

# Inhoud | Contents

## Artikelen | Features

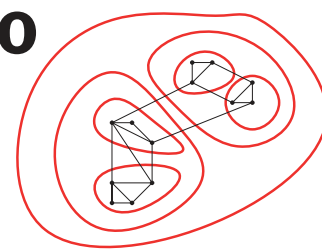
- 8 **De oplossing | The Solution**  
Het Cap Set-probleem *Aart Blokhuis, Dion Gijswijt*
- 11 **Nieuws | News**  
Compositio-prijs uitgereikt aan Bhargav Bhatt  
*Gerard van der Geer*
- 12 **Oratie | Inaugural Lecture**  
Enige beschouwingen over de waarde der wetenschapscommunicatie  
*Ionica Smeets*
- 20 **In Memoriam | Obituary**  
Ed de Moor: Didacticus met een energie om jaren op te zijn  
*Marjolein Kool, Willem Uittenbogaard*
- 23 **Column | Column**  
Slim tellen  
*Casper Albers*
- 25 **Onderzoek | Research**  
Cardioids and Morley's trisector theorem  
*Jan van de Craats, Jan Brinkhuis*
- 34 **Maatschappij | Society**  
Wiskunde helpt misdaad oplossen  
*Jan Peter van Zandwijk*
- 40 **Oratie | Inaugural Lecture**  
Hoe verbonden is je netwerk? *Hans L. Bodlaender*
- 49 **Onderzoek | Research**  
Phase transitions, Euclidean fields and self-similar random fractals  
*Federico Camia*
- 55 **In Memoriam | Obituary**  
Kobus Oosterhoff: Statistics as both a purely mathematical activity and an applied science  
*Piet Groeneboom, Jan van Mill, Aad van der Vaart*
- 60 **Onderzoek | Research**  
A glimpse into the asymptotics of polynomial identities  
*Geoffrey Janssens*
- 67 **Boekbespreking | Book Review**  
Hogere wiskunde omlaaghalen  
*Wim Caspers, Mark Timmer*
- 69 **Boekbespreking | Book Review**  
A global view of Poincaré's scientific work  
*Ferdinand Verhulst*

## Rubrieken | Departments

- 3 Redactioneel *Jan van Neerven*
- 4 Agenda *Nicolaos Starreveld*
- 6 Nieuws *Nicolaos Starreveld*
- 47 Het keerpunt van Rogier Swierstra  
*Jan Beuving*
- 65 In de verdediging *Geertje Hek*
- 72 Boekbesprekingen
- 78 Problemen

### Uitgelicht

40



In zijn oratie vertelt Hans Bodlaender over zijn onderzoek naar complexe netwerken.

49



Federico Camia geeft een inkijkje in zijn Vidi project 'Phase transitions, Euclidean fields and self-similar random fractals'.

## 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 eenieder 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

### Hoofredactie

Robbert Fokkink (r.j.fokkink@tudelft.nl)  
Jan van Neerven (j.m.a.m.vanneerven@tudelft.nl)

### Eindredactie/Bladmanager

Fred Drissen (fred@nieuwarchief.nl)

### Redactie

Hans Cuypers (TU/e), Jan Draisma (TU/e), Geertje Hek (UvA), Jan Hogendijk (UU), Teun Koetsier (VU), Sjoerd Rienstra (TU/e), Nicolaos Starreveld (UvA), Hans Sterk (TU/e), Mark Timmer (UT)

### Problemenrubriek

Gabriele Dalla Torre (UL), Christophe Debry (KU Leuven), Jinbi Jin (UL), Marco Streng (UL), Wouter Zomervrucht (UL)

### Bureauredactie

Lætitia Diemer, Claartje Ottens

### Illustraties

Ryu Tajiri

### Vormgeving

Susanne Laws (omslag, begeleiding binnenwerk)  
Kitty Molenaar (ontwerp)

### Druk

Veldhuis Media

### Advertenties

advertenties@nieuwarchief.nl

### Koninklijk Wiskundig Genootschap

www.wiskgenoot.nl

### Platform Wiskunde Nederland

www.platformwiskunde.nl

### European Mathematical Society

www.euro-math-soc.eu

### International Mathematical Union

www.mathunion.org

## Koninklijk Wiskundig Genootschap

Het Koninklijk Wiskundig Genootschap (KWG) is een landelijke vereniging van beoefenaars 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.

### Bestuur

Erik van den Ban (UU, voorzitter), Joke Blom (CWI, penningmeester), Sonja Cox (UvA, secretaris), Fetsje Bijma (VU), Theo van den Bogaart (HU), Geurt Jongbloed (TUD), André Ran (VU), Herman te Riele (CWI), Mark Veraar (TUD)

### Secretariaat

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

### Ledenadministratie KWG / Abonnementen

Postbus 1064, 7940 KB Meppel  
admin@wiskgenoot.nl, tel. 085-0160252

### Contributie

De contributie voor gewone leden bedraagt in 2016 € 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.

Instituten in Nederland en buitenland kunnen zich abonneren op het Nieuw Archief voor Wiskunde voor € 100 (Nederland), € 110 (Europa) of € 112,50 (buiten Europa).

### Members of foreign societies

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.

### Exchange subscriptions

Koninklijk Wiskundig Genootschap Library  
Centrum Wiskunde & Informatica  
P.O. Box 94079, NL-1090 GB Amsterdam  
KWG-exchange@cw.nl

## Informatie voor auteurs

### Kopij

Het Nieuw Archief voor Wiskunde is een geredigeerd tijdschrift. Kopij dient per e-mail te worden aangeboden aan de hoofdredactie. Uitgebreide informatie en auteursinstructies kunt u vinden op [www.nieuwarchief.nl](http://www.nieuwarchief.nl).

### Volgende nummers

| nummer | maand     | verschijnen | deadline kopij |
|--------|-----------|-------------|----------------|
| 18(2)  | juni      | 2017        | verstreken     |
| 18(3)  | september | 2017        | 01-05-2017     |
| 18(4)  | december  | 2017        | 01-08-2017     |
| 19(1)  | maart     | 2018        | 01-11-2017     |

## 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

Elke kaart in het kaartspel SET heeft vier eigenschappen die elk in drie varianten voorkomen. Drie kaarten vormen een SET als ze voor elke eigenschap alle drie gelijk zijn óf alle drie verschillend. Lees het artikel van Aart Blokhuis en Dion Gijsijt op pagina 8 over het Cap Set-probleem.

# Alternatieven

Met een goed gevoel voor timing koos de *Oxford Dictionary* het woord *post-truth* (“relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief”) tot Word of the Year 2016. Dat was nog voordat *alternative facts* hun intrede deden.

De actualiteit van de laatste maanden zal menig wetenschapper de wenkbrauwen hebben doen fronsen. Met het aantreden van de regering-Trump staat het primaat van de waarheid van de ene dag op de andere danig onder druk. In een aan het surrealistische grenzend schouwspel worden feiten ontkend, naar de hand gezet, of simpelweg afgedaan als ‘nep’. Intussen zijn klimaatwetenschappers aan overheidsinstituten in de Verenigde Staten koortsachtig hun datasets aan het veiligstellen op servers in het buitenland, vrezend dat die anders misschien gewist worden. Een March for Science staat op til.

Maar deze column is niet de plaats voor politieke beschouwingen. Laten we het dus over waarheid hebben. Wiskundigen hebben een bijzondere relatie met waarheid. *Post-truth* bestaat niet in de wiskunde. Eenmaal bewezen is een stelling voor altijd waar en één tegenvoorbeeld is genoeg om een vermoeden voor altijd naar de prullenbak te verwijzen. In theorie althans, want in de praktijk valt hier wel wat op af te dingen.

Zijn onze stellingen allemaal waar? Er zijn bewijzen die zo lang en complex zijn dat vrijwel niemand alle details overziet. Ons vertrouwen in de classificatie van eindige simpele groepen, de laatste stelling van Fermat, het Poincaré-vermoeden, wat meer is dat dan de constatering dat een kleine groep experts consensus heeft bereikt dat het bewijs klopt? Zullen we ooit zeker weten dat ze niets over het hoofd hebben gezien? En al jaren worstelt een hele gemeenschap met het bewijs van het *abc*-vermoeden van Mochizuki. Hier lijken de experts er voorlopig nog niet uit te zijn.

Iets dichterbij huis: heeft u de bewijzen uitgepluisd van alle stellingen die u toepast? Ook in de wiskunde is fact-checking niet onbelangrijk. Iedereen kent immers voorbeelden van gepubliceerde artikelen met incorrecte bewijzen. Wiskunde is mensenwerk, en simpelweg vertrouwen op het refereeproces is riskant. Er wordt steeds meer gepubliceerd en we hebben het steeds drukker, en dan krijgt de peer review al snel lagere prioriteit. Wie editor is van een tijdschrift kent ook het omgekeerde probleem: het is vaak knap lastig iemand bereid te vinden een artikel te refereëren.

Het systeem van peer review zoals we dat nu kennen is overigens nog niet zo oud. In 1936 ging het nog heel anders. Bekend is het verhaal van een boze Albert Einstein die onverwachts met een

refereerapport geconfronteerd werd, terwijl hij “had sent you our manuscript for publication and had not authorized you to show it to specialists before it is printed.” In haar interessante artikel ‘In referees we trust?’ laat Melinda Baldwin zien dat de peer review pas in de jaren zestig van de vorige eeuw gemeengoed werd. Daarvoor lag de taak van het beoordelen vaak bij de editor, al dan niet bijgestaan door een referee wiens rapport buiten het zicht van de auteur bleef.

Met de opkomst van arXiv lijken we weer een beetje naar dat oude model terug te keren. Steeds meer speelt de wiskundige actualiteit zich daar af. In sommige andere wetenschapsgebieden, zoals de deeltjesfysica, is het zelfs heel normaal dat de lijst met referenties van een artikel grotendeels uit arXiv-preprints bestaat. De officiële publicatie in een tijdschrift is dan enkel nog ‘for the record’. De rol van de peer review verschuift hiermee al naar de achtergrond. Een volgende stap zou kunnen zijn arXiv te voorzien van een MathOverflow-achtige functionaliteit. Die zou de peer review, en in feite het publiceren in dure tijdschriften als zodanig, gedeels overbodig kunnen maken.

Maar zo ver is het nog niet. Daarom tot slot de volgende vrolijke noot. Wiskundigen waren immers wiskundigen niet als zij niet ook leuke grappen maakten, en die zijn er momenteel volop. U heeft ze ongetwijfeld zien langskomen. Deze kreeg ik toegestuurd van Jacob Hoogenboom. Een *post-truth* rebuttal die handig van pas komt wanneer uw volgende artikel wordt afgewezen:

“Dear editor, You sent me four review reports. They are bad review reports. I have never seen such bad review reports. And I know this. Because I have seen a lot of good review reports. I write them. I write good review reports. I have also seen many good papers. It is easy. (...) They have my name on it. And my affiliation. A great affiliation. It is the best affiliation. And this paper is the best paper. It is the no. 1 paper. Do not listen to the reviewers. They are wrong. They are lying. They are the worst reviewers ever. They are probably Mexican. Or from the media. Yes, they are Mexican media!”

En mocht dat niet helpen, dan is er nog het *Journal of Alternative Facts*. Of is dat nepnieuws? Gelukkig bouwen wiskundigen liever bruggen dan muren. Laten we dat vooral blijven doen. En ze goed doorrekenen.

**Jan van Neerven**, hoofdredacteur  
Delft Institute of Applied Mathematics, TU Delft

# Agenda

## | Upcoming Events

### maart 2017

6–24 maart 2017

#### IMAGINARY, Wiskunde in beelden

Tijdens het schooljaar 2016–2017 trekt IMAGINARY door Nederland, een tentoonstelling georganiseerd door Platform Wiskunde Nederland (PWN) en de Nederlandse universiteiten.

**plaats** Leiden

**info** [imaginarymaths.nl](http://imaginarymaths.nl)

11 maart 2017

#### Transformatie

Een avond met kunst, wiskunde en wetenschap naar aanleiding van het 100-jarig bestaan van het boek *On Growth and Form* van D'Arcy Wentworth Thompson.

**plaats** De Balie, Amsterdam

**info** [meromorphpress.co.uk](http://meromorphpress.co.uk)

13 maart 2017

#### Studium Generale: Complex Networks

Een lezing over economische en financiële netwerken gegeven door Diego Garlaschelli.

**plaats** Lipsiusgebouw, Leiden

**info** [thenetworkcenter.nl](http://thenetworkcenter.nl)

13–17 maart 2017

#### Emerging Applications of Data Assimilation in the Geosciences

Een workshop over data-assimilatie mede georganiseerd door Svetlana Dubinkina (CWI), en Tristan van Leeuwen (UU).

**plaats** Lorentz Center@Oort, Leiden

**info** [lorentzcenter.nl](http://lorentzcenter.nl)

16 maart 2017

#### Onderwijsparade 2017

Met deze jaarlijkse 'optocht' van studenten en docenten wil de Onderwijsparade een substantiële bijdrage leveren aan onderwijsverbetering.

**plaats** Utrecht

**info** [uu.nl/agenda/onderwijsparade-2017](http://uu.nl/agenda/onderwijsparade-2017)

16 maart 2017

#### W4Kangoeroe

WereldWijde WiskundeWedstrijd voor basis- en middelbare scholieren.

**plaats** eigen scholen

**info** [w4kangoeroe.nl](http://w4kangoeroe.nl)

17 maart 2017

#### 2e ronde Nederlandse Wiskunde Olympiade

De tweede ronde van de Nederlandse Wiskunde Olympiade. De beste leerlingen gaan door naar de finale in september.

**plaats** Nederlandse universiteiten

**info** [wiskundeolympiade.nl](http://wiskundeolympiade.nl)

20 maart 2017

#### Studium Generale: Complex Networks

Een lezing over de verspreiding van epidemieën met titel 'Epidemics on Networks' gegeven door Frank den Hollander.

**plaats** Lipsiusgebouw, Leiden

**info** [thenetworkcenter.nl](http://thenetworkcenter.nl)

22 maart 2017

#### Grote Rekendag

Een dag voor alle basisscholen in Nederland en Vlaanderen (groep 1 t/m 8) die helemaal in het teken staat van rekenen. Het doel is kinderen onderzoekend te laten rekenen.

**plaats** basisscholen in Nederland en Vlaanderen

**info** [groterekendag.nl](http://groterekendag.nl)

27 maart 2017

#### Studium Generale: Complex Networks

Een lezing over het complexiteit van problemen in de analyse van netwerken gegeven door Michael Emerich.

**plaats** Lipsiusgebouw, Leiden

**info** [thenetworkcenter.nl](http://thenetworkcenter.nl)

29–31 maart 2017

#### Multi-Scale Features of Selection in Population Genetics

Een workshop georganiseerd door Frank den Hollander.

**plaats** Eurandom, Eindhoven

**info** [eurandom.nl](http://eurandom.nl)

### april 2017

6–12 april 2017

#### European Girls' Mathematical Olympiad

Europese wiskundewedstrijd speciaal voor meisjes

**plaats** Zürich, Zwitserland

**info** [egmo.org](http://egmo.org)

Suggesties voor deze agenda zijn welkom.

Redacteur: Nicolaos Starreveld

[agenda@nieuwarchief.nl](mailto:agenda@nieuwarchief.nl)

---

11–12 april 2017

**□ NMC 2017**

53ste Nederlands Mathematisch Congres, jaarlijkse bijeenkomst van wiskundig Nederland.

**plaats** Koningsbergergebouw, Uithof, Utrecht  
**info** [wiskgenoot.nl](http://wiskgenoot.nl)

---

## mei 2017

---

19 mei 2017

**□ ECENT/ELWier-conferentie**

Jaarlijkse conferentie voor lerarenopleiders van Expertisecentrum Lerarenopleidingen Natuurwetenschappen en Techniek en Expertisecentrum Lerarenopleidingen Wiskunde en Rekenen.

**plaats** Freudenthal Instituut, Utrecht  
**info** [uu.nl/agenda/ecentelwier-conferentie-2017-0](http://uu.nl/agenda/ecentelwier-conferentie-2017-0)

22 mei – 9 juni 2017

**□ IMAGINARY, Wiskunde in beelden**

Tijdens het schooljaar 2016–2017 trekt IMAGINARY door Nederland, een tentoonstelling georganiseerd door Platform Wiskunde Nederland (PWN) en de Nederlandse universiteiten.

**plaats** Groningen  
**info** [imaginarymaths.nl](http://imaginarymaths.nl)

29 mei – 2 juni

**□ Order Structures, Jordan Algebras, and Geometry**

Workshop mede georganiseerd door Onno van Gaans (UL).

**plaats** Lorentz Center@Snellius, Leiden  
**info** [lorentzcenter.nl](http://lorentzcenter.nl)

---

## juni 2017

---

7–9 juni 2017

**□ Complexe Netwerken**

Internationale conferentie georganiseerd door het wiskundeconsortium NETWORKS.

**plaats** CWI, Amsterdam  
**info** [thenetworkcenter.nl](http://thenetworkcenter.nl)

---

19 juni – 7 juli 2017

**□ IMAGINARY, Wiskunde in beelden**

Tijdens het schooljaar 2016–2017 trekt IMAGINARY door Nederland, een tentoonstelling georganiseerd door Platform Wiskunde Nederland (PWN) en de Nederlandse universiteiten.

**plaats** Nijmegen  
**info** [imaginarymaths.nl](http://imaginarymaths.nl)

26–30 juni 2017

**□ 10th Extreme Value Analysis Conference**

Een conferentie georganiseerd door J.J. Cai (TU Delft) en J. Einmahl (Universiteit Tilburg) over extreme value analysis en zijn toepassingen.

**plaats** TU Delft  
**info** [eva2017.nl](http://eva2017.nl)

---

29–30 juni 2017

**□ INFORMS Revenue Management and Pricing Section Conference**

Een jaarlijkse conferentie op het gebied van prijsanalyse en revenue management.

**plaats** CWI, Amsterdam  
**info** [cwi.nl](http://cwi.nl)

## juli 2017

---

3–8 juli 2017

**□ 32nd International Workshop on Statistical Modeling**

Een workshop over statistisch modelleren in verschillende vakgebieden.

**plaats** Rijksuniversiteit Groningen  
**info** [iws2017.webhosting.rug.nl](http://iws2017.webhosting.rug.nl)

---

## augustus 2017

---

25–26 augustus

**□ Vakantiecursus 2017**

Vakantiecursus georganiseerd door PWN, voor leraren, studenten van lerarenopleidingen en andere belangstellenden. Dit jaar is het thema ‘De computer in de wiskunde: breekijzer of oud schroot’.

**plaats** Technische Universiteit Eindhoven  
**info** [platformwiskunde.nl](http://platformwiskunde.nl)

---

28 augustus – 1 september 2017

**□ Bayesian and Nonlinear Inverse Problems**

Workshop mede gereorganiseerd door Johannes Schmidt-Hieber (LU) en Aad van der Vaart (LU).

**plaats** Lorentz Center@Oort, Leiden  
**info** [lorentzcenter.nl](http://lorentzcenter.nl)

---

## september 2017

---

1–2 september 2017

**□ Vakantiecursus 2017**

Vakantiecursus georganiseerd door PWN, voor leraren, studenten van lerarenopleidingen en andere belangstellenden. Dit jaar is het thema ‘De computer in de wiskunde: breekijzer of oud schroot’.

**plaats** CWI, Amsterdam  
**info** [platformwiskunde.nl](http://platformwiskunde.nl)

4–8, 11–15 en 18–22 september 2017

**□ Stochastic Activity Month**

Tijdens de Stochastic Activity Month (SAM) zullen er drie workshops zijn, beginnend met de Young European Probabilists workshop (YEP).

**plaats** Eurandom, Eindhoven  
**info** [eurandom.nl](http://eurandom.nl)

---

15 september 2017

**□ Finale Nederlandse Wiskunde Olympiade**

De prijswinnaars van deze finale worden uitgenodigd om deel te nemen aan een speciaal trainingsprogramma voor de internationale wiskundewedstrijden.

**plaats** Technische Universiteit Eindhoven  
**info** [wiskundeolympiade.nl](http://wiskundeolympiade.nl)

18–22 september 2017

**□ Applied Mathematics Techniques for Energy Markets in Transition**

Workshop mede georganiseerd door Cornelis Oosterlee (CWI).

**plaats** Lorentz Center@Oort, Leiden  
**info** [lorentzcenter.nl](http://lorentzcenter.nl)

---



---

19–22 september 2017

**□ Fifth International Conference on the History of Mathematics Education**

Internationale conferentie georganiseerd door de NVvW en de Werkgroep Geschiedenis van het Reken- WiskundeOnderwijs (WGRWO).

**plaats** Universiteit Utrecht  
**info** [ichme-5.nl](http://ichme-5.nl)

# 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, zo mogelijk met illustratie. De redactie behoudt zich het recht voor berichten te weigeren of in te korten.

Redacteur: Nicolaos Starreveld  
 nieuws@nieuwarchief.nl

## Koninklijke onderscheiding voor Frank den Hollander

Wiskundehoogleraar Frank den Hollander is op vrijdag 16 december 2016 onderscheiden als Ridder in de Orde van de Nederlandse Leeuw. De versierselen werden hem opgespeld door burgemeester Emile Jaensch van zijn woonplaats Oegstgeest. Dit gebeurde tijdens een congres aan de TU/e ter ere van Den Hollanders zestigste verjaardag. Frank den Hollander, hoogleraar kansrekening en statistische mechanica, is op dit moment verbonden aan de Universiteit Leiden.

*Monique van de Ven, TU/e*

## Koninklijke onderscheiding voor Martin Kindt

Wiskundige, onderwijsontwikkelaar en docent Martin Kindt is vrijdag 27 januari 2017 onderscheiden als Officier in de Orde van Oranje Nassau. Dit gebeurde tijdens een feestelijk symposium aan de Universiteit Utrecht ter ere van zijn tachtigste verjaardag en zijn inspanningen voor de vernieuwing van het wiskundeonderwijs gedurende ruim 45 jaar. De versierselen werden hem opgespeld door burgemeester Geert van Rumond van zijn woonplaats Wageningen.

*wiskundepersdienst.nl*

## Madelon de Kemp wint ASML Afstudeerprijs

Madelon de Kemp (Vrije Universiteit) heeft op 28 november de ASML Afstudeerprijs voor Wiskunde (5.000 euro) ontvangen bij de Koninklijke Hollandse Maatschappij der Wetenschappen (KHMW) te Haarlem. Ze onderzocht het optimaliseren van de afsprakenplanning van bijvoorbeeld medisch specialisten. Hiervoor combineerde ze zeer verschillende wiskundige technieken, met name uit

### Erratum

In het decembernummer is in het artikel "Vergeet 'zuivere en toegepaste' wiskunde" op p. 296 in de differentiaalvergelijking een minteken weggevalen. De vergelijking moet luiden:

$$-\varepsilon\phi'' + \phi = 1 + 2x^2.$$

*Ferdinand Verhulst*

### Rectificatie

In het decembernummer schreef ik dat Truida Wijthoff in 1907 erelid werd van het WG nadat zij de prijsvraag tien keer had gewonnen. Dat klopt niet, ik had de e-mail van Herman te Riele slecht gelezen. Truida Wijthoff werd lid van verdienste nadat ze vijf keer een eervolle vermelding had gekregen. De prijsvraag winnen kon niet. Een eervolle vermelding, middels het predicaat 'voldoende', was het hoogst haalbare. In 1923 kreeg Truida voor de tiende keer een eervolle vermelding.

Ook schreef ik dat Willem Wijthoff opgroeide aan de PC Hooftstraat. In werkelijkheid is hij daar pas later gaan wonen. Wim Pijls, die eerder schreef over Wijthoffs Nim in het *Nieuw Tijdschrift voor Wiskunde* 67 (1980), stuurde informatie die hij van Clark Kimberling had gekregen. De familie Wijthoff woonde op de Lauriergracht 111, net naast de suikerraffinaderij die in 1880 is afgebrand. Later dat jaar verongelukte Willems vader in Deventer en pas in 1881 verhuisde de familie naar de PC Hooftstraat. Eerst naar nummer 106 en vijf jaar later naar nummer 21.

*Robbert Fokkink*

de stochastische besliskunde en de combinatorische optimalisering. Bij het inplannen van afspraken voor artsen is de tijd die een behandeling gaat duren niet van tevoren bekend, maar onderhevig aan toeval. Doorgaans is door eerdere metingen wel bekend hoe lang een behandeling gemiddeld duurt en van hoeveel variatie sprake is. Het doel van Madelons onderzoek was om met deze beperkte informatie te bepalen hoeveel tijd per afspraak ingepland moest worden, zodat de patiënten niet lang hoeven te wachten, én de dokter niet te lang niets te doen heeft. *khw.nl*



### Wiskunde in het Horizon 2020 subsidieprogramma

In het voorjaar van 2016 heeft het directoraat-generaal DG CONNECT (Europese Commissie Directoraat-generaal Communicatienetwerken Inhoud en Technologie) een zogenaamde 'open consultation' gehouden over wiskunde. Wiskundigen in heel Europa konden zo hun zegje doen over het belang van wiskunde, en welke grote thema's er met wiskunde zouden moeten worden aangepakt, zoals bijvoorbeeld 'netwerken' en 'complexiteit'. Hierop is massaal gereageerd, en de eindrapportage hiervan is beschikbaar via [www.ec.europa.eu/futurium/en/content/mathematics-europe-report-open-consultation](http://www.ec.europa.eu/futurium/en/content/mathematics-europe-report-open-consultation). Als follow-up is er een sessie over wiskunde tijdens de 'ICT Proposer's Day' in Bratislava op dinsdag 27 september. Samen met de Europese organisatie EU-MATHS-IN ([www.eu-maths-in.eu](http://www.eu-maths-in.eu)) is een programma samengesteld waarin onderzoekers worden geïnformeerd over kansen voor wiskunde en wiskundigen binnen het Europese subsidieprogramma Horizon 2020. Naast de reguliere en bekende kanalen zoals ERC en Marie Skłodowska Curie (ITN, EID, EJD) zijn er ook vele mogelijkheden om enerzijds interessante en nieuwe wiskunde te ontwikkelen, anderzijds een belangrijke bijdrage te leveren aan het oplossen van industriële en maatschappelijke uitdagingen. *Wil Schilders, PWN*

### CWI ontwikkelt nieuwe rekenmethoden met ING Bank en UvA

Het Centrum Wiskunde & Informatica (CWI) heeft samen met de Universiteit van Amsterdam (UvA) nieuwe rekenmethoden ontwikkeld voor het inschatten van financiële risico's. Het project 'Advanced Estimation of Credit Valuation Adjustment', waarin de instellingen samenwerkten met ING, EY (Ernst & Young) en VORtech, wordt dit najaar afgerond. In het project werkten CWI en UvA onder

leiding van prof. Kees Oosterlee aan het ontwikkelen van rekenmethoden die een zo betrouwbaar mogelijke inschatting maken van de benodigde buffers om financiële risico's te dekken. *cwi.nl*

### Volledig simulatie van een menselijke slagader

Onderzoekers uit Nederland en Rusland hebben een computermodel van de menselijke slagader ontwikkeld. Hun model combineert een aantal kleinere submodellen van het cardiovasculair systeem. "We willen een methode ontwikkelen die wetenschappers kan helpen om nieuwe middelen te ontwikkelen zonder op dieren te hoeven experimenteren. En misschien ook klinisch onderzoek kan verminderen", aldus Alfons Hoekstra, hoogleraar aan het UvA en leider van dit project. *rsta.royalsocietypublishing.org*

### De keten zonder zwakste schakel

De blockchain-techniek werd bekend doordat hij een onderdeel is van bitcoin, de cybermunt waar veel ophef over was. Maar blockchain is veel breder toepasbaar. De ware blockchaineders koesteren visioenen hoe deze techniek alle banken en centrale autoriteiten overbodig gaat maken. Maar onder meer ABN AMRO gaat er nu zelf mee aan de slag, samen met computerwetenschappers van het Blockchain Lab aan de TU Delft. *nemokennislink.nl*

### Koninklijk Wiskundig Genootschap

#### Algemene Ledenvergadering

De Algemene Ledenvergadering wordt gehouden tijdens het NMC 2017, op dinsdag 11 april van 12.15 tot 13.15 uur. De vergaderstukken vindt u vanaf 14 maart op [wiskgenoot.nl/mijn-kwg](http://wiskgenoot.nl/mijn-kwg) (gebruiksnaam: KWG-leden; wachtwoord: K2003WG1778).

#### Brouwermedaille

Tijdens het NMC 2017 wordt de Brouwermedaille uitgereikt aan Kenneth Ribet, die dan ook de Brouwerlezing zal geven.

#### Pythagoras

Esther van Vroonhoven is toegetreden tot het managementteam van *Pythagoras*. Zij is tevens werkzaam bij Noordhoff Uitgevers.

#### Subsidie NMC 'nieuwe stijl'

NWO heeft een subsidie van 25.000 euro toegekend voor het organiseren van een pilot NMC 'nieuwe stijl' in 2018. Een commissie onder leiding van André Ran beraadt zich nu over de vormgeving van het NMC 'nieuwe stijl'.

Recent verschenen:

#### Indagationes Mathematicae ([www.elsevier.com/locate/indag](http://www.elsevier.com/locate/indag))

*Special Issue on Dynamics and Geometry*, Heinz Hanßmann, Igor Hoveijn, Sebastian van Strien, Holger Waalkens (eds.), Volume 27, Issue 5, 2016.  
*Special Issue on Automatic Sequences, Number Theory, and Aperiodic Order*, Alex Clark e.a. (eds.), Volume 28, Issue 1, 2017.

#### Epsilon Uitgaven ([www.epsilon-uitgaven.nl](http://www.epsilon-uitgaven.nl))

87. *Mathematics that Works – volume 4*, Maarten de Gee, € 37, 2017.  
Zebra 48. *Het Avontuur dat Algebra heet*, Martin Kindt en Henk Hietbrink, € 10, 2017.  
Zebra 49. *Een verkenning van krommen*, Jan Aarts en Swier Garst, € 10, 2017.

## Aart Blokhuis

Faculteit Wiskunde en Informatica  
Technische Universiteit Eindhoven  
a.blokhuis@tue.nl

## Dion Gijswijt

Delft Institute of Applied Mathematics  
TU Delft  
d.c.gijswijt@tudelft.nl

### De oplossing

# Het Cap Set-probleem

Mei vorig jaar zorgde de Nederlandse wiskundige Dion Gijswijt van de TU Delft samen met Jordan Ellenberg (University of Wisconsin) voor een doorbraak in het Cap Set-probleem. In dit artikel bespreken Aart Blokhuis en Dion Gijswijt het probleem en de gevonden oplossing aan de hand van het kaartspel SET.

Vijftien jaar geleden verscheen er in dit tijdschrift een artikel van N.G. de Bruijn [4] over het nog immer populaire kaartspel SET. Elke kaart in dit spel heeft vier eigenschappen (vorm, kleur, aantal en invulling) en elke eigenschap komt voor in drie varianten. Drie kaarten vormen een SET als ze voor elke eigenschap ofwel alle drie gelijk zijn, ofwel alle drie verschillend. Het doel van het spel is om binnen de kaarten die op tafel liggen zo snel mogelijk een SET te vinden. Wanneer er geen SET is, worden extra kaarten neergelegd.

Zoals in het artikel van De Bruijn wordt uitgelegd, kunnen de 81 kaarten worden geïdentificeerd met de punten van  $AG(4,3)$ , de vierdimensionale (affiene) ruimte over het lichaam  $\mathbb{F}_3 = \{0,1,2\}$  met drie elementen en met optelling en vermenigvuldiging modulo 3. Hierbij vormen drie kaarten een SET precies dan wanneer de bijbehorende punten op één lijn liggen (en dus een lijn vormen). Het grootste aantal kaarten zonder SET is dus precies de maximale kardinaliteit van een *cap set* of *cap* in  $AG(4,3)$ , een verzameling punten met hooguit twee

op een lijn. Op affiene transformaties na is er precies één grootste cap: de ‘Pellegrino cap’ afgebeeld in Figuur 1.

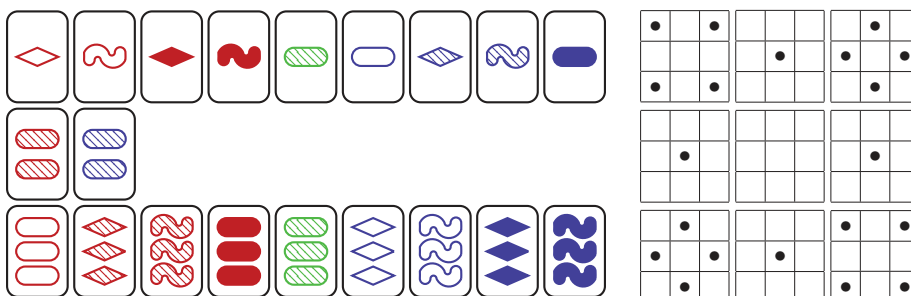
Voor een toegankelijke introductie tot het Cap Set-probleem en de relatie met SET verwijzen we de lezer graag naar het artikel van Davis en Maclagan [7].

### Het Cap Set-probleem

Een van de centrale problemen in de eindige meetkunde, de studie van affiene en projectieve ruimten over een eindige lichaam met  $q$  elementen, is het bepalen van

onder- en bovengrenzen op de kardinaliteit van zulke ‘caps’ of ‘cap sets’: een collectie punten met de eigenschap dat er geen drie op één lijn liggen. Voor een projectief vlak van de orde  $q$  is hier de bovengrens  $q+2$  als  $q$  even is, en  $q+1$  voor  $q$  oneven.

Een van de hoogtepunten in de geschiedenis van de eindige meetkunde is het resultaat uit 1954 van Beniamino Segre [14] dat voor oneven  $q$  een cap van de maximale grootte  $q+1$  in  $PG(2,q)$  (of  $AG(2,q)$ ), het projectieve (affiene) vlak van orde  $q$ , altijd een kegelsnede is (en dus voldoet aan een homogene tweedegraadsvergelijking). Voor even  $q$  is de classificatie van caps met het maximale aantal van  $q+2$  punten, zogenaamde hyperovalen nog een open probleem. Om een idee te krijgen



**Figuur 1** De twintig kaarten links zonder SET representeren een cap in  $AG(4,3)$  van maximale grootte. Puzzeltje: vind een correspondentie tussen de kaarten links en de vierkantjes met een punt rechts.

van de ontwikkelingen op dit gebied kan men bijvoorbeeld de hyperovaalpagina op de homepage van Bill Cherowitzo [5] raadplegen.

In dimensie 3 is de zaak ook nog redelijk duidelijk. Hier is de grens  $q^2 + 1$  voor de projectieve ruimte  $PG(3, q)$  en  $q^2$  voor de affiene ruimte  $AG(3, q)$  (voor  $q > 2$ ). Een cap van deze afmeting heet een ovoïde. Voor oneven  $q$  is dit noodzakelijk een elliptische kwadriek (in het affiene geval met een raakvlak op oneindig), voor even  $q$  is er één andere familie bekend. Het vermoeden is, dat dat alles is. Vanaf dimensie 4 is alles onduidelijk. Het laatste grote overzicht over deze en soortgelijke problemen is te vinden op de homepage van Leo Storme [12] en dateert van 2001. Een meer recent (2012), beperkter, maar voor ons verhaal belangrijker overzicht is te vinden in [2] en op de pagina van Yves Edel [9].

Voor  $q = 2$  is het bovenstaande probleem niet interessant, hier geldt de triviale bovengrens  $2^n$  in  $PG(n, 2)$  en  $AG(n, 2)$ , in het laatste geval door simpelweg alle punten te nemen. Ons Cap Set-probleem is het geval  $q = 3$  en meer speciaal  $AG(n, 3)$  of  $\mathbb{F}_3^n$ . We kunnen in dit geval het probleem als volgt herformuleren. Laat  $\mathbb{Z}/3\mathbb{Z}$  de groep zijn met drie elementen, dus  $\{0, 1, 2\}$  met optellen modulo 3 (ofwel de optelgroep van het lichaam  $\mathbb{F}_3$ ). Onze cap is nu een deelverzameling  $C \subset (\mathbb{Z}/3\mathbb{Z})^n$  met de eigenschap dat  $a + b + c = \underline{0}$  geen oplossingen heeft met  $a, b, c \in C$ , of beter gezegd, de enige oplossingen zijn die met  $a = b = c$ . Nog een manier om dit te formuleren is dat  $C$  geen rekenkundige rij ter lengte 3 bevat. In termen van de  $n$ -dimensionaal variant van SET bestuderen we verzamelingen kaarten zonder SET.

### Asymptotisch gedrag

Laat  $c_n$  het maximale aantal punten van een cap in  $AG(n, 3)$  zijn. De enige waarden die exact bekend zijn, zijn  $c_1 = 2$ ,  $c_2 = 4$ ,  $c_3 = 9$ ,  $c_4 = 20$ ,  $c_5 = 45$  en  $c_6 = 112$  (rij A090245 van de OEIS). Naast specifieke waarden zijn we vooral geïnteresseerd in het asymptotisch gedrag van de rij  $(c_n)$ . Wanneer  $C_1$  een cap is in  $AG(n, 3)$  en  $C_2$  een cap is in  $AG(m, 3)$ , dan is het product  $C_1 \times C_2 = \{(c_1, c_2) \mid c_1 \in C_1, c_2 \in C_2\}$  een cap in  $AG(n + m, 3)$ . Uit onze observatie volgt (via Fekete's lemma) dat

$$c = \lim_{n \rightarrow \infty} \sqrt[n]{c_n}$$

bestaat en dat  $c \geq \sqrt[n]{c_n}$  voor elke  $n$ . De cap uit Figuur 1 levert dus op dat  $c \geq \sqrt[4]{20} \geq 2,11$ . De beste ondergrens die bekend is komt van een constructie van een cap in  $AG(480, 3)$  door Edel [8] en geeft  $c \geq 2,21739$ .

Een overduidelijke bovengrens is  $c \leq 3$ , want  $AG(n, 3)$  heeft immers maar  $3^n$  punten. Brown en Buhler [3] gaven in 1982 de eerste niet-triviale bovengrens  $c_n = o(3^n)$ . Door gebruik te maken van ideeën uit de Fourieranalyse verbeterde Meshulam dit in 1995 tot  $c_n = O(3^n/n)$ . De beste bovengrens, tot vorig jaar, was een resultaat van Bateman en Katz [1]. Het bewijs is wederom gebaseerd op Fourieranalyse en geeft  $c_n = O(3^n/(n^{1+\varepsilon}))$  voor een zekere positieve (maar heel kleine)  $\varepsilon$ . Voor de constante  $c$  geven deze bovengrenzen echter geen verbetering. Het volgende probleem (geformuleerd door Brown en Buhler) bleef dan ook open.

**Probleem 1** (Cap Set-probleem). Geldt er dat  $c < 3$ ?

### De doorbraak

Een doorbraak kwam in mei 2016 toen Ernie Croot, Seva Lev en Péter Pach [6] een verwant probleem oplosten. Zij bewezen dat elke deelverzameling van  $(\mathbb{Z}/4\mathbb{Z})^n$  zonder rekenkundige rij van lengte 3 niet meer dan  $\gamma^n$  elementen kan bevatten voor een zeker constante  $\gamma \approx 3,60$  kleiner dan 4. Binnen een paar weken wisten Jordan Ellenberg en de tweede auteur van dit stuk, onafhankelijk van elkaar, het bewijs van Croot, Lev en Pach aan te passen voor  $\mathbb{F}_q^n$ , met  $q$  een oneven priemmacht. Het geval  $q = 3$  correspondeert precies met het Cap Set-probleem, waarvoor ze de volgende expliciete grens vonden:

**Stelling 2.** Laat  $A \subseteq \mathbb{F}_3^n$  een cap set zijn. Dan geldt  $|A| \leq 3 \cdot 2,756^n$  (en dus  $c \leq 2,756$ ).

Het resulterende artikel [10] en het artikel van Croot, Lev en Pach zijn, zij aan zij, verschenen in hetzelfde nummer van *Annals of Mathematics*.

### Hypermatrices en slice-rank

Kort na deze oplossing van het Cap Set-probleem gaf Terence Tao in een blogpost [15] een elegante herformulering van het bewijs in termen van de 'slice-rank' van hypermatrices. Hieronder geven we een schets van de ideeën in het bewijs.

Een hypermatrix van orde  $k \geq 2$  is een functie

$$H : J_1 \times J_2 \times \dots \times J_k \rightarrow \mathbb{F}$$

voor zekere eindige verzamelingen  $J_1, \dots, J_k$  en een lichaam  $\mathbb{F}$ . In het geval  $k = 2$  is  $H$  een  $|J_1| \times |J_2|$ -matrix en in het algemeen kunnen we  $H$  zien als een  $k$ -dimensionale tabel waarbij de 'plakken' in de  $t$ -de richting gelabeld zijn met de elementen uit  $J_t$ .

We zeggen dat  $H$  simpel is als in een van de richtingen  $t \in \{1, \dots, k\}$  geldt dat de  $|J_t|$  plakken waaruit  $H$  bestaat lineaire veelvouden van elkaar zijn. In formules:

$$H(x_1, \dots, x_k) = f(x_t) \cdot G(x_1, \dots, x_{t-1}, x_{t+1}, \dots, x_k)$$

voor zekere functies  $f : J_t \rightarrow \mathbb{F}$  en  $G : J_1 \times \dots \times J_{t-1} \times J_{t+1} \times \dots \times J_k \rightarrow \mathbb{F}$ . In dit verhaal is de rang (maar eigenlijk 'slice-rank') van een hypermatrix  $H$  het kleinste getal  $r$  waarvoor  $H$  te schrijven is als de som van  $r$  simpele hypermatrices. Voor  $k = 2$  komt dit overeen met de gebruikelijke matrixrang. Veel eigenschappen van matrixrang gelden ook in het algemene geval. Analooq aan het geval van diagonaal matrices geldt het volgende belangrijke lemma (waarvan het bewijs een leuke oefening is).

**Lemma 3.** Laat  $H : J \times \dots \times J \rightarrow \mathbb{F}$  een hypermatrix van orde  $k$  zijn met  $H(x_1, \dots, x_k) \neq 0$  dan en slechts dan als  $x_1 = x_2 = \dots = x_k$ . Dan geldt  $\text{rang } H = |J|$ .

### Verboden deelstructuren

Zoals veel andere problemen uit de extreme combinatoriek kan het Cap Set-probleem worden geformuleerd als het bepalen van de maximale grootte van een verzameling waarin een gegeven structuur ontbreekt. Preciezer gezegd: laat  $V$  een eindige verzameling zijn en  $E \subseteq V^k$  een verzameling 'verboden'  $k$ -tallen. We zijn geïnteresseerd in de maximale kardinaliteit  $\alpha(V, E)$  van een deelverzameling  $S \subseteq V$  zonder verboden  $k$ -tal (dat wil zeggen waarvoor geldt:  $S^k \cap E = \emptyset$ ). In het Cap Set-probleem is  $V = \mathbb{F}_3^n$ , bestaat  $E$  uit alle drietallen collineaire punten en is  $c_n = \alpha(V, E)$ .

We zeggen dat een hypermatrix  $H : V^k \rightarrow \mathbb{F}$  past bij  $(V, E)$  als alle diagonaal-elementen  $H(a, \dots, a)$  ongelijk aan nul zijn en de elementen  $H(a_1, \dots, a_k)$  buiten de diagonaal alleen ongelijk nul kunnen zijn voor  $(a_1, \dots, a_k) \in E$ . De volgende stelling geeft een bovengrens voor  $\alpha(V, E)$  die analoog is aan de Haemers-grens [11].

**Stelling 4.** *Laat  $H$  een hypermatrix zijn die past bij  $(V, E)$ . Dan geldt  $\alpha(V, E) \leq \text{rang } H$ .*

Het bewijs volgt direct uit Lemma 3 en het feit dat de rang van  $H$  niet kleiner is dan die van zijn beperking tot  $S^k$  wanneer  $S \subseteq V$  een deelverzameling is zonder verboden  $k$ -tal.

#### Het bewijs

We schetsen nu het bewijs van Stelling 2. Laat  $H$  de volgende hypermatrix zijn:

$$H: \mathbb{F}_3^n \times \mathbb{F}_3^n \times \mathbb{F}_3^n \rightarrow \mathbb{F}_3, \\ H(a, b, c) := \begin{cases} 1 & \text{als } a + b + c = 0, \\ 0 & \text{anders.} \end{cases} \quad (1)$$

Wegens Stelling 4 is het nu voldoende om

een goede bovengrens te vinden voor de rang van  $H$ . Hiertoe schrijven we  $H$  als een polynoom in de  $3n$  variabelen  $x = (x_1, \dots, x_n)$ ,  $y = (y_1, \dots, y_n)$  en  $z = (z_1, \dots, z_n)$ :

$$H(x, y, z) = \prod_{i=1}^n [1 - (x_i + y_i + z_i)^2]. \quad (2)$$

Hier gebruiken we dat  $1^2 = 2^2 = 1$  in  $\mathbb{F}_3$  en dus  $H(x, y, z) = 0$  zodra  $x_i + y_i + z_i \neq 0$  voor zekere  $i$ .

Voor elk monoom

$$x^\alpha y^\beta z^\gamma = x_1^{\alpha_1} \dots x_n^{\alpha_n} \cdot y_1^{\beta_1} \dots y_n^{\beta_n} \cdot z_1^{\gamma_1} \dots z_n^{\gamma_n}$$

in de expansie van  $H$  geldt dat elke variabele exponent ten hoogste 2 heeft, en de totale graad ten hoogste  $2n$  is. Dit impliceert dat  $H$  geschreven kan worden als

$$H(x, y, z) = \sum_{\alpha} x^\alpha F_{\alpha}(y, z) \\ + \sum_{\beta} y^\beta G_{\beta}(x, z) \\ + \sum_{\gamma} z^\gamma H_{\gamma}(x, y) \quad (3)$$

voor zekere functies  $F_{\alpha}$ ,  $G_{\beta}$ ,  $H_{\gamma}$ , waarbij  $\alpha, \beta, \gamma$  in de som lopen over de verzameling  $T = \{t \in \{0, 1, 2\}^n \mid t_1 + \dots + t_n \leq \frac{2n}{3}\}$ . Aangezien elke term in (3) een simpele hypermatrix is, volgt dat  $H$  rang ten hoogste  $3|T|$  heeft. Met behulp van de ongelijkheid van Chernoff kan  $|T|$  worden afgeschat als  $|T| \leq 2,756^n$ . We vinden dus dat de maximale kardinaliteit van een cap set in dimensie  $n$  niet meer is dan  $3 \cdot 2,756^n$ .  $\square$

#### Referenties

- 1 M. Bateman en N. Katz, New bounds on cap sets, *Journal of the American Mathematical Society* 25(2) (2012), 585–613.
- 2 J. Bierbrauer en Y. Edel, Large caps in projective Galois spaces, in Jan de Beule en Leo Storme, eds., *Current Research Topics in Galois Geometry*, Nova Science Publishers (2012), 87–104.
- 3 T.C. Brown en J. P. Buhler, A density version of a geometric Ramsey theorem, *Journal of Combinatorial Theory, Series A* 32(1) (1982), 20–34.
- 4 N.G. de Bruijn, Set!, *Nieuw Archief voor Wiskunde* 5/3(4) (2002), 320–325.
- 5 B. Cherowitzo, Bill Cherowitzo's hyperoval page (2004), <http://math.ucdenver.edu/~wcherowi/research/hyperoval/hypero.html>.
- 6 E. Croot, V.F. Lev en P.P. Pach, Progression-free sets in  $\mathbb{Z}_4^n$  are exponentially small, *Annals of Mathematics* 185(1) (2017), 331–337.
- 7 B.L. Davis en D. MacLagan, The card game SET, *The Mathematical Intelligencer* 25(3) (2003), 33–40.
- 8 Y. Edel, Extensions of generalized product caps, *Designs, Codes and Cryptography* 31(1) (2004), 5–14.
- 9 Y. Edel, Caps (2010), [www.mathi.uni-heidelberg.de/~yves/Matritzen/CAPS/CAPMatIndex.html](http://www.mathi.uni-heidelberg.de/~yves/Matritzen/CAPS/CAPMatIndex.html).
- 10 J.S. Ellenberg en D. Gijswijt, On large subsets of  $\mathbb{F}_3^n$  with no three-term arithmetic progression, *Annals of Mathematics* 185(1) (2017), 339–343.
- 11 W. Haemers, An upper bound for the Shannon capacity of a graph, *Colloq. Math. Soc. Janos Bolyai*, 1978, pp. 267–272.
- 12 J.W.P. Hirschfeld en L. Storme, *The Packing Problem in Statistics, Coding Theory and Finite Projective Spaces: Update 2001*, Developments in Mathematics, Vol. 3, Eds. A. Blokhuis, J.W.P. Hirschfeld, D. Jungnickel en J.A. Thas, eds., *Finite Geometries, Proceedings of the Fourth Isle of Thorns Conference*, Kluwer, 2001, pp. 201–246.
- 13 R. Meshulam, On subsets of finite abelian groups with no 3-term arithmetic progressions, *Journal of Combinatorial Theory, Series A* 71(1) (1995), 168–172.
- 14 B. Segre, Sulle ovali nei piani lineari finiti, *Atti Accad. Naz. Lincei Rend.* 17(1-2) (1954).
- 15 T. Tao, A symmetric formulation of the Croot–Lev–Pach–Ellenberg–Gijswijt capset bound (2016), <http://terrytao.wordpress.com/2016/05/18/a>.

## Gerard van der Geer

voorzitter Stichting Compositio Mathematica  
Korteweg-de Vries Instituut, Universiteit van Amsterdam  
g.b.m.vandergeer@uva.nl

### Nieuws Stichting Compositio Mathematica

# Compositio-prijs uitgereikt aan Bhargav Bhatt

Op vrijdag 9 december 2016 is de Compositio-prijs uitgereikt aan Bhargav Bhatt voor zijn artikel 'Derived splinters in positive characteristic' dat verschenen is in *Compositio Mathematica* 148 (2012), 1757–1786. De Compositio-prijs is een prijs die eens in de drie jaar door de Stichting Compositio Mathematica toegekend wordt voor het beste artikel dat in een driejarige periode in *Compositio Mathematica* verschenen is. De prijs die nu is toegekend betrof de periode 2011–2013 en de jury onder leiding van Frits Beukers is uitgekomen op een lijst van drie genomineerde artikelen:

- Bhargav Bhatt, Derived splinters in positive characteristic, *Compositio Mathematica* 148(6) (2012), 1757–1786.
- D. Kotschick en S. Schreieder, The Hodge ring of Kaehler manifolds, *Compositio Mathematica* 149(4) (2013), 637–657.
- June Huh, The maximum likelihood degree of a very affine variety, *Compositio Mathematica* 149(8) (2013), 1245–1266.

Uiteindelijk is de prijs toegekend aan Bhargav Bhatt. De prijs is overhandigd tijdens een feestelijke middag met lezingen door alle genomineerde auteurs. Alle drie genomineerde artikelen kenmerken zich door een zeer hoog niveau en alle drie komen met een prachtig resultaat. In hun voordrachten op 9 december gaven de genomineerden een overzicht van hun recente werk. De eerste voordracht van June Huh uit Princeton behandelde een combinatorisch probleem dat voortborduurt op een stelling van De Bruijn en Erdős die zegt dat een verzameling  $E$  van punten in een projectief vlak tenminste  $\#E$  lijnen bepaalt tenzij alle punten bevat zijn in een lijn. June Huh bewijst samen met Botong Wang dat voor een deelverzameling  $E$  van een  $d$ -dimensionale vectorruimte  $V$  die  $V$  opspant er in de partieel geordende verzameling van lineaire deelruimten van  $V$  voortgebracht door  $E$  tenminste evenveel  $(d-k)$ -dimensionale deelruimten zijn als er  $k$ -dimensionale deelruimten zijn als  $k \leq d/2$ . Dit bewijst het zogenaamde 'top-heavy'-vermoeden van Dowling en Wilson uit 1974. Het verrassende van het bewijs van Huh en Wang is dat dit eenvoudig te formuleren resultaat zwaar cohomologisch geschut gebruikt — de decompositie-stelling voor  $\ell$ -adische intersectiecomplexen. In een uiterst levendige en aantrekkelijke voordracht wist June Huh dit bewijs te schetsen. In de tweede voordracht die in gedeelten werd gegeven door beide genomineerde auteurs, Dieter Kotschick uit München en Stefan Schreieder uit Bonn, stond een probleem van Hirzebruch uit 1954 centraal: welke Chernse getallen en welke Hodge-getallen van gladde algebraïsche variëteiten zijn topologische invarianten, die dus niet van de complexe structuur, maar alleen van de topologie afhangen. Deze vraag bleef zestig jaar onbeantwoord, maar is door



Bhargav Bhatt

Kotschick en Schreieder in hun genomineerde artikel op een elegante manier opgelost. Voor de oplossing van het probleem construeren ze de zogenaamde Hodge-ring, een analogon van de cobordisme-ring. De derde voordracht was de voordracht van prijswinnaar Bhargav Bhatt uit Ann Arbor. Hij is een leerling van Johan de Jong en behoort met Peter Scholze tot een nieuwe generatie van hemelbestormers die in zeer korte tijd een golf van nieuwe re-

sultaten over de cohomologie van algebraïsche variëteiten hebben tweegebracht. Bhargav Bhatt geeft extreem heldere en energieke voordrachten. In zijn prijswinnend artikel voert hij het begrip afgeleide splinter in: een ring (of schema) die in de afgeleide categorie afsplitst van de cohomologie van een propere overdekking. Het artikel van Bhatt wordt gemotiveerd door het zogenaamde directe sommand-vermoeden van Hochster uit de commutatieve algebra. Het zegt dat voor een reguliere ring  $R$  die een deelring is van een ring  $S$  met  $S$  eindig voortgebracht als  $R$ -moduul, dat  $R$  afsplitst als directe sommand. Hochster bewees het in geval  $R$  een lichaam bevat. Maar het vermoeden stond meer dan veertig jaar open in het gemengde karakteristiek geval. Bhatt's prijswinnend artikel leidde afgelopen zomer tot een bewijs door Yves André van Hochster zijn vermoeden en Bhatt heeft direct een eenvoudiger bewijs gegeven. In de voordracht schetste hij zijn bewijs van het feit dat reguliere ringen afgeleide splinters zijn, zoals vermoed door Johan de Jong. Het bewijs maakt gebruik van de onlangs in zwang gekomen perfectioide technieken zoals door Scholze en Bhatt ingevoerd en ontwikkeld.

