties die vraagpunt j dekken door voertuigtype k. De vraag of vraagpunt j wordt gedekt door basislocatie $i \in W$ en voertuigtype $k \in K$ hangt af van twee aspecten: (1) de dekkingsnorm r_{jk}, de maximaal toelaatbare aanrijtijd voor vraagpunt $j \in V$ en voertuigtype $k \in K$, en (2) t_{ij}, de rijtijd van i naar j inclusief de pre-trip delay. We nemen aan dat de rijtijden niet afhangen van het voertuigtype en dat de pre-trip delay vast is. Dus geldt voor $j \in V$ en $k \in K$ dat $i \in W_{jk}$ als en alleen als $t_{ij} \leq r_{jk}$.

In het model onderscheiden we drie typen variabelen. De variabele x_{ik} is het aantal voertuigen van type k dat gesitueerd is op potentiële basislocatie i; omdat we in dit model geen meervoudige dekking meenemen geldt dat $x_{ik} \in \{0,1\}$. De variabele $y_{jk} \in \{0,1\}$ neemt waarde 1 aan als vraagpunt j gedekt is door voertuigtype k. Tot slot definiëren we $z_i := 1$ als en alleen als basislocatie i wordt gebruikt door tenminste één voertuigtype, en $z_i := 0$ anders. Als een basislocatie i wordt gebruikt moeten we kosten β betalen. Het gebruik van hetzelfde basisstation voor verschillende voertuigtypen wordt hiermee gestimuleerd.

Het planningsprobleem kan als volgt worden geformuleerd als een geheeltallig lineair optimaliseringsprobleem:

$$
\begin{aligned}
\max \quad & \sum_{j \in V} \sum_{k \in K} d_{jk} y_{jk} - \beta \sum_{i \in W} z_i \\
\text{s.t.} \quad & \sum_{i \in W_{jk}} x_{ik} \geq y_{jk}, \quad j \in V, k \in K, \\
& \sum_{i \in W} x_{ik} \leq c_k, \quad k \in K, \\
& x_{ik} \leq z_i, \quad i \in W, k \in K, \\
& y_{jk}, z_i, x_{ik} \in \{0,1\}.
\end{aligned}
$$

De doelfunctie bestaat uit twee delen: het gewogen aantal meldingen die 'gedekt' zijn door het juiste voertuigtype, en kosten voor het aantal gebruikte basisstations. De eerste restrictie geeft aan dat vraagpunt j alleen gedekt is door voertuigtype k als tenminste één voertuig van type k voldoet aan de dekkingsnorm. De tweede restrictie beperkt het aantal voertuigen van elk type. Als gevolg van de derde restrictie kunnen we alleen maar een voertuig hebben op een basisstation dat geopend is.

In de gegeven formulering modelleert β de kosten van het in gebruik nemen van een basisstation. Deze parameter definieert de trade-off tussen de extra dekking als gevolg van het openen van een nieuwe basislocatie enerzijds en de daarmee gepaard gaande onderhoudskosten anderzijds. De waarde van β kan grofweg worden geïnterpreteerd als het aantal voorheen niet-gedekte incidentmeldingen die wel gedekt zouden zijn als een extra basislocatie zou worden geopend. In de praktijk is het soms lastig om een geschikte waarde voor β te bepalen. In dat geval kunnen we dit deel van de doelfunctie weglaten en in plaats daarvan een restrictie toevoegen op het aan basislocaties dat kan worden geopend.

Het boven gegeven model is NP-moeilijk. Echter, commerciële solvers zoals CPLEX kunnen voor realistische modelinstanties tot een optimale oplossing komen binnen redelijke tijd. Het model zoals beschreven is gebruikt voor het optimaliseren van de locaties en bezetting van de standplaatsen voor de Brandweer Amsterdam-Amstelland. In deze dataset worden vier voertuigtypen onderscheiden: tankautospuiten (TS), redvoertuigen (RV), hulpverleningsvoertuigen (HV) en waterongevallenwagens (WO). Het aantal beschikbare voertuigen van de vier typen bedragen respectievelijk 21, 9, 3 en 2. We nemen het aantal inzetten van deze voertuigen in het jaar 2011 als gegeven. In deze periode zijn deze 39.516 keer ingezet, per voertuigtype respectievelijk 29.016, 9.182, 615 en 703. We onderscheiden 2643 vraagpunten, die overeenkomen met de secties die in de praktijk gebruikt worden. Daarnaast onderscheiden we 2223 potentiële basislocaties. De rijtijden t_{ij} tussen de basislocaties en de vraagpunten zijn verkregen van de brandweer en zijn gebaseerd op geschatte reistijden op het wegennet.

Voertuigtype	TS	RV	HV	WO
Aantal beschikbaar	21	9	3	2
Aantal incidenten	29.016	9.182	615	703
Minimum dekkingsnorm	6	6	15	15
Gemiddelde dekkingsnorm	7,98	14,68	15	15
Maximum dekkingsnorm	10	15	15	15

Tabel 1 Overzicht van de gebruikte dataset.

	Bedekkingsgraad per voertuigtype				
# wijzigingen	TS	RV	HV	WO	**Totaal**
0	87,68	98,23	96,84	88,64	90,83
1	89,99	98,23	96,84	88,64	92,29
2	91,76	99,64	96,84	88,64	93,74
3	93,20	99,64	97,27	89,78	94,76
4	94,38	99,64	96,84	90,68	95,53
Ongelimiteerd	98,62	99,86	98,10	93,37	98,53

Tabel 2 Dekkingsgraad bij verschillend aantal wijzigingen (in procenten).

Figuur 3 Illustratie van fictieve verplaatsingen in de Regionale Eenheid Amsterdam.

Voor voertuigtypes TS en RV gelden wettelijke dekkingsnormen, die variëren tussen 5, 6, 8 en 10 minuten. afhankelijk van het type en de functie van de gebouwen bij de verschillende vraagpunten. Voor type HV geldt een maximale dekkingsnorm van 15 minuten, voor type WO is dit gesteld op 18 minuten. In onze berekeningen zetten we de dekkingsnorm op 15 minuten voor elk vraagpunt voor beide voertuigen. De pre-trip delay is 150 seconden. Tabel 1 geeft een overzicht over de gebruikte data.

Het hierboven beschreven model stelt ons in staat goede locaties van de standplaatsen te berekenen voor een willekeurig gegeven aantal potentiële standplaatsen. Daarnaast bestaat de mogelijkheid, gegeven de negentien kazernes die op dit moment in gebruik zijn, om te bepalen wat het zou opleveren om een gelimiteerd aantal kazernes te wijzigen. Als we één wijziging toestaan bestaat de oplossing uit achttien van de huidige kazernes en één kazerne is mogelijk anders. Tabel 2 laat per voertuigtype de dekkingsgraad zien bij een verschillend aantal wijzigingen. Eén wijziging levert al een winst op van 1,46 procentpunt, de twee daaropvolgende eveneens ook telkens ruim één procentpunt. De situatie bij vier wijzigingen is weergegeven in Figuur 3.

De resultaten in Tabel 2 laten zien dat grote winsten in de dekkingsgraad te behalen zijn door het wijzigen van een klein aantal kazernelocaties en het optimaal verdelen van de voertuigen over de kazernes.

Dynamisch Ambulance Management

In de traditionele ambulancedienstverlening heeft elk ambulancevoertuig een vaste standplaats: elke wagen rijdt, wanneer deze vrijkomt, terug naar zijn *eigen* thuisbasis. Tegenwoordig wordt dit echter vaak gezien als een verouderde aanpak, en beseft men dat het terugsturen van de ambulance naar zijn thuisbasis niet altijd de beste keuze is. Afhankelijk van de situatie op dat moment, bijvoorbeeld de locaties en status van de huidige incidenten en van de andere voertuigen, kan het verplaatsen naar een andere basis beter zijn. Dit is de kern van het *Dynamisch Ambulance Management*, kortweg DAM.

In Nederland is een verandering in paradigma merkbaar: steeds vaker wordt de statische planning voor een, al dan niet door de ambulancevoorzieningen zelf ontworpen, dynamische planning ingeruild. Meestal wordt daarbij gebruik gemaakt van een eenvoudige *schuifregeltabel*, die voor elk aantal wagens aangeeft op welke basislocaties zij zich dienen te bevinden. De schuifregeltabel kan leiden tot *voorwaardescheppende ritten*, kortweg VWS-ritten, waarbij één of meer ambulancevoertuigen naar een andere locatie worden gedirigeerd.

Het grote voordeel van DAM is dat door de proactieve verplaatsingen van voertuigen veel flexibeler kan worden geanticipeerd op toekomstige incidenten, waardoor de responstijden, en daarmee ook de mortaliteit en morbiditeit, kunnen worden gereduceerd. Een veelgenoemd nadeel van DAM is dat ambulances vaker moeten verplaatsen, en dat de werkdruk daardoor wordt verhoogt. Bovendien geldt voor gebieden met veel incidenten dat het aantal vrije ambulances zeer snel kan wijzigen, vaak zo snel dat de situatie al wijzigt voordat de ambulances de nieuwe gewenste configuratie hebben bereikt.

De schuifregeltabellen zijn gebaseerd op *enkelvoudige dekking* van de regio, waarbij ambulances zodanig zijn verspreid over de regio dat de kans dat het *eerstvolgende* incident binnen de vijftienminutenstraal is geoptimaliseerd. Het probleem is echter dat in de praktijk in veel gevallen meerdere incidenten tegelijkertijd moeten worden afgehandeld, waardoor de 'statische' schuifregeltabel niet goed functioneert: de schuifregels houden geen rekening met de actuele locaties van de voertuigen, en voorzien niet in een meervoudige dekking.

Om dat probleem op te lossen wordt binnen het REPRO-project onderzoek verricht naar *online optimaliseringsalgoritmen*, waarbij optimale relocaties worden bepaald, bijvoorbeeld voor iedere mogelijke 'toestand' van het systeem (zoals incident- en voertuiglocaties). Het probleem hierbij is dat het aantal mogelijke toestanden van het systeem exponentieel snel stijgt in het aantal voertuigen en incidentlocaties, waardoor de rekentijden voor het bepalen van 'optimale' relocaties van voertuigen veel te lang worden voor praktische toepassingen. Uit onderzoek binnen REPRO blijkt dat het mogelijk is om online optimaliseringsalgoritmen te ontwikkelen, die in real-time bijna-optimale relocatieadviezen geven. Een voorbeeld is het *Dynamic Maximum Expected Coverage Location Problem* (D-MEXCLP)-algoritme, geïnspireerd door het MEXCLP-model [8], waarin de bezettingsgraad q wordt meegenomen vroor optimaliseren van meervoudige dekking.

Het D-MEXCLP-algoritme [14] beperkt zich in eerste instantie tot het verplaatsen van een ambulance wanneer deze zich vrij meldt. Hierbij wordt slechts een beperkt deel van de real-time data meegenomen, namelijk enkel de *bestemmingen* van andere vrije ambulances. Dit wordt afgezet tegen de verwachte behoefte aan ambulances.

We bepalen de gewenste locatie voor de zojuist vrijgekomen ambulance door te berekenen op welke standplaats de ambulance de *grootste toename in bedekking* zou geven volgens het MEXCLP-model. Voor de notatie, zie de eerste paragraaf. Het model definieert de waarde van het toevoegen van een k-de ambulance die vraagpunt $j \in V$ op tijd kan bereiken als $d_j(1-q)q^{k-1}$. De zo gedefinïeerde strategie $\pi(\underline{n})$, met $\underline{n} = (n_1, \ldots, n_{|W|})$, kan genoteerd worden aan de hand van het aantal vrije ambulances n_x dat bestemming x heeft. Merk op dat vrije ambulances altijd stilstaan op een basis, of daar naar onderweg zijn, dus het volstaat om $x \in W$ te beschouwen. Voor gegeven voertuigconfiguratie $\underline{n}$ stuurt het D-MEXCLP-algoritme een vrijkomende ambulance naar de bestemming $w \in W$ waarvoor de verwachte toename in de totale dekkingsgraad

$$\sum_{j \in V} d_j(1-q)q^{k(j,w,\underline{n})-1}$$

maximaal is, waarbij

$$k(j,w,\underline{n}) := \sum_{i=1}^{|W|} n_i \cdot \mathbf{1}(t_{ij} \leq r) + \mathbf{1}(t_{wj} \leq r),$$

het aantal voertuigen dat knoop j kan bereiken binnen r minuten, voor gegeven voertuigconfiguratie $\underline{n}$, aannemende dat de vrijkomende ambulance instantaan op de knoop van bestemming $w \in W$ is.

Doordat de keuzemogelijkheden beperkt zijn (in de praktijk is $|W|$ meestal niet groter dan 30 of 40), kunnen we de optimale

bestemming 'brute force' uitrekenen. De D-MEXCLP-oplossing is uitgetest voor een realistische dataset van de provincie Utrecht. De resultaten laten zien dat door het gebruik van D-MEXCLP het aantal tijdsoverschrijdingen met ongeveer $15–20$ procent kan worden gereduceerd ten opzichte van de optimale statische oplossing [14].

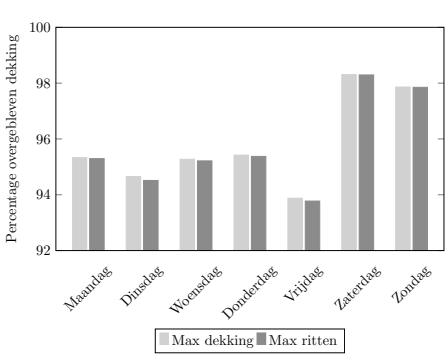
De modellen voor DAM worden momenteel uitgetest in een pilot die wordt uitgevoerd in samenwerking met GGD Flevoland, CityGIS en het CWI.

Besteld vervoer

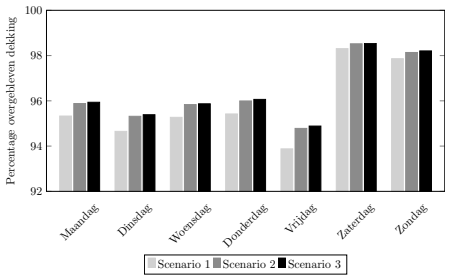
Ambulances worden naast spoedvervoer ook gebruikt voor besteld vervoer. Bij besteld vervoer wordt er een afspraak gemaakt voor vervoer tussen het woon- of verblijfadres van een patiënt of zorginstellingen. In het geval van een behandeling of onderzoek in een ziekenhuis wordt deze afspraak vaak ruim van tevoren gemaakt. Bij ontslag uit of opname in het ziekenhuis wordt deze afspraak pas kort van tevoren gemaakt. Dit betekent dat een deel van de afspraken ruim op tijd kan worden ingepland, maar dat deze planning verstoord wordt door afspraken die later binnenkomen. Hierdoor moet deze planning gedurende de dag worden aangepast zodat alle afspraken kunnen worden nagekomen.

Als er een afspraak voor besteld vervoer gemaakt wordt, wordt ook een gewenste ophaaltijd van de patiënt doorgegeven. Bij het maken van de planning wordt zoveel mogelijk rekening gehouden met deze gewenste tijd, maar hier kan niet altijd aan worden voldaan. Vaak proberen ambulancediensten in het tijdsinterval van een uur voor en een uur na de gewenste tijd de patiënt op te halen.

De meeste ambulancediensten in Nederland maken gebruik van verschillende soorten ambulances, waaronder *Advanced Life Support* (ALS)-ambulances en *Basic Life Support* (BLS)-ambulances. Voor spoedvervoer kunnen alleen ALS-ambulances worden ingezet welke bemand worden door een ambulancechauffeur en een ambulanceverpleegkundige. Voor besteld vervoer kunnen zowel ALS- en BLS-ambulances worden ingezet waarbij een BLS-ambulance eenvoudiger uitgerust is dan een ALS-ambulance en wordt bemand door een chauffeur en een basisverpleegkundige. Voor sommige patiënten is de medische indicatie zodanig dat een ALS-ambulance nodig is bij besteld vervoer. Als de patiënt geen ALS-ambulance nodig heeft probeert men zoveel mogelijk besteld vervoer uit te voeren met BLS-ambulances, maar soms is deze capaciteit niet toereikend. Omdat voldoende ALS-ambulances nodig zijn om een korte responstijd voor spoedvervoer te garanderen, is het noodzakelijk om het gebruik van ALS-ambulances voor besteld vervoer goed te plannen. Dit betekent dat het bepalen van de routes voor de BLS-ambulances en de inzet van ALS-ambulances voor besteld vervoer op zo'n manier moet gebeuren dat de verwachte dekking van de ALS-ambulances voor spoedvervoer zo hoog mogelijk blijft.



Figuur 4 Overgebleven verwachte dekking voor spoedvervoer.



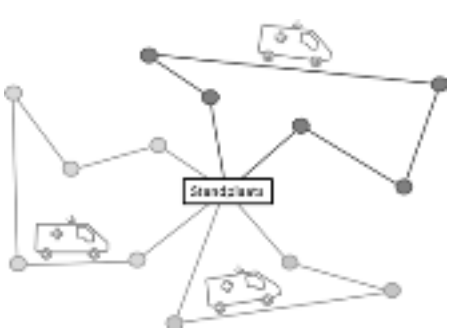
Figuur 5 Overgebleven verwachte dekking voor spoedvervoer.

Aangezien een groot deel van de afspraken pas gedurende de dag wordt gemaakt, hebben we een *geheeltallig lineair optimaliseringsprobleem* geformuleerd [4] dat wordt opgelost zodra er een nieuwe afspraak binnenkomt. Dit model neemt alle tot dan toe beschikbare informatie mee en bepaalt de planning voor het besteld vervoer. De planning van de ritten die afgerond of bezig zijn staat vast, maar de planning van toekomstige afspraken mag worden gewijzigd. De doelfunctie van het model is het maximaliseren van de overgebleven dekking voor spoedvervoer. Het model is ingewikkeld en we geven daarom hier een beschrijving van de beslissingen die volgen uit het oplossen van het model:

- Toewijzing van besteld vervoer aan een ALS- of BLS-ambulance.
- Bij gebruik van ALS-ambulance voor besteld vervoer: wijs de afspraak toe aan een bepaalde standplaats en een bepaalde tijd.
- Bij gebruik van BLS-ambulance voor besteld vervoer: bepaal de routes voor de verschillende BLS-ambulances waarbij rekening wordt gehouden met de gewenste ophaaltijd van de patiënt.

Een voor de hand liggende aanpak is om het aantal ritten uitgevoerd met een BLS-ambulance te maximaliseren om zo de overgebleven dekking voor spoedvervoer te maximaliseren. Een andere aanpak is om de dekking expliciet mee te nemen, bijvoorbeeld met behulp van de doelfunctie van MEXCLP [8]. Om deze twee aanpakken met elkaar te vergelijken, kijken we naar de overgebleven dekking voor spoedvervoer als percentage van de dekking wanneer er geen besteld vervoer wordt uitgevoerd met ALS-ambulances. De berekeningen zijn gemaakt voor RAV Utrecht. Figuur 4 laat zien dat het expliciet meenemen van de dekking resulteert in een hogere overgebleven dekking, terwijl het aantal ritten uitgevoerd met een BLS-ambulance afneemt van $91{,}7$ procent naar $90{,}2$ procent vergeleken met het maximaliseren van het aantal met BLS uitgevoerde ritten.

Figuur 5 laat zien wat de invloed op de overgebleven dekking is van het feit dat niet alle afspraken 's ochtends bekend zijn. Hiertoe bekijken we drie verschillende scenario's. Het eerste scenario is de reële setting waarbij afspraken gedurende de dag bekend worden en daarna pas kunnen worden ingepland. Bij dit scenario kan $90{,}2$ procent van de afspraken met een BLS-ambulance worden uitgevoerd. Het tweede scenario is de situatie waarin we veronderstellen dat alle afspraken 's morgens al bekend zijn en dus al ingepland kunnen worden. In dit scenario zit nog wel de restrictie dat afspraken pas na het tijdstip waarop ze aangevraagd worden mogen worden uitgevoerd. Dit betekent dat we alsnog flexibiliteit verliezen omdat het voor kan komen dat een afspraak niet een uur voor de gewenste ophaaltijd kan worden ingepland omdat de rit toen nog niet was aangevraagd.



Figuur 6

Doordat er wordt verondersteld dat alle afspraken 's ochtends al bekend zijn, neemt het percentage ritten uitgevoerd met een BLS-ambulance toe tot $95{,}5$ procent. In het derde scenario hebben we de meeste flexibiliteit omdat in dit scenario niet alleen wordt verondersteld dat alle afspraken bekend zijn in de morgen, maar dat we ook voor alle afspraken onbeperkte flexibiliteit hebben. Dit betekent dat alle afspraken in het interval van een uur voor tot een uur na de gewenste ophaaltijd kunnen worden ingepland. Dit verhoogt het aantal ritten dat uitgevoerd kan worden met een BLS-ambulance tot $97{,}1$ procent.

Uit Figuur 5 blijkt dat het eerder maken van afspraken een grote invloed heeft op de overgebleven dekking van ALS-ambulances voor spoedvervoer. Deze resultaten kunnen door de ambulancediensten gebruikt worden om de ziekenhuizen te overtuigen om eerder te bellen als er een ontslag of opname gepland staat. Een volgende stap is om de gebruikte oplossingsmethode te implementeren in de praktijk om zo de efficiëntie van de BLS-ambulances en daarmee de overgebleven dekking voor spoedvervoer te verhogen. Ons model is tot stand gekomen na intensieve samenwerking met RAV Utrecht.

Simulatie

Aanbieders van ambulancediensten moeten veel strategische beslissingen nemen. "Hoeveel beter presteert de ambulancedienst in een gebied als we daar één extra ambulance inzetten?", "Wat is de invloed door het toevoegen van de beste plek voor een nieuwe standplaats die uit locatieproblemen volgt?" of "Wat is nu echt de toegevoegde waarde van het introduceren van DAM-algoritmen op de meldkamer?" Dit soort vragen kunnen vaak beantwoord worden met behulp van simulaties.

In het kader van REPRO is een simulatiemodel ontwikkeld onder de naam *Testing Interface For Ambulance Research*, kortweg TIFAR [5]. In deze sectie wordt de werking van TIFAR toegelicht. Eerst kijken we naar de gegevens die we als in- en uitvoer gebruiken, en daarna beschrijven we de diverse objecten van het simulatiemodel en hoe die samen werken.

Vanwege onze intensieve samenwerking met verscheidene ambulancediensten in Nederland zijn er veel bronnen beschikbaar gesteld die gebruikt kunnen worden als *invoer* van het simulatiemodel. Een van de belangrijkste bronnen voor de nulmeting zijn de coördinaten van de huidige standplaatsen en de huidige ambulance- en meldkamerroosters. Door het meenemen van de roosters wordt op elk tijdstip de juiste hoeveelheden ambulances in de berekening meegenomen.

De *zorgvraag* wordt geaggregeerd op vier positie postcodeniveau (4pp); dit zijn de cijfers van het postcodegebied waarin het incident plaatsheeft. Voor elk postcodegebied weten we de coördinaten van de centroïde. Het RIVM heeft een rijtijdentabel beschikbaar gesteld die aangeeft hoe lang de reistijd voor ambulances is tussen elk koppel van 4pp-centroïdes bij spoedvervoer in de spits. Deze tabel is nuttig bij het berekenen van locatieproblemen, en DAM-algoritmes die onderdeel uitmaken van relocatievoorstellen. CityGIS levert een andere *routeplanner* die speciaal gemaakt is voor ambulancediensten. Met deze routeplanner valt de reistijd, de afstand en het traject tussen twee coördinaten op de kaart op te vragen. Dat is handig, want in de regel bevinden ambulances zich niet op de centroïde van een postcodegebied. Deze twee routeplanners zijn speciaal afgestemd op ambulancezorg, en houden rekening met onder meer aangepaste snelheid, busbanen, fietspaden en speciaal voor hulpdiensten aangelegde afritten.

Naast deze geografische eigenschappen en roosters hebben de ambulancediensten de logistieke keuzes van de meldkamercentralisten en ambulances in beslisregels gevat en beschikbaar gesteld. Hierover wordt later in de tekst nog teruggekomen.

De laatste en waarschijnlijk belangrijkste dataset zijn de *historische ritgegevens*. Vanaf 2008 weten we van elke rit de urgentie die toegekend is, het ambulancenummer, de vertrekplaats van de ambulance, de incidentlocatie, de bezorglocatie en de standplaats waarnaar de ambulance teruggekeerd is. Deze dataset bevat ook de tijdstippen, rijtijden, en behandel- en bezorgduur voor elke rit. Vrijwel alle parameters die nodig zijn om modellen te vullen vallen hieruit te halen, in het bijzonder de zorgvraag per 4pp. Merk op dat een alternatief hiervoor dat beter kan zijn bij groeiende steden, zijn demografische voorspellingen door het CBS. Uit de ritgegevens worden ook de prestatiematen berekend, waaronder de fractie A1-ritten met een responstijd van onder de vijftien minuten.

De *uitvoer* van de simulatiesoftware is een nieuwe ritdatabase die precies dezelfde structuur heeft als de historische ritdatabase. Hierdoor kunnen de simulatieresultaten goed vergeleken worden met historische prestaties. Het is zelfs mogelijk om naderhand andere prestatiematen te introduceren en vergelijken, wat vooral voordelen heeft in de communicatie met managers die soms aanvullende informatie vragen.

Figuur 7 Illustratie van het gebruikersinterface van simulatie-tool TIFAR.

De *werking* van het simulatiemodel is als volgt. Initieel worden alle ambulances leeg op hun standplaats gezet en zijn er geen incidenten in het systeem. Vanaf hier wordt een *discrete event simulation* (DES) uitgevoerd. De computer houdt een geordende lijst bij met alle tijdstippen waarop een beslissing genomen moet worden. Vervolgens wordt van een beslismoment naar het eerstvolgende beslismoment vooruit in de tijd gesprongen, en alleen op deze discrete tijdstippen wordt de toestand van het systeem herberekend. Een herberekening bestaat uit de volgende opeenvolgende stappen: incidentgeneratie, ritvernieuwing, meldkamervernieuwing, ambulancevernieuwing, en uiteindelijk de relocatieprocedure. Direct na het updaten van een parameter begint het hele herberekeningsproces op hetzelfde tijdstip van voor af aan. Uiteraard zijn er in de implementatie wel optimalisatieslagen gedaan om de rekentijd kort te houden. Figuur 7 geeft een illustratie van de gebruikersinterface van TIFAR, waarbij een proactieve relocatie wordt voorgesteld.

Een *incidentengenerator* maakt incidenten aan op 4pp's, wanneer er geen toekomstige incidenten in de wachtrij staan en voor zolang de simulatieduur nog niet verstreken is. Zowel de incidentlocatie, het tijdstip van binnenkomst, als de diverse andere tijdsduren zijn stochastisch bepaald. Hierbij is de historische ritdatabase een goede bron voor de parameterbepaling van de verdelingen.

Bij de *ritvernieuwing* worden ritten die op dat moment binnenkomen als actief aangemerkt, en afgesloten ritten in de outputdatabase weggeschreven en het bijbehorende werkgeheugen schoongemaakt.

Indien ervoor gekozen is om de *meldkamerprocessen* in de simulatie mee te nemen, worden telefoongesprekken in dit proces gekoppeld aan of ontkoppeld van de centralisten. Vaak wordt ervoor gekozen om de meldkamer niet in dit detail mee te nemen, aangezien de focus dan ligt op het wegdomein. In dat geval beperkt de meldkamer zich tot het uitgeven van een rit, en het ogenblikkelijk beantwoorden van vragen door ambulances: normaliter wordt bij een uitgifte de dichtstbijzijnde beschikbare ambulance gestuurd, gaat de ambulance naar het dichtstbijzijnde ziekenhuis, en keert ze terug naar de standplaats waar de dienst begonnen is. Later in deze tekst zullen we een meldkamermodel uitvoerig behandelen.

In de *ambulancevernieuwing* wordt de locatie van elke ambulance vernieuwd. De ambulance is verantwoordelijk voor het statussen van een rit, net als in de werkelijkheid. Dit betekent dat hij aangeeft wanneer hij aankomt op zijn bestemming, wanneer een patiënt behandeld is, of wanneer de overdracht bij de bestemming van de patiënt voltooid is. De meldkamer blijft verantwoordelijk voor het aansturen van alle logistieke processen, dus voor het nemen van dit type beslissingen neemt een ambulance altijd contact op met de meldkamer. Elke vernieuwingsslag wordt beëindigd met een *relocatieberekening*. Hiervoor kunnen onder andere DAM-methodes gebruikt worden.

In alle stappen heeft degene die de simulatie uitvoert veel vrijheid. Hierdoor kunnen dus veel vragen nauwkeurig beantwoord worden, zonder dat de ambulancedienst eerst geheel op een andere werkwijze hoeft over te schaken. Een keuze die gemaakt kan worden is om een simulatie *trace-driven* uit te voeren; in dat geval gebeuren de incidenten op dezelfde tijdstippen als in de historie, waarbij ook de behandelduur, overdrachtstijd en bestemming uit de historische ritdatabase gehaald wordt. Vergelijken met een historische nulmeting is goed voor een duidelijke beeldvorming omtrent de implicaties van het wijzigen van een beleidsregel.

Simulatie is een beproefde methode die een belangrijke ondersteunende rol in een zorgvuldig beslisproces vervult. Door in detail te simuleren wordt het effect van diverse beleidsaanpassingen inzichtelijk gemaakt.

