Bert Zwart
Research Group Stochastics
CWI, Amsterdam
bert.zwart@cwi.nl

Onderzoek Vidi-project

De macroscoop

In 2008 ontvangt Bert Zwart een Vidi-subsidie voor het project 'High dimensional service systems'. Dit fundamentele onderzoek richt zich op analyse van complexe systemen en procedures waarin zoveel factoren een rol spelen dat traditionele analysemethoden onbruikbaar zijn. Daarbij kan bijvoorbeeld gedacht worden aan de bezetting in een callcenter met honderden telefonisten. Dat soort systemen valt niet door te rekenen. Als je meer ongeduldige klanten hebt, dan verandert ook het aantal benodigde medewerkers.

Onlangs liep mijn periode als Vidi-laureaat af, een mooie periode waarin ik dankzij NWO in staat werd gesteld me te richten op enkele fundamentele vragen in mijn onderzoeksrichting. Hoewel mijn Vidi-beurs werd toegekend in augustus 2008, begonnen de voorbereidingen achteraf gezien al in 2005. Toen liep mijn Veni-beurs af, en besloot ik om geen Vidi-voorstel in te dienen. In plaats daarvan koos ik voor een verblijf in het buitenland.

Ik kwam als associate professor terecht bij het Georgia Institute of Technology in Atlanta. Daar had ik het naar mijn zin en kreeg ik in februari 2008 ook een vaste aanstelling (tenure). Toch besloot ik terug te keren naar Nederland. Ik had, om alle opties open te houden, al een Vidi-voorstel ingediend. Dit voorstel was geïnspireerd door diverse ontwikkelingen in mijn vakgebied die zich in die tijd grotendeels in de Verenigde Staten afspeelden, en waar ik in Atlanta ook aan werkte. Op 1 augustus 2008, mijn eerste werkdag op het CWI, ontving ik het heuglijke nieuws dat mijn Vidi-voorstel was toegekend.

Waar ging het voorstel over? Op de NWO-website staat nog steeds de volgende accurate informatie bij de Vidi-toekenningen in 2008:

Grootschalige bedieningssystemen onder de macroscoop
Dr. A.P. (Bert) Zwart (m) 16-04-1974, CWI - Wiskunde
Wiskundige modellen voor computer- en communicatienetwerken, callcenters en productieprocessen zijn vaak hoogdimensionaal en daarom moeilijk analyseerbaar. Dit onderzoeksproject beoogt de complexiteit van dergelijke modellen te reduceren door zich te richten op het macroscopische gedrag.