Na de voordrachten werd de prijs aan Bhargav Bhatt overhandigd in de nabijliggende ruimte waar de reizende tentoonstelling 'Imaginary' te vinden was. Dat was toepasselijk want de prijs zelf bestaat uit een model van het Cayley-oppervlak, een oppervlak van graad drie met vier gewone dubbelpunten als singulariteiten en de tentoonstelling Imaginary toonde een model van dit oppervlak. Cayley vond dit bijzondere oppervlak, gegeven door  $\sigma_3 = 0$  met  $\sigma_3$  de derde elementair-symmetrische functie in de vier variabelen van de driedimensionale projectieve ruimte, in 1844. ☼

## Ionica Smeets

Wetenschapscommunicatie  
Universiteit Leiden  
i.smeets@biology.leidenuniv.nl

### Oratie

# Enige beschouwingen over de waarde

In 2010 promoveerde Ionica Smeets in de wiskunde aan de Universiteit Leiden. In die tijd schreef zij samen met Jeanine Daems het blog Wiskundemeisjes. Na haar promotie deed ze een post-doc bij Publiek Begrip van Wetenschap en werkte een paar jaar als zelfstandig wetenschapsjournalist. Smeets is columnist bij onder andere *de Volkskrant*, maakte TV-programma's als *Eureka* en de *Nationale Wetenschapsquiz* en schreef boeken als *Het exacte verhaal – Wetenschapscommunicatie voor bèta's*. Voor ons eigen blad verzorgde zij vijf jaar de rubriek 'Het keerpunt'. Uiteindelijk verkoos Smeets wetenschapscommunicatie boven wiskunde en aanvaarde zij in 2015 een benoeming tot hoogleraar wetenschapscommunicatie aan de Universiteit Leiden. In haar oratie, uitgesproken op 13 mei 2016, zet zij uiteen waarom goede wetenschapscommunicatie zo belangrijk is.

Vandaag begin ik met het drie-deuren-probleem. Dat is berucht onder wiskundigen en het werd bekend onder een groter publiek toen Marilyn Vos Savant, de vrouw met het hoogste IQ ter wereld, erover schreef in 1990 [Vos]. Ze beantwoordde destijds in haar rubriek voor *PARADE Magazine* de volgende lezersvraag.

Bij de finale van een televisiequiz mag de kandidaat kiezen uit drie deuren. Achter één deur staat de hoofdprijs: een gloednieuwe auto. Achter de twee andere deuren staan geiten. De kandidaat wijst één van de deuren aan. Dan opent de presentator, die precies weet waar de auto staat, één van de overgebleven deuren waar een geit achter staat. Nu zijn er nog twee deuren over: één met een geit en één met de felbegeerde auto. Als elke week stelt de presentator nu zijn vraag: wil de kandidaat misschien wisselen van deur? Hij mag nog de andere gesloten deur kiezen. Heeft het zin voor de kandidaat om te wisselen?

Vos Savant antwoordde dat je in deze situatie moet wisselen: De eerst gekozen deur heeft een kans van  $1/3$  om te winnen,

de andere een kans van  $2/3$ . Ze gaf ook een manier om dit te visualiseren. Stelt u zich eens voor dat er een miljoen deuren zijn en dat u deur nummer één kiest. vervolgens maakt de presentator, die weet waar de auto staat en die altijd zal vermijden, alle andere deuren open, behalve nummer 777.777. U zou snel naar die deur wisselen, of niet?

Vos Savant kreeg duizenden boze brieven van lezers die schreven dat haar antwoord niet klopte, waaronder een heleboel van wiskundigen en andere academici. Zoals deze:

"[...] U heeft het verpest! Ik zal het u uitleggen. Zodra een deur is onthuld als verliezer, verandert die informatie de kans van elk van de overgebleven keuzen naar  $1/2$ . Geen van die twee deuren heeft ook maar een enkele reden om meer waarschijnlijk te zijn. Als beroepswiskundige maak ik me veel zorgen over het gebrek aan wiskundige vaardigheden bij het algemeen publiek. Helpt u alstublieft door uw fout toe te geven en voortaan zorgvuldiger te zijn."

Van de academische lezers die een brief naar Vos Savant stuurden was 65 procent het oneens met de door haar gegeven oplossing, onder de rest van de lezers was dat zelfs 92 procent. Zelfs na herhaalde uitleg en experimenten die lieten zien dat wisselen de winstkans verdubbelt, geloofden mensen haar niet. Eén lezer schreef dat er blijkbaar zoiets bestaat als 'vrouwelijke logica'. Enfin, in april 2011 dacht ik dat het aardig zou zijn om voor *de Volkskrant* een column te schrijven over dit klassieke probleem. Ik was een beetje bezorgd dat het zo'n bekend voorbeeld was, dat veel lezers het al zouden kennen en na drie zinnen zouden mompelen 'wisselen'. Daarom gebruikte ik het drie-deuren-probleem om te illustreren hoe de menselijke intuïtie het faliekant mis kan hebben [Drie1].

In mijn column legde ik op twee manieren uit hoe je kunt zien dat wisselen loont. Op eenzelfde manier als Vos Savant met honderd deuren én met de redenering die mij overtuigde toen ik dit probleem voor het eerst hoorde tijdens een eerstejaarscollege kansrekening. Als de kandidaat in het begin een deur kiest, dan heeft hij  $1/3$  kans om de auto aan te wijzen. Dus bij niet-wisselen, is en blijft de winstkans  $1$  op  $3$ . Bij wél wisselen, wint de kandidaat in alle gevallen waar hij in eerste instantie de geit aanwees. De kans op zo'n situatie is  $2/3$ . Kortom: wisselen verdubbelt de kans om te winnen van  $1/3$  naar  $2/3$ .

*De Volkskrant* werd bedolven onder de lezerspost. Zelden kwam er zo'n lawine van brieven over een onderwerp op de wetenschapspagina's. En ze waren bijna net



Foto: Ype Driessen

Ionica Smeets

# der wetenschapscommunicatie



Illustratie: Kazuma Fukumoto

zo boos van toon als de brieven die Vos Savant kreeg. Een lezer vroeg waarom de krant niet iemand inhuurde die wiskunde had gestudeerd in plaats van die domme Ionica Smeets.

Drie dagen na de publicatie van mijn column plaatste de opinieredactie een boze lezersbrief [Driez]:

“De guitige sneer in de rubriek Wiskundemeisjes (Wetenschap, 16 april) in de richting van al die domme mensen die hun intuïtie verkeerd gebruiken, is mij in het verkeerde keelgat geschoten. Ionica Smeets heeft wel gelijk, maar haar argument deugt niet. Dat komt wel meer voor bij wiskundigen. Ze kunnen aan anderen soms niet goed uitleggen hoe ze tot hun resultaat zijn gekomen.”

Vervolgens betoogt deze briefschrijver dat het glashelder is dat elk van de twee overgebleven deuren een kans van  $2/3$  heeft om de auto te bevatten en dat wisselen dus niet uitmaakt. Blijkbaar was het hem, en de opinieredactie, ontgaan dat de totale kans dat de auto ergens zit daarmee op een wonderbaarlijke  $4/3$  komt.

Op deze reactie volgde weer een reeks nieuwe lezersbrieven die de dagen daarna op de opiniepagina's verschenen. Ook daarna kwamen er weer boze reacties binnen en de redactie besloot dat ze wel genoeg inkt aan dit wiskundige probleem hadden besteed en drukten de reacties niet meer af. In mijn mailbox stapelden de boze reacties zich op met vragen als: “Wil je volhouden dat de kans verdubbelt door

van deur te wisselen? Als jij ook beschikt over gezond verstand kan ik me dat niet voorstellen.”

In mijn volgende column kwam ik nog eens terug op het probleem en haalde alles uit de kast om het idee uit te leggen [Drie3]. Ik eindigde met de verwijzing naar een simulatie en de suggestie om het spel herhaaldelijk na te spelen. Dan zouden lezers zelf ervaren dat wisselen op de lange termijn vaker winst oplevert. Bij elke honderd spelen levert wisselen ongeveer 67 keer winst op en niet-wisselen pakweg 33.

Toen kreeg ik de mooiste lezersbrief van allemaal. Een man schreef dat hij het drie-deuren-probleem avondenlang met zijn vrouw aan de keukentafel had nagespeeld met bekertjes en muntjes. Hij zag dat wisselen inderdaad meer winst opleverde. Hij begreep er nog steeds niets van, maar hij gelóofde me nu wel. Toch weer een zieltje gewonnen voor de wetenschap.

Is het überhaupt erg als iemand niet begrijpt hoe het drie-deuren-probleem werkt? Welnee, al verklaart een gebrekkige intuïtie voor kansen wel waarom zoveel mensen zich een woekerpypothek laten aansmeren of meedoen aan tamelijk hopeloze loterijen. Het is ook geen ramp als het algemeen publiek niet weet wat priemgetallen zijn, of wat een Higgs-deeltje is. Maar er zijn andere gebieden waar gebrekkige wetenschapscommunicatie ernstige gevolgen heeft.

Mijn voorganger Jos van de Broek noemde in zijn oratie bijvoorbeeld therapietrouw als een groot probleem: slechts de helft van de patiënten neemt hun medicijnen in zoals voorgeschreven [Broek1]. Slechte communicatie is één van de oorzaken dat mensen hun pillen verkeerd gebruiken. In dit soort voorbeelden draait wetenschapscommunicatie niet om leuke weetjes, maar om noodzakelijke kennis en die moeten we delen. Of zoals de titel van Jos' oratie het samenvatte: 'Nice to know, Need to know, Sharing to know'.

### Titelverklaring

De titel van mijn oratie is 'Enige beschouwingen over de waarde der wetenschapscommunicatie'. Toen ik die titel aankondigde, mopperde een collega dat dit een voorbeeld is van slechte wetenschapscommunicatie. Nu is een oratie überhaupt een enorm ingewikkelde vorm van wetenschapscommunicatie. De doelgroep is zeer inhomogeen met je trotse ouders op de

eerste rij, je kritische collega's in de zijbanen en daartussen een bont gezelschap van allerlei soorten toehoorders. Tot wie richt je je? Vanuit wetenschapscommunicatieperspectief had ik beter de helft van u níet kunnen uitnodigen, dan had ik tenminste een heldere doelgroep om me tot te richten.

Mijn titel is dan ook allerm minst bedoeld om mensen naar deze oratie te lokken, als ik dat had gewild, dan had ik iets clickbait-achtigs gekozen als: 'Ik werd hoogleraar wetenschapscommunicatie en wat er gebeurde zal je verbazen'. Nee, mijn titel is een verwijzing naar de inaugurele rede van de onvolprezen wiskundige N.G. de Bruijn. In 1946 sprak hij in Delft onder de titel 'Eenige beschouwingen over de waarde der wiskunde' [Bruijn]. De Bruijn kondigt daarin aan dat hij zijn toeschouwers niet zal lastig vallen "met het trekken van reeksen logische conclusies uit van tevoren nauwkeurig geformuleerde onderstellingen". In plaats daarvan belicht De Bruijn verschillende kanten van de wiskunde en onderwerpt die aan een persoonlijke waardebeoordeling. Eén van de onderwerpen die hij bespreekt is de drang van wiskundigen om te zoeken naar mooie bewijzen:

"Het is niet elegant om op musschen te schieten met een kanon, of visch te vangen met handgranaten. Zoo wordt het ook niet elegant geacht om met partieel differentieëren los te trekken op een sommetje dat met doodgewone algebra gemakkelijk kan worden opgelost."

De Bruijn eindigt zijn rede met dat we wiskunde de ruggengraat van onze cultuur mogen noemen. Hij vertelt over een schooljongen die de ruggengraat omschrijft als "op den eenen kant zit het hoofd, op den anderen kant zitten we zelf". Zo kunnen we volgens De Bruijn de geestelijke en praktische waarde van wiskunde symboliseren.

Ook in de wetenschapscommunicatie is er een dubbele waarde: de praktijk van de uitvoering en de theorie van het onderzoek. Mijn column over het drie-deuren-probleem was een voorbeeld van wetenschapscommunicatie dóen. Maar nu sta ik voor u als hoogleraar en niet als wetenschapsjournalist. Onze groep Science Communication & Society krijgt van diverse collega's verzoeken om over hun onderzoek een persbericht te schrijven, een festival

te organiseren of een site te maken. Allemaal prachtige ideeën, maar de hoofdtaak van onze vakgroep is niet de praktische kant van het uitvoeren van wetenschapscommunicatie, maar de geestelijke kant van het onderzoeken.

Daarover gaan de beschouwingen die volgen: Wat leert de theorie van de wetenschapscommunicatie ons? Laten we analyseren wat er tot nu toe is gebeurd in deze oratie.

### Beginnen met een anekdote

Om bij het begin te beginnen: Deze oratie opende met een anekdote en dat was voor sommige wetenschappers misschien wat ongemakkelijk. Waar wilde ik naartoe? Waarom duurde dat verhaal zo lang? En wat zei één zo'n voorbeeld nu? Wetenschappers hebben immers liever harde feiten en daarop gebaseerde redeneringen. Maar onderzoek laat zien dat het algemeen publiek juist eerder overtuigd raakt door verhalen dan door de "reeksen logische conclusies uit van tevoren nauwkeurig geformuleerde onderstellingen" die N.G. de Bruijn vermeed in zijn oratie [Dahlstrom].

Dat werd vorig jaar weer eens pijnlijk duidelijk in een Brits onderzoek naar het beeld van antibiotica-resistentie [Wellcome]. Allereerst bleek het algemeen publiek veel



Illustratie: Kazuma Eekman

minder daarover te weten dan experts aannamen. Zo dachten alle ondervraagden dat mensen resistent worden, zodat antibiotica niet meer werken. Terwijl het werkelijke probleem is dat bacteriën resistent raken.

Het tweede deel van dit onderzoek bekeek hoe het probleem dan wél overge-

bracht kon worden. Wetenschappers hadden een aantal manieren bedacht om de ernst van de situatie te laten zien, met veel cijfers en feiten. Dat bleek weinig indruk te maken. Bij een uitspraak als “Antibiotica-resistentie kost nu wereldwijd jaarlijks 700.000 doden en dat kunnen er in de toekomst tien miljoen worden”, vroegen mensen zich af wat die getallen over hun leven zeiden en mopperden dat al die nullen een beetje belachelijk klonken.

Bij uitleg dat resistente bacteriën zoals MRSA niet meer te behandelen zijn, reageerde iemand met: “Dat is wel zorgwekkend, maar zolang ik geen persoonlijk verhaal hoor, betrek ik het niet op mezelf.” Eén goed gekozen verhaal heeft meer impact dan een reeks feiten. In die zin was ik dus lekker bezig door vandaag te beginnen met een persoonlijke anekdote.

### De journalisten snappen er niets van

Door naar de inhoud van die anekdote waarmee ik begon. Wat mij het meeste verbaasde aan de drie-deuren-geschiedenis, is dat de opinieredactie van *de Volkskrant* een brief afdruckte die beweerde dat er twee deuren zijn die elk  $\frac{2}{3}$  kans hebben om de hoofdprijs te bevatten. De totale kans om te winnen zou daarbij  $\frac{4}{3}$  zijn en dat kán helemaal niet. Dat winkansen optellen tot één is geen kwestie van opinie, dat is nu eenmaal hoe het is. Het is alsof een journalist schrijft dat Berlijn de hoofdstad van Duitsland is, waarop een woeste lezer schrijft dat dit natuurlijk Parijs moet zijn en de opinieredactie dit met droge ogen afdruckt, omdat het óók een mening is.

Wetenschappers klagen graag dat journalisten er niets van begrijpen. En soms is dat terecht. Bij dieetadviezen in Britse kranten blijkt bijvoorbeeld 70 procent nergens op gebaseerd [Cooper]. John Oliver had deze week een geweldige reeks voorbeelden van ‘bullshit masquerading as science’ [Oliver]. Bijvoorbeeld een studio vol uitgelaten presentatoren die hadden geleerd dat een paar glazen champagne per week drinken je beschermt tegen dementie. Alleen was de achterliggende studie gedaan op ratten, en is het resultaat helemaal niet te vertalen naar mensen.

Alleen blijkt het verstandig om hierbij als universiteit ook de hand in eigen boezem te steken, want onze eigen persberichten nemen het niet altijd even nauw met de feiten. De wetenschapscommunicatie-onderzoeksgroep uit Cardiff verzamelde

honderden persberichten die in 2011 uitgestuurd werden door de twintig Britse topuniversiteiten naar aanleiding van een peer-reviewed publicatie [Summer].

Bij elk persbericht zochten ze de oorspronkelijke wetenschappelijke publicatie én de erover verschenen nieuwsberichten in kranten en op nieuwssites. Ze turfden op verschillende soorten overdrijving. Eén



Illustratie: Kazuma Ekiyama

daarvan was het presenteren van een gevonden verband als oorzaak-en-gevolg. Het klassieke voorbeeld van deze fout is dat iemand ontdekt dat er een verband is tussen ijsverkoop en het aantal verdrinkingen: hoe meer ijs er verkocht wordt, hoe meer mensen er verdrinken. Je mag hieruit niet concluderen dat ijs die verdrinkingen veroorzaakt. Toch maakte één derde van de persberichten zo'n soort sprong. En het overdrijven van advies en dierenonderzoek rechtstreeks vertalen naar mensen ging nog vaker mis.

Er bleek een sterke relatie tussen een overdreven persbericht en onzorgvuldige berichtgeving in de media. Ging het persbericht de fout in met verbanden, dan deed 81 procent van de nieuwsberichten dat ook. Was het persbericht in dit opzicht correct, dan was 82 procent van de nieuwsberichten dat ook. Eenzelfde effect was zichtbaar bij de andere bekeken overdrijvingen. De grootste vertekening zat niet bij de journalisten, maar bij de persberichten van de universiteit.

Gelukkig zat er ook nog een beetje goed nieuws bij de studie: keurige persberichten zonder overdrijvingen werden door journalisten net zo vaak opgepakt als de opgeklapte berichten. Het lijkt dus helemaal niet nodig om te overdrijven om media-aandacht te krijgen.

De studie zelf legt natuurlijk ook alleen een verband en de onderzoekers benadrukken dat er meer nodig is om aan te to-

nen wat de oorzaak is. Misschien horen de persberichten die niet overdreven zijn bij studies met spectaculairdere resultaten en zijn alleen de berichten over suffe studies opgeklapt en komen ze daardoor evenveel in het nieuws.

Daarom loopt er nu een vervolgstudie die kijkt welk effect bepaalde formuleringen in persberichten hebben. Onderzoekers en wetenschapsvoorlichters maken hun persbericht zoals gebruikelijk, waarna het naar de onderzoeksgroep wetenschapscommunicatie gaat. Zij brengen in de persberichten veranderingen aan, waarbij voor elk bericht willekeurig wordt toegewezen of het een verandering krijgt en zo ja in welke categorie en op welke manier. Vervolgens gaan de persberichten de wijde wereld in en wordt bijgehouden welke formulering wat voor invloed heeft op de verschenen nieuwsberichten. Heel spannend en ik ben zeer benieuwd naar de resultaten van dit project.

### Wetenschappers lezen de krant

De derde beschouwing: het was opvallend bij het drie-deuren-probleem dat er zoveel brieven van wetenschappers kwamen, zowel bij Marilyn Vos Savant als bij mij. Overigens kreeg ik geen enkele brief van wiskundigen die mijn oplossing in twijfel trokken, al schreven een paar van hen dat je eigenlijk moet zeggen dat de procedure van tevoren vaststaat en dat de presentator als de kandidaat de auto aanwijst willekeurig kiest welke van de andere twee deuren hij opent. Waarvan akte.

Wat je uit al die brieven kunt concluderen is dat wetenschappers de krant lezen. Dat geeft wetenschappers een zelfzuchtige reden om aan wetenschapscommunicatie te doen: hun werk wordt opgemerkt door collega's. Frans de Waal vertelde laatst in een lezing dat hij tegenwoordig het meeste wetenschappelijke nieuws ontdekt via bevriende onderzoekers op Facebook [Waal].

Een klassieke studie toonde aan dat in de krant komen daadwerkelijk tot meer citaties leidt. Hierbij vergeleek men wetenschappelijke artikelen uit de *New England Journal of Medicine* die al dan niet de *New York Times* haalden [Philips]. Het resultaat: een studie die de krant haalde, werd in het eerste jaar gemiddeld zeventien keer aangehaald. Een studie die niet in de krant kwam bleef steken op pakweg tien citaties. Tien jaar na publicatie was het verschil nog steeds zichtbaar. De in de krant genoem-

de artikelen werden daarbij ook vaker in hoog-impacttijdschriften als *Nature* en *Science* geciteerd. Overigens ging deze studie over publicaties tussen 1971 en 1979, dus voordat alles voor iedereen op internet te vinden was. U vraagt zich nu misschien af of deze studie niet oorzaak en gevolg verwart. Is het niet domweg zo dat belangrijkere studies een grotere kans hebben om in de krant te komen? Zouden die niet ook zonder publiciteit meer citaties hebben gehaald dan andere, minder spectaculaire resultaten?

Bij dit soort vragen zou je het liefste een gerandomiseerd experiment opzetten: elke maand wijs je puur willekeurig een paar publicaties aan die in de krant komen en de rest houd je in de anonimiteit. Maar vind maar eens een redactie die aan zo'n experiment wil meewerken en dan mooie stukken schrijft over een onderzoek dat ze zelf nooit geselecteerd zouden hebben. Het slimme van de *New York Times*-studie is dat ze een zogenaamd natuurlijk experiment gebruikten.

Tijdens een drie maanden durende staking in 1978 werd de *New York Times* niet verspreid, maar maakte de redactie wél een krant (voor de archieven). In die tijd werd net als altijd besloten welke wetenschappelijke artikelen in aanmerking kwamen voor verslaggeving, de redactie

invalshoeken is als het door N.G. de Bruijn genoemde streven naar elegante wiskundige bewijzen.

### Jargon

Mijn vierde beschouwing gaat over jargon. Inmiddels zijn er hier een aantal woorden voorbij gekomen waarvan u misschien niet precies weet wat ze betekenen: zoals Higgs-deeltje of inhomogeen. Die woorden heb ik bewust gebruikt, omdat ik het met u over jargon wil hebben. Ik heb dus expres een paar onnodige vaktermen toegevoegd. Maar misschien was dat helemaal niet nodig en gebruikte ik onbewust allerlei woorden waarvan ik dacht dat ze bekend waren, terwijl u geen idee had wat ik bedoelde.

Het is namelijk ongelooflijk makkelijk om te overschatten wat bekend is bij anderen. Dat geldt niet alleen voor vaktermen, maar ook voor ideeën, zoals ik zag toen ik dacht dat iedereen het drie-deuren-probleem al lang kende.

De Organisatie voor Economische Samenwerking en Ontwikkeling onderscheidt vijf niveaus van taalbeheersing. Wereldwijd scoren volwassen gemiddeld op niveau twee, wat vergelijkbaar is met de taalbeheersing van een elfjarig schoolkind [Oecd]. Nederland doet het gemiddeld iets beter, maar ook hier zit ruim 40 procent van de bevolking op niveau twee of lager. Meer dan 20 procent van de bevolking is laaggeletterd. En daarbij zijn ook hooggeleerden een leek zodra het over termen buiten hun eigen vakgebied gaat.

Bij de voorbereiding van een artikel over wetenschapscommunicatie in de aardwetenschappen stuurde een co-auteur een televisiefragment waar volgens haar geen enkel jargon in zat [Hut]. In een korte reportage over bosbranden komen echter in rap tempo de volgende termen voorbij: water-erosie, wind-erosie, minerale bodem, humuslaag, vegetatie, arme bodem en suppressie.

Zelfs als kijkers een vaag idee hebben dat erosie het afslijten van land is, zullen ze wat tijd nodig hebben om te bedenken wat het verschil hier is tussen wind-erosie en water-erosie. Terwijl zij daarover nadenken is de spreker al snel vier zinnen verder.

Het fragment bevat ook een tweede soort jargon: arme grond. Dit is een woord dat in het dagelijks leven iets anders betekent dat binnen een specialisme. We weten wat arm is, maar wat betekent dit bij grond? De aardwetenschappers hebben

er ongetwijfeld een nauwkeurige definitie van, maar voor de leek is het raden wat het exact is.

Het meeste onderzoek naar jargon gebeurt in de medische hoek, omdat spraakverwarring daar de grootste gevolgen heeft. Zelfs in eenvoudige gesprekken blijkt een groot deel van de patiënten de dokter niet goed te begrijpen. Bij laaggeletterde diabetespatiënten gebruikten artsen bij 81 procent van de bezoeken onuitgelegde vaktermen [Castro]. Gemiddeld ging het om vier vaktermen per gesprek. Daarbij ging het vooral om de tweede vorm van jargon: de artsen vermeden ingewikkelde Latijnse termen, maar zeiden iets als "uw gewicht is stabiel". Als daarna aan de patiënt gevraagd werd wat de dokter hiermee bedoelde, dan hadden ze geen idee. Betekende stabiel dat het goed was? Of bedoelde hij dat ze moesten aankomen of misschien juist afvallen?

Een andere klassieke studie laat zien dat artsen en patiënten gangbare medische termen niet hetzelfde interpreteren [Boyle].

Hoe zou u diarree omschrijven?

- a: dunne ontlasting
- b: meer dan eenmaal per dag ontlasting
- c: veel ontlasting in een korte tijd
- d: moeite met ontlasting produceren
- e: winderigheid

De meerderheid van de patiënten koos optie c: veel ontlasting in korte tijd. De meeste dokters typeerden diarree echter als a: dunne ontlasting. Een nuanceverschil, maar het is belangrijk om precies te weten wat een patiënt bedoelt als hij een klacht noemt.

Het is dus logisch dat het meeste onderzoek in deze hoek uit de medische wereld komt, daar heeft miscommunicatie de meeste gevolgen. Dit is die *need to know* in plaats van *nice to know* waar Jos van den Broek over sprak. Maar ik zou ook graag onderzoeken hoe het zit met jargon bij andere vakken.

### Wiskundemeisjes

Terug naar de boze brieven. De eerste briefschrijver refereerde aan mijn column onder de naam wiskundemeisjes. Die naam komt van de blog die Jeanine Daems en ik in 2006 begonnen [Wm]. Een gezamenlijke vriend noemde ons al zo voor wij elkaar voor het eerst ontmoetten: wij studeerden destijds wiskunde en we wa-



Illustratie: Kazuma Eekman

schreef hun stukken, maar geen enkele lezer zag daar iets van. Bij de wetenschappelijke artikelen die in deze periode uitkwamen was er qua aantal citaties geen verschil zichtbaar tussen de groep die wel én niet door de *New York Times* was gekozen. De toegenomen citaties zaten dus niet in de aard van de artikelen, maar in de publiciteit die ze kregen.

Het aardigste van dit onderzoek vind ik dat ze zo slim gebruik hebben gemaakt van die staking. Het zoeken naar dit soort

ren allebei meisjes. Toen we zijn grappige verzinzel als naam voor onze blog kozen, hadden we geen idee hoeveel effect dat zou hebben. Eén van de factoren van ons succes is dat voor veel mensen de combinatie wiskunde en meisjes verrassend bleek.

Er is een test die laat zien hoe sterk dit soort onbewuste vooroordelen zijn: de impliciete-associatietest. Deze test doe je op een computer en je moet steeds met twee verschillende toetsen aangeven in welke categorie woorden vallen [Imp]. Je krijgt bijvoorbeeld achter elkaar woorden te zien die mannelijk of vrouwelijk zijn, zoals 'tante' of 'vader'. Bij elk nieuw woord moet je zo snel mogelijk op de juiste toets drukken. Daarna volgt een ronde met woorden die over alfa-of-bètawetenschap gaan zoals 'Engels' of 'natuurkunde'. Weer moet je zo snel mogelijk de woorden in de juiste categorie indelen. En dan komt de truc: daarna krijg je de categorieën gecombineerd en de ene toets is voor 'vrouwelijk en bèta' en de ander voor 'mannelijk en alfa'. Daarna volgt nog een ronde met de combinaties omgekeerd. (Voor de volledigheid: de volgorde van de combinaties wordt per gebruiker willekeurig gekozen, dus je kunt ook beginnen met 'mannelijk en bèta' en 'vrouwelijk en alfa'.)

Combinaties die tegen je ingebakken vooroordelen ingaan blijken moeilijker om snel in te delen. Inmiddels deden meer dan een half miljoen mensen uit 34 landen deze test. En 70 procent van hen blijkt bèta onbewust met mannelijk te associëren.

Pijnlijk om toe te geven: ik ben één van hen. Zelfs ik die zelf in de wiskunde promoveerde, heb onbewust toch het idee dat wiskunde toch meer iets voor mannen is. Er blijkt een verband te bestaan tussen de ingebakken nationale vooroordelen op dit gebied en de prestaties van meisjes bij de exacte vakken. Hoe sterker het in de cultuur zit dat exacte wetenschap voor mannen is, hoe slechter meisjes presteren op school bij de bètavakken [Nosek].

Een Franse studie illustreerde dat fenomeen prachtig. Honderden 11-jarigen moesten een ingewikkeld figuur bestuderen en daarna uit hun hoofd natekenen [Huguet]. De helft van de kinderen kreeg te horen dat dit een tekenopdracht was, bij de andere helft heette het een meetkundeopdracht. Bij de tekenopdracht scoorden de meisjes beter dan de jongens, bij de

meetkunde-opdracht konden de meisjes er ineens niets meer van en deden de jongens het beter. Meisjes gaven ook aan dat ze de meetkunde-opdracht moeilijker vonden dan de tekenopdracht. Terwijl het exact dezelfde opgave was. Klein detail: als meisjes en jongens de test in aparte groepen deden, verdampten de verschillen.



Illustratie: Kazuma Eekman

Wetenschapscommunicatie gaat ook over de vraag hoe de academische wereld eruitziet voor de buitenwereld. Zoals mijn lievelingsschrijver Kurt Vonnegut schreef: "We are what we pretend to be, so we must be careful about what we pretend to be." In het publieke imago is wetenschap nog steeds een mannenvak [Dudo]. In Amerikaanse series is bijvoorbeeld 42 procent van de personages een vrouw, maar als het om wetenschappers gaat, zakt dat percentage naar 30 procent. Alles wat ik tot nu toe over vrouwen heb gezegd geldt trouwens minstens zo sterk voor minderheden. Zo zijn minderheden in wetenschappelijke kinderprogramma's minder vaak de wetenschapper en als er al eens een gekleurde wetenschapper is, dan komt deze minder lang in beeld dan zijn witte collega's [Long].

Deze stereotypen zorgen ervoor dat het voor hele groepen kinderen minder vanzelfsprekend is om voor wetenschap te kiezen. Zelf ben ik nog elke dag dankbaar dat ik wiskunde heb gestudeerd. Het is een akelige gedachte dat er nog steeds te veel kinderen zijn die ook zeer gelukkig zouden worden van een wetenschappelijke studie, maar die in een omgeving zitten waar niemand ze vertelt dat dit een optie is en waar niemand hen aanmoedigt. Eén populair-wetenschappelijke lezing op school, één artikel dat je als kind leest, één toevallige ontmoeting met een wetenschapper: zulke kleine dingen kunnen het verschil maken in een mensenleven. Dat is mijn persoonlijke reden om wetenschapscommunicatie boven de wiskunde te kiezen, hoe prachtig ik mijn oude vak ook vind.

De overstap is wel even wennen. Mijn promotieonderzoek ging over ergodische eigenschappen van niet-reguliere kettingbreuken [Smeets]. De definities waren glashelder, de vragen duidelijk en de bewijzen spijkerhard. Dat is in de wetenschapscommunicatie anders. Veel begrippen zijn er helemaal niet eenduidig te definiëren. Vraag tien experts om een definitie van *outreach* en je zult tien antwoorden met verschillende invalshoeken en nuances krijgen [Illingworth]. Het vak wetenschapscommunicatie is vergeleken bij wiskunde piepjong. Waar je bij wiskunde struikelt over de resultaten die door de oude Grieken bewezen zijn, of als het wat modernere wiskunde is door Gauss in de negentiende eeuw, zijn de onderzoeken die ik vandaag noemde hooguit vijftig jaar oud. Veel ideeën zijn nieuw en nog niet helemaal uitgekristalliseerd. Er zijn nog grote lege plekken op de kaart en dat maakt het een opwindend vakgebied om in te werken.

Daarbij vind ik het zeer stimulerend om samen te werken met collega's uit andere vakgebieden. Toen ik aan dat artikel over wetenschapscommunicatie bij aardwetenschappen werkte, leerde ik van alles over bosbranden en brughooften meten. Bij zulke interdisciplinaire samenwerkingen ontdek je daarnaast dat dingen die in jouw vakgebied vanzelfsprekend zijn elders totaal anders gaan. En dat dat misschien wel beter is dan dat wat je gewend bent. Zo geven wiskundigen hun referenties domweg een nummer, terwijl anderen veel handiger een eerste auteur en jaartal noemen bij een verwijzing.

Die eerste auteur is ook zo iets: bij wiskunde zetten we auteurs domweg op alfabetische volgorde, maar bij andere vakken blijken er andere mores te gelden. Bij het aardwetenschap-paper ontstond er een discussie over de juiste volgorde. Wie had het initiatief genomen? Wie had het meeste werk gedaan? Wie had voor zijn tenure-track het meeste aan deze publicatie? Na wat vriendschappelijke onderhandelingen kwamen we uit op de ideale volgorde, die grappig genoeg precies de alfabetische bleek te zijn. De kans daarop is overigens één op 24.

### Imagoprobleem

Voor het laatste deel van mijn oratie wil ik terugkomen op het beginnen met een anekdote. Ik vertelde u dat ik daarmee

goed bezig was, omdat mensen van verhalen houden. We weten óók dat het publiek meer geboeid is door een spreker die zijn lezing begint met een anekdote, dan door iemand die in een zogenaamde 'ethos'-introdunctie vooral bezig is met het neerzetten van zijn eigen expertise [Andeweg]. Wat ik u niet vertelde is dat sprekers die met een anekdote beginnen minder hoog scoren op geloofwaardigheid.

Ai. Sowieso is wetenschapscommunicatie slecht voor je imago. Britse wetenschappers gaven in interviews aan dat popularisering gedaan wordt door degenen die niet goed genoeg zijn voor een academische carrière [Survey]. Sterker nog: hoe vaker een wetenschapper in de media komt, hoe slechter diens kwaliteit als onderzoeker moet zijn. Dit heet ook wel het *Carl Sagan-effect* [Martinez] naar de astronoom die met zijn televisieprogramma's en boeken hele generaties warm maakte voor de sterrenkunde. Hij was daarnaast ook een toponderzoeker, maar werd weggestemd toen hij in 1991 voorgedragen werd als lid van de U.S. National Academy of Sciences. Terwijl hij een betere publicatielijst had dan anderen die wél lid waren. Bizar genoeg kende diezelfde academie hem drie jaar later wel hun meest prestigieuze prijs toe: de Public Welfare Medal.

Die dubbelheid kom ik vaak tegen. Onderzoekers mopperen tegen me dat het niets voorstelt wat ik doe: een beetje met mijn hoofd op televisie komen en kletsen over wetenschap. Om daarna te vragen of ik kan regelen dat zij ook een keer bij Humberto Tan mogen aanschuiven.

Daarbij blijkt het vooroordeel dat wetenschapscommunicatie voor de sukkels is nergens op gebaseerd. Integendeel: wetenschappers die actief zijn op het gebied van popularisering, presteren ook beter op academisch gebied dan hun collega's die in de ivoren toren blijven. Ze maken langere werkweken, publiceren meer artikelen en hun werk wordt vaker geciteerd [Jensen, Bentley].

Er is een lange traditie van excellente wetenschappers die ook excellente wetenschapscommunicatie deden. Neem Simon Stevin die in 1585 zijn pleidooi voor het gebruik van het tientallig stelsel in het Nederlands publiceerde in plaats van in het in wetenschappelijke kringen gebruikelijke Latijn [Stevin]. Aan het begin van *De Thien-*de schrijft hij:

"Den Sterrekyckers, Landt meters, Tapijtmeters, Wijnmeters, Lichaammeters int ghemeene, Muntmeesters, ende allen Cooplieden, wenscht Simon Stevin Gheluck."

Hij wilde dat al die mensen zijn idee zouden gebruiken en Stevin gaf aan het eind van zijn boek ook allerlei praktische tips voor verschillende beroepsgroepen. Prachtig.

Of neem natuurkundige Michael Faraday. In 1825 startte hij een reeks populair-wetenschappelijke lezingen voor kinderen met veel experimenten: The Christmas Lectures [Xmas]. Sinds 1966 zendt de BBC deze reeks jaarlijks uit en inmiddels zijn horden kinderen dankzij deze lezingen enthousiast geraakt over wetenschap.

Natuurlijk zijn er ook talrijke fameuze wetenschappers die helemaal niets met de buitenwereld deden en lekker als een kluisenaar aan hun ideeën werkten. Dat is ook prima. Onderzoekers zeiden in een recent promotieonderzoek naar valorisatie dat het de gezamenlijke verantwoordelijkheid van een vakgebied is om naar buiten te treden



Illustratie: Kazuma Eekman

[Jong]. 'De gans met de gouden eieren' wil je daarbij sparen: de briljante collega aan het eind van de gang wil je niet primetime op televisie zetten. Dus dan moet een collega van die briljante, maar wereldvreemde gans naar *De wereld draait door*.

Maar het huidige systeem belooft onderzoekers die naar buiten treden niet. Bij beoordelingen telt een gepubliceerd peer reviewed artikel met twee citaties vele malen zwaarder mee dan een televisieoptreden voor honderdduizenden kijkers. Maar wat is nu waardevoller voor de wetenschap als geheel? Robbert Dijkgraaf stelde ooit

voor om bij popularisering de overgedragen kennis te meten als het product van de hoeveelheid informatie en de grootte van het bereikte publiek [Dijkgraaf]. Zo gemeten geeft een paar zinnen over een wetenschappelijk idee op nationale televisie meer kennisoverdracht dan een specialistische lezing voor een klein zaaltje vol experts.

We zijn er als geheel van de wetenschap verantwoordelijk voor dat we de maatschappij betrekken bij ons werk. Wetenschapscommunicatie is geen hobby die je als onderzoeker in je eigen tijd doet in ruil voor een boekenbon, het is een wezenlijk onderdeel van het academische werk. Het zou goed zijn als er naast de gebruikelijke *principal investigators* (hoofdonderzoekers) aan de universiteit ook ruimte kwam voor een aantal *principal educators* en *principal communicators* die uitblinken in onderwijs en communicatie.

### Als onderzoeksgebied

Onze vakgroep wil met haar onderzoek een stevige theoretische basis leggen voor die communicatie. Zoals u gezien heeft is dat onderzoek nogal divers. De gebruikte methoden komen veelal uit de sociale wetenschappen en geesteswetenschappen. In het buitenland zit de vakgroep wetenschapscommunicatie dan ook vaak bij sociologie, psychologie of mediastudies. En de voorbeelden die ik vandaag noemde zijn een kleine, persoonlijke greep uit dit brede landschap.

Mijn voorganger Jos van den Broek is bijvoorbeeld gespecialiseerd in visuele communicatie en dit jaar zullen bij hem twee promovendi een proefschrift verdedigen over beeldgebruik binnen de gezondheidszorg [Broek2]. Mara van Beusekom ontwikkelt pictogrammen in bijsluiters voor laaggeletterden en Ronella Grootens-Wiegers onderzoekt materiaal dat kinderen ondersteunt bij beslissingen over deelname aan wetenschappelijk onderzoek [Beusekom, Grootens].

Mijn collega Anne Land-Zandstra is als universitair docent gespecialiseerd in informeel leren, zij werkt daarbij met musea als Naturalis en Museon en doet ook onderzoek naar *citizen science*-projecten zoals De Grote GriepMeting en iSPEX [Land1, Land2]. Pedro Russo onderzoekt hoe je met internationale sterrenkundige projecten als Universe Awareness juist kwetsbare groepen kinderen kunt bereiken [Russo]. En dat is dan

alleen maar hier in Leiden, in de rest van Nederland doen collega's bij wetenschapscommunicatie onder andere onderzoek naar dialoog mét het publiek, wetenschapswinkels, het ontwerpen van communicatieprocessen en risicocommunicatie.

En wat ga ik de komende jaren doen? Dat weet ik nog niet precies. Ik ben minder dan een jaar geleden benoemd en tot maart was ik slechts één dag per week verbonden aan de universiteit. Inmiddels is

mijn aanstelling uitgebreid naar vier dagen per week, maar netto ben ik hier nog geen vier maanden. Dus ik zit nog in de fase van kennismaken, plannen smeden en projectaanvragen indienen. Mijn rondleiding van vandaag laat zien wat mijn interesse heeft: onderzoek naar berichtgeving over wetenschap in de massamedia, zoals het persberichtenonderzoek en de *New York Times*-studie. Het ontwikkelen van effectievere wetenschapscommunicatie in sa-

menspraak met de doelgroep, zoals bij het antibioticaonderzoek. Het bestuderen van jargon bij aardwetenschappen of misschien wel natuurkunde. Ook zou ik graag een *citizen science*-project opzetten rond mijn oude liefde wiskunde.

Ik hoop vooral nauw samen te werken met onderzoekers uit verschillende vakgebieden en onderzoek te doen waar wetenschappers iets aan hebben bij hun *outreach*. ☺

## Referenties

- [Andeweg] Bas Andeweg en Jaap De Jong, Start with a story: theory versus practice of the anecdote as speech introduction technique, in *The Map and the Landscape: Norms and Practices in Genre*, Peter Lang, Bern, 2006, pp. 169–196.
- [Bentley] Peter Bentley en Svein Kyvik, Academic staff and public communication: a survey of popular science publishing across 13 countries, *Public Understanding of Science* 20 (2010), doi:10.1177/0963662510384461.
- [Beusekom] Mara van Beusekom e.a., Patients' preferences for visuals: Differences in the preferred level of detail, type of background and type of frame of icons depicting organs between literate and low-literate people, *Patient Education and Counseling* 98(2) (2015), 226–233.
- [Boyle] Charles Murray Boyle, Difference between patients' and doctors' interpretation of some common medical terms, *British Medical Journal* 2.5704 (1970), 286.
- [Broek1] Jos M. van den Broek, Nice to know – Need to know – Sharing to know, *Biomedische wetenschapscommunicatie in transitie*, (2007), 14.
- [Broek2] Jos M. van den Broek, W. Koetsenruijter, J. de Jong en L. Smit, *Beeldtaal: perspectieven voor makers en gebruikers*, Boom onderwijs, 2010.
- [Bruijn] N.G. de Bruijn, Eenige beschouwingen over de waarde der wiskunde, oratie TU Delft, 1946.
- [Castro] Cesar M. Castro e.a., Babel babble: physicians' use of unclarified medical jargon with patients, *American Journal of Health Behavior* 31, Supplement 1 (2007), S85–S95.
- [Cheryan] Sapna Cheryan e.a., Do female and male role models who embody STEM stereotypes hinder women's anticipated success in STEM?, *Social Psychological and Personality Science* 2(6) (2011), 656–664.
- [Cooper] Benjamin E.J. Cooper e.a., The quality of the evidence for dietary advice given in UK national newspapers, *Public Understanding of Science* 21(6) (2012), 664–673.
- [Dahlstrom] Michael F. Dahlstrom, Using narratives and storytelling to communicate science with nonexpert audiences, *Proceedings of the National Academy of Sciences* 111, Supplement 4 (2014), 13614–13620.
- [Dijkgraaf] R. Dijkgraaf, De kip in New York, *NRC Handelsblad*, 11 november 2006, www.nrc.nl/nieuws/2006/11/11/de-kip-in-new-york-11226595-a825915.
- [Drie1] Ionica Smeets, Na 998 deuren kun je beter wisselen, om de auto te winnen, *de Volkskrant*, 16 april 2011, www.volkskrant.nl/archief/na-998-deuren-kun-je-beter-wisselen-om-de-auto-te-winnen~a1876170.
- [Drie2] Opinie en Debat, *de Volkskrant*, 19 april 2011.
- [Drie3] Ionica Smeets, Nogmaals die drie deuren, *de Volkskrant*, 30 april 2011, www.volkskrant.nl/archief/nogmaalsdie-drie-deuren~a2925866.
- [Dudo] Anthony Dudo e.a., Science on television in the 21st century: Recent trends in portrayals and their contributions to public attitudes toward science, *Communication Research* 38 (2011), doi:10.1177/0093650210384988.
- [Grootens] Petronella Grootens-Wiegers e.a., Comic strips help children understand medical research: Targeting the informed consent procedure to children's needs, *Patient Education and Counseling* 98(4) (2015), 518–524.
- [Handley] Ian M. Handley e.a., Quality of evidence revealing subtle gender biases in science is in the eye of the beholder, *Proceedings of the National Academy of Sciences* 112(43) (2015), 13201–13206.
- [Huguet] Pascal Huguet en Isabelle Regner, Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances, *Journal of Educational Psychology* 99(3) (2007), 545.
- [Hut] Rolf Hut e.a., Geoscience on television: a review of science communication literature in the context of geosciences, *Hydrology and Earth System Sciences* 20 (2016), 2507–2518.
- [Illingworth] Sam Illingworth e.a., What's in a Name? Exploring the Nomenclature of Science Communication in the UK, *F1000Research* 4 (2015).
- [Imp] Project Implicit, <http://implicit.harvard.edu>.
- [Jensen] Pablo Jensen e.a. Scientists who engage with society perform better academically, *Science and Public Policy* 35(7) (2008), 527–541.
- [Jong] Stefan de Jong, *Engaging Scientists: Organising Valorisation in the Netherlands*, Leiden University, 2015.
- [Land1] Anne M. Land-Zandstra e.a., Citizen science on a smartphone: Participants' motivations and learning, *Public Understanding of Science* 25 (2015), doi:10.1177/0963662515602406.
- [Land2] Anne M. Land-Zandstra e.a. Motivation and learning impact of Dutch flu-trackers, *Journal of Science Communication* 15 (2016).
- [Long] Marilee Long, Greg Boiarsky en Greg Thayer, Gender and racial counter-stereotypes in science education television: A content analysis, *Public Understanding of Science* 10(3) (2001), 255–269.
- [Martinez] Susana Martinez-Conde, Has Contemporary Academia Outgrown the Carl Sagan Effect? *The Journal of Neuroscience* 36(7) (2016), 2077–2082.
- [Nosek] Brian A. Nosek e.a., National differences in gender-science stereotypes predict national sex differences in science and math achievement, *Proceedings of the National Academy of Sciences* 106(26) (2009), 10593–10597.
- [OECD] Outlook, *OECD Skills: First Results from the Survey of Adult Skills*, 2013.
- [Oliver] Last Week Tonight with John Oliver: Scientific Studies, 8 mei 2016, [www.youtube.com/watch?v=oRnq1NpHdmw](http://www.youtube.com/watch?v=oRnq1NpHdmw).
- [Phillips] David P. Phillips e.a., Importance of the lay press in the transmission of medical knowledge to the scientific community, *New England Journal of Medicine* 325(16) (1991), 1180–1183.
- [Russo] Pedro Miguel Rodrigues dos Santos Russo, *Design, Implementation and Evaluation of Transnational Collaborative Programmes in Astronomy Education and Public Outreach*, PhD Thesis, Leiden Observatory, Faculty of Science, Leiden University, 2015.
- [Smeets] Ionica Smeets, *On Continued Fraction Algorithms*, PhD Thesis, Mathematical Institute, Faculty of Science, Leiden University, 2010.
- [Stevin] Simon Stevin, *De Thiende*, [www.dbnl.org/tekst/stevoo1thieo1\\_01](http://www.dbnl.org/tekst/stevoo1thieo1_01), 1585.
- [Sumner] Petroc Sumner e.a., The association between exaggeration in health related science news and academic press releases: retrospective observational study, *BMJ* 349 (2014), g7015.
- [Survey] Survey of factors affecting science communication by scientists and engineers, Royal Society (2006), <https://royalsociety.org/topics-policy/publications/2006/science-communication>.
- [Vos] [marilynvossavant.com/game-show-problem](http://marilynvossavant.com/game-show-problem).
- [Waal] 2016.wtcvakconferentie.nl/84276/wiki/141075/frans%20de%20waal.
- [Welcome] Exploring the consumer perspective on antimicrobial resistance (2015), <http://wellcomelibrary.org/item/b24978000>.
- [Wm] [wiskundemeisjes.nl](http://wiskundemeisjes.nl).
- [Xmas] [www.rigb.org/christmas-lectures](http://www.rigb.org/christmas-lectures).
- [Zahra] Zahra Hazari e.a., Factors that affect the physical science career interest of female students: Testing five common hypotheses, *Physical Review Special Topics-Physics Education Research* 9(2) (2013), 020115.

**Marjolein Kool***Pabo Instituut Theo Thijssen en Kenniscentrum Educatie  
Hogeschool Utrecht  
marjolein.kool@hu.nl***Willem Uittenbogaard***Freudenthal Instituut  
Universiteit Utrecht  
w.uittenbogaard@uu.nl***In Memoriam Ed de Moor (1933–2016)**

# Didacticus met een energie om jaloers op te zijn

Op 6 december 2016 overleed Ed de Moor, eminent reken-wiskundedidacticus op de leeftijd van 83 jaar. Twee mensen die nauw met hem samengewerkt hebben blikken terug op zijn leven en werk. Marjolein Kool schreef samen met Ed de Moor de boeken *Alledaags rekenen* en *Rekenen is leuker dan/als je denkt*. Haar herinnering verscheen eerder in de nieuwsbrief van de NVvW. De tweede herinnering is van Willem Uittenbogaard, vriend en collega van Ed de Moor vanaf 1971. In de boekbesprekingen verderop in dit nummer volgt een pleidooi van Rob van der Waall voor heruitgave van het boek *Vroeger* van Ed de Moor, een bundel met 40 columns uit het *Tijdschrift voor Reken- en Wiskundedidactiek* Willem Bartjens.

Ed de Moor, de man van geschiedenis van het reken-wiskundeonderwijs, is zelf geschiedenis geworden. Eigenlijk heeft hij tijdens zijn leven al geschiedenis geschreven door zijn onophoudelijke, grote en bevlogen inzet. Hij verdient een bladzijde in het geschiedenisboek van het reken-wiskundeonderwijs, een bladzijde met een gouden

rand. Wat heeft hij gewerkt en geknokt om zijn onderwijsidealen te verwezenlijken! Eerst als wiskundeleraar en schoolleider op het Barlaeus- en het Vondelgymnasium, en sinds 1971 als medewerker van Wiskobas op het IOWO, de voorganger van het Freudenthal Instituut. Hij was een wiskundige die zich inzette voor het basisonder-

wijs. Daarin leek hij op Freudenthal, wiens stemgeluid hij feilloos kon imiteren: “Meneer de Moor, je kunt alles altijd nog weer beter begrijpen, zelfs de meest eenvoudige dingen.” Dat werd de lijfspreuk van Ed, naast het onderwijsdoel dat hij voor alle leerlingen wilde bereiken. De 21st-century skills vond hij moderne fratsen, één doel is genoeg: Leer kinderen denkend rekenen en rekenend denken.

Ed was een man uit één stuk, consequent en dapper, tijdens jaarvergaderingen en studiedagen ging zijn vinger altijd weer omhoog: “Ja maar, wacht eens even... is dat nou wel zo?” Hij hield nooit zijn mond, legde vaak de vinger op de zere plek, maakte niet altijd vrienden, maar had ook geen vijanden, want hij speelde nooit op de man en wist zijn mening altijd sterk te onderbouwen. Met zijn grote gevoel voor humor wist hij uiteindelijk iedereen voor zich te winnen.

Hij had een energie om jaloers op te zijn. Hij was oprichter van Panama en *Panama-Post*, mede-oprichter en bestuurslid van de Nederlandse Vereniging voor Reken- en Wiskundeonderwijs, oprichter van de Historische Kring voor Reken- en Wiskunde-



Ed de Moor in gesprek met leerlingen op het Barlaeus Gymnasium

onderwijs. En dat allemaal naast zijn werk aan het Freudenthal Instituut en de Stichting Opleiding Leraren. Toen hij 66 jaar was promoveerde hij nog 'even' op een vuistdik historisch-didactisch proefschrift (694 pagina's) over het meetkundeonderwijs.

Op 73-jarige leeftijd ging hij eindelijk officieel met pensioen, maar hij noemde zich sinds die tijd 'reken-wiskundedidacticus-

in-onruste' en ging gewoon door op de barricade, met het schrijven van artikelen en boeken en met het geven van lezingen en workshops over rekenen, wiskunde, onderwijs en geschiedenis. Hij fietste roelloos door Amsterdam en trok elke dag zijn baantjes in het zwembad. Ed was mijn paranimf, mijn mede-auteur van twee boeken, maar vooral mijn vriend. En daar ben

ik trots op. Met Ed kon je lachen, bomen, kunst kijken, gedichten lezen, maar ook haring eten en over de Schalkwijkse fanfare praten.

Rust zacht, reken-wiskundedidacticus-in-onruste, hopelijk gaan we steeds beter begrijpen hoeveel jij voor het Nederlandse reken- en wiskundeonderwijs betekend hebt.

Marjolein Kool

Op 12 december 2016 hebben we met een indrukwekkende bijeenkomst in de Duif in Amsterdam afscheid genomen van Ed de Moor.

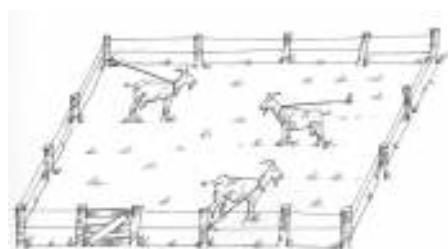
Ed begint zijn carrière in de zestiger jaren als wiskundeleraar aan zijn eigen school: het Barlaeus Gymnasium in Amsterdam. Hij vraagt zich op het Barlaeus al af waarom kinderen op de kleuterschool wel aan verkenning van de ruimte doen zoals hij zelf had gedaan met Fröbelmateriaal en waarom de lagere school geen aandacht besteedt aan eenvoudige meetkunde. "Wiskunde is toch een wetenschap van getal én ruimte?"

## IOWO

In 1970 wordt het IOWO (Instituut voor Ontwikkeling van het Wiskunde Onderwijs) opgericht met als taak een leerplan rekenen/wiskunde te maken voor vier- tot achttienjarige leerlingen (inclusief de lerarenopleidingen). Freudenthal wordt hoofdleraar-directeur.

Het instituut start met twee afdelingen: Wiskobas (wiskunde op de basisschool) voor vier- tot twaalfjarigen en Wiskivon (wiskunde in het voortgezet onderwijs) voor twaalf- tot achttienjarigen.

Er is een vacature bij Wiskobas en Ed solliciteert. Ed heeft geen ervaring, behalve zijn eigen ervaring met de lagere school en de kleuterschool. Men is van mening dat hij meer zou passen in de afdeling voortgezet onderwijs. Maar dat wil Ed niet. Hij



Welke geit is het beste af? (een lievelingssom van Ed uit het *Wiskobas-Bulletin* leerplandeel 7, 1977)

wordt bij Wiskobas benoemd. Er gaat — zoals hij zelf altijd zei — een wereld voor hem open: de didactiek van het rekenonderwijs met een zeer rijke traditie.

Er volgen tien hectische jaren. Met studie, ontwerp, ontwikkeling, het *Wiskobas-Bulletin*, conferenties, werkgroepen, nascholing van leerkrachten basisonderwijs en nascholing van pabo-docenten. Ook dan al werkt hij nadrukkelijk aan meetkundeonderwijs voor vier- tot twaalfjarigen.

Ed krijgt ook een belangrijke rol in de opzet en uitvoering van de nascholing van leerkrachten basisonderwijs. Talloze pedagogische academies doen daar aan mee.