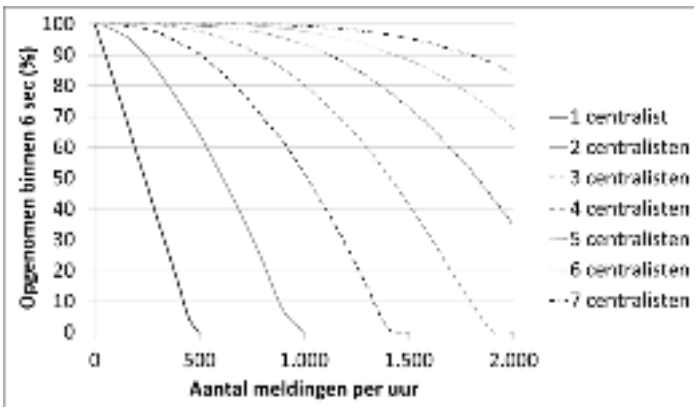
Meldkamermodel

De *meldkamer ambulancezorg* (MKA) is de plek waar de aanvragen voor ambulancezorg binnenkomen, waarvandaan de voertuigen worden aangestuurd. De MKA vervult de rol als communicatieknooppunt tussen ambulanceverpleegkundigen, ziekenhuizen en andere professionals die betrokkenen zijn bij de operationele ambulancezorg. Op de meldkamer zijn er meerdere inkomende lijnen via welke ritten worden aangevraagd. Ongeveer 40 tot 50 procent van de aanvragen voor spoedeisende ambulancezorg betreft een melding via het landelijke telefoonnummer 1-1-2. Daarnaast zijn er veel aanvragen die via een huisarts lopen. Het besteld vervoer wordt door ziekenhuizen of andere zorginstellingen aangevraagd.

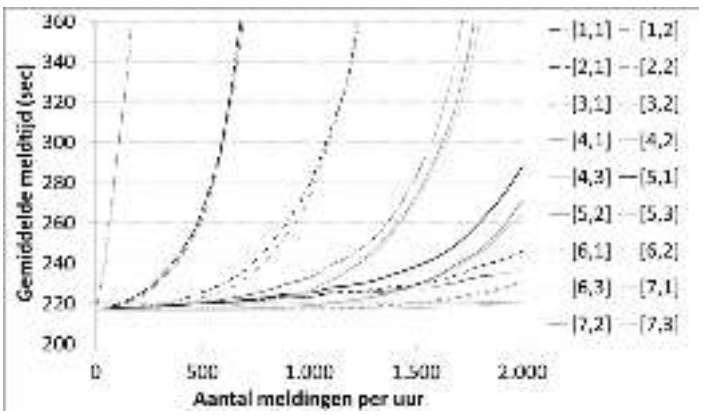
In de Nederlandse MKA's vindt sinds 2010 schaalvergroting plaats waarbij meerdere kleinere meldkamers samengaan tot een grotere organisatie. Ook hebben nieuwe triage-systemen hun intrede gedaan waardoor de uitvraag en beoordeling van spoedeisende ambulancezorg meer geprotocolleerd plaatsvindt. Door deze ontwikkelingen is er een behoefte gekomen aan een simulatiemodel dat de processen op de meldkamer nabootst en waarmee het presteren van de MKA kan worden onderzocht. Ons simulatiemodel TIFAR kan gebruikt worden om de processen op de meldkamer te simuleren. Het model omvat de processen rondom de aanvragen en uitgiftes voor zowel spoedeisende ambulancezorg als besteld vervoer.

Het voert te ver om het model in detail te bespreken. Daarom geven we in deze paragraaf een kort overzicht van de belangrijkste elementen in het binnen REPRO ontwikkelde

Foto: Till Krech



Figuur 8 Aantal meldingen binnen zes seconden bij toenemend aantal meldingen per dag.



Figuur 9 Doorlooptijd als functie van het gemiddelde aantal meldingen per dag ([centralisten, uitgifte-centralisten]).

MKA-model. De geïnteresseerde lezer wordt verwezen naar [6] voor meer details.

Op een MKA werken *aanname- en uitgifte-centralisten*. Een aanvraag voor ambulancezorg wordt door een aanname-centralist beoordeeld op urgentie, en bij een indicatie voor het verlenen van ambulancezorg wordt een ritopdracht uitgegeven door een uitgifte-centralist. Deze geeft het meest geschikte ambulanceteam de opdracht om naar het incident te rijden, ter plaatse zorg te verlenen en, indien nodig, vervoer naar een ziekenhuis te bieden. Veelal is dit het ambulanceteam dat het snelste ter plaatse kan zijn.

In het geval van een spoedeisende inzet gaat een ambulance met optische en geluidssignalen (OGS) naar de locatie van het incident en verleent daar levensreddende zorg. Beide centralisten kunnen op meerdere momenten tijdens de inzet communiceren: de aanname-centralist met de melder, bijvoorbeeld om instructie voor eerste hulp te geven of om de patiënt in veiligheid te brengen. De uitgifte-centralist kan een contact hebben om vragen van het ambulanceteam te beantwoorden of de komst van de patiënt met het ziekenhuis te communiceren. We noemen deze vervolgcontacten van de uitgifte-centralist 'terugkoppelmomenten'.

Indien het druk wordt op de meldkamer, wordt er geprioriteerd. Voor de aanname-centralist hebben de 1-1-2-aanvragen en spoedeisende aanvragen van huisartsen de hoogste prioriteit aangezien het dan om een levensbedreigende situatie kan gaan. Het besteld vervoer heeft de laagste urgentie omdat deze aanvragen vaak niet direct respons nodig hebben. We gaan er in ons model van uit dat alle taken 'geduldig' zijn en het systeem niet zullen verlaten tot ze afgehandeld zijn. In werkelijkheid kan een brandweer- of politie-centralist een aanvraag voor ambulancezorg oppakken als alle ambulance-centralisten bezet zijn. Voor de uitgifte-centralist heeft het uitgeven van spoedritten een hogere prioriteit dan de terugkoppelmomenten. Bij gelijke prioriteit hebben langer wachtenden altijd voorrang.

Het MKA-model onderscheidt zeven verschillende inkomende lijnen voor ambulancezorg. Afhankelijk van de lijn heeft de aanvraag een hoge, middelmatige of lage prioriteit. Gedurende de loop van het gesprek bepaalt de aanname-centralist of ambulancezorg gewenst is en kent aan de melding een urgentie toe. In sommige gevallen is de urgentie a priori bekend doordat de inkomende lijn alleen voor een bepaald soort meldingen gebruikt wordt. We veronderstellen dat nieuwe aanvragen middels een Poisson-proces optreden. De frequentie $\lambda_{i,u}$ en gespreksduur $\mu_{i,u}$ van de aanvragen zijn afhankelijk van de binnenkomende lijn i en urgentie u, en kunnen ook in de tijd variëren. Tijdens het uitvoeren van de inzet door het ambulanceteam zijn er verschillende contactmomenten met de meldkamer. We gaan ervan uit dat er maximaal één contactmoment is bij iedere status van de ambulance, waaronder aanrijdend, behandelend, bezorgend en overdragend. Het contactmoment laten we altijd plaatsvinden aan het einde van de status. Door deze keuzes wordt de kans op een contactmoment tijdens elke status een Bernouilli-proces, waarinn we uitgaan van aannames van meldkamerexperts voor de parameterwaarden. De verdeling voor de duur van elke status en bijbehorende parameters worden verkregen uit de ritdatabase.

De bedieningsduurverdelingen zijn geschat aan de hand van een dataset die is opgebouwd uit een koppeling tussen gegevens uit de telefooncentrale van de meldkamer, de arbi-data, en de ritgegevens van de ambulance-inzetten. De arbi-data bevatten geen kenmerken van de inzetten, zoals de aanvrager van de inzet en of bij een aanvraag daadwerkelijk een inzet heeft plaatsgevonden. Gegevens over de periode april–juni 2012 waren beschikbaar voor ons onderzoek; dit betrof gegevens van 109.000 telefonische contacten waaruit 41.000 ambulance-inzetten hebben plaatsgevonden.

Met de uitkomsten van het simulatiemodel kan het functioneren van de meldkamer worden geanalyseerd onder verschillende werklasten en met verschillende centralisten-aantallen. Belangrijke prestatie-indicatoren hierbij zijn de wachttijd voordat de telefoon wordt opgenomen, de totale doorlooptijd van de meldkamer voor elk van de verschillende soorten telefoongesprekken, en de tijd die nodig is voor de uitgifte van een inzet.

Ter illustratie is het simulatiemodel TIFAR gebruikt om het presteren van een meldkamer te analyseren bij een veranderende werkdruk. Het aantal aanvragen is stapsgewijs verhoogd van $\lambda = 100$ naar 2000 aanvragen per dag, waarbij de verhouding tussen het aantal aanvragen dat elke inkomende lijn levert gelijk blijft. Hierbij is per λ in een aantal stappen het aantal centralisten verhoogd. Bij een vast aantal centralisten en een toenemend aantal aanvragen neemt de wachttijd voor de aanvrager toe: deze zal vaker in een wachtrij komen en de melding zal niet direct kunnen worden beantwoord. Als vuistregel geldt een prestatienorm dat een binnenkomende 1-1-2-aanvraag binnen zes seconden moet worden opgenomen. Wij hanteren dat tenminste 95 procent van de hulpverzoeken aan deze aanvragen moet voldoen.

Door de scheiding in functie ondervinden de aanname-centralisten geen nadeel door eventuele vertragingen bij uitgifte-centralisten. In Figuur 8 zien we dat tussen de 50 en 700 meldingen per dag twee aanname-centralisten voldoende zijn, terwijl je tot 1600

meldingen per dag met drie aankan. Hierin zijn schaalvoordelen al zichtbaar.

De meldkamer-*doorlooptijd* van een aanvraag is de tijdsduur vanaf het opnemen van een telefoon tot het moment van een opdracht aan een ambulance. De doorlooptijd is sterk afhankelijk van het aantal centralisten en de verhouding tussen het aantal aanname- en het aantal uitgifte-centralisten. Bij een gegeven λ kan gezocht worden naar het optimale aantal centralisten waarbij de doorlooptijd nog acceptabel is. In Figuur 9 zien we dat tot ongeveer 1200 meldingen per dag één uitgifte-centralist voldoende zal zijn. Dit valt te zien doordat de uitgifte-centralist hier niet de beperkende factor is op de doorlooptijd. Tot 2000 meldingen per dag zien we dat een derde niet nodig is.

Door enkele keuzes in het modelleren wordt er een minimale doorlooptijd van 220 seconden weergegeven, in plaats van ongeveer twee minuten die in de praktijk geldt. Dit komt doordat enkele parallelle processen in ons model sequentieel uitgevoerd worden. Deze keuze heeft echter geen invloed op de werklast van de centralisten en daarmee de toename in de doorlooptijd, waardoor conclusies over de toename in werklast en doorloop veilig getrokken kunnen worden.

Het simulatiemodel is een krachtig middel om het presteren van een MKA te analyseren zonder dat er grootschalige organisatorische veranderingen nodig zijn. Door simulatie kunnen fictieve situaties worden doorgerekend, bijvoorbeeld de situatie dat meldkamers worden samengevoegd en het volume van aanvragen toeneemt. De relatie tussen het aantal aanvragen en het aantal benodigde centralisten kan inzichtelijk gemaakt worden.

Referenties

1 L. Aboueljinane, E. Sahin en Z. Jemaï, A review on simulation models applied to emergency medical service operations, *Computers & Industrial Engineering* 66(4) (2013), 734–750.

2 R. Alanis, A. Ingolfsson en B. Kolfal, A Markov chain model for an EMS system with repositioning, *Production and Operations Management* 22(1) (2013), 216–231.

3 P.L. van den Berg en K.I. Aardal, Time-dependent MEXCLP with start-up and relocation cost, *European Journal of Operational Research* 242(2) (2015), 383–389.

4 P.L. van den Berg en J.T. van Essen, Scheduling non-urgent patient transport while maximizing emergency coverage, submitted, 2015.

5 M. van Buuren, R.D. van der Mei, K.I. Aardal en H. Post, Evaluating Dynamic Dispatch Strategies for Emergency Medical Services: TIFAR Simulation Tool, in *Proceedings of the Winter Simulation Conference*, 2012.

6 M. van Buuren, G.J. Kommer, R.D. van der Mei en S. Bhulai, A simulation model for Emergency Medical Service call centers, te verschijnen in *Proceedings Winter Simulation Congress, Los Angeles*, 2015.

7 R. Church en C. S. ReVelle, The maximal coverage location problem, *Papers in Regional Science* 32(1) (1974), 101–118.

8 M.S. Daskin, A maximum expected covering location model: formulation, properties and heuristic solution, *Transportation Science* 17(1) (1983), 48–70.

9 M. Dzator en J. Dzator, An effective heuristic for the p-median problem with application to ambulance location, *OPSEARCH* 50(1) (2013), 60–74.

10 E. Erkut, A. Ingolfsson en G. Erdoğan, Ambulance location for maximum survival, *Naval Research Logistics* 55(1) (2008), 42–58.

11 M. Gendreau, G. Laporte en F. Semet, Solving an ambulance location model by tabu search, *Location Science* 5(2) (1997), 75–88.

12 K. Hogan en C.S. ReVelle, Concepts and applications of backup coverage, *Management Science* 32(11) (1986), 1434–1444.

13 A. Ingolfsson, S. Budge en E. Erkut, Optimal ambulance location with random delays and travel times, *Health Care Management Science* 11(3) (2008), 262–274.

14 C.J. Jagtenberg, S. Bhulai en R.D. van der Mei, An efficient heuristic for real-time ambulance redeployment, *Operations Research for Health Care* 4 (2015), 27–35.

15 V.A. Knight, P.R. Harper en L. Smith, Ambulance allocation for maximal survival with heterogeneous outcome measures, *Omega* 40(6) (2012), 918–926.

16 X. Li, Z. Zhao, X. Zhu en T. Wyatt, Covering models and optimization techniques for emergency response facility location and planning: a review, *Mathematical Methods of Operations Research* 74(3) (2011), 281–310.

17 M.S. Maxwell, S.G. Henderson en H. Topaloglu, Tuning approximate dynamic programming policies for ambulance redeployment via direct search, *Stochastic Systems* 3(2) (2013), 322–361.

18 M.S. Maxwell, M. Restrepo, S.G. Henderson en H. Topaloglu, Covering models and optimization techniques for emergency response facility location and planning: a review, *Mathematical Methods of Operations Research* 74(3) (2011), 281–310.

19 W.B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionalit*, John Wiley & Sons, 2011.

20 J.F. Repede en J.J. Bernardo, Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky, *European Journal of Operational Research* 75(3) (1994), 567–581.

21 C.S. ReVelle en K. Hogan, The maximum availability location problem, *Transportation Science* 23(3) (1989), 192.

22 V. Schmid en K.F. Doerner, Ambulance location and relocation problems with time-dependent travel times, *European Journal of Operational Research* 207(3) (2010), 1293–1303.

23 O. Zhang, A.J. Mason en A.B. Philpott, Simulation and optimisation for ambulance logistics and relocation, presentatie op *INFORMS 2008 Conference*.

Onno Boxma
Department of Mathematics and Computer Science
Eindhoven University of Technology
o.j.boxma@tue.nl

Stella Kapodistria
Department of Mathematics and Computer Science
Eindhoven University of Technology
s.kapodistria@tue.nl

Michel Mandjes
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
m.r.h.mandjes@uva.nl

Performance analysis of stochastic networks

Queues are everywhere, and have significant impact on how we experience everyday's life. The mathematical analysis of queueing, rooted in the interface of probability theory and operations research, is a strongly developed branch of research. Onno Boxma, Stella Kapodistria, Michel Mandjes give an overview.

In 1909 the Danish mathematician Agner Krarup Erlang published the paper 'The theory of probabilities and telephone conversations' [16]. In this paper, which is commonly viewed as the birth of queueing theory, Erlang studied dimensioning issues for traditional circuit-switched telephone systems. More specifically, a procedure was developed to determine the number of telephone lines which are needed between two villages so that the probability that, at some random time epoch, all lines are simultaneously busy is less than some specified small number.

The essential feature of Erlang's model, and of queueing theory in general, is that there are *customers* who are competing for access to a *scarce resource*. In his model there was no waiting — if all lines are busy, a newly incoming call is 'lost'. One comes across many situations in which models of this type apply, for instance in the context of wireless communication [6] and computer science [26]. One can, however, also think of variants in which customers who cannot be accommodated directly are sent to a 'waiting room', thus forming a genuine queue. Examples abound; one could think of the checkout counter of a supermarket, an elevator, a traffic light intersection, a machine that produces parts, a computer processor processing jobs, or a communication channel with a buffer for packets which still need to be transmitted. In some situations customers are initially willing to wait, but might become impatient at some

Figure 1 Queueing for on-the-day tickets at Wimbledon.

point — think for instance of the customers of a call center.

It is far from an easy task to model such a wide range of situations in which customers compete for access to a scarce resource. One has to model the service facility (the number of servers, their service speeds, the assignment of priorities, the size of waiting room, et cetera) as well as the customer behavior (the arrival process of customers at the service facility, the service requirements of the individual customers, the amount of patience they have, the choice which server to join, et cetera). Still, in the century following Erlang's pioneering work, queueing theory has been remarkably successful in capturing the essential features of the congestion phenomena of a staggeringly wide range of extremely complicated real-life systems with relatively simple models — models which have shown to lend themselves to a detailed mathematical analysis. It has resulted in a set of techniques with which accurate predictions can be made of the global behavior of intricate stochastic systems, and which facilitate their optimization and control.

To a considerable extent, the success of queueing theory is due to the fact that one can distinguish a few basic building blocks, which have been studied in much detail and which time and again pop up in the analysis of new congestion phenomena. For instance, with the advent of wireless communications, sensor networks, and peer-to-peer networks, queueing models could be used to describe their performance. The building blocks most frequently used are the *Erlang loss system* (the system studied by Erlang in 1909; a system without queueing, calls being lost when all lines are busy) and the *single server queue*. We shall describe the latter system in some detail, as it also plays a crucial role in the product-form networks that we shall discuss in the next section.

The single server queue

Customers arrive at a service facility, where they would like to receive a certain amount of service. There is a single server, who serves customers in order of arrival (that is, First-Come-First-Served, usually abbreviated to FCFS). If a customer can not immediately be served, then it joins a queue, and waits patiently until its turn comes. The waiting room is assumed to have infinite capacity. The interarrival times of customers, and also the required service times, are assumed to be random variables. This captures the fact that these times are usually a priori unknown to us, and fluctuate over time. The consequence is that our main performance measures, like waiting times and queue lengths, are also random variables and that we have to settle for probabilistic statements about them. Examples of such statements are : $\mathbb{P}(W > 5) = 0.3$, i.e., the probability that an arbitrary customer waits longer than 5 minutes is 0.3; or: $\mathbb{E}(W) = 1.4$, i.e., the expectation (that is, the mean) of the waiting time equals 1.4. Now we look a bit closer at the two stochastic ingredients that we identified above, the arrival process and the service requirements.

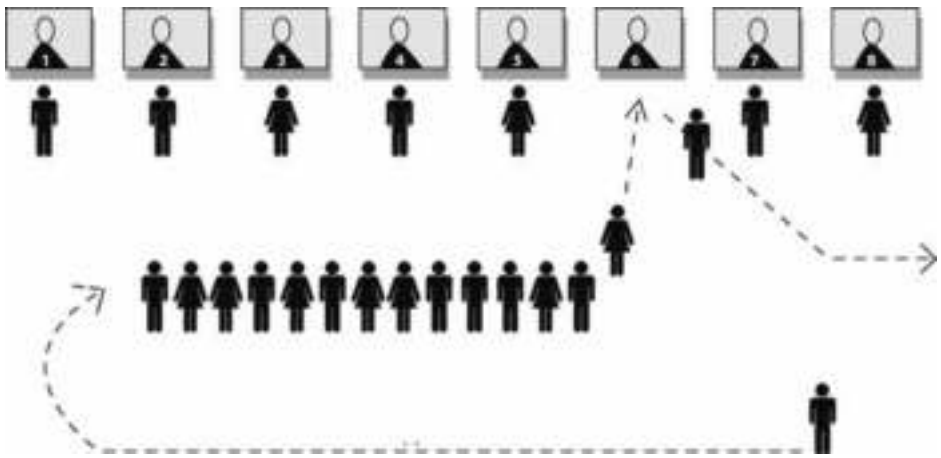


Figure 2 A schematic picture of a multi-server system $G/G/c$.

It is often assumed that the arrival process of customers is a *Poisson process*. This means that the intervals between successive arrivals are independent, identically distributed random variables, generically indicated by A, with as probability distribution the so-called *exponential distribution*:

$$\mathbb{P}(A > x) = \mathrm{e}^{-\lambda x}, \quad x \geq 0, \tag{1}$$

with λ some positive number. The parameter λ is called the arrival rate, since the mean time between two arrivals equals $1/\lambda$. The exponential distribution is unique in having the appealing *memoryless* property:

$$\begin{aligned} &\mathbb{P}(A > x + y | A > x) \\ &= \frac{\mathrm{e}^{-\lambda(x+y)}}{\mathrm{e}^{-\lambda x}} = \mathrm{e}^{-\lambda y} \\ &= \mathbb{P}(A > y), \quad \forall\, x, y \geq 0. \end{aligned} \tag{2}$$

This means that, at any arbitrary time t_0, no matter how long ago the last arrival took place, the remaining time (after this time t_0) until the next arrival is again exponentially distributed with parameter λ. The memoryless property is mathematically attractive and also quite natural for arrival intervals. It is mathematically attractive because there is no need to keep track of the time since the last arrival — it gives no information whatsoever that can help us predict the remaining time until the next arrival. In addition it is quite natural, for the following reason. In many arrival processes, like those of customers at a supermarket, hits of a website or orders at a factory, there is a huge number of *potential* customers. If we receive the information that a website has been visited five times in the last ten minutes, and that the last visit took place seventeen seconds ago, this gives us hardly any information about the behavior of all those other potential customers (and the likelihood of them arriving some time soon): the interarrival times are memoryless, and hence have to be exponential.

Now that we have had a look at the arrival process, we consider the customers' service requirements. The service requirements of successive customers are typically also assumed to be independent, identically distributed random variables. Unlike the interarrival times, there is no particular reason — apart from perhaps mathematical convenience — to assume that the service requirements follow the exponential distribution. Let that mathematical convenience prevail for the moment; assume that the service requirements are exponentially distributed with parameter (rate) μ, so with mean $1/\mu$. The above described single server queue is then called the $M/M/1$ queue. The first and second M respectively indicate that the interarrival times and the service times are Memoryless (or Markovian); the 1 indicates that there is one server (along the same lines, $M/G/c$

indicates Memoryless interarrival times, Generally distributed service requirements, and $c \in \mathbb{N}$ servers).

A very attractive feature of the $M/M/1$ queue is that the stochastic process of numbers of customers $\{X(t), t \geq 0\}$, with $X(t)$ the number of customers present (waiting plus in service) at time t, is a Markov process. The powerful machinery of Markov processes now can be used, and readily yields elegant, explicit results. Suppose we assume that $\lambda < \mu$, implying that the amount of work arriving per unit of time is smaller than the amount that can be served, thus guaranteeing that our queueing system is stable. Then it turns out that the steady-state distribution $\lim_{t\to\infty} \mathbb{P}(X(t) = n \mid X(0) = i)$ exists, and is given by the geometric distribution:

$$\mathbb{P}(X = n) = (1-\rho)\rho^n, \quad n = 0, 1, \ldots, \quad (3)$$

with $\rho := \lambda/\mu < 1$ representing the offered traffic load per time unit.

So far we have focused on single queueing facilities in isolation. In many practical contexts, however, the underlying stochastic systems can be seen as networks of multiple interrelated nodes. In the next section we consider these.

Networks of queues

Open networks of queues

Around 1950, the mathematical theory of stochastic processes had reached a certain maturity. Several monographs were published, including the landmark book of Feller [19], which not only gave a systematic discussion of a number of important stochastic processes, but also showed in a lucid way how to model many biological and physical phenomena by various stochastic processes like Markov chains and birth-and-death processes. The year 1954 saw the publication of the first study on networks of queues. R.R.P. Jackson [23] considered an $M/M/1$ queue Q_1, with arrival rate λ and service rate μ_1, and assumed that each served customer immediately enters a second single server facility Q_2, again with infinite waiting room capacity and FCFS service, and again with independent, exponentially distributed service requirements; μ_2 denotes the service rate in the downstream queue. He observed that the two-dimensional process of numbers of customers at Q_1 and Q_2, $\{(X_1(t), X_2(t)), t \geq 0\}$, again is a Markov process — now a two-dimensional one. If $\lambda < \mu_1$ *and* $\lambda < \mu_2$ then this Markov process has a steady-state (limiting) distribution, and that distribution is *unique*. Jackson guessed that the steady-state distribution is given by

$$\begin{aligned}\pi(n_1, n_2) &= \mathbb{P}(X_1 = n_1, X_2 = n_2) \\ &= (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}, \quad (4) \\ & \quad n_1, n_2 = 0, 1, \ldots,\end{aligned}$$

with $\rho_i := \lambda/\mu_i$, $i = 1, 2$. Then he set up the balance equations for this two-dimensional Markov chain: for $n_1, n_2 = 1, 2, \ldots$,

$$\begin{aligned}&(\lambda + \mu_1 + \mu_2)\pi(n_1, n_2) \\ &\quad = \lambda\,\pi(n_1 - 1, n_2) + \mu_1\,\pi(n_1 + 1, n_2 - 1) \\ &\qquad + \mu_2\,\pi(n_1, n_2 + 1), \\ &(\lambda + \mu_1)\pi(n_1, 0) \\ &\quad = \lambda\,\pi(n_1 - 1, 0) + \mu_2\,\pi(n_1, 1), \\ &(\lambda + \mu_2)\pi(0, n_2) \\ &\quad = \mu_1\,\pi(1, n_2 - 1) + \mu_2\,\pi(0, n_2 + 1), \\ &\lambda\,\pi(0, 0) = \mu_2\,\pi(0, 1).\end{aligned}$$

Probably much to his surprise, Jackson observed that (4) indeed satisfies all the balance equations, and he had actually found the unique steady-state distribution!

Jackson's results had a wide set of implications, of which we now mention a few.

i. The steady-state numbers of customers in Q_1 and Q_2 are *independent*, since the joint distribution is the product of the marginal distributions:

$$\begin{aligned}&\mathbb{P}(X_1 = n_1, X_2 = n_2) \\ &\quad = \mathbb{P}(X_1 = n_1)\mathbb{P}(X_2 = n_2), \quad (5) \\ &\qquad n_1, n_2 = 0, 1, \ldots .\end{aligned}$$

ii. Q_2 actually behaves like an $M/M/1$ queue with Poisson(λ) arrival process (as follows by summing the expression in (4) over all $n_1 = 0, 1, \ldots$).

For obvious reasons, formula (4) has become known as a *product-form* result. Triggered by the above implications, Jackson's results immediately gave rise to a frantic research effort. In 1956 Paul Burke [9], working at Bell Labs, proved what has later become known as the *Output Theorem*. This states that (i) the departure process of an $M/M/c$ queue is again a Poisson process with (if the arrival rate λ is less than c times the service rate μ) the same rate as the arrival process, and (ii) the number of customers in an $M/M/c$ queue at some arbitrary time t_0 is independent of the departure process *before* t_0.

Statement (i) immediately shows that Q_2 in Jackson's two-queue model has a Poisson arrival process, and hence indeed behaves like an $M/M/1$ queue. Statement (ii) readily implies that the steady-state numbers of customers in Q_1 and Q_2 are independent.

A year later Edgar Reich [32] gave a very simple proof of the output theorem, exploiting the observation that the queue length process in an $M/M/c$ queue is *reversible*. Intuitively speaking, a reversible process is a stochastic process with the following property: if one would take a film of such a process and run the film backwards, then the resulting process is, statistically speaking, indistinguishable from the original process. Statement (i) of the output theorem immediately follows because, in the time-reversed process, the departure process becomes the arrival process — and hence is a Poisson process. Statement (ii) of the output theorem becomes after time reversal: the number of customers at some arbitrary time t_0 is independent of the arrival process *after* t_0. The memoryless property of that (Poisson) arrival process immediately implies that the latter statement is true.

J.R. Jackson [24], inspired by the results of R.R.P. Jackson, Burke and Reich, considered the following network of N single server queues $Q_1, \ldots, Q_N$. New customers arrive at the queues according to independent Poisson processes, with rate λ_i at Q_i. Service requirements at Q_i are independent, exponentially distributed with rate μ_i, $i = 1, \ldots, N$. All servers operate under FCFS, and all waiting rooms have infinite capacity. If a customer has been served at Q_i, then it is routed to Q_j with probability p_{ij} and leaves the network with probability p_{i0}, $i, j = 1, \ldots, N$; obviously, one assumes that $\sum_j p_{ij} = 1$. All external interarrival times and service times are assumed to be independent.

Because of all the exponential, memoryless, assumptions, the process

$$\{(X_1(t), X_2(t), \ldots, X_N(t)), t \geq 0\}$$

of numbers of customers at $Q_1, \ldots, Q_N$ is a Markov process. Jackson [24] verified that the balance equations for its steady-state distribution are satisfied by

$$\begin{aligned}&\mathbb{P}(X_1 = n_1, \ldots, X_N = n_N) \\ &\quad = \prod_{i=1}^{N}(1-\rho_i)\rho_i^{n_i}, \quad n_1, \ldots, n_N = 0, 1, \ldots,\end{aligned} \quad (6)$$

with, for $i = 1, \ldots, N$: $\rho_i := \Lambda_i/\mu_i$, the offered load at Q_i, and where the so-called throughputs Λ_i are the solution of the set of equations

$$\Lambda_i = \lambda_i + \sum_{j=1}^{N} \Lambda_j p_{ji}, \quad i = 1, \ldots, N$$

(in vector-matrix notation: $\Lambda = \lambda + \Lambda P$), which turns out to be unique as a direct consequence of the Perron-Frobenius theorem. The interpretation of Λ_i is that it equals the external arrival rate λ_i plus the sum of all the internal flows going into Q_i. It can be shown that the steady-state queue length distribution exists if and only if $\rho_i < 1$ for all $i = 1, \ldots, N$.