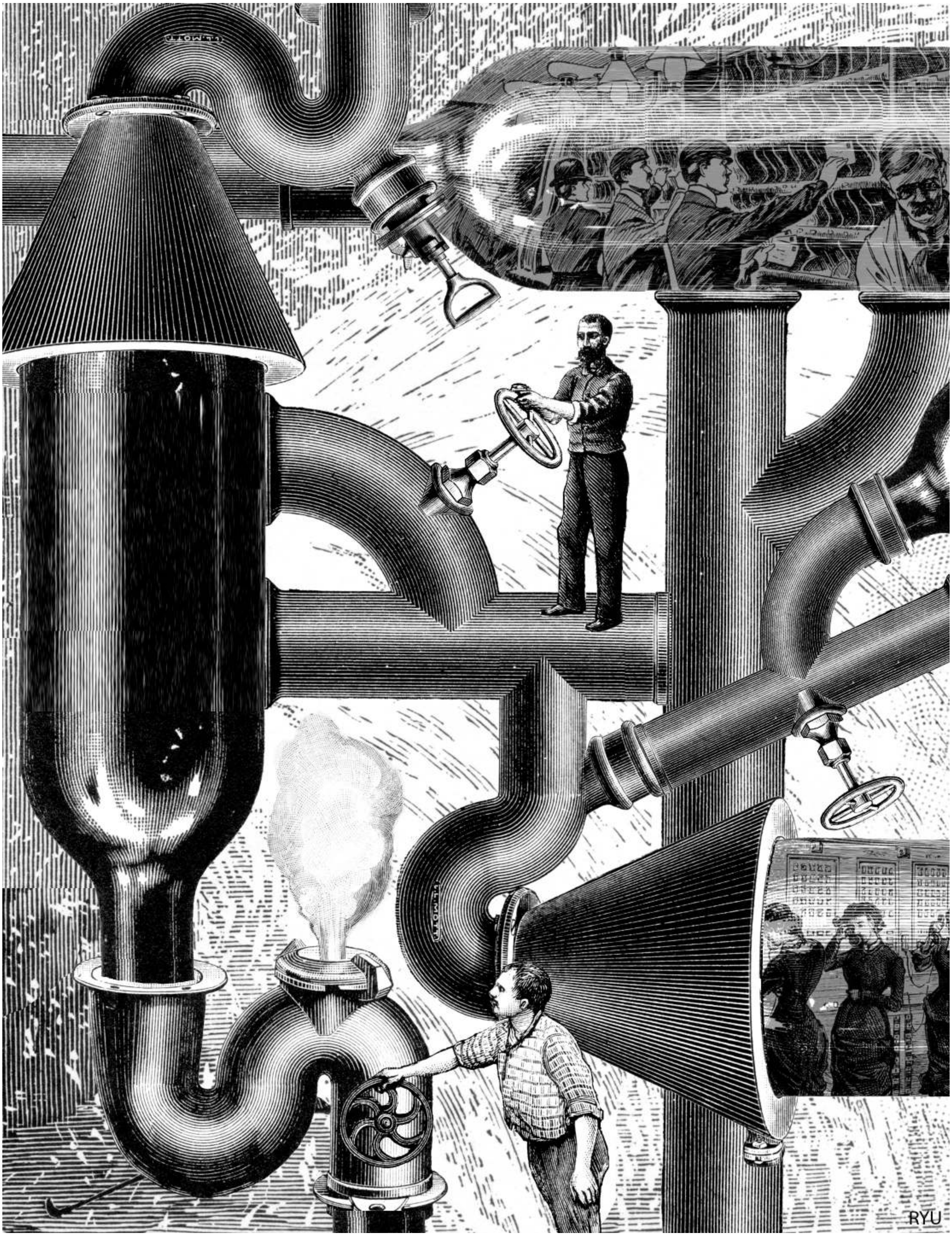
Belangrijke standaardwerken binnen mijn vakgebied zijn [2] en [8], maar de technieken beschreven in die standaardwerken zijn vaak alleen geschikt voor modellen op een eindig-dimensionale toestandsruimte. Een belangrijke categorie van modellen welke modernere technieken vereisen zijn wachtrijen waarbij het aantal bediendes zeer groot is, hetgeen relevant is voor bijvoorbeeld callcenters en cloud computing. Hoewel ik hier binnen mijn Vidi-project ook aan gewerkt heb, zal ik me in dit artikel verder richten op een ander model. Dit model beschrijft congestie in communicatienetwerken, en vereist ook een oneindig-dimensionale toestandsruimte. De resultaten zijn beschreven in [6], alsmede in Hoofdstuk 3 van het proefschrift van Maria Remerova [5].

Wanneer je het internet gebruikt voor het bekijken van een video, het versturen van een file, of het voeren van een telefoongesprek, dan maak je doorgaans gebruik van meerdere knooppunten, ook wel links genoemd. Deze links kunnen bottlenecks vormen, waardoor vertraging op kan treden. Dit kan leiden tot ongewenst gedrag, zoals het afbreken van een sessie door ongeduldige gebruikers. Om deze reden is het belangrijk om inzichten te verwerven in het gedrag van communicatienetwerken. Dit soort inzichten kunnen worden ontwikkeld met behulp van wiskundige modellen, hetgeen de motivatie is voor het gedane onderzoek.

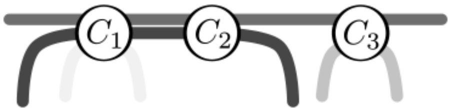
Modelbeschrijving

In Figuur 1 is elke link in een netwerk, die je in traditionele wachtrij-terminologie bediende zou kunnen noemen, weergegeven door een cirkel. Een communicatiesessie (denk bijvoorbeeld aan een datatransmissie of telefoongesprek) in het netwerk volgt een bepaalde route die uit meerdere links bestaat. Zo'n sessie zou je in traditionele wachtrij-terminologie klant kunnen noemen. Er zijn I routes en J links. A is een $0-1$ matrix waarbij $A_{ji}=1$ als route i gebruik maakt van link j. De totale hoeveelheid data die link j per tijdseenheid kan verwerken is C_j.