## OW&OC

In 1980 is het afgelopen met het IOWO. Omdat er te weinig onderzoek in universitaire zin wordt gedaan vindt de rijksoverheid dat het IOWO niet aan de universiteit thuis hoort. Toetsontwikkeling gaat naar het Cito, ontwikkeling van het wiskundeonderwijs en het tijdschrift naar de SLO (Stichting Leerplan Ontwikkeling). Iedere medewerker krijgt de mogelijkheid naar de SLO over te stappen. Er blijft een kleine onderzoeksgroep over: het OWOC (Onderzoek Wiskundeonderwijs en Onderwijs Computercentrum) met vijf inhoudelijk medewerkers, drie personen in de ondersteuning en nog wel Freudenthal aan het hoofd. Een aantal medewerkers gaat over naar de SLO, de overige medewerkers van het voormalige IOWO zoeken en vinden een baan elders.

Ed kan ook naar de SLO in Enschede maar dat wil hij niet. Hij vindt een baan als lerarenopleider aan de SOL (Stichting Opleiding Leraren) in Utrecht (niet zo ver van z'n oude werkplek). Hij zoekt contact met een groep dolenden die eveneens het dak boven hun hoofd verloren hebben, zoals de responsgroep. Samen met alle Wiskobassers besluiten zij in 1982 op



aanstichting van Freudenthal een nieuwe vereniging op te richten: de NVORWO (Nederlandse Vereniging tot Ontwikkeling van Reken Wiskunde Onderwijs). Er komt een voorlopig bestuur: Ed wordt secretaris met een beetje vrijwillige ondersteuning.

## Panama

Wat Ed dwars zit was dat er voor bijna alle taken een nieuwe plek is gevonden maar niet voor de nascholing van leerkrachten basisonderwijs en pabodocenten. Ed schrijft de toenmalige Staatssecretaris van Onderwijs Hermes elke dag een brief met de vraag waar de nascholing blijft. Hij eindigt zijn brief steeds met: Morgen Schrijf Ik U Weer! Na enige tijd gloort nieuw licht aan de horizon. Het ministerie vraagt z'n ideeën op papier te zetten. Ed gaat aan de slag. Al gauw ligt daar het Panama-plan (Pedagogische Academie NAScholing Mathematische Activiteiten).

Ed is aanvankelijk nogal teleurgesteld in de voorlopige goedkeuring van zijn plannen. Hij verwacht meer. Ed krijgt een aanstelling om de plannen uit te voeren plus 100.000 gulden projectgeld. Het duurt niet lang of het gonst weer van de activiteiten.

De Panama-responsgroep komt maandelijks bij elkaar. De *Panama-Post* ziet het levenslicht: Jaargang 1, nummer 1 is een mededelingenblaadje van één kantje.

We denken al snel aan een voortzetting van de voorjaarsconferentie van het IOWO. En zo komt er met vereende krachten de eerste Panama-najaarsconferentie (reiskosten voor eigen rekening en 100 gulden eigen bijdrage: mooier kunnen we het niet maken).

Ed is baas over zijn eigen project en dat geeft hem de mogelijkheid om zijn bureau weer langzaam maar zeker terug te verhuizen van de SOL naar het oude IOWO-gebouw. Toch weer terug op het oude nest. Men vroeg mij eens op één van de Panama-conferenties waar het Panama-gebouw staat en hoeveel mensen daar werken. Ik heb het zo maar gelaten.

De samenwerking met Cito, SLO en natuurlijk ook met het OWOC komt goed op gang. Dat is ook een basis voor het houden van de jaarlijkse conferentie, de uitbouw van het tijdschrift en Eds werk aan de Proeve van een Nationaal Programma voor het reken-wiskundeonderwijs op de basisschool en de TAL-brochures die uiteindelijk hebben geleid tot de totstandkoming van eindtermen en kerndoelen voor het vak rekenen/wiskunde voor de basisschool. Die goede samenwerking is ook voor een groot deel te danken aan de persoon van Ed: je kunt grondig met hem van mening verschillen maar je blijft altijd (zeker op langere termijn) met hem in gesprek.

En die eindtermen en kerndoelen bevatten meetkunde. Wat hij zo graag wil. De meetkunde is in 1889 op een achternamiddag uit de wet geschrapt en precies honderd jaar later dreigt de meetkunde in 1989 niet in de eindtermen terecht te komen. Dat is goed gekomen. We hebben nu basisschoolmethoden met meetkunde. Ik zeg vaak: de mooiste methoden van de hele wereld. Ook met dank aan Ed.

Panama staat als een huis. De subsidie wordt structureel. Ed kan eindelijk een beetje terugtreden. Opvolgers komen bij Panama en Ed kan met pensioen.

### Eds klapstuk

Oud-Wiskobasmaatjes van wie de meesten inmiddels gepromoveerd waren, zeggen: "Waarom zou je niet ook promoveren?" En zo gaat Ed opnieuw aan de slag. Het is dan begin jaren negentig. Een jaar of



Ed de Moor en Willem Uittenbogaard kort voor een avondlezing najaar 2000 over meetkundeonderwijs tijdens een Panama-conferentie in de aula van de Universiteit Leiden.

zeven heeft hij er aan gewerkt. Met niet aflatende ijver.

Hij heeft een historisch-didactisch onderzoek gedaan naar het meetkundeonderwijs aan kinderen van vier tot veertien jaar. Eerst de inperking: wat wel en wat niet? Hij heeft z'n werk beperkt tot drie perioden.

Over de eerste periode geeft hij een beeld van de opkomst van de vormleer (1800–1889) als vak op de lagere school. Hij noemt en roemt Jan Versluys als degene die Pestalozzi's idee dat wiskunde meer is dan de leer van het getal, het meest heeft vormgegeven. Het blijkt toch lastig om los van de euclidische traditie informele meetkunde als vak te introduceren. In 1889 wordt het vak vormleer op de lagere school vervangen door het vak tekenen. Het zal dan bijna honderd jaar duren voor meetkunde als vak voor de lagere school weer ter hand wordt genomen.

De tweede periode (1890–1970) noemt hij: Van Fröbel tot Freudenthal, didactische denkers en doeners. Fröbels pedagogische opvattingen waren: 'leren door doen' en 'zelf ontdekken'.

Fröbels werk vindt ingang in West-Europa en in Nederland vooral door toedoen van de Leidse Kweekschool voor Fröbel-onderwijzeressen. Het programma was in Nederland wijd verbreid. Het meetkunde-programma van Fröbel heeft in Nederland geen invloed gehad op het rekenprogramma voor de lagere school. De didactische discussie tussen Dijksterhuis (logisch-deductieve stroming) en Ehrenfest (intuïtieve stroming) wordt voortgezet in de Wiskunde Werkgroep voor Vernieuwing van Opvoeding en Onderwijs. Dat wordt een constructief platform waarop de uiteindelijke veranderingen uit de volgende periode

stoelen. Freudenthal was sinds de oorlog voorzitter.

De derde periode (1970–1995) heet 'ontstaan van de realistische meetkunde'. Hieronder volgt een beschrijving van de periode, waarbij Ed zelf ook een grote rol speelt in de totstandkoming van onder andere drie concrete kerndoelen voor meetkunde voor de basisschool.

- De leerlingen beschikken over eenvoudige noties en begrippen waarmee zij ruimte kunnen ordenen en beschrijven.
- De leerlingen kunnen ruimtelijk redeneren. Zij bedienen zich daarbij van bouwsels, plattegronden, kaarten, foto's en gegevens over plaats, richting, afstand en schaal.
- De leerlingen kunnen schaduwbeelden verklaren, figuren samenstellen, bouwplaten van regelmatige objecten ontwerpen en identificeren.

Terwijl hij druk bezig is met zijn proefschrift richt Ed in 1995 ook nog de Historische Kring Reken en Wiskundeonderwijs (HKRW) op.

### Decorum

Ed hield altijd van decorum. In de vereniging (NVORWO) werd al gauw gesproken (vooral door Ed) over erelidmaatschap. We hadden ook een logo dat nog altijd in gebruik is (een opengewerkte kubus). Een goudsmid in Alkmaar kon daar wel speldjes van maken. En zo ontstonden de drie speldjes: een gewone, een zilveren en een gouden. Ook wisten we dat Ed een koninklijke onderscheiding erg op prijs zou stellen. Dat lukte aanvankelijk niet met iemand die verdiensten had op één gebied: het reken-wiskundeonderwijs. Maar door ook het Barlaeus Gymnasium in de aanvraag te betrekken konden we Ed voordragen.

Op zijn doodskleed droeg Ed het gouden speldje van de NVORWO en z'n koninklijke onderscheiding.

Willem Uittenbogaard

### Bronnen

- H.J. van Die, De betekenis van de kerndoelen van de vernieuwing van het reken-wiskundeonderwijs, *Panama-Post* 29(4) (2010), 13–22.
- Marjolein Kool en Ed de Moor, *Alledaags rekenen*, Bert Bakker, 2016, pp. 110–111.
- E.W.A. de Moor, *Van vormleer naar realistische meetkunde*, dissertatie, CD-beta Press, 1999.
- A. Treffers, *Weg van het cijferen*. Universiteit Utrecht, 2015.

## Casper Albers

Psychometrie & Statistiek  
Rijksuniversiteit Groningen  
c.j.albers@rug.nl



## Column Casper grijpt een kans

# Slim tellen

Caspers Albers schrijft op regelmatige basis in dit blad een column over alledaagse statistische onderwerpen.

Rekenen is de basis van de wiskunde. En tellen is de basis van rekenen. Daarnaast is tellen ook gewoon leuk (zoals Sir Francis Galton zei: “Whenever you can, count” [4]). Je zou dus verwachten dat we ondertussen dat ook gewoon kunnen. Maar tellen is soms nog best moeilijk. Bijvoorbeeld wanneer je niet (goed) kan zien wat je moet tellen.

Een klassiek voorbeeld van veertig jaar geleden is het *proofreading problem*. Aan twee reviewers wordt gevraagd om, onafhankelijk van elkaar, een tekst te lezen en de (spel)fouten te corrigeren. Geen van beide is perfect en ze zien mogelijk fouten over het hoofd. Uiteindelijk worden er  $a$  fouten gevonden door de ene reviewer en  $b$  door de andere. Er worden  $c$  fouten gezamenlijk gevonden. Uiteindelijk worden dus  $a + b - c$  fouten gevonden. De vraag is of nu alle fouten gevonden zijn. Via een elegante berekening komt Pólya [5] op

$$\frac{(a-c)(b-c)}{c}$$

als puntschatter voor het aantal ongevonden fouten. De puntschatter is aan te vullen met standaardfouten via de delta-methode, of zelfs om te zetten in een verdelingsschatter [1]. Het aantal fouten dat niet gezien is, valt dus te tellen.

Een nog ouder voorbeeld is de volgende ouderwetse kermis-attractie. Een grote pot is gevuld met knikkers en tegen betaling van een stuiver mag je raden hoeveel knikkers er in de pot zitten. Aan het eind van de dag wint degene die het dichtste bij het juiste aantal zit een rollade (of wat vroeger ook maar de prijzen op de kermis waren). Wat bij dit soort problemen vaak blijkt, is dat de mediaan van alle pogingen verrassend dicht bij de echte waarde zit. Galton beschreef dit probleem al ruim een eeuw geleden in zijn stuk ‘Vox Populi’ [3]. Hier ging het om het raden van het gewicht van een os op de markt van Plymouth. De

mediaan van de schattingen van  $N = 787$  deelnemers bleek op 1 procent nauwkeurig te zijn. (Galton pleitte voor de mediaan in plaats van het gemiddelde, omdat de schattingen niet symmetrisch verdeeld waren.)

In zijn resultaat zag Galton bewijs van de kracht van democratische besluiten door een groep mensen: afwijkende meningen strepen tegen elkaar weg en gemiddeld gesproken weet men wel hoe het moet. Dit concept is later *the wisdom of the crowd* gaan heten. In essentie zegt dit concept eigenlijk weinig meer dan dat het steekproefgemiddelde doorgaans een betere schatting voor de (onbekende) populatiewaarde is dan individuele schattingen. Het werkt echter niet altijd en men is er de afgelopen jaren achter gekomen dat het vaker dan verwacht niet werkt. De belangrijkste eis is dat men — gemiddeld genomen — wel de juiste waarde schat. De schattingen moeten *zuiver* zijn (of, op zijn minst statistisch consistent): als dat zo is dan geldt, onder milde voorwaarden, dat de *mean squared error* van het steekproefgemiddelde een factor  $N$  kleiner is dan die van de individuele meting. Bij een pot met knikkers of het gewicht van een os gaat die zuiverheid nog wel op, maar bij complexere situaties niet.

In het dagelijks leven maken veel mensen gebruik van dezelfde informatiebronnen — kranten, tv, et cetera — en lopen daarmee het risico dezelfde vertekende informatie te ontvangen, die een verstoring in het wereldbeeld kan opleveren. Onlangs is een internationale peiling uitgevoerd [2] met onder meer vragen als “Vindt u abortus moreel verwerpelijk?” en “Van welk percentage landgenoten verwacht u dat ze abortus moreel verwerpelijk vinden?” In Nederland gaf 8 procent van de deelnemers aan abortus verwerpelijk te vinden, maar het gemiddelde van de schattingen voor de tweede vraag was 37 procent, ruim vier keer zo veel. Ook schattingen van het percentage inwoners dat moslim is (schatting: 18 procent, werkelijkheid: 6 procent) en of homoseksualiteit verwerpelijk is (5 procent gaf aan dat te vinden, de inschatting was dat 34 procent van de landgenoten het vinden) lagen ver van de juiste waarde: de wisdom van de crowd is hier ver te zoeken.



Systematische verstoringen komen niet alleen voor bij het volk (de *vox populi*), maar ook bij de elite (de professionals). Zo hadden bijna alle peilingbureau's de uitslag van de Amerikaanse presidentsverkiezingen fout ondanks dat de peilmethoden opgezet waren door slimme en ervaren politicologen en methodologen. Een van de grootste oorzaken hiervoor was dat men in alle staten dezelfde systematische fout maakte (en daarmee de grootte van het Trump-kamp onderschatte). De grootte van die onderschatting zat nog wel in de foutenmarge per staat, maar bij het samenstellen

van de resultaten voor het gehele land niet meer. Hiervoor werd het (gewogen) gemiddelde genomen van de individuele staten en werd ervan uitgegaan dat de meetfouten ongeveer een factor  $\sqrt{50}$  kleiner zouden worden. Die vlieger gaat echter alleen op wanneer de meetfouten onafhankelijk zijn (in de ene staat overschat je de Trump-aanhang, in de andere onderschat je deze, en uiteindelijk streept alles grotendeels tegen elkaar weg). Die onafhankelijkheid ging hier niet op: in elke staat lagen dezelfde redenen achter de onderschatting.

Een recenter voorbeeld uit Amerika over telproblemen is de discussie rond de opkomst bij de inauguratie van Donald Trump op 20 januari. Op social media verschenen foto's die de opkomst bij Obama en Trump vergeleken: bij Obama leek het overvol te staan, bij Trump waren grote stukken leeg. Bij die foto's zijn wat vraagtekens te zetten (zijn ze bijvoorbeeld wel op hetzelfde tijdstip genomen; waren de weersomstandigheden bij Obama beter, et cetera), maar ook de experts waren duidelijk: de opkomst acht jaar geleden was (flink) groter. Watson en Yip [6] leggen verschillende methoden uit om goede schattingen te krijgen van de opkomst en hoe je bij deze methoden de standaardfout kan schatten. Met informatie over de oppervlakte ( $\hat{O}$ ; aantal  $m^2$ ) alsmede de gemiddelde dichtheid ( $\hat{D}$ ; aantal personen per  $m^2$ ) in de vorm van schattingen aangevuld met standaardfouten (SE), kan je de totale opkomst  $\hat{N} = \hat{O} \times \hat{D}$  alsmede de standaardfout van de opkomst

$$\frac{SE(\hat{N})}{\hat{N}} = \sqrt{\frac{SE^2(\hat{O})}{\hat{O}^2} + \frac{SE^2(\hat{D})}{\hat{D}^2}}$$

schatten. Deze ligt doorgaans op circa 10 procent van  $\hat{N}$ . Ook via andere methoden van schatten van de opkomst kom je op soortgelijke standaardfouten.

De technologie en wiskunde is er dus om redelijk nauwkeurige schattingen te geven van de grootte van een menigte. Helaas zitten koppige politici de werkelijkheid in de weg: de voorlichter van het Witte Huis stond er op dat de opkomst bij Trump toch echt groter was dan die bij Obama — iets dat door Trumps hoogste assistent een 'alternatief feit' genoemd werd. Trump heeft geen alleenrecht op het glashard ontkennen van de waarheid. In 2011 omschreven Watson en Yip al "However, it seems there is too much politics in the mix for crowd estimation to be made precise in the near future. The public has a view of the truth that is coloured by their beliefs. This applies particularly to crowd estimation." ☿

## Referenties

- 1 C.J. Albers, G.Th. de Roos en W. Schaafsma, Estimating a frequency unseen: an example from ornithology, *Statistica Neerlandica* 59(3) (2005), 397–413.
- 2 P. Duncan, Europeans greatly overestimate Muslim population, poll shows, *The Guardian*, 13 december 2016.
- 3 F. Galton, Vox Populi, *Nature* 1949(75) (1907), 450–451.
- 4 F. Galton, in J.R. Newman, ed., *The World of Mathematics – Volume 2*, Simon and Schuster, 1956, p. 1169.
- 5 G. Pólya. Probabilities in proofreading, *The American Mathematical Monthly* 83(1) (1976), 42.
- 6 R. Watson en P. Yip, How many were there when it mattered? *Significance*, 2011.

Jan van de Craats

Korteweg-de Vries Institute for Mathematics  
University of Amsterdam  
j.vandecraats@uva.nl

Jan Brinkhuis

Econometric Institute, Erasmus School of Economics  
Erasmus University, Rotterdam  
brinkhuis@ese.eur.nl

## Research

# Cardioids and Morley's trisector theorem

At the end of the nineteenth century the algebraic geometer Frank Morley discovered a nice little theorem on the trisectors of a triangle, known as 'Morley's trisector theorem'. Since the beginning of the twentieth century, many proofs of this theorem have been published. In this article Jan van de Craats and Jan Brinkhuis elucidate the role of cardioids in Morley's discovery of his theorem.

The algebraic geometer Frank Morley (1860–1937) came from England to the United States in 1887. He was the editor of the *American Journal of Mathematics* from 1900 to 1921 and served as President of the American Mathematical Society from 1919 to 1920. In 1899, while studying properties of general configurations of  $n$  lines in the Euclidean plane by means of complex numbers, Morley discovered a nice lit-

tle theorem on the trisectors of a triangle. He mentioned it to friends, who spread it over the world in the form of mathematical gossip (Coxeter [1, p.23]). The simple version of Morley's trisector theorem, as it later became known, reads as follows (see Figure 1): *in any triangle, the three points of intersection of the adjacent angle trisectors form an equilateral triangle.*

Morley never bothered to publish an elementary proof. In the book *Inversive Geometry* from 1933, which he wrote with his son Frank Vigor Morley, the authors, after stating this version of the theorem as a corollary of some rather intricate results on cardioids (certain heart-shaped curves, see Figure 5) simply ask: *Exercise 10: Verify this by trigonometry* [2, p.244]. Indeed, it is straightforward to verify that, in the notation of Figure 1,  $OP = 8r \sin \alpha \sin \beta \sin \gamma$ , where  $r$  equals the circumradius of triangle  $ABC$  (see [3, p.740]). This, by symmetry, proves that  $OP = PQ = QO$ , as desired.

Since the beginning of the twentieth century, many other proofs of Morley's theorem have been published. An article in the November 1978 issue of *The American Mathematical Monthly* by Cletus Oakley and Justine C. Baker (with supplements by Charles W. Trigg), see [3], lists no less than 150 references. Some only give a proof of the simple version of the theorem. But many proofs not only consider inner trisectors, but also their outer counterparts. In fact, if all trisectors are extended to full lines, there are precisely eighteen trisectors: six for each vertex of  $ABC$ , the outer ones making angles of  $\pm\pi/3$  with the inner trisectors through the same vertex. Note that the directed angle from one line to another is determined modulo  $\pi$ , so trisectors are determined modulo  $\pi/3$ . From these eighteen trisectors many more equilateral Morley triangles can be constructed, as will be indicated in the next section. Since 1978, numerous other articles and notes on Morley's theorem have appeared in print.

### A preview of Morley's analysis

At this point, the reader might like to get a preview of Morley's analysis. Let a triangle  $ABC$  be given. Extend its sides to full

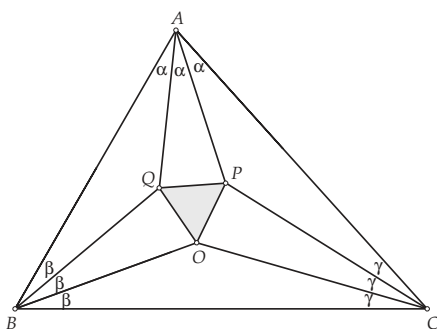


Figure 1 Morley's trisector theorem.

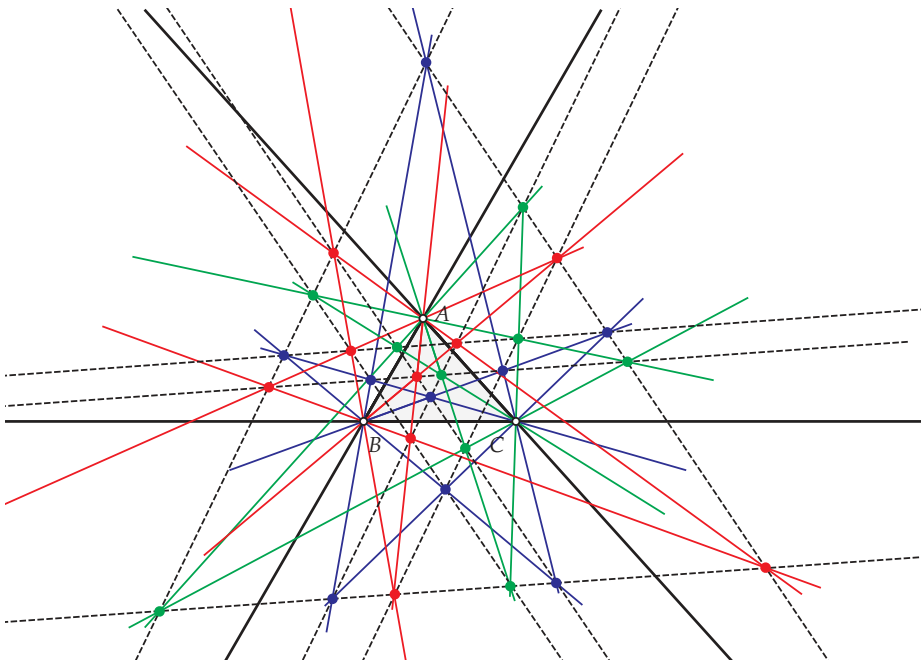


Figure 2 The full Morley trisector configuration.

lines. All eighteen trisectors of the triangle  $ABC$  from Figure 1, extended to full lines, are shown in Figure 2, colored in red, blue and green. To be more precise, the two inner trisectors through the endpoints of side  $AB$  and adjacent to  $AB$  are red, for  $BC$  this is blue and for  $CA$  this is green; each other trisector has the same color as the inner trisector through the same vertex of triangle  $ABC$  with which it makes an angle  $\pm\pi/3$ . Thus, six trisectors are red, six are blue and six are green. Trisectors of the same color but not at the same vertex intersect in points of the same color indicated by small dots: nine red, nine blue and nine green intersection points. The reader is invited to inspect the resulting collection of 27 colored intersection points of pairs of trisectors in Figure 2.

Morley considered cardioids that are tangent to each of the extended sides of triangle  $ABC$ . He discovered that the set of the so-called centers of these cardioids is very special: it consists of nine lines in three directions, three in each direction, intersecting in 27 points at angles  $\pm\pi/3$ . These lines are called the *axes* of the triangle. In Figure 2, the nine axes are drawn as dashed black lines. Furthermore, Morley proved the astonishing fact that *the 27 intersection points of the axes are precisely the 27 colored intersection points of pairs of trisectors constructed above*.

Now we are ready to explain the construction of the Morley triangles. Choose

an axis in each of the three directions. This can be done in 27 ways. Each of these creates an equilateral triangle, called a *Morley triangle*. From the Morley triangles, 18 are *proper*, having vertices in three different colors, and the remaining 9 triangles turn out to be *monochromatic*, having vertices in one color only. Triangle  $OPQ$  from Figure 1 is one of the proper Morley triangles.

In the book *Inversive Geometry* [2], it takes less than four pages (§137, §138 and §140 on pp.239–244) to elucidate the role of cardioids in the discovery of Morley's results, but this is not easy reading because of the idiosyncratic notation and style. The present paper offers self-contained explanations of this role of cardioids with full details of the proofs and with figures to illustrate the arguments. In this way, we hope to make this elegant, and almost forgotten fragment of analytic Euclidean geometry more accessible to modern readers.

### Clinants, bisectors and trisectors

Throughout this paper, the points in the plane will be viewed as complex numbers in the Argand plane (the Euclidean plane co-ordinated by complex numbers). Complex numbers will be represented by lower case letters. We will always use the letter  $t$  for points on the unit circle, so  $t\bar{t} = 1$ , in other words,  $\bar{t} = 1/t$ . Sometimes, we will also use  $\tau$  (the Greek letter 'tau') instead of  $t$ . Recall that, for given lines  $l$  and  $m$ ,

the directed angle from  $l$  to  $m$  is the angle through which a variable line has to be turned in the counterclockwise sense, in order to pass from the position  $l$  to the position  $m$ . This angle is considered modulo  $\pi$ . For parallel lines, this angle can be taken 0.

The *clinant* of a line is the complex number on the unit circle that has as its argument *twice* the directed angle from the real axis to the line. The clinant uniquely describes the direction of the line: two lines are parallel if and only if they have the same clinant. Furthermore, two lines are perpendicular if and only if their clinants differ by a factor  $-1$ .

The clinant will always be calculated — and arises in a natural way — as the quotient  $\tau = (z - x) / (\bar{z} - \bar{x})$ , where  $x$  and  $z$  are two distinct points on the line. Indeed, the vectors  $z - x$  and  $x - z$  both indicate the direction of the line, but in opposite sense. The quotient  $(z - x) / (\bar{z} - \bar{x})$ , however, remains the same if we interchange  $z$  and  $x$ . In fact, it is a complex number on the unit circle that only depends on the line, and not on the particular choice of  $z$  and  $x$  on this line. For any fixed point  $x$  on a line with clinant  $\tau$ , the equation  $(z - x) / (\bar{z} - \bar{x}) = \tau$ , or

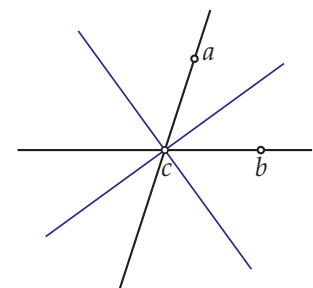
$$z - x = \tau(\bar{z} - \bar{x})$$

represents, for variable  $z$ , the points on the line.

Clinants are very convenient to define trisectors. First the idea is explained for the simpler case of bisectors.

**Definition of bisectors.** Let  $a$ ,  $b$  and  $c$  be distinct points, determining lines  $ac$  and  $bc$ . Let the clinants of the lines  $ac$  and  $bc$  be  $t_a$  and  $t_b$ , respectively. Then any line through  $c$  with clinant  $t$  satisfying  $t^2 = t_a t_b$  is called a *bisector* at  $c$  of the lines  $ac$  and  $bc$ .

Note that, by this definition, the directed angle at  $c$  from the line  $bc$  to the line  $ac$

Figure 3 Bisectors of lines  $ac$  and  $bc$ .

has precisely two bisectors, intersecting at right angles, see Figure 3. Their clinants are  $\pm\sqrt{t_a t_b}$ . Recall that each complex number different from 0 has two complex square roots, differing by a factor  $-1$ . Multiplying the directed angle from  $bc$  to either one of the two bisectors by a factor 2, yields the directed angle from  $bc$  to  $ac$ , modulo  $\pi$ . This justifies the terminology *bisector*.

In relation to Morley's theorem, trisectors are important. They are defined in a similar manner.

**Definition of trisectors.** Let  $a$ ,  $b$  and  $c$  be distinct points, determining lines  $ac$  and  $bc$ . Let the clinants of the lines  $ac$  and  $bc$  be  $t_a$  and  $t_b$ , respectively. Then any line through  $c$  with clinant  $t$  satisfying  $t^3 = t_a t_b^2$  is called a *trisector* at  $c$  of the lines  $ac$  and  $bc$ , adjacent to the line  $bc$ .

See Figure 4. The intuition behind this definition is that the direction of the line  $bc$  pulls twice as hard as the direction of the line  $ac$ . Note that there are *three* trisectors adjacent to  $bc$ , intersecting at angles  $\pm\pi/3$ . If  $\sqrt[3]{t_a t_b^2}$  is one of the (complex) cube roots of  $t_a t_b^2$ , the others are  $\omega\sqrt[3]{t_a t_b^2}$  and  $\omega^2\sqrt[3]{t_a t_b^2}$ , where  $\omega = e^{2\pi i/3}$ . These cube roots are the clinants of the trisectors adjacent to  $bc$ .

Multiplying the directed angle from  $bc$  to any trisector adjacent to  $bc$  by a factor 3, yields the directed angle from  $bc$  to  $ac$ , modulo  $\pi$ . This justifies the terminology *trisector*. By the definition above, any angle determined by two distinct intersecting lines, has *six* trisectors, three adjacent to one leg and three adjacent to the other. Note that the use of the word 'adjacent' in the definition of trisectors agrees with its use in the formulation of the simple version of Morley's theorem. Also note that in our definition the case that  $a$ ,  $b$  and  $c$  are collinear is not excluded: even in this case there are three trisectors at  $c$ , one of these being the line through  $a$ ,  $b$  and  $c$  itself.

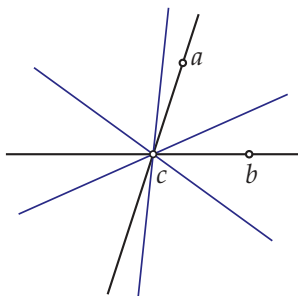


Figure 4 Trisectors adjacent to  $bc$ .

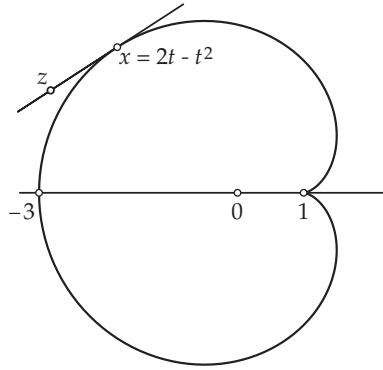


Figure 5 The standard cardioid.

### The standard cardioid and its tangent lines

When  $t$  runs through the unit circle, the equation

$$x = 2t - t^2 \quad (1)$$

is a parametric representation of a closed curve, called the *standard cardioid* (see Figure 5). More generally, a *cardioid* is a curve that is similar to the standard cardioid. Its name is derived from its heart-like shape. The point 0 is called the *center* of the cardioid. Centers of cardioids will play an important role in the sequel. The cardioid has a *cusp* when  $dx/dt = 0$ , which occurs for parameter value  $t = 1$  at the point  $x = 1$ . The real axis, characterized by  $z = \bar{z}$ , will be called the *cusp tangent line*. The point  $x = -3$ , taken for parameter value  $t = -1$ , is called the *apse* of the cardioid. Its tangent line is perpendicular to the cusp tangent line.

Now we will derive an equation for the tangent line at the point on the cardioid with parameter value  $t \neq 1$ . Let us first find its clinant. Taking differentials from the equation  $x = 2t - t^2$  yields

$$dx = 2(1-t)dt = 2(1-t)itd\theta$$

where the real number  $\theta$  is the argument of  $t = e^{i\theta}$ , always taken, as usual, modulo  $2\pi$ . It follows that

$$\begin{aligned} \overline{dx} &= 2(1-\bar{t})\overline{dt} = 2(1-\frac{1}{t})(-i)\frac{1}{t}d\theta \\ &= \frac{1}{t^3}2(1-t)itd\theta = \frac{1}{t^3}dx. \end{aligned}$$

Therefore,  $dx/\overline{dx} = t^3$ . Since in terms of differentials the clinant of the tangent line equals  $dx/\overline{dx}$ , we have proved: *the clinant of the tangent line at the point with parameter value  $t$  of the standard cardioid  $x = 2t - t^2$  equals  $t^3$ . Moreover, for any point  $z$  on the tangent line, we have*

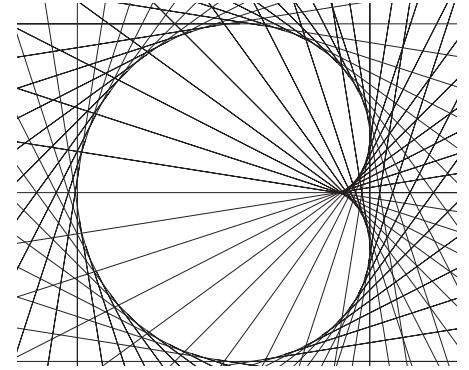


Figure 6 Tangent lines to the cardioid.

$$t^3 = \frac{z-x}{\bar{z}-\bar{x}} = \frac{z-(2t-t^2)}{\bar{z}-(2\bar{t}-\bar{t}^2)}$$

which, using  $\bar{t} = 1/t$ , can be simplified to

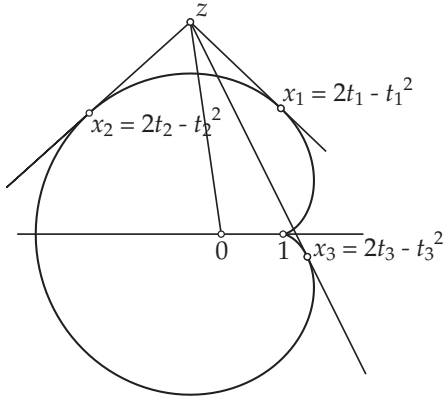
$$z - 3t + 3t^2 - \bar{z}t^3 = 0. \quad (2)$$

This is the promised equation for the tangent line for parameter value  $t$ . It also holds for  $t = 1$ : the tangent line at the cusp is  $z - \bar{z} = 0$  with clinant  $t^3 = 1$ .

As above, we will use the Greek letter  $\omega$  ('omega') to denote the complex number  $\omega = e^{2\pi i/3}$ . Note that the tangent lines to the cardioid for parameter values  $t$ ,  $\omega t$  and  $\omega^2 t$  are parallel. Indeed, their clinants are  $t^3$ ,  $(\omega t)^3$  and  $(\omega^2 t)^3$ , and so are all equal, as  $\omega^3 = 1$ .

Figure 6 shows the cardioid as the envelope of its tangent lines. The curve divides the plane into two regions: an outer region and an inner region. For each point  $z$  in the outer region, there are precisely three distinct points on the cardioid for which the tangent line runs through  $z$ . This is obvious from Figure 6, but it can also be explained as follows. Note that for any solution  $t = u$  of the cubic equation (2), conjugating the equation shows that also  $\bar{t} = 1/\bar{u}$  is a solution. Since  $u = 1/\bar{u}$  precisely when  $u$  is on the unit circle, (2) has either one or three solutions on the unit circle, counted with multiplicity. Multiple roots can be found by differentiating (2) with respect to  $t$  and conjugating, yielding  $z = 2t - t^2$ , so they occur precisely if  $z$  is on the cardioid. If  $z$  is in the outer region of the cardioid, then there are at least two distinct points on the cardioid with tangent lines through  $z$ . Hence, then there must always be a third point on the cardioid with tangent line through  $z$ .

Taking, as in Figure 7, a point  $z$  in the outer region of the cardioid, the cubic equation (2) has three distinct solutions  $t_1$ ,  $t_2$  and  $t_3$ , all situated on the unit circle.

Figure 7 Three tangent lines through  $z$ .

After dividing (2) by  $-\bar{z}$ , we see that the constant term is  $-z/\bar{z}$ , so

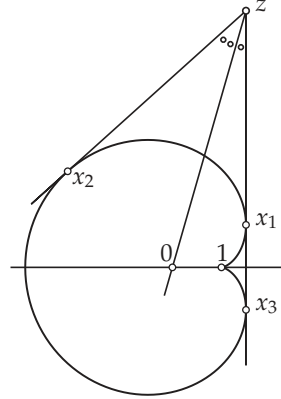
$$z/\bar{z} = t_1 t_2 t_3. \quad (3)$$

Taking cubes gives  $(z/\bar{z})^3 = t_1^3 t_2^3 t_3^3$ . This can be seen as a relation between clinants of certain lines. Indeed,  $z/\bar{z}$  is the clinant of the line connecting  $z$  to the center 0 of the cardioid, and  $t_k^3$  is the clinant of the tangent line for parameter value  $t_k$  for  $k = 1, 2, 3$ . Therefore, we get that the cube of the clinant of the line connecting  $z$  to the center 0 of the cardioid equals the product of the clinants of the three tangent lines from  $z$ . In other words, *the clinant of the line  $z0$  is one of the cube roots of the product of the clinants of the three tangent lines from  $z$* . Note that the cube roots of a complex number different from zero differ by a factor  $\omega^k$ , where  $k = 0, 1$  or  $2$ .

Equation (3) will play an important role in the next section, where the relation between tangent lines and trisectors will be explained. To help the reader to get an intuitive feeling for equation (3) in the case depicted in Figure 7, we reformulate it in terms of the arguments of the vectors  $0-z$ ,  $x_1-z$ ,  $x_2-z$  and  $x_3-z$ , pointing from  $z$  to 0,  $x_1$ ,  $x_2$  and  $x_3$ , respectively. Since the argument of a direction vector of a line is, modulo  $\pi$ , half the argument of the clinant of the line, we get

$$\arg(0-z) = \frac{1}{3}(\arg(x_1-z) + \arg(x_2-z) + \arg(x_3-z)) \pmod{\pi/3} \quad (4)$$

so the argument of  $0-z$  equals the *arithmetic mean* of the arguments of the three complex numbers  $x_k-z$  ( $k = 1, 2, 3$ ), modulo  $\pi/3$ . The reader is invited to check this in Figure 7 by verifying, using a protractor, that  $\angle x_2 z 0 = \angle 0 z x_1 + \angle 0 z x_3$ .

Figure 8  $z$  on the double tangent line.

### Trisectors and double lines

There is a unique line that is tangent to the standard cardioid in more than one point (see Figure 8): the double tangent line, or *double line* for short. It is vertical, so its clinant is  $-1$ . The parameter values  $t$  yielding a vertical tangent line satisfy  $t^3 = -1$ . The value  $t = -1$  gives the tangent line at the apse, so each of the other two possibilities,  $t = -\omega$  and  $t = -\omega^2$ , must give the double line. Indeed, substitution into (2) in both cases, using  $\omega^2 + \omega + 1 = 0$ , yields the same equation:

$$z + \bar{z} = 3.$$

Therefore, this equation represents the double line and the two tangency points  $x_3$  and  $x_1$  on the double line have parameter values  $-\omega$  and  $-\omega^2$ , respectively.

From any point  $z$  on the double line different from the two tangency points, there is precisely one other tangent line to the cardioid. Take  $z$  on the double line, but not on the closed interval between  $x_1$  and  $x_3$ , as in Figure 8. Then it follows from (4) that  $\angle x_2 z 0 = \angle 0 z x_1 + \angle 0 z x_3 = 2\angle 0 z x_1$ . Hence, the line connecting  $z$  to 0 is the *inner trisector* of the angle  $\angle x_1 z x_2$  adjacent to the

double line, i.e.,  $\angle x_1 z x_2 = 3\angle x_1 z 0$ . This is a first hint of the connection between cardioids and Morley's theorem.

To further explore this connection, let  $z_1 z_2 z_3$  be any triangle. From now on, we assume, as we may without loss of generality, that the standard cardioid  $x = 2t - t^2$  is situated inside triangle  $z_1 z_2 z_3$ , touching each of its sides, and that  $z_2 z_3$  is the double line, as in Figure 9. Note that this implies that the given triangle  $z_1 z_2 z_3$  is drawn in a way that is different from the triangle in the Figures 1 and 2, where it was denoted  $ABC$ . However, the choice of the coloring in Figure 9 (and later figures) for triangle  $z_1 z_2 z_3$  has been chosen according to the coloring in Figure 2 for triangle  $ABC$ .

The lines  $z_2 0$  and  $z_3 0$  (drawn in blue) are inner trisectors adjacent to  $z_2 z_3$ . They have been given the same color, as they are adjacent to the same side of the triangle. The other inner trisectors of the angles at  $z_3$  and  $z_2$  in triangle  $z_1 z_2 z_3$  are drawn in green and red, respectively.

Clearly, there are precisely two other cardioids situated inside triangle  $z_1 z_2 z_3$  that touch each side and have one side of the triangle as a double line. In Figure 10, the three cardioids with their accompanying inner trisectors are drawn in blue, green and red. Their centers are  $c_1 = 0$ ,  $c_2$  and  $c_3$ . Therefore, in order to prove the simple version of Morley's theorem, it suffices to show that the centers of the three cardioids form an equilateral triangle. However, from now on, we will aim directly at proving Morley's results in their general form. To this end, we first have to consider cardioids in general position.

### Cardioids in general position

In this section, the formulas (2) and (3), the clinant interpretation of (3) and the

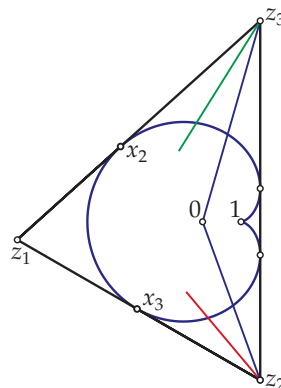
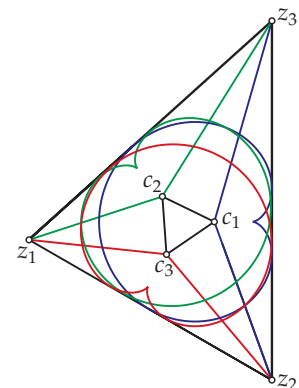
Figure 9 Inner trisectors adjacent to  $z_2 z_3$ .

Figure 10 Inner trisectors and cardioids.

relation between centers and trisectors for the standard cardioid are extended to cardioids in general position. For given complex numbers  $c$  and  $a$  with  $a \neq 0$ , the equation

$$x = c + 2a\tau - \bar{a}\tau^2 \quad (5)$$

represents, for parameter  $\tau$  running through the unit circle, a cardioid with center  $c$ . Indeed, with  $a = |a|e^{i\alpha}$ ,  $b = |a|e^{3i\alpha}$ ,  $\tau = e^{2i\alpha}t$ , equation (5) can be rewritten as

$$\frac{x-c}{b} = 2\frac{a}{b}\tau - \frac{\bar{a}}{b}\tau^2 = 2t - t^2. \quad (6)$$

Therefore, equations (5) and (6) both represent a cardioid with center  $c$  of any size in any orientation.

To find an equation of the tangent line to a cardioid with equation (5) for parameter value  $\tau$  with  $\tau \neq a/\bar{a}$  (the cusp value, given by  $dx/d\tau = 0$ ), we first determine its clinant by means of differentials, as we did for the standard cardioid. With  $\tau = e^{i\theta}$ , we have

$$dx = (2a - 2\bar{a}\tau)d\tau = (2a - 2\bar{a}\tau)i\tau d\theta$$

so

$$\begin{aligned} \overline{dx} &= (2\bar{a} - 2a\bar{\tau})\bar{i}\bar{\tau}d\theta \\ &= (2\bar{a} - 2a\frac{1}{\tau})(-i)\frac{1}{\tau}d\theta \\ &= \frac{1}{\tau^3}(2a - 2\bar{a}\tau)i\tau d\theta \\ &= \frac{1}{\tau^3}dx. \end{aligned}$$

Therefore, the clinant  $dx/\overline{dx}$  of the tangent line to the cardioid in general position  $x = c + 2a\tau - \bar{a}\tau^2$  at the point with parameter value  $\tau$  equals  $\tau^3$ . It follows that the tangent line is given by the equation  $z - x = \tau^3(\bar{z} - \bar{x})$  or, using  $x = c + 2a\tau - \bar{a}\tau^2$ ,

$$z - c - 2a\tau + \bar{a}\tau^2 = \tau^3(\bar{z} - \bar{c} - 2\bar{a}\frac{1}{\tau} + a\frac{1}{\tau^2})$$

which can be simplified to

$$(z - c) - 3a\tau + 3\bar{a}\tau^2 - \tau^3(\bar{z} - \bar{c}) = 0. \quad (7)$$

This is the equation of the tangent line to (5) at the point with parameter value  $\tau$ . It also holds for the cusp parameter  $\tau = a/\bar{a}$ . Then the equation of the tangent line is  $(z - c) - (a/\bar{a})^3(\bar{z} - \bar{c}) = 0$  with clinant  $\tau^3 = (a/\bar{a})^3$ . This completes the promised extension of equation (2) to cardioids in general position.

Special attention deserves the case that the double line of the cardioid is vertical. This occurs if and only if the complex number  $b$  in equation (6) is *real*, as this equation can be written as  $x = c + b(2t - t^2)$  and, moreover, the double line of the stan-

dard cardioid is vertical. However, if  $b$  is real (and  $b \neq 0$ ), we can write  $a$  instead of  $b$ , thus getting  $x = c + 2at - at^2$ , which is of the form (5) with  $a$  real and  $\tau = t$  (if  $b$  is not real, this is not the case!).

If  $a$  is real, substitution of the parameter values  $\tau = -\omega$  and  $\tau = -\omega^2$  in (7) yields the same equation, which therefore must be the equation of the double line. Using  $\omega^3 = 1$  and  $\omega^2 + \omega + 1 = 0$ , this equation can be simplified to

$$z + \bar{z} = c + \bar{c} + 3a \quad (a \text{ real}).$$

We thus have proved: *if a cardioid has a vertical double line, then its equation can be taken as  $x = c + 2a\tau - a\tau^2$  with  $a$  real. The tangency points of the cardioid on the double line then occur for parameter values  $\tau = -\omega$  and  $\tau = -\omega^2$ .*

Returning to the case of a cardioid  $\Gamma$  in general position, take a point  $z$  in the outer region of  $\Gamma$ . Then there are three distinct parameter values  $\tau_1, \tau_2, \tau_3$  for which  $z$  is on the tangent line. Hence, the cubic equation (7) in  $\tau$  has the three solutions  $\tau_1, \tau_2$  and  $\tau_3$ . After dividing (7) by  $-(\bar{z} - \bar{c})$ , we see that the constant term is  $-(z - c)/(\bar{z} - \bar{c})$ , so

$$(z - c)/(\bar{z} - \bar{c}) = \tau_1\tau_2\tau_3. \quad (8)$$

This is the extension of equation (3) to cardioids in general position. It follows from (8) that the clinant of the line connecting  $z$  to the center  $c$  is one of the cube roots of the product  $\tau_1^3\tau_2^3\tau_3^3$  of the clinants of the three tangent lines through  $z$ .

An important consequence of (8) is the following lemma, which describes the relation between centers and trisectors for cardioids in general position.

**Lemma 1.** *If  $z$  is a point on the double line of a cardioid with center  $c$ , then the line  $zc$  is a trisector at  $z$ , adjacent to the double line, of the angle from the double line to the other tangent line through  $z$  to the cardioid.*

*Proof.* Let again  $\tau_1, \tau_2, \tau_3$  be the parameter values of the three points on the cardioid for which the tangent line runs through the point  $z$ . Two of these values are the parameter values of the tangency points on the double line. Let these values be  $\tau_1$  and  $\tau_3$ . Then  $\tau_1^3 = \tau_3^3$  is the clinant of the double line, while  $\tau_2^3$  is the clinant of the other tangent line through  $z$ . It follows from (8) that the cube of the clinant of the line  $zc$  equals

$\tau_2^3(\tau_1^3)^2$ , so by the definition of trisectors,  $zc$  is one of the trisectors adjacent to the double line of the angle from the double line to the other tangent line through  $z$ .  $\square$

Note that Lemma 1 holds for all points  $z$  on the double line, even for the two tangency points. Then the other tangent line through  $z$  coincides with the double line, but also in that case it is true that  $zc$  is one of the three trisectors of the (zero) angle at  $z$ . Indeed, then the directed angle from the double line to  $zc$  is either  $\pi/3$  or  $-\pi/3$ , so multiplication by a factor 3 gives 0 modulo  $\pi$ .

The converse of Lemma 1 also holds:

**Lemma 2.** *If a cardioid  $\Gamma$  is tangent to each of two distinct intersecting lines  $za$  and  $zb$  and if the center  $c$  of  $\Gamma$  is on one of the trisectors adjacent to  $zb$  of the angle at  $z$  from  $zb$  to  $za$ , then  $zb$  is the double line of  $\Gamma$ .*

*Proof.* Let  $t_a, t_b$  and  $t_c$  be the clinants of  $za, zb$  and  $zc$ , respectively. Then  $t_c^3 = t_a t_b^2$ . Let  $z_d$ , with clinant  $t_d$ , be the third tangent line from  $z$  to  $\Gamma$ . It then follows from (8) that  $t_c^3 = t_a t_b t_d$ , so  $t_b = t_d$ . Hence, the lines  $zb$  and  $zd$  coincide, in other words,  $zb$  is the double line of  $\Gamma$ .  $\square$

### Trisectors and doubly inscribed cardioids

Morley discovered his results by studying the infinitely many cardioids touching the sides, extended to full lines, of a given triangle  $z_1 z_2 z_3$ . That is, he considered all cardioids touching each one of the three given lines  $z_2 z_3, z_3 z_1$  and  $z_1 z_2$ . We will call such cardioids *inscribed cardioids*. Note that inscribed cardioids need not be situated inside the triangle. Morley paid special attention to inscribed cardioids for which, moreover, one of the given lines is the double line (such as, for example, the three cardioids in Figure 10 and the nine cardioids in Figure 11). We will call such cardioids *doubly inscribed cardioids*.

**Lemma 3.** *For each line of triangle  $z_1 z_2 z_3$  there are nine doubly inscribed cardioids with that line as double line. The centers of these cardioids are precisely the nine points of intersection of pairs of trisectors adjacent to the double line at the two vertices on the double line.*

*Proof.* We assume without loss of generality that the chosen line is  $z_2 z_3$  and that

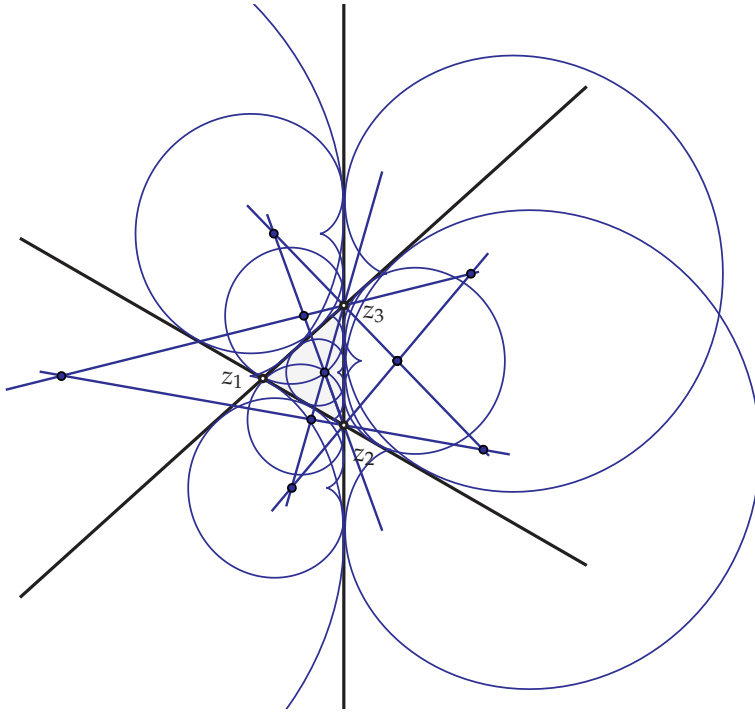


Figure 11 The 6 trisectors adjacent to  $z_2z_3$  and the 9 inscribed cardioids with double line  $z_2z_3$ .

the doubly inscribed cardioid with  $z_2z_3$  as double line that is situated inside triangle  $z_1z_2z_3$ , is the standard cardioid  $x = 2t - t^2$ .

We start our proof by investigating pairs of trisectors adjacent to  $z_2z_3$ , of course with one trisector at  $z_2$  and the other at  $z_3$ . At each of the vertices  $z_2$  and  $z_3$  there are three trisectors adjacent to  $z_2z_3$ , as shown in Figure 11, where they have been drawn in blue. Note that no pair of these trisectors is parallel. Indeed, in the notation of Figure 9, for the *inner* trisectors  $z_20$  and  $z_30$ , we have  $0 < \angle 0z_3z_2 + \angle 0z_2z_3 < \pi/3$ . Furthermore, for any other pair of trisectors adjacent to  $z_2z_3$ , we have to add integer multiples of  $\pi/3$  to these angles, but this never leads to an angle sum that is an integer multiple of  $\pi$ , so we never get parallel trisectors. Therefore, the  $2 \times 3 = 6$  trisectors give  $3 \times 3 = 9$  pairs of trisectors adjacent to  $z_2z_3$  and each of these pairs intersects in a point, which is indicated in Figure 11 by a small blue dot.

Any cardioid  $\Gamma$  with double line  $z_2z_3$  is determined as soon as its center is given. This follows from the definition of a cardioid as a curve in the plane that is similar to the standard cardioid. Therefore, each of the nine intersection points  $c$  of pairs of trisectors mentioned above defines a unique cardioid  $\Gamma$  with center  $c$  and double line  $z_2z_3$ . Since  $z_2c$  is a trisector adjacent to

$z_2z_3$  of the angle from  $z_2z_3$  to  $z_1z_2$ ,  $\Gamma$  is also tangent to  $z_1z_2$ , and since  $z_3c$  is a trisector adjacent to  $z_2z_3$  of the angle from  $z_2z_3$  to  $z_1z_3$ ,  $\Gamma$  is also tangent to  $z_3z_1$ . Hence  $\Gamma$  is a doubly inscribed cardioid and Lemma 3 is proved.  $\square$

#### The axes of triangle $z_1z_2z_3$

Perhaps the most surprising aspect of Morley's results on inscribed cardioids is his discovery that the set of their centers consists of nine lines in three directions, three in each direction, intersecting at angles  $\pm\pi/3$  in 27 points. These lines are the *axes* of triangle  $z_1z_2z_3$ . In Figure 2, the axes have been drawn as dashed black lines. The axes of a triangle are defined as follows.

**Definition of axes.** Let  $\Gamma$ , given by  $x = c + 2at - \bar{a}t^2$ , be an inscribed cardioid of a triangle  $z_1z_2z_3$ , and let  $t_1, t_2, t_3$  be parameter values of points where  $\Gamma$  touches the lines of the triangle. Then the line through the center  $c$  of  $\Gamma$  with clinant  $t_1t_2t_3$  is called an *axis* of the triangle.