We observe that the steady-state distribution (6) exhibits a *product form*, which again implies that in steady state the numbers of customers at the various queues are independent, and again each queue behaves like an $M/M/1$ queue in isolation. Actually, Jackson [24] believed that the $M/M/1$ behavior of each Q_i is not surprising, and can be seen as a immediate implication of the output theorem. He argued that if one merges two independent Poisson arrival processes, one obtains another Poisson process; and if one splits a Poisson process with fixed probabilities, a fraction p_{ij} entering Q_j, then the resulting processes are independent Poisson processes. However, he overlooked the fact that his routing probabilities allow the possibility of *feedback*: a customer may visit a queue where it has been before. It is easily shown that the resulting dependency destroys the Poisson property of the flows. That makes the product-form result (6) all the more remarkable: the marginal queue length distribution at Q_i is geometric, as if Q_i were an $M/M/1$ queue (cf. (3)), but its arrival process does not have to be Poisson! Thanks to the work of Kelly [29] and others, much insight has been obtained into the phenomenon that each queue in the above-described *Jackson network* behaves *as if* it is an $M/M/1$ queue. The concept of *quasi-reversibility* plays a crucial role here: a service facility is quasi-reversible if it has the property that the departure process would be a Poisson process if the arrival process were a Poisson process.

Closed networks of queues

In 1963, J.R. Jackson [25] extended the results of his paper [24] to *closed* networks of $\cdot/M/1$ (actually, $\cdot/M/c$) queues. The only changes with respect to his above-described open network were that all external arrival rates $\lambda_i \equiv 0$ and that all $p_{i0} \equiv 0$, and that the system starts with K customers. Since no customers can enter or leave, those K customers stay in the network forever. At first sight this may seem not only cruel but also artificial, but actually it may well represent, e.g., having a fixed number K of pallets in a factory, or having window flow control with window size K in a communication network (i.e., at most K packets may be transmitted without yet having received acknowledgment of receipt). Jackson [25] proved that the steady-state distribution of the numbers of customers at the various queues is once more given by a product form: for $n_1, \ldots, n_N = 0, 1, \ldots$ such that $n_1 + \cdots + n_N = K$,

$$\mathbb{P}(X_1 = n_1, \ldots, X_N = n_N) = \frac{1}{G(N,K)} \prod_{i=1}^{N} \rho_i^{n_i}, \tag{7}$$

with $\rho_i := \Lambda_i/\mu_i$ and $\Lambda_i = \sum_{j=1}^{N} \Lambda_j p_{ji}$. The quantity $G(N,K)$ is a normalizing constant, obtained by summing the numerator of (7) over all possible combinations of $(n_1, \ldots, n_N)$, and realizing that the sum over all probabilities should equal 1. Notice that the Λ_i are now determined up to a multiplicative constant (i.e., if Λ_i is a solution, then so is $a\Lambda_i$ for any scalar a). The probability $\mathbb{P}(X_1 = n_1, \ldots, X_N = n_N)$, however, still is uniquely determined. Indeed, multiplying all Λ_i by a amounts to multiplying both the numerator and denominator of (7) by a^K. It should also be observed that the product form now does *not* imply independence; in fact, the numbers of customers have an obvious dependence due to $X_1 + \cdots + X_N = K$.

Generalizations

Spurred by the elegance of the above product-form results, but also by the rapidly increasing need to study the performance of advanced computer and communication networks, a stream of papers was produced in the seventies and eighties, in which the product-form results of [23–25] were generalized. Some of the key publications are [5], [12] and [29]; several Dutch researchers have made important contributions to the field, including Boucherie, Cohen, van Dijk (who also published a monograph [15] on the topic) and Hordijk.

Thanks to all these efforts we now know that the steady-state joint queue length distribution in a small but significant class of queueing networks (open, closed, and mixed) has a product form. To mention some extensions: (i) Service facilities may have multiple servers; put differently, the service rate at a service facility may depend on the number of customers present. (ii) The service discipline at some nodes may be Last-Come-First-Served Preemptive-Resume, or Processor Sharing, instead of FCFS; here Processor Sharing is particularly relevant in a broad range of computer-communication applications. (iii) A network may have multiple classes of customers, with different routing probabilities for different classes (but not different service rates at FCFS nodes). For more details and information on the topic of queueing networks the interested reader is referred to [11, 34].

While these product-form results are of huge importance, as they allow a relatively simple performance analysis and optimization of a model that may reasonably accurately describe the behavior of a complex real-life system, they are also quite limited in the following sense. If one of the conditions for having a product-form network is violated, then most likely an exact analysis is extremely complicated, or — more often than not — completely out of reach. However, there is a class of, mainly, two-dimensional models — for example, two queues in series, or two queues and one arrival stream, customers joining the shorter queue — for which an exact analysis is possible. This is the topic we turn to in the next section. But beforehand, we first describe two interesting special systems.

Remark 1. A special case of a closed product-form network is a two-queue model with K customers, Q_1 being an infinite server system (or, equivalently, a K-server system, as that would be sufficient to prevent any waiting; mean service time is $1/\mu_1$) and Q_2 being a FCFS single server with exponentially(μ_2) distributed service times. In addition we assume that $p_{12} = p_{21} = 1$, meaning that the customers hop between both queues.

This model has become known as the *computer-terminal model*: K active terminal users alternate between a 'think mode' in which they generate a job for the central processor, and a 'wait mode' in which they stay until the processor has handled the job. It also has become known as the *machine-repair model*: K machines all alternate between an operational mode and a mode in which they are broken and stay in the repair shop Q_2, to be repaired by a single repairman.

As observed in, e.g., [5], one has a product form even if the service times in Q_1 are generally distributed. Since $n_1 + n_2 = K$, the prod-

uct form degenerates into a one-dimensional result: with $\nu := \mu_2/\mu_1$,

$$\mathbb{P}(X_1 = n_1) = \frac{\nu^{n_1}}{n_1!} \Big/ \sum_{j=0}^{K} \frac{\nu^j}{j!}, \qquad n_1 = 0, 1, \ldots, K. \tag{8}$$

Interestingly, the same distribution holds for the so-called Erlang loss system, viz., calls arrive according to a Poisson(μ_2) process at a system of K telephone lines, and the lengths of calls are generally distributed with mean $1/\mu_1$. In fact, it is not hard to see that the machine-repair model indeed is probabilistically equivalent with the Erlang loss system — a system that we introduced above as one of the basic building blocks of queueing.

Remark 2. In this second special case we consider the class of *loss networks*. In this model there are N types of customers; customers of class i arrive according to a Poisson process of rate λ_i and remain in the system during a random time with mean $1/\mu_i$. The N customer types use R resources: a type i customer uses an amount A_{ir} at resource r. There is a total amount C_r of resources of type r, entailing that a customer of type i is blocked (and therefore lost) if upon arrival the remaining amount of resources available is less than the required A_{ir}. The resulting model is usually referred to as a loss network. Clearly, the numbers of customers in this system only attain values in the polyhedron

$$H = \left\{ (n_1, \ldots, n_N) : \sum_{i=1}^{N} A_{ir} n_i \le C_r \right\}.$$

The steady-state distribution of the numbers of customers is again of product form: due to the detailed analysis in e.g. Kelly [30], with $\nu_i := \lambda_i/\mu_i$,

$$\mathbb{P}(X_1 = n_1, \ldots, X_N = n_N) = \frac{1}{G} \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!};$$

here the normalizing constant $G = G(N, C_1, \ldots, C_R)$ is given by

$$\sum_{(n_1,\ldots,n_N)\in H} \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!},$$

which can be efficiently computed using Buzen's algorithm [10]. Because of the high relevance of this type of models, a substantial research effort was spent on developing computational techniques for loss networks, culminating in the elegant recursive techniques published essentially simultaneously by Kaufman [27] and Roberts [33].

The loss network attracted substantial attention in the 1990s, where it was used in the context of *multi-service communication networks*. Till then networks were service-specific: there was a telephone network, a separate network for data traffic, et cetera. From about 1990 on, however, networks were increasingly organized in such a way that they could support multiple services over a common infrastructure — as we know it from the current internet. For example, a voice call typically requires less of the network's capacity (perhaps a few tens of kilobits, or even less) than a video connection (a few hundreds of kilobits), which can be nicely incorporated in the loss network model described above.

In a way the loss model can be considered as a very advanced version of the Erlang loss model that we introduced at the very beginning of this paper.

Stability

It was mentioned earlier that the steady-state distribution (6) exists if and only if the offered load of each station in the network is strictly less than one, i.e. $\rho_i := \frac{\lambda_i}{\mu_i} < 1$, $i = 1, 2, \ldots, N$. Such conditions are, in the queueing context, referred to as stability conditions and can be viewed, informally speaking, as indications of whether a network has enough resources to handle incoming work. The stability analysis of queueing networks was perhaps thought to be a moot subject, in the sense that, based on the pioneering work of Jackson [24] and Kelly [28], it initially seemed that stability depends only on the offered load of each station in the network. Essentially, this simplistic analysis would imply that the stability of the network can be derived by looking individually at each station in the network.

However, a series of counterexamples demonstrated that the station traffic intensities may not be sufficient to determine the stability of the network. In [7], Bramson gave an example of a two-station network that is unstable, even though the offered load of each station in the network is strictly less than one. In particular, Bramson assumed a network consisting of two stations in tandem, to which customers arrive to station 1 according to a Poisson arrival process at rate 1 and follow a prescribed route $1 \to 2 \to 2 \to \cdots \to 2 \to 1$ at which point they exit the network. In total, any arriving customer to the network will visit station 1 twice and station 2 J times according to the prescribed route. Furthermore, Bramson assumed that the service rate at each station depends on the number of times the customer has already visited the station, say $\mu_{i,j}$, where i denotes the station (taking value 1 for station 1 and value 2 for station 2) and j denotes the number of times this station has been visited up to then (taking values 1 and 2, if $i = 1$, and values $1, 2, \ldots, J$, if $i = 2$). For instance, if one chooses

$$\mu_{1,2} = \mu_{2,1} = \frac{400}{399}, \quad \mu_{1,1} = \mu_{2,j} = 10^{11}, \quad j = 2, 3, \ldots, J \text{ and } J = 1600, \tag{9}$$

then,

$$\rho_1 = \frac{1}{\mu_{1,1}} + \frac{1}{\mu_{1,2}} < 1 \text{ and } \rho_2 = \sum_{j=1}^{J} \frac{1}{\mu_{2,j}} < 1.$$

Bramson showed, see [7, Theorem 1], that for $\mu_{i,j}$ chosen according to (9), this system is unstable with the number of customers in the system growing unboundedly as $t \to \infty$.

This result, while looking counterintuitive at first sight, can be explained as follows, see [8, Section 3.2]. Assume that at time $t = 0$ there are M customers (with M a very large number) in station $(1,1)$ and a few more in the rest of the system. Moreover, let S_1 denote the time at which the last of the original jobs i.e., the jobs present at time $t = 0$, at station 1 is served. Let $S_2, S_3, \ldots$ denote the successive times at which the last jobs at station 2 are served. Since $\mu_{1,1} \gg 1$, one has that $S_1 \ll M$ except on a set of small probability. Also, $\mu_{2,1} = 1/c \approx 1$, and so at time S_1 nearly all of the original jobs in the network are still at $(2,1)$. Next, over this time interval $(S_1, S_2]$, the (approximately) M jobs at $(2,1)$ all move to $(2,2)$. Since $\mu_{2,1} = 1/c$, the time it takes to serve these jobs is (approximately) cM. The time required to serve other jobs is minimal, so $S_2 - S_1 \approx cM$. During this time, (approximately) cM new jobs enter the system, which quickly move to $(2,1)$. Thus, at $t = S_2$, there are (comparatively) few jobs in the system except at $(2,2)$ and $(2,1)$, where there are (approximately) M and cM jobs, respectively. Continuing our reasoning along the same lines, we observe that over $(S_2, S_3]$, the jobs at $(2,1)$ and $(2,2)$ advance to $(2,2)$ and $(2,3)$, respectively. Since $\mu_{2,2} \gg 1$, the time required to serve the jobs at $(2,2)$ is negligible; the time required for the jobs at $(2,1)$ is c^2M, so $S_3 - S_2 \approx c^2M$. Over this time,

c^2M new jobs enter the system, which quickly move to $(2,1)$. So, at time S_3, there are few jobs in the system except at $(2,3)$, $(2,2)$, and $(2,1)$, where there are M, cM and c^2M jobs, respectively. Proceeding inductively, we obtain that at time S_J, there are M jobs at $(2,J)$, cM jobs at $(2,J-1)$, and so on down to $(2,1)$, where there are $c^{J-1}M$ jobs. At station 1, there are few jobs. The elapsed time is $S_J - S_{J-1} \approx c^{J-1}M$. Note that c^J was chosen to be very small, and so there are about

$$\sum_{\ell=0}^{J-1} c^\ell M \approx \frac{M}{1-c} \tag{10}$$

jobs in the system. Likewise, $S_J \approx cM/(1-c)$. Over the short period of time $(S_J, S_{J+1}]$, the evolution of the system changes. The M jobs from $(2,J)$ arrive at $(1,2)$. Since $\mu_{2,1} = 1/c$, these jobs require time cM to be served at station 1, during which time new arrivals at $(1,1)$ will not be served. By time S_{2J}, the jobs that were at station 2 at time S_J have already arrived at $(1,2)$; because of (10), there are essentially $M/(1-c)$ such jobs. So, at time S_{2J}, there are essentially $M/(1-c)$ jobs at $(1,2)$ and no jobs elsewhere. Of course, here and elsewhere, we are taking liberties in ignoring 'negligible' quantities of jobs and probabilities. Let now T denote the time that these last jobs will exit the system. The time required to serve these jobs is $cM/(1-c)$. So, $T - S_{2J} \approx cM/(1-c)$. During this time, $cM/(1-c)$ jobs enter the system. These new jobs are obliged to remain at $(1,1)$ until time $T = S_{2J} + (T - S_{2J}) \approx 2cM/(1-c)$. At this time, there are few jobs elsewhere in the system. So at time T, the state of the system is a 'multiple', by the factor $c/(1-c)$, of the state at time 0.

Of course, since we are working with random events here, the above behavior is sometimes violated. However, such exceptional events occur with probabilities that are exponentially small in M, and one can show they can be ignored without affecting the basic nature of the evolution of number of customers in the system. Needless to say, a rigorous proof requires accurate bookkeeping of such exceptional probabilities, but we were only interested in presenting here an intuitive argument with which the interested reader can grasp why this system is unstable.

Such counter examples inspired further investigations into the stability regions of queueing models under various scheduling policies and also spurred work on the development of a theory for the determination of the stability region for a wide range of queueing networks, see e.g. [21].

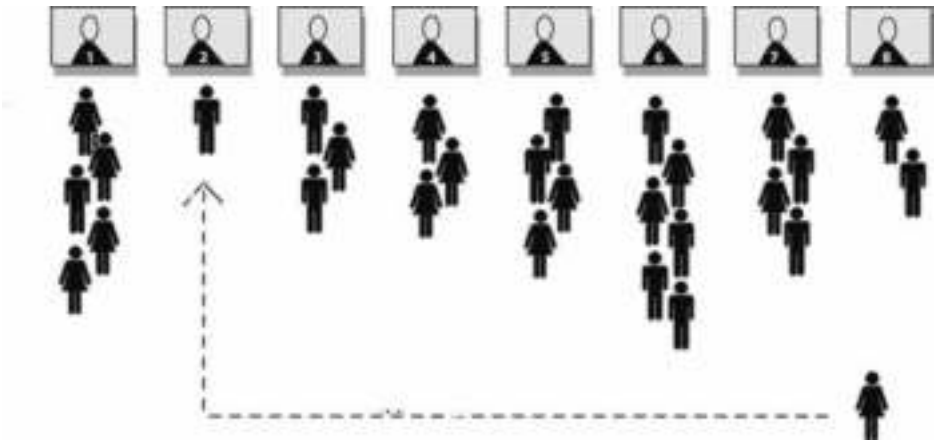


Figure 3 A schematic picture of a JSQ-system.

Routing policies

In the contexts previously described, we assume that customers are routed to the various stations of the network independently of the number of customers already waiting in these stations. However, in practice when a rational customer makes a decision on which station to join, then typically this decision is influenced by the number of customers waiting in queue in each one of the stations. Think for example of the structure of a supermarket: there are multiple cashiers each with their own waiting line, these constitute the various stations in our 'supermarket' network. In the context of supermarkets customers typically join the station with the smallest number of waiting customers. The steady-state distribution of this type of network has received the attention of various researchers and some of the area's important contributions were achieved by several Dutch researchers, including Adan and Cohen (who also published two monographs [13–14] on the related topic of two-dimensional random walks). We will further elaborate on the topic of the steady-state analysis of the *join the shortest queue* (JSQ) policy in the case of two stations in the next section.

Mathematical analysis of 2D models

We have seen that single server queues and specific classes of multi-dimensional queueing systems, such as Jackson networks, can be analyzed in great detail. When slightly changing the mechanics, however, the analysis may become substantially harder. For example, in the case of two stations in parallel where customers are routed according to the JSQ policy, the steady-state distribution does not obey a product-form solution. The steady-state solution can still be found, as we demonstrate in this section.

Model description, steady-state distribution

In the basic version of the model customers arrive to the system according to a Poisson process at rate λ. There are two queues; a new arrival is routed to the shorter one (in the case of a tie, the queue is selected at random). The service times at each of the queues are exponential with mean $1/\mu$. It was argued that this system is stable if $\rho := \lambda/2\mu < 1$. This model was first introduced by Haight [22], and was analyzed by Flatto and McKean [20] and Kingman [31]. We now describe the approach followed by the latter, identifying the probability generating function of the numbers of customers in steady-state.

First, with X_i denoting the number of customers in station i in stationarity, we define

$$\pi(n_1,n_2) = \mathbb{P}(X_1 = n_1, X_2 = n_2),$$
$$n_1, n_2 = 0,1,2,\ldots.$$

By symmetry, $\pi(n_1,n_2) = \pi(n_2,n_1)$. Then, write the balance equations of the system: for $n_1, n_2 = 0,1,\ldots$ such that $n_1 \le n_2$,

$$\begin{aligned}(2\rho + \mathbf{1}_{\{n_1>0\}} + \mathbf{1}_{\{n_2>0\}})\pi(n_1,n_2) \\ = (2\rho\mathbf{1}_{\{n_2=n_1\}} + \rho\mathbf{1}_{\{n_2=n_1+1\}}) \\ \cdot \pi(n_1,n_2-1) + 2\rho\pi(n_1-1,n_2) \\ + \pi(n_1+1,n_2) + \pi(n_1,n_2+1),\end{aligned} \tag{11}$$

where $\mathbf{1}_{\{A\}}$ is the delta Kronecker taking value 1 when event A occurs and 0 otherwise. Let

$$P(x,y) := \sum_{n_1=0}^{\infty} \sum_{n_2=n_1}^{\infty} \pi(n_1,n_2) x^{n_1} y^{n_2-n_1}, \quad |x|,|y| < 1,$$

be the bivariate probability generating function of the minimum queue length (represented by the variable n_1) and of the difference of the two queues (represented by the variable $n_2 - n_1$). Then, multiplying equation (11) with $x^{n_1} y^{n_2-n_1}$ and summing for all $n_1 \le n_2$ yields a functional equation for the probability generating function:

$$\begin{aligned} &\left(x(2\rho x+1) - 2(\rho+1)xy + y^2\right) P(x,y) \\ &\quad = \left(x(2\rho x+1) - (\rho+1)xy \right. \\ &\qquad \left. - \rho x y^2\right) P(x,0) + y(y-x)P(0,y), \\ &|x|,|y| < 1. \end{aligned} \tag{12}$$

This functional equation can be solved as follows. First define the *zero tuples* (x,y), i.e., the (x,y) satisfying

$$\begin{aligned} &x(2\rho x+1) - 2(\rho+1)xy + y^2 = 0, \\ &|x|,|y| < 1. \end{aligned} \tag{13}$$

Then, along the curve (13), equation (12) becomes

$$\begin{aligned} &y(y-x)P(0,y) + \left(x(2\rho x+1)\right. \\ &\left. - (\rho+1)xy - \rho x y^2\right) P(x,0) = 0. \end{aligned} \tag{14}$$

Note that (13) defines a 2-sheeted Riemann surface over the x- and y- planes, which, for any value of x, gives rise to a smooth and closed contour, say $\mathcal{L}$. Thus, equation (14) can be solved as a Riemann–Hilbert boundary value problem: determine a function $P(0,y)$ that is regular for y in the interior of the contour $\mathcal{L}$, continuous on the closure of the contour $\mathcal{L}$ and that satisfies equation (14) on the boundary of the contour $\mathcal{L}$.

Malyshev pioneered this approach of transforming the functional equation to a boundary value problem in the 1970s. The idea to reduce the functional equation for the generating function to a standard Riemann-Hilbert boundary value problem stems from the work of Fayolle and Iasnogorodski [17] on two parallel $M/M/1$ queues with coupled processors (the service speed of a server depends on whether or not the other server is busy). Extensive treatments of the boundary value technique for functional equations can be found in Cohen and Boxma [14] and Fayolle, Iasnogorodski and Malyshev [18].

In the setting of the JSQ model, Kingman noticed that for a given x there are two zero tuples, say (x,y) and (x,Y), that satisfy (13). After tedious calculations he showed that

$$\frac{YP(0,Y)}{yP(0,y)} = \frac{(2+\rho)Y - \rho y}{(2+\rho)y - \rho Y}.$$

With this equation he could calculate the unknown probability generating function $P(0,y)$, and he also concluded that $P(0,y)$ can be continued into a meromorphic function over the whole y-plane. As a result, $P(0,y)$ is holomorphic on the entire y-plane except for a set of isolated points (the poles of the function), at each of which the function must have a Laurent series. Hence, the corresponding probabilities, $\pi(0,n_2)$, can be written as an infinite sum of product forms. Meromorphicity extends also to $P(x,0)$ and eventually $P(x,y)$. As a consequence, for $n_1 \le n_2$ and (x_i, y_j) being roots of (13),

$$\pi(n_1,n_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} c_{ij} x_i^{-n_1} y_j^{n_1-n_2}, \tag{15}$$

for constants c_{ij}. With the solution being still rather implicit, Kingman also tried to look at the asymptotic behavior of $\pi(n_1,n_2)$. He proved

$$\pi(n_1,n_2) \sim c\, \rho^{2n_2} (2+\rho)^{n_1-n_2} \tag{16}$$

as $n_1, n_2 \to \infty$ (while $n_1 \le n_2$), for some constant $c > 0$. Furthermore, it is worth noting that the dominant singularity (x_0, y_0) appearing in (15) is capturing the asymptotic behavior of the steady state distribution, i.e., $1/x_0 = \rho^2$ and $1/y_0 = \rho^2/(2+\rho)$.

An alternative approach

An approach which is not based on generating functions, is developed by Adan et al. in [2–3]. The idea is to directly solve the balance equations, thus leading to an explicit solution for the sub-class of two-dimensional models having a meromorphic generating function. The essence of the approach is to first characterize the products satisfying the balance equations for states in the inner region (i.e., $0 < n_1 \le n_2$), by putting $\pi(n_1,n_2) = x^{-n_1} y^{n_1-n_2}$ into the balance equations for the interior, cf. (11), and simplifying all common terms. This results in a kernel equation for the parameters x and y associated with these product forms, cf. (13). Next it is required that they also satisfy the balance equations on the boundaries (i.e., $n_1 = 0$ and/or $n_2 = n_1$). For JSQ it can be checked that there is no single product form satisfying simultaneously the balance equations in the interior and on the boundaries. To remedy this, a product form, called the 'initial solution', is chosen to satisfy the balance equations in the interior and on one of the boundaries (say $n_1 = 0$), but not necessarily on the other boundary ($n_2 = n_1$). Then a second product form is added to deal with this other boundary, now violating the balance equations for $n_1 = 0$. For this reason, new product forms are alternatingly added in order to compensate for the errors on the boundaries, eventually leading to an infinite series of the form (15). The structure of the alternating compensations gives the method its name: the *compensation approach*.

The difficulty of the approach lies in proving that the series of product forms converges, due to the fact that typically there exists no closed-form expression for the terms of the infinite series. It is interesting to note that the initial solution of the compensation approach is the dominant term of the boundary value problem given in equation (16).

For more details on the various methods that have been developed for two-dimensional models, the interested reader is referred to [1].

Concluding remarks

In this paper we have discussed techniques for identifying the steady-state distribution of the numbers of customers in various elementary queueing systems. We have seen that sometimes elegant closed-form solutions exist, but that the analysis typically substantially complicates when slightly changing the underlying dynamics. It is fair to say that the queueing systems presented in this paper often serve as useful baseline models, but in applications their assumptions (Poisson arrivals, exponentially distributed service times, et cetera) tend to be rather restrictive.

Considering more realistic systems, in terms of size, underlying dynamics, and assumptions on arrival processes and service times, results in most cases in no explicit results being available. In those situations, one typically resorts to approximations [11] (which are sometimes exact in specific asymptotic regimes), or alternatively computational techniques such as Laplace inversion and Monte Carlo simulation [4].

References

1 I.J.B.F. Adan, O.J. Boxma and J.A.C. Resing, Queueing models with multiple waiting lines, *Queueing Systems* 37(1–3) (2001), 65–98.

2 I.J.B.F. Adan, S. Kapodistria and J.S.H. van Leeuwaarden, Erlang arrivals joining the shorter queue, *Queueing Systems* 74(2–3) (2013), 273–302.

3 I.J.B.F. Adan, J. Wessels and W.H.M. Zijm, Analysis of the symmetric shorter queue problem, *Stochastic Models* 6(4) (1990), 691–713.

4 S. Asmussen and P. Glynn, *Stochastic Simulation: Algorithms and Analysis*, Springer, 2007.

5 F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios, Open, closed and mixed networks of queues with different classes of customers, *Journal of the Association for Computing Machinery* 22 (1975), 248–260.

6 S. Borst, J. van Leeuwaarden and P. van de Ven, Stochastische modellen voor random-access-netwerken, *Nieuw Archief voor Wiskunde* 5/16(3) (2015), 201–206, this issue.

7 M. Bramson, Instability of FIFO queueing networks, *The Annals of Applied Probability* 4(2) (1994), 414–431.

8 M. Bramson, Stability of queueing networks, *Probability Surveys* 5 (2008), 169–345.

9 P.J. Burke, The output of a queueing system, *Operations Research* 4 (1956), 699–704.

10 J.P. Buzen, Computational algorithms for closed queueing networks with exponential servers, *Communications of the Association for Computing Machinery* 16 (1973), 527–531.

11 H. Chen and D.D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, Stochastic Modelling and Applied Probability, Vol. 46, Springer, 2013.

12 J.W. Cohen, The multiple phase service network with generalized processor sharing, *Acta Informatica* 12 (1979), 245–284.

13 J.W. Cohen, *Analysis of Random Walks*, IOS Press, 1992.

14 J.W. Cohen and O.J. Boxma, *Boundary Value Problems in Queueing System Analysis*, Elsevier, 2000.

15 N.M. van Dijk, *Queueing Networks and Product Forms*, Wiley, 1993.

16 A.K. Erlang, The theory of probabilities and telephone conversations, *Nyt Tidsskrift for Matematik* 20 (1909), 33–41.

17 G. Fayolle and R. Iasnogorodski, Two coupled processors: the reduction to a Riemann-Hilbert problem, *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* 47 (1979), 325–351.

18 G. Fayolle, R. Iasnogorodski and V. Malyshev, *Random Walks in the Quarter Plane*, Springer, 1999.

19 W. Feller, *An Introduction to Probability Theory and its Applications*, Wiley, 1950.

20 L. Flatto and H.P. McKean, Two queues in parallel, *Communications on Pure and Applied Mathematics* 30(2) (1977), 255–263.

21 S. Foss and T. Konstantopoulos, An overview of some stochastic stability methods, *Journal of the Operations Research Society of Japan* 47(4) (2004), 275–303.

22 F.A. Haight, Two Queues in Parallel, *Biometrika* 45(3–4) (1958), 401–410.

23 R.R.P. Jackson, Queueing systems with phase-type service, *Operational Research Quarterly* 5 (1954), 109–120.

24 J.R. Jackson, Networks of waiting lines, *Operations Research* 5 (1957), 518–521.

25 J.R. Jackson, Jobshop-like queueing systems, *Management Science* 10 (1963), 131–142.

26 M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, Cambridge University Press, 2013.

27 J.S. Kaufman, Blocking in a shared resource environment, *IEEE Transactions on Communications* 29 (1981), 1474–1481.

28 F.P. Kelly, Networks of queues with customers of different types, *Journal of Applied Probability* 12 (1975), 542–554.