Het aantal klanten dat gebruikmaakt van route i geven we weer met z_i, en we noemen $z=(z_1,\dots,z_I)$ de populatievector. Ge-



Illustratie: Ryu Tajiri



Figuur 1 Een netwerk met drie links en vier routes.

geven een populatie z definiëren we $\Lambda(z)$ als de vector met bedieningssnelheden per route, deze worden ook wel bandbreedtes genoemd. Elke klant op route i wordt bediend met snelheid $\Lambda_i(z)/z_i$ — klanten op dezelfde route worden bediend met dezelfde snelheid. Daarnaast hebben klanten op route i ook een bovengrens m_i op de individuele bandbreedte.

Verder nemen we aan dat een klant op route i een nut $U_i(x)$ toekent aan een bedieningssnelheid x, en dat de allocatie van toegewezen bandbreedtes $\Lambda(z)$ in elke toestand z wordt bepaald door het totale, gezamenlijk nut te maximaliseren. De functie U_i voldoet aan enkele regulariteitseigenschappen (stijgend, tweemaal differentieerbaar, strikt concaaf). De vector $\Lambda(z)$ wordt nu gedefinieerd als de oplossing van het probleem

$$\max \sum_i z_i U_i(\Lambda_i/z_i)$$

onder de condities $A\Lambda \leq C$ en $\Lambda_i/z_i \leq m_i$, $i = 1, \ldots, I$.

Een belangrijke eigenschap van de functie $\Lambda(\cdot)$ is dat deze differentieerbaar is in elke richting. Dit kan worden aangetoond met behulp van algemene theorie voor continue optimaliseringsproblemen zoals is beschreven in [1], en uitgewerkt in [4].

Dit allocatiemechanisme lijkt op het eerste gezicht ver te staan van de werkelijkheid maar, indien U_i goed wordt gekozen, blijkt het toch diverse eigenschappen te hebben die de werkelijkheid goed beschrijven. Voor meer informatie en referenties verwijs ik naar [9].

We beschouwen dit model in stochastische dynamische context:

- Het aantal aankomsten van gebruikers/klanten bij route i tot en met tijdstip t is $E_i(t)$.
- De hoeveelheid werk per klant op route i is verdeeld als de stochast B_i.
- Elke klant heeft daarnaast een deadline D_i, die kan afhangen van B_i. Een klant vertrekt als zijn hoeveelheid werk is afgehandeld, of als zijn deadline is verstreken.
- We nemen aan dat klanten die continu op hun maximale snelheid bediend worden, niet ongeduldig zijn: $D_i \geq B_i/m_i$.

Om de dynamiek van dit systeem te kunnen analyseren moeten we voor elk tijdstip en elke klant zowel de resterende hoeveelheid werk, alsmede de resterende tijd tot het verstrijken van de deadline bijhouden. Dit resulteert in een toestandsbeschrijving welke grafisch is weergegeven in Figuur 2.

Schaling en resultaten

Omdat er geen bovengrens op het aantal gebruikers in het netwerk is, is de toestand oneindig-dimensionaal, en is het soms nodig om een maatwaardig proces te analyseren. Het lijkt onmogelijk om prestatiematen voor dit soort processen, zoals de fractie klanten die vertrekt wegens het aflopen van hun deadline, te berekenen. Om de complexiteit te reduceren kijken we daarom naar het realistische geval waarbij de C_j's veel groter zijn dan de m_i's, en waarbij het aantal aankomsten per route ook hoog is. Formeel vervangen we C door Cr en $E_i(t)$ door $E_i(rt)$ en laten we r willekeurig groot worden. We nemen aan dat $E_i(rt)/r \to \nu_i t$ en bekijken het stochastische proces $Z^{(r)}(t)$ dat voor elk tijdstip t het totale aantal klanten per route weergeeft. Het blijkt nu dat, als $r \to \infty$,

$$\frac{1}{r} Z^{(r)}(t), \quad t \geq 0 \to z(t), t \geq 0,$$

voor een proces $z(\cdot)$ welke voldoet aan de vergelijking

$$z_i(t) = \nu_i \int_0^\infty P(D_i > t - u,$$
$$B_i > S_i(u,t))du, \quad i = 1, \ldots, I,$$