Note that  $t_1, t_2, t_3$  are not uniquely defined for a doubly inscribed cardioid  $\Gamma$ . In that case there are two possibilities for the parameter value of the point where  $\Gamma$  touches the double line, giving two axes through its center.

**Lemma 4.** The clinant of any axis  $l$  of a triangle  $z_1z_2z_3$  is a cube root of the product of the clinants of its sides. Each point of  $l$  is the center of an inscribed cardioid.

*Proof.* The first statement follows from the definition of axes and the property of cardioids that the clinant of the tangent line for parameter value  $t$  is  $t^3$ .

Let  $c$  be the center of an inscribed cardioid  $\Gamma$  given by  $x = c + 2at - \bar{a}t^2$ . Suppose that  $t_1, t_2$  and  $t_3$  are parameter values for which  $\Gamma$  touches the lines  $z_2z_3, z_3z_1$  and  $z_1z_2$ , respectively, of triangle  $z_1z_2z_3$ . Let  $l$  be the axis through the center  $c$  of  $\Gamma$  with clinant  $t_1t_2t_3$ .

On account of (7), the tangent line at  $\Gamma$  for parameter value  $t$  is given by

$$(z - c) - 3at + 3\bar{a}t^2 - t^3(\bar{z} - \bar{c}) = 0. \quad (9)$$

For  $t = t_1, t = t_2, t = t_3$  this equation represents the lines of triangle  $z_1z_2z_3$ .

We will prove next that every point of the axis  $l$  is the center of an inscribed cardioid. But first, define the following symmetric expressions in  $t_1, t_2$  and  $t_3$ :

$$\begin{aligned} s_1 &= t_1 + t_2 + t_3, \\ s_2 &= t_1t_2 + t_2t_3 + t_3t_1, \\ s_3 &= t_1t_2t_3. \end{aligned}$$

Then

$$\bar{s}_1 = \frac{1}{t_1} + \frac{1}{t_2} + \frac{1}{t_3} = \frac{s_2}{s_3}.$$

Note that  $s_1 = 0$  would imply, as  $t_1, t_2$  and  $t_3$  are on the unit circle, that  $t_1, t_2$  and  $t_3$  are the vertices of an equilateral triangle. Consequently,  $t_1^3 = t_2^3 = t_3^3$  would hold, so the lines of triangle  $z_1z_2z_3$ , with equal clinants, would be parallel, contradiction. Hence, we may suppose that  $s_1 \neq 0$ .

Let  $c_1$  be any point on the axis  $l$  other than  $c$ . Then

$$(c_1 - c) / (\bar{c}_1 - \bar{c}) = t_1t_2t_3 = s_3.$$

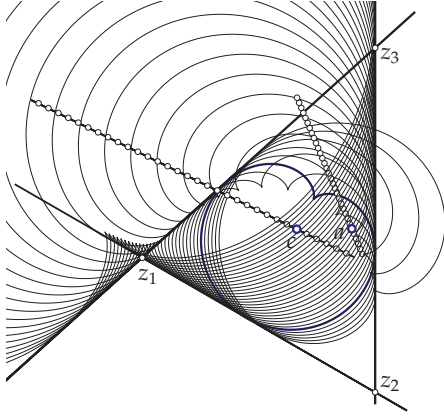
Define

$$a_1 = a - \frac{1}{3}(c_1 - c)\bar{s}_1. \quad (10)$$

We claim that  $a_1 \neq 0$  holds and that the cardioid  $\Gamma_1$ , given by the equation  $x = c_1 + 2a_1t - \bar{a}_1t^2$ , is an inscribed cardioid. Let, for each parameter value  $t$ , a line be given by the equation

$$(z - c_1) - 3a_1t + 3\bar{a}_1t^2 - t^3(\bar{z} - \bar{c}_1) = 0. \quad (11)$$

If  $a_1 \neq 0$ , this is the equation of the tangent line of  $\Gamma_1$  at the point with parameter value  $t$ . We will prove that, without any assump-



**Figure 12** A sequence of inscribed cardioids with their centers on an axis of  $z_1z_2z_3$ , together with the sequence of corresponding collinear  $a$ -values.

tion on  $a_1$ , equation (11) for  $t = t_1, t = t_2, t = t_3$  represents the lines  $z_2z_3, z_3z_1$  and  $z_1z_2$ , respectively. Then, as a consequence,  $a_1 \neq 0$  must hold, since otherwise, by (11), the point  $z = c_1$  would be on each of the three lines of  $z_1z_2z_3$ , which is impossible.

To prove our claims, note that, with the notations  $\Delta c = c_1 - c$  (so  $\Delta c / \Delta c = s_3$ ) and  $\Delta a = a_1 - a$ , we have, on account of (10)

$$\Delta a = -\frac{1}{3} \Delta c s_1 = -\frac{1}{3} \Delta c \frac{s_2}{s_3} = -\frac{1}{3} \frac{\Delta c}{\Delta c} s_2$$

so

$$\Delta c s_2 = -3 \Delta a$$

and, since  $\Delta a = -\frac{1}{3} \Delta c s_1$ ,

$$\Delta c s_1 = -3 \Delta a.$$

Hence, for  $t = t_1, t = t_2$  or  $t = t_3$ , we have

$$\begin{aligned} 0 &= \Delta c (t - t_1)(t - t_2)(t - t_3) \\ &= \Delta c (t^3 - t^2 s_1 + t s_2 - s_3) \\ &= \Delta c t^3 + 3 \Delta a t^2 - 3(\Delta a)t - \Delta c. \end{aligned}$$

Expanding and rearranging the last expression yields that, for  $t = t_1, t = t_2$  or  $t = t_3$ ,

$$(c_1 - c) + 3(a_1 - a)t - 3(\bar{a}_1 - \bar{a})t^2 - (\bar{c}_1 - \bar{c})t^3 = 0. \quad (12)$$

Subtracting equation (12) from equation (9) gives equation (11). Hence, for  $t = t_1, t = t_2$  and  $t = t_3$  equation (11), just like (9), represents the lines  $z_2z_3, z_3z_1$  and  $z_1z_2$ , respectively. Therefore,  $a_1 \neq 0$  holds and the cardioid  $\Gamma_1$  is an inscribed cardioid, as desired.  $\square$

The proof above shows that, as  $c_1$  runs over the axis  $l$ , the inscribed cardioid  $\Gamma_1$  with equation  $x = c_1 + 2a_1t - \bar{a}_1t^2$  that we have defined in the proof of Lemma 4 has the property that the parameters of the tangency points to the triangle  $z_1z_2z_3$  do

not depend on  $c_1$ . In particular, they are equal to the parameters  $t_1, t_2, t_3$  of the tangency points of  $\Gamma$  to  $z_1z_2z_3$ . Moreover, by (10), the point  $a_1$  also runs over a line. We will call this line the  $a$ -line.

As an illustration of Lemma 4 and the  $a$ -line, we have drawn in Figure 12 an inscribed but not doubly inscribed cardioid  $x = c + 2at - \bar{a}t^2$  with center  $c$  and point  $a$  (in blue), together with a sequence of inscribed cardioids with centers on the (uniquely determined) axis through  $c$ . The corresponding part of the  $a$ -line is also marked.

**Lemma 5.** *If a cardioid is doubly inscribed, then its center lies on precisely two axes.*

*Proof.* If  $\Gamma$  is a doubly inscribed cardioid with, say,  $z_2z_3$  as double line, then there are two parameter values  $t_1$  and  $t'_1$  for which  $\Gamma$  touches the double line. The clinants of the corresponding axes through the center  $c$  of  $\Gamma$  then are  $t_1t_2t_3$  and  $t'_1t_2t_3$ . By Lemma 4, a third axis through  $c$  would only be possible if  $c$  would also be the center of an inscribed cardioid  $\Gamma_1 \neq \Gamma$ . On account of Lemma 3,  $z_2c$  and  $z_3c$  are trisectors adjacent to  $z_2z_3$ , so on account of Lemma 2 also  $\Gamma_1$  must be doubly inscribed with  $z_2z_3$  as double line. It follows that  $\Gamma$  and  $\Gamma_1$ , as doubly inscribed cardioids with the same center and the same double line, must be the same. Contradiction.  $\square$

### Monochromatic Morley triangles

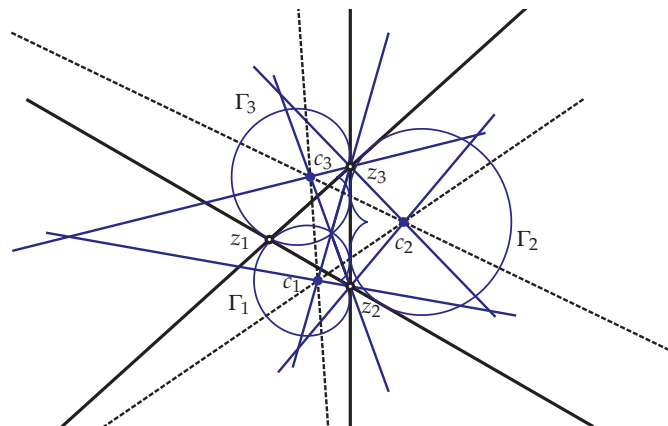
In this section, we will prove the following lemma on monochromatic Morley triangles, i.e., triangles with their sides along axes of triangle  $z_1z_2z_3$  and with vertices that are centers of doubly inscribed cardioids with the same double line.

**Lemma 6.** *For each center  $c_1$  of a doubly inscribed cardioid there exist two more centers  $c_2$  and  $c_3$  of doubly inscribed cardioids with the same double line such that  $c_1c_2c_3$  is a monochromatic Morley triangle.*

*Proof.* First, we will define  $c_2$  and  $c_3$ . Suppose that  $c_1$  is the center of a doubly inscribed cardioid  $\Gamma_1$  and that  $z_2z_3$  is its double line, which, as before, we assume to be vertical. By Lemma 3, its center  $c_1$  is the intersection of two trisectors  $z_3c_1$  and  $z_2c_1$  adjacent to  $z_2z_3$ . Let  $c_2$  and  $c_3$  be the intersections of the pairs of trisectors adjacent to  $z_2z_3$  that are obtained from the former pair by multiplying their clinants simultaneously by  $\omega$  or  $\omega^2$ , respectively (see Figure 13). We will prove that the (elongated) sides of triangle  $c_1c_2c_3$  are axes of triangle  $z_1z_2z_3$ . This will establish the lemma.

Let  $\Gamma_1$  be given by the equation  $x = c_1 + 2a_1t - \bar{a}_1t^2$  for some real number  $a_1$  (cf. section 'Cardioids in general position'). As  $a_1$  is real, the parameter values of the tangency points of  $\Gamma_1$  on the double line  $z_2z_3$  are  $t = -\omega$  and  $t = -\omega^2$ . Let  $t_2$  and  $t_3$  be the parameter values of the points of tangency of  $\Gamma_1$  on  $z_3z_1$  and  $z_1z_2$ , respectively. Then, on account of equation (8), the clinants of the trisectors  $z_3c_1$  and  $z_2c_1$  are  $(-\omega)t_2(-\omega^2) = t_2$  and  $(-\omega)t_3(-\omega^2) = t_3$ , respectively. Take  $t_1 = -\omega$ .

Let  $\Gamma_2$  be the doubly inscribed cardioid with center  $c_2$ , double line  $z_2z_3$  and equation  $x = c_2 + 2a_2\tau - \bar{a}_2\tau^2$  for some real  $a_2$ . Then, as above, the parameter values for the points of tangency on the double line are  $\tau = -\omega$  and  $\tau = -\omega^2$ . Let  $\tau_2$  and  $\tau_3$  be the parameter values of the points of tangency on  $z_3z_1$  and  $z_1z_2$ , respectively. Then, again by (8), the parameter values  $\tau_2$  and  $\tau_3$  are also the clinants of the trisectors  $z_3c_2$



**Figure 13** The monochromatic Morley triangle  $c_1c_2c_3$ .

and  $z_2c_2$ , which are equal to  $\omega t_2$  and  $\omega t_3$  by the definition of  $c_2$ . Therefore,  $\tau_2 = \omega t_2$  and  $\tau_3 = \omega t_3$ . This time, we take  $\tau_1 = -\omega^2 = \omega t_1$ , so  $\tau_k = \omega t_k$  for all  $k = 1, 2, 3$ .

The equation of the tangent line to the cardioid  $\Gamma_1$  for parameter value  $t$  (cf. (7)) is given by

$$(z - c_1) - 3a_1t + 3a_1t^2 - t^3(\bar{z} - \bar{c}_1) = 0. \quad (13)$$

The tangent line to  $\Gamma_2$  for  $\tau = \omega t$  is given by

$$(z - c_2) - 3a_2\omega t + 3a_2\omega^2t^2 - \omega^3t^3(\bar{z} - \bar{c}_2) = 0. \quad (14)$$

For  $t = t_1$ ,  $t = t_2$  or  $t = t_3$ , equations (13) and (14) both represent the lines  $z_2z_3$ ,  $z_3z_1$  or  $z_1z_2$ , respectively. Subtracting (13) from (14) and using  $\omega^3 = 1$  yields the following cubic equation in  $t$ :

$$(c_1 - c_2) + 3(a_1 - \omega a_2)t - 3(a_1 - \omega^2 a_2)t^2 - (\bar{c}_1 - \bar{c}_2)t^3 = 0. \quad (15)$$

Its three solutions are  $t_1$ ,  $t_2$  and  $t_3$ , so  $(c_1 - c_2)/(\bar{c}_1 - \bar{c}_2) = t_1 t_2 t_3$ . Hence, the clinant of the line  $c_1c_2$  equals  $t_1 t_2 t_3$ , which implies that the line  $c_1c_2$  is an axis of triangle  $z_1z_2z_3$ .

By cyclically permuting  $c_1$ ,  $c_2$  and  $c_3$ , it follows that also the lines  $c_2c_3$  and  $c_3c_1$  are axes of triangle  $z_1z_2z_3$ . Indeed, starting with  $c_2$  instead of  $c_1$ , in the calculation above we only have to replace  $t_2$  and  $t_3$  by  $\omega t_2$  and  $\omega t_3$ , respectively, while keeping  $t_1 = -\omega$ . Then it follows that the clinant of  $c_2c_3$  equals  $t_1(\omega t_2)(\omega t_3) = \omega^2 t_1 t_2 t_3$ . Similarly, starting with  $c_3$  yields  $t_1(\omega^2 t_2)(\omega^2 t_3) = \omega t_1 t_2 t_3$  as the clinant of  $c_3c_1$ . Therefore,  $c_1c_2c_3$  is a monochromatic Morley triangle.  $\square$

Figure 11 shows the nine centers of doubly inscribed cardioids with  $z_2z_3$  as double line as small blue dots. We have shown that three of these centers form a blue monochromatic Morley triangle. From the remaining six blue centers in a similar way two more blue monochromatic Morley triangles can be constructed. The reader is invited to identify them in Figure 11.

### Axes and doubly inscribed cardioids

In this section, we will prove the following lemma:

**Lemma 7.** *Each axis of triangle  $z_1z_2z_3$  contains precisely six centers of doubly inscribed cardioids, two for each side as double line.*

*Proof.* Choose an inscribed cardioid and let  $\tau_1$ ,  $\tau_2$  and  $\tau_3$  be parameter values of its tangency points with the lines  $z_2z_3$ ,  $z_3z_1$  and  $z_1z_2$ , respectively. Let  $l$  be the axis of  $z_1z_2z_3$  with clinant  $\tau_1\tau_2\tau_3$  through the center of the cardioid. In the section 'The axes of triangle  $z_1z_2z_3$ ' we have shown that for each point  $c$  on  $l$  it is possible to choose an inscribed cardioid  $\Gamma$  with center  $c$  and equation  $x = c + 2a\tau - \bar{a}\tau^2$  such that the parameter values of the tangency points of  $\Gamma$  with the lines  $z_2z_3$ ,  $z_3z_1$  and  $z_1z_2$  are  $\tau_1$ ,  $\tau_2$  and  $\tau_3$ , respectively, independently of the choice of  $c$  on  $l$ . Moreover, using equation (10), it was shown that, as  $c$  runs over the axis  $l$ , the point  $a$  also runs over a line, which we called the  $a$ -line.

As observed in the section 'Cardioids in general position', the  $a$ -line cannot pass through the origin, since for  $a = 0$  the cardioid degenerates into a point, and a point cannot touch all lines of triangle  $z_1z_2z_3$ . Therefore, if the center  $c$  runs over the axis  $l$  while  $a$  runs over the  $a$ -line, the argument of  $a$  runs over an open interval of length  $\pi$ . To be more precise, let  $a = |a|e^{i\alpha}$ , then we may take  $\alpha_0 < \alpha < \alpha_0 + \pi$  for a certain  $\alpha_0$ .

Now we will investigate for which points  $c$  on the axis  $l$  the chosen inscribed cardioid  $\Gamma$  with  $c$  as its center is doubly inscribed. In the section 'Cardioids in general position', we have rewritten the equation of  $\Gamma$  as  $x = c + b(2t - t^2)$  using  $b = |a|e^{3i\alpha}$  and  $\tau = e^{2i\alpha}t$ , thus showing in an explicit way its similarity to the standard cardioid  $x = 2t - t^2$  (cf. equation (6)). For the standard cardioid, the two points of contact with the double line occur for parameter values  $t = -\omega$  and  $t = -\omega^2$ , so for the cardioid  $\Gamma$  the two points of contact with its double line occur for parameter values  $\tau = -e^{2i\alpha}\omega$  and  $\tau = -e^{2i\alpha}\omega^2$ . If for some  $\alpha$  one of the two parameter values  $-e^{2i\alpha}\omega$  and  $-e^{2i\alpha}\omega^2$  equals  $\tau_1$ ,  $\tau_2$  or  $\tau_3$ , then  $\Gamma$  is a doubly inscribed cardioid with  $z_2z_3$ ,  $z_3z_1$  or  $z_1z_2$ , respectively, as double line. To be more precise, let  $\tau_k = e^{i\theta_k}$  for  $k = 1, 2, 3$ , and suppose that for a certain  $\alpha$  we have  $-e^{2i\alpha}\omega = \tau_k = e^{i\theta_k}$ . Then  $e^{2i\alpha} = e^{i\pi}e^{-2\pi i/3}e^{i\theta_k}$  so  $\alpha = \frac{1}{2}(\pi/3 + \theta_k)$  modulo  $\pi$ . Similarly, if we have  $-e^{2i\alpha}\omega^2 = \tau_k = e^{i\theta_k}$  then  $e^{2i\alpha} = e^{i\pi}e^{-4\pi i/3}e^{i\theta_k}$  so  $\alpha = \frac{1}{2}(-\pi/3 + \theta_k)$  modulo  $\pi$ . Therefore, as a rule, there are precisely six values of  $\alpha$  in the open interval  $\langle \alpha_0, \alpha_0 + \pi \rangle$ , given by  $\alpha = \frac{1}{2}(\pm \pi/3 + \theta_k)$  modulo  $\pi$ , corresponding to six centers  $c$  on the axis  $l$  for which  $\Gamma$  is doubly in-

scribed, two centers for each line of triangle  $z_1z_2z_3$  as double line.

The only exception would occur if one of the six  $\alpha$ -values would be equal to  $\alpha_0$  modulo  $\pi$ . In that case, for one of the lines of triangle  $z_1z_2z_3$ , say for  $z_2z_3$ , there would be only one center  $c_1$  of a double inscribed cardioid with  $z_2z_3$  as double line on  $l$ . However, on account of Lemma 6,  $c_1$  is a vertex of a monochromatic Morley triangle  $c_1c_2c_3$ . The axes through  $c_1$  then are the lines  $c_1c_2$  and  $c_1c_3$ , and since by Lemma 5 no other axes through  $c_1$  are possible,  $l$  must be one of these, contradicting our assumption that  $c_1$  is the only center of a doubly inscribed cardioid with  $z_2z_3$  as double line on  $l$ .  $\square$

### The full Morley trisector configuration

In the section 'Monochromatic Morley triangles', we identified three monochromatic blue Morley triangles. Their vertices are the nine centers of doubly inscribed cardioids with  $z_2z_3$  as double line. On account of Lemma 7, the nine axes along their sides must all be different. We claim that this set of nine axes is complete: there can be no other axes. Indeed, choose any axis. Then by Lemma 7 it contains a blue point. By Lemma 6 this point is a vertex of a monochromatic blue Morley triangle, so it lies on two of the nine axes, and by Lemma 5 it cannot lie on any other axis. Therefore the chosen axis is one of the nine.

The nine axes occur in three directions, three in each direction, intersecting in  $3 \times 3 \times 3 = 27$  points, nine blue points, nine red points and nine green points, with two points of each color on each axis and two axes through each colored point. Note that two points of the same color on an axis are vertices of a unique monochromatic Morley triangle. Since there are  $3 \times 3 = 9$  monochromatic Morley triangles, the remaining  $27 - 9 = 18$  Morley triangles are 'proper', having vertices in three different colors.

It would be nice to produce an animation based on Figure 2 in which one could trace the nine axes, while for each center  $c$  on an axis the corresponding cardioid touching the sides of triangle  $z_1z_2z_3$  were shown (cf. Figure 12). For reasons of clarity, we have abstained from adding cardioids to Figure 2.

This concludes our elaboration of the text on cardioids and Morley's full trisector configuration on pages 239–244 in [2]. We summarize our lemmas and other results in the following theorem.

**Theorem.** Let  $z_1 z_2 z_3$  be an arbitrary triangle with sides extended to full lines. Let  $t_0$  be the product of the clinants of these lines. Then:

1. The set of all centers of inscribed cardioids — cardioids touching each of the three lines of triangle  $z_1 z_2 z_3$  — consists of nine lines, called the axes of the triangle, in three directions under angles  $\pm\pi/3$ , three axes in each direction. The clinants of the axes are the three cube roots of  $t_0$ .
2. The 27 intersection points of the axes are the centers of the 27 doubly inscribed cardioids, i.e., the inscribed cardioids having one of the lines of  $z_1 z_2 z_3$  as double line, nine cardioids for each line of  $z_1 z_2 z_3$ .
3. Each axis contains exactly six centers of doubly inscribed cardioids, two for each line of  $z_1 z_2 z_3$  as double line.
4. For each side of the triangle, the nine centers of the doubly inscribed cardi-

oids with this side as double line are the nine points of intersection of pairs of trisectors adjacent to this side, at different endpoints.

Consider the 27 equilateral triangles that are created by making a choice of three axes, one in each direction.

5. For nine of these equilateral triangles, the vertices are intersection points of pairs of trisectors adjacent to the same side of triangle  $z_1 z_2 z_3$ , three triangles for each side of  $z_1 z_2 z_3$ .
6. For the remaining 18 equilateral triangles, the vertices are intersection points of pairs of trisectors adjacent to the three sides of  $z_1 z_2 z_3$ , one vertex for each side.

It seems appropriate to end with a quotation from a letter by Frank Vigor Morley to Cletus Oakley, who had asked him about the origins of his father's theorem [3, p.741]:

“Now my father did not lack warmth for any geometric property so simple and startling as this one. I never asked him outright the question, though it is a proper one, that Professor Oakley now asks me, namely, why at the time of the discovery my father kept his cool about promoting the ‘gem’—there might have been some bit of hoo-ha if he had removed the cover and sent it to the showroom as a separate static cut stone. I think the way the theorem is presented in the book *Inversive Geometry* [2] may answer the question. Attention to the detached theorem was not, for him, to interfere with the pleasure of watching his ‘mobile’ of cardioids and their tangents: it was the cardioids which led him to, and provided for him the most elegant proof of, the trisector theorem. Proof and theorem were pleasing in their togetherness.”

## References

- 1 H.S.M. Coxeter, *Introduction to Geometry*, 2nd ed., Wiley, 1969.
- 2 Frank Morley and F.V. Morley, *Inversive Geometry*, Ginn, 1933. [Reissued by Chelsea, 1954 and by Dover Books on Mathematics, 2013.]
- 3 Cletus O. Oakley and Justine C. Baker, The Morley trisector theorem, *The American Mathematical Monthly* 85(9) (1978), 737–745.

## Jan Peter van Zandwijk

Team Forensische Digitale Technologie  
Afdeling Digitale en Biometrische Sporen  
Nederlands Forensisch Instituut  
j.p.van.zandwijk@nfi.minvenj.nl

### Maatschappij

# Wiskunde helpt misdaad oplossen

Het Nederlands Forensisch Instituut (NFI) in Den Haag heeft tot taak door onafhankelijk forensisch onderzoek de waarheidsvinding bij strafrechtelijk onderzoek te bevorderen. Bij het team Forensische Digitale Technologie (FDT) wordt forensisch onderzoek gedaan naar diverse typen digitale gegevensdragers, zoals harde schijven en mobiele telefoons. Een van de diensten die het team DT aan haar klanten levert, is het uitlezen van gegevens die aanwezig zijn op een geheugenchip die zich bijvoorbeeld in een smartphone, navigatieapparatuur of tabletcomputer bevindt. Vaak is het mogelijk gegevens aan te leveren die direct geschikt zijn voor verdere verwerking, maar soms is extra werk nodig om de uitgelezen gegevens inzichtelijk te maken. Jan Peter van Zandwijk legt uit hoe dit met wat wiskunde mogelijk is.

#### NAND-flash geheugenchips

Vrijwel alle moderne elektronische apparatuur bevat geheugenchips. Een veelgebruikt type geheugenchip is het zogenaamde NAND-flash, in de jaren tachtig bedacht door de Japanner Fujio Masuoka. Dit type geheugenchip is goedkoop en slaat gegevens op in de vorm van elektrische lading in cellen die zich in de chip bevinden. Door deze wijze van opslag blijven de gegevens ook bewaard als het apparaat uit staat of de batterij leeg is. Om gegevens eenvoudig toegankelijk te maken, is een NAND-flashgeheugen hiërarchisch georganiseerd: het is opgedeeld in blokken, die elk weer uit een aantal geheugenpagina's bestaan. In elektronische apparatuur wordt het

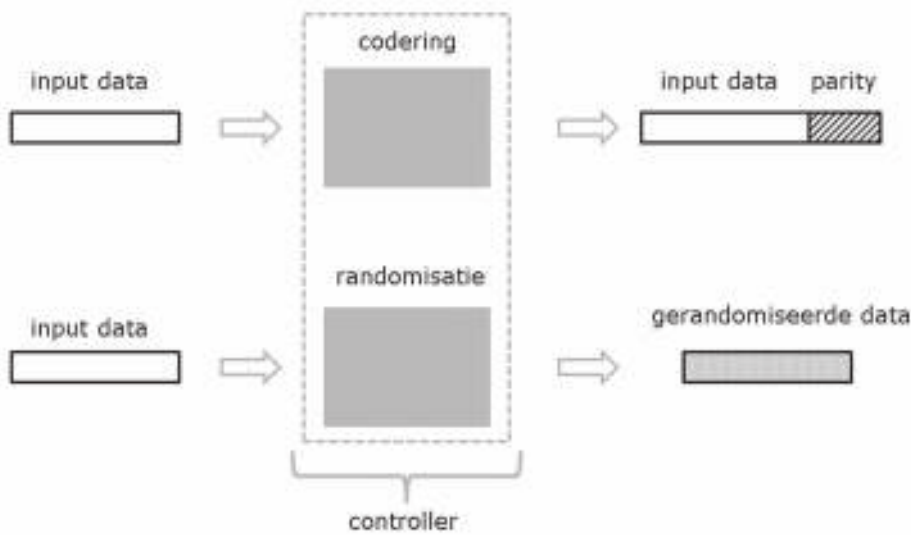
NAND-flashgeheugen beheerd door de controller. Dit is een speciaal stukje elektronica dat de communicatie tussen de geheugenchip en de buitenwereld voor zijn rekening neemt.

Iedere keer als gegevens naar het geheugen worden geschreven of uit het geheugen worden gelezen voert de controller een aantal taken uit om ervoor te zorgen dat dit foutloos gebeurt. Dit is belangrijk omdat NAND-flash een enigszins onbetrouwbaar opslagmedium is: teruggelezen gegevens hoeven niet hetzelfde te zijn als de oorspronkelijk opgeslagen gegevens. We zeggen dan dat er bitfouten in het geheugen kunnen optreden.

Fabrikanten gebruiken specifieke technieken om de betrouwbaarheid van NAND-

flashgeheugens te verhogen. Veelgebruikte technieken zijn randomisatie en foutverbeterende codes (Error-Correcting-Codes, of ECCs). Bij randomisatie wordt een pseudo-random bitrij opgeteld bij gegevens voordat deze in het NAND-flash opgeslagen worden. Hierdoor worden patronen gemaskeerd die zich mogelijk in de gegevens bevinden en wordt voorkomen dat NAND-flashgeheugencellen elkaar beïnvloeden. Bij teruglezen wordt de pseudo-random bitrij er weer afgetrokken, waardoor de oorspronkelijke gegevens weer tevoorschijn komen. Bij het gebruik van foutverbeterende codes wordt extra informatie ('parity bits') aan de gegevens toegevoegd, waardoor het mogelijk is na te gaan of deze correct zijn en het mogelijk wordt deze eventueel te corrigeren. Het toevoegen van randomisatie en toepassen van ECCs om parity bits te berekenen gebeurt door de controller op het moment dat gegevens in het geheugen opgeslagen worden. De buitenwereld merkt hier helemaal niets van. Figuur 1 illustreert de processen die plaatsvinden bij het opslaan van gegevens in een NAND-flashgeheugenchip.





**Figuur 1** Schematische weergave van processen die plaatsvinden wanneer gegevens op een NAND-flashgeheugen worden opgeslagen. De controller voert operaties uit om integriteit van de gegevens te waarborgen. Dit zijn het toevoegen van randomisatie en het berekenen van parity bits van een foutverbeterende code. De volgorde waarin deze operaties worden uitgevoerd is onderling verwisselbaar: de controller kan eerst randomisatie toevoegen en dan parity bits berekenen voor de gerandomiseerde gegevens of eerst parity bits berekenen en dan pas randomisatie toevoegen.

### Veiligstellen van gegevens

Gegevens die zich op een NAND-flashchip bevinden kunnen in een digitaal forensisch onderzoek op verschillende manieren veiliggesteld worden. Het eenvoudigst is om deze via de controller op te vragen. De controller zorgt er dan voor dat randomisatie verwijderd wordt en gebruikt ECC-informatie om eventuele fouten in de gegevens te corrigeren. In dit geval zijn de gegevens die zich op de NAND-flashgeheugenchip bevinden, direct beschikbaar voor verder forensisch onderzoek.

Ingewikkelder wordt het als het niet mogelijk is de controller te gebruiken om gegevens veilig te stellen. Dit kan bijvoorbeeld het geval zijn wanneer de controller kapot is of toegang tot de controller beveiligd is met een wachtwoord. In dergelijke gevallen is het vaak wel mogelijk de NAND-flashchip uit het apparaat los te solderen en vervolgens met specialistische apparatuur uit te lezen. Er wordt dan een ruwe dump van gegevens verkregen zoals die door de controller in het geheugen zijn geplaatst. Dit betekent dat deze nog de door de controller toegevoegde randomisatie bevatten. Daarnaast kunnen de gegevens nog een onbekend aantal bitfouten bevatten omdat ECC-informatie nog niet is toegepast. Om deze geschikt te maken voor verder onderzoek moet er eerst een nabewerking plaatsvinden waarbij randomisatie verwijderd wordt en ECC-informatie toegepast wordt, twee taken die normaal door de controller

afgehandeld worden. Complicatie hierbij is dat informatie over de voor randomisatie gebruikte pseudo-random bitrij en over de gebruikte ECC doorgaans niet door de fabrikant gespecificeerd wordt en ook niet op andere wijze beschikbaar is.

Dit betekent gelukkig niet dat de zaak hopeloos is: in bepaalde gevallen is het mogelijk op basis van de ruwe dump te achterhalen wat de pseudo-random bitrij en ECC geweest zijn. Deze informatie kan vervolgens gebruikt worden om zelf gegevens uit de ruwe dump van randomisatie te ontdoen en eventuele bitfouten te corrigeren, waardoor de ruwe dump toch inzichtelijk gemaakt kan worden.

### Achterhalen van randomisatie

Om de voor randomisatie gebruikte pseudo-random bitrij te achterhalen, kun je veronderstellen dat deze rij door een schuifregister (*linear feedback shift register* of LFSR) geproduceerd is. Zie het kader hieronder voor meer uitleg. Het is niet gek deze veronderstelling te maken: schuifregisters zijn eenvoudig in hardware te realiseren, zijn snel en produceren bitrijen met goede statistische eigenschappen. Alle kans dus, dat de controller er een aan boord heeft. Het is mogelijk snel te testen of een gegeven bitrij met een schuifregister geproduceerd is met het beroemde Berlekamp-Massey-algoritme (BM-algoritme). Dit algoritme is uit de coderingstheorie afkomstig en vond zijn oorsprong in een methode

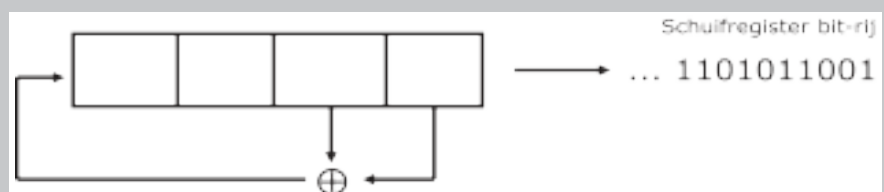
### Schuifregisters

Schuifregisters zijn structuren die uit een aantal cellen bestaan. Iedere keer dat het register geklokt wordt, schuift de inhoud van elke cel één positie naar rechts. De inhoud van de laatste cel is de output van het register. De nieuwe vulling van de eerste cel wordt bepaald op basis van de vulling van het register voordat deze een stap neemt. Het terugkoppelpolynoom bepaalt welke cellen in het register bij deze berekening betrokken zijn. Is de terugkoppeling lineair, dan spreekt men van een *Linear Feedback Shift Register* (LFSR).

Parameters van een LFSR zijn de lengte  $L$ , terugkoppelpolynoom  $c(x)$  en initiële vulling. Zijn deze allen bekend, dan ligt de bitrij  $\{s_i\}_{i=0}^N$  die door het register geproduceerd wordt, volledig vast. De eerste  $L$  bits van de rij  $\{s_i\}_{i=0}^N$  zijn gelijk aan de beginvulling van het register en de overige bits worden uniek bepaald door de lineaire recursie

$$s_j = \sum_{i=1}^L c_i s_{j-i} \quad \text{voor } j = L, L+1, \dots \quad (L1)$$

Hierin zijn de getallen  $c_i$  de coëfficiënten van het terugkoppelpolynoom  $c(x)$ . Schuifregisters produceren bitrijen die periodiek zijn. Maximale periode van de rij geproduceerd door een LFSR van lengte  $L$  is  $2^{L-1}$ .



Voorbeeld van een lineair schuifregister met vier cellen.

```

1
1 + x^2
1 + x^2
1
1 + x^3
1 + x^3
1 + x^2 + x^3
1 + x^1 + x^2 + x^3 + x^4
1 + x^1 + x^3 + x^4 + x^5
1 + x^1 + x^3 + x^4 + x^5
1 + x^1 + x^3 + x^4 + x^5
1 + x^1 + x^6 + x^7
1 + x^1 + x^6 + x^7
1 + x^1 + x^6 + x^7
1 + x^1 + x^3 + x^4 + x^8
1 + x^1 + x^3 + x^4 + x^8
1 + x^1 + x^3 + x^4 + x^8
1 + x^1 + x^3 + x^4 + x^8
1 + x^1 + x^3 + x^4 + x^8
1 + x^1 + x^3 + x^4 + x^5 + x^6 + x^8 + x^11 + x^12
1 + x^1 + x^3 + x^4 + x^5 + x^6 + x^8 + x^11 + x^12
1 + x^1 + x^2 + x^4 + x^8 + x^10 + x^11 + x^12
1 + x^1 + x^2 + x^4 + x^8 + x^10 + x^11 + x^12
1 + x^1 + x^2 + x^5 + x^7 + x^10 + x^11
1 + x^1 + x^2 + x^5 + x^7 + x^10 + x^11
1 + x^1 + x^2 + x^5 + x^6 + x^9 + x^11 + x^14
1 + x^3 + x^5 + x^8 + x^9 + x^12 + x^14
1 + x^3 + x^5 + x^8 + x^9 + x^12 + x^14
1 + x^3 + x^5 + x^8 + x^9 + x^12 + x^14
1 + x^3 + x^5 + x^8 + x^9 + x^12 + x^14
1 + x^3 + x^5 + x^8 + x^9 + x^12 + x^14

```

**Figuur 2** Voorbeeld van het zoeken naar een lineair schuifregister in stuk data afkomstig uit een ruwe dump van een NAND-flashgeheugen. Elke regel bevat het schuifregister dat voor de desbetreffende positie in de bitrij door het BM-algoritme bepaald is. Merk op dat het berekende register voor de laatste vijf posities hetzelfde is. Dit is eveneens het geval voor de daarop volgende posities, die niet in de figuur getoond zijn. Op basis hiervan is vastgesteld dat dit het register is dat gebruikt is voor randomisatie van de gegevens in de ruwe dump.

van Elwyn Berlekamp om BCH-codes te decoderen. Later is het verder aangepast door James Massey [1], vandaar dat het nu hun beider naam draagt. Het BM-algoritme berekent voor elke positie in een bitrij het kortste LFSR dat de rij tot die positie kan produceren. Zie het kader ‘Berlekamp–Massey-algoritme’ voor meer uitleg.

Op basis van het BM-algoritme kan eenvoudig een test op een schuifregister geconstrueerd worden. De truc is, om het BM-algoritme op achtereenvolgende posities in de te analyseren bitrij toe te passen en bij te houden welk LFSR het algoritme oplevert. Indien de bitrij inderdaad door een LFSR geproduceerd is, zul je zien dat na wat initiële variaties, het berekende LFSR steeds hetzelfde blijft. Zie je dat het berekende LFSR voor een groot aantal opeenvolgende posities in de bitrij niet verandert, dan kan je ervan uitgaan dat dit degene is die gebruikt is om de bitrij te produceren.

Indien voor deze methode een stuk data uit de ruwe dump als input van een NAND-flashgeheugen gebruikt wordt, wordt impliciet de veronderstelling gemaakt dat de onderliggende inhoud van deze gegevens uit nullen bestaat. Dit hoeft natuurlijk niet te kloppen en de methode zal geen

resultaat in de vorm van een schuifregister opleveren indien de impliciete aanname onjuist is. In dat geval kun je het proces herhalen met data van andere locaties uit de ruwe dump totdat je geluk hebt en de berekening toepast op een stuk data waarvan de onderliggende gegevens wél uit nullen bestaan. Een alternatief is, om een andere veronderstelling over de inhoud van de onderliggende gegevens te maken, deze eerst van de data uit de ruwe dump af te trekken en daarna de berekening uit te voeren. In de praktijk blijkt dit niet echt nodig: veel ruwe dumps van NAND-flash-geheugens blijken gerandomiseerde stuk-

ken nullen te bevatten, die eenvoudig met bovenstaande methode geanalyseerd kunnen worden. Het overgrote deel van de op deze wijze op het NFI geanalyseerde ruwe NAND-flashdumps blijkt randomisatie met een door een LFSR geproduceerde bitrij te bevatten. Figuur 2 geeft een voorbeeld van het zoeken naar een LFSR door middel van het BM-algoritme op de wijze die zojuist is beschreven.

Is eenmaal vastgesteld dat de voor randomisatie gebruikte pseudo-random bitrij door een schuifregister is geproduceerd, moet nog voor elke geheugenpagina vastgesteld worden wat de initiële vul-

### Berlekamp–Massey-algoritme

Het Berlekamp–Massey-algoritme (BM-algoritme) bepaalt recursief voor alle posities in een gegeven bitrij  $\{s_i\}_{i=0}^N$  de terugkoppelcoëfficiënten en de lengte van een LFSR die de rij tot en met die positie produceert. Stel, vanwege de recursie, dat we voor het  $n$ -de bit in de bitrij een LFSR met lengte  $L_n$  gevonden hebben. Dan geldt op basis van (L1) in kader ‘Schuifregisters’:

$$s_j = \sum_{i=1}^{L_n} c_i s_{j-i} \quad \forall j: L_n \dots (n-1). \quad (1)$$

Alle variabelen in deze formule zijn binair (0 of 1), dus (1) is equivalent met

$$s_j + \sum_{i=1}^{L_n} c_i s_{j-i} = 0 \quad \forall j: L_n \dots (n-1). \quad (2)$$

De taak is nu om een LFSR voor de volgende positie,  $n$ , in de rij te bepalen. Daartoe berekenen we

$$d_n = s_n + \sum_{i=1}^{L_n} c_i s_{n-i} \quad (3)$$

Hierin is  $d_n$  het verschil tussen het volgende door het LFSR geproduceerde bit en het volgende bit in de rij. Is dit verschil 0, dan hoeven we natuurlijk niets aan te passen en gaan we naar de volgende positie, waar we dan  $d_{n+1}$  uitrekenen. Is het verschil 1, dan moet het LFSR aangepast worden. De elegante truc van het BM-algoritme is om hiervoor een LFSR te gebruiken dat eerder bij analyse van de rij gevonden is. Veronderstel dat we op de eerdere positie  $m < n$  in de rij hetzelfde probleem zijn tegengekomen, namelijk dat  $s_m + \sum_{i=1}^{L_m} \hat{c}_i s_{m-i} = d_m = 1$ , terwijl

$$s_j + \sum_{i=1}^{L_m} \hat{c}_i s_{j-i} = 0 \quad \forall j: L_m \dots (m-1) \quad (4)$$

met de bij dat LFSR horende terugkoppelcoëfficiënten.

Informatie over dit eerder gevonden register kan met (2) gecombineerd worden tot de volgende uitdrukking:

$$s_j + \sum_{i=1}^{L_n} c_i s_{j-i} + s_{j-(n-m)} + \sum_{i=1}^{L_m} \hat{c}_i s_{j-(n-m)-i} \quad (5)$$

Uitdrukking (5) is 0 voor waarden van de index  $j < n$  op basis van (2) en (4) en  $j = n$  invullen levert de waarde  $d_n + d_m = 1 + 1 = 0$ . Met andere woorden, we hebben op deze wijze impliciet een LFSR geconstrueerd dat de eerste  $n+1$  bits van de rij kan genereren. Hierna kunnen we door naar de volgende positie en daar weer  $d_{n+1}$  uitrekenen. Uiteraard is bovenstaande niet meer dan een schets van het BM-algoritme, zonder al te veel wiskundige striktheid. Een complete beschrijving en alle bewijzen zijn te vinden in het originele artikel van Massey [1].

ling van het register is. Dit is meestal nogal wat gepuzzel, maar blijkt in veel gevallen wel mogelijk.

Als bovenstaande informatie bekend is, is het mogelijk alle pagina's in een ruwe dump van een NAND-flashgeheugen te ontdoen van randomisatie, waardoor de inhoud van de pagina's inzichtelijk wordt. Dit is al een hele stap, maar nog niet voldoende, aangezien de zo geproduceerde gegevens nog niet gecorrigeerde bitfouten kunnen bevatten. Als de inhoud van de pagina gewone tekst bevat, hoeven bitfouten geen grote problemen op te leveren en kunnen zelfs handmatig gecorrigeerd worden. Maar als de pagina bijvoorbeeld gecompriëerde gegevens bevat (denk bijvoorbeeld aan een WinZip-bestand of een jpg-plaatje), dan kan een enkele fout in de data desastreuze gevolgen hebben en er toe leiden dat data niet verder verwerkt kan worden. Het is daarom van belang ook te achterhalen welke ECC gebruikt is om de parity bits te berekenen die in de ruwe dump van een NAND-flashgeheugen aanwezig zijn.

### Achterhalen van ECC-parameters

Zie het kader hiernaast voor een korte introductie over ECCs. In de context van NAND-flashgeheugens voeren we nu de volgende notatie in. De controller krijgt een berichtwoord  $\mathbf{m} = (m_1 \dots m_k)$  aangeleverd en produceert codewoorden  $\mathbf{c} = (c_1 \dots c_n)$  door op basis van  $\mathbf{m} = (m_1 \dots m_k)$  en de gebruikte ECC parity bits te berekenen. Deze codewoorden  $\mathbf{c} = (c_1 \dots c_n)$  worden vervolgens in het geheugen opgeslagen. Bij teruglezen ontvangt de controller een datawoord  $\mathbf{d} = (d_1 \dots d_n)$  uit het NAND-flashgeheugen. Indien het datawoord  $\mathbf{d} = (d_1 \dots d_n)$  geen bitfouten bevat, is het natuurlijk een geldig codewoord  $\mathbf{c} = (c_1 \dots c_n)$ . Zijn er wel bitfouten, dan kan de controller deze corrigeren door een decodeermethode voor de gebruikte ECC toe te passen en zo een datawoord  $\mathbf{d} = (d_1 \dots d_n)$  te herleiden tot het 'dichtstbijzijnde' codewoord  $\mathbf{c} = (c_1 \dots c_n)$ .

Het probleem waar wij ons hier mee bezighouden is, om parameters van de gebruikte ECC te achterhalen op basis van een collectie datawoorden  $\{\mathbf{d}_i\}$ . Immers, dit zijn de enige gegevens die uit een ruwe dump van een NAND-flashgeheugen beschikbaar zijn. Op voorhand is onbekend of er in de collectie  $\{\mathbf{d}_i\}$  datawoorden zijn die bitfouten bevatten. Een voor de hand liggende aanpak van het probleem is, om

een aantal (zeg  $t > k$ ) datawoorden rijgewijs in een matrix te verzamelen, deze matrix vervolgens te vegen en te kijken of je  $k$  lineair onafhankelijke rijen overhoudt en  $t - k$  nulrijen. Is dit het geval, dan is de generatormatrix  $G$  van de gebruikte ECC gevonden en zijn we in principe klaar. Groot probleem hierbij is selectie van datawoorden uit de collectie  $\{\mathbf{d}_i\}$ . Voor deze aanpak is het nodig dat alle geselecteerde datawoorden  $\mathbf{d} = (d_1 \dots d_n)$  geen bitfouten bevatten en dus geldige codewoorden  $\mathbf{c} = (c_1 \dots c_n)$  zijn. Zit er een verkeerd exemplaar tussen, dan zal deze door de bitfouten lineair onafhankelijk van de andere codewoorden geworden zijn, waardoor de

dimensie van de generatormatrix na het vegen te hoog zal zijn. Natuurlijk kun je proberen de selectie zo te doen, dat de kans op keuze van foutloze datawoorden  $\mathbf{d} = (d_1 \dots d_n)$  zo groot mogelijk is (bijvoorbeeld door alleen datawoorden te gebruiken die meerdere malen in de ruwe dump voorkomen), maar de inherente gevoeligheid van deze aanpak voor selectie van foutloze datawoorden  $\mathbf{d} = (d_1 \dots d_n)$  blijft.

Het is daarom beter een methode te ontwikkelen om ECC-parameters te achterhalen, die robuust zijn voor bitfouten en die idealiter zelf in staat zijn vast te stellen welke datawoorden in de voor analyse gebruikte collectie  $\{\mathbf{d}_i\}$  geen bitfouten bevat

### Foutverbeterende codes

Een lineaire code  $C$  wordt volledig bepaald door de generator matrix  $G$ . De rijen van  $G$  vormen een basis voor de deelruimte opgespannen door de codewoorden van  $C$ . Parameters van een lineaire code zijn de lengte  $n$ , de dimensie  $k$  en het aantal fouten  $t$ , dat de code maximaal kan verbeteren. Bij coderen wordt een  $k$ -dimensionaal bericht  $\mathbf{m} = (m_1 \dots m_k)$  afgebeeld op een codewoord  $\mathbf{c} = (c_1 \dots c_n)$  door vermenigvuldiging met de generator matrix  $G$ :

$$\mathbf{c} = \mathbf{m}G.$$

Bij decoderen wordt voor een gegeven datawoord  $\mathbf{d} = (d_1 \dots d_n)$  het 'dichtstbijzijnde' codewoord  $\mathbf{c} = (c_1 \dots c_n)$  uit  $C$  gezocht. Er zijn verschillende noties van dichtstbijzijnd mogelijk, maar een veelgebruikte is de Hamming-afstand, dat wil zeggen het aantal posities waarin  $\mathbf{d} = (d_1 \dots d_n)$  en  $\mathbf{c} = (c_1 \dots c_n)$  van elkaar verschillen. Door het decoderen kunnen fouten die zich eventueel in  $\mathbf{d} = (d_1 \dots d_n)$  bevinden, gedetecteerd en gecorrigeerd worden. Een algemene lineaire code kan op verschillende manieren gedecodeerd worden, bijvoorbeeld door middel van de parity check-matrix  $H$ , die uit de generator matrix  $G$  bepaald kan worden. Voor specifieke codes, zoals bijvoorbeeld BCH- of RS-codes, bestaan efficiëntere decodeermethodes, die gebruik maken van speciale eigenschappen die deze codes hebben.

Een belangrijke klasse lineaire codes betreft de zogenaamde cyclische codes. Cyclische codes hebben de eigenschap dat elke cyclische rotatie van de elementen in een codewoord ook weer een codewoord oplevert. Dus als  $\mathbf{c} = (c_1, c_2, \dots, c_n)$  een codewoord is, geldt dit ook voor  $\mathbf{c}_2 = (c_n, c_1, \dots, c_{n-1})$ .

Bij cyclische codes wordt de informatie in een codewoord gerepresenteerd als de coëfficiënten van een code polynoom:

$$\mathbf{c} = (c_1 \dots c_n) \iff c(x) = c_1 + c_2x + \dots + c_nx^{n-1}.$$

Deze codes hebben een rijke algebraïsche structuur die gebaseerd is op berekeningen in polynoomringen. Naast een generatormatrix  $G$  hebben cyclische codes een zogenaamd generatorpolynoom  $g(x)$ . Alle codewoorden zijn veelvouden van dit generatorpolynoom. Dus als  $c(x)$  een codewoord is van een cyclische code, dan impliceert dit dat

$$c(x) = m(x)g(x)$$

voor een zeker berichtpolynoom  $m(x) = m_1 + m_2x + \dots + m_kx^{k-1}$ . De structuur van cyclische codes maakt efficiëntere decoding mogelijk dan bij gewone lineaire codes. Door het generator polynoom van de code op een speciale manier te kiezen kunnen cyclische codes geconstrueerd worden met mooie eigenschappen, zoals een gegarandeerd minimaal aantal fouten  $t$  dat door de code gecorrigeerd kan worden. Voorbeelden van cyclische codes met speciale eigenschappen zijn BCH- en RS-codes. Voor deze speciale codes bestaan, zoals gezegd, superefficiënte decodeermethodes.

```
f70c6217d67e32f9c5a0501d085129e51f65eec3291721d47d601d867b2d97772859294ead6d0010ac960e
44
04
3573b7c9a7c65bd16ed9552d2fb093a0715244d72a0277d5c6c5e6809a583b22cf7a76b34f7db92c138201
3573b7c9a7c65bd16ed9552d2fb093a0715244d72a0277d5c6c5e6809a583b22cf7a76b34f7db92c138201
3573b7c9a7c65bd16ed9552d2fb093a0715244d72a0277d5c6c5e6809a583b22cf7a76b34f7db92c138201
07
3573b7c9a7c65bd16ed9552d2fb093a0715244d72a0277d5c6c5e6809a583b22cf7a76b34f7db92c138201
3573b7c9a7c65bd16ed9552d2fb093a0715244d72a0277d5c6c5e6809a583b22cf7a76b34f7db92c138201
3573b7c9a7c65bd16ed9552d2fb093a0715244d72a0277d5c6c5e6809a583b22cf7a76b34f7db92c138201
05
03
03
```

**Figuur 3** Voorbeeld van het zoeken naar het generatorpolynoom van een cyclische code door het berekenen van de GCD van paren codewoorden uit een ruwe dump van een NAND-flashgeheugen. Elke regel bevat de GCD van één paar codewoorden, weergegeven als een reeks hexadecimale getallen. De meermalen voorkomende grote waarde blijkt het generatorpolynoom  $g(x)$  van een BCH-code te zijn met parameters  $n = 8560$ ,  $k = 8224$ ,  $t = 24$ . Merk op dat de GCD op de eerste regel afwijkt. Deze GCD is een klein veelvoud van het generatorpolynoom  $g(x)$ . Geanalyseerde data is van dezelfde ruwe dump als getoond in Figuur 2.

ten. Dit is mogelijk door de extra aanname te maken, dat de in het NAND-flashgeheugen gebruikte ECC een cyclische code is. Veelgebruikte codes, zoals BCH-codes en RS-codes zijn cyclische codes. Zoals uitgelegd in het kader op de vorige pagina zijn alle codewoorden  $c = (c_1 \dots c_n)$ , opgevat als polynoom over een eindig lichaam, veelvouden van het generatorpolynoom  $g(x)$  van de cyclische code. Het is mogelijk dit generatorpolynoom  $g(x)$  te bepalen op basis van twee codewoorden, zeg  $c_1(x)$  en  $c_2(x)$ . Noodzakelijkerwijs geldt dan dat:

$$\begin{aligned} c_1(x) &= m_1(x)g(x), \\ c_2(x) &= m_2(x)g(x), \end{aligned}$$

voor zekere originele data  $m_1(x)$  en  $m_2(x)$ , weer opgevat als polynomen over een eindig lichaam. Dit betekent dat  $g(x)$  een grote gemeenschappelijke factor van  $c_1(x)$  en  $c_2(x)$  is. Als we dus de grootste gemene deler (GCD) van  $c_1(x)$  en  $c_2(x)$  uitrekenen, zullen we als antwoord  $g(x)$  vinden, of een klein veelvoud ervan. Dat laatste kan gebeuren als  $m_1(x)$  en  $m_2(x)$  onderling ook een factor gemeen hebben. De GCD van twee codewoordpolynomen kan efficiënt berekend worden met het stokoude algoritme van Euclides van Alexandrië, die het ongeveer 300 v.C. bedacht om de GCD van twee getallen te bepalen.

GCD-berekening kunnen we ook gebruiken om het generatorpolynoom  $g(x)$  te bepalen op basis van een collectie datawoorden  $\{d_i\}$ , waarvan op voorhand niet bekend is of deze bitfouten bevatten. De truc is in dat geval om de GCD te bepalen van al ongeordende paren datawoorden  $(d_i, d_j)$  uit de collectie. Indien beiden geen bitfouten bevatten, levert dit meteen het generatorpolynoom  $g(x)$  op. In geval een of beide wel bitfouten bevatten, zal deze

hierdoor geen veelvoud van  $g(x)$  meer zijn. Daarom zal de GCD van het paar dan een klein polynoom zijn, dat makkelijk te onderscheiden is van de veel grotere  $g(x)$ . Figuur 3 toont een voorbeeld van het zoekproces naar het generatorpolynoom  $g(x)$  van een cyclische code uit de praktijk. Er zijn nog wel een paar technische details van belang om op deze manier naar het generatorpolynoom van een cyclische code te zoeken. Met name betreft dit de wijze waarop de datawoorden tot een mogelijk codewoord herleid worden. Hier kan mogelijk herformatting optreden, waardoor wat extra mogelijkheden geprobeerd moeten worden. Voor het principe van deze zoekmethode is dit niet echt van belang. Geïnteresseerden worden naar [2] verwezen, waar alle details beschreven zijn.