29 F.P. Kelly, *Reversibility and Stochastic Networks*, Cambridge University Press, 2011.

30 F.P. Kelly, Loss networks, *Annals of Applied Probability* 1 (1991), 319–378.

31 J.F.C. Kingman, Two similar queues in parallel, *The Annals of Mathematical Statistics* 32(4) (1961), 1314–1323.

32 E. Reich, Waiting times when queues are in tandem, *The Annals of Mathematical Statistics* 28 (1957), 768–773.

33 J.W. Roberts, A service system with heterogeneous user requirement, in *Performance of Data Communications Systems and Their Applications*, G. Pujolle (ed.), North-Holland, 1981, pp. 423–431.

34 R. Serfozo, *Introduction to Stochastic Networks*, Applications of Mathematics, Vol. 44. Springer, 2012.

Sem Borst
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
s.c.borst@tue.nl

Johan van Leeuwaarden
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
j.s.h.v.leeuwaarden@tue.nl

Peter van de Ven
Research Group Stochastics
Centrum Wiskunde & Informatica, Amsterdam
ven@cwi.nl

Stochastische modellen voor random-access-netwerken

Draadloze communicatienetwerken spelen een cruciale rol in het verbinden van laptops, smartphones, sensoren en talloze fysieke apparaten, en zijn van vitaal belang voor het uitwisselen van data tussen personen, computerbreinen en andere elementen in onze informatiemaatschappij. Wegens de immense schaalgrootte en fijnmazige infrastructuur, is het onmogelijk centraal de regie te voeren over het dataverkeer in deze netwerken. Doorgaans worden dan ook zogenaamde random-access-algoritmen gebruikt, waar individuele gebruikers de datastromen op autonome wijze aansturen. Sem Borst, Johan van Leeuwaarden en Peter van de Ven beschrijven in dit artikel stochastische deeltjesmodellen om het macroscopische gedrag van deze netwerken te begrijpen en de prestaties te verbeteren.

Data wordt in toenemende mate draadloos verzonden, wegens de voordelen in het gebruiksgemak, bereik en flexibiliteit vergeleken met bekabelde netwerken. Een kenmerkende eigenschap van draadloze communicatie is dat de signalen in alle richtingen worden verzonden, in plaats van slechts naar de beoogde ontvanger. Zodoende kunnen de talrijke signalen in grootschalige netwerken elkaar verstoren als gevolg van interferentie. Om zulke interferentie te voorkomen of te verminderen, is het dan ook van belang om ervoor te zorgen dat nabije gebruikers van het draadloze netwerk niet gelijktijdig data verzenden. Draadloze netwerken beschikken daartoe over algoritmen die de activiteit van de gebruikers regelen.

Wanneer we een groep draadloze gebruikers beschouwen, bijvoorbeeld apparaten zoals laptops, tablets en smartphones, dan vormen de posities en activiteiten van deze apparaten het netwerk. Ieder apparaat heeft een eigen databuffer en genereert in de loop van de tijd data die de gebruiker naar een bepaalde ontvanger wil verzenden, via draadloze communicatie. Het netwerk is stabiel als alle individuele databuffers stabiel zijn, en het globale perspectief op het gehele netwerk is zodoende sterk verweven met de lokale blik op de individuele gebruikers.

We nemen aan dat de activiteit van de gebruikers wordt aangestuurd door lokale algoritmen, waarbij elke gebruiker in het netwerk zelf bepaalt wanneer deze data verzendt, slechts gebruikmakend van lokale informatie. Een voorbeeld van een dergelijk lokaal algoritme is geïmplementeerd in het zogeheten IEEE 802.11-protocol, dat het kloppend hart is van wifi-netwerken zoals we die elke dag gebruiken op het werk, in de trein en thuis. Grofweg gaat het algoritme als volgt te werk. De capaciteit wordt verdeeld volgens een loterij tussen apparaten in de directe nabijheid van elkaar. Het kansspel wordt, zeg, iedere milliseconde gespeeld en het apparaat met het winnende lot mag gedurende die milliseconde data versturen.

Een interessante verfijning van dit protocol, die we later zullen bespreken, bestaat uit een gewogen loterij. Hoe voller de buffer van het apparaat, hoe meer lotjes het apparaat krijgt om de loterij mee te winnen. Op deze manier stabiliseert het systeem zich effectiever. De minder volle buffers winnen minder vaak en worden wat voller, terwijl de volle buffers de kans krijgen om leeg te lopen.

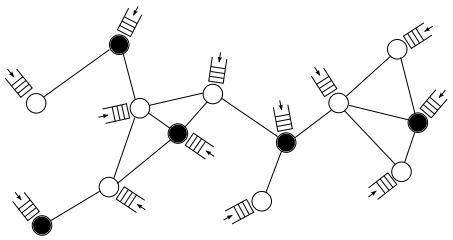
Wanneer we draadloze netwerken en lokale algoritmen iets abstracter bekijken, dan komen we tot een probabilistisch deeltjesmodel op een graaf. De graaf bestaat uit N nodes (de gebruikers in het netwerk) en twee nodes zijn verbonden door een kant wanneer deze nodes niet gelijktijdig actief kunnen zijn (data kunnen verzenden). Denk hierbij aan gebruikers die zich dicht bij elkaar bevinden en elkaars signalen dus kunnen verstoren. Deze graaf noemen we de interferentiegraaf. Een lokaal algoritme zou, lichtelijk gestileerd, als volgt kunnen opereren op deze graaf:

1. Iedere node krijgt een klok die afgaat na een exponentiële tijd met gemiddelde $1/\nu$.
2. Zodra zijn klok afgaat, probeert een node actief te worden en dus data te verzenden. De node zal echter eerst moeten nagaan of geen van zijn buren in de graaf reeds actief is.

3a. Indien geen enkele buur actief is, zal de node data verzenden voor een exponentiële tijd met gemiddelde 1.

3b. Indien tenminste één naburige node actief is, dan blijft de node waarvan de klok is afgegaan inactief en probeert het na een nieuwe exponentiële tijd met gemiddelde $1/\nu$ nogmaals.

Op deze manier verkrijgen we een stochastisch netwerk bestaande uit nodes die door de tijd afwisselend actief en inactief zijn. Het bovenstaande lokale en tevens random algoritme zorgt ervoor dat alle nodes regelmatig



Figuur 1 Een interferentiegraaf met draadloze gebruikers en hun buffers.

kans krijgen iets te verzenden, en belet dat naburige nodes tegelijk actief zijn. De parameter $\nu > 0$ geeft de agressiviteit van het algoritme weer: wanneer ν klein is, zullen nodes lang wachten tussen opeenvolgende activeringspogingen en zullen dus maar relatief weinig nodes tegelijk actief zijn.

Wegens de randomness van het hierboven beschreven algoritme worden deze netwerken in de literatuur aangeduid als *random-access-netwerken*. Dankzij de exponentiële aannamen zijn random-access-netwerken te beschrijven in termen van stochastische modellen die bekend staan als *deeltjesmodellen met interactie*. Deze deeltjesmodellen hebben veel toepassingen in onder andere de wiskunde, natuurkunde en scheikunde (zie ook [7] in dit nummer). Zo wordt in de statistische mechanica magnetische werking beschreven in termen van het Ising-model, een deeltjesmodel waarbij de deeltjes magnetische dipolen zijn waarvan de spin ofwel opwaarts ofwel neerwaarts gericht is, dus toestand $+1$ of -1. De deeltjes in random-access-netwerken zijn in de toestand 0 (inactief) of 1 (actief). In beide deeltjesmodellen vertonen de deeltjes een bepaalde interactie. Waar magnetische werking aanleiding geeft tot ferromagnetische interactie (deeltjes in de nabijheid van elkaar hebben een voorkeur voor dezelfde richting), is de interactie in random-access-netwerken anti-ferromagnetisch: de buren van een deeltje in toestand 1 moeten wel in toestand 0 zitten. Deze zeer strikte vorm van interactie wordt ook wel *hardcore* interactie genoemd en vindt naast random-access-netwerken ook toepassingen in moleculaire gassen waarin de moleculen elkaar niet in de nabijheid dulden [10].

Laat $X_i(t) \in \{0,1\}$ de toestand weergeven van node i op tijdstip t van het deeltjesmodel met deze hardcore interactie. We introduceren $X(t) = (X_1(t), \ldots, X_N(t))$ en merken op dat $\{X(t)\}_{t\geq 0}$ een Markov-proces is. Het langetermijngedrag van dit proces is uitvoerig bestudeerd in de literatuur, in het bijzonder de limietverdeling $\pi(\omega)$. Deze verdeling beschrijft voor elke mogelijke netwerktoestand ω de frequentie waarmee deze toestand op de lange termijn optreedt. Een toestand ω is een combinatie van actieve en inactieve nodes, zodat $\omega = (\omega_1, \ldots, \omega_N) \in \{0,1\}^N$. De limietverdeling kan op elegante wijze worden uitgedrukt in de parameters van het netwerk:

$$\pi(\omega) = Z^{-1} \prod_{i=1}^{N} \nu^{\omega_i} = Z^{-1} \nu^{|\omega|}, \qquad (1)$$

waarbij Z de normalisatieconstante is die ervoor zorgt dat de kansen $\pi(\omega)$ sommeren tot 1. We zien dat $\pi(\omega)$ exponentieel groeit in het aantal actieve nodes in de betreffende toestand (de kardinaliteit $|\omega|$ van ω).

De verzameling van mogelijke toestanden die het netwerk kan aannemen bestaat uit alle onafhankelijke verzamelingen van de interferentiegraaf. Een onafhankelijke verzameling is een deelverzameling van de knopen van die graaf die geen van alle verbonden zijn door een kant. Voor het gemak nemen we aan dat slechts één frequentiekanaal beschikbaar is in het draadloze netwerk. Wanneer het netwerk over meerdere frequentiekanalen beschikt, dan wordt de rol van de onafhankelijke verzamelingen overgenomen door kleuringen van de interferentiegraaf, zie bijvoorbeeld [1] elders in dit nummer.

Het vinden van de onafhankelijke verzamelingen van grafen, en in het bijzonder de grootste onafhankelijke verzameling, is een van de meest klassieke problemen in de grafentheorie en NP-volledig. Ondanks de elegante en ogenschijnlijk eenvoudige structuur, kan de limietverdeling in (1) dus zeer lastig te berekenen zijn, omdat in grote netwerken met grote interferentiegrafen voor het bepalen van Z in principe alle onafhankelijke verzamelingen in kaart moeten worden gebracht. Het random-access-netwerk kan weliswaar lokaal worden aangestuurd, maar het gedrag en de prestatie van het netwerk worden beschreven in termen van globale eigenschappen zoals de onafhankelijke verzamelingen. In dit artikel bespreken we een tweetal cruciale prestatiematen voor random-access-netwerken: doorzet en vertragingen. De wiskundige analyse van het deeltjesmodel verschaft inzicht in deze prestatiematen en berust op technieken uit de kansrekening voor de klasse van Markov-processen waartoe de deeltjesmodellen behoren.

Doorzet van draadloze netwerken

De doorzet meet de hoeveelheid data die gemiddeld kan worden verzonden per tijdseenheid, uitgedrukt in bijvoorbeeld bits of aantallen pakketten per seconde. Uiteraard geldt hierbij dat gebruikers vaak gebaat zijn bij een hoge doorzet. We nemen vooralsnog aan dat alle nodes in een 'verzadigde' toestand verkeren. Dat betekent dat de nodes per definitie datapakketten hebben om te verzenden en altijd activeren wanneer zij, gedicteerd door de kloktijden en de activiteit van hun buren, de gelegenheid krijgen. Later zullen we deze aanname loslaten.

We nemen aan dat nodes per activering één pakketje verzenden. De doorzetwaarden van de verschillende nodes in het netwerk vertonen sterke afhankelijkheden, vanwege het feit dat naburige nodes in de graaf niet tegelijk actief kunnen zijn. Namelijk, wanneer een node een hoge doorzet heeft en veel pakketjes verzendt, zullen diens directe buren minder gelegenheid hebben om actief te worden en als gevolg zullen deze een lage doorzet ervaren.

Deze concurrentie tussen de nodes staat centraal bij het analyseren en ontwerpen van draadloze netwerken. De nodes moeten immers samen de capaciteit van het draadloze netwerk delen en het is van belang dat dit op een eerlijke of acceptabele manier gebeurt. Zo is het natuurlijk om de vraag te stellen: "Hoe moeten we een draadloos netwerk ontwerpen zodanig dat alle nodes dezelfde doorzet hebben?"

Om deze en soortgelijke vragen te beantwoorden, kunnen we gebruikmaken van de limietverdeling (1). De doorzet van node i kan als volgt worden uitgedrukt in termen van de limietkansen:

$$\theta_i := \sum_{\omega:\omega_i=1} \pi(\omega) = Z^{-1} \sum_{\omega:\omega_i=1} \nu^{|\omega|}, \qquad (2)$$

met $\theta = (\theta_1, \ldots, \theta_N)$ de vector met de doorzetwaarden van alle nodes. De sommatie in (2) is dus over alle toestanden waarin node i actief is.

In [9] wordt met behulp van gedetailleerde simulaties aangetoond dat de doorzet (2) van het deeltjesmodel een verrassend goede benadering vormt voor de doorzet van dit soort draadloze netwerken in de praktijk. Dit ondanks het feit dat signaalsterkte, botsingen en fluctuaties van het draadloze signaal niet worden meegenomen in het deeltjesmodel op de graaf. In de wetenschap dat het deeltjesmodel de werkelijkheid goed beschrijft, kan het worden ingezet om de verschillen tussen de doorzetwaarden van de nodes te verminderen.

Ongelijke of oneerlijke doorzetten zien we in vrijwel alle interferentiegrafen. Beschouw bijvoorbeeld een netwerk van $N = 9$ nodes op een lijn, waarbij elke node zijn directe buren blokkeert wanneer hij actief wordt. Figuur 2 toont de doorzet van elke node berekend met (2), voor verschillende waarden van ν. De buitenste nodes van het netwerk hebben de hoogste doorzet, aangezien zij slechts één naburige node hebben, en dus vaker kunnen activeren dan de overige nodes. De verschillende lijnen komen overeen met bepaalde waarden van ν, waarbij de oscillatie groter wordt voor grotere waarden van ν. Het valt eenvoudig in te zien met behulp van (2) dat $\theta_i \uparrow 1$ wanneer $\nu \to \infty$ voor alle oneven nodes, en $\theta_i \downarrow 0$ voor alle even nodes.

Een manier om deze ongelijkheid op te heffen is door nodes verschillende activeringsintensiteiten toe te wijzen. Indien node i een activeringsintensiteit ν_i heeft, en $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$, dan kan (2) worden herschreven als

$$\theta(\boldsymbol{\nu}) := \Big(Z(\boldsymbol{\nu})^{-1} \sum_{\boldsymbol{\omega}:\omega_i=1} \prod_{j=1}^{N} \nu_j^{\omega_j} \Big)_{i=1,\ldots,N}. \quad (3)$$

We beschouwen wederom een lineair netwerk met N nodes, waarin actieve nodes hun directe buren blokkeren. Noteer met $\gamma(i)$ het aantal buren van node i. Het is eenvoudig aan te tonen (zie [15]) dat wanneer we node i activeringsintensiteit

$$\nu_i = \alpha(1+\alpha)^{\gamma(i)-\gamma(1)}, \quad i = 1, \ldots, N,$$

toewijzen, met α een willekeurige positieve constante, elke node een doorzet zal krijgen van

$$\theta_i = \frac{\alpha}{1+2\alpha}, \quad i = 1, \ldots, N.$$

Omdat $\gamma(1) = \gamma(N) = 1$, en $\gamma(i) = 2$ voor $i = 2, \ldots, N-1$, zien we dat de buitenste nodes een lagere activeringsintensiteit wordt toegewezen dan de overige nodes. Dit is intuïtief duidelijk, aangezien dit compenseert voor de hoge doorzet van de buitenste nodes in het geval dat de activeringsintensiteiten gelijk zijn.

Het kiezen van activeringsintensiteiten om te zorgen dat alle nodes in een lineair netwerk dezelfde doorzet zullen krijgen, is slechts één instantie van een veel bredere klasse problemen: gegeven een algemene interferentiegraaf, hoe moeten we de activeringsintensiteiten kiezen om een gewenste vector van doorzetwaarden te verkrijgen? Om deze vraag te kunnen beantwoorden, introduceren we de verzameling Γ, die alle mogelijke combinaties van doorzetwaarden bevat die kunnen worden bereikt door het selecteren van verschillende activeringsintensiteiten. Deze verzameling hangt uiteraard af van de structuur van de interferentiegraaf, en kan worden geïnterpreteerd als het bereik van de functie θ geïntroduceerd in (3). Dit bereik kan worden gekarakteriseerd als het inwendige van het convex omhulsel van alle mogelijke onafhankelijke verzamelingen van de interferentiegraaf.

Laat nu $\boldsymbol{\gamma} \in \Gamma$ een vector van doorzetwaarden zijn die we willen bereiken. We zijn dan op zoek naar de oplossing $\boldsymbol{\nu} = \boldsymbol{\nu}(\boldsymbol{\gamma})$ van het stelsel vergelijkingen

$$\theta(\boldsymbol{\nu}) = \boldsymbol{\gamma}. \quad (4)$$

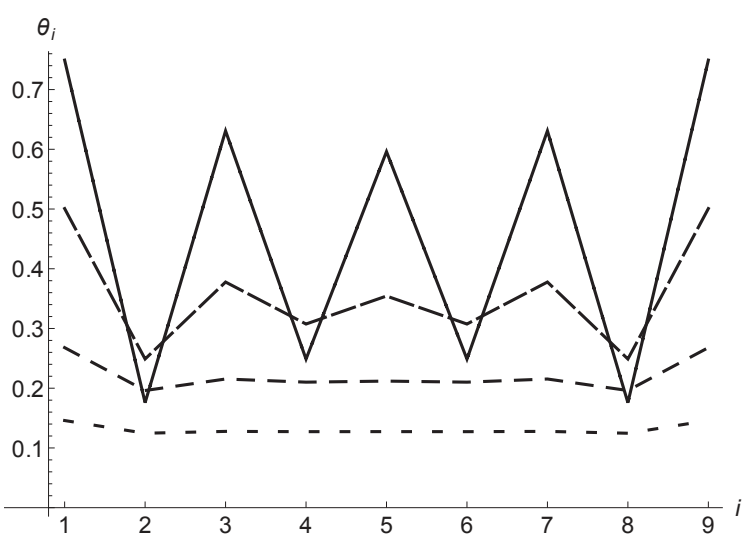
Dit is in het algemeen geen eenvoudig probleem, wegens de complexe vorm van de functie θ.

De eerste stap in het oplossen van (4) is om aan te tonen dat θ globaal inverteerbaar is, zodat we weten dat voor elke vector $\boldsymbol{\gamma} \in \Gamma$ er precies één oplossing $\boldsymbol{\nu}(\boldsymbol{\gamma})$ bestaat van (4), zie [14]. In tegenstelling tot het voorbeeld van een lineair netwerk, zal het in het algemeen niet mogelijk zijn om een uitdrukking voor $\boldsymbol{\nu}(\boldsymbol{\gamma})$ in gesloten vorm te vinden. In plaats daarvan kunnen we bijvoorbeeld numerieke methoden toepassen om dit stelsel op te lossen.

Er is echter nog een manier om $\boldsymbol{\nu}(\boldsymbol{\gamma})$ te bepalen, waarbij de nodes hun eigen activeringsintensiteit aanpassen aan de hand van het verschil tussen hun gerealiseerde doorzet en de gewenste doorzet. Hiertoe wordt de tijd opgedeeld in intervallen van lengte $\beta(j)$ $j =, 1, 2, \ldots$, waarbij aan het eind van elk interval de activeringsintensiteiten $\nu_i(j)$ kunnen worden aangepast. Laat $r_i(j) = \log \nu_i(j)$, en pas deze als volgt recursief aan:

$$r_i(j+1) = \Big(r_i(j) + \beta(j)\big(\gamma_i - \hat{\theta}_i(j)\big)\Big)^+, \quad (5)$$

met $\hat{\theta}_i(j)$ de empirische doorzet in het $j$de interval en $x^+ = \max\{x, 0\}$. Het kan vervolgens worden aangetoond dat voor de juiste keuze van $\beta(j)$, de recursieve reeks (5) inderdaad convergeert naar een oplossing $\mathbf{r} = (r_1, \ldots, r_N)$ zodanig dat $\theta_i(\exp(\mathbf{r})) \geq \gamma_i$, $i = 1, \ldots, N$ [8]. Dit is een opmerkelijk resultaat, aangezien de aanpassing (5) door elke gebruiker afzonderlijk kan worden berekend, en de nodes dus ondanks de complexe interacties in het netwerk op volledig gedistribueerde wijze op de juiste doorzetwaarden kunnen uitkomen. Net als het eenvoudige algoritme beschreven in de vorige paragraaf is dus ook (5) een lokaal algoritme.



Figuur 2 Doorzet van een linear netwerk van 9 nodes voor verschilende waarden van ν.

Congestie en vertragingen

Het basismodel zoals besproken in de voorgaande paragrafen berust op de aanname dat in de buffers van de diverse nodes altijd datapakketten beschikbaar zijn voor verzending. In werkelijkheid zullen de buffers geen 'verzadigde' conditie vertonen, maar een variërend aantal pakketten bevatten als gevolg van de natuurlijke fluctuatie in de generatie en transmissie van datapakketten in de loop van de

tijd. In het bijzonder zullen de buffers van tijd tot tijd geheel leeg raken. Het is belangrijk om het stochastische gedrag van de bufferinhoud te analyseren, alsmede de tijdsduur die een datapakket in de buffer doorbrengt tussen generatie en transmissie. Deze verblijftijd, of vertraging, biedt een cruciale maatstaf voor de prestatie van draadloze netwerken zoals ervaren door gebruikers.

In het vervolg van deze paragraaf veronderstellen we dat datapakketten worden gegenereerd bij node i volgens een Poissonproces met intensiteit λ_i. Om het stochastische gedrag van de bufferinhoud te modelleren, introduceren we $Y(t) = (Y_1(t), \ldots, Y_N(t))$, waar $Y_i(t)$ het aantal pakketten voorstelt dat wacht op transmissie of in transmissie is bij node i op tijdstip t. Ondanks het feit dat buffers nu soms geen pakketten bevatten, houden we in eerste instantie vast aan de aanname dat nodes zich voortdurend in de concurrentie om het medium mengen, en altijd activeren wanneer zij daarvoor de gelegenheid krijgen. Deze aanname impliceert dat het activiteitsproces $\{X(t)\}_{t\geq 0}$ van de diverse nodes niet wordt beïnvloed door de toestand van de buffers, en derhalve de evenwichtsverdeling in (1) behoudt. Het kan dan worden aangetoond dat het Markov-proces $\{(X(t), Y(t))\}_{t\geq 0}$ positief-recurrent is wanneer $\lambda_i < \theta_i$ voor alle nodes $i = 1, \ldots, N$. Het proces $\{(X(t), Y_i(t))\}_{t\geq 0}$ is eveneens Markov, en gedraagt zich als een zogenaamd quasi-geboorte-sterfteproces, waarvoor de evenwichtsverdeling langs numerieke weg is te bepalen met behulp van zogeheten matrix-geometrische methoden.

Om enig kwalitatief inzicht te verkrijgen, beschouwen we nu een regime waar de aankomstintensiteiten $\lambda_1, \ldots, \lambda_N$ relatief hoog zijn. Wil stabiliteit gegarandeerd zijn voor dergelijke aankomstintensiteiten, dan moeten ook de doorzetwaarden $\theta_1, \ldots, \theta_N$ relatief hoog zijn, wat op zijn beurt weer vereist dat de activeringsintensiteiten $\nu_1, \ldots, \nu_N$ relatief hoog zijn. Een dergelijk regime is niet alleen in theoretisch opzicht interessant, maar ook relevant vanuit praktisch oogpunt. Een relatief zware belasting kan namelijk resulteren in lange vertragingen en mindere kwaliteit van de draadloze dienstverlening, terwijl vertragingen voor lagere aankomstintensiteiten geen essentieel punt van zorg zullen zijn.

Om het globale beeld te belichten, bekijken we een symmetrisch scenario waar alle nodes dezelfde aankomstintensiteiten $\lambda_i \equiv \lambda$ en activeringsintensiteiten $\nu_i \equiv \nu$ hebben. Bovendien richten we de aandacht specifiek op een interferentiegraaf die uit een vierkant rooster bestaat van dimensie $N = 2L \times 2L$, zoals geïllustreerd in Figuur 3.

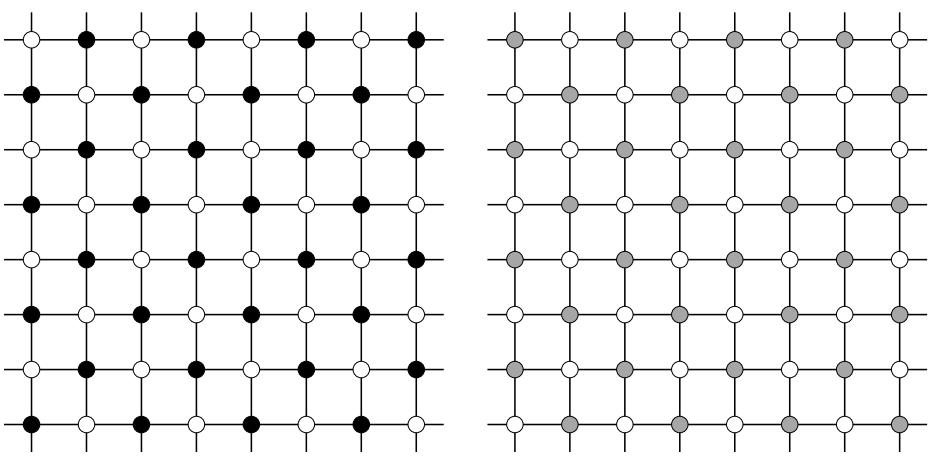
In het geval van 'periodieke' randen, waarbij het rooster een torus wordt, ondervinden de nodes aan de periferie van het rooster niet alleen interferentie van directe buren, maar ook van de corresponderende node aan de tegenovergestelde kant van het rooster. Voor zowel 'open' als periodieke randen, is eenvoudig in te zien dat er twee maximumconfiguraties zijn van grootte $N/2 = 2L^2$, die corresponderen met de witte en zwarte velden van een schaakbord, zie Figuur 3. Gemakshalve zullen we beide maximumconfiguraties en corresponderende nodes respectievelijk als 'even' of 'oneven' aanduiden.

Het systeem zal zich in geval van hoge activeringsintensiteiten vrijwel voortdurend in één van beide maximumconfiguraties bevinden. Transities tussen beide maximumconfiguraties treden slechts zelden op, namelijk op een tijdsschaal van de orde $O(\nu^{\kappa(L)})$ voor grote waarden van ν, waar de exponent $\kappa(L)$ respectievelijk gelijk is aan L of $2L$ in het geval van een open dan wel periodieke rand. Dientengevolge zullen individuele nodes die niet tot de huidige actieve configuratie behoren, gedurende lange tijd geen enkele gelegenheid krijgen om pakketten te verzenden. Wanneer bijvoorbeeld de even configuratie actief is, zullen de aankomende pakketten bij de oneven nodes zich opeenhopen, totdat op zeker moment een transitie optreedt naar de oneven configuratie, en de oneven nodes op hun beurt de gelegenheid krijgen om een fors aantal pakketten achtereenvolgens te verzenden. De opeenvolging van langdurige periodes van accumulatie en transmissie betekent dat het aantal pakketten bij een individuele node zich ontwikkelt volgens een soort van zaagtandpatroon zoals geïllustreerd in Figuur 4. De amplitude van het zaagtandpatroon is evenredig met de geassocieerde tijdsschaal, en derhalve ruwweg van de orde $O(\nu^{\kappa(L)})$ voor grote waarden van ν.

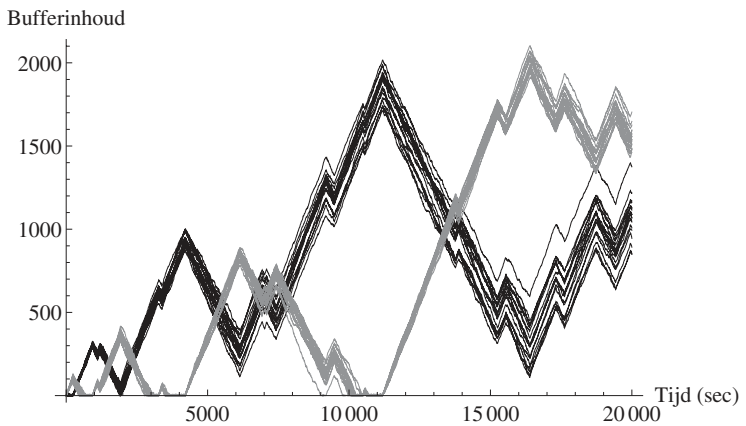
Het is eenvoudig te verifiëren dat zowel voor open als periodieke randen $\lambda < \frac{1}{2}$ en $\nu > \lambda/(1-2\lambda)$ een noodzakelijke voorwaarde is voor stabiliteit van het systeem. We concluderen derhalve dat het gemiddelde aantal pakketten dat zich in de buffer van een individuele node bevindt minstens van de orde $O((\lambda/(1-2\lambda)^{\kappa(L)})$ is wanneer de aankomstintensiteit λ de kritieke waarde $\frac{1}{2}$ nadert [16]. Wegens de zogenaamde wet van Little, heeft de gemiddelde vertraging van een pakket dezelfde orde van grootte. Het is belangrijk deze schaling te vergelijken met de gebruikelijke situatie in conventionele wachtrijsystemen waar de gemiddelde vertraging van de orde van grootte $O(\lambda/(1-2\lambda))$ zou zijn. Zie bijvoorbeeld [5] voor een inleiding tot wachtrijtheorie elders in deze uitgave. Met andere woorden, de langzame transities tussen de rigide maximumconfiguraties hebben een sterk negatief effect op de schaling van het aantal pakketten en de vertraging, waar dat negatieve effect bovendien verergert met de dimensie van het netwerk zoals bepaald door de exponent $\kappa(L)$.