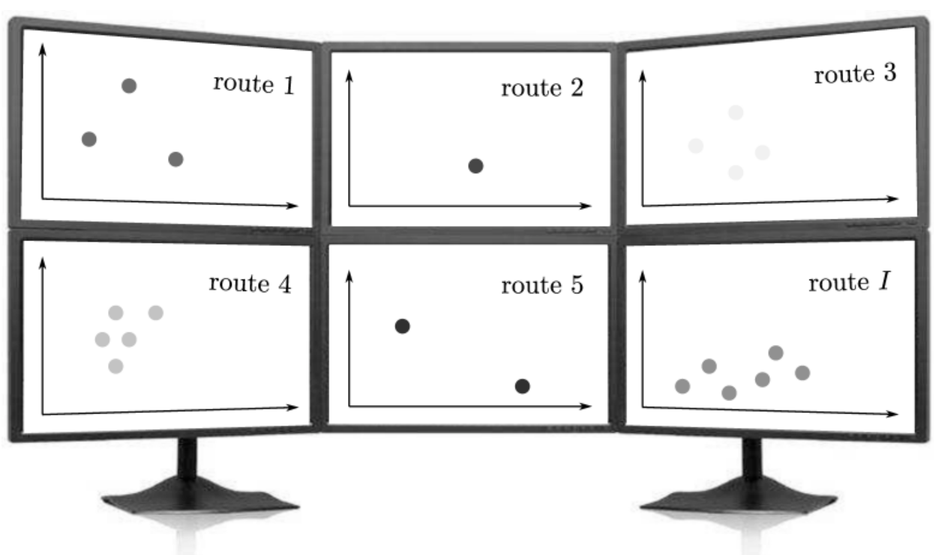
waarbij $S_i(u,t) = \int_u^t \Lambda_i(z(v))/z_i(v)dv$. We zien dat het limietproces $z(t), t \geq 0$, deterministisch is, maar volledig afhangt van de gezamenlijke verdeling van B_i en D_i. Het type vergelijking dat $z(t), t \geq 0$, karakteriseert is equivalent aan een functionaaldifferentiaalvergelijking. De vergelijking voor $z(t)$ is intuïtief als volgt te verklaren: vloeistof van type i komt op tijdstip u het systeem binnen met intensiteit ν_i. De fractie van die vloeistof die nog steeds aanwezig is op tijdstip t is $P(D_i > t - u; B_i > S_i(u,t))$, waarbij $S_i(u,t)$ de totale hoeveelheid bediening is per klant van type i in het interval $[u,t]$.

Ten slotte convergeert $z(t)$ naar een vector z^* als $t \to \infty$. Deze vector voldoet aan de vergelijking

$$z_i^* = \nu_i E[\min\{D_i, B_i z_i^*/\Lambda_i(z^*)\}], i = 1, \ldots, I.$$

Ook deze vergelijking heeft een fysische interpretatie. Volgens de wet van Little is het gemiddeld aantal klanten in een wachtrijsysteem (z_i) in evenwicht gelijk aan het gemiddeld aantal aankomsten per tijdseenheid (ν_i), vermenigvuldigd met de gemiddelde verblijftijd van een klant. In ons geval is de verblijftijd het minimum van de deadline D_i en de potentiële verblijftijd. In evenwicht zal de gemiddelde bedieningssnelheid gelijk blijven; deze is gelijk aan $\Lambda_i(z)/z_i$, waarmee de potentiële verblijftijd gelijk is aan $B_i/(\Lambda_i(z)/z_i)$.

De vraag is of de vergelijking waaraan z^* voldoet een unieke oplossing heeft. Dit blijkt het geval te zijn, al duurde het even voor het bewijs gevonden was. Het is bijvoorbeeld niet eenvoudig om een argument gebaseerd op contracties te construeren. Het door ons ge-



Figuur 2 Snapshot van de gedetailleerde toestandsbeschrijving. Elk punt geeft de resterende hoeveelheid werk en de resterende tijd tot de deadline weer voor een klant, en verschuift richting het zuidwesten met snelheid $(z_i/\Lambda_i(z), 1)$.

vonden bewijs is algoritmisch en geeft ook een procedure om de vector z^* uit te rekenen. Het idee is om eerst $\Lambda^* = \Lambda(z^*)$ te karakteriseren, waarmee z^* kan worden gevonden door I eendimensionale bisecties. Definieer $g_i(x) = \nu_i E[\min\{xD_i, B_i\}]$ en laat G_i een functie zijn zodanig dat $G_i'(x) = U_i'(g_i^{-1}(x))$. Dan is Λ^* de unieke oplossing van het concave optimaliseringsprobleem

$$\min_{\Lambda} \sum_i G_i(\Lambda_i)$$

onder de voorwaarden $\Lambda \leq \rho$ en $A\Lambda \leq C$, met $\rho_i = \nu_i E[B_i]$. Dankzij de structuur van dit optimaliseringsprobleem (de functie G_i is strikt concaaf) volgt niet alleen uniciteit van z^*, maar is het ook mogelijk om Λ^* te vinden in polynomiale tijd (als functie van I en J).