Is eenmaal een grote GCD op basis van datawoorden uit de collectie  $\{d_i\}$  gevonden, dan zal dit met grote waarschijnlijkheid het generatorpolynoom  $g(x)$  van de gebruikte ECC zijn en is het probleem bijna volledig opgelost. Het is nog de moeite waard om na te gaan of  $g(x)$  het generatorpolynoom is van een BCH-code. Dit omdat voor BCH-codes snellere decodeermethoden bestaan dan de methodes die voor een algemene cyclische code gebruikt kunnen worden. Deze stap kan eenvoudig uitgevoerd worden door, gegeven de bekende lengte van de codewoorden, alle BCH-codes op basis van het aantal fouten dat zij kunnen verbeteren te produceren en te kijken of het bijbehorende generator polynoom overeenkomt met de eerder bepaalde GCD. Enige complicatie hierbij is dat er een keuze moet worden gemaakt voor het irreducibele polynoom  $p(x)$  dat gebruikt wordt om het eindige lichaam te genereren waarin de berekeningen worden

uitgevoerd. Tot nu toe bleek in de praktijk het simpelweg proberen van een aantal kandidaten  $p(x)$  met een beperkt aantal termen voldoende om deze hobbel te nemen.

### Tot slot

Dit werk is een mooie illustratie van de kracht van een multidisciplinaire aanpak in (forensisch) onderzoek. Door de beschreven wiskundige technieken is het team Forensische Digitale Technologie van het NFI nu in staat zaakonderzoek succesvol uit te voeren, waar dat voorheen niet mogelijk was. Maar het is bovenal interessant te zien hoe verrassend en onverwacht toepassingen van de wiskunde kunnen zijn: wie had ooit kunnen denken dat een meer dan 2000 jaar oud algoritme, zoals dat van Euclides, ooit emuloor zou vinden bij het ontsluiten van gegevens die zich in een geheugenchip in een moderne tabletcomputer bevindt?  $\square$

### Biografie

Jan Peter van Zandwijk heeft natuurkunde gestudeerd aan de Vrije Universiteit in Amsterdam en is daarna gepromoveerd bij de faculteit Bewegingswetenschappen, ook aan de Vrije Universiteit, op het modelleren van skeletspieren. Na een periode als onderzoeker bij het Ministerie van Defensie is hij nu aan het Nederlands Forensisch Instituut verbonden als forensisch onderzoeker en deskundige in opleiding op het gebied van digitale technologie. Zijn onderzoeksinteressen zijn onder meer forensische data-analyse, coderingstheorie, cryptografie en het modelleren van complexe dynamische systemen, zoals het menselijk lichaam.

### Referenties

- 1 J. L. Massey, Shift-register synthesis and BCH decoding, *IEEE Transactions on Information Theory* 15 (1969), 122–127.
- 2 J. P. van Zandwijk, A mathematical approach to NAND flash-memory descrambling and decoding, *Digital Investigation* 12 (2015), 41–52.

Hans L. Bodlaender

Departement Informatica, Universiteit Utrecht, en,  
 Faculteit Wiskunde en Informatica, Technische Universiteit Eindhoven  
 h.l.bodlaender@uu.nl

## Oratie

# Hoe verbonden is je

In 2014 werd Hans Bodlaender benoemd tot hoogleraar Algorithms and Complexity aan de Universiteit Utrecht. Zijn onderzoek richt zich op complexe netwerken, waarbij de eigenschappen ‘samenhang’ en ‘verbondenheid’ een belangrijke rol spelen. Dit artikel is een verkorte en bewerkte versie van zijn oratie, uitgesproken op 12 oktober 2015.

In dit verhaal wil ik het hebben over algoritmen, netwerken en complexiteit, en de vraag “Hoe verbonden is je netwerk?” loopt daar op verschillende manieren doorheen. Eerst kijken we naar het onderwerp van de netwerken met de bijbehorende wiskundige structuren die *grafen* worden genoemd.

## De bruggen van Königsberg

Het begin van het wetenschappelijk onderzoek aan grafen en netwerken wordt geplaatst rond het jaar 1736 in het Russische stadje Königsberg, het tegenwoordige Kaliningrad. Midden door de stad loopt de rivier de Pregel, en midden in de rivier lagen een aantal eilanden. De eilanden en oevers waren verbonden met een aantal bruggen.

Het verhaal gaat dat de inwoners van Königsberg de volgende vraag aan de wiskundige Euler stelden: is het mogelijk een rondwandeling te maken zodat elke brug precies één keer gepasseerd wordt? Tegenwoordig noemen we zo’n route een Eulertour. Samen met Euler kunnen we snel zien dat voor het achttiende-eeuwse Königsberg zo’n Eulertour niet bestaat: in een Eulertour gaan we even vaak naar een

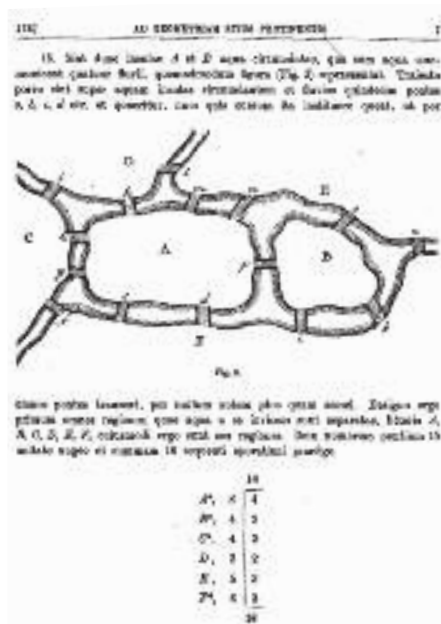
eiland toe als dat we het verlaten, dus een noodzakelijke voorwaarde voor het bestaan van een Eulertour is dat elk eiland (of oever) een even aantal bruggen heeft, en Königsberg heeft een eiland met drie

bruggen. Als we het netwerk als een graaf modelleren, dan betekent dit dat elke knoop even graad moet hebben (een even aantal grenzende kanten).

Er is nog een tweede makkelijk te controleren noodzakelijke voorwaarde voor het bestaan van een Eulertour. In IJsland wil wel eens een vulkaanuitbarsting een vloedgolf in een rivier teweegbrengen, die bruggen wegspoelt. Als nu de rivier het bruggenstelsel in twee delen splitst die onderling niet verbonden zijn en beide gedeeltes hebben minstens één brug, dan zal een wandeling altijd aan één kant van de rivier moeten blijven en er daarom geen Eulertour zijn. Dat levert het tweede criterium op: het netwerk van bruggen moet samenhangend zijn.

Eulers vondst [10] was dat deze twee voorwaarden ook voldoende zijn: een samenhangend netwerk van bruggen waarbij elke locatie een even aantal bruggen heeft, heeft altijd een Eulertour. Eulers oorspronkelijke artikel bevat overigens niet het bewijs van deze stelling; het eerst bekende bewijs is van de Duitse wiskundige Carl Hierholzer uit 1871 [13]. Hierholzers bewijs is constructief: het levert ook een snelle methode op om de route te construeren als deze bestaat.

Een aardig voorbeeld van een graaf die een Eulertour laat zien is een schematische



Figuur 1 Een bladzijde uit Eulers artikel.

# netwerk?



Hans Bodlaender



Figuur 2 Kaart van het Veluwe Zwerfpad.

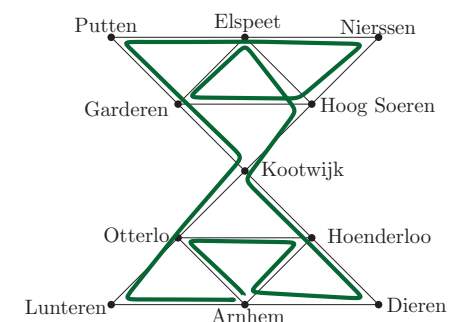
weergave van het Veluwe Zwerfpad [17], zie Figuur 2 en 3. Het Veluwe Zwerfpad is een van de streekpaden in Nederland, een

prachtige wandelroute van totaal 357 kilometer. Als we een paar delen waar trajecten overlappen buiten beschouwing laten,

dan legt de wandelaar tijdens deze 357 kilometer een Eulertour door de Veluwe af. (De onlangs verschenen derde druk heeft een iets gewijzigde route, maar volgt nog steeds een Eulertour.)

De methode van Hierholzers bewijs is als volgt: begin met een rondwandeling die een deel van de kanten eenmaal gebruikt; zolang niet alle kanten gebruikt zijn, kan je ergens beginnen, net zo lang doorgaan tot je weer op dat punt begonnen bent, en die tweede rondwandeling invoegen in de eerste, dit herhalend tot alle kanten gebruikt zijn.

Uit dit bewijs krijgen we een algoritme: een stappenplan dat op een precieze manier voorschrijft hoe met een gegeven invoer de gewenste uitvoer kan worden berekend; in dit geval bepaalt het voor



Figuur 3 Schematische weergave van het Veluwe Zwerfpad met een Eulertour.

een gegeven graaf of deze een Eulertour heeft en zo ja, levert het zo'n tour op. Het woord *algoritme* is een verwijzing naar de Perzische wetenschapper Mohammed ibn Moesa al-Chwarizmi die rond het jaar 800 leefde. Een algoritme kan door een mens worden uitgevoerd, maar liever maken we tegenwoordig een programma zodat een computer voor ons het algoritme uitvoert.

De formalisering van het begrip algoritme vond met name plaats in het midden van de twintigste eeuw, waarbij het werk van Alan Turing voor de algoritmiek van essentieel belang is. De film *The Imitation Game* brengt op een prachtige manier zowel veel aspecten van zijn leven als van zijn wetenschappelijk werk in beeld. Het onderwerp van de *complexiteit* wordt ook op een heel mooie manier in die film verbeeld; hierover later meer.

Het vakgebied van de algoritmiek legt zich toe op het ontwerpen en analyseren van algoritmen, voor allerlei soorten van vragen. Belangrijk is natuurlijk dat een algoritme correct is, en efficiënt. Efficiëntie wordt met name gemeten in de tijd die het algoritme gebruikt, maar kan ook gemeten worden op andere manieren, zoals bijvoorbeeld de hoeveelheid geheugen die het algoritme nodig heeft. We meten dan de tijd die een algoritme gebruikt als het aantal stappen dat het algoritme doet, in het algemeen als functie van het formaat of andere parameters van de invoer. Het Königsberger bruggenprobleem heeft een lineair algoritme: het aantal stappen dat nodig is om vast te stellen of een Eulertour bestaat, en zo ja deze te construeren, is hooguit een aantal keren het aantal knopen plus kanten in de graaf. Terwijl het Eulertourprobleem een efficiënt algoritme heeft, is dit niet het geval voor veel andere problemen. Vaak wordt de grens tussen doenbaar en ondoenbaar gelegd bij wanneer het algoritme tijd gebruikt die begrensd is door een polynoom van de grootte van de invoer, zogenaamde polynomiale algoritmen. Stel dat de inwoners van Königsberg vragen om een wandeling waarbij ze zo veel mogelijk bruggen willen gebruiken, elke brug nog steeds één keer. Dit probleem is NP-volledig — dat wil zeggen, dat het onwaarschijnlijk is dat er een polynomiaal algoritme voor is. (Wat preciezere terminologie: het probleem om een rondwandeling met een maximum aantal bruggen te vinden is *NP-moeilijk*; het is *NP-volledig* om, gegeven een netwerk en

een getal  $K$ , te bepalen of er een rondwandeling met minstens  $K$  bruggen bestaat.) Het probleem is makkelijk op te lossen voor een klein netwerk — voor de bruggen van Königsberg is het makkelijk alle mogelijkheden na te gaan, maar het aantal verschillende mogelijkheden groeit als een exponentiële functie met het formaat van het netwerk. Dit effect heet wel de *combinatorische explosie*.

Die combinatorische explosie wordt mooi geïllustreerd in de film *The Imitation Game*, die ik al eerder noemde. In de Tweede Wereldoorlog geeft Turing leiding aan een team dat de geheime code van de Enigma, een Duitse codeermachine, probeert te kraken. Elke ochtend is er een nieuwe code, waarmee die dag boodschappen van het Duitse leger worden versleuteld. In de film zien we een door Turing en zijn team ontworpen machine, die dit probleem moet oplossen. Het aantal door te rekenen oplossingen is echter zo groot, dat de machine door blijft rekenen zonder iets te antwoorden; en de berekening wordt de volgende ochtend nutteloos als de machine nog rekent op de code van de vorige dag. Een grotere machine helpt niet. Wat wel blijkt te helpen is gebruik te maken van een speciale eigenschap van de invoer: elke ochtend wordt het weerbericht gecodeerd doorgegeven, en door dit gegeven te gebruiken kan het team van Turing de machine de code laten breken.

De verschillende vragen die we over een combinatorische structuur, zoals een netwerk, kunnen stellen, kunnen dus heel verschillende complexiteit hebben: een kleine variatie op het Eulertourprobleem verandert van een snel oplosbare vraag in een vraag waar berekeningen veel tijd kosten. Het vaststellen van de complexiteit van een probleem kan op verschillende manieren: een algoritme voor het probleem kan op wiskundige manier worden geanalyseerd, waarbij vooral (maar zeker niet als enige) gekeken wordt naar de asymptotische looptijd van een algoritme als functie van de invoergrootte; de complexiteitstheorie geeft een heel arsenaal van hulpmiddelen om ondergrenzen te verkrijgen — deze laten ons zien wat voor soort algoritmen waarschijnlijk niet gemaakt kunnen worden.

Ik noemde al het begrip NP-volledigheid. Dit begrip werd in de zeventiger jaren van de twintigste eeuw ontwikkeld

door Cook en Levin [6, 14], en is een van de belangrijkste voorbeelden, misschien wel het belangrijkste voorbeeld van dit soort gereedschap. Daarnaast is de experimentele analyse belangrijk: uitprogrammeren en kijken voor geschikte testinvoeren hoe het algoritme zich in de praktijk gedraagt. Het vakgebied van de theoretische informatica heeft hier niet altijd voldoende oog voor; mijn ervaring is dat de theoretische/wiskundige analyse en het experiment elkaar — in beide richtingen — versterken.

Naast de combinatorische explosie hebben we tegenwoordig ook de *data-explosie*. Zoveel met elkaar verbonden apparaten en mensen produceren een gigantische hoeveelheid gegevens — dan wordt het vinden van goede antwoorden op een vraag over die gegevens ook een hele klus.

### Algoritmen en netwerken

We zagen dat een netwerk van eilanden en bruggen kan worden gemodelleerd als een (eindige) graaf. We gebruiken de term *netwerk* als we extra informatie hebben voor de knopen en kanten in het netwerk, bijvoorbeeld lengtes of geografische locaties.

In veel toepassingen komen we structuren tegen die met grafen of netwerken kunnen worden gemodelleerd. Vanuit de toepassingen komt een veelvoud van vragen over de netwerken — soms willen we inzicht krijgen in de structuur van het netwerk; soms helpt het netwerk een heel andere vraag uit een toepassing te beantwoorden; soms vraagt een planingsprobleem om een oplossing.

### Netwerkmodellen en vragen

Bij veel van die vragen spelen samenhang en verbondenheid van het netwerk een belangrijke rol.

Een van de manieren waarop samenhang in een netwerk wordt gemeten is de minimumsnede: voor twee knopen, hoeveel kanten moet je weghalen zodat de ene knoop niet te bereiken is vanaf de andere? In 1956 bereikten Ford en Fulkerson [7, 12] een klassiek resultaat: je kan dit efficiënt bepalen met behulp van stromingstechnieken. Het resultaat kan je ook zien als een wiskundige illustratie bij het feit dat verkeersopstoppen vaak bij bruggen en tunnels zijn: de maximumhoeveelheid die je van  $A$  naar  $B$  kan vervoeren is gelijk aan de minimumsnede, en zo'n minimumsnede zal vaak door een rivier gevormd worden, want het bouwen van bruggen en

tunnels is veel duurder dan het bouwen van wegen.

Samenhang en verbondenheid in netwerken kom je in veel toepassingen tegen. In een aantal voorbeelden uit het werk van informatici en wiskundigen uit Utrecht en Eindhoven zien we die begrippen een belangrijke rol spelen in onderzoek naar kennisuitwisseling in bedrijven (Remko Helms), belissingsondersteuning met Bayesiaanse netwerken (Linda van der Gaag), frequentietoewijzing en andere vragen rond mobiele (telefoon)netwerken (Mark de Berg, Erik Jan van Leeuwen), optimale planning voor elektriciteitsnetwerken (Marjan van den Akker, Han Hoogeveen), software ecosystemen (Slinger Jansen), en natuurkundig onderzoek naar polymeren vertaalt zich naar het tellen van cycli in een gridnetwerk (Gerard Barkema, Rob Bisseling, Raoul Schram).

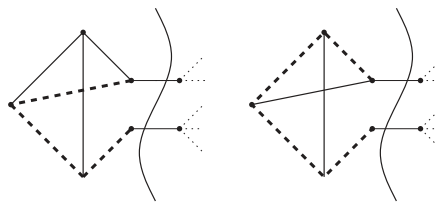
Bij deze en soortgelijke vragen is de uitdaging aan de algoritmicus om te komen met efficiënte algoritmen — “kunnen de vragen binnen redelijke tijd worden beantwoord?” Inzicht in de complexiteit van het probleem helpt daarbij.

Naast het ontwerpen van algoritmen voor concrete problemen willen we ook nieuwe technieken ontwerpen om op die manier de gereedschapskist van de algoritme-ontwerper te vergroten. Ik bespreek nu een paar van zulke technieken.

### Dynamisch programmeren, boomstructuren

Een belangrijke algoritmische techniek is *dynamisch programmeren*. De techniek en term komen uit het werk van Richard Bellman in de veertiger jaren van de vorige eeuw [1]. Het basisidee is om deelproblemen te identificeren, de antwoorden op deze deelproblemen op te slaan en zo te vermijden dat deze antwoorden vaker worden uitgerekend, en op die manier rekentijd te besparen.

Laten we eens opnieuw kijken naar de vraag van een zo lang mogelijke rondwandeling die elke brug hooguit één keer gebruikt. We merken eerst op dat Eulers resultaat ons toestaat om het anders te formuleren: vind een zo groot mogelijk samenhangend deel van de kanten van het netwerk waarbij elke knoop een even aantal kanten ziet. Als we die eis van samenhangendheid niet hadden, zou het probleem vrij makkelijk zijn — het is een variatie op het koppelingsprobleem en kan met Edmonds algoritme uit 1965 [9] wor-



Figuur 4 Twee deelroutes.

den opgelost. Met die samenhangendheid is het, zoals gezegd, NP-moeilijk.

Maar stel nu dat midden door ons landschap een grote rivier stroomt met twee bruggen. We kunnen nu kijken wat de route doet aan de linkerkant van de rivier en aan de rechterkant: het is voldoende om te weten wat de langste rondwandeling is aan de linkerkant, de langste rondwandeling aan de rechterkant, en voor beide kanten van de rivier, het langste pad van de ene brug naar de andere dat aan één kant van de rivier blijft en steeds elke brug maar hooguit één keer gebruikt. In Figuur 4 zien we twee mogelijke routes aan de linkerkant van een rivier: de eerste zal nooit leiden tot een optimale oplossing, want hoe we die ook uitbreiden, de tweede kan op dezelfde manier worden uitgebreid tot een langere route. Die eerste route kan dus vergeten worden. Je zou een tabel kunnen maken van alle essentieel verschillende deelroutes aan de linkeroever, of, beter, voor elke karakterisering van een deelroute aan de linkeroever, de maximale lengte opslaan. Als er twee bruggen over de rivier zijn, zijn er twee waarden om op te slaan; bij een groter aantal bruggen groeit dit aantal, maar we kunnen — als het aantal bruggen niet te groot is — nog steeds ons beperken tot een tabel met veel minder gegevens dan het totaal aantal mogelijke routes aan de linkeroever.

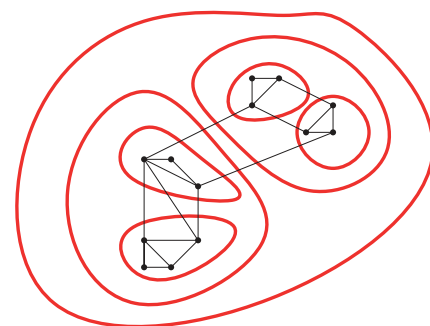
We zien hier een netwerk met een lage samenhangendheid — door twee kanten weg te halen valt het netwerk in meerdere delen uiteen. Dat gebrek aan samenhang kan worden uitgebuit door het probleem te splitsen in kleinere delen. Als het aantal verbindingen tussen de twee delen groter dan twee is, dan stijgt het aantal gevallen dat we moeten bekijken, maar met een klein aantal verbindingen blijft het aantal gevallen beperkt genoeg. Sommige probleeminstanties laten toe dat we de delen weer verder opsplitsen in weer nog kleinere gedeeltes. Op deze manier kan je een geneste structuur van deelnetwerken krijgen, waarbij de essentiële informatie

van een deelnetwerk te bepalen valt uit de essentiële informatie van de direct daar onderliggende delen. Zie Figuur 5.

Dit idee kan op een aantal verschillende manieren vorm krijgen. Een heel succesvolle is de *boomdecompositie* met het daarbij horende begrip *boombreedte* (of, in het Engels, ‘treewidth’). Het begrip generaliseert de bomen, en de series-parallelle netwerken die sommige van ons op de middelbare school hebben gezien (bijvoorbeeld bij het bepalen van de weerstand van een elektrisch netwerk). Een boomdecompositie is een alternatieve representatie van een graaf of netwerk, die ons in staat stelt om het probleem met behulp van dynamisch programmeren op te lossen. Hoe kleiner de boombreedte, hoe sneller het algoritme, hoe minder tijd.

Het gaat te ver om hier de wat technische definitie van boombreedte te geven; de intuïtie is de volgende: in de voorbeelden hierboven splitsten we het netwerk op de kanten — door bruggen over rivieren weg te denken; bij boombreedte doen we iets soortgelijks, maar hakken we knopen door.

In de afgelopen decennia zijn boomdecomposities in veel gebieden met succes toegepast, zowel in de algoritmische grafentheorie, maar ook in veel toepassingsgebieden, waaronder operations research, en ook computationele biologie en logica. Een belangrijk gebied waar ze ook worden gebruikt zijn de probabilistische netwerken, zoals die door Linda van der Gaag en haar groep bestudeerd worden (bijvoorbeeld [15]). Een centraal probleem in deze netwerken is het uitrekenen van een kansverdeling, gegeven een aantal observaties (het *inferentie-probleem*): dit probleem is in zijn algemeenheid moeilijk (NP-moeilijk, en zelfs moeilijk voor de gro-



Figuur 5 Geneste dissectie.

tere complexiteitsklasse  $\#P$ ), maar als de boombreedte van een geassocieerde graaf begrensd is door een constant getal dan is het inferentie-probleem in lineaire tijd op te lossen.

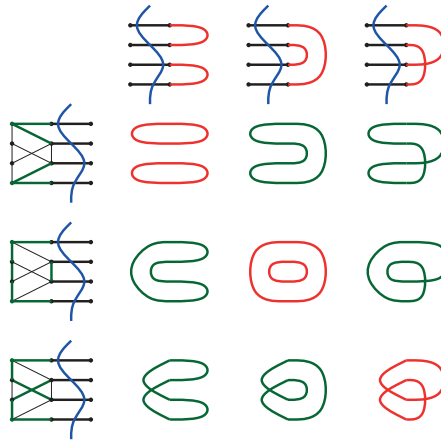
Voor ons probleem van de langste rondwandeling die elke brug hooguit één keer gebruikt, geldt ook dat als de boombreedte van de graaf begrensd is door een constant getal, we het probleem op kunnen lossen in tijd die slechts lineair is in de grootte van het netwerk. Daar staat wel tegenover dat de tijd exponentieel groeit met de boombreedte.

Belangrijke vragen om te beantwoorden zijn: hoe vind je boomdecomposities en hoe kan je algoritmen met behulp van deze structuur zo efficiënt mogelijk maken? Het algoritmisch onderzoek in Utrecht heeft de laatste jaren daar veel aan bijgedragen. Naast efficiënte algoritmen voor het vinden van boomdecomposities en het bepalen van de boombreedte van grafen, het oplossen van allerlei concrete problemen met behulp van deze structuren, hebben we ook een aantal nieuwe technieken ontworpen die dit soort algoritmen versnellen.

Het gaat er bij dynamisch programmeren om om te zorgen dat je geen zaken voor de tweede keer uitrekt. Door goed te kijken naar de structuur van het probleem dat je op wil lossen en naar de structuur van het netwerk, kan je verdere winst boeken. In het werk van Johan van Rooij zien we hoe snellere algoritmen bereikt kunnen worden door representaties van de deelproblemen te transformeren met behulp van *convoluties* [16]. Een prachtige techniek zien we in het werk met Marek Cygan, Jesper Nederlof en Stefan Kratsch [3]. We kijken eens naar een rivier met vier bruggen. Een dynamisch programmeeralgoritme rekent voor de linkeroever van de rivier een tabel uit. De tabel zou kunnen bestaan uit een collectie van deelroutes die zich geheel aan de linkeroever bevinden, met als eigenschap dat er een optimale route in het hele netwerk bestaat die een van de elementen van de tabel als deel bevat.

Als de tabel de volgende drie deeloplossingen bevat, dan kunnen we schematisch kijken naar alle essentieel verschillende manieren waarop die tot een volle tour uitgebreid kunnen worden.

In Figuur 6 ziet u de deeloplossingen, met de uitbreidingen. De uitbreidingen op



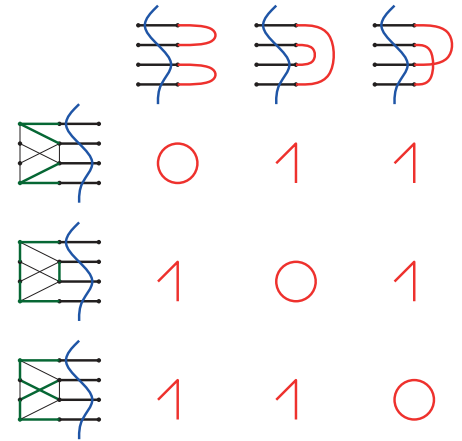
Figuur 6 Deeloplossingen en uitbreidingen.

de hoofddiagonaal geven geen correcte oplossingen, want ze bestaan uit meerdere cycli; de andere zes uitbreidingen geven wel een correcte oplossing. We zien dat de derde deeloplossing niet nodig is: elke uitbreiding van de derde deeloplossing die een tour geeft, werkt ook met de eerste of de tweede. Die derde kan uit de tabel weggelaten worden bij het dynamisch programmeren.

Het aardige is dat een *representatieve verzameling* van deeloplossingen gevonden kan worden door Gauss-eliminatie (modulo twee rekenend) los te laten op een matrix als getoond in het voorbeeld, waarbij we een goede combinatie met 1 en een slechte combinatie met 0 representeren, zie Figuur 7.

Het formaat van de tabel wordt dan begrensd door de *rang* van deze matrix en deze matrix blijkt een rang te hebben die veel kleiner is dan het aantal mogelijke combinaties. We zien een onverwacht verband tussen algoritmen op boomstructuren, en algebra, en voor allerlei problemen, inclusief het langste-cyclusprobleem, krijg je zo een significant sneller algoritme als de boombreedte klein genoeg is [3, 11].

Naast boombreedte en boomdecomposities zijn een aanzienlijk aantal andere soortgelijke structuren en graafparameters voorgesteld en bestudeerd, met hun steeds



Figuur 7 Matrix met binaire waarden voor deeloplossingen en uitbreidingen.

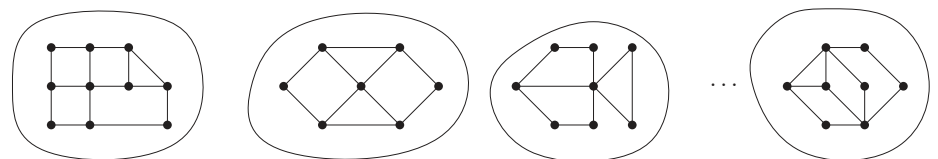
wisselende eigenschappen, toepassingen, mogelijkheden, algoritmische vragen en oplossingen. Een zo'n graafparameter is *gonaliteit*, bestudeerd door Gunther Cornelissen [8], met belangrijke toepassingen in de getaltheorie.

### Vaste-parametercomplexiteit

Het onderzoek aan boomstructuren is een voorbeeld van wat tegenwoordig *fixed parameter tractability* genoemd wordt. Uit de invoer lichten we een parameter, bijvoorbeeld de boombreedte, en kijken hoe snel we het probleem kunnen oplossen als die parameter klein is. Zo wordt de analyse van de complexiteit van een probleem veel diverser: voor verschillende parameters kan het probleem zich anders gedragen.

Soms kan een probleem dat in zijn algemeenheid niet snel genoeg is op te lossen dus wél efficiënt worden opgelost als er een speciale eigenschap van de invoer gebruikt kan worden. Bij het ontcijferen van de codes van de Enigma door Turing en zijn team was dat het vertaalde weerbericht; hier nemen we aan dat één van de parameters van de invoer klein is.

Als de wandelaars uit Königsberg vragen of er een rondwandeling is met minstens  $k$  bruggen, en we weten dat  $k$  een klein getal is, dan heeft het probleem een snel algoritme: we kunnen het oplossen in



Figuur 8 Compositioneleit van het langste-cyclusprobleem.

tijd, die exponentieel is in  $k$ , maar lineair in het aantal knopen van de graaf. In dit geval is de doelwaarde, het aantal bruggen  $k$ , de uitgelichte parameter, en noemen we het probleem *fixed parameter tractable*.

Het vakgebied van de *vaste-parametercomplexiteit* kijkt naar dit soort algoritmen: algoritmen die efficiënt zijn als de bekeken parameter klein is.

Het gaat erom om van problemen met een uitgelichte parameter vast te stellen hoe snel ze oplosbaar zijn; we krijgen daarbij looptijden die een functie zijn van zowel de grootte van de invoer als van de betreffende parameter.

Het gebied van de vaste-parametercomplexiteit kent een ruim aantal technieken, zowel om vast te stellen dat onder complexiteitstheoretische aannames bepaalde algoritmen niet bestaan, als om vaste-parameteralgoritmen te ontwerpen. Eén van die technieken, *kernelisation* ('verkering'), heeft daarbij een heel eigen plek, door de brede toepasbaarheid voor allerlei optimaliseringsproblemen.

### Kernelisation

Als we een moeilijk probleem willen oplossen met een exact, maar wel langzaam algoritme, dan is het vaak zinvol om het probleem voor te bewerken; dat wil zeggen, te kijken of op een relatief eenvoudige manier de probleeminstantie omgevormd kan worden tot een equivalente maar wel kleinere instantie. In ons langste-cyclusprobleem kunnen we bijvoorbeeld alle doodlopende wegen weglaten zonder het antwoord te beïnvloeden.

Zo'n voorbewerkingsalgoritme noemen we een *kernel* als het aan een aantal eisen voldoet: veiligheid, polynomiale tijd, en een grens op het formaat van resulterende instanties die alleen van de parameter afhangt. Een centraal resultaat is dat een beslisbaar geparameteriseerd probleem een kernel heeft, dan en slechts dan als het *fixed parameter tractable* is. Belangrijke vragen zijn hoe snel het *kernelisation*-algoritme is, en welke grens op de formaten er resulteert.

Opnieuw, naast veel resultaten voor concrete problemen zijn er ook een aantal algemene technieken ontwikkeld, waarbij ik hier vanwege de eigen bijdrage de *meta-kernelisation* [5], en ondergrenzen voor kernels [4] wil noemen.

Het centrale idee van die ondergrenzen wil ik hier uitlichten: het is een aardig voor-

beeld van het exploiteren van het gebrek aan samenhangendheid van een netwerk.

Stel we hebben een instantie van het langste-cyclusprobleem, waarbij we een cyclus met  $k$  kanten willen vinden. Maar stel nu dat onze instantie bestaat uit een archipel met heel veel, zeg  $l$ , niet verbonden eilanden, met op elk eilandje een netwerkje, met  $l$  veel groter dan  $k$ .

Er is een cyclus met  $k$  kanten, dan en slechts dan als er tenminste één eiland is met zo'n cyclus. Die verschillende instanties op de eilanden hoeven niets met elkaar te maken te hebben. Het is daarom onwaarschijnlijk dat we snel deze invoer kunnen reduceren tot één met een aantal bits dat minder is dan het aantal eilanden dat we eerst hadden. Met zo'n archipelconstructie hebben we een voorbeeld van *compositionality*: in dit geval laat het zien dat het langste-cyclusprobleem geen kernel met polynomiaal formaat heeft, onder een bepaalde complexiteitstheoretische aanname. Hier gebruiken we het gebrek aan verbondenheid tussen delen van de invoer om te beargumenteren dat bepaalde algoritmische oplossingen niet voorhanden zijn.

Een variatie op het begrip kernel is de polynomiale Turing-kernel; hier vertaal je niet een invoer naar één kleine, equivalente invoer, maar mag je dit een aantal keren doen, op een interactieve manier. Wat preciezer: je hebt een polynomiaal algoritme dat een orakel mag raadplegen dat het antwoord geeft voor instanties met formaat polynomiaal in de parameter. Aardig is dat het langste-cyclusprobleem wél een polynomiale Turing-kernel heeft: in een email-uitwisseling met Bart Jansen zagen we dat dit kan worden verkregen door gebruik te maken van een stelling van Bilinski en medeauteurs van 2011 [2], waaruit volgt dat als we een zogenaamde 3-kant samenhangende component in de graaf hebben met meer dan  $k^{1.33}$  kanten dan heeft die component een cyclus met  $k$  kanten.

We zien hier dat algoritmische technieken hand in hand gaan met inzichten die komen uit de structuur van het netwerk — vaak moet je om een snel algoritme te ontwerpen eerst een graaftheoretische stelling bewijzen, of in de literatuur vinden — en opnieuw, omgekeerd is een algoritmische vraag vaak het startpunt van een studie naar de structuur van grafen.

### Fundamenteel onderzoek in verbinding

Ik heb net twee nieuwe technieken aangestipt: een om snellere algoritmen te vinden, en een om beter de complexiteit van een probleem te kunnen vaststellen. Het vakgebied van de algoritmen en complexiteit heeft een flink arsenaal van zulke technieken, en het ontwikkelen van nieuwe algoritmische technieken of deze op nieuwe manieren toepassen of combineren, geeft belangrijke stappen in de ontwikkeling van het gebied.

Zulke fundamentele resultaten zijn erg het nastreven waard, maar laten die zich plannen...? Nu, het begint altijd met het *stellen van een goede vraag*, en daarna het goed kijken naar (de essentiële elementen in) het antwoord op die vraag.

Verschillende gebieden waaruit zulke goede (algoritmische) vragen kunnen komen zijn: de grafentheorie; de computationele geometrie; de algebraïsche geometrie; toepassingen zoals ik eerder besprak, maar ook softwaretechnologie; planningsproblemen uit de operations research, bijvoorbeeld uit de planning van openbaar vervoer; uit de analyse van complexe systemen; of vanuit de gametechnologie.

Fundamentele vragen liggen bijvoorbeeld op het gebied van de parameterisering van polynomiaal oplosbare problemen op heel grote netwerken: kunnen technieken uit de vaste-parameteralgoritmiek leiden tot nieuwe fundamentele inzichten in de 'big data'?

Een sterke koppeling met het wetenschappelijk onderwijs is iets waar groot belang aan moet worden gehecht. Het terrein van de algoritmiek is uitstekend geschikt om studenten te laten ervaren wat het is om te werken aan de frontlinie van de wetenschap, en daar belangrijke bijdragen te leveren.

In het fundamenteel onderzoek in de algoritmiek wordt het belang van experimentele verificatie van theoretisch verkregen resultaten vaak onderschat. Mijn ervaring is dat experimenteel en fundamenteel onderzoek elkaar kunnen versterken en inspireren.

Onderzoek doe je niet op een spreekwoordelijk eiland: onderzoeksonderwerpen hebben — vaak verrassende — verbanden met elkaar, en de verbanden tussen onderzoekers zijn essentieel om die verbanden tot hun volle recht te laten komen; goede samenhang van het netwerk van onderzoekers doet het onderzoek aan kracht en belang winnen.

## Referenties

- 1 R.E. Bellman en S.E. Dreyfus, *Applied Dynamic Programming*, Princeton University Press, 1962.
- 2 M. Bilinski, B. Jackson, J. Ma en X. Yu. Circumference of 3-connected claw-free graphs and large Eulerian subgraphs of 3-edge-connected graphs, *Journal of Combinatorial Theory, Series B* 101(4) (2011), 214–236.
- 3 H.L. Bodlaender, M. Cygan, S. Kratsch en J. Nederlof, Deterministic single exponential time algorithms for connectivity problems parameterized by treewidth, *Information and Computation* 243 (2015), 86–111.
- 4 H.L. Bodlaender, R.G. Downey, M.R. Fellows en D. Hermelin, On problems without polynomial kernels, *Journal of Computer and System Sciences* 75(8) (2009), 423–434.
- 5 H.L. Bodlaender, F.V. Fomin, D. Lokshtanov, E. Penninkx, S. Saurabh en D.M. Thilikos, (Meta) kernelization, in *Proceedings of the 50th Annual Symposium on Foundations of Computer Science, FOCS 2009*, pp. 629–638. Te verschijnen in *Journal of the ACM*.
- 6 S.A. Cook, The complexity of theorem-proving procedures, in *Proceedings of the 3rd Annual Symposium on Theory of Computing, STOC'71*, ACM, 1971, pp. 151–158.
- 7 T.H. Cormen, C.E. Leiserson, R.L. Rivest en C. Stein, *Introduction to Algorithms, Second Edition*, MIT Press, 2001.
- 8 G. Cornelissen, F. Kato en J. Kool, A combinatorial Li-Yau inequality and rational points on curves, *Mathematische Annalen* 361 (2012), 211–258.
- 9 J. Edmonds, Paths, trees, and flowers, *Canadian Journal of Mathematics* 17 (1965), 445–467.
- 10 L. Euler, Solutio problematis ad geometriam situs pertinentis, *Commentarii Academiae Scientiarum Petropolitanae* 8 (1736), 128–140.
- 11 S. Fafianie, H.L. Bodlaender en J. Nederlof, Speeding up dynamic programming with representative sets – an experimental evaluation of algorithms for steiner tree on tree decompositions, in *Proceedings IPEC 2013*, Lecture Notes in Computer Science, Vol. 8246, Springer, 2013, pp. 321–334.
- 12 L.R. Ford Jr. en D.R. Fulkerson, Maximal flow through a network, *Canadian Journal of Mathematics* 8 (1956), 399–404.
- 13 C. Hierholzer, Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren, *Mathematische Annalen* VI (1873), 30–32.
- 14 L. Levin, Universal search problems (in Russian), *Problems of Information Transmission* 3 (1973), 115–116.
- 15 A. Pastink en L.C. van der Gaag, Multi-classifiers of small treewidth, in *Proceedings ECSQARU 2015*, Lecture Notes in Computer Science, Vol. 9161, Springer, 2015, pp. 199–209.
- 16 *Streekpad 16 – Veluwe Zwerfpad*, Nivon, 2006, 1e druk.
- 17 J.M.M. van Rooij, H.L. Bodlaender en P. Rossmanith, Dynamic programming on tree decompositions using generalised fast subset convolution, in A. Fiat en P. Sanders, eds., *Proceedings of the 17th Annual European Symposium on Algorithms, ESA 2009*, Lecture Notes in Computer Science, Vol. 5757, Springer, 2009, pp. 566–577.

Jan Beuving

cabaretier, Zeist  
janbeuving@gmail.com



Het keerpunt van Rogier Swierstra

# Als jongetje droomde ik van een Nobelprijs

Rogier Swierstra promoveerde in de wiskunde, in de algebraïsche meetkunde. Toch bleef hij niet op de universiteit, maar ging hij rekenen aan pensioengelden bij PGGM in Zeist. “Heel goed zijn in wiskunde is niet de enige manier om geluk te vinden.”

*Wat voor werk doe je nu precies?*

“PGGM is een pensioenuitvoeringsorganisatie. Wij beheren het geld van meerdere pensioenfondsen, onder andere in de zorgsector. In totaal hebben we zo’n 200 miljard euro onder onze hoede. Zelf zit ik bij de afdeling strategie, waarbij het mijn hoofdtak is om na te denken over de allocatie van het vermogen. Dat geld zit deels in aandelen, maar ook in vastgoed, binnen en buiten Nederland, liquide en illiquide. Over de keuzes die daarin gemaakt worden denk ik na en adviseer ik, met name over de risico’s die we ermee lopen.

*Dat zijn belangrijke vragen in deze tijd.*

“Ja, maar de belangrijkere vraag voor of we genoeg pensioen hebben, is of we genoeg premie sparen, omdat we als maatschappij langer leven dan we hadden ingecalculeerd. De rendementen die wij op beleggingen halen, dat is voor PGGM heel interessant, maar voor het pensioen van de mensen is dat niet het allerbelangrijkste. Maar, dat gezegd, het gaat om grote bedragen, dus een beslissing gaat al snel over honderden miljoenen.”

*Is het fijn om zoiets concreets te doen?*

“Ha, vergeleken met mijn collega’s doe ik juist vrij abstract werk. Maar vergeleken met mijn promotieonderzoek is het inderdaad concreet. Ik vind het heel leuk werk. Van alle financiële diensten is pen-



Rogier Swierstra

sioen de meest sociale. We verdelen als maatschappij samen het risico, en goede inrichting van de pensioenen is goed voor iedereen.

Bij mijn promotie heb ik wel eens gezegd: mijn ideale baan zou zijn dat je elke ochtend op kantoor een puzzel zou aantreffen voor die dag. Een puzzel die je aan het denken houdt, waarvan je weet dat jij hem kunt oplossen. Dat lijkt eigenlijk heel erg op mijn baan nu. Het is geen hogere wiskunde, of nieuwe wiskunde, maar wel creatief nadenken over nieuwe toepassingen van de wiskunde. Ik lees nog veel over waar wiskundigen mee bezig zijn.”

*Kun je een voorbeeld geven van nieuwe wiskunde die jij in je werk gebruikt?*

“Dat zijn met name inzichten. Dat hoeft niet heel diep te zijn. Bijvoorbeeld: beleggen door de tijd heen is niet een ergodisch proces. Veel beleggingstheorieën gaan ervan uit dat je het gemiddelde over je ensemble kunt vervangen door het gemiddelde over de tijd. Maar zeker als lange-termijnbelegger kan dat niet, door het effect van rente op rente.”

*Ben jij de enige die dit werk doet bij PGGM?*

“Nee. We hebben een afdeling actuarissen, die met echt heel lange termijn bezig

is. Ik ben de schakel tussen hun plannen en het beleggingsbeleid. Op onze afdeling zijn meer mensen; voornamelijk econometristen. Maar hier en daar tref je een wiskundige.”

*Wat is het verschil met de universitaire wereld?*

“De omgangsvormen zijn hier iets formeler. Je wordt bij de universiteit op een andere manier op je werk afgerekend. Wat ik wel mis, is dat in de universitaire wereld veel mensen dezelfde interesses hebben als jij. Als ik hier iets leuks tegenkom, moet ik zoeken met wie ik dat ga delen.”

*Toch heb je heel bewust de academia vaarwel gezegd.*

“Zeker. Ik ben niet echt een onderzoeker. Het abstractieniveau waarin academische wiskunde plaatsvindt, daar was ik niet zo goed in als ik had gehoopt. Onderzoeken is echt iets anders dan studeren. Problemen waar ik goed in ben, dat zijn vaak kleine problemen. Geen grote nieuwe stellingen of overkoepelende theorieën. Het zien dat er een verband is tussen de vraag die iemand stelt en al een bekend feit, dat heb ik nu veel vaker. En wat je hier ook meemaakt is dat er problemen opgelost worden op een heel niet-wiskundige manier.

Toen ik begon met studeren, wilde ik zo slim zijn als Einstein. Na een jaar leek Gauss me haalbaarder. Daarna werd mijn doel zo goed te worden als mijn begeleider. Maar ook dat heb ik niet gehaald. Mijn belangstelling verschoof ook van het pure wiskundige naar het leven eromheen. Heel goed zijn in wiskunde is niet de enige manier om geluk te vinden. Maar, het moet gezegd: als jongetje wilde ik nog graag een Nobelprijs winnen.”

*Dat is geen alledaagse droom voor een kind!*

“Ha ha, nee. Ik was best ambitieus. Ik moest het maar proberen, vond ik. Dat veranderde al toen ik ging studeren, want toen moest ik veel harder werken om het niveau bij te houden dan op de middelbare school.

Toen ik mijn promotie afrondde, was het wel even heroriënteren. In mijn opleiding en promotie, en dat vind ik achteraf

wel jammer, ben ik niet aangespoord om breder te kijken. Mijn promotie was helemaal gericht op het onderzoek. Als je dat niet kan of wilt, moet je gaan bedenken: wat dan? Het duurt best even voordat je bedenkt wat je allemaal nog meer kunt.”

*Ligt daar een taak voor universiteiten? Of zijn ze juist zo goed in onderzoek omdat hun studie daar helemaal op gericht is?*

“Veel minder dan de helft vindt na zijn of haar promotie echt een plek aan de universiteit. Dus de universiteit heeft wel een zekere verantwoordelijkheid richting degenen die niet tot de uitverkorenen behoren. Hetzelfde had je bijvoorbeeld bij consultancybedrijven. Daar kwamen elk jaar 40 knappe koppen binnen, maar er werd ieder jaar maar één iemand partner. Dus 39 mensen verlieten het bedrijf. Die bedrijven zijn er heel goed over na gaan denken hoe al die uitstromers iets voor hen konden blijven betekenen, als netwerk.”

*Wat zou de universiteit kunnen doen?*

“Tijdens de studie meer nadruk op presenteren, zowel aan wiskundigen als aan niet-wiskundigen, en niet alleen maar op tentamens maken. Het samenwerken met niet-wiskundigen is heel belangrijk. En laten zien waar je allemaal terecht kunt. Ik ben na mijn promotie zelf andere oud-aio's gaan opzoeken, om te zien wat ze deden. Daar heb ik heel veel aan gehad, en het hielp me in de bevestiging dat de wereld echt wel op ons zat te wachten. De universiteit zou zo'n netwerk van alumni moeten hebben.

De selectie naar hoogleraar toe is zo streng, dat je daar alles voor moet geven. Dat betekent ook dat maar weinig hoogleraren weten van een carrière buiten de universiteit. Die mensen weten heel veel van hun vak, maar zijn vaak niet de beste adviseurs voor mensen die geen hoogleraar willen worden.”

*Ik neem aan dat je als jongetje al wel wist dat er geen Nobelprijs in de wiskunde is?*

“Ja, ik ben ook begonnen met natuurkunde. Ik deed een TWIN-studie met wiskunde. Ik kon moeilijk kiezen, en dat stelde de keuze nog even uit. Gaandeweg werd ik gegrepen door de meetkunde. Ik stu-

deerde af op een onderwerp in de K3-opervlakken. Net als elliptische krommen slaan die een soort brug tussen de meetkunde en de algebra. Leuke bijkomstigheid is dat er lekker veel K3-merchandise is, ik had een map met de K3-dames voorop. Helaas was de kleding er alleen in kindermaten.

Zowel mijn afstudeeronderzoek als mijn promotieonderzoek probeerden die brugfunctie uit te breiden naar een hogere dimensie.”

*Mis je de K3-oppervlakken nog wel eens?*

“Nee. Ik doe nu ook iets heel anders. Ik ben overgestapt van de platonische perfectie van de meetkunde, naar de onzekerheid van risico's. Ik sta nu in de klei van de grot van Plato. Je ziet soms iets, maar je weet vaak niet wat het is. Ik ben ook in een interessante tijd ingestapt. In 2008 klapten alle financiële en economische modellen. In zekere zin zijn we nu in een andere economische wereld terechtgekomen. Academisch economen zijn op zoek naar nieuwe modellen.

Er was veel wat ik niet begreep, dus ik ben me erin gaan verdiepen, en zag dat er veel is wat niemand begrijpt. In de economie is ook heel veel spraakverwarring. In de wiskunde ga je er altijd van uit dat iets waar is als iemand het zegt. In de economie ga je ervan uit dat iemand een belang heeft.”

*Economie is dus niet per se overzichtelijker?*

“Nee. Max Planck kon ooit kiezen tussen economie en natuurkunde, en koos toen natuurkunde omdat hij economie te moeilijk vond. Het is heel ingewikkeld om oorzaak en gevolg uit elkaar te houden. De echte wereld is moeilijker dan de wiskundige. Maar, ik kijk met een wiskundige bril naar die economie, dat helpt misschien. Zo heb ik het vermoeden dat er manieren zijn om economische processen te vangen met behulp van de evenwichtstheorieën uit de statistische mechanica. Of in elk geval om die economische processen uit te leggen. Daar houd ik me nu als hobby mee bezig.”

Goede suggesties voor een Nederlandse wiskundige met een keerpunt in zijn of haar carrière zijn welkom via [keerpunt@nieuwarchief.nl](mailto:keerpunt@nieuwarchief.nl).

## Federico Camia

New York University Abu Dhabi, UAE, and,  
Department of Mathematics, VU University Amsterdam  
federico.camia@nyu.edu

### Research Vidi project

# Phase transitions, Euclidean fields and self-similar random fractals

In his Vidi project ‘Phase transitions, Euclidean fields and self-similar random fractals,’ Federico Camia worked on the mathematical theory of phase transitions in two dimensions, focusing in particular on its connections to Euclidean field theory and on the emergence of geometric structures that are both random and fractal. This article will not attempt to reproduce or describe in detail the results of his research; it will rather take inspiration from that research to explore how combining self-similarity and randomness produces interesting objects whose analysis requires a mixture of techniques from different areas of mathematics.

Fractal-like structures are common in nature. They have a property called self-similarity, which means that each portion looks like the whole. Some of the most famous mathematical fractals, such as the Cantor, Julia and Mandelbrot sets or the Sierpinski carpet, are produced by a deterministic process and contain identical, scaled-down copies of themselves. On the contrary, natural fractal-like objects usually look random and are self-similar only in a statistical sense, which means that each portion of the object looks similar but not identical to the whole. Besides appearing naturally in the theory of phase transitions, random fractals have many applications to various fields of mathematics, the natural sciences and economics, and they appear in such diverse contexts as the modeling of financial markets and cosmology.

#### Critical phenomena and scale invariance

A physical system is said to have *correlation length*  $l$  if two portions of the system at distance larger than  $l$  are roughly uncorrelated. The correlation length of a system is typi-

cally determined by the strength of thermal fluctuations, which are random and tend to ‘wash out’ the correlation that would normally exist between separate parts of the system. This concept is central in the study of certain phase transitions that lead to the emergence of self-similar patterns.

*Phase transition* is the term used to denote phenomena such as the melting of ice and the freezing or the evaporation of water. In certain types of phase transitions, the correlation length diverges and a sudden ‘structural’ change takes place as a parameter (e.g., temperature or density) governing the system’s behavior goes through a *critical point*. What is meant by the divergence of the correlation length is a special balance between thermal fluctuations and correlations that makes the system effectively scale-free, or scale-invariant. Scale invariance can thus be considered a new symmetry acquired by the system at the critical point. This symmetry implies that the thermodynamic behavior of critical systems must be described by power laws, since other types of laws

would not be compatible with scale invariance. (In other words, at the critical point, the thermodynamic functions are powers.) It also implies that certain geometric objects that are sometimes useful to describe some aspects of certain systems, such as clusters and interfaces, must have a fractal nature. Examples of such clusters and interfaces will be provided later, through the introduction of models such as percolation and the Ising model. But, in general, their shapes are determined by random fluctuations, so that, at the critical point, they are both random and fractal-like.

The term *critical phenomena* describes the collection of phenomena, such as power law behavior of thermodynamic functions and fractal behavior of certain geometric objects, that take place at the critical point; they are mainly due to the divergence of the correlation length and



**Figure 1** Romanesco broccoli. Fractal-like structures are abundant in nature.

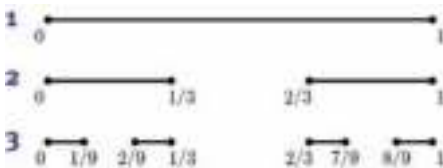
have some universal features that are independent of the specificities of the system or model under consideration. Because of their universality, critical phenomena are well-suited to be analyzed mathematically, since they can be studied using simplified models.

Examples of fractal-like objects are not difficult to find in nature, as evidenced by Figure 1. In mathematics, the prototypical example of a fractal is the triadic Cantor set, depicted in Figure 2. At each iteration, the middle third of each segment is removed according to a scale-invariant construction that looks the same at all scales. For example, looking at Figure 2, what lies below each of the two equal segments produced by removing the middle third of the original segment is a copy of the whole construction scaled down by a factor  $1/3$ . The total length removed in the construction of the triadic Cantor set is

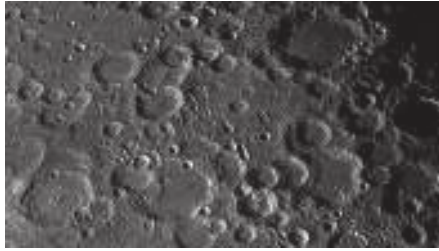
$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \dots + \frac{2^{n-1}}{3^n} + \dots = 1.$$

This tells us that the remaining set has total length zero, but it is by no means empty, since it contains, for example, all points of the type  $1/3^n$ , and (uncountably) many more. What is then the ‘size’ of the triadic Cantor set? To answer the question we need to decide what concept of size we want to use. Length is clearly not appropriate, since it gives a trivial answer that does not reflect the fact that the Cantor set is not empty.

Let’s compare the triadic Cantor set with a ‘full’ segment, which is clearly a one-dimensional object. If we want to cover a segment of length 1 with smaller segments of length, say,  $\ell_k = 1/3^k$ , we need exactly  $N_k = 3^k$  such segments. In the case of the triadic Cantor set instead, we need  $N_k = 2^k$  segments of length  $\ell_k = 1/3^k$ . One way of thinking about the dimension  $D$  of an object is via the relation  $N_k \ell_k^D = 1$ . For the segment this gives the familiar dimension  $D = 1$ , but for the Cantor set it gives  $D = \frac{\log 2}{\log 3} \approx 0.631$ .<sup>1</sup>



**Figure 2** The first few steps in the construction of the triadic Cantor set.



**Figure 3** Benoit B. Mandelbrot noticed that the distribution of the sizes of lunar craters looks approximately scale-invariant [7].

A canonical example of fractality associated with randomness is Brownian motion. Named after the botanist Robert Brown, Brownian motion refers to the jittery motion of particles suspended in a fluid. In 1905, Albert Einstein published a paper explaining how the random motion observed by Brown is a result of the particles colliding with the atoms or molecules in the fluid. Einstein’s explanation provided a confirmation of the atomic theory of matter.