Meer geavanceerde algoritmen

De bovenstaande beschrijving berust op de vereenvoudigende aanname dat nodes altijd activeren wanneer zij daarvoor de gelegenheid krijgen, zelfs wanneer de buffer geen pakketten bevat. Vanuit praktisch oogpunt ligt het vanzelfsprekend eerder voor de hand dat nodes zich in dat geval tijdelijk niet mengen in de concurrentie om het medium, totdat een volgend pakket is gegenereerd dat moet worden verzonden.



Figuur 3 Even en oneven configuraties voor een netwerk met een roosterstructuur.



Figuur 4 Het verloop van de bufferinhoud in de tijd, voor een roosternetwerk.

De consequentie is dat de toestand van de buffers zoals gerepresenteerd door het proces $\{Y(t)\}_{t\geq 0}$, niet alleen wordt bepaald door het activiteitsproces $\{X(t)\}_{t\geq 0}$, maar omgekeerd ook $\{X(t)\}_{t\geq 0}$ beïnvloedt, dat daardoor de evenwichtsverdeling in (1) verliest. Hoewel het proces $\{(X(t), Y(t))\}_{t\geq 0}$ de Markoveigenschap behoudt, geeft de interactie tussen beide componenten aanleiding tot ernstige complicaties in de wiskundige analyse.

Zelfs noodzakelijke en voldoende voorwaarden voor positief-recurrent gedrag zijn in het algemeen niet eenvoudig te karakteriseren. Aangezien het terugtrekken van nodes met lege buffers uit de concurrentie om het medium zinnig lijkt, zou het voor de hand liggen te veronderstellen dat de eerdere voorwaarde $\lambda_i < \theta_i$ voor alle nodes $i = 1, \ldots, N$ voldoende zou blijven om stabiliteit te garanderen. Verrassend genoeg blijkt dat echter niet het geval te zijn. Om dat te illustreren, beschouwen we een netwerk met $N = 4$ nodes en interferentie tussen nodes 1 en 2, 2 en 3, 3 en 4, en 4 en 1. Als bijvoorbeeld node 4 een zeer lage aankomstintensiteit heeft, dan zal deze zich hoogst zelden mengen in de concurrentie om het medium. Node 2 staat er daardoor nagenoeg alleen voor in de concurrentie met nodes 1 en 3, en bevindt zich in een nadelige positie. Dat betekent dat zelfs wanneer node 2 altijd pakketten te verzenden heeft, en nodes 1 en 3 vrij hoge aankomstintensiteiten hebben, de ontvangen doorzet voor node 2 substantieel lager kan zijn dan de waarde θ_2 in geval de buffers van alle vier de nodes in een verzadigde toestand zouden verkeren. Dientengevolge kan instabiliteit optreden bij node 2 wanneer de aankomstintensiteit λ_2 slechts iets lager is dan de waarde θ_2 [13].

De subtiele burenrelaties in het bovenstaande eenvoudige voorbeeld zijn illustratief voor de complexe interactie tussen de diverse nodes. Als consequentie is het karakteriseren van noodzakelijke en voldoende voorwaarden voor positief-recurrent gedrag in zekere zin vrijwel even lastig als het volledig bepalen van de evenwichtsverdeling van het proces $\{(X(t), Y(t))\}_{t\geq 0}$. Voor Markov-processen zonder specifieke (product-vorm-)structuur en met een toestandsruimte die oneindig is in meerdere dimensies, zijn weinig algemene resultaten beschikbaar. Het blijkt dat voor volledige interferentiegrafen wel expliciete voorwaarden voor zogenaamde 'rate' stabiliteit zijn te formuleren [13]. Voor alle overige interferentiegrafen is het lastig hanteerbare voorwaarden voor algehele stabiliteit af te leiden. Wel impliceren de resultaten voor volledige interferentiegrafen partiële voorwaarden voor stabiliteit van bepaalde subsets van nodes [6].

De bovenstaande beschrijving veronderstelt dat de activeringsintensiteit van node i nul is wanneer de buffer leeg is en gelijk aan een constante waarde ν_i wanneer de buffer niet leeg is. Zogenaamde 'queue-based' algoritmen bieden een interessante generalisatie waar de activeringsintensiteit van node i op tijdstip t een algemene (typisch stijgende) functie $f_i(\cdot)$ is van het aantal pakketten $Y_i(t)$ dat zich op tijdstip t in de buffer bevindt. Een gerelateerde variatie op het basisalgoritme behelst dat node i niet slechts één pakket verstuurt maar wanneer een transmissie eindigt op tijdstip t direct een volgend pakket verstuurt met kans $1 - g_i(Y_i(t^+))$, waar $g_i(\cdot)$ een typisch dalende functie is en $g_i(0) = 1$. Ook voor deze modelvarianten blijft de Markoveigenschap van het proces $\{(X(t), Y(t))\}_{t\geq 0}$ behouden.

In het speciale geval van $N = 1$ node blijkt het mogelijk de evenwichtsverdeling van het aantal pakketten te bepalen voor enkele speciale keuzes voor de functies $f(\cdot)$ en $g(\cdot)$ [2]. De ratio van de functies $f(\cdot)$ en $g(\cdot)$ is sterk bepalend voor het kwalitatieve gedrag in een regime waar de aankomstintensiteit λ de kritieke waarde 1 nadert. Hoewel het aantal pakketten in een dergelijk regime onvermijdelijk groeit, blijkt zich een scherpe driedeling in het asymptotische groeigedrag voor te doen, afhankelijk van de vraag of de ratio $f(q)/g(q)$ super-lineair, lineair of sub-lineair stijgt in q. Wegens de wet van Little, treedt dezelfde driedeling in het groeigedrag op voor de orde van grootte van de vertraging van een pakket.

Het behoeft geen betoog dat in het geval van $N > 1$ nodes de evenwichtsverdeling van $\{(X(t), Y(t))\}_{t\geq 0}$ nog minder analytisch toegankelijk is dan in het eerder besproken scenario $f_i(q_i) = \nu_i \mathbf{1}_{\{q_i>0\}}$ en $g_i(q_i) = 0$. In het licht van deze observatie is het des te opmerkelijker dat er wel zeer expliciete voorwaarden voor stabiliteit zijn geïdentificeerd, ook al zijn de ontwikkelde bewijsargumenten hiervoor uitermate geavanceerd. Voor een brede klasse van $f_i(\cdot)$- en $g_i(\cdot)$-functies is aangetoond dat de (praktisch) noodzakelijke voorwaarde dat de vector van aankomstintensiteiten tot het inwendige van het convex omhulsel Γ behoort, voldoende is voor stabiliteit [11]. Dit baanbrekende resultaat demonstreert dat gedistribueerde algoritmen, mits zorgvuldig geconstrueerd, de potentie hebben om de optimale doorzet en maximale stabiliteit van complexe gecentraliseerde mechanismen te evenaren.

Dit belangrijke gegeven roept de vraag op of deze queue-based algoritmen ook in staat zijn de eerdergenoemde zorgelijke schalingen van vertragingen te overwinnen. Helaas blijkt juist voor de klasse van functies waarvoor optimale doorzet en maximale stabiliteit is aangetoond, de schaling niet aanzienlijk te verbeteren [3]. Tot op enige hoogte is deze keerzijde ook een onvermijdelijk feit, aangezien het beperken van vertragingen impliciet vereist dat een maximumconfiguratie snel is te vinden. Aangezien dat laatste probleem NP-lastig is, moet het weinig plausibel worden geacht dat de ongunstige schaling van vertragingen significant is te verbeteren via queue-based algoritmen.

Uitbreidingen en kanttekeningen

Deeltjesmodellen worden al sinds de jaren tachtig gebruikt om random-access-netwerken te analyseren [4] en hebben diverse waardevolle inzichten verschaft in het gedrag van deze netwerken. De adaptieve algoritmen zoals eerder beschreven bieden interessante mogelijkheden om de huidige generatie random-

access-netwerken flexibeler en robuuster te maken. Hoewel de beschreven modellen de meest relevante karakteristieken van draadloze netwerken weten te vatten, zijn verscheidene praktische factoren en verschijnselen niet expliciet gemodelleerd. We besluiten dit artikel met het bespreken van twee van deze verschijnselen die, zodra ze in een wiskundig hanteerbare vorm zouden worden gebracht, zowel de praktijk van de draadloze netwerken als de stand van de theorievorming verder zouden brengen.

Mobiliteit en gebruikersdynamiek. De interferentiegraaf vormt de basis van de wiskundige studie van random-access-netwerken, en deze wordt altijd als statisch verondersteld. In de praktijk zien we echter dat het netwerk beweegt. Althans, de gebruikers van de apparaten bewegen, en daarmee is de structuur van het netwerk dynamisch. Daarnaast zullen gebruikers ook tot het netwerk toetreden en na een tijdje het netwerk weer verlaten. Een vaste interferentiegraaf is dan ook niet meer dan een momentopname van de toestand van het netwerk. Dit vormt een extra uitdaging voor de adaptieve algoritmen. Immers, in plaats van langzaam te convergeren naar een gewenste vector van doorzetwaarden, zal deze vector veranderen met de tijd. De vraag is dan niet of we algoritmen kunnen vinden die naar bepaalde doorzetwaarden convergeren, maar hoe goed we de steeds veranderende gewenste doorzetwaarden kunnen volgen, en dus hoe snel de algoritmen kunnen inspelen op veranderende omstandigheden.

Doorsturen van pakketjes. De eerder beschreven wachtrijmodellen betreffen zogenaamde *single-hop*-scenario's, waarbij iedere node zijn eigen aankomstproces van pakketjes heeft, en een pakket het netwerk verlaat zodra het verzonden is. In werkelijkheid zullen verzonden pakketjes vaak naar andere nodes worden doorgestuurd, om zo bijvoorbeeld het netwerk te kunnen doorkruisen. Dit versterkt de afhankelijkheid tussen de nodes, aangezien de aankomstprocessen van nieuwe pakketjes bij een node nu ook afhankelijk zijn van het activeringsproces van zijn buren. Het lijkt nauwelijks haalbaar om dit soort *multi-hop*-scenario's exact te analyseren, en men zal zich moeten richten op benaderingen en schalingslimieten om wiskundige hanteerbare resultaten te verkrijgen.

Referenties

1 M. de Berg, Conflictvrije kleuringen voor frequentietoekenning in draadloze netwerken, *Nieuw Archief voor Wiskunde* 5/16(3) (2015), 179–182, dit nummer.

2 N. Bouman, S.C. Borst, O.J. Boxma en J.S.H. van Leeuwaarden, Queues with random back-offs, *Queueing Systems* 77(1) (2014), 33–74.

3 N. Bouman, S.C. Borst en J.S.H. van Leeuwaarden, Delay performance in random-access networks, *Queueing Systems* 77(2) (2014), 211–242.

4 R.R. Boorstyn, A. Kershenbaum, B. Maglaris en V. Sahin, Throughput analysis in multihop CSMA packet radio networks, *IEEE Trans. Commun.* 35(3) (1987), 267–274.

5 O. Boxma, S. Kapodistria en M. Mandjes, Performance analysis of stochastic networks, *Nieuw Archief voor Wiskunde* 5/16(3) (2015), 193–200, dit nummer.

6 F. Cecchi, S.C. Borst en J.S.H. van Leeuwaarden, Throughput of CSMA networks with buffer dynamics, *Perf. Eval.* 79 (2014), 216–234.

7 D. Garlaschelli, F. den Hollander en A. Roccaverde, Complexe netwerken vanuit fysisch perspectief, *Nieuw Archief voor Wiskunde* 5/16(3) (2015), 207–209, dit nummer.

8 L. Jiang en J. Walrand, A distributed CSMA algorithm for throughput and utility maximization in wireless networks, *IEEE/ACM Trans. Netw.* 18(3) (2010), 960–972.

9 S.C. Liew, C.H. Kai, H.C.J. Leung en P.B. Wong, Back-of-the-envelope computation of throughput distributions in CSMA wireless networks, *IEEE Trans. Mobile Comput.* 9(9) (2010), 1319–1331.

10 J. Sanders, Atomaire gassen en draadloze netwerken, *STAtOR* 15(1) (2014), 9–12.

11 D. Shah, J. Shin en P. Tetali, Medium access using queues, in: *Proc. FOCS 2011*.

12 D. Shah, D. Tse en J. Tsitsiklis, Hardness of low delay network scheduling. *IEEE Trans. Inf. Theory* 57(12) (2011), 7810–7818.

13 P.M. van de Ven, S.C. Borst, J.S.H. van Leeuwaarden en A. Proutière, Insensitivity and stability of random-access networks, *Perf. Eval.* 67(11) (2010), 1230–1242.

14 P.M. van de Ven, A.J.E.M. Janssen, J.S.H. van Leeuwaarden en S.C. Borst, Achieving target throughputs in random-access networks, *Perf. Eval.* 68(11) (2011), 1103–1117.

15 P.M. van de Ven, J.S.H. van Leeuwaarden, D. Denteneer en A.J.E.M. Janssen, Spatial fairness in linear random-access networks, *Perf. Eval.* 69(3) (2012), 121–134.

16 A. Zocca, S.C. Borst, J.S.H. van Leeuwaarden en F.R. Nardi, Delay performance in random-access grid networks, *Perf. Eval.* 70(10) (2013), 900–915.

Diego Garlaschelli
Instituut-Lorentz voor Theoretische Fysica
Universiteit Leiden
garlaschelli@lorentz.leidenuniv.nl

Frank den Hollander
Mathematisch Instituut
Universteit Leiden
denholla@math.leidenuniv.nl

Andrea Roccaverde
Mathematisch Instituut
Universteit Leiden
a.roccaverde@math.leidenuniv.nl

Complexe netwerken vanuit fysisch perspectief

Voor veeldeeltjessystemen geldt in de regel dat de *microcanonieke* beschrijving in termen van energie equivalent is aan de *canonieke* beschrijving in termen van temperatuur, in de thermodynamische limiet wanneer het aantal deeltjes naar oneindig gaat. Echter, wanneer de deeltjes over lange afstanden met elkaar in wisselwerking staan, dan kan deze equivalentie worden gebroken. Voor complexe netwerken — grote toevallige grafen met topologische randcondities — doet zich eenzelfde verschijnsel voor. In dit artikel geven Diego Garlaschelli, Frank den Hollander en Andrea Roccaverde diverse voorbeelden en kwantificeren zij de mate van niet-equivalentie met behulp van het begrip entropie.

In de statistische fysica wordt voor het bestuderen van de eigenschappen van veeldeeltjessystemen in evenwicht gebruik gemaakt van zogenaamde *ensembles*, dat wil zeggen kansverdelingen op de verzameling van deeltjesconfiguraties [1]. Welk ensemble wordt gekozen hangt af van de informatie die over het systeem beschikbaar is [3]. Zo correspondeert het *microcanonieke ensemble* met de uniforme kansverdeling op de deelverzameling van al die deeltjesconfiguraties waarvoor de energie een voorgeschreven waarde heeft. De keuze voor de uniforme kansverdeling drukt daarbij uit dat van het systeem geen andere informatie beschikbaar is dan dat het de voorgeschreven energie heeft, zodat elke configuratie met deze energie een a priori gelijke kans heeft. Echter, het rekenen met een 'harde randconditie' is in het algemeen lastig. In de praktijk is het handiger om te werken met een 'zachte randconditie'. Het *canonieke ensemble* correspondeert met de kansverdeling met maximale entropie op de verzameling van alle deeltjesconfiguraties, zonder restrictie op de energie, maar zodanig dat de *gemiddelde* energie een voorgeschreven waarde heeft. De keuze voor de maximale entropie drukt daarbij uit dat van het systeem geen andere informatie beschikbaar is dan dat het de voorgeschreven gemiddelde energie heeft. Er wordt een geschikte *temperatuur* gekozen, die wiskundig gezien de rol speelt van een Lagrangemultiplicator die de zachte randconditie realiseert. Het canonieke ensemble kan gezien worden als de oplossing van een probleem van *statistische inferentie* op basis van de *partiële informatie* die verpakt zit in deze randconditie [6].

Voor systemen waarin de deeltjes over *korte afstanden* met elkaar in wisselwerking staan, is aangetoond dat de beide ensembles equivalent zijn in de thermodynamische limiet, dat wil zeggen wanneer het aantal deeltjes naar oneindig gaat. Het idee hierachter is dat in het canonieke ensemble de energie weliswaar fluctueert, maar op een schaal die verwaarloosbaar klein is ten opzichte van de gemiddelde energie, zodat het zich effectief gedraagt als het microcanonieke ensemble. Er zijn echter voorbeelden van systemen waarin de deeltjes over *lange afstanden* met elkaar in wisselwerking staan, waarvoor de equivalentie wordt gebroken. De achtergrond van dit verschijnsel is nog niet goed begrepen.

Vanuit wiskundig perspectief zijn er diverse manieren om niet-equivalentie te kwantificeren [2, 10]. Een recente suggestie is om te kijken naar de *relatieve entropie van de twee ensembles per deeltje* en te laten zien dat die niet naar nul convergeert in de thermodynamische limiet [11]. Relatieve entropie is een pseudo-afstand op de ruimte van kansverdelingen en kan derhalve kansverdelingen van elkaar onderscheiden. Onder bepaalde aannamen kan worden aangetoond dat deze vorm van kwantificeren van niet-equivalentie de scherpste is.

Intermezzo: relatieve entropie

Zij χ een eindige verzameling. Laten $\vec{p} = (p_i)_{i\in\chi}$ en $\vec{q} = (q_i)_{i\in\chi}$ kansverdelingen zijn op χ. De relatieve entropie van $\vec{p}$ ten opzichte van $\vec{q}$ is gedefinieerd als

$$S(\vec{p} \mid \vec{q}) = \sum_{i\in\chi} p_i \log\left(\frac{p_i}{q_i}\right).$$

Omdat de functie $x \mapsto h(x) = x\ln x$ strikt convex is op $[0,\infty)$, geldt dat

$$\begin{aligned} S(\vec{p} \mid \vec{q}) &= \sum_{i\in\chi} q_i\, h\left(\frac{p_i}{q_i}\right) \\ &\geq h\left(\sum_{i\in\chi} q_i \left(\frac{p_i}{q_i}\right)\right) \\ &= h\left(\sum_{i\in\chi} p_i\right) = h(1) = 0. \end{aligned}$$

Met andere woorden de relatieve entropie is niet-negatief. Merk op dat gelijkheid geldt dan en slechts dan wanneer $\vec{p} = \vec{q}$. Merk verder op dat $S(\vec{p} \mid \vec{q})$ niet symmetrisch is onder verwisseling van $\vec{p}$ en $\vec{q}$.

Complexe netwerken

In dit artikel bestuderen we ensembleequivalentie voor *complexe netwerken*: grote toevallige grafen met topologische randcondities [9]. De rol van de deeltjesconfiguratie wordt daarbij overgenomen door de graaf zelf, en de energie van de deeltjesconfiguratie door een geschikte functie van de graaf.

Voor $N \in \mathbb{N}$, zij $\mathcal{G}_N$ de verzameling van alle grafen met N knooppunten. Zij $\vec{C}$ een vectorwaardige functie op $\mathcal{G}_N$. De *microcanonieke kansverdeling* met *harde randconditie* $\vec{C}^\star$ is gedefinieerd als

$$P_{\text{mic}}(\mathbf{G}) = \begin{cases} 1/\Omega_{\vec{C}^\star}, & \vec{C}(\mathbf{G}) = \vec{C}^\star, \\ 0, & \vec{C}(\mathbf{G}) \neq \vec{C}^\star, \end{cases} \qquad \mathbf{G} \in \mathcal{G}_N,$$

waar

$$\Omega_{\vec{C}^\star} = |\{\mathbf{G} \in \mathcal{G}_N \colon \vec{C}(\mathbf{G}) = \vec{C}^\star\}|$$

het aantal grafen is dat $\vec{C}^\star$ realiseert. De *canonieke kansverdeling* wordt gedefinieerd als de kansverdeling op $\mathcal{G}_N$ die de *Shannon entropy*

$$S_N(P_{\text{can}}) = -\sum_{\mathbf{G}\in\mathcal{G}_N} P_{\text{can}}(\mathbf{G}) \ln P_{\text{can}}(\mathbf{G})$$

maximaliseert onder de *zachte randconditie* $\langle \vec{C} \rangle = \vec{C}^\star$, waarbij $\langle \cdot \rangle$ het gemiddelde is met betrekking tot P_{can}. Deze 'duale' kansverdeling wordt gegeven door

$$P_{\text{can}}(\mathbf{G}) = \frac{\exp[-H(\mathbf{G}, \vec{\theta}^\star)]}{Z(\vec{\theta}^\star)}, \quad \mathbf{G} \in \mathcal{G}_N,$$

waar

$$H(\mathbf{G}, \vec{\theta}) = \vec{\theta} \cdot \vec{C}(\mathbf{G})$$

de zogenaamde *Hamilton-functie* op $\mathcal{G}_N$ is en

$$Z(\vec{\theta}) = \sum_{\mathbf{G}\in\mathcal{G}_N} \exp[-H(\mathbf{G}, \vec{\theta})]$$

de normalisatieconstante. De parameter $\vec{\theta}$ moet gelijk worden gekozen aan de waarde $\vec{\theta}^\star$ waarvoor $\langle \vec{C} \rangle = \vec{C}^\star$. Deze waarde maximaliseert de 'statistical likelihood' van de 'random data' $\mathbf{G}$.

Specifieke relatieve entropie van ensembles

De relatieve entropie van P_{mic} ten opzichte van P_{can} is

$$S_N(P_{\text{mic}} \mid P_{\text{can}}) = \sum_{\mathbf{G}\in\mathcal{G}_N} P_{\text{mic}}(\mathbf{G}) \ln \frac{P_{\text{mic}}(\mathbf{G})}{P_{\text{can}}(\mathbf{G})}.$$

Net als in [11], zeggen we dat de twee ensembles equivalent zijn dan en slechts dan als de *specifieke relatieve entropie* nul is, dat wil zeggen

$$s = \lim_{N\to\infty} \frac{S_N(P_{\text{mic}} \mid P_{\text{can}})}{N} = 0.$$

Vanwege de vorm van $H(\mathbf{G}, \vec{\theta})$ geldt dat $P_{\text{can}}(\mathbf{G}_1) = P_{\text{can}}(\mathbf{G}_2)$ zodra $\vec{C}(\mathbf{G}_1) = \vec{C}(\mathbf{G}_2)$. De canonieke kans is dus hetzelfde voor alle configuraties met dezelfde waarde van de randconditie. Derhalve geldt dat

$$S_N(P_{\text{mic}} \mid P_{\text{can}}) = \ln \frac{P_{\text{mic}}(\mathbf{G}^\star)}{P_{\text{can}}(\mathbf{G}^\star)},$$

waarbij $\mathbf{G}^\star$ een *willekeurige* graaf in $\mathcal{G}_N$ is waarvoor $\vec{C}(\mathbf{G}^\star) = \vec{C}^\star$. Ensemble-equivalentie betekent dus dat

$$\lim_{N\to\infty} \frac{1}{N} \Big[\ln P_{\text{mic}}(\mathbf{G}^\star) - \ln P_{\text{can}}(\mathbf{G}^\star) \Big] = 0,$$

met andere woorden het asymptotisch gedrag van de kans van die configuraties die aan de harde randconditie voldoen is voor beide ensembles hetzelfde [5]. Deze observatie is niet alleen van theoretisch belang, het vereenvoudigt ook aanzienlijk de berekeningen, zoals we zullen laten zien.

Configuratie Model

Het *Configuratie Model* genereert een toevallige graaf bestaande uit N punten met van tevoren vastgelegde graden $\vec{k}^\star = (k_1^\star, \dots, k_N^\star)$, waarbij $k_i^\star \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$ het aantal lijnen is dat met knooppunt i is verbonden. In dit voorbeeld is de randconditie dus $\vec{C}^\star = \vec{k}^\star$. Het aantal grafen $\Omega_{\vec{k}^\star}$ dat aan deze randconditie voldoet is niet bekend, maar er zijn wel asymptotische resultaten voor het geval dat $N \to \infty$ en

$$(\star) \qquad k_{\max} = \max_{1\le i\le N} k_i^\star = o(\sqrt{N}).$$

Namelijk, onder conditie $(\star)$ geldt dat [7]

$$\Omega_{\vec{k}^\star} = \frac{\sqrt{2}\,(\frac{2L^\star}{e})^{L^\star}}{\prod_{i=1}^N k_i^\star!}\, e^{-[\overline{k^{\star 2}}/2\overline{k^\star}]^2 + \frac{1}{4} + o(N^{-1}\overline{k^\star}^3)},$$

waar

$$\overline{k^\star} = N^{-1}\sum_{i=1}^N k_i^\star \quad \text{(gemiddelde graad)},$$

$$\overline{k^{\star 2}} = N^{-1}\sum_{i=1}^N k_i^{\star 2} \quad \text{(gemiddelde kwadratische graad)},$$

$$L^\star = \tfrac{1}{2}N\overline{k^\star} \quad \text{(totaal aantal lijnen)}.$$

De canonieke kansverdeling correspondeert met $H(\mathbf{G}, \vec{\theta}) = \vec{\theta} \cdot \vec{k}(\mathbf{G})$, en $\vec{\theta}^\star$ is de oplossing van de vergelijking

$$\sum_{j\neq i} \frac{e^{-\theta_i^\star - \theta_j^\star}}{1 + e^{-\theta_i^\star - \theta_j^\star}} = k_i^\star \quad \forall\, i$$

[8]. Met de afkorting $p_{ij}^\star = e^{-\theta_i^\star - \theta_j^\star}/(1 + e^{-\theta_i^\star - \theta_j^\star})$ hebben we dan

$$P_{\text{can}}(\mathbf{G}) = \prod_{i=1}^N \prod_{j<i} (p_{ij}^\star)^{g_{ij}} (1 - p_{ij}^\star)^{1-g_{ij}},$$

waar g_{ij} het (i, j)-element is van de zogenaamde nabuurmatrix van de graaf $\mathbf{G}$ (dat wil zeggen $g_{ij} = 1$ wanneer i en j verbonden zijn door een lijn, en $g_{ij} = 0$ anders). De conditie in $(\star)$ garandeert dat $k_{\max} = o(\sqrt{L})$, zodat

$$p_{ij}^\star \sim e^{-\theta_i^\star - \theta_j^\star} = \frac{k_i^\star k_j^\star}{2L^\star} = o(1),$$

waar $\sim$ betekent dat het quotiënt van de linker- en de rechterzijde naar 1 convergeert. Derhalve geldt dat $\theta_i^\star \sim -\ln(k_i^\star/\sqrt{2L^\star})$, en een eenvoudige berekening leert dat

$$\ln P_{\text{can}}(\mathbf{G}^\star) \sim \sum_{i=1}^N k_i^\star \ln k_i^\star - L^\star \ln(2L^\star) - L^\star.$$

Dit leidt op zijn beurt tot de formule

$$S_N(P_{\text{mic}} \mid P_{\text{can}}) \sim \sum_{i=1}^N g(k_i^\star) + \Big[\, \overline{k^{\star 2}}/2\overline{k^\star} \,\Big]^2 - \tfrac{1}{4} + o\left(N^{-1}\overline{k^\star}^3\right)$$

met

$$g(k) = \ln\left(\frac{k!}{k^k e^{-k}}\right).$$

De *empirische kansverdeling van de graden* is

$$f_N^\star = N^{-1}\sum_{i=1}^N \delta_{k_i^\star},$$

waarbij δ_k de puntkansverdeling is gedefinieerd door

$$\delta_k(l) = 1_{\{l=k\}}, \quad l \in \mathbb{N}_0.$$

Onder de veronderstelling dat f_N voor $N \to \infty$ convergeert naar een kansverdeling f op $\mathbb{N}_0$, en wel zodanig dat $\lim_{N\to\infty} \sum_{k\in\mathbb{N}_0} |f_N^\star(k) - f(k)|g(k) = 0$ (convergentie in $\ell_1(g)$), vinden we dat

$$s = \sum_{k\in\mathbb{N}_0} f(k)g(k).$$

Omdat $g(k) > 0$ voor alle $k \in \mathbb{N}_0\backslash\{0\}$, volgt dat $s > 0$ zodra $f \neq \delta_0$, en dus zijn de ensembles niet-equivalent. Met andere woorden in de limiet $N \to \infty$ geldt dat onder de canonieke kansverdeling de meeste grafen *niet* een gradenrij hebben die in de buurt ligt van de gemiddelde gradenrij.

De laatste formule heeft een interessante interpretatie. Namelijk,

$$g(k) = s\left(\delta_k \mid \mathrm{POI}[k]\right)$$

is de relatieve entropie van de punt-kansverdeling δ_k ten opzichte van de Poisson-kansverdeling POI[k] gedefinieerd door

$$\mathrm{POI}[k](l) = e^{-k}\,\frac{k^l}{l!}, \quad l \in \mathbb{N}_0.$$

Dit zegt dat elk knooppunt met graad k een relatieve entropie $g(k)$ bijdraagt aan de totale relatieve entropie. Met andere woorden de totale relatieve entropie is een *lineaire functie* van het aantal knooppunten dat aan een (harde dan wel zachte) randconditie op de graad moet voldoen. Het is daarom dat de specifieke relatieve entropie in het Configuratie Model strikt positief is.