Samenvattend zien we dat het macroscopisch gedrag van dit oneindig-dimensionale stochastische systeem te beschrijven valt door een functionaaldifferentiaalvergelijking, die een uniek invariant punt heeft. Dit invariante punt kan gekarakteriseerd worden door middel van de oplossing van een concaaf optimaliseringsprobleem. Cruciaal is dat de functie $\Lambda(\cdot)$ voldoet aan de nodige regulariteitseigenschappen. Deze kunnen worden afgeleid met technieken uit de continue optimalisering.

De convergentie van $Z^{(r)}(t)/r, t \geq 0$, naar $z(t), t \geq 0$, volgt uit de meer gedetailleerde maatwaardige beschrijving van het systeem, gebaseerd op Figuur 2. Om relatieve compactheid van het rijtje processen $Z^{(r)}(\cdot)/r$ vast te stellen is het belangrijk aan te tonen dat het aantal punten in L-vormige verzamelingen in Figuur 2 afgeschat kan worden in termen van de dikte van L. Hierdoor is het dan niet mogelijk dat veel klanten vertrekken in een kort tijdsbestek, waarmee het systeem ook geen oscillaties kan vertonen. Dit impliceert dan uiteindelijk de gewenste compactheidseigenschap. De benodigde afschattingen volgen onder meer uit de theorie van empirische processen, welke terug te vinden is in [7].

Excursies

Bovengenoemde studie is illustratief voor mijn Vidi-onderzoeksproject. Een ander resultaat waar ik trots op ben is verkregen in samenwerking met Amaury Lambert en mijn voormalige postdoc Florian Simatos. We hebben een nieuwe methode ontwikkeld voor het afleiden van convergentie van stochastische processen, gebruikmakend van excursietheorie. Het idee is dat wachtrijen vaak relatief eenvoudig te beschrijven zijn zolang het systeem niet leeg raakt. De zogenaamde bezige perioden zijn voorbeelden van wat in de kansrekening excursies worden genoemd. Om nu aan te tonen dat een rijtje stochastische processen $Z^{(r)}$ convergeert naar een stochastisch proces Z blijkt het voldoende te zijn dat (i) het rijtje $(Z^{(r)})$ relatief compact is; (ii) de excursies van $Z^{(r)}$ zwak convergeren naar de excursies van Z, en (iii) Z gekarakteriseerd wordt door zijn excursies. Voor meer details verwijs ik naar [3]. Dit resultaat is krachtig genoeg om één van de moeilijkste problemen in mijn Vidi-voorstel op te lossen, een probleem dat ik tijdens het schrijven van het voorstel nogal speculatief vond en ik ergens helemaal op het eind noemde. Ik prijs me gelukkig met het gegeven dat Simatos een uitstekende postdoc was die bereid was om het risico te nemen aan dit probleem te werken.

Tot slot

Terugziend heeft de Vidi-beurs me in staat gesteld enkele talentvolle jonge onderzoekers aan te stellen, en me te richten op uitdagende open problemen in mijn vakgebied. Hierbij kon ik ook de tijd nemen om me te verdiepen in onderwerpen uit aangrenzende vakgebieden. Ik ben NWO daar zeer dankbaar voor.

Referenties

1 J. Bonnans en A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, 2001.

2 H. Chen en D. Yao, *Stochastic Networks: Performance, Asymptotics & Optimization*, Springer, 2003.

3 A. Lambert, F. Simatos en B. Zwart, Scaling limits via excursion theory: Interplay between Crump–Mode–Jagers branching processes and Processor-Sharing queues, *Annals of Applied Probability* 23 (2013), 2357–2381.

4 J. Reed en B. Zwart, Fluid and diffusion approximations for Markovian bandwidth sharing networks with rate constraints, te verschijnen in *Operations Research* (2014).

5 M. Remerova, *Fluid Limit Approximations of Stochastic Networks*, proefschrift, VU, Amsterdam, 2014.

6 M. Remerova, J. Reed en B. Zwart, Fluid limits for bandwidth-sharing networks with impatience, te verschijnen in *Mathematics of Operations Research* (2014).

7 A. van der Vaart en J. Wellner, *Weak Convergence and Empirical Processes*, Springer, 1996.

8 W. Whitt, *Stochastic Process Limits*, Springer, 2002.

9 Y. Yi en M. Chiang, Stochastic network utility maximization – A tribute to Kelly's paper published in this journal a decade ago, *European Transactions on Telecommunications* 19 (2008), 421–442.