The mathematical model of Brownian motion has numerous real-world applications, one of the best-known being in finance, where it is used to model stock market fluctuations (e.g., in the Black–Scholes model). From a mathematical perspective, Brownian motion is a continuous-time stochastic process with continuous trajectories and with increments that are independent and identically distributed. A salient feature of a Brownian trajectory is its scale invariance, which is responsible for its fractal nature. If one takes a portion of a Brownian trajectory and scales it up to the size of the whole trajectory, the result looks the same as the original, but only *in a statistical sense*. Indeed, because of randomness one cannot expect a closer resemblance. The same type of phenomenon can be found on the surface of the moon, as shown in Figure 3.

### The Ising model and its phase transition

In order to introduce random fractal objects more relevant to the Vidi project discussed in this article, we need to turn to statistical mechanics. Historically, statistical mechanics originates from attempts to use the laws of mechanics to derive the laws and equations of thermodynamics. The latter can be seen as either empirical laws (e.g., the ideal gas law:  $PV = nRT$ ) or axiomatic principles that form the basis of the theory (the three laws of thermodynamics). All of them contain only macro-

scopic variables such as energy, entropy, temperature, volume, and pressure, and make no reference to the microscopic constituents of matter, such as molecules and atoms, and their (quantum-)mechanical interactions. The idea behind statistical mechanics is that one can arrive at the laws of thermodynamics by way of a statistical analysis of the collective behavior of the microscopic constituents of matter. Thus, in statistical mechanics one typically studies systems with a large number of elementary components, and often resorts to a useful mathematical idealization, called the *thermodynamic limit*, in which the number of elementary components is sent to infinity. In doing this, the focus is on the average behavior of the system and on possible deviations from it.

One of the most interesting and challenging topics in statistical mechanics is the study of phase transitions.<sup>2</sup> A *phase* is a state of a thermodynamic system with spatially uniform physical properties, and a *phase transition* is the transformation of a thermodynamic system from one phase to another by heat transfer. A classic example is that of a liquid turning to vapor or vice versa. Another example of a phase transition is what happens to a magnet when warmed to a sufficiently high temperature (depending on the material). If the temperature is high enough, the magnet loses its ability to attract metals; this signals a transition from the *ferromagnetic* to the *paramagnetic* phase.

Let’s now explore this phase transition more in detail using a simplified model based on the observation that the atoms inside a magnet act as little magnets themselves, and that it is energetically advantageous for nearby atoms to have their ‘north poles’ aligned.<sup>3</sup> More precisely, the atoms of a magnet have magnetic moments which naturally tend to align with each other because of magnetic interactions. This tendency is contrasted by random thermal fluctuations which tend to disrupt the ‘magnetic order’. The *Ising model* was introduced in 1920 by physicist Wilhelm Lenz, who asked his student, Ernst Ising, to study its behavior at different temperatures. I will describe the two-dimensional version of the model based on a square grid, where the vertices of the grid represent the positions of the atoms (see Figure 4). The three-dimensional version of the model is analogous to the two-dimension-

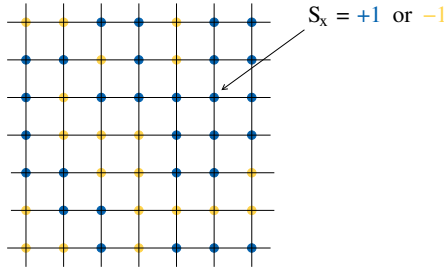


Figure 4 The two-dimensional Ising model.

al one except that it uses a cubic grid, but much less is known about its behavior. The one-dimensional version, defined on subsets of  $\mathbb{Z}$ , is the one studied by Ising, but it does not have a phase transition and is therefore uninteresting for our purposes.

To each vertex  $x$  of a square grid, one associates a *spin* variable  $S_x$  that takes values  $+1$  (represented by a darker color, blue, in Figure 4 and 5) or  $-1$  (represented by a lighter color, yellow). An assignment of a spin  $+1$  or  $-1$  to each vertex is called a *spin configuration* and is denoted by  $\mathbf{S}$ , while  $\mathcal{S}$  will denote the set of all spin configurations. On a finite grid, to a spin configuration  $\mathbf{S} \in \mathcal{S}$ , one assigns probability

$$P(\mathbf{S}) = \frac{1}{Z} \exp \left( \beta \sum_{x, y: |x-y|=1} S_x S_y \right), \quad (1)$$

where  $\beta \geq 0$  represents the inverse of the temperature  $T$  ( $\beta = 0$  corresponds to the idealization of ‘infinite temperature’), the sum is over all pairs of vertices at distance one from each other, and

$$Z = Z(\beta) = \sum_{\mathbf{S} \in \mathcal{S}} \exp \left( \beta \sum_{x, y: |x-y|=1} S_x S_y \right)$$

is a normalizing factor which ensures that (1) defines a probability distribution.

The exponential form of the probability distribution (1) is very important in statistical mechanics, where it is called a *(Boltzmann-)Gibbs distribution*.  $Z$  is called the *partition function* of the model and, despite being introduced here simply as a normalization factor, plays a crucial role in statistical mechanics. The partition function depends in general on the (inverse) temperature and possibly on other quantities, depending on the physical system or the mathematical model under consideration. From this dependence one can derive quantities, such as the (average) energy, that describe the thermodynamic properties of the system.

It’s not difficult to see that  $P$  assigns greater probability to spin configurations where neighboring vertices have the same spin value (or color). Indeed, the two configurations where all spins are  $+1$  or  $-1$  are always the most probable ones. Informally, one could say that vertices prefer to have the same color, or that a vertex has to pay a ‘price’, in terms of probability, for ‘disagreeing’ with its neighbors. But the extent to which disagreement is penalized depends on the value of  $\beta = 1/T$ . For instance, in the idealized case of infinite temperature, where  $\beta = 0$ ,  $P$  is constant and all configurations are equally likely. In such a situation, a ‘typical’ configuration would look very different from those with all spins  $+1$  or all spins  $-1$ , simply because there are only two such ‘orderly’ configurations and many more ‘mixed’ configurations. In this case, what is ‘typical’ is determined uniquely by chance (or entropy, to be more precise). However, if one increases the value of  $\beta$ , the ‘price’ to pay for ‘disagreement’ increases, giving an advantage to more ‘orderly’ configurations.

Thus, the model captures, at least qualitatively, the tendency of the magnetic moments of the single atoms to be aligned with each other, as well as the observation that this tendency is less pronounced at high temperatures. But can the model reproduce the transition between the ferromagnetic phase and the paramagnetic phase, marked by the loss of the ability to attract metals as the temperature becomes high enough?

To answer that question we should look at the total magnetization, the sum of all magnetic moments within a finite region. (A more precise definition of the magnetization will be given in the next section, together with a discussion of some results concerning the magnetization and its important role in the analysis of the Ising model.) If thermal fluctuations win over

the tendency of neighboring atoms to align their magnetic moments, the latter will average out at large scales, leaving zero total magnetization in any large finite region. The presence of a non-zero magnetization, on the contrary, implies long-range order. It is the presence of a non-zero total magnetization that gives ferromagnets the property of attracting metals.

Ising discovered that in one dimension thermal fluctuations dominate at every non-zero temperature. Hence, the one-dimensional Ising model does not have a phase transition. But the one-dimensional case turns out to be special,<sup>4</sup> and the Ising model in any dimension higher than one exhibits a phase transition. This was first demonstrated by Rudolf Peierls, who showed that in two dimensions the total magnetization in any large region averages to zero at sufficiently high temperatures, and takes a non-zero value at sufficiently low temperatures. The same is true in dimensions three and higher. This is illustrated by Figure 5, which depicts the outcomes of three simulations of the model on a large square grid with a small mesh size, so that the figure can fit the page. (The simulations were carried out by Wouter Kager.)

The simulation on the left corresponds to low temperature and shows the ferromagnetic phase in which the spins tend to be aligned with each other. In this phase, the system ‘chooses’ one of the two colors with equal probability, and most of the spins take that color. The high-temperature or paramagnetic phase is shown on the right. The contrast is very clear, with the spin configuration on the right looking much more ‘chaotic’ than that on the left: no order can be seen and there is no preferred color. The reason for this ‘chaos’ is that entropy wins over energy. If one were to choose a spin configuration uniformly at random (i.e., at infinite temperature, or  $\beta = 0$ ),

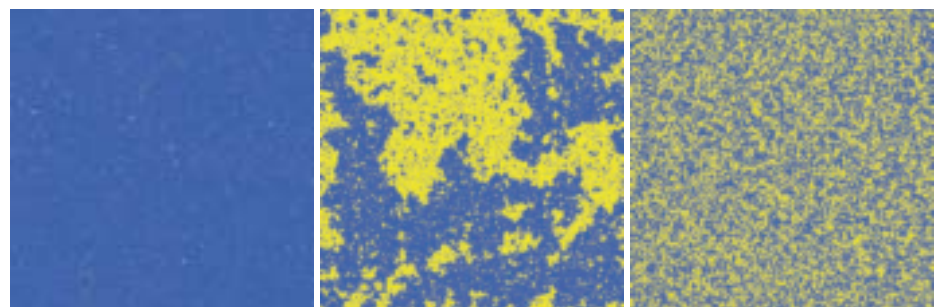


Figure 5 The Ising model phase transition.

with very high probability the configuration would look like the one on the right rather than that on the left, simply because there are many more configurations that look like that. This is still true at high enough temperatures, when  $\beta$  is sufficiently small. But when  $\beta$  becomes large, energy considerations become more important and the system tends to ‘choose’ a more ‘ordered’ configuration. Intuitively, one can think of energy as a currency, and of disordered configurations as being more expensive than ordered ones. At high temperatures the system has a lot of energy and can afford to be in an ‘expensive’ configuration, but at low temperatures it is forced to choose a ‘cheap’ configuration.

We can conclude that the Ising model is successful in reproducing the paramagnetic-ferromagnetic phase transition. This ability has made it one of the most studied models of statistical mechanics. Indeed, the Ising model has played a crucial role in the development of statistical mechanics itself and the theory of phase transitions, and it continues to be a topic of active research.

So far we have focused on the high and low temperature configurations and we haven’t said anything about the configuration in the middle of Figure 5, which looks very different from the other two. If one were to start at high temperature and then gradually lower the temperature, one would see the system move from disordered to more ordered configurations. In doing so, the system would have to make a choice of color, breaking the symmetry between blue and yellow, or  $+1$  and  $-1$ .<sup>5</sup> This very important phenomenon is called *spontaneous symmetry breaking* and plays a crucial role not only in the theory of phase transitions but also in particle physics and cosmology.<sup>6</sup>

The configuration in the middle of Figure 5, in which both colors are roughly equally represented, corresponds to the *critical temperature*, when the phase transition happens and the  $\pm$  symmetry is about to be broken; it represents the watershed between order and disorder, and has very special properties. In fact, I will argue that it bears some resemblance with Figure 3. In that figure, if you take a small part of the whole and blow it up to the size of the original, what you see looks roughly (statistically) similar to the original. The same is true for the middle con-

figuration in Figure 5 because it contains blue and yellow clusters of all possible sizes between the size of the grid’s mesh and that of the whole system, so that a scale transformation will not make a significant difference, as long as the scale factor is small compared to the inverse of the mesh size. The same is not true in the left and right configurations because in both the yellow clusters have a definite mean size, which would be visibly changed by a scale transformation. As a consequence of the ‘democracy of scales’ of the middle configuration, one can see that the interfaces between blue and yellow clusters look fractal-like, particularly around large clusters, due to their ‘peninsulas’ and ‘fjords’. This is precisely what one would expect from the theory of critical phenomena briefly discussed at the beginning of the introduction. But one cannot have full scale invariance and real fractals on a grid: the presence of the grid will eventually be revealed if one blows up the system by a sufficiently large scale factor.

In order to bring the fractal-like aspects of the critical spin configuration to full-blown scale invariance, one can send the mesh size of the grid to zero. This procedure is called a *scaling limit* and is similar to the thermodynamic limit mentioned earlier in that the number of elementary components of the system (in this case, the number of vertices of the grid) tends to infinity. The difference is that the result of a scaling limit is a *continuum model*, not a model defined on an infinite grid. The advantage of this formulation over the thermodynamic limit is that a continuum model can potentially possess more symmetries than a model based on a grid, such as full scale invariance and rotation invariance. In the 1970s–80s, Polyakov observed [9], based on heuristic physics arguments, that a continuum model obtained by taking the scaling limit at the critical point should not only be invariant under translations and rotations, but under all transformations that leave angles unchanged, technically called *conformal transformations*. This observation provides a tremendously powerful tool to study continuum models, and led to the development of *Conformal Field Theory* (CFT). (I will say more about *conformal fields* in the next section.)

Although simulations and physics arguments give a convincing picture of what to expect from a scaling limit at the criti-

cal point, a complete mathematical understanding of scaling limits has proved difficult. The main goal of my Vidi project was to explore the theory of scaling limits at or near the critical point for the two-dimensional Ising model and related models of statistical mechanics, with a new approach that relies on the study of certain random geometric objects such as the clusters and interfaces shown in Figure 5. One particular goal of the research was to provide a rigorous proof of the connection between scaling limits performed at or near the critical point and *Euclidean field theory*, a close relative of the *relativistic quantum field theory* developed by physicists to describe the nature of elementary particles and their interactions.

Despite being one of the most interesting aspects of the theory of scaling limits, and well supported by heuristic physics arguments, until recently there was little mathematically rigorous evidence for this connection beyond some well-understood but relatively trivial examples. One of the main achievements of my Vidi project was to rigorously prove that connection in the case of the two-dimensional Ising model. This is conceptually important because of the crucial role played by the Ising model in the development of statistical mechanics and the theory of phase transitions. In the next and final section, I will describe some of the results about the Ising model obtained during my Vidi project.

## Discussion of some of the main results

This section is going to be somewhat more technical than the previous ones, but I will try to keep the discussion as simple as possible. One of the quantities analyzed in my Vidi project is the Ising *magnetization*, already mentioned in the previous section. Mathematically, the magnetization inside a domain  $D$  is the sum of the spin variables  $S_x$  for all  $x$  in  $D$ . Physically, this corresponds to the total magnetic moment generated by all the atoms in  $D$ . It is an important quantity because its behavior is different in the paramagnetic and ferromagnetic phases, so it can be used to ‘detect’ the phase transition. In physics jargon it is referred to as the *order parameter* associated with the phase transition. The difference in behavior should be present also after taking the scaling limit described above. But what does it mean to take the scaling limit of the magnetization?

To answer that question, let's consider the Ising model on a large (possibly infinite) square grid. We can think of the vertices of the grid as (a subset of)  $a\mathbb{Z}^2$ , where  $a > 0$  is the mesh size of the grid. Consider now a bounded domain  $D \subset \mathbb{R}^2$ ; the *total magnetization* in  $D$  is given by

$$M_D^a = \sum_{x \in a\mathbb{Z}^2 \cap D} S_x,$$

where the sum is over all vertices contained in  $D$ , so that all spins contained in that region, and only those, contribute to  $M_D^a$ . We can now imagine keeping  $D$  fixed and making  $a$  smaller; we will then have more vertices inside  $D$  and more terms in the sum. In order to take the scaling limit, we need to send the mesh size to zero, but it is not difficult to see that as  $a$  tends to zero, the sum defining  $M_D^a$  diverges. The scaling limit of the magnetization  $M_D^a$  is therefore meaningless.

To understand how to overcome this problem, let's first consider the infinite temperature Ising model. In this case, according to (1), all the spin configurations have the same probability. In other words, the spins behave completely independently of each other (we say that their marginal distributions are independent) and each spin  $S_x$  has the same probability (1/2) of being +1 or -1, independently of all other spins. This means that the magnetization  $M_D^a$  is the sum of independent random variables (the spin variables  $S_x$ ), each with the same distribution ( $S_x = 1$  with probability 1/2 and  $S_x = -1$  with probability 1/2 for every  $x$ ). Since all variables are independent and equally distributed, we can apply the Central Limit Theorem, one of the most important and useful theorems of probability theory. Because the number of vertices in  $D$  is of the order of  $(1/a)^2$ , the Central Limit Theorem implies that the random variable  $aM_D^a$  converges to a normal (Gaussian) random variable as  $a \rightarrow 0$ .

Motivated by this result, let's introduce the *renormalized magnetization*

$$m_D^a = a^\alpha \sum_{x \in a\mathbb{Z}^2 \cap D} S_x,$$

where  $\alpha = \alpha(\beta)$  is a parameter that may depend on the temperature. We've already seen that in the infinite-temperature case ( $\beta = 0$ ), the 'correct' choice is  $\alpha(0) = 1$ ; any value smaller than 1 would lead to no limit, while any value larger than 1 would lead, in the limit, to a very uninteresting

random variable identically equal to 0. The fact that we have to choose  $\alpha = 1$  tells us something about the fluctuations of the spins, which in this case are normal (or Gaussian) in the limit, as one would expect.

When  $\beta > 0$ , the spin random variables  $S_x$  are no longer independent, but as long as  $\beta$  is sufficiently small, it turns out that  $\alpha = 1$  is still the correct choice for the renormalized magnetization, and that the scaling limit still gives a normal random variable [1] (see also [8]). This is true for all values of  $\beta < \beta_c = \frac{1}{2} \log(1 + \sqrt{2})$ . All those values correspond to the high-temperature phase depicted on the right in Figure 5. But what happens at the critical value  $\beta_c = \frac{1}{2} \log(1 + \sqrt{2})$ , corresponding to the middle picture in Figure 5?

Looking at the figure suggests that the fluctuations in the system are very different at the critical point than in the high-temperature regime. According to physics wisdom, the behavior of the renormalized magnetization should differ in an interesting way at the critical point with respect to the high-temperature regime. The scaling limit of the magnetization at the critical point is studied in detail in [5] and [3], where it is proved that, in order to obtain a nontrivial limit (i.e., not identically 0), one needs to choose the value  $\alpha(\beta_c) = 15/8$ . With that choice, the renormalized magnetization has a limit as  $a \rightarrow 0$ :

$$m_D^a = a^{15/8} \sum_{x \in a\mathbb{Z}^2 \cap D} S_x \xrightarrow{a \rightarrow 0} m_D.$$

The distribution of the limiting random variable  $m_D$  is still an open question, but it is known [4] that there is a positive constant  $c < \infty$  such that

$$P(m_D > x) \sim e^{-cx^{16}} \text{ as } x \rightarrow \infty.$$

In particular, this shows that  $m_D$  does not have a normal distribution.

It is also shown in [3] that in the scaling limit one can define a sort of magnetization density  $\Phi$  such that  $m_D = \int_D \Phi(z) dz$ . The reader should be warned that in writing  $\Phi(z)$  I'm committing an abuse of notation since  $\Phi$  is not a function and cannot be evaluated pointwise; only its integral over bounded domains or against sufficiently smooth functions of bounded support is well defined. Such objects are called generalized functions or Schwartz distributions (not to be confused with probability distributions).

The magnetization density  $\Phi$  has the following property [3]:

$$\int_{[0, sL]^2} \Phi(z) dz \stackrel{\text{dist.}}{=} s^{15/8} \int_{[0, L]^2} \Phi(z) dz \quad (2)$$

where the equality holds in distribution since both sides of the equation are random variables. If  $\Phi$  were not present, the integrals would compute the areas of the squares  $[0, sL]^2$  and  $[0, L]^2$ , and the relation would obviously be

$$\int_{[0, sL]^2} dz = s^2 \int_{[0, L]^2} dz,$$

where the exponent 2 in the  $s^2$  factor reflects the two-dimensional nature of the area. Based on this simple observation, equation (2) reveals that the total magnetization behaves like a fractal object with dimension  $15/8 = 2 - 1/8$ . The field  $\Phi$ , which represents a magnetization density and should therefore have dimension equal to *magnetization/area*, scales with exponent  $15/8 - 2 = -1/8$ .

A transformation  $x \mapsto sx$  like the one considered above is called a *scale transformation*. It belongs to the family of *conformal transformations*, which includes all transformations that preserve angles, at least in a local, 'infinitesimal' sense. Scale transformations, translations and rotations are all transformations that preserve angles globally. Other conformal transformations can be thought of as suitable combinations of those three basic transformations. In the 1970s–80s, some physicists made the very interesting and somewhat mysterious prediction that, after taking a scaling limit, objects like the Ising magnetization at criticality should scale like fractal objects not only under scale transformations, but under all conformal transformations [2, 9]. Such objects were named *conformal fields*. The prediction of the appearance of conformal fields in the scaling limit was based on heuristic considerations and could not be proved rigorously. Despite the lack of mathematically rigorous examples, the theoretical study of conformal fields, named Conformal Field Theory (CFT), expanded very rapidly and proved extremely interesting and useful, becoming a central topic in theoretical physics, and producing important applications to various other fields, from statistical mechanics to string theory.

Until recently, however, a rigorous mathematical connection between statistical

mechanics and CFT was largely lacking, due to difficulties in the mathematical analysis of scaling limits. The situation started changing at the end of the 1990s, and evolved very rapidly in the first decade of this century. As already mentioned, the scaling limit of the critical Ising magnetization and of the magnetization density  $\Phi$  were first shown to exist in [3]. The proof that the Ising magnetization behaves like a fractal object with dimension  $15/8$  under all conformal transformations was also proved in [3]. This provides a rigorous example of a conformal field obtained from the critical scaling limit of a model of statistical mechanics. It is also particularly pleasing that the model under consideration is perhaps the most studied model of statistical mechanics, and certainly one

that played a crucial role in the development of the theory.

Because of the way they transform under conformal transformations, the critical Ising magnetization and the magnetization density  $\Phi$  are called *conformally covariant*. They are also called *Euclidean fields* because they are invariant under translations and rotations, transformations which leave invariant the Euclidean distance, as opposed to *relativistic fields* which are invariant under the *Lorentz transformations* of relativity theory. Their transformation laws are examples of the power laws discussed at the beginning of this paper. The exponents  $15/8$  and  $1/8$  that appear in those power laws are examples of *critical exponents*.

We started with a general discussion about random fractals and ended with two

specific random fractal objects, the renormalized Ising magnetization and its density, whose existence and properties were proved rigorously almost one hundred years after the Ising model was introduced by Lenz. Those years have witnessed the birth and development of important fields of mathematics deeply connected to the Ising model, such as rigorous statistical mechanics, Euclidean field theory and, more recently, *conformal probability*, the study of conformally-invariant and conformally-covariant random objects.  $\square$

#### Acknowledgments

The author thanks Wouter Kager for Figure 5 and Erik Broman, Arman Mimar and Faisal Saied for useful comments on a draft of the paper.

#### Notes

- 1 The discussion above is not restricted to objects that ‘live’ in a one-dimensional space. One can think, for instance, of covering the unit square with smaller squares of side length  $\ell_k = 1/3^k$ , or the unit cube with smaller cubes with the same side length. The factor  $1/3$  is chosen here for convenience, since the discussion focuses on the triadic Cantor set.
- 2 The 2016 Nobel Prize in Physics was awarded for the study of certain phase transitions in two dimensions [6].

- 3 The real picture is more complex, but the microscopic details are not important in this discussion, due to the universality of critical phenomena mentioned earlier.
- 4 This is somehow related to the fact that the boundary of a segment has the same size (two points) regardless of the size of the segment. In higher dimensions instead, the boundary of square or a (hyper-)cube grows with the size of the square or (hyper-)cube.

- 5 Note that the symmetry is broken at the level of configurations but not at the distributional level since the system chooses each color with equal probability.
- 6 The 2008 Nobel Prize in Physics was awarded to Yoichiro Nambu “for the discovery of the mechanism of spontaneous broken symmetry in subatomic physics” and to Makoto Kobayashi and Toshihide Maskawa “for the discovery of the origin of the broken symmetry which predicts the existence of at least three families of quarks in nature”.

#### References

- 1 G.A. Baker Jr. and S. Krinsky, Renormalization group structure for translationally invariant ferromagnets, *J. Math. Phys.* 18 (1977), 590–607.
- 2 A. Belavin, A.M. Polyakov and A.B. Zamolodchikov, Infinite conformal symmetry in two-dimensional quantum field theory, *Nuclear Physics B* 241 (1984), 333–380.
- 3 F. Camia, C. Garban and C.M. Newman, Planar Ising magnetization field I. Uniqueness of the critical scaling limit, *Ann. Probab.* 43 (2015), 528–571.
- 4 F. Camia, C. Garban and C.M. Newman, Planar Ising magnetization field II. Properties of the critical and near-critical scaling limits, *Ann. Inst. Henri Poincaré Probab. Stat.* 52 (2016), 146–161.
- 5 F. Camia and C.M. Newman, Ising (conformal) fields and cluster area measures, *Proc. Natl. Acad. Sci. USA* 106 (2009), 5547–5463.
- 6 Class for Physics of the Royal Swedish Academy of Sciences, *Topological Phase Transitions and Topological Phases of Matter*, 2016, available at [www.nobelprize.org](http://www.nobelprize.org).
- 7 B.B. Mandelbrot, How long is the coast of Britain? Statistical self-similarity and fractional dimension, *Science* 156 (1967), 636–638.
- 8 C.M. Newman, Normal fluctuations and the FKG inequalities, *Commun. Math. Phys.* 74 (1980), 119–128.
- 9 A.M. Polyakov, Conformal symmetry of critical fluctuations, *JETP Letters* 12 (1970), 381–383.

Piet Groeneboom

*Delft Institute of Applied Mathematics  
Delft University of Technology  
p.groeneboom@tudelft.nl*

Jan van Mill

*KdV Institute for Mathematics  
University of Amsterdam  
j.vanmill@uva.nl*

Aad van der Vaart

*Mathematical Institute  
Leiden University  
avdvaart@math.leidenuniv.nl*

**In Memoriam** Kobus Oosterhoff (1933–2015)

# Statistics as both a purely mathematical activity and an applied science

On 27 May 2015 Kobus Oosterhoff passed away at the age of 82. Kobus was employed at the Mathematisch Centrum in Amsterdam from 1961 to 1969, at the Roman Catholic University of Nijmegen from 1970 to 1974, and then as professor in Mathematical Statistics at the Vrije Universiteit Amsterdam from 1975 until his retirement in 1996. In this obituary Piet Groeneboom, Jan van Mill and Aad van der Vaart look back on his life and work.

Kobus (officially: Jacobus) Oosterhoff was born on 7 May 1933 in Leeuwarden, the capital of the province of Friesland in the north of the Netherlands. He was the eldest of three sons of Wijbo Jacobus Oosterhoff and Gerarda Elisabeth Johanna Oosterhoff-Hoefer Wijehen. The Second World War broke out during his early school days, but Kobus escaped famine and other war miseries, as his father had sent him to his uncle's farm, while himself continuing his practice as a lawyer in Leeuwarden.

After the war Kobus attended the Stedelijk Gymnasium in Leeuwarden (a top level high school with a curriculum that includes ancient Greek and Latin), where he graduated in 1951 in the science ('beta') track. He next started studies in geography at the University of Amsterdam, in line with a passion for traveling and collecting minerals that he maintained throughout his life. However, he switched to mathematics, in which he obtained his 'kandidaatsdiploma' in 1959 and his 'doctoraal-

diploma' (comparable to a masters) in 1963. His favorite lecturer was the topologist J. de Groot, who seems to have deeply influenced Kobus outlook on mathematics. During his student days he married Käthe Hötte, with whom he was to share the rest of his life and to have two sons, Wijbo and Rutger; his eldest was born before his graduation.

## *Career*

From 1961 on Kobus was employed at the Mathematisch Centrum in Amsterdam (now CWI), as an assistant until he obtained his master in 1963, then as a 'medewerker', and from 1967 to 1969 as the deputy-director ('sous-chef') of the department of Mathematical Statistics. The head of this department was J. Hemelrijk, and the Mathematisch Centrum was a thriving place for both applied and mathematical statistics. Kobus was involved in consulting projects and also worked on a PhD thesis. In his thesis Kobus acknowledges the regular

contact with Hemelrijk and the encouragement received from him, but he did his thesis under the direction of Willem van Zwet who, one year younger than Kobus, had been a professor at Leiden University since 1965. Kobus became Willem's first PhD student, defending his dissertation *Combination of One-sided Test Statistics* on 26 June 1969 at Leiden University.

After his graduation Kobus spent a year as a visiting assistant professor at the University of Oregon in Eugene, where he became friends with Don Truax (who was professor of Statistics there) and his family. In 1970 he was appointed as 'lector' at the (then) Roman Catholic University of Nijmegen, and then in 1975 as full professor in Mathematical Statistics at the Vrije Universiteit Amsterdam, a position he held until his retirement in 1996. It is ironic that Kobus switched from a catholic to a protestant university, being not at all religious himself. Nowadays both universities have a general outlook, but this was different in the 1970s. The Vrije Universiteit had to grant Kobus official (and rare) dispensation from endorsing the Christian objectives of the university, and demanded him to write a statement that he would respect the protestant character of the university.



Kobus Oosterhoff in 1998

Kobus always kept to this promise, which was made easy by the fact that his political ideas were in tune with Christian ethics and his colleagues in the mathematics department were liberal, even if in majority adhering to the religious character of the university. On occasions such as a PhD defense, where the chair person had to pronounce a prayer, Kobus would have himself replaced for this part of the ceremony.

The Mathematisch Centrum in Amsterdam remained a meeting place for statisticians in the Netherlands, and Kobus

returned to the Centrum as an advisor in 1971, a position he kept until 1981. In this capacity he became PhD advisor to Richard Gill and Piet Groeneboom, who were appointed at the Centrum and both obtained their PhD in 1979. They were the second and third of eight doctoral students under Kobus guidance, all defending at the Vrije Universiteit. The first student was Wilbert Kallenberg, who had started with Kobus in Nijmegen and defended in Amsterdam

in 1977. Kobus enjoyed working with students, and was an encouraging teacher. Although he left his students free to pursue whatever subject they chose, he read every word of their dissertations, and his preference for exact and clear expression made him want to rewrite every sentence. Later he more than once jokingly claimed that perhaps Richard Gill had taught him survival analysis, but certainly he had taught Richard (a native speaker of course) to write correct English. A typical feedback might start with saying “You make people work so hard. Why don’t you give more explanation!”, but would quickly turn into positive comments. The Mathematisch Centrum also organised courses, usually one week long, for mathematicians, also from Belgium. Kobus was involved in two such courses: one on contiguity and efficiency of rank statistics, and one on efficiency of tests and large deviations (with Piet Groeneboom and Rob Potharst as lecturers). These were important topics of research at the time; the second was the topic of the theses of three of Kobus students.

#### Music

Kobus played the violin in his youth, and was interested in music, both classical music and jazz. Piet knows this because Kobus told him, but his sons do not remember having heard him play. Piet still remembers the following events, which he considers somehow typical for his relationship with Kobus. During the big ISI congress in Amsterdam in 1985 a ‘gathering of friends’ (Kobus, Henry Daniels, Ronald Pyke, Jon Wellner and Piet) was planned at Piet’s house in the town Utrecht, which is about an hour’s drive from Amsterdam. It was a day of extreme rainfall, which made driving difficult. The plan was that Kobus would take some of the friends in his car and Piet the remaining persons, and that they would meet at the first gasoline station on the highway from Amsterdam to Utrecht, after which Kobus would follow Piet to his house in Utrecht. Piet waited for a long time at the first gasoline station, but did not see his PhD supervisor appear. He reluctantly got out of the car into the heavy rain to call Kobus’ wife from the station (pre mobile phone time!), but she also had absolutely no idea where her husband could be. What had happened was that Kobus had taken a second entrance to the highway, missing the first gasoline station.

Anyway, the group gathered at the second gasoline station (close to Utrecht), and everybody finally arrived at Piet's house. Ronald Pyke then said: "Henry and Piet, why don't you play something for us?" And Henry Daniels immediately set himself at the piano, after which it was decided that Händel's famous and beautiful D major sonata for violin and continuo would be played. Henry's favourite instrument was a British invention, the concertina (a kind of accordeon), but he also was an excellent pianist. Piet met Kobus again the next morning and Kobus told him: "I did not exactly have high expectations of how this was going to be with Henry and you, but to my relief it was actually pretty nice, I enjoyed it!"

#### *A modest man*

Kobus suffered from asthmatic bronchitis, which he kept under control by regularly sniffing from a small inhalator. For the rest he was a strong man, and would never complain about health issues. He loved rowing and walking in the mountains. He also loved to cook and enjoyed fine dining. He took his wife and sons at least once a year to a Michelin star(s) restaurant, and when taking out scientific guests for dinner, it was usually Kobus who knew the best restaurants. One everyday memory of Kobus is that he always carried a black purse, in which he kept his money and also the inhalator. At a statistical meeting in Palermo, Sicily, this was snatched away by boys on a passing scooter, but this did nothing to change this habit. One also remembers his interest in Chinese porcelain, especially the soft blue-green qingbai, made during the Song and Yuan dynasties, of which he had a collection in a display cabinet in his bedroom, shown to special house guests.

Kobus was a modest man, probably too modest. He once observed that many hard working individuals end up failing repeatedly in life, through their inability to accept their own shortcomings. It is much easier, Kobus said, to know yourself and take it from there. Kobus was honest with others too, and of the highest integrity. Although he was aware that being honest is not the same as being right, he was sometimes direct in giving his honest opinions. He generously supported the young assistant professors in statistics (first Wilbert Kallenberg and later Aad van der Vaart and Math-

isca de Gunst), enabling them to travel and to develop in teaching and research, and pointing out less obvious details of academic life.

#### *'Algemene Statistiek'*

At the time of his appointment at the Vrije Universiteit in 1975, the department of mathematics there was on its way to finding a balance between theoretical and applied mathematics. There had been recent appointments in applied analysis and numerical mathematics, and computer science was developing. Although he also taught pure mathematical statistics courses, some advanced, Kobus naturally weighed in on the applied side. A large proportion of students would conclude their studies with an internship in industry on some statistical project. A course in applied data analysis, using state-of-the-art computing with the GLIM package was unique in the Netherlands. The lecture notes of Kobus undergraduate course, named 'Algemene Statistiek' to reach out also to non-mathematics students, became the basis for a Dutch language book that appeared in the Epsilon series in 2013. The richness of statistics as both a purely mathematical activity and an applied science was central to Kobus thinking, as is also expressed in his farewell lecture ('Wiskunde of niet', 1996; the title is also a pun on his own initial choice not to study mathematics, perhaps, as he recalls, in response to his high school rector's opinion that mathematics and Greek and Latin have much in common: eminent value and total absence of usefulness).

#### *Administrator*

Soon after his appointment at the Vrije Universiteit it became clear that Kobus had strong administrative skills. He served several times as chairman of the Subfaculteit Wiskunde, and later became dean of the Faculty of Mathematics and Computer Science. He was an effective leader, respected and trusted not only by the mathematicians, but also by the computer scientists. When in 2002 the Division of Mathematics and Computer Science was split into two independent departments, Kobus (emeritus professor by then) drafted the financial plan. Although nobody understood the figures, everybody agreed within days, also the computer scientists. Given the strained relationship between the two parties, mostly caused by financial disputes, this

was one of the best compliments to him ever given.

Kobus was full of ideas, especially on new education plans. He was dean when in 1990 the study program *Bedrijfs-wiskunde en Informatica* (now Business Analytics) was founded at the Vrije Universiteit. It became a great success, although co-founder Gerke Nieuwland once jokingly complained that Kobus together with Bert Kersten had turned his concept of a modern information-technology-based mathematics into mere applied statistics. In later years Kobus was thinking about how mathematics could be integrated with life sciences in order to attract more students, and worked out the financial basis to the successful exchange program with Eastern-European universities.

Kobus retired early at age 62 to make room for the younger generation, after having first put pressure on other professors to do so too. In the years to follow he remained involved and became a source of information for newly appointed and much younger administrators, and also for the board of the newly established Faculty of Science (comprising mathematics, computer science, physics and chemistry). At one time Kobus produced a financial formula that the dean did not understand, but as the 'Oosterhoff formula' became a matter of heated discussion during board meetings of the Faculty of Science.

Kobus was also active as an administrator outside the Vrije Universiteit. He was president of the Netherlands Statistical Society from 1976–1979 and associate editor of the *Annals of Statistics* in the 1980s. During his term as treasurer of the Bernoulli Society in the 1990s he helped establish the journal *Bernoulli*, now among the best journals in statistics. He was the main advisor when the board of the Koninklijk Wiskundig Genootschap took over the Epsilon publishing house. Kobus wisdom and responsibility were highly valued in such administrative matters. Kobus was also involved in statistical education in Indonesia and in setting up a program in 'Bedrijfs-wiskunde en Informatica' in Potchefstroom, South Africa, after the end of apartheid.

Kobus was elected as a member of the International Statistical Institute in 1975 and received the Commemorative Medal of the Faculty of Mathematics and Physics of Charles University, Prague, in 1988.

He truly was a man of many talents.

## Work

Kobus research output was modest, but spans his career and includes significant contributions, mostly to statistical testing theory.

### Statistical decision theory

His first major paper was joint with W.R. van Zwet (see [14]) and was published in the *Annals of Mathematical Statistics*, the premier journal in the area (nowadays split into the *Annals of Statistics* and the *Annals of Probability*). In the reprint of his thesis [10], finished two years later, Kobus was to write “The present study is a continuation of previous work ([14] in our references) by professor W.R. van Zwet and the author”. The paper and the main chapter of the thesis deal with a problem in *statistical decision theory* given a  $k$ -dimensional Gaussian vector  $T$  with mean vector  $\mu$  and unit covariance matrix, test the null hypothesis  $H_0: \mu = 0$  versus the alternative  $H_1: \mu > 0$ , where the inequality on  $\mu = (\mu_1, \dots, \mu_k)$  is understood coordinatewise. There is no ‘best’ solution to this problem, because the most powerful level  $\alpha$  test for testing  $H_0$  versus  $H_1: \mu = \mu_1$ , given by the Neyman–Pearson lemma, depends on  $\mu_1$ . To overcome this Oosterhoff and van Zwet undertook to find a *most stringent* test: one that minimizes the maximum short-coming to the Neyman–Pearson test. If  $\beta^+(\mu_1)$  is the power of the Neyman–Pearson test, this means to find for a given level  $\alpha \in (0, 1)$  the measurable set (critical region)  $K \subset \mathbb{R}^k$  with  $\Pr_{\mu=0}(T \in K) = \alpha$  that minimizes

$$\sup_{\mu > 0} [\beta^+(\mu) - \beta_K(\mu)],$$

$$\beta_K(\mu) = \Pr_{\mu}(T \in K).$$

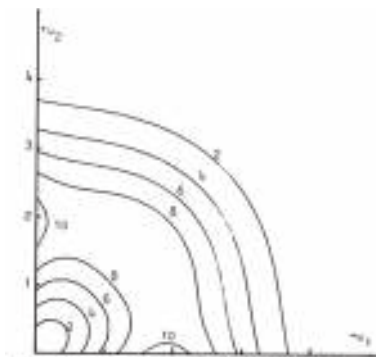


Figure 1 (left) Contour plot of shortcoming as a function of the alternative  $\mu_1 \in \mathbb{R}^2$  of the minimax exponential combination test

It turns out that the solution depends strongly on the value of  $\alpha$ . One highlight in the paper is that for  $\alpha = 0.05$  a region of the form

$$K = \{(t_1, t_2): e^{r(\alpha)t_1} + e^{r(\alpha)t_2} \geq c(\alpha)\}$$

gives a most stringent test. Such an exponential combination of tests was and is unusual. Its optimality results from the fact that it is Bayes relative to a certain prior that is supported at two points of maximal shortcoming. In his dissertation Kobus manages to extend a number of key results to dimension  $k > 2$ , but solves the full problem only up to a conjecture (see p.73), which is presumably still open in 2016. For  $\alpha$  decreasing to zero the least favourable prior will have more support points, and analytical solutions are difficult. Kobus proves that in this case the maximal short-coming of the likelihood ratio test tends to zero, and other tests with the same property are essentially equivalent to this test. Two other chapters are on the combination of  $t$ -tests, and on multinomial tests, and the thesis also contains numerical investigations. The two panels in Figure 1 are reprinted from Figures 2.5.3 and 2.5.5, and Kobus thankfully acknowledges the programmer of the ‘plotter of the EL-X1’. The thesis makes a solid impression and gained Kobus a cum laude PhD.

The Gaussian testing problem arises as a limit when the sample size tends to infinity. Another joint paper with Willem van Zwet (see [12]) is concerned with *contiguity*, a concept introduced by Lucien le Cam, which can be used to make this connection mathematically rigorous. The paper gives necessary and sufficient conditions for connecting contiguity of product measures

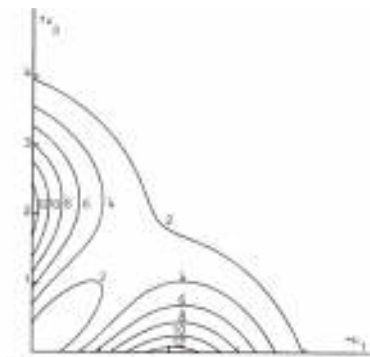


Figure 1 (right) Contour plot of shortcoming as a function of the alternative  $\mu_1 \in \mathbb{R}^2$  of Fisher's combination test

to local asymptotic normality and the Hellinger distance. It is known for its clarity of results and exposition.

### Large deviations and efficiencies

Testing problems with level tending to zero were also the motivation of the theses of Kobus' students Wilbert Kallenberg, Piet Groeneboom and Arnold Kester, on *large deviations* and Bahadur efficiency and deficiency; Kobus published several joint papers with Piet on this subject [6–8]. The *relative efficiency* of a sequence of test statistics  $\{T_n^{(2)}\}$  with respect to another such sequence  $\{T_n^{(1)}\}$  at given  $0 < \alpha < \beta < 1$  and a parameter  $\theta$  from the alternative hypothesis is defined as

$$\text{eff}(T_n^{(2)}, T_n^{(1)}; \alpha, \beta, \theta) = \frac{N^{(1)}(\alpha, \beta, \theta)}{N^{(2)}(\alpha, \beta, \theta)},$$

if  $N^{(j)}(\alpha, \beta, \theta)$  is the minimal sample size such that the test has level  $\alpha$  (probability of falsely rejecting the null hypothesis), and power at  $\theta$  at least  $\beta$  (probability of correctly rejecting the null hypothesis). A large value indicates that  $\{T_n^{(2)}\}$  is superior to  $\{T_n^{(1)}\}$ , since fewer observations are needed to achieve the same power. The relative efficiency is usually difficult to determine exactly, but one can simplify by taking a limit, sending  $\alpha$  to 0,  $\beta$  to 1, or letting  $\theta$  approach the null hypothesis, making the testing problem harder in each case. Letting  $\alpha$  tend to 0, meanwhile keeping  $\beta$  and  $\theta$  fixed, yields the *Bahadur efficiency*, which is directly connected to large deviation theory. Because the alternative is kept fixed, the  $p$ -values  $L_n^{(j)}$  of the tests (defined as the probability of a more extreme value of the test statistic than observed calculated under the assumption that the null hypothesis is true) usually tend to zero exponentially fast in the number of observations, almost surely under the alternative, and  $-(2/n) \log L_n^{(j)}$  tends to a positive limit, called the *Bahadur slope*. Then the Bahadur efficiency is the ratio of the Bahadur slopes: under the probability measure defined by the alternative  $\theta$ , almost surely:

$$\lim_{\alpha \downarrow 0} \text{eff}(T_n^{(2)}, T_n^{(1)}; \alpha, \beta, \theta) \stackrel{\text{a.s.}}{=} \lim_{n \rightarrow \infty} \frac{\log L_n^{(2)}}{\log L_n^{(1)}}.$$

Many test statistics can be written as a functional  $T(\mathbb{P}_n)$  of the empirical distribution  $\mathbb{P}_n$  of the data points. This is the uniform distribution on the data points:

$\mathbb{P}_n(B)$  is the fraction of data points falling in a set  $B$ . For such test statistics calculation of the Bahadur slope comes down to establishing limits of the type or more generally  $\lim_{n \rightarrow \infty} n^{-1} \log \Pr\{\mathbb{P}_n \in \Omega\}$ , for a set  $\Omega$ , with the data points a random sample from the distribution  $P$  determined by the alternative  $\theta$ . *Sanov's theorem* (see [13] which has, however, some difficulties, a version on Polish Spaces can be deduced from [2]) guarantees the existence of such a limit for Polish sample spaces and sets  $\Omega$  such that  $K(\text{int}(\Omega), P) = K(\text{cl}(\Omega), P)$ . Here  $K(Q, P)$  is the infimum over  $Q \in \Omega$  of the Kullback–Leibler divergence  $K(Q, P)$  between the probability measures  $Q$  and  $P$ , and  $\text{int}(\Omega)$  and  $\text{cl}(\Omega)$  the interior and closure of the set  $\Omega$  with respect to the weak topology on the set of all Borel probability measures (which is the topology generated by all maps  $Q \mapsto \int f dQ$ , for bounded continuous functions  $f$ ). The limit is then given by

$$\lim_{n \rightarrow \infty} n^{-1} \log \Pr\{\mathbb{P}_n \in \Omega\} = -K(Q, P).$$

In [8] this was improved to sample spaces that are just Hausdorff and the topology on the set of probability measures generated by all maps  $Q \mapsto Q(B)$ , for Borel sets  $B$ , called  $\tau$  topology, which seems to be the right topology for the Sanov theorem, although it is non-metrizable. Because

this is stronger than the weak topology, the result is obtained for a wider class of sets  $\Omega$ , which is essential to its application to statistical functionals. The  $\tau$  topology has later been used in the more general probabilistic theory on large deviations as well (see, e.g., [1]), but has some peculiar properties. For example, in contrast with the weak topology a sequence of discrete distributions cannot converge to a continuous non-atomic distribution, but a *net* of discrete distributions can.

#### Goodness-of-fit tests

Towards the end of his career Kobus concentrated on *goodness-of-fit tests*. Such tests have the purpose of investigating whether a dataset can be a sample from a given probability distribution. A *chi square statistic* is a certain quadratic form in the vector of counts of the numbers of observations belonging to the sets of a given partition of the sample space. It compares this to the expected number if the null hypothesis were true. Such statistics are popular, but are sensitive to the choice of the partition, which is up to the user. Kobus (with co-authors Wilbert Kallenberg and Bert Schriever, see [9]) studied this in the large sample setting, allowing the size of the partition to tend to infinity when the number of observations increases indefinitely. One finding was that for con-

tamination alternatives the partition size should remain finite if the expectation of the quotient of the null and contamination densities is finite for some  $r > \frac{4}{3}$  and tend to infinity if this norm is infinite for some  $r < \frac{4}{3}$ . This precise criterion contradicts the naive idea that the partitions should always be enriched when more and more observations become available.

Another problem connected to goodness-of-fit testing is that the null hypothesis typically consists of more than one distribution, making it necessary to estimate the expected null counts. Kobus (with co-authors, see [5]) studied this problem for general statistics that compare the empirical distribution of the data to its expected value under the null hypothesis, finding that non-robust estimators of location and scale of the null distribution can improve the power for heavy-tailed alternatives. With the same co-authors Kobus also studied power approximations for a class of tests encompassing the likelihood ratio and chi square tests (see [3, 4]), with a main finding that the usual non-central chi square approximation can be much improved, in particular for the likelihood ratio test.

Kobus published his last paper in 1994 [11], two years before his retirement: a comparison of the merits of trimmed means versus the median. ☞

#### References

- 1 E. Bolthausen, Markov process large deviations in  $\tau$ -topology, *Stochastic Process. Appl.* 25(1) (1987), 95–108.
- 2 M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. III, *Comm. Pure Appl. Math.* 29(4) (1976), 389–461.
- 3 F.C. Drost, W.C.M. Kallenberg, D.S. Moore and J. Oosterhoff, Asymptotic error bounds for power approximations to multinomial tests of fit, in L.J. Gleser et al., eds., *Contributions to Probability and Statistics*, Springer, 1989, pp. 429–446.
- 4 F.C. Drost, W.C.M. Kallenberg, D.S. Moore and J. Oosterhoff, Power approximations to multinomial tests of fit, *J. Amer. Statist. Assoc.* 84(405) (1989), 130–141.
- 5 F.C. Drost, W.C.M. Kallenberg and J. Oosterhoff, The power of EDF tests of fit under nonrobust estimation of nuisance parameters, *Statist. Decisions* 8(2) (1990), 167–182.
- 6 P. Groeneboom and J. Oosterhoff, Bahadur efficiency and probabilities of large deviations, *Statistica Neerlandica* 31(1) (1977), 1–24.
- 7 P. Groeneboom and J. Oosterhoff, Bahadur efficiency and small-sample efficiency, *Internat. Statist. Rev.* 49(2) (1981), 127–141.
- 8 P. Groeneboom, J. Oosterhoff and F. H. Ruymgaart, Large deviation theorems for empirical probability measures, *Ann. Probab.* 7(4) (1979), 553–586.
- 9 W.C.M. Kallenberg, J. Oosterhoff and B.F. Schriever, The number of classes in chi-squared goodness-of-fit tests, *J. Amer. Statist. Assoc.* 80(392) (1985), 959–968.
- 10 J. Oosterhoff, *Combination of One-sided Statistical Tests*, Mathematical Centre Tracts, Vol. 28, Mathematisch Centrum, Amsterdam, 1969.
- 11 J. Oosterhoff, Trimmed mean or sample median? *Statist. Probab. Lett.* 20(5) (1994), 401–409.
- 12 J. Oosterhoff and W.R. van Zwet, A note on contiguity and Hellinger distance, in J. Jurečková, ed., *Contributions to Statistics*, Reidel, 1979, pp. 157–166.
- 13 I.N. Sanov, On the probability of large deviations of random magnitudes, *Mat. Sb. N.S.* 42(84) (1957), 11–44.
- 14 W.R. van Zwet and J. Oosterhoff, On the combination of independent test statistics, *Ann. Math. Statist* 38 (1967), 659–680.

Geoffrey Janssens

Mathematics Department  
Vrije Universiteit Brussel  
geoffjans@vub.ac.be

Research KWG Prize for PhD students

# A glimpse into the asymptotics of polynomial identities

Many daily phenomena can be modelled by algebraic structures such as algebras. But what do the ‘rough shapes’ of these abstract objects look like? Commutativity, associativity, nilpotency,... all these important properties can be expressed in polynomials with non-commutative variables. At the 2016 BeNeLux Mathematical Congress the KWG Prize was awarded to Geoffrey Janssens. In this article he writes about his research on algebras satisfying polynomial identities and how they may be distinguished using asymptotic theory.

There is an abundance of examples of functions that arise in our everyday lives and in nature such as cooking and tasting food, washing and drying clothes, going to left or right while driving, stock exchange, et cetera. Note that in the first three examples the order in which the actions take place matters. In other words these functions do not commute. Algebras can be used to model their behaviour. Another, more advanced, example is momentum and position of subatomic particles in quantum mechanics. By the fundamental equation of quantum mechanics they are related by  $PM - MP = i\hbar$  where  $\hbar$  is Planck’s constant. In this case the algebraic model corresponding to this is the so called Weyl Algebra which is generated by two variables  $x$  and  $y$  and satisfying  $xy - yx = 1$ .

As a starter we will explain more precisely what is an algebra. After that, we shall address such questions as “What is a polynomial identity?”, “What do they tell us?” and “Can we classify all algebras corresponding to a given set of polynomial identities?”

## What is an algebra?

A set  $A$  is called an algebra over  $\mathbb{C}$  if it is a vector space over  $\mathbb{C}$  (i.e., one can add elements in  $A$  and do scalar multiplication with scalars from  $\mathbb{C}$ ), it is a (associative) ring with unit element (i.e., we can not only add but also multiply two elements from  $A$  in a compatible way which is expressed by distribution) and finally  $\alpha ab = a\alpha b = ab\alpha$  for all  $a, b \in A$  and  $\alpha \in \mathbb{C}$ .<sup>1</sup> The latter property simply express that also the both types of multiplication are compatible with each other. The easiest example is  $\mathbb{C}^n$ , the  $n$ -dimensional complex space. A more enlightening example is the set consisting of the  $n \times n$  square matrices  $M_n(\mathbb{C})$  with entries in  $\mathbb{C}$ . Also  $\mathbb{C}[x_1, \dots, x_n]$ , the set of polynomials in commutative variables  $x_1, \dots, x_n$ , is an example.

All the examples are instances of *finitely generated algebras over  $\mathbb{C}$* . This signifies that there exists a finite number of elements  $a_1, \dots, a_n$  such that  $A = \mathbb{C}\langle a_1, \dots, a_n \rangle$  is generated as an algebra over  $\mathbb{C}$  by the elements  $a_1, \dots, a_n$  (i.e., by starting with a finite number of elements  $a_1, \dots, a_n$  and by only mul-

tiplying, adding and taking scalar multiples one recovers all of  $A$ ). In case of  $\mathbb{C}^n$  these elements are typically the unit vectors  $\vec{e}_i$  for  $1 \leq i \leq n$ , and in the case of  $M_n(\mathbb{C})$  an easy example of a basis is provided by the elementary matrices  $E_{ij}$  (having 1 on place  $(i, j)$  and 0 elsewhere). Note that in these examples even more holds, because it is enough to take only linear combinations (i.e., only addition and scalar multiplication) of the elements  $a_1, \dots, a_n$ , called basis elements, to reconstruct all of  $A$ . If such elements with this stronger property exist,  $A$  is said to be *finite-dimensional over  $\mathbb{C}$* .

Now we have some examples, we could go on and try to describe all possible algebras over  $\mathbb{C}$  up to isomorphism (say up to ‘renaming symbols’). However, this is completely hopeless. The goal will rather be to describe the possible ‘rough shapes’, such as commutativity, of algebras. To make this more precise some definitions are needed. Recall that an algebra  $A$  is commutative if  $ab = ba$  for all  $a, b \in A$ . Or put otherwise if and only if  $f(a, b) = 0$  for all  $a, b \in A$  where  $f(x, y) = [x, y] := xy - yx$ . Such a polynomial is called a polynomial identity of  $A$ . For instance  $\mathbb{C}^n$  and  $\mathbb{C}[x_1, \dots, x_n]$  satisfy this commutativity polynomial identity.

## A first step into polynomial identities

Throughout this article,  $X$  will denote an infinite (countable) set of variables, say

$X = \{x_i \mid i \in \mathbb{N}\}$ . Further  $\mathbb{C}\langle X \rangle$  is the set consisting of all non-commutative polynomials in the variables from  $X$ . This is an algebra over  $\mathbb{C}$  for the usual addition and multiplication of polynomials.<sup>2</sup>

**Definition.** A non-zero polynomial  $f(x_1, \dots, x_n) \in \mathbb{C}\langle X \rangle$ , in some non-commutative indeterminates  $x_1, \dots, x_n$ , is called a *polynomial identity of  $A$*  if  $f(a_1, \dots, a_n) = 0$  for any  $(a_1, \dots, a_n) \in A^n$ , notation:  $f \equiv_A 0$ . The set of all polynomial identities of  $A$  is denoted

$$\text{id}(A) = \{f \in \mathbb{C}\langle X \rangle \mid f \equiv_A 0\}.$$

Such polynomials need not exist in general (as is the case for the Weyl algebra<sup>3</sup>). In many cases however, it does, and then  $A$  is called a *PI algebra*. To start, consider  $A = M_2(\mathbb{C})$ , then  $f(x, y, z) = [[x, y]^2, z] \equiv_A 0$ . To see this we need the Cayley–Hamilton theorem, which asserts that a matrix satisfies its own characteristic equation. For a  $2 \times 2$ -matrix  $C$  this equation has the form  $x^2 - \text{Tr}(C)x + \det(C) = 0$  where  $\text{Tr}(-)$  denotes the trace of a matrix. Since  $\text{Tr}([C_1, C_2]) = 0$  for two matrices  $C_1, C_2 \in M_2(\mathbb{C})$  and because  $\det(A)I_2$  is a scalar matrix (in particular it commutes with all the other matrices), we indeed get that  $[[x, y]^2, z]$  is a polynomial identity of  $M_2(\mathbb{C})$ .