Twee voorbeelden

We bekijken twee bijzondere gevallen.

Reguliere netwerken

Elk punt heeft dezelfde graad, dat wil zeggen $k_i^\star = k^\star$ met $k^\star = o(\sqrt{N})$. Dan geldt

$$s = g(k^\star).$$

Merk op dat wanneer $k^\star = k^\star(N)$ divergeert als $N \to \infty$, er een extreme vorm van niet-equivalentie optreedt omdat $g(k^\star) \sim \ln(\sqrt{2\pi k^\star})$, $k^\star \to \infty$.

Schaalvrije netwerken

We kiezen de gradenrij $\vec{k}^\star$ zodanig dat

$$f_N^\star(k) = \begin{cases} A_{k_c} k^{-\gamma}, & 1 \le k < k_c, \\ 0, & k \ge k_c, \end{cases}$$

met $\gamma \in (1, \infty)$ en met $k_c = k_c(N)$ zodanig gekozen dat $\lim_{N\to\infty} k_c(N) = \infty$ en $k_c(N) = o(\sqrt{N})$. Wanneer we $f_N^\star$ benaderen door een continue kansverdeling, dan vinden we dat $A_{k_c} \approx \gamma - 1$. Omdat $g(k) \approx \ln\sqrt{2\pi k}$ leidt dit tot de approximatieve formule

$$s \approx \frac{1}{2(\gamma-1)} + \ln\sqrt{2\pi}.$$

Merk op dat wanneer γ afneemt, de mate van niet-equivalentie toeneemt. Hoe vaker knooppunten met grote graad voorkomen ('hubs' genoemd), hoe sterker de niet-equivalentie is.

Bipartite grafen

Gegeven twee verzamelingen van knooppunten U en V, ter grootte M en N. Voor de eerste verzameling leggen we de gradenrij vast op $\vec{k}^\star = (k_1^\star, \dots, k_M^\star)$, voor de tweede op $\vec{l}^\star = (l_1^\star, \dots, l_N^\star)$. Lijnen zijn alleen toegestaan tussen knooppunten in U enerzijds en V anderzijds. In het bijzonder geldt dat

$$\sum_{i=1}^{M} k_i^\star = \sum_{j=1}^{N} l_j^\star = L^\star,$$

met $L^\star$ het totaal aantal lijnen. Zij $\mathcal{G}_{M,N}$ de verzameling van alle bipartite grafen. Voor P_{mic} kiezen we weer de uniforme verdeling op de deelverzameling van alle grafen in $\mathcal{G}_{M,N}$ die gradenrijen $\vec{k}^\star$ en $\vec{l}^\star$ hebben. Voor P_{can} kiezen we de kansverdeling

$$P_{\mathrm{can}}(\mathbf{G}) = \frac{\exp[-H(\mathbf{G}, \vec{\theta}^\star, \vec{\zeta}^\star)]}{Z(\vec{\theta}^\star, \vec{\zeta}^\star)}, \quad \mathbf{G} \in \mathcal{G}_{M,N},$$

met Hamilton-functie

$$H(\mathbf{G}, \vec{\theta}^\star, \vec{\zeta}^\star) = \vec{\theta}^\star \cdot \vec{k}(\mathbf{G}) + \vec{\zeta}^\star \cdot \vec{l}(\mathbf{G})$$

en normalisatieconstante $Z(\vec{\theta}^\star, \vec{\zeta}^\star)$, waarbij $\vec{\theta}^\star$ en $\vec{\zeta}^\star$ zo dienen te worden gekozen dat de gemiddelde gradenrijen onder P_{can} gelijk zijn aan $\vec{k}^\star$ en $\vec{l}^\star$.

Onder de veronderstelling dat $L^\star \to \infty$ en

$$(\star) \qquad k_{\max} l_{\max} = o\left(L^{\star\,2/3}\right),$$
$$k_{\max} = \max_{1\le i\le M} k_i^\star, \quad l_{\max} = \max_{1\le j\le N} l_j^\star,$$

geldt [4]

$$\Omega_{\vec{k}^\star, \vec{l}^\star} = \frac{L^\star!}{\prod_{i=1}^M k_i^\star! \prod_{j=1}^N l_j^\star!} \exp[o(M+N)],$$
$$M, N \to \infty.$$

(Een meer precieze uitdrukking is beschikbaar voor de exponent, maar die is niet nodig voor onze berekening.) Met behulp van de formule

$$S_{M,N}(P_{\mathrm{mic}} \mid P_{\mathrm{can}}) = \ln \frac{P_{\mathrm{mic}}(\mathbf{G}^\star)}{P_{\mathrm{can}}(\mathbf{G}^\star)},$$

waarbij $\mathbf{G}^\star$ een *willekeurige* graaf in $\mathcal{G}_{M,N}$ is met gradenrijen $\vec{k}^\star$ en $\vec{l}^\star$, leidt dit uiteindelijk tot de uitdrukking

$$s = \lim_{M,N\to\infty} \frac{S_{M,N}(P_{\mathrm{mic}} \mid P_{\mathrm{can}})}{M+N}$$
$$= \alpha \sum_{k\in\mathbb{N}_0} f_1(k) g(k) + (1-\alpha) \sum_{k\in\mathbb{N}_0} f_2(k) g(k),$$

waarbij f_1, f_2 de limieten zijn van de empirische kansverdelingen $f_{1,M}^\star, f_{2,N}^\star$ van de graden, de convergentie naar deze limieten geschiedt in de $\ell_1(g)$-norm, en de limiet $M, N \to \infty$ zo genomen wordt dat

$$\lim_{M,N\to\infty} \frac{M}{M+N} = \alpha \in [0,1].$$

De berekening is analoog aan die voor het Configuratie Model. Opnieuw zien we dat de ensembles niet-equivalent zijn, en dat de totale relatieve entropie op te vatten is als een som over knooppunten van de relatieve entropie per knooppunt.

Conclusie

Niet-equivalentie van ensembles is een natuurlijke eigenschap van complexe netwerken. De achterliggende reden is dat *lokale randcondities een langedrachteffect* hebben wanneer hun aantal extensief is in het aantal knooppunten. Het is een uitdaging om een volledige *classificatie* te geven van alle randcondities die leiden tot niet-equivalentie.

Referenties

1 L. Boltzmann, *Wiener Berichte* 2 (1877), 373.

2 R.S. Ellis, H. Touchette en B. Turkington, *Physica A* 335 (2004), 518.

3 J.W. Gibbs, *Elementary Principles of Statistical Mechanics*, Yale University Press, New Haven, CT, 1902.

4 C. Greenhill, B.D. McKay en X. Wang, *J. Combin. Theory A* 113 (2004), 291.

5 F. den Hollander, *Large Deviations*, Fields Institute Monographs 14, American Mathematical Society, Providence, RI, 2000.

6 E.T. Jaynes, *Phys. Rev.* 106 (1957), 4.

7 B.D. McKay en N.C. Wormald, *Combinatorica* 11 (1991), 369.

8 T. Squartini en D. Garlaschelli, *New Journal of Physics* 13 (2011), 083001.

9 T. Squartini, J. de Mol, F. den Hollander en D. Garlaschelli, Breaking of ensemble equivalence in networks, arXiv:1501.00388.

10 H. Touchette, R.S. Ellis en B. Turkington, *Physica A* 340 (2004), 138.

11 H. Touchette, arXiv:1403.6608.

Derk Pik
hoofdredacteur Pythagoras
Amsterdam
derk@pyth.eu

Rob Tijdeman
Mathematisch Instituut
Universiteit Leiden
tijdeman@math.leidenuniv.nl

In Memoriam Jan Turk (1951–2015)

Veelzijdig en kritisch

Op 26 maart 2015 is de wiskundige Jan Turk overleden. De laatste jaren heeft hij zich met veel enthousiasme ingezet voor het jongerentijdschrift *Pythagoras*. Zijn promotor Rob Tijdeman en hoofdredacteur van Pythagoras Derk Pik gedenken een gedreven collega. Na dit In Memoriam volgt een artikel over priemgetallen, dat hij enkele maanden voor zijn dood ter publicatie heeft aangeboden.

Jan Turk werd op 24 maart 1951 te Nieuwer Amstel (thans Amstelveen) geboren. Van 1963 tot 1968 bezocht hij het Keizer Karel College te Amstelveen. Aansluitend studeerde hij tot 1975 wiskunde aan de Universiteit van Amsterdam waar hij met lof doctoraalexamen deed. Van 1975 tot 1979 deed hij promotieonderzoek aan de Rijksuniversiteit Leiden. Dit resulteerde in het proefschrift *Multiplicative properties of neighbouring integers*. Een typerend resultaat uit het proefschrift is dat er een constante c is zo dat er voor elk natuurlijk getal N in het interval $[N, N + N^{1/2}]$ ten hoogste $c(\log N)^{1/2}$ machten van natuurlijke getallen liggen. Hierna volgden enkele postdocjaren waarin Jan onder andere een tijd aan de Universiteit van Michigan in Ann Arbor verbleef. In deze periode schreef hij het artikel 'Products in short intervals' samen met Paul Erdős, wat zijn Erdősgetal op 1 bracht.

Weer terug in Nederland ontwikkelde Jan zich tot een veelgevraagd ICT-adviseur. Hij werkte eerst als research scientist en system architect bij Philips. In de jaren negentig werd hij consultant bij PricewaterhouseCoopers. Vanaf 2003 had hij een eigen bedrijf, ITSucces. Hij deed veel adviseurswerk voor de overheid. Hij behield al die tijd zijn interesse in de wiskunde en bezocht af en toe conferenties en lezingen. Na zijn zestigste verjaardag besloot hij zijn ICT-werkzaamheden geleidelijk af te bouwen en meer tijd aan zijn grote liefde wiskunde te besteden. Hij startte weer met onderzoek en verdiepte zich in wat op zijn gebied sinds 1984 gebeurd was. Het hierna volgende artikel is daar een vrucht van. Ook spande hij zich in voor het scholierentijdschrift *Pythagoras*.

Het begon in 2012 toen Jan een artikel voor publicatie in Pythagoras aanbood. Pythagoras is inhoudelijk een goed lopend blad met een enthousiaste redactie. Als bedrijf is het echter kwetsbaar. Bij het KWG wilde men het blad beter organiseren en daarvoor een Managementteam (MT) vormen. Toen we Jan daarvoor vroegen, aarzelde hij geen moment en bood hij bovendien aan dit voor niets te doen. Hij vormde samen met Mark Veraar van het KWG en de hoofdredacteur het MT Pythagoras. Onder zijn leiding zijn de financiën op orde gebracht, is er een jaarverslag en een jaarplan gerealiseerd en zijn de relaties met onze leveranciers genormaliseerd. Onder Jans leiding zijn verder een aantal belangrijke projecten gestart: het digitaliseren van Pythagoras, Junior Pythagoras en het in kaart brengen van het abonneebestand. Jan deed dit met grote inzet. Hij onderkende het grote belang van een variant van Pythagoras voor leerlingen in de bovenbouw van de basisschool.

Jan toonde zich iemand die graag sprak over wiskunde, over hoe je op een goede manier met bedrijven moet omgaan en over muziek. Ook reageerde hij vaak op artikelen in Pythagoras. Hij zag in het bijzonder de wiskundige waarde van het programmeren en reageerde soms scherp en kritisch als daar niet genoeg aandacht aan werd besteed. Jan was zeer enthousiast over de Pythagorasserie over problemen van Erdős. Deze serie bestaat inmiddels twee jaar en heeft veel Pythagoraslezers aan het werk gezet.

Op zijn 64ste verjaardag kreeg Jan zware hoofdpijn. Het bleek een fatale hersenbloeding te zijn. Twee dagen later overleed hij. Tijdens een indrukwekkende uitvaartplechtigheid namen velen afscheid van hem. Daar bleek hoe veelzijdig hij was. We zullen hem missen, als persoon, vriend, collega en als deskundige bij het tijdschrift Pythagoras. Bij Pythagoras zullen we ons inzetten om te voltooien wat onder zijn leiding zo goed is gestart.

Jan Turk
Den Haag

Artikelbespreking **Micheal Rubinstein en Peter Sarnak, Chebyshev's bias**

Subtiel scheve verdeling van priemgetallen in rekenkundige rijen met gegeven reden

In januari ontving de redactie deze artikelbespreking van Jan Turk. In het begeleidende bericht schrijft hij: "Voor eigen gebruik heb ik een bijzonder artikel van Peter Sarnak bestudeerd. Het artikel is weliswaar uit 1994, maar Sarnak heeft in 2014 de Wolf-prijs gekregen, ik kan me voorstellen mede vanwege dit artikel." Helaas heeft Jan de publicatie van dit stuk niet meer mee mogen maken. Hij overleed op 26 maart 2015. Een In Memoriam is te vinden op de pagina hiervoor.

Het aantal priemgetallen ten hoogste $x \geq 2$ dat 1 plus een veelvoud van 4 is, wordt genoteerd met $\pi_{4;1}(x)$, evenzo $\pi_{4;3}(x)$ en algemener $\pi_{q;a}(x)$ voor veelvouden van q plus a met $\text{ggd}(a,q) = 1$. Voor vaste q zijn er $\phi(q)$ van zulke a's met $1 \leq a \leq q$, waarbij ϕ de bekende functie van de Zwitser Euler is. We nemen $q \geq 3$. Het is duidelijk dat $\phi(q) = 2$ voor $q \in \{3,4,6\}$ en $\phi(q) \geq 4$ voor $q \notin \{1,2,3,4,6\}$. Een getal a met $0 < a < q$ heet *kwadraatrest* als de congruentie $x^2 \equiv a \pmod q$ een oplossing heeft, notatie $a \in R$, en anders een *niet-kwadraatrest*, notatie $a \in N$. Voor $q = 4, p^v, 2p^v$ met p oneven en v een natuurlijk getal zijn er evenveel kwadraatresten als niet-kwadraatresten.

De Duitser Dirichlet bewees in 1837 dat $\pi_{q;a_1}(x)$ en $\pi_{q;a_2}(x)$ beide naar oneindig gaan als x naar oneindig gaat en dat hun quotiënt dan naar 1 gaat. De Fransman Hadamard en de Belg de la Vallée Poussin preciseerden dit aan het eind van de negentiende eeuw tot de *priemgetalstelling voor rekenkundige rijen*

$$\pi_{q;a}(x) \sim \frac{1}{\phi(q)} \frac{x}{\ln x}.$$

De Duitser Riemann stelde in de negentiende eeuw formules op voor $\pi_{q;a}(x)$ in termen van complexe nulpunten van Dirichletfuncties. Latere wiskundigen formuleerden de *Gegeneraliseerde Riemann Hypothese* (GRH), en de Hongaarse Amerikaan Wintner, de Brit Hooley en de Amerikaan Montgomery formuleerden de *Grand Simplicity Hypothesis* (GSH) over die nulpunten van Dirichletfuncties.

De Rus Chebyshev maakte in 1853 de observatie dat $\pi_{4;3}(x)$ groter is dan $\pi_{4;1}(x)$ voor alle $x \geq 3$ waarvoor hij het berekende. Wat is de betekenis van Chebyshevs observatie?

De Engelsman Littlewood bewees in 1914 dat $P_{4;3,1} := \{n \in \mathbb{N} \mid \pi_{4;3}(n) < \pi_{4;1}(n)\}$ niet leeg is, zelfs oneindig is, en $P_{4;3,1} = \{n \in \mathbb{N} \mid \pi_{4;3}(n) > \pi_{4;1}(n)\}$ evenzo. De Engelsman Leech berekende in 1957 dat 26861 het kleinste getal in $P_{4;3,1}$ is. Chebyshev was niet tot 26861 gekomen is de nuchtere conclusie na dit resultaat van Leech.

Bij de verdere bestudering wordt gebruik gemaakt van dichtheden. De *natuurlijke dichtheid* van een verzameling V van reële getallen groter dan 1 is gedefinieerd als

$$\mathrm{d}(V) = \lim_{n\to\infty} X^{-1} \int_2^X \chi_V(t)\,\mathrm{d}t,$$

als de limiet bestaat. De *logaritmische dichtheid* $\delta(V)$ is gedefinieerd als

$$\delta(V) = \lim_{x\to\infty} (\ln X)^{-1} \int_2^X \chi_V(t)\,\mathrm{d}(\ln t),$$

als de limiet bestaat. Hier is $\chi_V(t)$ de karakteristieke functie van V met waarde 1 als $t \in V$ en waarde 0 als $t \notin V$.

Wintner, die met de kosmopolitische Hongaar Erdős en de Poolse Amerikaan Kac aan de wieg van de probabilistische getaltheorie stond, had al in 1941 laten zien dat $P_{4;1,3}$ en $P_{4;3,1}$ geen natuurlijke dichtheid hebben, maar dat je met de logaritmische dichtheid wel een maat kunt geven. De Zuid-Afrikaanse Amerikaan Peter Sarnak en de Amerikaanse Michael Rubinstein berekenden in 1994 dat $\delta(P_{4;3,1}) \doteq 0{,}9959$ en $\delta(P_{4;1,3}) \doteq 0{,}0041$ [1]. Inderdaad een erg scheve verdeling in deze

maat. Het geeft een andere betekenis aan Chebyshevs observatie. Rubinstein en Sarnak introduceerden de term *Chebyshev-bias* in hun artikel van 1994 en bewezen enkele prachtige stellingen, onder aanname van GRH en GSH, over de logaritmische dichtheden $\delta(P_{q:a_1,a_2,\ldots,a_r})$ van priemgetallen in r verschillende rekenkundige rijen $a_i \pmod q$, $i = 1,\ldots,r$ met reden q, waar $2 \le r \le \phi(q)$. Hier is

$$P_{q:a_1,a_2\ldots,a_r} = \Big\{x \in \mathbb{N}, x \ge 2 \mid \pi_{q;a_1}(x) < \pi_{q;a_2}(x) < \cdots < \pi_{q;a_r}(x)\Big\}.$$

Gegeven $a_1,\ldots,a_r$, zijn er natuurlijk $r!$ van zulke verzamelingen natuurlijke getallen.

Mijn selectie van hun stellingen voor deze bespreking is de volgende:

1. $\delta(P_{7;1,2,4}) = \frac{1}{6}$ en evenzo voor de andere vijf permutaties van $1,2,4$. Met andere woorden, in de drie rekenkundige rijen $1,2,4 \pmod 7$ is er geen Chebyshev-bias. Ook niet in de andere drie, $3,5,6 \pmod 7$. Wel is er sterke Chebyshev-bias van $3,5,6 \pmod 7$ ten opzichte van $1,2,4 \pmod 7$: $\delta(P_{7;3,5,6,1,2,4}) \doteq 0,9782$, $\delta(P_{7;1,2,4,3,5,6}) \doteq 0,0217$.
2. Er is geen Chebyshev-bias in drie rekenkundige rijen met reden q als ze de machten zijn van ρ met $\rho^3 \equiv 1 \pmod q$, $\rho \ne 1$, zoals bij $q = 7$, $\rho = 2$ want $2^3 \equiv 1 \pmod 7$. Zulke trio's zijn er bijvoorbeeld voor elke q die priem en 1 plus een veelvoud van 3 is, een leuke echo van ons onderwerp.
3. Er is geen Chebyshev-bias in twee rekenkundige rijen $a_1 \pmod q$ en $a_2 \pmod q$ als het aantal oplossingen van $X^2 \equiv a_1 \pmod q$ gelijk is aan het aantal oplossingen van $X^2 \equiv a_2 \pmod q$. Zulke paren zijn er voor alle $q \in \mathbb{N} \setminus \{1,2,3,4,6\}$. Bijvoorbeeld geldt:

$$\begin{aligned}\delta(P_{5;1,4}) = \delta(P_{5;4,1}) = \delta(P_{5;2,3}) \\ = \delta(P_{5;3,2}) = \tfrac{1}{2},\end{aligned}$$

$$\begin{aligned}\delta(P_{8;3,5}) = \delta(P_{8;5,3}) = \delta(P_{8;3,7}) \\ = \delta(P_{8;7,3}) = \delta(P_{8;5,7}) \\ = \delta(P_{8;7,5}) = \tfrac{1}{2},\end{aligned}$$

$$\begin{aligned}\delta(P_{10;1,9}) = \delta(P_{10;9,1}) = \delta(P_{10;3,7}) \\ = \delta(P_{10;7,3}) = \tfrac{1}{2}.\end{aligned}$$

Foto: School of Mathematics, Shandong University

Peter Sarnak in januari 2014 op een conferentie over getaltheorie die ter ere van hem gehouden werd aan de universiteit van Shandong in China.

4. Afgezien van bovenstaande (oneindig vele) gevallen van rekenkundige rijen zonder Chebyshev-bias is er altijd Chebyshev-bias in r rekenkundige rijen met reden q. Bijvoorbeeld $\delta(P_{3;2,1}) \doteq 0,9990$ en $\delta(P_{3;1,2}) \doteq 0,0010$. Dit is overigens de maximale Chebyshev-bias in rekenkundige rijen. Er is altijd Chebyshev-bias als $r = \phi(q)$ met $q \notin \{1,2,3,4,6\}$, want dan is $r \ge 4$.
5. Er is Chebyshev-bias naar de niet-kwadraatresten $N \pmod q$ ten opzichte van de kwadraatresten $R \pmod q$ voor alle $q = 4, p^\nu, 2p^\nu$ met p oneven priem en ν natuurlijk. Het maakt niet uit hoe de N en R daarbij onderling geordend zijn, zie punt 3. Bijvoorbeeld geldt voor $q = 10$ dat $\delta(P_{10;3,7,1,9}) = \delta(P_{10;7,3,1,9}) > \frac{1}{2}$.
6. De Chebyshev-bias naar niet-kwadraatresten wordt willekeurig klein als q naar oneindig gaat: $\delta(P_{q;N,R}) \to \frac{1}{2}$, $\delta(P_{q;R,N}) \to \frac{1}{2}$ als $q \to \infty$.
7. De Chebyshev-bias wordt willekeurig klein als q naar oneindig gaat: voor elke $r \ge 2$ en alle $a_1,\ldots,a_r$ convergeert $\delta(P_{q;a_1,\ldots,a_r})$ naar $\frac{1}{r!}$ als $q \to \infty$.

In het artikel staat nog een mooi resultaat dat niet over Chebyshev-bias gaat. In 1941 had Wintner al bewezen, onder aanname van GRH, dat de logaritmische dichtheid $\delta(P_1)$ bestaat, waarbij

$$P_1 = \{n \in \mathbb{N} \mid \pi(n) > \mathrm{Li}(n)\}$$

met

$$\mathrm{Li}(n) = \int_2^n \frac{dx}{\ln x}$$

de eerste term in de formule van Riemann voor $\pi(n)$ is. Littlewood bewees in 1914 dat P_1 oneindig groot is. De Zuid-Afrikaan Skewes berekende in 1933 een hoge bovengrens voor het kleinste getal in P_1 onder aanname van GRH, en in 1955 een nog hogere zonder aanname van een hypothese. De Nederlander Herman te Riele bewees in 1986 dat de bovengrens ten hoogste 10^{370} is. De Sloveen Kotnik toonde in 2008 aan dat de bovengrens ten minste 10^{14} is. De berekening van de logaritmische dichtheid van P_1,

$$\delta(P_1) \doteq 0,00000026$$

door Rubinstein en Sarnak is spectaculair te noemen.

Lees het artikel, het bevat meer moois, waaronder bewijzen. Het is prachtig geschreven. In 2014 is de prestigieuze Wolf-prijs toegekend aan Peter Sarnak.

Referentie

1 Micheal Rubinstein en Peter Sarnak, Chebyshev's bias, *J. Experimental Math.* 3 (1994), 173–197.

Ionica Smeets
wetenschapsjournalist, Leiden
i@ionica.nl

Het keerpunt van Bert Peletier

De intellectuele bevrediging is anders, maar net zo groot

Bert Peletier (1937) werkte zijn hele carrière in de zuivere wiskunde, maar gooide bij zijn emeritaat het roer om en dook in een vakgebied dat hij per toeval ontdekte: wiskundige modellen voor medicijnen. "De vragen zijn zó ontzettend moeilijk, en zó leuk."

Voor de jonge Bert Peletier was de studiekeuze snel gemaakt. "Mijn vader was ingenieur en ik ging graag mee naar zijn lab bij Shell om proeven te bekijken. Ik vond techniek altijd ontzettend leuk, ik maakte graag modelvliegtuigen en wilde een technische studie doen. Eindhoven en Twente waren er nog niet, dus ging ik vanzelfsprekend naar Delft. Daar koos ik voor natuurkunde. Mijn wiskundeleraar zei namelijk dat het verstandig was om de context van wiskunde te snappen en dat kon je maar beter doen bij natuurkunde, dan bij wiskunde zelf. Later heb ik altijd veel plezier gehad van die context. Al deed ik heel zuivere wiskunde, het was handig om te weten waar die vergelijkingen vandaan komen."

Heeft u na uw studie overwogen om het bedrijfsleven in te gaan?
"Via mijn afstudeerhoogleraar kon ik na mijn afstuderen een jaar naar MIT, dat was een mini-keerpunt. Daar werd ik zó enthousiast over het academische leven. Mensen hadden er een missie om iets uit te vinden: je moet de eerste zijn en je moet erachteraan gaan. Onderzoek was er spannend: je snapt iets niet en je moet en zult eruit komen. Die sfeer hing overal, ook op party's praatten ze niet over snelle auto's, maar over hun vak. Die gemeenschap sprak me enorm aan. Het was volledig duidelijk dat ik moest proberen om me die kant op te ontwikkelen."

Peletier vertrok naar de gloednieuwe universiteit van Eindhoven, waar hij promoveerde op golfvergelijkingen. Later werkte hij onder andere in Sussex voor hij uiteindelijk neerstreek als hoogleraar wiskunde in Leiden. Eind jaren negentig, vlak voor zijn emeritaat, kwam hij ineens in aanraking met farmacologie: hoe modeleer je de werking van een medicijn?

Hoe belandt een zuiver wiskundige bij dit soort vraagstukken?
"Dat is een typisch *serendipity*-verhaal. De zus van mijn schoonzoon is apotheker. Zij las in haar vakblad iets over Nederlandse onderzoekers die werkten met wiskundige modellen uit Zweden. Toen belde ze me op en zei: 'Als ze dit in Zweden kunnen, dan moet jij het toch ook kunnen.' Ze wist wel dat ik van dit soort dingen houd. Ik ging op onderzoek uit: wie zou ik hiervoor moeten hebben in Leiden? Ik belde als eerste maar eens met Meindert Danhof, de directeur van LACDR (Leiden Academic Center for Drug Research). Hij vond het wel leuk om een lokaal iemand te hebben om mee te praten."

Deed u dat vaker, zomaar een onbekende opbellen omdat u met hem wilde werken?
"Ja, dat is toch niet zo moeilijk? Je moet achter de problemen aan gaan. Een van mijn vriendjes in Chicago had als lijfspreuk: 'Science is the pursuit of knowledge in the company of friends.' Je moet het altijd met

Bert Peletier

mensen doen, dat is toch het leukst. Meindert en ik hadden aan het begin ook veel geluk: Klaas Zuideveld, de promovendus die op dit onderwerp zat, was briljant en had een goed gevoel voor wiskunde. We hebben toen met een groep mensen een model gemaakt en omdat Klaas zo goed was met computers maakte hij daar een prachtig plaatje van. Het was een 3D-model waarin alles heel mooi bewoog. Klaas stal de show op een groot congres van farmacologen toen hij dat liet zien."

Waar was dat plaatje van?
"Een stelsel van drie differentiaalvergelijkingen waarvan de oplossing kan worden weergegeven door een baan in drie dimensies."

Ik bedoelde eigenlijk waar die vergelijkingen voor stonden.
"Dan moet ik beginnen bij de achterliggende data. Hier ging het om het effect van een bepaalde stof op de lichaamstemperatuur. Wat de farmacologen zagen, was dat bij een kleine dosis van de toegediende stof de temperatuur eerst daalde, en tijdens terugkeer naar het oude niveau behoorlijk schommelde; bij een grote dosis waren er in de terugkeer geen schommelingen te bespeuren. Ze wilden een model hebben dat bij elke willekeurige dosis kon voorspellen hoe de kromme van de temperatuur eruit zou zien. Dat hebben we toen gemaakt. Daarin kwamen twee vergelijkingen voor, een voor de temperatuur en een voor een, hypothetische, signaalstof. De derde dimensie was voor de concentratie van het stofje, die ook veranderde gedurende het experiment."

Waar komen die schommelingen bij een lage dosis vandaan?
"Dat was het punt: dat wisten de farmacologen niet. Ze wilden een model dat dit rare gedrag vertoonde en dat hebben wij gemaakt. Als je aan dit soort problemen werkt, moet je je eroverheen zetten dat het onbevredigend is als je iets niet helemaal snapt. Als je een brug bouwt, dan gebruik je balken die voldoen aan allerlei wetten: mechanische, elastische, noem het maar op. In wezen is alles bekend. Maar als je een medicijn maakt is het veel ingewikkelder. Iemand eet een pil en die doet vervolgens iets ergens in het lichaam. Het is een hele keten van reacties en die is maar zeer partieel bekend. Een volledig dekkend model waarbij je alles snapt is niet realistisch. Toch wil je nieuwe medicijnen op de markt brengen en weten wat de juiste dosis is. Dus moet je met onvolledige kennis een model maken.

Het leuke is dat er in deze hoek steeds meer wiskunde nodig is. De simpele oplossingen zijn inmiddels wel gevonden, dus nu is er meer onderzoek nodig: Welk stukje van de keten moet je hebben? Sommige delen zijn buitengewoon stabiel: daar kun je aan rukken en er gebeurt niets. Andere plekken zijn een kantelpunt en enorm gevoelig voor veranderingen. Wat is de optimale plek om in te grijpen met je medicijn? Daar kun je met wiskundige modellen achter komen.

In de farmacologie is het criterium voor een goed model dat het een correcte voorspelling geeft. Ik heb ontdekt dat dit een even hard criterium is als alles willen snappen: je wordt afgestraft als het fout is. Niet intellectueel, maar door de consequenties. De wetenschappelijke strengheid is heel groot in dit vakgebied, maar hij ligt op een ander terrein dan pure kennis. Daar moest ik erg aan wennen."

Wat maakt deze modellen voor u de moeite waard?
"De uitdaging. Je hebt een dataset en wat zit er in hemelsnaam achter? Dit is een uitstekend vak voor een emeritus, want je moet heel veel praten met die farmacologen. Je kunt dit niet op je studeerkamer doen en *ins Blaue hinein* fantaseren, je moet wel een beetje weten wat er in het lichaam gebeurt. Met Meindert heb ik veel gepraat over zijn werk. Bovendien had ik het grote geluk dat Johan Gabrielsson naar Leiden kwam om een cursus te geven, hij is de Zweedse onderzoeker uit dat artikel waarmee het allemaal begon. Wij bleken het ontzettend goed te kunnen vinden. Gabrielsson is mijn promotor geworden, om het maar zo te noemen. Hij weet gigaveel van farmacologie en beseft goed hoe belangrijk wiskunde daarin kan zijn. Ik heb eindeloos veel van hem geleerd. Zo iemand heb je nodig als je in een nieuw vakgebied komt, dat kun je allemaal niet uit artikelen halen.

Hoewel mijn eigen achtergrond ontzettend zuiver is, vind ik het erg leuk om hieraan te werken. Kloosterman had altijd een minachting voor toepassingen en ik ben het daar helemaal niet mee eens. Het is heel spannend en de intellectuele bevrediging is anders, maar net zo groot. Je hebt vragen die zó ontzettend moeilijk zijn, alleen heb je ze niet zelf bedacht."

Heeft u spijt dat u dit niet eerder hebt ontdekt?
"Nee, ik heb altijd ongelooflijk leuk werk gedaan. Ik ben door mijn ongedurigheid wel vaker een beetje veranderd van onderwerp. En toen kwam dit voorbij en het was aardig om van richting te veranderen ter ere van mijn emeritaat. Het heeft het sociale nadeel dat ik van mijn oude vriendjes niet meer zo op de voet volg wat ze wetenschappelijk doen, maar het grote voordeel is dat ik in een heel leuke nieuwe sociale omgeving ben beland. De mensen zijn iets makkelijker om mee te praten dan wiskundigen. Discussies gaan vaak over interpretatie van modellen en meningen spelen een rol. Het is voor mij een nieuw vakgebied met iets meer politiek dan ik vanuit mijn wiskundige achtergrond gewend ben."

Doet het u pijn als u negatieve dingen over de farmaceutische industrie leest?
"Nee, ik begrijp het wel. Alleen heb ik een iets bredere blik en zie ik dat er in die wereld ook een heleboel idealisten zijn die dit vak juist kozen omdat ze mensen willen helpen. De onderzoekers uit de farmaceutische industrie waarmee ik samenwerk zijn niet te onderscheiden van mensen aan de universiteit, ze werken op dezelfde manier. Maar de CEO's en de manier waarop die bedrijven met mensen omgaan, daar word je niet vrolijk van."

Wat heeft u verbaasd over farmacologen?
"Ze hebben ongelooflijk hoge verwachtingen van wiskunde en computers. Ze kunnen inmiddels vrij goed vergelijkingen opstellen, maar die gaan vaak ongezien de computer in. Ze zijn verbaasd als er niet onmiddellijk iets uitkomt. Terwijl het rekenwerk makkelijk twee weken kan duren. Eén van mijn missies, je moet altijd een missie hebben, is om mensen duidelijk te maken dat het verstandig is als je eerst eens even kijkt naar dat ding. Dat je aan een vergelijking heus wel wat kunt zien zonder dat je een toetsenbord hebt aangeraakt."

Heeft u een goede suggestie voor een (Nederlandse) wiskundige met een bijzonder keerpunt in zijn of haar carrière? Stuur dan een e-mail naar *keerpunt@nieuwarchief.nl*.

Wim Caspers
Adelbert College, Wassenaar, en
Faculteit EWI en Lerarenopleiding, TU Delft
w.t.m.caspers@tudelft.nl

Onderwijs Bespreking examen vwo wiskunde B 2015

Over regels en ontregeling

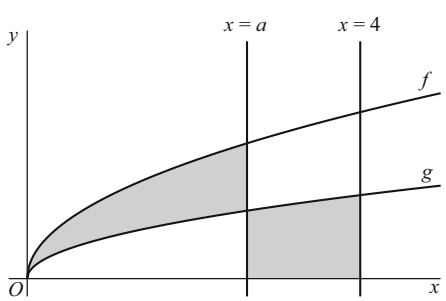
Afgelopen voorjaar werd er weer een nieuwe editie toegevoegd aan de rij centrale examens wiskunde B voor het vwo. Nog maar een paar te gaan voordat in 2018 het nieuwe examenprogramma centraal geëxamineerd gaat worden. Het examen uit het eerste tijdvak wordt besproken door Wim Caspers, met daarbij aandacht voor typisch examenidioom, vwo-regels en ontregelende vragen. En ook: werpen de wiskundige denkactiviteiten hun schaduw vooruit?

Als u werkzaam bent in het hoger onderwijs, en u heeft van doen met de eerstejaarsstudenten, dan is het centraal examen wiskunde B (zie [1]) dat ze afgelegd hebben misschien niet het eerste gespreksonderwerp dat u te binnen schiet. Toch is het aardig om te weten welke hindernis ze geslecht hebben om toegang te verkrijgen tot uw collegezaal of klaslokaal.

Examenidioom

Waarschijnlijk kunnen ze zich met enige moeite de openingsopgave nog wel herinneren. In Figuur 1 is de kern van de opgave weergegeven: "bereken exact voor welke waarde van a deze vlakdelen gelijke oppervlakte hebben", waarbij gegeven was dat $f(x) = \sqrt{x}$ en $g(x) = \frac{1}{2}\sqrt{x}$. Hier komen we al meteen examenidioom tegen. "Bereken exact" betekent op het vwo: "Stap voor stap, zonder gebruik te maken van specifieke opties van de grafische rekenmachine; de antwoorden mogen niet benaderd worden." Dit in tegenstelling tot "Bereken", want dan is de wijze van berekenen vrij. Iets dergelijks is het onderscheid tussen "Bewijs" en "Toon aan". Als u in hetzelfde dialect wiskunde wilt spreken, raadpleeg dan de totale lijst van examenwerkwoorden in de 'Syllabus centraal examen 2015 wiskunde B vwo' [2] (zie Figuur 2 voor een gedeelte ervan). Ook handig wanneer u studenten hun accent juist wilt afleren.

Een ander typisch examenfenomeen openbaart zich in opgave 4 en 6 en ook in opga-



Figuur 1 Figuur bij opgave 1.

woord	toelichting
aantonen	een redenering en/of berekening waaruit de juistheid van het gestelde blijkt In het algemeen geldt dat het gestelde controleren door middel van een of meer voorbeelden niet voldoet.
afleiden (van een formule)	een redenering en/of berekening waaruit de juistheid van een formule blijkt In het algemeen geldt dat de formule controleren door middel van een of meer voorbeelden niet voldoet .
aflezen	het antwoord is voldoende
algebraïsch	stap voor stap, zonder gebruik te maken van specifieke opties en de grafische mogelijkheden van de grafische rekenmachine; tussenantwoorden en eindantwoord mogen benaderd worden
bepalen	de wijze waarop het antwoord gevonden wordt is vrij; een toelichting is vereist
berekenen	de wijze van berekenen is vrij; een toelichting is vereist de De toevoeging 'algebraïsch' of 'exact' legt beperkingen op aan de wijze van berekenen.
bewijzen	een redenering en/of exacte berekening waaruit de juistheid van het gestelde blijkt In het algemeen geldt dat het gestelde controleren door middel van een of meer voorbeelden niet voldoet.
exact	stap voor stap, zonder gebruik te maken van specifieke opties en de grafische mogelijkheden van de grafische rekenmachine; de antwoorden mogen niet benaderd worden

Figuur 2

Helderheid van sterren

Aan de sterrenhemel bevinden zich heldere en minder heldere sterren. De **helderheid** van een ster werd in de oudheid reeds aangegeven met een getal, de **magnitude** van de ster. Zeer heldere sterren kregen magnitude 1. Nauwelijks zichtbare sterren kregen magnitude 6. Een kleine waarde betekent dus een grote helderheid. In deze opgave is m de magnitude.

Tegenwoordig meet men de hoeveelheid licht die van een ster wordt ontvangen. De helderheid van een ster wordt dan vaak uitgedrukt in **lux** (een eenheid voor verlichtingssterkte). In deze opgave is L de helderheid in lux.

In de tabel staan voor een aantal helderheden de waarden van m en L.

tabel

m	1,0	2,0	3,0	4,0	5,0	6,0
L	$1{,}0\cdot10^{-6}$	$4{,}0\cdot10^{-7}$	$1{,}6\cdot10^{-7}$	$6{,}3\cdot10^{-8}$	$2{,}5\cdot10^{-8}$	$1{,}0\cdot10^{-8}$

Tussen L en m bestaat een exponentieel verband van de vorm $L=10^{p+qm}$.

4p **4** Leid uit de tabelgegevens bij $m=1{,}0$ en $m=6{,}0$ af dat $p=-5{,}6$ en $q=-0{,}4$.

Figuur 3 Opgave 4 van het examen.

ve 7. Afgezien van het gedoe met significante cijfers, die nog nooit een duidelijke plek in het examenprogramma hebben weten te veroveren, wordt er in die opgaven een mededeling gedaan, gevolgd door de vraag om het bewijs of de afleiding te geven. Het antwoord wordt gegeven, gevolgd door de opdracht het antwoord te vinden. Dat is begrijpelijk omdat er doorgerekend moet worden met de antwoorden. Zonder een dergelijk 'voorgezegd' antwoord zou het gevaar groot zijn dat de hele opdracht misloopt in geval van een klein foutje aan het begin. Naar het antwoord toewerken lijkt bij de huidige vraagstelling echter heel legitiem. Met name in opgave 4 (zie Figuur 3) ligt het voor de hand om eerst maar eens de gegeven p en q in te vullen om zo na te gaan of het klopt. Maar in een examencontext levert dat geen punten op.

Het geven van een getallenvoorbeeld wordt overigens over het algemeen niet op prijs gesteld in correctievoorschriften, behalve natuurlijk wanneer er een tegenvoorbeeld gevraagd wordt. Het geven van een bewijs voor een specifieke situatie wordt echter niet gezien als een waardevolle stap om het algemene geval te bewijzen. "Als voor p een waarde is ingevuld voor deze vraag geen scorepunten toekennen", staat er dan bijvoorbeeld dreigend bij de modeluitwerking. Het specifieke bewijs zal inderdaad niet volledig zijn in het algemeen, maar er kan toch zeer zinvol werk verzet zijn [3]. Terug naar het examen van afgelopen voorjaar.

Ontregelende elementen

De openingsopgave was als zodanig zeer geschikt: de te integreren functies redelijk eenvoudig, een duidelijke en herkenbare vraagstelling. De manier waarop leerlingen de opgave uitwerkten verschilde enorm. Niet het aantal fouten onderscheidde de leerlingen, maar vooral het aantal regels dat gebruikt werd voor het geven van een oplossing. Leerlingen die het hele jaar door beter gepresteerd hadden, schreven in een paar regels naar een oplossing toe terwijl anderen hele bladzijdes nodig hadden. Landelijk scoorden de leerlingen gemiddeld 90 procent van de te behalen punten.

De tweede opgave kende wat minder vertrouwde elementen en zal nogal wat leerlingen ontregeld hebben. In Figuur 4 ziet u hem afgebeeld. Te herkennen valt een opgave uit het herexamen van 2004, waar de bewegingsvergelijkingen van (het puntje van) de grote en de kleine wijzer van een klok werden gegeven. In deze nieuwe variant is de grote wijzer twee keer zo lang als de kleine wijzer en de grote wijzer draait twee keer zo snel rond als de kleine. Ditmaal wordt er natuurlijk niet gesproken van een klok. Bij een klok ligt het voor de hand dat dezelfde t gebruikt wordt voor grote en kleine wijzer. In deze opgave voelt het ongemakkelijker. Leerlingen zijn in hun voorbereiding op het examen niet vaak tegengekomen dat de t die gevonden wordt uit de ene set bewegingsvergelijkingen, gebruikt moet worden in de andere. Nog een ontregelend aspect is het opschrijven van een sinus achter $x(t)$ en een cosinus achter $y(t)$. Dat is naar, als je het andersom gewend bent.

Zo waren er nog wel meer ontregelende elementen; logaritmen die sowieso voor bescheiden paniek zorgen, notaties die net iets minder in de lesstof verankerd zitten en het onverwacht moeten toepassen van bekende regeltjes. In opgave 6 kwam de formule

$$m(x) = -14{,}0 - 2{,}5\log C + 5{,}0\log x$$

voor (inderdaad geen haakjes om C en x). De $5{,}0$ werd nogal eens voor grondtal van de logaritme versleten. En in opgave 7 werd de kettingregel verklapt in de vorm

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \frac{\mathrm{d}m}{\mathrm{d}x}\cdot\frac{\mathrm{d}x}{\mathrm{d}t}.$$

Blijkbaar werd gevreesd dat leerlingen niet op eigen kracht de regel zouden toepassen. Die helpende hand heeft niet iedereen herkend. Andere vragen waren weer juist verontrustend elementair. Een leerling zal de avond tevoren ingewikkelde integralen berekend hebben en het eenvoudig integreren van de constante functie niet verwacht hebben de volgende dag. Ook het moeten aanroepen van de stelling van Pythagoras is onverwacht eenvoudig, evenals een parabool gecombineerd met een bissectrice. Door deze minder gebruikelijke elementen bleven er weinig direct herkenbare vragen over voor leerlingen die vooral op routine varen; als er 'dit' staat moet ik 'dat' doen. Het uitwerken van de opgaven viel mee, maar de herkenbaarheid was wat minder groot.

De ontregeling bereikt een hoogtepunt in opgave 9 over gelijke hoeken (zie Figuur 6). De lijst met stellingen en definities die aangeroepen mogen worden, is in het plaatje bijna volledig toepasbaar. Er is een koordenvierhoek, meerdere zelfs, er zijn congruente driehoeken, constante hoeken, middelpuntshoeken. De stelling van Thales is toepasbaar en het lijkt er ook op dat de hoekensom van een driehoek uitkomst kan bieden. Nu ziet de leerling die routinematig te werk gaat zich voor een probleem geplaatst. In de oefenopgaven is er vaak sprake geweest van een duidelijke openingszet en vaak ook een voor de hand lig-

Cirkels en lijnstuk

Over de cirkel met middelpunt $(0, 0)$ en straal 1 beweegt een punt A met bewegingsvergelijkingen:

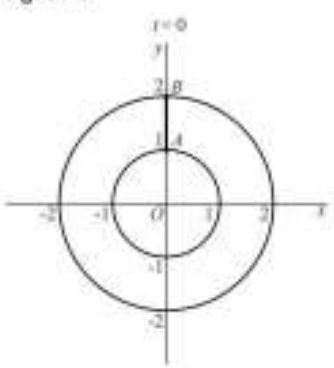
$$\begin{cases} x(t) = \sin t \\ y(t) = \cos t \end{cases} \text{ met } 0 \le t \le 2\pi$$

Over de cirkel met middelpunt $(0, 0)$ en straal 2 beweegt een punt B met bewegingsvergelijkingen:

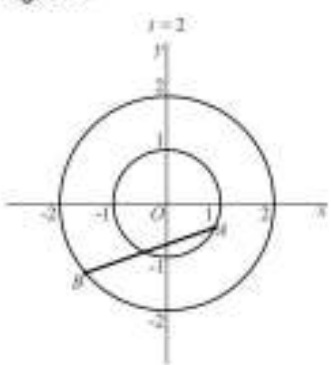
$$\begin{cases} x(t) = 2\sin(2t) \\ y(t) = 2\cos(2t) \end{cases} \text{ met } 0 \le t \le 2\pi$$

In de figuren 1 en 2 zijn de twee cirkels en het lijnstuk AB getekend voor de tijdstippen $t = 0$ en $t = 2$.

figuur 1



figuur 2



Op de tijdstippen waarop B zich op de x-as bevindt, bevindt A zich op de lijn met vergelijking $y = x$ of op de lijn met vergelijking $y = -x$.

2 Bewijs dit.

Figuur 4 Opgave 2 van het examen.

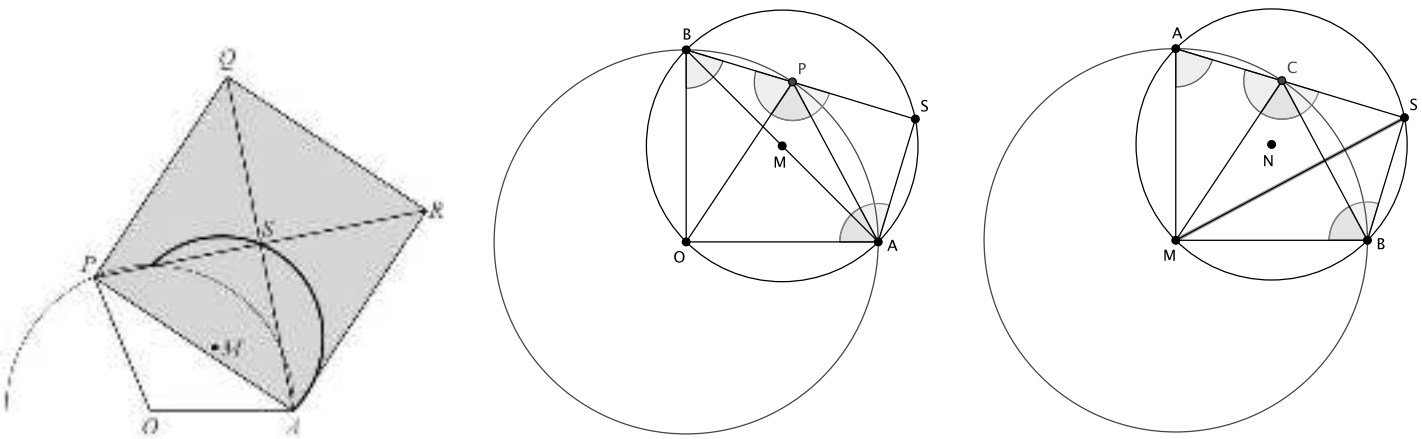
gend slotargument. Zo speelden in geval van een gelijkbenige driehoek de gelijke basishoeken altijd een rol en als er moest worden aangetoond dat punten op eenzelfde cirkel liggen, dan was de koordenvierhoekstelling onvermijdelijk. Het begin en einde van het bewijs stonden daarmee redelijk vast. Die zekerheid wordt bij deze opgave uit handen geslagen. Het plaatje is vergeven van de bijzondere eigenschappen en gelijke hoeken. Veel wegen leiden naar Rome, en er zijn heel veel omwegen en wegen die nergens toe leiden. Wat nu? De ontregeling is een feit.

Wiskundige denkactiviteiten

In het kader van de wiskundige denkactiviteiten (WDA's) is dit natuurlijk een prettige opgave. WDA's zullen in het nieuwe examenprogramma meer aandacht krijgen, dus misschien is dit wel een voorproefje. Maar het onderwerp meetkunde zal in het nieuwe examenprogramma veel meer met coördinaten beoefend worden. Het redeneren en bewijzen in de meetkunde zoals dat nu aan de orde komt, verdwijnt dan grotendeels. Het maakt wel nieuwsgierig naar de manier waarop dat dan in de examens tot uitdrukking komt.

Er worden nu al pilotexamens afgenomen en in het pilotexamen van 2014 vinden we een meetkundeopgave die niet in het reguliere examen stond (zie Figuur 5). Het linkerplaatje werd gegeven en gevraagd werd aan te tonen dat wanneer P beweegt over de cirkel, dat S dan ook een cirkel beschrijft. Met behulp van coördinaten gaat dat redelijk eenvoudig. Het is niet helemaal een routineopgave, maar hoe verloopt het bewijs zonder gebruik te maken van coördinaten? Met andere woorden, zou de opgave ook onder het huidige examenprogramma gevraagd kunnen worden? GeoGebra biedt uitkomst.

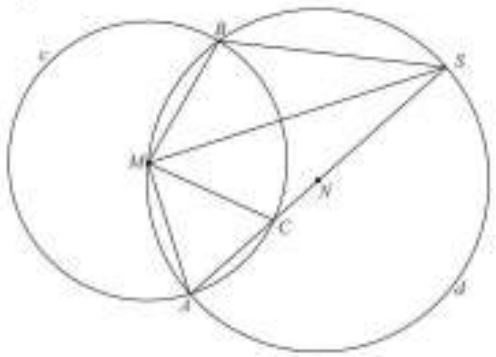
Het plaatje teruggebracht tot zijn essentie levert het middelste plaatje op. In het rechterplaatje zijn de punten van een andere naam voorzien. En daar verschijnt een plaatje dat in wezen de situatie van de meetkundeopgave uit 2015 weergeeft. Met dat verschil dat N nu binnen de cirkel ligt in plaats van erbuiten zoals nadrukkelijk vermeld werd. Een jaar geleden was de opgave in principe dus al te vinden. Misschien loont het de moeite



Figuur 5 Afbeelding bij meetkundeopgave in het pilotexamen van 2014 (links) en uitwerkingen met GeoGebra.

De hierboven beschreven situatie geldt ook in figuur 2. Punt S is nu zo gekozen dat lijnstuk AS door N gaat. Het snijpunt van AS en cirkel c is het punt C. Figuur 2 staat ook op de uitwerkbijlage.

figuur 2

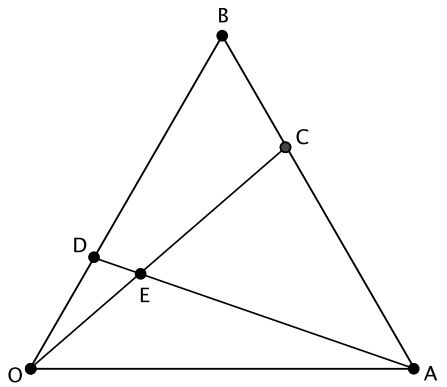


9p 9 Bewijs dat $\angle AMC = \angle ASB$.

Figuur 6 Opgave 9 van het examen.

voor kandidaten die in 2016 examen doen, om de meetkundeopgave uit het pilotexamen 2015 met GeoGebra grondig binnenstebuiten te keren.

Het levert de volgende voorspelde opgave op voor het examen 2016: Gegeven is een gelijkzijdige driehoek OAB. Op zijde AB ligt het punt C zo dat $AC = \frac{2}{3} \cdot AB$, en op zijde BO ligt het punt D zo dat $BD = \frac{2}{3} \cdot BO$. Punt E is het snijpunt van de lijnstukken OC en AD. Zie Figuur 7. Bewijs dat de punten D, E, C en B op een cirkel liggen met middellijn BD. Over een jaar weten we of deze voorspelling uitkomt.



Figuur 7

De afgedrukte opgaven en losse figuren zijn afkomstig uit het eindexamen vwo wiskunde B, eerste tijdvak, 13 mei 2015, en uit het pilot-eindexamen vwo wiskunde B, eerste tijdvak, 20 mei 2014.

Referenties

1 www.examenblad.nl/examen/wiskunde-b-vwo/2015.
2 www.examenblad.nl/examenstof/syllabus-2015-wiskunde-b-vwo/2015.
3 Jeroen Spandaw en Wim Caspers, Voorbeeldige bewijzen (artikel in voorbereiding).

Gewijzigde beoordeling wiskunde A en C: breien en spookhaakjes

Eindexamens worden altijd nagekeken volgens een zogeheten correctievoorschrift (CV). Het CV bevat een beoordelingsmodel waarin per opgave één of meerdere juiste oplossingen worden genoemd, met daarbij een weergave van de hoeveelheid scorepunten die elk onderdeel van deze oplossing waard is. Daarnaast bevat het CV enkele algemene regels en vakspecifieke regels, die uitleggen hoe er met het beoordelingsmodel moet worden omgegaan. De vakspecifieke regels bij wiskunde vermelden bijvoorbeeld dat iedere rekenfout tot 1 scorepunt aftrek leidt, en dat het gebruik van de grafische rekenmachine om tot een antwoord te komen altijd moet worden toegelicht.

Op basis van dit recept zou iedere docent in het land op gelijke wijze kunnen nakijken, zo lijkt het tenminste. Toch blijkt dat soms tegen te vallen. Met name bij redeneeropgaven is het maar de vraag in hoeverre een gebrekkig geformuleerde argumentatie nog wat oplevert. Ook notatiefouten leiden nogal eens tot discussie: moet een leerling bijvoorbeeld een scorepunt verliezen als hij het aantal manieren om 3 objecten uit 10 objecten te kiezen noteert als $\left(\frac{10}{3}\right)$, ondanks dat hij vervolgens het antwoord 120 geeft en dus duidelijk niet de breuk $\frac{10}{3}$ maar de binomiaalcoëfficient $\binom{10}{3}$ bedoelt? En wat te doen bij de berekening $\frac{1}{6}^2 = \frac{1}{36}$, terwijl de correcte notatie uiteraard $\left(\frac{1}{6}\right)^2 = \frac{1}{36}$ is?

Met het oog op deze problematiek is er met ingang van het schooljaar 2014–2015 voor wiskunde A en C bij zowel havo als vwo een nieuwe vakspecifieke regel toegevoegd aan het CV: “Als de kandidaat bij de beantwoording van een vraag een notatiefout heeft gemaakt en als gezien kan worden dat dit verder geen invloed op het eindantwoord heeft, wordt hiervoor geen scorepunt in mindering gebracht.”

Toelichting op de regel

Voor ieder vak is een *vaststellingscommissie* van het College voor Toetsen en Examens verantwoordelijk voor het vaststellen van de opgaven en correctievoorschriften van de centrale examens. In een recent artikel in vakblad *Euclides*, getiteld ‘Gelijke monniken, gelijke kappen’, verduidelijkte de vaststellingscommissie havo wiskunde A en vwo wiskunde A en C onder andere de achtergrond en interpretatie van deze nieuwe regel (zie www.nvvw.nl/19661/90-3-euclides-p16-p20). Het doel van de nieuwe regel en de toelichting erop lijkt tweeledig. Enerzijds wordt vermeld dat men vindt dat leerlingen een gelijke

beoordeling verdienen. Anderzijds wordt aangevoerd dat leerlingen bij wiskunde A en C niet worden opgeleid om actief wiskundige notaties te kunnen gebruiken, maar om (wiskundige) problemen in betekenisvolle contexten op te kunnen lossen en hun oplossing te kunnen onderbouwen. Het foutloos gebruik van wiskundetaal wordt daarbij als ondergeschikt beschouwd.
Het gaat dus bij wiskunde A en C voortaan voornamelijk om de *gedachtegang* van de leerling om een correct antwoord te bereiken, niet zozeer om een correcte formele beschrijving hiervan. "Uitgangspunt is dat er geen scorepunten in mindering gebracht moeten worden, als een leerling een notatiefout gemaakt heeft bij de beantwoording van een vraag, terwijl gezien kan worden dat hij correct gehandeld heeft bij de daaropvolgende stappen", aldus de commissie.
Op basis van deze nieuwe richtlijn dienen docenten vanaf het afgelopen examen bijvoorbeeld geen punten meer in mindering te brengen als een leerling de vermenigvuldiging van $p = 3$ en $q = x + 1$ noteert als $p \cdot q = 3 \cdot x + 1 = 3x + 3$, aangezien immers uit het antwoord kan worden afgeleid dat de leerling wel heeft doorgerekend "alsof er haakjes staan". In wiskundeland wordt al gesproken over 'spookhaakjes'. Ook het bekende 'breien', waarbij bijvoorbeeld de berekening van $1 + 2 + 4$ genoteerd wordt als $1 + 2 = 3 + 4 = 7$, wordt niet meer fout gerekend. En bij het berekenen van een groeipercentage per uur van een kolonie bacteriën die in twee uur verdubbelt komt de leerling ongeschonden weg met de notatie $2^{\frac{1}{2}} = 1,41 = 41\%$. Wenselijk, of niet?

Aanhoudende problemen
Hoewel een gelijke beoordeling voor alle leerlingen natuurlijk een prima uitgangspunt is, lijkt de discussie over het al dan niet aanrekenen van (notatie)fouten er niet veel minder om. De grens tussen wel of geen puntenaftrek is wellicht verschoven, maar lijkt in sommige gevallen nog steeds even onduidelijk.
Zo is het met name bij opgaven waar het antwoord al is verwerkt in de vraagstelling ("Bewijs dat $f(x) = \ldots$") lastig om te bepalen of een leerling een notatiefout of een denkfout heeft gemaakt. Bij vwo wiskunde A moest dit jaar bijvoorbeeld van een piramide worden aangetoond dat de inhoud gelijk is aan $I = 6x - \frac{2}{3}ax^2$, op basis van de gegeven formule $I = \frac{1}{3} \cdot$ *oppervlakte grondvlak* $\cdot$ *hoogte*, een grondvlak van 2 bij x en een hoogte van $h = 9 - ax$. Het correctievoorschrift is weergegeven in Figuur 8.