More generally,  $A = M_n(\mathbb{C})$  satisfies some polynomial identity, e.g.

$$f_{n^2+1} := \sum_{\sigma \in \text{Sym}_{n^2+1}} \text{sgn}(\sigma) x_{\sigma(1)} \cdots x_{\sigma(n^2+1)},$$

called the standard polynomial.<sup>4</sup> As an important consequence *any finite-dimensional algebra satisfies a polynomial identity*. Indeed the regular representation,

$$\rho: A \rightarrow GL(A): a \mapsto (\rho_a: A \rightarrow A: b \mapsto a \cdot b),$$

embeds any finite-dimensional algebra into matrices of size  $\dim(A)$ . All this is to say that the class of PI-algebras is a large one.

#### Enriched structure of $\text{id}(A)$

So PI theory is a story about algebras and their corresponding set of polynomial identities  $\text{id}(A)$ . This set is actually a (two-sided) ideal of  $\mathbb{C}\langle X \rangle$  (i.e., if  $f, g \in \text{id}(A)$ , then also  $f + g$ ,  $h \cdot f$  and  $f \cdot h$  are in  $\text{id}(A)$  for any  $h \in \mathbb{C}\langle X \rangle$ ). Further it possess one more important property, namely if  $f(x_1, \dots, x_n) \in \text{id}(A)$  and  $g_1, \dots, g_n$

are arbitrary polynomials in  $\mathbb{C}\langle X \rangle$ , then also  $f(g_1, \dots, g_n) \in \text{id}(A)$ . So polynomial identities remain polynomial identities after eventual substitutions. In more sophisticated words,  $\text{id}(A)$  is closed under endomorphisms  $\phi \in \text{End}_{\mathbb{C}}(\mathbb{C}\langle X \rangle)$ . An ideal with this property is called a *T-ideal*. Now it is not hard to check that all T-ideals of  $\mathbb{C}\langle X \rangle$  are actually of this type. In fact, if  $I$  is a T-ideal, it is easily proved that  $\text{id}(\mathbb{C}\langle X \rangle / I) = I$ .<sup>5</sup>

The algebra  $\mathbb{C}\langle X \rangle / \text{id}(A)$  is called a *relatively free algebra*.<sup>6</sup> Our story is one about  $\text{id}(A)$ , but now we see that equivalently, since there is a 1-1 correspondence, it is a story about understanding the different possible relatively free algebras  $\mathbb{C}\langle X \rangle / I$  with  $I$  a T-ideal in  $\mathbb{C}\langle X \rangle$ .

#### Specht's problem and representability

Let  $A$  be a finitely generated algebra over  $\mathbb{C}$ .

**Specht's problem.** Do there exists polynomials  $f_1, \dots, f_l \in \text{id}(A)$  such that  $\text{id}(A) = (f_1, \dots, f_l)_{T\text{-id}}$  is finitely generated as a T-ideal?<sup>a</sup>

This is a variant of Hilbert's basis theorem which gives an affirmative answer for the two-sided ideals of the commutative polynomial ring.

It is important to add 'as a T-ideal', since otherwise the result do not hold, e.g.  $I = (yx^n y \mid n \in \mathbb{N})$  is a non-finitely generated ideal of  $\mathbb{C}\langle x, y \rangle$ .

In his seminal work from 1991, Kemer proved Specht's problem [15]. Actually he proved a stronger statement, called the representability theorem.

**Representability theorem.** Let  $A$  be a finitely generated algebra over a field  $F$ . Then there exists a finite-dimensional algebra  $B$  that is PI-equivalent to  $A$ . Moreover,  $F\langle y_1, \dots, y_m \rangle / \text{id}(A)$  can be embedded in a matrix algebra  $M_n(L)$  where  $L$  is a field extension of  $F$ .

Thus it is not possible to distinguish finitely generated algebras from finite dimensional ones solely using polynomial identities. Also this will enable to assume  $A$  is finite-dimensional.

<sup>a</sup> In the T-ideal also all substitutions into the  $f_i$  are added.

#### The classification problem

We now have all the ingredients to refine our thoughts into the following problem:

"Classify all finitely generated algebras over  $\mathbb{C}$  up to PI equivalence"

where we say that  $A$  is PI-equivalent to  $B$  if  $\text{id}(A) = \text{id}(B)$ .<sup>7</sup> By the explanations in the previous section we could equally ask to classify up to isomorphism all relatively free algebras.

So, in the above question, we do not take into account 'small relations' only relating certain elements of  $A$ . In a way we try to describe the possible rough shapes delivered by polynomial identities to an algebra.

Let us reformulate the problem. First fix some set  $S = \{f_i \in \mathbb{C}\langle X \rangle\}$  of polynomials in non-commutative variables from the set  $X$ . Associated to it we can consider the set

$$\mathcal{V}(S) = \{A \in \text{Alg}_{\mathbb{C}} \mid S \subseteq \text{id}(A)\},$$

called *the variety<sup>8</sup> corresponding to  $S$* , consisting of all algebras (over  $\mathbb{C}$ ) having at least all  $f_i \in S$  as polynomial identities. For example if  $S = \{xy - yx\}$  then  $\mathcal{V}(S)$  is the set of all commutative algebras. Now, in other words, *the goal is to describe  $\mathcal{V}(S)$  for all possible sets  $S$* .

How does one tackle such a problem? One way is to find a full list of invariants distinguishing (and thus determining) all varieties. Unfortunately, this is (yet) completely out of reach. An invariant is a number that one associates to any algebra (and variety) and which does not change under PI-equivalence.<sup>9</sup> For example, the area of a triangle is an invariant with respect to isometries of the Euclidean plane. Also the determinant of a matrix associated to a certain linear map  $\phi: V \rightarrow V$  is invariant under change of basis of  $V$ . Moreover this invariant fits perfectly in our mindset. More precisely, the determinant notices the difference between invertible linear maps ( $\det \neq 0$ ) and non-invertible linear maps ( $\det = 0$ ).

In our setting, with any algebra  $A$  (and variety) we associate a function  $c(A): \mathbb{N} \rightarrow \mathbb{N}$  which will turn out to look asymptotically (i.e., for  $n$  big enough) as the function  $f(n) = qn^t d^n$ . The numbers  $t$  and  $d$  will be the invariants one is looking for. Moreover the numbers  $t$  and  $d$  will be (half)-integers and be connected (in a precise way) to the algebraic structure of  $A$ . Let us be more concrete.

### Asymptotics behind polynomial identities

In order to describe an algebra up to PI-equivalence we have to completely determine  $\text{id}(A)$ . Unfortunately, determining when an arbitrary polynomial is in  $\text{id}(A)$  can be very hard and painful. Luckily, in characteristic 0, it is enough to determine which multilinear polynomials are polynomial identities. A polynomial is called multilinear if the power of each variable occurring is exactly one in each monomial. More formally,

$$P_n(\mathbb{C}) = \text{span}_{\mathbb{C}} \{X_{\sigma(1)} \cdots X_{\sigma(n)} \mid \sigma \in \text{Sym}_n\}$$

is the set of all multilinear polynomials over  $\mathbb{C}$ . Now let  $f \in \text{id}(A)$ . Then it is possible, by a multilinearization process, to replace  $f$  by a set of multilinear polynomials  $g_i$  that are polynomial identities if and only if  $f$  is and such that  $f$  is in the  $T$ -ideal generated by the  $g_i$  (e.g. if  $f(x) = x^2$ , then  $g(x, y) = f(x + y) - f(x) - f(y) = 2xy$  and  $f(x) = \frac{1}{2}g(x, x)$ ). Thus if we know  $\text{id}(A) \cap P_n(\mathbb{C})$  for all  $n$ , then we can also reconstruct whole of  $\text{id}(A)$  and consequently the PI-equivalence class of  $A$ . Without real surprise, the story would have been too short otherwise, for only in very few cases generators for  $\text{id}(A)$  and  $\text{id}(A) \cap P_n(\mathbb{C})$  are known. Even for  $M_n(\mathbb{C})$  with  $n \geq 3$  this is an open problem. Instead one could try to compute only  $\dim_{\mathbb{C}} \text{id}(A) \cap P_n(\mathbb{C})$  for large  $n$ , which a priori is more tractable (but provides less information). Note that  $\dim_{\mathbb{C}} P_n(\mathbb{C}) = |\text{Sym}_n| = n!$ . It turns out that for  $n$  big enough also  $\dim_{\mathbb{C}} \text{id}(A) \cap P_n(\mathbb{C}) \approx n!$  which is asymptotically a wild function.

In the light of all this, much research in the field of asymptotic PI-theory is focused on the sequence  $(c_n(A))_n$ , where

$$c_n(A) = \dim_{\mathbb{C}} \frac{P_n(\mathbb{C})}{P_n(\mathbb{C}) \cap \text{id}(A)},$$

is called the  $n$ -th codimension of the algebra  $A$ . Notice that the function  $c(A): \mathbb{N} \rightarrow \mathbb{N}: n \mapsto c_n(A)$  depends on  $\text{id}(A)$  rather than  $A$ , thus it is constant on PI-equivalence classes and can therefore be used as an invariant. Unfortunately determining  $c_n(A)$  in any point is, even for concrete examples, out of reach. Nevertheless for large  $n$  it becomes tractable. The reason for this is the pioneering result of Regev who proved in 1972 that this function is exponentially bounded [16], i.e.,

$$\exists d \in \mathbb{R}: c_n(A) \leq d^n \text{ for all } n.$$

Historically Regev introduced codimensions and used above exponential bound in order to solve in the affirmative sense an (at that time) sixty years old conjecture asserting that the tensor product of two PI algebras is again PI.

At this stage of the story we can associate to any PI algebra a sequence of numbers  $c_n(A)$  such that  $\limsup_{n \rightarrow \infty} \sqrt[n]{c_n(A)}$  exists. But much more has to come.

### Conjecture of Regev and Amitsur

Concerning the exact asymptotics of the codimension sequence of a finitely generated algebra, Regev conjectured at the end of the seventies the following.

**Conjecture** (Regev, 70's). *There exist numbers  $t \in \frac{1}{2}\mathbb{Z}, d \in \mathbb{N}$  and  $c \in \mathbb{Q}[\sqrt{2\pi}, \sqrt{b}]$  for some  $b \in \mathbb{N}$  such that*

$$c_n(A) \simeq cn^t d^n$$

where  $f \simeq g$  iff  $\lim_{n \rightarrow \infty} \frac{f}{g} = 1$ .

Historically one should mention an even earlier conjecture of Amitsur asserting that the limit  $d = \lim_{n \rightarrow \infty} \sqrt[n]{c_n(A)}$  exists and is an integer. This number, which represents the exponential growth rate of  $c(A): \mathbb{N} \rightarrow \mathbb{N}: n \mapsto c_n(A)$ , is called the *PI-exponent* of  $A$ , denoted  $\exp(A)$ .

This integrality conjecture is very surprising if one thinks of other growth functions in algebra. As an illustration, one aspect of this conjecture is that the codimension sequence would (asymptotically) never be a function between a polynomial and an exponential function such as  $f(n) = e^{\sqrt{n}}$ . This is in contrast to other growth functions such as the word growth in group theory. Also the polynomial growth of the word growth function of a finitely generated algebra (which gives rise to the so called Gelfand–Kirillov dimension) can be any real value greater than 3. Thus the conjectures of Amitsur and Regev are really strong ones.

It was only in 1998, in their breakthrough paper [6], that Giamburno and Zaicev proved that indeed

$$d = \lim_{n \rightarrow \infty} \sqrt[n]{c_n(A)} \in \mathbb{N}.$$

This is an amazing fact, but nevertheless one could wonder whether this number contains any useful information... The answer is yes! They proved the integrality by delivering a surprisingly transparent

and computable algebraic formula relating  $d$  to the Wedderburn–Malcev decomposition of  $A$  (in case  $A$  is finite-dimensional). Philosophically,  $d$  relates how the ‘nice elements’ (i.e., semisimple) and the ‘bad elements’ (i.e., the Jacobson radical  $J(A)$ ) in  $A$  interact with each other.

It is time for two small examples. First suppose  $A$  is abelian and thus  $xy - yx \in \text{id}(A)$ . In this case  $d = 1$ . To prove this we must find a basis of  $\frac{P_n(\mathbb{C})}{P_n(\mathbb{C}) \cap \text{id}(A)}$  as vector space over  $\mathbb{C}$ . In this quotient space, one has  $x_i x_j - x_j x_i = 0$  (and thus  $x_i x_j = x_j x_i$ ) for any variables  $x_i$  and  $x_j$ . Consequently, in this case  $\frac{P_n(\mathbb{C})}{P_n(\mathbb{C}) \cap \text{id}(A)} = \text{span}_{\mathbb{C}} \{x_1 \cdots x_n\}$ .

### Wedderburn–Malcev decomposition

Representation theory aims to represent groups and algebras inside matrix algebras  $M_n(\mathbb{C})$  which we understand well from our first courses in linear algebra. These are, in a ring theoretic sense, simple and direct summands  $M_{n_1}(\mathbb{C}) \oplus \cdots \oplus M_{n_l}(\mathbb{C})$  are called semisimple algebras.<sup>a</sup> Unfortunately in general an algebra is not semisimple. One can collect all ‘bad elements’ due to which  $A$  fails to be semisimple. This set, which actually is an ideal, is called the Jacobson radical, denoted  $J(A)$ .<sup>b</sup>

**Theorem** (Wedderburn–Malcev). *Let  $A$  be a finite-dimensional algebra over  $\mathbb{C}$ . Then*

$$A = B_1 \oplus \cdots \oplus B_l \oplus J(A)$$

where  $B_i$  is a simple subalgebra (say  $B_i \cong M_{n_i}(\mathbb{C})$ ),  $B_1 \oplus \cdots \oplus B_l$  a maximal semisimple subalgebra,  $J(A)$  is nilpotent (i.e., there exists a number  $s \in \mathbb{N}$  such that  $J(A)^s = 0$ ) and  $\oplus$  the direct sum of vector spaces.

Thus one can properly decompose the elements into a set of ‘nice elements’ (the semisimple part) and of ‘bad elements’ (the radical part).

<sup>a</sup> To be more precise, an algebra  $A$  is semisimple if and only if it is a direct sum of minimal left ideals. By a theorem of Wedderburn–Artin semisimple  $\mathbb{C}$ -algebras are isomorphic to a direct sum (of rings)  $M_{n_1}(\mathbb{C}) \oplus \cdots \oplus M_{n_l}(\mathbb{C})$ .

<sup>b</sup> Concretely  $J(A)$  is the intersection of all (left) maximal ideals. Moreover,  $J(A)$  is the smallest ideal  $I$  such that  $A/I$  is semisimple.

Therefore  $c_n(A) = 1$  for any  $n$  and indeed  $d = \lim_{n \rightarrow \infty} \sqrt[n]{c_n(A)} = 1$ . In the case of our other main example  $M_m(\mathbb{C})$ , the exponential growth rate  $d = m^2$  equals the dimension of the algebra.<sup>10</sup>

Next, using topological methods, Berele and Regev proved in 2008 the full conjecture except for the part  $c \in \mathbb{Q}[\sqrt{2\pi}, \sqrt{b}]$ , see [4]. Unfortunately this time no concrete information concerning  $t$  can be extracted from the proof. Therefore it remained as a main open problem in asymptotic PI theory to understand this black box. Finally, in October 2015, joint with Yakov Karasik and Eli Aljadeff, we found a concrete formula for the polynomial growth rate  $t$  [2].

Time has come to go in detail on the concrete information contained in the invariants  $t$  and  $d$ .

### The first invariant $d$

Let  $A$  be a finitely generated algebra over  $\mathbb{C}$ . By the representability theorem we may even assume  $A$  to be finite-dimensional. In particular we can decompose it nicely according to the theorem of Wedderburn–Malcev  $A = B_1 \oplus \dots \oplus B_q \oplus J(A)$ . Then

**Theorem** (Giambruno–Zaicev [6]). *With notations as before,*

$$d = \max \dim \{B_{i_1} \oplus \dots \oplus B_{i_r} \mid B_{i_1}J(A) \cdot \dots \cdot J(A)B_{i_r} \neq 0\}$$

where  $r \geq 1$  and all  $B_{i_j}$  are different simple components.

So, as announced, the number  $d$  is tightly connected to the algebraic structure and the way ‘bad’ and ‘good’ elements interact with each other. With this formula at hand it is clear that, as mentioned earlier, if  $A = M_m(\mathbb{C})$  then  $d = m^2$  (since then  $J(A) = 0$ ). More generally, if  $A$  is the algebra  $UT(d_1, \dots, d_q)$  consisting of upper block triangular matrices of the type

$$\begin{pmatrix} M_{d_1}(\mathbb{C}) & & * \\ 0 & \ddots & \\ \vdots & & \\ 0 & \dots & 0 & M_{d_q}(\mathbb{C}) \end{pmatrix},$$

then  $d = d_1^2 + \dots + d_q^2$ . Finally, if  $A$  is a so-called quiver algebra, then  $d$  tells us what is the largest path in the quiver that does not pass the same vertex twice.

Giambruno and Zaicev, in [7], also handled the case when  $A$  is not a finitely generated algebra (by using a generalization of the representability theorem).

### The second invariant $t$

As before we may decompose  $A = B_1 \oplus \dots \oplus B_q \oplus J(A)$  according to Wedderburn–Malcev’s theorem. Then:

**Theorem** (Aljadeff, Janssens and Karasik [2]). *If  $A$  is a finite-dimensional basic algebra (see definition below), then*

$$t(A) = \frac{d-q}{2} + (s-1)$$

where  $d = \sum_{i=1}^q \dim B_i$  and  $s \in \mathbb{Z}_+$  the smallest integer such that  $J(A)^s = 0$ .

The proof uses, among other things, Kemer Theory (i.e., techniques and objects central in the solution by Kemer of Specht’s problem and his representability theorem) and introduced the so-called basic algebras, which can serve as building blocks for decomposing algebras up to PI-equivalence.

### Basic algebras

With any finite-dimensional algebra, due to the theorem of Wedderburn–Malcev, we can associate two numbers  $d$  and  $s$  where  $d$  and  $s$  are as in the theorem above. The tuple  $\text{Par}(A) = (d, s)$  is called the *parameter* of  $A$ . Basic algebras are minimal models for a certain given tuple  $(d, s)$  of numbers  $d, s \in \mathbb{N}$ . More precisely:

**Definition.** A finite-dimensional algebra  $A$  is called *basic* if  $A$  is not PI equivalent to an algebra  $B = B_1 \times \dots \times B_r$  where  $B_i$  are finite-dimensional algebras such that  $\text{Par}(B_i) < \text{Par}(A)$  for any  $i = 1, \dots, r$ .

These algebras have the advantage to yield geometric and combinatorial translations. I will not go further in detail and rather refer to [3]. Prime examples of basic algebras are matrix algebras  $M_m(\mathbb{C})$  and upper block triangular matrices  $UT(d_1, \dots, d_l)$ . Also, any (finite-dimensional) algebra is PI-equivalent to a direct product of basic subalgebras. Since the polynomial growth rate  $t$  behaves well towards direct products above result gives an interpretation to  $t$  for any finitely generated algebra.

### Classifying varieties

In the next stage of the story, now that we have these invariants containing useful algebraic information on  $A$ , it is time to use them for the problem of classifying varieties (cf. section ‘The Classification problem’).

For this we start by distributing all varieties into layers according to their PI-exponent (see the figure). Let now  $S$  be a fixed set of polynomials, consider  $\mathcal{V}(S)$  and suppose that its exponential growth is  $d$ .<sup>11</sup>

It could be that by adding polynomials to  $S$  we get a strictly smaller variety (since not all algebras in  $\mathcal{V}(S)$  have to satisfy this extra polynomial) with a strictly smaller invariant  $d$ . If this always happens,  $\mathcal{V}(S)$  is called a *minimal variety* (intuitively in this case  $\mathcal{V}(S)$  lies at ‘the bottom of the layer’). They have been classified in [8] and the answer surprisingly turns out to be very elegant. Namely:

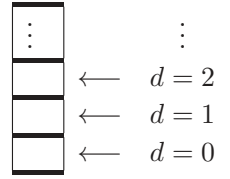
**Theorem** (Giambruno–Zaicev). *Let  $\mathcal{V}$  be a variety with  $\exp(\mathcal{V}) \geq 2$ .<sup>12</sup> Then  $\mathcal{V}$  is a minimal variety if and only if there exists a upper block triangular matrix algebra  $UT(d_1, \dots, d_q) \in \mathcal{V}$  such that  $\mathcal{V} = \mathcal{V}(\text{id}(UT(d_1, \dots, d_q)))$ .*

In a next step one tries to differentiate varieties in a fixed layer (i.e., we fix the invariant  $d$ ). Again we can distribute them into smaller layers depending on the polynomial growth, thus the invariant  $t$ . Investigations are being done into classifying the varieties minimal with respect to the invariant  $t$  with a fixed exponential growth  $d$ . There is evidence that again an elegant answer pops up.

### Other and further research

At this point we have a beautiful story starting with a set of algebras sharing the same ‘rough shape’ (delivered by a common set of polynomial identities) to which we can associate two invariants (which basically are the exponential and polynomial growth rate of the so called codimension sequence) that delivers precise information. Due to these we are able to differentiate certain classes of algebras. But what next?

Two remarks can be made about the story. To begin with, by the representability theorem it is sufficient to consider finite-dimensional algebras. The proof of this theorem, however, is not constructive. Thus an important, and completely open, problem is to find an algorithm whose input is a finitely generated algebra and the output



some finite-dimensional one PI-equivalent to it.

In the same spirit, for the polynomial part, we use a decomposition into basic algebras. Again it would be interesting to have a more constructive proof of this. Actually more examples of basic algebras would already be welcome.

The ultimate goal remains to classify all algebras up to PI-equivalence. In full generality this problem seems to be completely out of reach. However, as Steve Jobs said: “Stay hungry stay foolish!” Towards a full-

er answer researchers have (successfully) started to take more refined information, such as certain (algebraic) group or Hopf algebra actions or (semi)group gradings, into account. Some of the most recent results can be found in [1, 10, 11, 14] and the references therein. First things first, a good starting reference to learn asymptotic PI-theory is [9].

Going beyond the above associative setting, we note that all definitions make sense for non-associative algebras, such as Lie algebras. For a survey of the progress

on the analogous conjectures in this setting we refer to [5, 12, 13, 17] and the references therein. Last but not least, to find methods for finding generators of  $\text{id}(A)$  is one of the main open challenges.

As a brief summary, in spite of many nice results and joyful partial answers concerning the classification problem, a long and interesting road towards describing all varieties (and thus all algebras up to PI-equivalence) is ahead of us. But there is nothing as nice as a walk on a sunny day! ☺

## Notes

- 1 Actually everything holds over an arbitrary field of characteristic 0. But for sake of clarity we simply consider  $\mathbb{C}$ .
- 2 In other words,  $\mathbb{C}\langle X \rangle$  is the free algebra over  $\mathbb{C}$  generated by the elements  $x \in X$ .
- 3 If it would satisfy a polynomial identity then the Weyl algebra would have finite dimensional simple representations. However, the Weyl algebra has no finite dimensional representations. The latter can be seen by taking the trace of the equation  $xy - yx = 1$ .
- 4 This roughly follows from the following two observations: on the one hand,  $f_{n^2+1}$  is multilinear, so it is sufficient to substitute the basis elements of  $M_n(\mathbb{C})$ , and there are  $n^2$  such elements. On the other hand, the polynomial is alternating, so if we substitute two times the same element the polynomial vanishes.
- 5 By  $\mathbb{C}\langle X \rangle/I$  is meant the quotient  $\mathbb{C}\langle X \rangle$  by the ideal  $I$ . Intuitively this means that all elements in  $I$  are made equal to zero in  $\mathbb{C}\langle X \rangle$ .
- 6 The name is not a coincidence. Actually  $\mathbb{C}\langle X \rangle/\text{id}(A)$  is the free object in the category consisting of the algebras  $B$  with  $\text{id}(A) \subseteq \text{id}(B)$ .
- 7 PI-equivalence is really an equivalence relation.
- 8 This is a variety in the sense of Birkhoff. Also note that these varieties encompass the equivalence classes of the PI-equivalence relation. Indeed  $A \sim_{PI} B$  iff  $\mathcal{V}(\text{id}(A)) = \mathcal{V}(\text{id}(B))$  where  $\mathcal{V}(\text{id}(A)) = \{C \in \text{Alg}_{\mathbb{C}} \mid \text{id}(A) \subseteq \text{id}(C)\}$ .
- 9 Thus some number that is constant on each equivalence class.
- 10 Even more holds. Recall that an algebra  $A$  is simple if and only if the only two-sided ideals are  $\{0\}$  and  $A$ . By a theorem of Wedderburn such  $\mathbb{C}$ -algebras are isomorphic to some  $M_n(\mathbb{C})$ . It is known that  $A$  is simple if and only if  $\exp(A) = \dim(A)$ . Thus the PI-exponent detects simple algebras out of a set of algebras.
- 11 Note that by the affirmative answer on Specht's problem we may assume  $S$  to be a finite set. If  $\mathcal{V}$  is some variety then  $\text{id}(\mathcal{V})$  is defined as  $\bigcap_{A \in \mathcal{V}} \text{id}(A)$ . Thus in our case  $\text{id}(\mathcal{V}(S)) = (S)_{T-\text{id}}$  is the  $T$ -ideal generated by  $S$ . Using this definition it makes sense to look at  $c_n(\mathcal{V}) = \dim_{\mathbb{C}} \frac{P_n(\mathbb{C})}{P_n(\mathbb{C}) \cap \text{id}(\mathcal{V})}$  and its exponential and polynomial growth rates.
- 12 Note that  $\exp(\mathcal{V}) = 1$  is the same as saying that  $\mathcal{V}$  has polynomial growth.

## References

- 1 E. Aljadeff, A. Giambruno and D. La Mattina, Graded polynomial identities and exponential growth, *J. reine angew. Math.* 650 (2011), 83–100.
- 2 E. Aljadeff, G. Janssens and Y. Karasik, The polynomial part of the codimension growth of affine PI algebras, *Adv. Math.* 309 (2017), 487–511.
- 3 E. Aljadeff, A. Kanel-Belov and Y. Karasik, Kemer's theorem for affine PI algebras over a field of characteristic zero, arXiv:1502.04298 (2015), 41 pp.
- 4 A. Berele and A. Regev, Asymptotic behaviour of codimensions of p.i. algebras satisfying Capelli identities, *Trans. Amer. Math. Soc.* 360(10) (2008), 5155–5172.
- 5 O.A. Bogdanchuk, S.P. Mishchenko and A.B. Verevkin, On Lie algebras with exponential growth of the codimensions, *Serdica Math. J.* 40 (2014), 209–240.
- 6 A. Giambruno and M.V. Zaicev, On codimension growth of finitely generated associative algebras, *Adv. Math.* 140(2) (1998), 145–155.
- 7 A. Giambruno and M.V. Zaicev, Exponential codimension growth of PI algebras: an exact estimate, *Adv. Math.* 140(2) (1998), 221–243.
- 8 A. Giambruno and M.V. Zaicev, Minimal varieties of algebras of exponential growth, *Adv. Math.* 174 (2003), 310–323.
- 9 A. Giambruno and M.V. Zaicev, *Polynomial Identities and Asymptotic Methods*, AMS Mathematical Surveys and Monographs 122, American Mathematical Society, 2005.
- 10 A. Giambruno and D. La Mattina, Graded polynomial identities and codimensions: computing the exponential growth, *Adv. Math.* 225 (2010), 859–881.
- 11 A. Gordienko, G. Janssens and E. Jespers, Semigroup graded algebras and graded PI-exponent, to appear in *Israel J. Math.* (2017), 41 pp.
- 12 A. Gordienko and G. Janssens,  $\mathbb{Z}S_n$ -modules and polynomial identities with integer coefficients, *Int. J. Algebra Comput.* 23(8) (2013), 1925–1943.
- 13 A. Gordienko, Amitsur's conjecture for polynomial  $H$ -identities of  $H$ -module Lie algebras, *Trans. Amer. Math. Soc.* 367(1) (2015), 313–354.
- 14 A. Gordienko, On codimension growth of  $H$ -identities in algebras with a not necessarily  $H$ -invariant radical, arXiv:1505.02893 (2015), 11 pp.
- 15 A.R. Kemer, *Ideals of Identities of Associative Algebras*, Translations of Mathematical Monographs 87, American Mathematical Society, 1991.
- 16 A. Regev, Existence of identities in  $A \otimes B$ , *Israel J. Math.* 11 (1972), 131–152.
- 17 M.V. Zaicev, Integrality of exponents of growth of identities of finite-dimensional Lie algebras, *Izv. Math.* 66 (2002), 463–487.

# 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](mailto:verdediging@nieuwarchief.nl)



## **Images of Numeracy – Investigating the effects of visual representations of problem situations in contextual mathematical problem solving**

Kees Hoogland

Op 8 november 2016 promoveerde Kees Hoogland aan de Technische Universiteit Eindhoven bij prof.dr. Birgit Pepin en dr. Arthur Bakker (UU). De titel van zijn proefschrift, *Images of Numeracy – Investigating the effects of visual representations of problem situations in contextual mathematical problem solving*, geeft al aan dat zijn onderzoek geen *hardcore* wiskunde betreft, maar veeleer de didactiek van de wiskunde. Hoogland was ook geen standaard promovendus. Hij werkt bij de stichting leerplanontwikkeling (SLO) en heeft al een lange carrière achter de rug: hij was eerder werkzaam als wiskundeleraar, lerarenopleider, nascholer, leermedelenauteur, hoofdredacteur van *Euclides* en directeur van het Algemeen Pedagogisch Studiecentrum (APS).

Het onderzoek dat ten grondslag ligt aan zijn proefschrift werd in 2010 geselecteerd als een van de 19 prijswinnaars van de 74 inzendingen voor het programma Onderwijs Bewijs II van het ministerie van OCW. Een programma dat ernaar streefde om 'bewezen' antwoorden te krijgen op actuele onderwijsvragen.

### **Fascinatie voor wat rekenen en wiskunde doen met mensen**

In 1984 werd Hoogland student-assistent wiskundendidactiek in Leiden bij Gerben Bulthuis en vanaf 1985 was hij wiskundeleraar. Hij was toen al geïnteresseerd in de wereldwijde gemeenschap rond wiskundeonderwijs, met namen als Hans Freudenthal, Alan Schoenfeld en Alan Bishop, en in de internationale trends rond wiskundeonderwijs. Als jonge wiskundeleraar wilde hij al per se naar het International Congress on Mathematical Education (ICME) in 1988 in Budapest.

Die fascinatie is altijd gebleven: "Wat doen rekenen en wiskunde en het leren daarvan met mensen?" Later werd het wat specifiek: "Hoe gaan met name zwakkere leerlingen om met de kwantitatieve wereld om ons heen? Hoe kunnen praktische rekenkennis en zelfvertrouwen hen *empoweren*? Hoe kunnen rekenangst en een goed/fout- en deficiëntiecultuur hen juist belemmeren?"

Ruim tien jaar geleden ontstond het idee om daar ook zelf actief onderzoek naar te gaan doen. Dat heeft uiteindelijk via het Onderwijs Bewijs-traject tot zijn dissertatie geleid.

### **Beschrijvende versus beeldende contextopgaven**

Probleemoplossen en modelleren hebben steeds meer terrein gewonnen als doelen van het reken- en wiskundeonderwijs. In vrijwel elk curriculum wereldwijd zijn dergelijke onderwijsdoelen opgenomen. Daarbij spelen als vanzelfsprekend 'probleemsituaties uit de werkelijkheid' ofwel contexten een belangrijke rol. Er wordt veel



onderzoek gedaan naar goede manieren om die contexten in de schoolsituatie te (re)presenteren.

Hoogland heeft een aspect daarvan onderzocht door middel van de volgende onderzoeksvraag: “Wat is bij wiskundige contextopgaven het effect op de score van leerlingen wanneer we de representatie van de probleemsituatie veranderen van beschrijvend naar voornamelijk beeldend?” De figuur illustreert een voorbeeld van zo’n verandering.

### Verhaaltjessommen authentiek aanbieden

Uit onderzoek is bekend dat veel leerlingen contextopgaven zien als ‘verhaaltjessommen’ waarin procedurele sommen zijn verstopt. Het gaat hierbij meestal om talige representaties van contexten. Leerlingen plukken dan lukraak getallen uit het verhaaltje en laten er een bewerking op los. Dit gedrag wordt vaak *suspension of sense-making* genoemd, grofweg te vertalen als ‘het verstand op nul zetten’. De te onderzoeken aanname was dat leerlingen tot betere resultaten zouden komen als de probleemsituaties meer beeldend — en daardoor wellicht authentiek — aangeboden zouden worden. Dit zou een indicatie kunnen zijn dat ze in die variant hun verstand wat minder op nul zouden zetten.

### Het hoofdonderzoek

Echt een hoogtepunt tijdens het traject was de periode van het hoofdonderzoek. Om conclusies te kunnen trekken had Hoogland ruim 1000 deelnemers nodig. Maar binnen enkele weken stroomden de potentiële deelnemers binnen. De teller stond uiteindelijk stil bij ruim 32.000 geschikte leerlingen in basisscholen, voortgezet onderwijs en mbo.

Het onderzoek was opgezet als een *randomized controlled trial*. In de digitale toets die aan de deelnemers werd voorgelegd, werd voor elke opgave random een van de twee representaties gekozen. Alle antwoorden, de bestede tijd, de scores en de leerlingkenmerken werden anoniem in een onderzoeksdatabase opgeslagen. Op piekuren waren meer dan 4000 leerlingen tegelijk bezig met de digitale toets en dat zonder haperingen. Hoogland kon letterlijk zien hoe de database zich vulde.

Toen kwam de spanning van de eerste analyses: zou de hypothese bevestigd worden? Die spanning maakte het onderzoeken mooi. Het werd duidelijk hoe complex het is om iets te meten en te beweren over onderwijssituaties. “De analyse leerde mij zeer zorgvuldig en precies te zijn in uitspraken over onderwijs en ook waardering te hebben voor andere mensen die zorgvuldig en genuanceerd praten over reken- en wiskundeonderwijs”, aldus Hoogland.

Het resultaat van het onderzoek was dat over de hele toets beschouwd leerlingen net wat hoger op de beeldende opgaven scoorden. Gedetailleerdere analyses lieten zien dat dit niet gold voor alle

opgaven: voor opgaven uit het domein Meten & Meetkunde was het effect groter dan bij opgaven uit domeinen als Getallen en Verhoudingen, en er waren opgaven waarbij het omgekeerde het geval was.

### De waan van de dag

Hoogland vond het vaak lastig om voldoende focus te houden op zijn onderzoek. Rond het reken- en wiskundeonderwijs heb je nogal eens met de waan van de dag te maken en dat combineert slecht met een wat meer afstandelijke en diepe analyse van de problematiek. Hoogland houdt zich bovendien ook graag bezig met allerlei aspecten van reken- en wiskundeonderwijs die hem nogal eens van zijn dissertatieonderzoek afhielden: ontwikkelen, scholen bezoeken, advieswerk in binnen- en buitenland, nascholing van docenten, schrijven van columns en populair wetenschappelijke artikelen, en het volgen van de internationale trends in het reken- en wiskundeonderwijs. Hoogland geeft een beeld van de overweldigende omvang daarvan: op de laatste ICME-conferentie in 2016 in Hamburg waren 3500 deelnemers die ruim 1900 artikelen over recent onderzoek in reken- en wiskundeonderwijs presenteerden. Het leert hem dat simpele analyses en simpele oplossingen ernstig gewantrouwd moeten worden als het gaat over verbeteren van dat onderwijs.

### Conclusies

Heeft dit onderzoek nu iets bewezen? Persoonlijk gelooft Hoogland daar niet zo in als het gaat om sociaal-wetenschappelijk onderzoek. Wel vindt hij het belangrijk dat goed onderzochte resultaten meegenomen worden in afwegingen die mensen maken in hun werk: leraren die contextopgaven aan hun leerlingen voorleggen, methodeschrijvers die nadenken over hoe ze probleemsituaties uit de wereld om ons heen representeren in de boeken, toetsmakers die meer zouden kunnen experimenteren met verschillende manieren om contexten in een test te representeren. Tot slot beleidsmakers, die moeten beseffen dat een digitale toets met goed/fout-antwoorden lang niet het hele spectrum dekt van de doelen uit het referentiekader Taal en Rekenen. Als het effect van zijn onderzoek is dat deze mensen inderdaad beter gefundeerde afwegingen maken, is Hoogland zeer tevreden.

Het onderzoek heeft voor hem ook nog een bredere relevantie. Voor reken- en wiskundeonderwijs op alle niveaus is de vraag hoe de combinatie van verbale en visuele componenten een rol speelt bij het ontwikkelen van wiskundige cognities. Het is bekend dat bij het duurzaam verwerven van wiskundige kennis en vaardigheden visualisaties, patronen, structuren en beeldende en dynamische metaforen een belangrijke rol kunnen spelen. Hoogland vraagt zich dus af waarom men op alle niveaus van het reken- en wiskundeonderwijs zo blijft hangen in talige representaties van de werkelijkheid bij probleemoplossen, modelleren en toepassen van rekenen en wiskunde. Er is al zoveel meer mogelijk, bijvoorbeeld op het gebied van *virtual and augmented reality*. Er is dus nog genoeg te ontwikkelen en te onderzoeken.

Zijn gewone werk gaat ook door. Bij SLO wordt geprobeerd de curriculumontwikkeling in Nederland blijvend te voeden met kennis van internationale trends en recente wetenschappelijke onderzoeksresultaten. Hoogland heeft al gemerkt dat zijn nieuwe ‘status’ mogelijkheden biedt om bijvoorbeeld als co-promotor jonge mensen in binnen- en buitenland te begeleiden in onderzoek dat vergelijkbaar is met het zijne. En dat inspireert enorm.

## Wim Caspers

Lyceum Ypenburg, Den Haag, en  
Faculteit EWI en Lerarenopleiding, TU Delft  
w.t.m.caspers@tudelft.nl

## Mark Timmer

Carmel College Salland, Raalte, en  
ELAN, Vakgroep Docentontwikkeling, Universiteit Twente  
m.timmer@utwente.nl

**Boekbespreking** Ronald Meester en Joost Hulshof: Wiskunde in je vingers

# Hogere wiskunde omlaaghalen

“Een oppepcursus voor liefhebbers en andere freaks”, zo omschrijven Ronald Meester en Joost Hulshof hun boek  $W=\{kunde\} \frac{in}{je}$  vingers. Wim Caspers en Mark Timmer, allebei docent aan een middelbare school en lerarenopleider, bekeken het boek ieder vanuit beide rollen en gingen op bezoek bij de auteurs. Directe aanleiding is het gratis online beschikbaar stellen van het boek door de uitgever.

De eerste zin van het voorwoord vat het boek goed samen: “In dit boek zetten we een aantal basisbegrippen uit de wiskunde in een wat ander daglicht dan gebruikelijk”. Het betreft een breed scala aan onderwerpen: algebra, differentiaalrekening, integraalrekening,  $e$ ,  $\pi$ , oneindigheid, combinatoriek, kansrekening, maattheorie en de normale verdeling; het komt allemaal aan de orde.

Het boek is gebaseerd op twee readers die de auteurs maakten voor een nascholing voor wiskundedocenten, maar de uiteindelijke keuze van onderwerpen is ingegeven door vragen die bij de eerste cursus gesteld werden door de deelnemers. Hulshof geeft een voorbeeld: “Het stuk over toevalsgetallen in de eerste cursus leidde tot allerlei vragen over de wet van grote aantallen en de centrale limietstelling. Dus voor de tweede reader ging Ronald formuletjes uitschrijven, om het uit te rekenen voor een voorbeeldje. Hij had de formule van Stirling nodig en vroeg zich af hoe dat nou precies zat. Ik had net een college gegeven over asymptotische methoden voor integralen en daar was ik juist met die formule begonnen. Dus dat kwam ook in het boek terecht. Dan doe je die berekening, je hebt een integraal voor  $n!$  en dan schaal je zijn maximum en je schuift het op en dan zie je dat de functie die je dan krijgt, in feite de  $e$ -macht gedeeld is door de rente-op-rentebenadering die was behandeld in de eerste reader. Je ziet dan de overeenkomsten met de rekenpartijen van Ronald voor zijn centrale limietstelling. Zo ontstaat een samenhangend beeld van

die onderwerpen dat je niet ziet als je ze apart behandelt.”

### Informatiedichtheid

De auteurs hebben de moeite genomen om veel materie keurig vanaf de basis op te bouwen. In het eerste hoofdstuk vinden we bijvoorbeeld eerst rationale getallen en hun operaties, dan een mooi verhaal over het wel en wee van decimale getallen (inclusief irrationaliteit, repeterende breuken en het feit dat  $1 = 0,999\dots$ ), afgesloten met de vraag hoe oneindig  $\mathbb{Q}$  en  $\mathbb{R}$  nu eigenlijk zijn. De een-op-eenkoppeling met de natuurlijke getallen voor aftelbaarheid en een diagonaalargument voor de overaftelbaarheid van de reële getallen; het is allemaal onderdeel van het verhaal. Een feest van herkenning voor de lezer die dit inderdaad als ‘oppepcursus’ leest, maar wellicht wel een wat hoge informatiedichtheid voor een nog wat minder gevorderde, nietsvermoedende lezer.

Aanvankelijk bestond de doelgroep van de nascholing waar het boek op is gebaseerd niet alleen uit eerstegraadsdocenten wiskunde, zo vertellen Hulshof en Meester. “We hebben ook geprobeerd om tweedegraadsdocenten naar die eerste nascholing toe te krijgen, maar die vonden het wel aan de moeilijke kant. De bedoeling was echt om het toegankelijk te krijgen voor iedereen die de weggezakte wiskunde, niet alleen B maar ook A, wil ophalen. Het begint dus gewoon met breuken. Natuurlijk hebben we teruggehoord dat er bepaalde passages zijn die mensen gewoon moeilijk vinden. Heel veel positieve reacties vooral: vermakelijk, interessant, leerzaam.”

Uit de wiskundewereld zelf komt ook wel kritiek. “De filosofie van het boek was: niet meteen zorgen maken over de details, maar wel nauwkeurig te werk gaan. Er zijn vakgenoten die vinden dat je dat anders moet doen. Eerst netjes alles axiomatisch introduceren en dan pas ermee aan het werk gaan. Maar je kunt ook zeggen: we weten uit ervaring hoe we rekenen met die getallen, we weten wat breuken zijn en we weten wat reële getallen zijn. Het roept controverses op als je het woord breuk in plaats van het woord rationaal getal gebruikt of wanneer je die dingen als hetzelfde ziet. Mag je zeggen dat van een derde de noemer drie is? Daar doen we in dit boek niet moeilijk over. Het is wel onze bedoeling om grondig te zijn, maar sommige zaken worden naar achter geschoven.”

Iets dergelijks signaleert Meester in de leerboeken over kansrekening. “Als je nu



Ronald Meester en Joost Hulshof met  $W=\{kunde\} \frac{in}{je}$  vingers, VU University Press, 2015, 184 p., ISBN 9789086597154, prijs € 19,95. Gratis pdf, zonder afbeeldingen, is te vinden op [www.vuuniversitypress.com/wiskunde-in-je-vingers](http://www.vuuniversitypress.com/wiskunde-in-je-vingers).

een standaardboek over kansrekening ziet, dan begint dat met definities, uitkomstenruimte, kansmaat, de axioma's van Kolmogorov en dan ga je rekenen en op een gegeven moment komen de stochastische grootheden. Is dat allemaal meteen al nodig? Ik erger me daar al langere tijd aan. Het gaat er namelijk niet om alleen het proces te beschrijven waarmee je kansen uit kan rekenen, maar meer wat het nou betekent om een kans uit te rekenen. Hoe moet ik het interpreteren en hoe en wanneer kan ik het gebruiken?"

### Limieten?

Meester en Hulshof proberen zo ver mogelijk bij limieten weg te blijven, en bepalen zo bijvoorbeeld de helling van een polynomiale functie niet door middel van differentiequotienten maar door slimme algebraïsche herschrijvingen. Zo wordt de raaklijn aan de grafiek van  $f(x) = x^5$  in het punt  $(a, a^5)$  bepaald door te stellen dat

$$f(x) = x^5 = a^5 + (x^4 + ax^3 + a^2x^2 + a^3x + a^4)(x - a),$$

waarin de grote voorfactor voor  $x = a$  reduceert tot vijf keer dezelfde term. Natuurlijk moet de formule die je dan krijgt wel iets met wat je de raaklijn in  $x = a$  zou willen noemen te maken hebben. Die raaklijn is de grafiek van een functie. Met een schatting voor de fout maak je precies wat je daar mee wilt bedoelen, aldus Meester en Hulshof. Evenzo worden oppervlaktes exact bepaald zonder te primitiveren, door het insluiten met onder- en bovensommen. Fraai en vermakelijk, hoewel niet altijd de meest eenvoudige wijze om tot antwoorden te komen (maar dat hoeft ook niet altijd).

Ook het getal  $e$  komt uitgebreid aan de orde (in wat de auteurs het 'centrale hoofdstuk' noemen), waarbij via een rustige opbouw op basis van rente wordt gekozen voor de definitie

$$e^a := \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n.$$

Hier wel een limiet overigens dus, van een stijgende begrensde rij, maar dat zulke limieten bestaan is niet heel anders dan 1 als limiet van 0,999... Heel netjes worden vervolgens enige rekenregels bewezen evenals de overeenkomstigheid van de karakterisatie als oneindige som, maar ook wordt zeer uitgebreid stilgestaan bij de vraag of deze definitie wel in de haak is,

door te bewijzen dat  $e^a$  eindig is voor iedere  $a > 0$ . Ook het geval  $a < 0$  wordt apart en uitgebreid onder de loep genomen. Een heel zuiver en mooi verhaal.

### Humor

Het boek is naast interessant ook gewoon leuk om te lezen. Meester en Hulshof hebben weinig mogelijkheden onbenut gelaten tot een grappige toevoeging of een persoonlijke noot. Deze zijn met name te vinden in de grote hoeveelheid voetnoten die het boek rijk is. De hoofdstekst bevat slechts zo'n twintig pagina's zonder voetnoot, en op de meeste pagina's staan er twee of drie. Soms met nuttige wiskundige toevoegingen of verwijzingen voor de 'liefhebbers' uit de ondertitel, soms met nerdy grapjes voor de 'freaks' ("een goede definitie is per definitie een definitie"), soms met fijne flauwe humor (als je ervan houdt) zoals "de opmerking die je wist dat zou komen" of "meditatie hielp de verleiding van kettingbreuken voor dit boekje te weerstaan".

Ook de index achterin het boek is ongebruikelijk. Die bevat degelijke items als 'e-macht' en 'Banach-Tarski-paradox', maar ook minder voor de hand liggende zaken als 'bits', 'Engeland' en 'strafwerk'. De auteurs lichten toe: "Dan Henry schreef daarover dat je ants en grasshoppers kan onderscheiden onder je lezers:

"We may distinguish between the ants, who read page  $n$  before reading page  $(n+1)$ , and the grasshoppers who skim and skip until something of interest appears and only then attempt to trace its logical ancestry."

[citaat Daniël Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer, 1981]

Het voorwoord is een soort statement over hoe we tegen de dingen aankijken en daarin wordt beschreven wat er in het boek staat, voor de ants. De index is voor de grasshoppers en we wilden de index ook een beetje grappig maken. Je ziet bijvoorbeeld staan 'zo simpel mogelijk' bij de  $z$  en vraagt je af waar dat over gaat. Het nodigt uit om te kijken of het interessant is."

### Doelgroep

Naast uitleg is er ook een grote hoeveelheid aan opgaven door het boek heen gegeven. Erg leuk en grotendeels goed te

doen, maar toch jammer dat er geen oplossingen achterin te vinden zijn. Hulshof lachend: "Geen uitwerkingen achterin is misschien ook een beetje een statement in deze tijd. Er zitten wel veel hints in de opgaven zelf trouwens." De auteurs roepen de lezer in het nawoord bovendien nadrukkelijk op om contact te zoeken als ze vragen of suggesties hebben. Het wiskundig niveau van de theorie en de opgaven wisselt van eenvoudig tot behoorlijk pittig, wat het lastig maakt om heel precies een doelgroep van het boek te karakteriseren. Wie wiskunde of de eerstegraads lerarenopleiding wiskunde heeft gestudeerd, zal zeker met het boek uit de voeten moeten kunnen. De combinatie van herkenning, een nieuwe blik op bekende concepten en zo nu en dan ook nog wel eens wat nieuws, maakt het tot een vermakelijke verhandeling. De voorzitter van de Nederlandse Vereniging van Wiskundeleraars prijst het boek op de achterkant aan als een "mooie bijdrage aan de permanente scholing van docenten", en daar kan het zeker voor gebruikt worden.

De auteurs zouden dat ook toejuichen. "Als een alternatieve kijk op hoe de wiskunde in elkaar zit, zou dit voor lerarenopleidingen een ideaal boekje zijn om te gebruiken. Het geeft hopelijk inzicht in de stof die onderwezen moet worden en helpt ook om te zien hoe die samenhangt met andere zaken. En dat komt de didactiek ten goede. Het is niet uit de hoogte bedoeld om leraren opnieuw op te voeden. Het is echt de bedoeling om hogere wiskunde omlaag te halen. Laten ervaren dat iets wat je eerst heel moeilijk vond door een andere presentatie veel duidelijker wordt."

"Natuurlijk zit er ook een boodschap in dit boekje, daarvoor zijn wij provocerend en eigenwijs genoeg." Meester: "Een filosofie over de didactiek van de wiskunde en een filosofie over de wiskunde zelf. Dat we het niet over alles eens zijn, wordt ook wel duidelijk in het boek. Dat is belangrijk omdat het ingaat tegen de overtuiging dat wiskunde een objectief bouwwerk is waar je eigen mening geen rol meer speelt." Hulshof: "Dat niet iedereen het erover eens is dat 0,9999... gelijk is aan 1, om maar wat te noemen. En dan hebben we het nog niet eens over het keuzeaxioma."

Ondanks hun verschillen van inzicht zien de auteurs mogelijkheden genoeg voor een vervolg. En misschien hoeven we daar niet zo heel lang op te wachten. ☺

Ferdinand Verhulst

Mathematical Institute  
Utrecht University  
f.verhulst@uu.nl

**Book Review** Bertrand Duplantier and Vincent Rivasseau (eds.): *Henri Poincaré, 1912–2012*

# A global view of Poincaré's scientific work

On the occasion of the centennial of the death of Henri Poincaré in 1912 the monograph *Henri Poincaré, 1912–2012* appeared, edited by Bertrand Duplantier and Vincent Rivasseau. This book is published as the thirteenth volume of the Poincaré Seminar Series. Ferdinand Verhulst, author of *Henri Poincaré, impatient genius*, discusses each of the four chapters of the book.

In a long period around 1900 two giants of mathematics dominated the scene: David Hilbert and Henri Poincaré. It is nearly impossible to compare them, Hilbert who solved essential mathematical problems and who created fundamental generalizations of concepts, Poincaré who invented many new concepts to solve classical problems and who also made important contributions to physics and engineering. While Hilbert laid solid foundations of theory we are using to build on, the reaction to Poincaré has been completely different. We see each year new papers and books about Poincaré's work, he was a fountain of ideas and theories. As Vito Volterra said about him: "He was the impressionist among the mathematicians."

This book is the thirteenth in a series of Poincaré seminars aimed at a wide public of scientists. It consists of four chapters and a description of a movie focusing on scientific topics and experiments. To some extent it aims at presenting a global view of Poincaré's scientific work; we will discuss the chapters separately.

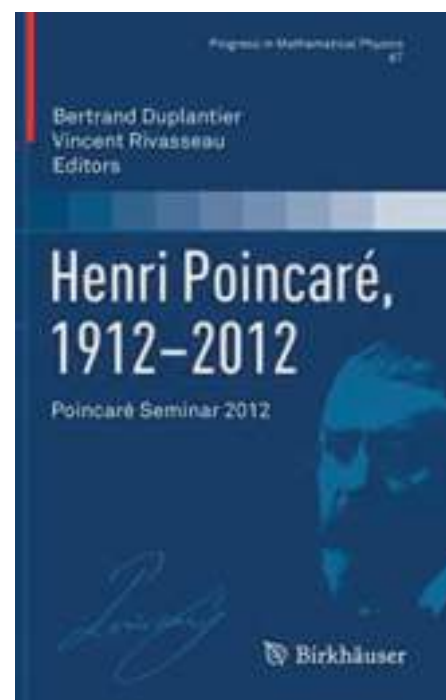
## Poincaré's optics

The first chapter is written by Olivier Darrigol (pp. 1–50). Interestingly and unusually, Darrigol is a historian of science who seems to know and understand the mathematical

physics of around 1900. In an interesting fifty pages he describes how in Poincaré's mathematical optics this leads Poincaré to diffraction, electrodynamics and relativity. Physics is intimately connected in this work with geometry. To avoid confusion to the reader: this connection does not reduce physics to mathematics, that would not be fruitful but mathematics is an efficient and very useful tool to describe the physical phenomena. Already much earlier, for Fresnel the concept of the ether was not necessary, but it often came up because of the inspiration to develop optical theory derived from the theory of elasticity where waves do need a medium. Maxwell unified optics and electromagnetism but the French did not like his results because of his non-axiomatic style. Poincaré however, admired Maxwell's results while deploing his lack of rigorous mathematical physics; the French took classical mechanics probably too much as a standard for theoretical physics. The misunderstanding that mechanics is part of mathematics has produced many beautiful results, think of Euler, Lagrange, Laplace and Riemann, it also obstructed the emergence of new physical ideas. In Poincaré's lectures of 1887–88 he reflects on a large number of optical theories while noting that it does not matter whether the ether exists or

not. The ether is just a convenient tool as electromagnetic waves act as if the ether exists. This seems also to have been the point of view of the most important physicist of that time, H.A. Lorentz. Poincaré mentions the ether but also states that at some time the ether will be discarded as useless.

Poincaré lectured in 1908 at the Institute of Post and Telegraph where he taught about the transmission of light and



Bertrand Duplantier and Vincent Rivasseau (eds.), *Henri Poincaré, 1912–2012*, Progress in Mathematical Physics, Vol. 67, Birkhäuser (Springer), 2015, 234 p., ISBN 9783034808330, prijs \$ 109.

wireless telegraph signals. His analysis of self-excited oscillations led to many applications in this engineering field.

One of the conclusions of the author is that Poincaré pushed theoretical physics from the study of particular models to a nuanced understanding of the organizing principles. His interest went further to the development of special relativity and Lorentz transformations in a publication shortly before Einstein's 1905 publication on relativity, see also [5].