12	**maximumscore 3**	
	• De oppervlakte van het grondvlak is $2x$	1
	• $I = \frac{1}{3} \cdot$ *oppervlakte grondvlak* $\cdot$ *hoogte* geeft $I = \frac{1}{3} \cdot 2x \cdot (9 - ax)$	1
	• Dit geeft $I = 6x - \frac{2}{3}ax^2$	1

Figuur 8 Correctievoorschrift bij wiskunde A 2015 (1e tijdvak), opgave 12.

Hoewel de meerderheid deze uitwerking gelukkig volledig correct noteert, geeft een enkeling als antwoord $I = \frac{1}{3} \cdot 2 \cdot x \cdot 9 - ax = 6x - \frac{2}{3}ax^2$. Aan de ene kant kan hier natuurlijk geredeneerd worden dat er is doorgerekend alsof de haakjes er staan, en het eindantwoord is correct. Gezien dat het antwoord al was gegeven in de opgave, zou het ook kunnen dat de leerling gewoon maar het antwoord heeft genoteerd waarvan hij wist dat hij ernaartoe moest werken, zonder te weten waarom dit volgt uit het voorgaande. Het lijkt voor de hand te liggen om hier punten in mindering te brengen, omdat uit het vervolg niet is af te leiden dat de leerling juist heeft gehandeld. Gezien de formulering van de nieuwe regel is het echter maar de vraag of elke docent hier ook zo zou redeneren.
Een ander soort onduidelijkheid doet zich voor als de notatie onderdeel is van het eindantwoord. Een tweetal voorbeelden uit het *Euclides*-artikel geeft aan hoe subtiel het verschil soms kan zijn. Uitgaande van een hypothetische opgave om $y = \frac{10}{x}$ te differentiëren worden twee uitwerkingen beschouwd:

$$y' = \tfrac{10}{x} = 10x^{-1} = -10x^{-2},$$
$$y = \tfrac{10}{x} = 10x^{-1} = -10x^{-2}.$$

De eerste uitwerking wordt volledig goed gerekend, in de categorie 'breien'. Hier wordt aangenomen dat het duidelijk is wat de leerling doet: hij herschrijft eerst en differentieert daarna, en heeft dit onjuist genoteerd; hij bedoelt $y = \frac{10}{x} = 10x^{-1}$ en dus $y' = -10x^{-2}$. De tweede uitwerking levert echter geen punten op, met de argumentatie dat "onduidelijk is of de leerling inderdaad de afgeleide berekend heeft." Gezien de vraag om te differentiëren lijkt het hier echter toch net zo duidelijk hoe de leerling gehandeld heeft als in het geval dat het accent juist te vroeg in plaats van niet genoteerd is, wat de keuze om verschil te maken tussen deze twee fouten op z'n minst discutabel maakt.

Conclusie
Al met al is het dus maar de vraag of er nu meer duidelijkheid en consistentie is gekomen. Het voordeel voor de leerling is dat een antwoord waaruit blijkt dat de oplossingsstrategie volledig is begrepen ook inderdaad leidt tot de volle punten. Bij notatiefouten zoals breien is het bovendien voor de corrector in sommige gevallen nu duidelijker dan voorheen dat er geen puntenaftrek hoeft plaats te vinden. De taak van de corrector verschuift in andere gevallen echter van het beoordelen wat iemand *noteert* naar het beoordelen wat iemand *bedoelt*. Met het formele en exacte karakter van wiskunde in het achterhoofd is het maar de vraag of we die kant op zouden moeten gaan. Het goed formuleren en gebruik van correcte wiskundige notatie is immers ook een belangrijk onderdeel van het curriculum, en het volledig loslaten hiervan zou een aanzienlijke verarming van het wiskundeonderwijs betekenen.
De examenmakers geven in hun toelichting overigens aan te veronderstellen dat docenten hun leerlingen wel leren om antwoorden wiskundig te formuleren, en het merendeel zal notatiefouten in de jaren voorafgaand aan het examen ook gewoon aanrekenen. Wat dat betreft wordt de wiskundige precisie die bij ons vak hoort dus alsnog wel aangeleerd. Door dit echter vervolgens op het eindexamen niet meer te beoordelen, ontstaat er ten eerste een discrepantie tussen het gegeven onderwijs en de uiteindelijke toetsing hiervan, en wordt ten tweede het belang van wiskundige notatie en precisie in formulering minder onderkend. Doe je daar niet de wiskunde mee tekort, evenals de leerling die zich in zijn schoolcarrière wel heeft vastgebeten in het aanleren van correcte wiskundige formulering?
Ter geruststelling: de meeste leerlingen waarvan het werk ons onder ogen is gekomen noteren hun antwoord gelukkig keurig, zodat het effect van de gewijzigde nakijkregels (vooralsnog) beperkt blijft.

Mark Timmer en Marjan Schutmaat
docenten wiskunde aan het Carmel College Salland te Raalte.

In de verdediging

| In defence

Pas gepromoveerden brengen hun werk onder de aandacht. Heeft u tips voor deze rubriek of bent u zelf pas gepromoveerd? Laat het weten aan onze redacteur.

Redacteur: Geertje Hek
la Voie-du-Coin 7
1218 Grand-Saconnex
Zwitserland
verdediging@nieuwarchief.nl

Intuitionistic Rules: Admissible Rules of Intermediate Logics
Jeroen Goudsmit

Jeroen Goudsmit behaalde bachelordiploma's wiskunde, informatica en cognitieve kunstmatige intelligentie en een masterdiploma wiskunde voor hij werd aangenomen in het Vidi-project 'The power of constructive proofs' van dr. Rosalie Iemhoff. Onder begeleiding van haar en van prof.dr. Albert Visser schreef hij het proefschrift *Intuitionistic Rules: Admissible Rules of Intermediate Logics* dat hij op 29 mei 2015 succesvol verdedigde. Hij werkte bij de theoretische filosofiegroep van de faculteit Geesteswetenschappen aan de Universiteit Utrecht, maar benadrukt dat zijn werk meer wiskundig dan filosofisch van aard is.

Tussen klassieke en intuïtionistische logica

Om een beschrijving te geven van de inhoud van zijn proefschrift vertelt Goudsmit eerst over verschillende logische stromingen: "Wiskundigen zien patronen en proberen deze te vangen in stellingen. Een stelling wordt sluitend beargumenteerd, waarmee ze als bewezen wordt beschouwd. Maar niet alle wiskundigen zijn het met elkaar eens over wat een sluitend argument is. Hoewel de meesten gebruikmaken van de klassieke logica, bestaan er verscheidene andere stromingen, elk met een eigen blik op wat een sluitend argument is. Een van die andere stromingen is het *intuïtionisme* waarvan Brouwer de grondlegger was."

Een logica wordt gegeven door axioma's en afleidingsregels. Een *axioma* is een beginaanname, een voor bewezen aangenomen uitspraak binnen de logica. Met een *afleidingsregel* bewijst men een uitspraak uit al eerder bewezen uitspraken. De axioma's en afleidingsregels van een logica geven tezamen een tweedeling tussen alle uitspraken: de uitspraken die bewezen kunnen worden, en de uitspraken die niet bewezen kunnen worden. Intuïtionistische logica en klassieke logica zijn aan elkaar verwant: ze delen dezelfde afleidingsregels, alleen kent klassieke logica één extra axioma, de 'wet van de uitgesloten derde'. De wet van de uitgesloten derde is een logische wet die inhoudt dat iedere uitspraak waar of onwaar is; een andere, derde, mogelijkheid is er niet. De 'uitgesloten derde' is dus iedere andere denkbare waarheidswaarde. Een logica die voldoet aan de wet heet klassiek.

Het al dan niet volgen van de wet van de uitgesloten derde heeft aanzienlijke implicaties voor de acceptatie van bepaalde bewijzen, met name bewijzen uit het ongerijmde. Goudsmit: "Een bewijs in intuïtionistische logica zegt niet alleen dat iets bestaat, maar ook hoe het te vinden is. Men zegt ook wel dat een bewijs computationele inhoud heeft: bewijzen zijn algoritmes. In klassieke logica is het veel lastiger om een algoritme te extraheren uit een bewijs."

Vanwege de verwantschap tussen intuïtionistische en klassieke logica vormen de bewezen uitspraken in de intuïtionistische logica een deelverzameling van die in de klassieke logica. Er bestaat een continuüm van logica's waarvan de bewezen uitspraken hier tussenin liggen, maar die ook precies dezelfde afleidingsregels kennen. Het proefschrift van Goudsmit richt zich op deze *intermediaire logica's*.

Toelaatbare regels

Gegeven een bepaalde logica heeft elke uitspraak een label: ofwel de uitspraak kan bewezen worden, ofwel dit kan niet. Als er een axioma toegevoegd zou worden aan de logica, dan kan de labeling veranderen. Immers, als dit axioma niet al een bewezen uitspraak (een stelling) was, is ze dat nu zeker wel. Als er een regel toegevoegd wordt aan de logica, dan kan de labeling eveneens veranderen. De *toelaatbare regels* van een logica zijn regels die aan de logica toegevoegd kunnen worden als afleidingsregel zonder dat dat nieuwe stellingen oplevert, ofwel regels die de labeling niet veranderen.

In klassieke propositielogica is de notie van een toelaatbare regel niet bijster interessant. Dit komt doordat klassieke logica structureel volledig is: elke toelaatbare regel is equivalent aan een stelling. In intuïtionistische logica ligt de situatie veel subtieler: tal van toelaatbare regels hebben niets van doen met de gegeven stellingen. Dat leidt tot allerlei vragen: is de verzameling van toelaatbare regels beslisbaar, zijn er meerdere logica's die dezelfde toelaatbare regels hebben als intuïtionistische logica, bestaan er basale regels waar alle toelaatbare regels uit volgen? In zijn proefschrift bestudeert Goudsmit de toelaatbare regels van intermediaire logica's in het licht van deze vragen.

Voor het feit dat in de klassieke propositielogica 'elke toelaatbare regel equivalent is aan een stelling' geeft Goudsmit hier een argument, gebruikmakend van waarheidstafels die in de meeste inleidingen tot de logica worden onderwezen: Bekijk een willekeurige regel A/B, waar A en B propositieformules zijn, en stel dat de implicatie $A \to B$ niet waar is in klassieke propositielogica. Dit betekent dat er een waarheidstafel v bestaat die $A \to B$ weerlegt. Zo'n waarheidstafel is een erg simpel semantisch object, dat je in de propositielogica kunt encoderen met een substitutie, zeg σ. Preciezer: voor elke formule C geldt dat '$\sigma(C)$ is bewijsbaar' equivalent is aan de waarheidswaarde van C onder v. Omdat v de implicatie $A \to B$ weerlegt, moet v de formule A waar maken en B onwaar ('als A dan B' is niet waar dan en slechts dan als A waar en B niet waar is). Nu volgt dat $\sigma(A)$ bewijsbaar is in klassieke logica, en $\sigma(B)$ dat niet is. Dus is de regel A/B niet toelaatbaar.

Hiermee uit het ongerijmde bewezen: als A/B toelaatbaar is, dan is de overeenkomstige formule $A \to B$ een stelling. De implicatie de andere kant op is triviaal. Stel $A \to B$ is een stelling (dus bewijsbaar), en we willen aantonen dat A/B toelaatbaar is. Neem een willekeurige substitutie σ, en stel $\sigma(A)$ is bewijsbaar. Omdat $A \to B$ bewijsbaar is, is $\sigma(A \to B) = \sigma(A) \to \sigma(B)$ dat ook. Omdat $\sigma(A) \to \sigma(B)$ en $\sigma(A)$ bewijsbaar zijn, is $\sigma(B)$ dat ook. Klaar.

Visser-regels en hun geschiedenis

Waarheidstafels zijn een systematische manier om een uitspraak in klassieke propositielogica (syntaxis) om te zetten naar een waarheidswaarde (semantiek). Zo kan men testen of een uitspraak waar is, simpelweg door de waarheidstafel te controleren. Er bestaat ook een verfijndere notie van semantiek voor intuïtionistische logica, in de vorm van Kripke-modellen. Goudsmit behandelt veralgemeniseringen van dit begrip, teneinde semantiek te geven, niet aan de bewijsbaarheid van een formule, maar *aan de toelaatbaarheid van een regel*. Deze noties komen van pas bij het beantwoorden van alle hierboven genoemde vragen. In grote tegenstelling tot semantiek voor stellingen, volstaat het bij semantiek voor toelaatbare regels niet om slechts eindige modellen te bekijken. Hiervoor geeft hij een elementair argument.

Goudsmit vindt stelling 7.33 van zijn proefschrift het leukst. Deze zegt dat de logica van een ten hoogste n-voudige vertakking de maximale intermediaire logica is waarvoor de eerste n Visser-regels toelaatbaar zijn. Minder technisch: de stelling karakteriseert een aftelbare reeks van intermediaire logica's door middel van hun toelaatbare regels.

De Visser-regels vormen een basis voor de toelaatbare regels van de intuïtionistische propositielogica. De n-de Visser-regel is:

$$\left(\bigvee_{i=1}^{n} x_i \to z\right) \to \bigvee_{j=1}^{n} x_j \Big/ \bigvee_{j=1}^{n} \left(\left(\bigvee_{i=1}^{n} x_i \to z\right) \to x_j\right).$$

De regels zijn genoemd naar Goudsmits promotor Albert Visser, maar zijn door meerdere mensen ontdekt, in verschillende contexten. De eerste ontdekker die Goudsmit kon vinden was Alex Citkin, die de regels in 1979 in een abstract voor een Russische conferentie omschreef. Albert Visser en Dick de Jongh dachten in de jaren tachtig over toelaatbare regels in de context van Heyting's rekenkunde, en in deze setting zijn de Visser-regels ontstaan. Omdat ze geen specifiek resultaat over deze regels hadden, waren ze toen niet gepubliceerd. Ook Skura en Rozière gebruikten varianten van de regels, voordat Rosalie Iemhoff ze in 2001 voor het eerst 'Visser-regels' noemde. Er waren tot kort geleden nauwelijks artikelen die het werk van Citkin en Skura noemden, omdat het eerste simpelweg onbekend was, en omdat het verband tussen Skura's werk en toelaatbaarheid nog niet bekend was.

Goudsmit kreeg het idee om genoemde stelling te bewijzen toen hij op Skura's artikel uit 1989 stuitte waarin intuïtionistische propositielogica op een zeer elementaire manier gekarakteriseerd wordt door middel van een zogeheten refutatiesysteem. Dit resultaat is best onbekend, en refutatiesystemen worden niet veel meer gebruikt. Wat het voor Goudsmit zo bijzonder maakte is dat het refutatiesysteem (in zekere zin een bewijssysteem voor onbewijsbaarheid) gebruikmaakte van regels die leken op de Visser-regels die hij kende, maar dan op hun kop. En dat terwijl Skura de Visser-regels nooit gezien had, of überhaupt gewerkt had aan toelaatbare regels. Goudsmit wilde begrijpen wat het verband was tussen refutatiesystemen en toelaatbare regels, en hieruit is zijn stelling voortgekomen.

Gesprekken en ontmoetingen rondom logica

Goudsmit genoot van de extreme autonomie die hij als promovendus had. Het ongebonden zoeken was erg leuk, maar soms ook eenzaam. Vooral in de eerste maanden, toen hij nog niet echt *zijn* probleem had gevonden, vond hij dat lastig. Toen hij eenmaal een pakkend probleem door Iemhoff aangedragen kreeg, kwam hij goed op gang. Dit probleem werd zijn eerste artikel en vormde het begin van zijn tocht door toelaatbaarheid. De begeleiding was hecht en de logici werkten binnen de theoretische-filosofiegroep, waarbinnen een prettige sfeer hing. Ze lunchten bijvoorbeeld ongeveer dagelijks met de hele groep. Ook buiten het onderzoek zocht en vond Goudsmit diversiteit. Hij sloot zich al snel aan bij het Promovendi-overleg Utrecht (Prout), en werd in september 2013 de eerste promovendus in de Utrechtse Universiteitsraad sinds een goed decennium. Goudsmit: "Het was bijzonder interessant om vanuit deze hoek de universiteit mee te maken, en hier ook wat bij te kunnen dragen."

Het mooiste aan zijn onderzoekstijd vond hij echter de gesprekken en ontmoetingen rondom logica. Bijvoorbeeld toen Tom Skura op bezoek kwam en voor het eerst sprak met de Nederlandse logici die werken aan toelaatbaarheid. Of tijdens zijn eigen bezoek aan Vladimir Rybakov, om met hem te spreken over de gedachten achter zijn oorspronkelijke bewijs voor de beslisbaarheid van de toelaatbare regels van intuïtionistische logica. Ook de vele inspirerende gesprekken met Dick de Jongh, die telkens weer tot nieuwe ideeën leidden, waren erg motiverend. En al met al heeft dit alles geleid tot een mooi proefschrift.

Problemen

Problem Section

This Problem Section is open to everyone; everybody is encouraged to send in solutions and propose problems. Group contributions are welcome.
For each problem, the most elegant correct solution will be rewarded with a book token worth €20. At times there will be a Star Problem, to which the proposer does not know any solution. For the first correct solution sent in within one year there is a prize of €100.
When proposing a problem, please either include a complete solution or indicate that it is intended as a Star Problem.
Please send your submission by e-mail (LaTeX is preferred), including your name and address to problems@nieuwarchief.nl.
The deadline for solutions to the problems in this edition is 1 December 2015.

Problem A (folklore)

Determine

$$\sum_p \tfrac{1}{p} \prod_{q<p} (1 - \tfrac{1}{q}),$$

where p ranges over all prime numbers, and q ranges over all prime numbers less than p.

Problem B (proposed by Wouter Zomervrucht)

Let $\mathbb{N} = \{0, 1, \ldots\}$ denote the set of natural numbers, and let $\mathbb{N}^{2015}$ denote the set of 2015-tuples $\big(a(1), a(2), \ldots, a(2015)\big)$ of natural numbers. We equip $\mathbb{N}^{2015}$ with the partial order $\preceq$ for which $a \preceq b$ if and only if $a(k) \leq b(k)$ for all $k \in \{1, 2, \ldots, 2015\}$. We say that a sequence $a_1, a_2, \ldots$ in $\mathbb{N}^{2015}$ is *good* if for all $i < j$ we have $a_i \not\preceq a_j$.

– Show that all good sequences are finite.

We say that a sequence $a_1, a_2, \ldots$ in $\mathbb{N}^{2015}$ is *perfect* if it is good and for all i and for all $k \in \{1, 2, \ldots, 2015\}$ we have $a_i(k) \leq 2015i$.

– Does there exists a positive integer N such that all perfect sequences have length at most N?

Problem C (proposed by Marcel Roggeband)

The *(first) Bernoulli numbers* B_n for integers $n \geq 0$ are defined by the following recursive formula.

$$B_0 = 1,$$
$$B_n = -\sum_{i=0}^{n-1} \binom{n}{i} \frac{B_i}{n-i+1} \qquad \text{for } n > 0.$$

Show that the Bernoulli numbers satisfy the following identity for all $n > 1$:

$$B_n = n! \sum_{i=1}^{n-1} \sum_{\sigma} \frac{(-1)^{i-1}}{\sigma_1! \cdots \sigma_i!} (\tfrac{1}{2} - \tfrac{1}{\sigma_i + 1}).$$

In this sum, σ runs through all i-tuples $(\sigma_1, \ldots, \sigma_i)$ of integers such that $\sigma_1 + \cdots + \sigma_i = n + i - 1$ and $\sigma_j \geq 2$ for all j.

Edition 2015-1 We received solutions from Rik Bos, Josse van Dobben de Bruyn, José María Giral, Alex Heinis, José Hernández Santiago, Thijmen Krebs, Robert van der Waall and Jeroen Winkel.

Redactie:
Gabriele Dalla Torre
Christophe Debry
Jinbi Jin
Marco Streng
Wouter Zomervrucht

Problemenrubriek NAW
Mathematisch Instituut
Universiteit Leiden
Postbus 9512
2300 RA Leiden
problems@nieuwarchief.nl
www.nieuwarchief.nl/problems

22

Oplossingen

Solutions

Problem 2015-1/A (proposed by Raymond van Bommel and Julian Lyczak)
A commutative ring R is *charming* if every ideal of R is an intersection of maximal ideals. Prove that a Noetherian charming ring is a finite product of fields. Does there exist a charming ring that is not a product of fields?

Solution We received solutions from Rik Bos, Josse van Dobben de Bruyn, José María Giral and Jeroen Winkel. The following solution is based on that of José María Giral, who also receives the book token. The example used in the following solution was given by Rik Bos, Josse van Dobben de Bruyn and José María Giral.
We first show that a Noetherian charming ring is a finite product of rings. The following lemma on general charming rings will be useful for this.

Lemma. *Let R be a charming ring. Then every prime ideal of R is maximal.*

Proof. First note that every ideal of R is an intersection of maximal — in particular radical — ideals, so every ideal of R is radical. In particular, for all $r \in R$ we have $(r^2) = (r)$, so for all $r \in R$ there exists some $s \in R$ such that $r = sr^2$.
Now let $\mathfrak{p}$ be a prime ideal of R, and let $r \in R$ be an element such that $r \notin \mathfrak{p}$. Then for $s \in R$ such that $r = sr^2$, we have $r(1 - sr) = 0 \in \mathfrak{p}$. Therefore $1 - sr \in \mathfrak{p}$, from which we deduce that $\mathfrak{p} + rR = R$, since $(1 - sr) + sr = 1$. Hence $\mathfrak{p}$ is maximal. □

Let R be a Noetherian charming ring. We show that every ideal of R is a *finite* intersection of maximal ideals. Suppose for a contradiction that not every ideal of R is a finite intersection of maximal ideals. Consider the (non-empty) collection $\mathcal{I}$ of ideals I that are not finite intersections of maximal ideals. As R is Noetherian, any ascending chain of ideals in $\mathcal{I}$ stabilises, so the union of any chain of ideals in $\mathcal{I}$ is again an ideal in $\mathcal{I}$. Therefore by Zorn's Lemma, the collection $\mathcal{I}$ contains a maximal element. Denote such an element by I.
Note that by Lemma, the ideal I is not prime. Therefore there exist $x, y \in R$ such that $x, y \notin I$ and $xy \in I$. Let $J_1 = I + xR$ and $J_2 = I + yR$. As all ideals of R are radical, we have $J_1 \cap J_2 = J_1 J_2 = xyR + xR \cdot I + yR \cdot I + I^2 = I$. By maximality of I, both J_1 and J_2 are finite intersections of maximal ideals. Therefore so is I, but this is a contradiction. So all ideals of R are finite intersections of maximal ideals.
In particular, 0 is a finite intersection of maximal ideals $\mathfrak{m}_1 \cap \cdots \cap \mathfrak{m}_n$, and maximal ideals are pairwise coprime, so by the Chinese Remainder Theorem, we have $R \cong \prod_{i=1}^{n} R/\mathfrak{m}_i$, which is a finite product of fields.
For the second part, we show that the answer to the question is yes. We first show that all *Boolean* rings — rings in which every element is an idempotent — are charming. Let R be a Boolean ring. First note that every prime ideal of R is maximal; if $\mathfrak{p} \subseteq R$ is a prime ideal and $r \in R - \mathfrak{p}$, then $r^2 = r$, so $r(1-r) = 0 \in \mathfrak{p}$, hence $1 - r \in \mathfrak{p}$ and therefore $1 = (1-r) + r \in \mathfrak{p} + rR$.
Now we note that every ideal in R is radical; if $r \in R$ such that $r^n \in I$ for some positive integer n, then $r = r^n \in I$. Therefore I is the intersection of the prime (hence maximal) ideals containing I, showing that R is charming.
Let R be the subring of $\prod_{i=1}^{\infty} \mathbb{F}_2$ consisting of the elements $(a_i)_{i=1}^{\infty}$ such that either all but finitely many a_i are zero or all but finitely many a_i are one. Note that all elements of R are idempotents, so R is Boolean, hence charming. Also, the only fields of which all elements are idempotents are isomorphic to $\mathbb{F}_2$, so all quotients of R by maximal ideals and all subfields of R are isomorphic to $\mathbb{F}_2$. So if R is a product of fields, then it must be isomorphic to a product of $\mathbb{F}_2$. Since products of $\mathbb{F}_2$ are either finite or uncountable, and since R is countable, it follows that R is not a product of fields. Therefore R is a charming ring that is not a product of fields, as desired.

Problem 2015-1/B (folklore)
Let S be a set of prime numbers with the following property: for all $n \geq 0$ and distinct $p_1, \ldots, p_n \in S$ the prime divisors of $p_1 \cdots p_n + 1$ are also in S. Show that S contains all primes.

Solution We received solutions from Thijmen Krebs and Jeroen Winkel. The following solution is based on that of Thijmen Krebs, who also receives the book token.

Oplossingen

Solutions

Fix any prime q. We say that a prime p in S is *q-recurring* if S contains infinitely many primes that are congruent to p modulo q. Note that there are only finitely many primes p in S that are not q-recurring. Let x denote the product of all primes p in S that are not q-recurring. Let y be any finite product of distinct q-recurring primes in S. Then note that all prime divisors of $xy+1$ lie in S. As $\gcd(xy+1, x) = 1$, it follows by definition of x that moreover all prime divisors of $xy+1$ are q-recurring.

We can now define a sequence $(y_i)_{i=0}^{\infty}$ of products of q-recurring primes recursively by setting $y_0 = 1$, and by setting for $n \geq 1$ the number y_n to be the number obtained from $xy_{n-1}+1$ in the following way. Let $p_1 p_2 \cdots p_s$ be the prime factorisation of $xy_{n-1}+1$ (in which primes can occur multiple times). Pick for each p_i a prime p_i' in S that is congruent to p_i modulo q, in such a way that all p_i' *are* distinct; this is possible as there are infinitely many such p_i'. Then set $y_n = p_1' p_2' \cdots p_s'$. An inductive argument quickly shows that for all non-negative integers n, we have

$$y_n \equiv x^n + \cdots + x + 1 \mod q.$$

We now show that $q \in S$. If $x \equiv 0$ modulo q, then $q \in S$ by definition of x. If $x \equiv 1$ modulo q, then by the above we conclude that $q \mid y_{q-1}$ so $q \in S$ by definition of y_{q-1}. Finally, in the other cases, we can apply Fermat's Little Theorem to see that since $(x-1)y_{q-2} \equiv x^{q-1} - 1$ modulo q, we have $q \mid y_{q-2}$, so $q \in S$ by definition of y_{q-2}. Therefore $q \in S$, as desired.

Problem 2015-1/C (proposed by Roberto Stockli)
Determine all pairs (p, q) of odd primes with $q \equiv 3 \mod 8$ such that $\frac{1}{p}(q^{p-1} - 1)$ is a perfect square.

Solution We received solutions from Alex Heinis, José Hernández Santiago, Thijmen Krebs, Robert van der Waall and Jeroen Winkel. The following solution is based on that of Alex Heinis, who also receives the book token. In addition, we thank Robert van der Waall for bringing our attention to the article [1], in which one of the results is that the equation $\frac{1}{p}(m^{p-1} - 1) = a^2$ has a unique integral solution $(m, p, a) = (3, 5, 4)$ with m odd.

Note that $(p, q) = (5, 3)$ is a solution. We show that it is the only one.

Suppose that (p, q) is a solution. Then $(q^{(p-1)/2} + 1)(q^{(p-1)/2} - 1)/p$ is a square. Both factors on the left hand side are even as q is odd, so their greatest common divisor is 2. Therefore the two factors are of the forms $2a$ and $2pb$ for certain positive integers a, b, in no particular order, such that $\gcd(a, pb) = 1$. As ab is a square, it follows that both a and b are squares. Note that 2 is not a square modulo q as $q \equiv 3$ modulo 8. Therefore $q^{(p-1)/2} + 1$ cannot be twice a square. It follows that $q^{(p-1)/2} - 1$ is twice a square.

Hence $q^{(p-1)/2} - 1 = 2a$ and $q^{(p-1)/2} + 1 = 2pb$. Writing $a = k^2$ and $b = l^2$ for integers k, l, we can rewrite the above system of equations as $q^{(p-1)/2} = k^2 + pl^2$ and $1 = pl^2 - k^2$. Note that precisely one of k and l is even as precisely one of $q^{(p-1)/2} - 1$ and $q^{(p-1)/2} + 1$ is divisible by 4. If l were even, then $1 \equiv -k^2$ modulo 4, which is a contradiction. So l is odd, k is even, and therefore $p \equiv 1$ modulo 4. Hence we have a factorisation $(q^{(p-1)/4} + 1)(q^{(p-1)/4} - 1)$ of $2k^2$ with integer factors.

In the same way as above, we see that the two factors on the left hand side are of the forms $2c^2$ and $4d^2$ for certain positive integers c, d, in no particular order, and that $q^{(p-1)/4} + 1$ cannot be twice a square. Therefore we write $q^{(p-1)/4} - 1 = 2c^2$ and $q^{(p-1)/4} + 1 = 4d^2$. Hence $q^{(p-1)/4} = (2d-1)(2d+1)$. Note that $\gcd(2d-1, 2d+1) = 1$. As q is prime, it follows that since d is positive, we must have $2d - 1 = 1$, so $d = 1$. Hence $q = 3$ and $p = 5$. This shows that $(p, q) = (5, 3)$ is the only solution.

Reference

1. Hiroyuki Osada, Nobuhiro Terai, *Generalization of Lucas' Theorem for Fermat's quotient*, Comptes Rendus de Mathématiques de l'Académie des Sciences 01/1989; 11(4).