### The three-body problem

In a second chapter (pp. 51–149) Alain Chenciner describes Poincaré's work on the 'three-body problem'. The basis for this chapter are the *Méthodes nouvelles de la mécanique céleste* and a three-year seminar on reading these books. It correctly emphasizes that the *Mécanique Céleste* of Poincaré is the first text on the modern theory of dynamical systems with main but not exclusive field of application the three-body problem. After describing the problem formulations and results of Lagrange and Laplace on the stability of the solar system, the author discusses various coordinate systems (a recurring and rather nasty aspect of celestial mechanics) to formulate the basic problem of perturbing simple solutions as periodic ones and tori that have been obtained by simplifying or truncating the complete equations of motion. For centuries, the main tool to handle such questions have been by series expansions, a topic discussed extensively by Poincaré with the two-fold motivation that series expansion should add precision and at the same time should prove existence of the objects described; both aspects are by no means trivial as most expansions in celestial mechanics are formal. Poincaré shows that most expansions are divergent, also the Lindstedt series which he prefers. Lindstedt introduced his series as a formal procedure to approximate a perturbed harmonic equation and a system of two coupled harmonic equations. Poincaré used many chapters to discuss series and their properties, a discussion to which the author of this chapter adds pictures of the type Poincaré could have had in mind. The inexperienced reader might lose his way in the various cases of fixed or varying frequencies, fixed or varying initial conditions, the role of singularities, degeneracies and resonances,



Henri Poincaré

but for people working in this field it is a beautiful summary. Integrability questions and integral invariants showing the relation with results by Lie and later Elie Cartan, are placed in the framework of modern theories. Other aspects are the use of normal forms and bifurcations to study periodic solutions and tori, clearly also anticipating singularity theory. It will have been quite an effort to compile this chapter, but notwithstanding the wealth of information we still have to add a few critical remarks.

First, the author does not seem to be aware of the modern theory of averaging and normal forms, see [4] with its results on accuracy and timescales; these basic aspects could have clarified a large number of calculational results. Then, somewhat more emphasis could have been given to the fact that Poincaré–Lindstedt series are convergent for periodic solutions, also for KAM-tori with fixed Diophantine frequencies. As a third item: Birkhoff normal forms were in modern times replaced by Birkhoff–Gustavson normal forms with many new results. Finally, on page 97 the method of Ziglin and other mathematicians is mentioned as the main tool to study non-integrability; this method has been very fruitful qualitatively but it misses out on evidence regarding the dynamics and the nature of non-integrability in which horseshoe maps often play a part. One may find for instance algebraically that a Hamiltonian system is non-integrable but one remains ignorant of the fact that this happens in thin subsets so that the non-integrable effects are negligible or find for

instance that non-integrability has significant effects but is very much localised to certain subsets. Other methods are in this respect important and helpful like those of Duistermaat and of Shilnikov–Devaney (for references see again [4] or [5]).

### Poincaré's odds

Chapter 3 by Laurent Mazliak (pp. 151–192) describes probability. This field was not one of the main topics of Poincaré's research, but he lectured and published about probability influencing in this way its development, for instance by Borel and Bachelier. Until the end of the nineteenth century, probability got little attention in France and it is typical that, although Poincaré formulated ideas about randomness, he still kept a position between the deterministic views of that time and the allowance of random phenomena. This becomes clear in his approach to thermodynamics and statistical mechanics, emerging fields in his time. Another problem in appreciating his contributions is that Poincaré considered probability problems and made calculations without stating assumptions and producing a general framework.

An early probabilistic statement can be found in [2] (the final text of his famous Prize Essay) where he formulates the Recurrence Theorem. Later, when he turns his attention to thermodynamics this recurrence yields basic problems in the description of systems of many particles. The Maxwell postulate of equipartition of kinetic energy of particles conflicts with recurrence. Is equipartition a transitory state lasting for an extremely long time?

And what about mixing of fluids? A drop of ink into a glass of water will mix, we do not expect that the drop of ink materialises again after some time. Without actually solving all these questions but certainly inspired by them, Poincaré started lecturing on probability. He considered the repetition of experiments that for a large number of times lead to a kind of asymptotic equilibrium. In this resulting distribution, the final state, the initial distribution plays no part. Poincaré applies this to playing roulette, the distribution of asteroids in the solar system and to card shuffling. His calculations are very explicit without reference to theorems known at that time; this is of course typical for Poincaré's style of working and writing.

An interesting discussion concerns the appreciation of the heritage of Poincaré's ideas on probability. His immediate successor in France in this field was Émile Borel who was a star of analysis around 1900. Unexpectedly, Borel started to work on probability, until that time a disreputable topic among mathematicians. He introduced quantification of randomness by using the Lebesgue integral and measure theory. In line with Poincaré's conventionalism he identified probability with the measure of a subset of a suitable set. Much later, the theory of Markov chains became known with contributions by the Czech Hostinsky who was inspired by Poincaré's treatment of the card shuffling problem and by discussions on statistical physics. The author

of this chapter summarizes Poincaré's influence on probability as follows:

He began to extract the domain from the grey zone to which it had been confined by almost all French mathematicians, he initiated methods that flourished when they integrated more powerful mathematical theories, he convinced Borel of the importance of certain questions, to the study of which he eventually devoted an enormous amount of energy. For a rather marginal subject in Poincaré's works, such a contribution appears far from negligible.

### Uniformisation of Riemann surfaces

Chapter 4 by François Béguin (pp. 193–230) is for a large part based on the book [3]. The author notes wisely that to understand Uniformisation theory one should ignore the modern formulations and consider the original discussions first. In the nineteenth century, uniformisation was a matter of parametrisation of Riemann surfaces for multi-valued complex functions. (I remember that as a student of complex functions I was baffled by the 'solution' for multi-valuedness by joining Riemann sheets at the problematic positions; how could this be interpreted geometrically, and in what space?) The nineteenth century idea to describe Riemann surfaces was to use parametrisation of curves and surfaces. Poincaré was the first to prove that it is always possible to reduce

any algebraic relation, say  $F(x, y) = 0$ , by expressing it as single-valued automorphic functions of one variable. The geometric interpretation of this parametrisation came much later. It is typical that Poincaré's interest in these questions (he was still a lecturer in Caen) came from learning about the treatment of Fuchs of linear ordinary differential equations. The coefficients of the equations have poles so that the complex solutions are naturally multi-valued. A little bit later, after 1880, Poincaré succeeded in proving a more general uniformisation theorem valid for all algebraic Riemann surfaces. It took a long time, until 1907, for Poincaré and Koebe to prove that all Riemann surfaces, algebraic or not, can be uniformised. The author of this chapter traces in detail the discussions and results of the various uniformisation theorems, also the part played in the initial phase by Felix Klein.

Many aspects of the four papers in this monograph can be found in the biographies [1] and [5]. However, the detailed discussions in all of them add a lot of material and understanding to these biographies; for me the third chapter on probability presented a clear and new view on Poincaré's interest in this field. Chapter 2 gives a relatively rare discussion on the problems of studying tori bifurcations by series expansions. The monograph can be a serious help to scientists aspiring to study Poincaré. ☞

### References

- 1 Jeremy Gray, *Henri Poincaré*, Princeton University Press, 2012.
- 2 Henri Poincaré, Sur le problème des trois corps et les équations de la mécanique, *Acta Mathematica* 13 (1890), 1–270.
- 3 H.-P. de Saint-Gervais, Uniformisation des surfaces de Riemann, retour sur un théorème centenaire, *École Normale Supérieure Éditions*, Lyon, 2010.
- 4 J.A. Sanders, F. Verhulst and J. Murdock, *Averaging methods in nonlinear dynamical systems*, Applied Mathematical Sciences, Vol. 59, Springer, 2007, 2nd ed.,
- 5 Ferdinand Verhulst, *Henri Poincaré, impatient genius*, Springer, 2012.

# Boekbesprekingen

| Book Reviews

Redactie: Hans Cuypers en Hans Sterk

Review Editors NAW - MF 7.092  
Faculteit Wiskunde & Informatica  
Technische Universiteit Eindhoven  
Postbus 513  
5600 MB Eindhoven  
reviews@nieuwarchief.nl  
www.win.tue.nl/wgreview



Ed de Moor

## Vroeger

40 historische columns over het reken-  
onderwijs

NVORWO/Uitgeverij Zwijsen, 1999  
96 p., uitverkocht  
ISBN 9789075586022

*Dit is een pleidooi voor een herdruk, naar aanleiding van het overlijden van Ed de Moor. Een In Memoriam vindt u elders in dit blad.*

In 1999 verscheen bij Uitgeverij Zwijsen in Tilburg onder de titel *Vroeger* een boekje met daarin 40 columns die Ed de Moor tussen 1991 en 1999 had geschreven in het *Tijdschrift voor Reken- en Wiskundendidactiek Willem Bartjens*. Het boek is alleen nog maar antiquarisch te verkrijgen, maar verdient het absoluut om herdrukt te worden.

De 40 bijdragen zijn prettig compact en allemaal voorzien van een foto of plaatje. Veel ideeën ontleende Ed aan oude schoolboeken en aan onderwijstijdschriften, waarbij vooral de negentiende eeuw een ware schatkamer bleek te zijn. Voorts leverde de geschiedenis van onderwijs en pedagogiek van reken- en wiskunde-onderwijs volop inspiratie. Ed laat zien hoe de geschiedenis van de wiskunde een nieuw licht kan werpen op hedendaagse didactische problemen. In een aantal van de bijdragen wordt ingegaan op invloedrijke figuren. Enkele namen: Willem Bartjens (1569–1638) met zijn boek *Cijfferinge*, hetwelk tot ver in de negentiende eeuw gebruikt is; Jan Versluijs (1845–1920) met zijn boeken over reken-onderwijs; Comenius (1592–1670); Pythagoras en Pierre de Fermat samen in één bijdrage (geschreven in 1993/94 toen het Fermat-vermoeden op het punt stond bewezen te worden); de meetkundige en historicus Dirk Struik (1894–2000); het rekenwonder Willem Klein (1912–1986), alias Pascal, Willy Wortel, Moosje Optel; Jan Ligthart (1859–1916) en Johann Heinrich Pestalozzi (1746–1827).

Leuk zijn ook beschouwingen over het boekje van J.F.L. Muller: *Nommerkransje, een geschenk voor kinderen die gaarne willen leren tellen* (1840); het verhaal over een drama in de klas bij het oplossen van het stelsel vergelijkingen  $a + 2 = 5$  en  $b + 2 = 5$ ; enkele (andere) bijdragen over rekenboekjes; honderd jaar denksommen; een (Groningse) ruzie over 'Meneer van Dalen wacht op Antwoord' en een stuk over rekenen op lijnen in de middeleeuwen. Er zijn ook technische zaken te vinden in het boekje, zoals: het ontstaan van de schrijfwijze van ons huidige procentteken. Ed behandelt kort *De Thiende* van Simon Stevin (1548–1620), de rekenmachine 'de wentelaar' in het Nationaal Schoolmuseum te Rotterdam en de uitvinder van de paperclip, Johan Valer (1866–1910), met beschrijving van het object. Er is ook een verhaal over de onderwijshervormer Theo Thijssen (1879–1943), met daarbij voorbeelden van 'metriekre krenge'. Dat zijn zeer gecompliceerd uitzienende breuken, veelal gelardeerd met inhouds/afstandsmaten, die na herleiding toch vaak een (mooi) overzichtelijk resultaat gaven. De negenproef wordt nader bekeken en het wel en wee van Charles Babbage (1791–1871) en zijn rekenmachine komt aan de orde. Enkele verhalende beschouwingen hier zijn: een vertelling over schoolinspecteur Wijnbeek (1772–1866), wat hij door Nederland reizende zoal aantrof; hoe Ed in Malden bij Nijmegen een zogenaamde

meetkunde-mozaïekdoos van Fröbel (1782–1852) wist op te sporen; hoe hij in het oosten van Duitsland het door Schmidt geschreven boek over meetkundig materiaal voor kleuters, voorzien van opmerkingen van Fröbels hand mocht inzien, en de inhoud teleurstellend vond. En dan is er een lang verhaal over Maurits Escher (1898–1972), zijn werk, en de verbindingen met meetkunde. Het boekje besluit met alleraardigste bijdragen over experimentele meetkunde in Frankrijk in de negentiende eeuw, over Pythagoreïsche driehoeken, het getal nul in oude culturen, en een beschouwing over de vijftiende- en zestiende-eeuwse rekenboekjes naar aanleiding van het proefschrift van Marjolein Kool dat in 1999 verscheen. Kortom, een opmerkelijk prachtig boek, met een rijke inhoud. En dan heb ik nog niet eens alle in het boek afgedrukte bijdragen genoemd.

Rob van der Waall



Simeon Ball

**Finite Geometry and Combinatorial Applications**  
**London Mathematical Society Student Texts 82**

Cambridge University Press, 2015  
 xii + 285 p., €35,99  
 ISBN 9781107518438

Laat ik beginnen met, in vertaling, wat kreten uit de flaptekst: “De projectieve en polaire ruimten die horen bij een vectorruimte over een eindig lichaam zijn uitzonderlijk nuttig bij de constructie van combinatorische objecten, zoals Latijnse vierkanten, designs, codes en grafen. Dit boek geeft een introductie in de theorie van deze meetkunden en hun vele toepassingen in andere gebieden binnen de combinatoriek [...] *Finite Geometry and Combinatorial Applications* is ideaal voor iedereen, van derdejaarsstudent tot researcher, die zichzelf vertrouwd wil maken en waardering wil krijgen voor het gebied van de eindige meetkunde.”

Wat kan ik hier aan toevoegen? Ik ken de schrijver Simeon Ball sinds zijn promotie in 1994 bij James Hirschfeld. Sindsdien hebben we altijd nauw samengewerkt en heeft hij een aantal (zeer) belangrijke resultaten bewezen in de eindige meetkunde, waaronder een bewijs van het *maximal arcs*-vermoeden en een volledige oplossing van het richtingenprobleem, twee prachtige resultaten die in dit boek overigens niet aan de orde komen.

Zijn meest spectaculaire resultaat is op het gebied van maximum distance separable codes en betreft een probleem dat ik ooit heb horen omschrijven als de Heilige Graal van de coderingstheorie. Het kan eenvoudig als volgt geformuleerd worden. Laat  $V = V(d, q)$  een  $d$ -dimensionale vectorruimte zijn over het lichaam met  $q$  elementen, en laat  $A \subset V$  een collectie vectoren zijn met de eigenschap dat elk  $d$ -tal een basis vormt. Dan geldt, als  $q$  priem is en  $d \leq q$ , dat  $|A| \leq q + 1$  (en deze grens is scherp). Het volledige bewijs van deze stelling vormt het laatste en langste hoofdstuk van dit boek en is op zichzelf voldoende reden om het boek aan te schaffen.

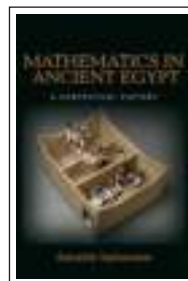
Een ander belangrijk gedeelte van het boek gaat over de combinatorische toepassingen. Voor dit onderdeel geldt zeker dat het voor een goede derdejaarsstudent te lezen is, zelf vind ik in het

bijzonder hoofdstuk 6 over verboden deelgrafen interessant. Dit materiaal is hier op een prettige manier bijeengebracht. Aan de basis van dit deel van de extremale combinatoriek ligt een resultaat van de Nederlander Mantel uit 1907 (oplossing van probleem 28 uit *Wiskundige Opgaven 10*): Het maximale aantal kanten in een graaf op  $n$  punten zonder driehoeken is  $\lfloor n^2/4 \rfloor$ . Het verband met eindige meetkunde is het duidelijkst wanneer we kijken naar bipartiete grafen zonder viercyclen: de punt-lijn-verbindingsgraaf van een projectief vlak van orde  $m$  is zo'n graaf op  $2(m^2 + m + 1)$  punten, met  $(m^2 + m + 1)(m + 1)$  kanten, het maximale aantal.

Wat het boek verder bijzonder en zeer nuttig maakt is een overvloed aan opgaven, allemaal voorzien van gedetailleerde oplossingen.

Welke voorkennis is er nodig om dit boek te kunnen lezen? We citeren uit het voorwoord: “Het boek staat op zichzelf in de zin dat de gebruikte stellingen en lemma's bewezen worden. Hoewel er vrijwel geen voorkennis wordt geëist, zou het toch erg nuttig zijn om basiscolleges algebra en lineaire algebra gevolgd te hebben, omdat anders de eerste paar hoofdstukken wel erg beknopt zijn.” Het voorwoord wordt beëindigd met suggesties welke onderdelen uit het boek geschikt zijn om te gebruiken in een cursus, iets wat ik zelf natuurlijk zeker van plan ben.

Aart Blokhuis



Annette Imhausen

**Mathematics in Ancient Egypt**  
**A Contextual History**

Princeton University Press, 2016  
 xi + 234 p., prijs \$45.00  
 ISBN 9780691117133

Na een studie wiskunde, geschiedenis van de wiskunde en Egyptologie, is de auteur nu hoogleraar wetenschapsgeschiedenis aan de Goethe Universiteit Frankfurt. Met dit boek hoopt de auteur de geïnteresseerde lezer in iets meer dan 200 pagina's een inleiding te geven in de geschiedenis van de Egyptische wiskunde. Het boek omspannt de periode vanaf het eind van de prehistorie tot en met de Grieks-Romeinse tijd. Hierbij wordt gebruik gemaakt van veel verschillende soorten historische bronnen, niet alleen van de relatief schaarse wiskundige teksten, maar ook van bouwkundige tekeningen, administratieve documenten en literaire teksten. De wiskundige ontwikkeling wordt daardoor in context gezet met de oude Egyptische cultuur. Elk onderdeel begint met een korte schets van de historische en politieke gebeurtenissen in die periode.

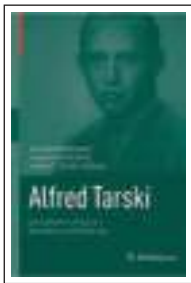
Het eerste deel beschrijft de prehistorie en de vroeg-dynastieke periode. In deze periode maakt Egypte de transitie door van losse nederzettingen naar één staat. De bewaarde bronnen van deze tijd stammen vooral van tombes en tempels, waar de bewaarestandigheden gunstig bleken. Er wordt geschat dat het schrift in Egypte ontstaan is rond 4000 v.Chr. In de tombe U-j zijn de oudste hiërogliefen van Egypte ontdekt: vazen met inscripties, en een verzameling ivoren, benen of stenen labels die voorzien zijn van een pictogram en iets dat geïnterpreteerd wordt als cijfers. Naast de

administratieve noodzaak voor schrift voor voorraadbeheersing, was het schrift ook een middel dat macht representeerde. Dat laatste uitte zich in harde, duurzame ondergronden waarin hiërogliefen werden gebeiteld die millennia meegaan. Voor de dagelijkse voorraadbeheersing was een snel schrift onontbeerlijk. Het hiëratische schrift ontstond. Het Egyptische getalsysteem is gebaseerd op het grondtal 10, maar kent geen positionele waarde. Het heeft een symbool voor elk tiental, en een getal wordt gerepresenteerd door een symmetrische schikking van de benodigde symbolen. Een apart symbool voor nul is daardoor niet nodig om een getal als 103 op te schrijven. Er was wel een symbool om een 0 als uitkomst (gebrek aan restant) te representeren. Imhausen geeft aan dat je de ontwikkelingen in bijvoorbeeld het omgaan met breuken vanuit perspectief moet bekijken, dan kunnen we pas de ingenieuze manieren waarderen die bedacht zijn om om te gaan met breuken. We missen veel als we alleen vanuit ons hedendaags perspectief kijken.

In de overige hoofdstukken worden de perioden Oude Rijk (meetsystemen, breuken), Midden Rijk (wiskundige teksten als Rhind-papyrus, vermenigvuldigen en delen, rekenen met breuken, architectuur en landmetingen), Nieuwe Rijk (administratieve teksten, papyrus Anastasi I) en de Grieks-Romeinse periode (demotische wiskundige papyrussen en rekenkunde) beschreven.

De eerste hoofdstukken zijn vrij compact geschreven, hier en daar een voorbeeld was wat vriendelijker geweest. Een overzichtstijl met jaartallen per periode ter oriëntatie zou een mooie aanvulling zijn. Het is een gedegen werk, dat veel inzicht biedt.

Martine Jansen



Andrew McFarland, Joanna McFarland, James Smith (eds.)

**Alfred Tarski**  
**Early Work in Poland – Geometry and Teaching**

Birkhäuser, 2014  
xxv + 499 p., prijs \$ 149.00  
ISBN 9781493914739

Het boek beschrijft het leven en werk van de wiskundige Alfred Tarski tot 1939. De schrijvers verdiepen zich voornamelijk in het onderzoek dat Tarski deed in de maattheorie en de logica evenals in zijn interesse in het wiskundeonderwijs.

Alfred Tarski werd in 1901 geboren in Warschau, Polen. In 1918 deed hij eindexamen aan het Mazowreckie Gymnasium en in 1919 schreef hij zich in als student wiskunde en logica aan de universiteit van Warschau. In 1921 publiceerde Tarski zijn eerste artikel, in vertaling, 'A Contribution to the Axiomatics of Well-Ordered Sets'. In dit artikel nam hij de postulaten van totaliteit ( $A_1$ ), antisymmetrie ( $A_2$ ) en transitiviteit ( $A_3$ ) aan evenals twee niet-equivalente postulaten ( $B$ ) en ( $C$ ), die 'eerste element'-postulaten genoemd werden. Vervolgens bewees hij dat de systemen  $\{A_1, A_2, A_3, B\}$  en  $\{A_1, A_2, A_3, C\}$  equivalent zijn. Verder bewees hij dat noch  $\{A_1, A_2, A_3, B\}$  noch  $\{A_1, A_2, A_3, C\}$  een onafhankelijk systeem van postulaten is. Zo kunnen bijvoorbeeld de postulaten  $\{A_2, A_3\}$  gederiveerd worden uit  $\{A_1, B\}$ .

Tijdens zijn promotieonderzoek, onder begeleiding van Stanislaw Lesniewski, verdiepte Tarski zich in de logica. Bovendien introduceerde hij de structurele definities van bewijs, stelling, logische stelling en logische consequentie. Gemotiveerd door de topoloog Waclaw Serpienski begon Tarski het werk van Felix Hausdorff te bestuderen. Een belangrijke mijlpaal in zijn carrière was de ontmoeting met Stefan Banach. Zijn interesse in de maattheorie leidde tot het wereldberoemde artikel 'On Decompositions of Point Sets into respectively Congruent Parts'. In dit artikel bewezen Banach en Tarski het paradoxale resultaat dat in een euclidische ruimte van dimensie groter dan twee, twee willekeurige, begrensde verzamelingen, wier inwendige niet leeg is, equivalent zijn onder eindige decomposities. Dit paradoxale resultaat is bekend als de Banach–Tarski-paradox.

In 1939 bracht Tarski een academisch bezoek aan de Verenigde Staten. Zijn terugkeer werd verhinderd door de invasie van Polen door Duitsland en de Sovjet-Unie. Vier jaar later werd hij hoogleeraar wiskunde aan de universiteit van Californië, Berkeley en hij bleef daar tot zijn sterfdag in 1983.

Dit boek levert een substantiële bijdrage aan de literatuur over Tarski's leven en onderzoek. Alle artikelen die voorheen niet in het Engels bestonden, zijn door de samenstellers uit het Pools vertaald. Deel één van het boek beschrijft Tarski's studiejaren; Tarski's eerste artikel staat er ook in. In deel twee wordt het artikel 'On Decompositions of Point Sets into respectively Congruent Parts' gepresenteerd en deel drie beschrijft Tarski's werk als docent wiskunde. Hij publiceerde veel artikelen bedoeld voor scholieren en wiskundedocenten en het grootste deel van die artikelen is voor het eerst in het Engels vertaald. Veel problemen die hij aan studenten en docenten voorstelde, kwamen voort uit zijn onderzoek! Dat is tegelijkertijd fascinerend en inspirerend, althans voor jonge wiskundigen, onderzoekers en wiskundedocenten.

Nicos Starreveld



George M. Zaslavsky

**The Physics of Chaos in Hamiltonian Systems**

World Scientific, 2nd edition 2007  
xiv + 315 p., prijs £107.00  
ISBN 9781860947957

De fysicus Zaslavsky werd in 1935 geboren in Odessa (Oekraïne) en ontwikkelde zich in de jaren zestig in Novosibirsk en Moskou tot een van de belangrijkste Sovjet-wetenschappers. Hij behoorde tot de excellente school van mathematische fysici die tot 1989 de Sovjet-wetenschap in de wereldtop hield. In de jaren negentig emigreerde hij naar de Verenigde Staten waar hij doceerde aan het New Yorkse Courant Institute of Mathematical Sciences. Hier kwamen in 1998 en 2007 de eerste en tweede editie van het besproken boek tot stand. Er moet een bijzondere reden zijn om een boek te bespreken dat tien jaar geleden is verschenen, daar kom ik op terug.

In de eerste hoofdstukken wordt bij Hamiltonsystemen de nadruk gelegd op het generiek samengaan van orde (regulier gedrag

op invariante verzamelingen zoals tori) en chaos. Voorbeelden zijn gestoorde slingers en biljart-dynamica. Samen met B. Chirikov ontwikkelde Zaslavsky de separatrix-afbeelding die veel van het chaotisch gedrag kan karakteriseren. Het ontstaan van chaos bij separatrix-verzamelingen is een fundamenteel mechanisme gebleken. Het reguliere gedrag speelt zich voornamelijk af in zogenaamde eilanden in de chaos; daarvan worden in hoofdstuk 3 een aantal verschillende types behandeld. Dat is ook de plaats om KAM-tori en de fractale verdeling van orde en chaos in de faseruimte te bespreken.

De onderwerpen van de eerste vijf hoofdstukken vindt men in een aantal fysicaboeken, maar meestal slordig geformuleerd, bijvoorbeeld in *Regular and Chaotic Dynamics* van A.J. Lichtenberg en M.A. Lieberman (Springer, 1992); in wiskundeboeken over dynamische systemen wordt vaak zo gegeneraliseerd dat je de fysische context niet herkent, zie bijvoorbeeld *Introduction to the Modern Theory of Dynamical Systems* van A. Katok en B. Hasselblatt (herziene uitgave, Springer, 1996). Onderwerpen die je in artikelen maar zelden in boeken vindt volgen na hoofdstuk 5. Allereerst de Poincaré-recurrentiestelling met fractale aspecten. Recurrentie komt tot stand door maatbehoudende afbeeldingen op begrensde verzamelingen (de fasestroming in Hamiltonsystemen is een voorbeeld) en leveren willekeurig vaak en willekeurig dichtbij de terugkeer bij het startpunt op. In de praktijk is dat een moeilijk hanteerbare stelling omdat het recurrentiegedrag heel erg afhankelijk is van het dynamisch systeem dat men bestudeert, integreerbaar, bijna-integreerbaar, volledig chaotisch, laag- of hoogdimensionaal, enzovoort. Poincaré was zich daar al bewust van. De stelling heeft echter consequenties voor de statistische mechanica. Zermelo gebruikte de discussie over de stelling om de theorie van Boltzmann aan te vallen; voor veel-deeltjessystemen met botsingen, de 'normale' gastheorie, is deze discussie wel afgerond maar voor allerlei veel-deeltjessystemen zoals niet-lineaire ketens zijn er nog veel vragen. Een fascinerende rol in de discussies over recurrentie is gespeeld door de analyse van biljart-dynamica, met name de voorbeelden van Sinai en Cassini. Deze voorbeelden leiden meestal tot zogenaamde 'weak chaos' (gemengd chaotisch en regulier gedrag).

Interessante dynamica ontstaat ook bij aanwezigheid van symmetrie leidend tot webstructuren, een soort roosters met chaotische stroming. Een verschijnsel dat essentieel is voor het begrip van weak chaos is 'quasi-trapping' (hoofdstuk 7). Dit ontstaat indien stabiele periodieke oplossingen bestaan in een omgeving van invariante verzamelingen zoals tori, waar oplossingen die van elders komen enige tijd doorbrengen. Echte invangst van oplossingen is niet mogelijk door de recurrentiestelling, maar het mengen van de deeltjes wordt door quasi-trapping effectief gehinderd. Expliciete voorbeelden zijn te vinden in mijn artikel in *Int. J. Bif. Chaos*, december 2016.

Het boek besluit met een uitvoerige discussie van de Fokker-Planck-Kolmogorov-vergelijking die een stochastische beschrijving geeft van veel-deeltjessystemen.

Het gebied van dynamische systemen en in het bijzonder Hamiltonsystemen is zeer uitgebreid. Wiskundigen houden zich noodzakelijkerwijs individueel bezig met een klein deelgebied waarbij, zeker bij de beginner, het overzicht en de grote lijn vaak ontbreekt. Het boek van Zaslavsky biedt een zeer goed overzicht van een aantal vragen en beschrijvingen van fundamentele mechanismen in Hamiltonsystemen, een reden om het als-

nog te bespreken. Voor verbanden met de algemene theorie van dynamische systemen en bifurcaties die hier nauwelijks aan de orde komen verwijs ik naar *Dynamical Systems and Chaos* van H.W. Broer en F. Takens (Springer, 2011), en V.I. Arnolds werken *Mathematical Methods of Classical Mechanics* (Springer, 1978) en *Geometric Methods of the Theory of Ordinary Differential Equations* (Springer, 1983).

Ferdinand Verhulst



Lewis Carroll

### **Alice's Adventures in Wonderland**

150th Anniversary Edition met illustraties van Salvador Dalí

Princeton University Press, 2015

xxx + 104 p., prijs \$24.95

ISBN 9780691170022

Wat is de reden van weer een nieuwe (Engelstalige) uitgave van het alom bekende (want afgezien van de honderden verschillende versies, ook al in 1951 door Disney verfilmd) verhaal van het meisje Alice dat via een konijnenhol terecht komt in een wonderbaarlijke wereld en daar allerlei avonturen beleeft? En waarom is juist deze uitgave de moeite waard voor de lezer van dit tijdschrift?

De aanleiding was simpel. In 2015 was het 150 jaar geleden dat Alice voor het eerst werd gepubliceerd (in 1865 dus) en een speciale herdenkingsuitgave lag voor de hand. De reden van het opnieuw uitbrengen van juist deze versie (met illustraties van de Spaanse, niet minder beroemde surrealist Salvador Dalí) is dat die eenmalige Random House-uitgave uit 1969 (oplage slechts 2700) inmiddels onvindbaar is en, indien toch gevonden, onbetaalbaar. Dalí is destijds gevraagd de laatste versie te illustreren die nog tijdens het leven van auteur Lewis Carroll (pseudoniem van wiskundige en fotograaf Charles Lutwidge Dodgson die leefde van 1832 tot 1898) is uitgekomen, want die wordt door kenners beschouwd als de meest authentieke en correcte versie. Dat juist Dalí werd gevraagd is wellicht vreemd (want die was op zijn minst controversieel) maar ook weer niet heel vreemd, aangezien Alice niet alleen een zeer gewaardeerd boek was in de kringen van de surrealisten, maar bovendien had Dalí in de jaren 1945 en 1946 acht maanden bij Disney gewerkt aan een film (*Destino*) in de periode dat men daar al bezig was met de filmversie (wellicht beïnvloed door Dalí?) van *Alice* die in 1951 uitkwam. Het getuigt overigens van enige lef *Alice* te willen illustreren, gezien de kwaliteit van de iconische illustraties van de eerste Alice-illustrator John Tenniel, maar aangezien Carroll weinig beschrijvend schrijft, had en heeft een illustrator altijd wel een grote mate van vrijheid. Dalí heeft er dan ook voor gekozen *Alice* op volstrekt eigen wijze te illustreren en je moet zelfs enige moeite doen om in de twaalf droomachtige platen elementen uit het boek te herkennen (behalve telkens een touwtje-springende Alice, het enige karakter trouwens dat in elk van de twaalf hoofdstukken voorkomt) en dat maakt het bestuderen van de illustraties (een per hoofdstuk) een kijkervaring op zich.

De lezer van dit blad komt, zonder iets af te doen aan de meerwaarde die de wonderlijke illustraties van Dalí aan het boek geven, zeer aan zijn of haar trekken aangezien *Alice* vol staat met

taalspelletjes, pure logica en intrigerende puzzels (waar onlangs ook een zeer aanbevelenswaardig boek over is uitgekomen, namelijk *Alice in Wonderland & Lewis Carroll's Games & Puzzles: The Surprising Connection* door Carroll-expert Christopher Morgan) en het verhaal is sowieso — zelfs bij herlezing — raadselachtig, heerlijk multi-interpretabel, soms verbijsterend maar altijd vermakelijk. In de inleiding stipt wiskundige (en vriend van Dalí) Thomas Banchoff nog wat wiskundige elementen in het werk van Dalí aan en beschrijft Carroll-kenner Mark Burstein enkele verbanden tussen Carroll en Dalí.

Het is een prachtige uitgave (mooi zuurvrij stevig papier met harde kft en stofomslag) die mij vooral interessant lijkt voor de liefhebber van het werk van Dalí, want als je dat niet bent volstaat zo'n beetje elke andere versie natuurlijk ook. Voor de lezers die echt willen duiken in de wereld van (of achter) *Alice* vermeld ik nog de bijbel op dat terrein, namelijk *The Annotated Alice* door de onvolprezen, onlangs overleden Martin Gardner. *Joop van der Vaart*



Jan van de Craats

### **Een passie voor symmetrie**

*Epsilon Uitgaven, deel 78, 2014*

100 p., prijs € 20,00

ISBN 9789050411431

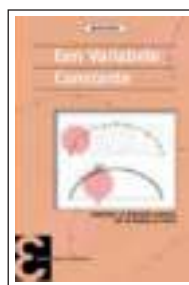
In de allereerste regel van de inleiding geeft de auteur al aan dat hij een passie voor symmetrie heeft. Impliciet geeft hij daarmee te kennen dat hij deze passie graag wil delen met de lezer van zijn boek. En dat gaat hij ook doen in dit boek. Hij begint met een korte opsomming van allerlei vormen van symmetrie die in de natuur en in het dagelijks leven voorkomen. Met voorbeelden uit de sportwereld, in de vorm van symmetrieën op diverse soorten ballen geeft Van de Craats daarmee al een beetje aan dat symmetrieën op een bol een prominente rol zullen gaan spelen in zijn boek. Verder laat hij aan de hand van fraaie foto's van vloer-, muur- en plafondversieringen in kerken, moskeeën en Romeinse huizen zien dat de ornamentale kunst een rijke bron van symmetrievormen is.

Van de Craats geeft aan dat het beoogde lezerspubliek niet uitsluitend bestaat uit wiskundigen, maar ook uit mensen met goede kennis van vwo-wiskunde en interesse in het onderwerp. Daarom zijn er in zijn boek bewust geen formules (met uitzondering van de codenamen) gebruikt en richt de auteur zich op het beschrijven van veertig soorten symmetrieën aan de hand van patronen. Op een zeer gestructureerde wijze worden de verschillende soorten symmetrieën beschreven die voorkomen in rosetpatronen, driedimensionale voorwerpen, patronen op een boloppervlak, strook- en behangpatronen. Een interessant aspect hierbij is dat diverse overeenkomsten in symmetrievormen van verschillende objecten duidelijk beschreven worden, zoals bijvoorbeeld de symmetrie-overeenkomsten tussen bol- en strookpatronen, tussen platonsche veelvlakken en die tussen 3D-voorwerpen en bolpatronen. Een cruciaal onderdeel van het boek is dat er voor elke soort symmetrie een identificatiecode (hier codenaam genoemd) aan-

gegeven wordt die het type van de symmetrie beschrijft. Tevens wordt er per soort patroon een duidelijke en handzame procedure vermeld waarmee de lezer de codenaam kan bepalen.

Het laatste deel (en hoogtepunt) van het boek is gewijd aan het classificeren van de tegel- en behangpatronen. De classificatie van de 17 verschillende behangpatronen wordt op een zeer doordachte wijze (met behulp van 'dolfijntjes') geïllustreerd, zodat de lezer zich (na voldoende oefening) de naamgeving eigen kan maken. Dat is ook wel nodig wanneer de lezer uitgedaagd wordt om de fraai versierde patronen op de foto's van de ornamentale kunst zelf te classificeren, maar gelukkig staan de antwoorden in het boek vermeld.

*Theo Beelen*



Martin Kindt

### **Een Variabele Constante**

**Historische en didactische aspecten van de Analyse op school**

*Epsilon Uitgaven, deel 82, 2015*

viii + 423 p., prijs € 36,00

ISBN 9789050411530

De titel roept associaties op met parameters: constant, maar toch ook variabel. De spanning tussen het veranderlijke en het constante is een kenmerk van de analyse in zijn geheel. Maar de titel verwijst toch in de eerste plaats naar de analyse als onderdeel van het schoolvak wiskunde: een blijver, maar in een gedaante die nog wel geregeld zal veranderen.

In het eerste deel van het boek is de insteek historisch. Aan de opname van analyse in het leerplan in 1958 is vijftig jaar aan discussie voorafgegaan. Nadien zorgden wijzigende opvattingen over wiskundeonderwijs voor tienjaarlijkse bijstellingen in de aanpak van de analyse, ontwikkelingen waar Martin Kindt mee vorm aan gegeven heeft, als leraar en medewerker van het Freudenthal Instituut. Kritische kanttekeningen brengen ons al snel in de didactiek (en in de wiskunde zelf) en maken dit deel ook interessant voor wie minder direct geïnteresseerd is in de geschiedkundige aspecten van het wiskundeonderwijs.

In het tweede deel van het boek is de verhouding tussen het historische en het didactische omgekeerd. Nu zijn insteek en klemtoon in de eerste plaats didactisch en staat de historische ontwikkeling, nu van de analyse zelf, en meer bepaald van de voorlopers ervan, ten dienste van deze didactische invalshoek.

Martin Kindt vraagt meer aandacht voor de begripsmatige aspecten van de analyse, wat in dit domein niet vanzelfsprekend is. Omdat de concepten moeilijk zijn, wordt de aandacht gemakkelijk verlegd naar berekeningen, en wordt het eerder calculus dan analyse. Nadruk op de relevantie van wiskunde voor vervolgstudies en andere disciplines, kan leraren (verkeerdelijk!) ertoe verleiden om zoveel mogelijk stof te behandelen en daarbij de diepgang uit het oog te verliezen (terwijl het opbouwen van een rijk conceptueel inzicht precies ook van belang is voor vervolgstudies en bij het toepassen van wiskunde).

Het boek biedt een schat aan materiaal dat deze klemtoon op het begripsmatige helpt stofferen. Analyse is ontstaan als mid-

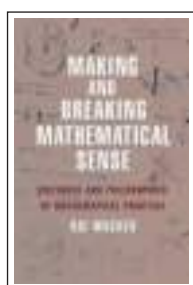
del om meetkundige problemen en problemen rond beweging op te lossen. Aan elk van beide contexten is een hoofdstuk uit het tweede deel gewijd. De andere hoofdstukken gaan over discrete analyse, extremumproblemen, de raaklijn en (conceptuele aspecten verbonden met) de algebraïsche kant van de analyse. Een deel van de uitgewerkte wiskundige thema's sluiten nauw aan bij de schoolwiskunde, terwijl andere verder van de klaspraktijk afstaan.

Iemand die zo'n lange tijd nagedacht heeft over de didactiek van de analyse heeft natuurlijk ook veel te vertellen, los van de uitgewerkte wiskundige thema's. Het boek is dan ook doorspekt met beschouwingen en opinies.

Martin Kindt heeft een inspirerend boek geschreven, dat je 'algemene cultuur' in verband met analyseonderwijs zeker zal verrijken.

Johan Deprez

*Recent verschenen publicaties. Als u een van deze boeken wilt bespreken of als u suggesties heeft voor andere boeken voor deze rubriek, laat dit dan per e-mail weten aan [reviews@nieuwarchief.nl](mailto:reviews@nieuwarchief.nl).*



Roi Wagner

**Making and Breaking Mathematical Sense**  
Histories and Philosophies of Mathematical Practice

Princeton University Press, 2017  
ISBN 9781400883783  
[press.princeton.edu/titles/10909.html](http://press.princeton.edu/titles/10909.html)



Mircea Pitici

**The Best Writing on Mathematics 2016**

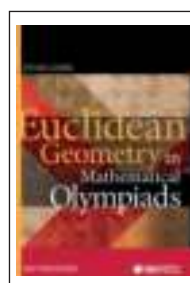
Princeton University Press, 2017  
ISBN 9781400885602  
[press.princeton.edu/titles/10953.html](http://press.princeton.edu/titles/10953.html)



Daniel Huybrechts

**Lectures on K3 Surfaces**

Cambridge University Press, 2016  
ISBN 9781316797570  
[www.cambridge.org/9781316797570](http://www.cambridge.org/9781316797570)



Evan Chen

**Euclidean Geometry in Mathematical Olympiads**

Cambridge University Press, 2016  
ISBN 9780883858394  
[www.cambridge.org/9780883858394](http://www.cambridge.org/9780883858394)



Ghislaine Gueudet et al.

**Transitions in Mathematics Education**

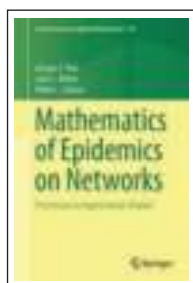
Springer, 2016  
ISBN 978331931621-5  
[www.springer.com/9783319316215](http://www.springer.com/9783319316215)



Keith Devlin

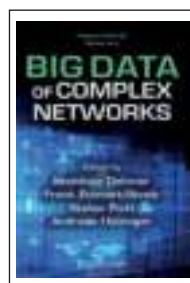
**Finding Fibonacci**  
The Quest to Rediscover the Forgotten Mathematical Genius Who Changed the World

Princeton University Press, 2017  
ISBN 9781400885534  
[press.princeton.edu/titles/10950.html](http://press.princeton.edu/titles/10950.html)



**Mathematics of Epidemics on Networks**

Springer, 2017  
ISBN 9783319508047  
[www.springer.com/9783319508047](http://www.springer.com/9783319508047)



Matthias Dehmer, Frank Emmert-Streib, Stefan Pickl, Andreas Holzinger (eds.)

**Big Data of Complex Networks**

CRC Press, 2016  
ISBN 9781498723619  
[www.crcpress.com/9781498723619](http://www.crcpress.com/9781498723619)

# 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. (To compete for the book token you should have a postal address in The Netherlands.)

Please send your submission by e-mail (LaTeX is preferred), including your name and address to [problems@nieuwarchief.nl](mailto:problems@nieuwarchief.nl).

The deadline for solutions to the problems in this edition is 1 June 2017.

---

## Problem A

Reconstruction. Suppose you get to know the  $n$  midpoints of the  $n$  edges of a polygon. Can you determine the polygon?

---

## Problem B

Large is odd. Let  $G$  be a graph with vertices  $V$  and edges  $E$ . A subset  $U \subset V$  is called large if every vertex that is not in  $U$  has a neighbor in  $U$ . Prove that the number of large subsets is odd.

---

## Problem C

Meanders. On our way to this problem, we first meander through a Martin Gardner eye test:



One of the two spirals in the illustration consists of a single piece of rope that has its ends joined. The other spiral consists of two separate pieces of rope, each with joined ends. Can you identify which is which using only your eyes?

Now we define a meander. Let  $n$  be an even number. Consider the integers 1 to  $n$  in the complex plane and connect them by semicircles centered around  $\frac{n+1}{2}$  in the upper half plane. Let  $n = a + b$  for even numbers  $a$  and  $b$ . Connect the first  $a$  integers by semicircles around  $\frac{a+1}{2}$  in the lower half plane. Similarly, connect the final  $b$  integers by semicircles around  $a + \frac{b+1}{2}$  in the lower half plane. The resulting curve, or set of curves, is a meander. For which  $a$  and  $b$  is it connected?

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](mailto:problems@nieuwarchief.nl)  
[www.nieuwarchief.nl/problems](http://www.nieuwarchief.nl/problems)

---

**Edition 2016-3** We received solutions from Raymond van Bommel, Rik Bos, Alexandros Efthymiadis, Alex Heinis, Thijmen Krebs, Toshihiro Shimizu, Djurre Tijsma and Traian Viteam.

---

## Problem 2016-3/A (folklore)

Show that a group  $G$  is torsion-free if and only if for all integers  $n \geq 2$  and finite subsets  $S, T \subseteq G$  with  $\#S = \#T = n$  we have  $\#\{st : s \in S, t \in T\} > n$ .

**Solution** We received solutions from Raymond van Bommel, Rik Bos, Alexandros Efthymiadis, Alex Heinis, Thijmen Krebs, Toshihiro Shimizu and Djurre Tijsma. The book token is

awarded to Rik Bos, the solution of whom the following is based on; the main idea below appeared in every solution received.

If  $G$  has a torsion element  $x$ , then  $S = T = \langle x \rangle$  have the property that  $\#(ST) = \#S = \#T$ . On the other hand, let  $n \geq 2$ , and let  $S, T$  be subsets of  $G$  with  $\#S = \#T = n$ , such that  $\#(ST) = n$ . Then, for every  $s \in S$ , we have  $sT \subseteq ST$ , and both sides have the same cardinality  $n$ , so  $sT = ST$  for all  $s \in S$ . Let  $s_1, s_2 \in S$  be distinct elements, and let  $a = s_1^{-1}s_2$ , so that  $a \neq 1$ . Then  $aT = s_1^{-1}s_2T = s_1^{-1}ST = s_1^{-1}s_1T = T$ , so  $a^kT = T$  for all  $k \geq 0$ . Hence, for any  $t \in T$ , we have  $\{a^kt : k \in \mathbb{Z}\} \subseteq T$  and therefore that  $a$  is torsion of order at most  $n$ , which contradicts  $G$  being torsion-free.

---

**Problem 2016-3/B** (proposed by Hendrik Lenstra)

Show that for all groups  $G$  the commutator subgroup  $[G, G] = \langle xyx^{-1}y^{-1} : x, y \in G \rangle$  of  $G$  has order at most 2 if and only if every conjugacy class in  $G$  has at most 2 elements.

**Solution** We received solutions from Raymond van Bommel, Rik Bos, Alexandros Efthymiadis, Alex Heinis, Thijmen Krebs, Toshihiro Shimizu, Djurre Tijsma and Traian Viteam. The book token is awarded to Alex Heinis, the solution of whom the following is mainly based on, with some elements based on other submissions; most of the received solutions are similar.

First suppose that the order of  $[G, G]$  is at most 2. Then there exists a unique  $a \in G$  such that  $[G, G] = \{1, a\}$ . Let  $x \in G$ . Then we have for all  $g \in G$  that  $gxg^{-1} = [g, x]x \in \{x, ax\}$ . Hence every conjugacy class in  $G$  has cardinality at most 2.

Now suppose that every conjugacy class in  $G$  has cardinality at most 2. Then for all  $x \in G$  there exists a unique  $\theta(x) \in G$  such that  $\{[g, x] : g \in G\} = \{1, \theta(x)\}$ , so that the conjugacy class of  $x$  is  $\{x, \theta(x)x\}$ . We see that  $\theta$  defines a map from  $G$  to  $[G, G]$  such that its image generates  $[G, G]$ . So it suffices to show that the image of  $\theta$  has cardinality at most 2, and that the image consists of elements of order at most 2; note that the image of  $\theta$  always contains 1.

First note that if  $x, y \in G$  do not commute, that then  $\theta(x)$  and  $\theta(y)$  are both non-trivial, since  $xyx^{-1} \neq x$  implies that  $xyx^{-1} = \theta(x)x$ , and  $xyx^{-1} \neq y$  implies that  $xyx^{-1} = \theta(y)y$ . Next, we show that if  $x, y \in G$  do not commute, that then  $\theta(x) = \theta(y)$ . Note that  $xy$  commutes with neither of  $x, y$ . Hence we have

$$\theta(x) = [xy, x] = [x, xy]^{-1} = \theta(xy)^{-1} = [y, xy]^{-1} = [xy, y] = \theta(y)$$

and

$$\theta(x) = [y, x] = [x, y]^{-1} = \theta(y)^{-1},$$

so in particular,  $\theta(x) = \theta(y)$  is of order 2.

Note that for  $x \in G$  we have  $\theta(x) = 1$  if and only if  $x \in Z(G)$ , so  $\theta$  is constant on  $Z(G)$ . So we are done once we show that  $\theta$  is constant on  $G - Z(G)$  of value of order 2. Let  $x, y \in G - Z(G)$ . Then the centralisers  $C_x, C_y$  of  $x, y$  respectively are proper subgroups of  $G$ , therefore  $C_x \cap C_y \neq G$ . Hence there exists  $g \in G$  that commutes with neither of  $x, y$ . We deduce that  $\theta(x) = \theta(g) = \theta(y)$ , and that this element is of order 2. Therefore  $\theta$  is constant on  $G - Z(G)$  of value of order 2, as required.

---

**Problem 2016-3/C** (proposed by Carlo Pagano and Mima Stanojkovski)

A subgroup  $H$  of a group  $G$  is said to be *solitary* if no other subgroup of  $G$  is isomorphic to  $H$ . A group  $G$  is said to be *totally solitary* if all of its subgroups are solitary. Show that a group  $G$  is totally solitary if and only if it is isomorphic to a subgroup of  $\mathbb{Q}/\mathbb{Z}$ .

**Solution** We received solutions from Rik Bos, Alex Heinis, Thijmen Krebs, Toshihiro Shimizu and Djurre Tijsma. The book token goes to Djurre Tijsma. All received proofs of the fact that any subgroup of  $\mathbb{Q}/\mathbb{Z}$  is totally solitary are similar, and the first part of the solution is based on these. The second part of following solution is based for the most part on the submission of Raymond van Bommel.

# Oplossingen

## | Solutions

First we show that  $\mathbb{Q}/\mathbb{Z}$  is totally solitary, thereby (by definition of totally solitary) showing that all of its subgroups are. Note that  $\mathbb{Q}/\mathbb{Z}$  is torsion, and that for any positive integer  $N$ ,  $\mathbb{Q}/\mathbb{Z}[N] = \frac{1}{N}\mathbb{Z}/\mathbb{Z}$ , which is cyclic of order  $N$ . So let  $H_1, H_2$  be two isomorphic subgroups of  $\mathbb{Q}/\mathbb{Z}$ , and let  $\phi: H_1 \rightarrow H_2$  be an isomorphism. Then, for any  $h \in H_1$ , we see that  $\phi(h)$  has the same order as  $h$ , say  $N$ . Hence  $h = a\phi(h)$  for some integer  $a$  coprime to  $N$ , so  $h \in H_2$  as well, showing that  $H_1 \subseteq H_2$ . The same argument shows that  $H_2 \subseteq H_1$ , so  $H_1 = H_2$ , thereby showing that  $\mathbb{Q}/\mathbb{Z}$  (and each of its subgroups) is totally solitary.

Now suppose that  $G$  is totally solitary. If  $G$  were to contain an element  $x$  of infinite order, then  $\langle x^2 \rangle \subset \langle x \rangle$  are isomorphic subgroups of  $G$  that are not equal, which contradicts  $G$  being totally solitary. Hence  $G$  is torsion. Moreover, for any positive integer  $N$  there is at most one cyclic subgroup of  $G$  of order  $N$ , in other words, there are either 0 or  $\phi(N)$  elements of  $G$  of order  $N$ , where  $\phi$  is the Euler totient function.

Now take any finite subgroup  $H$  of  $G$ , and let  $N$  be its order. Since elements of  $H$  must have order dividing  $N$ , it follows that  $N = \#H \leq \sum_{d|N} \phi(d) = N$ . Therefore, equality must hold, and  $H$  contains  $\phi(N)$  elements of order  $N$ , i.e.  $H$  is cyclic.

We use this to show that  $G$  is abelian. Let  $x, y \in G$  be of orders  $M, N$ , respectively. Then  $xyx^{-1}$  has the same order as  $x$ , and therefore  $xyx^{-1}$  generates the same subgroup of  $G$  as  $x$ . In other words, there exists some integer  $e$  with  $\gcd(e, M) = 1$  for which  $xyx^{-1} = x^e$ . Now every element of  $\langle x, y \rangle$  can be written in the form  $x^i y^j$  for some  $i = 0, 1, \dots, M-1$  and  $j = 0, 1, \dots, N-1$ , so  $\langle x, y \rangle$  is finite, hence cyclic by the above, and therefore  $x$  and  $y$  commute.

Now that we have shown that  $G$  is abelian, we will write the group operation on  $G$  additively from now on.

Let us construct an injective homomorphism  $G \rightarrow \mathbb{Q}/\mathbb{Z}$ . First, pick an enumeration  $p_1, p_2, p_3, \dots$  of the prime numbers, and let  $M_n = \prod_{i=1}^n p_i^{n_i}$  for  $n \geq 0$  (so  $M_0 = 1$ ). Moreover, let  $H_n = G[M_n]$ , so that every  $H_n$  is finite and cyclic, say of order  $N_n$ . Then  $H_m < H_n$  for  $m < n$ , so  $N_m | N_n$  for  $m < n$ , and  $G = \bigcup_{n=1}^{\infty} H_n$ .

Now we define recursively for every  $n \geq 0$ , an injective homomorphism  $\phi_n: H_n \rightarrow \mathbb{Q}/\mathbb{Z}$  such that for  $i < n$  and  $x \in H_i$  we have  $\phi_i(x) = \phi_n(x)$ . This suffices, since we can define an injective homomorphism  $\phi: G \rightarrow \mathbb{Q}/\mathbb{Z}$  by  $x \mapsto \phi_i(x)$  for any  $i$  such that  $x \in H_i$ ; such an  $i$  exists, and the value doesn't depend on the choice of  $i$ , by the above.

For  $n = 0$ , we define  $\phi_0: H_0 = 0 \rightarrow \mathbb{Q}/\mathbb{Z}$  to be the zero homomorphism. Now assume that for  $n \geq 1$  and  $i = 0, 1, \dots, n-1$ , injective homomorphisms  $\phi_i: H_i \rightarrow \mathbb{Q}/\mathbb{Z}$  are given such that for  $j < i$  and  $x \in H_j$  we have  $\phi_j(x) = \phi_i(x)$ . Let  $h_{n-1}$  be a generator of  $H_{n-1}$ , and let  $h_n$  be a generator of  $H_n$  such that  $\frac{N_n}{N_{n-1}}h_n = h_{n-1}$ . Then define  $\phi_n: H_n \rightarrow \mathbb{Q}/\mathbb{Z}$  to be the homomorphism sending  $h_n$  to any element  $h'_n \in \mathbb{Q}/\mathbb{Z}$  such that  $\frac{N_n}{N_{n-1}}h'_n = \phi_{n-1}(h_{n-1})$  if  $\phi_{n-1}(h_{n-1}) \neq 0$ , and to  $\frac{N_n-1}{N_n}$  otherwise. This is injective as a generator of a cyclic group of order  $N_n$  is always sent to an element of order  $N_n$  by assumption. Moreover, for all  $x \in H_{n-1}$ , we have  $\phi_{n-1}(x) = \phi_n(x)$ , so by assumption, for all  $i < n$  and  $x \in H_i$ , we have  $\phi_i(x) = \phi_n(x)$  as well, as required.

