Luistert u tussen Kerst en Nieuwjaar ook altijd naar de Top 2000? Heeft u zich wel eens gerealiseerd dat deze ultieme kerstbeleving ook aanleiding kan zijn tot wiskundige eindejaarsoverwegingen? Hoe kun je bijvoorbeeld ketens van songtitels maken waarvan het laatste woord van de vorige titel het eerste woord van de volgende vormt? Hierop gaan **Robert Kooij**, **Erik Boertjes** en **Joost de Wit** nader in, waarbij het P-NP-probleem om de hoek komt kijken.

Ketens van songtitels in de Top 2000

Introductie

Als minister van Onderwijs, Cultuur en Wetenschap in het kabinet Balkenende IV maakte Ronald Plasterk tijdens zijn ambtsperiode van iedereen die hem bezocht in zijn kantoor één foto. In totaal heeft hij op deze manier 670 fotoportretten gemaakt van mensen uit onder andere politiek, media, cultuur, wetenschap en journalistiek. Hierdoor geïnspireerd zijn de auteurs een enigszins vergelijkbaar project begonnen. Bij elk contact met een klant, opdrachtgever of collega, werd namelijk de volgende vraag gesteld: "Wat zijn je drie favoriete liedjes van The Beatles?"

Eén specifieke fout die werd gemaakt met een songtitel vormde de motivatie voor het onderzoek dat beschreven staat in dit artikel. De foutieve songtitel luidt *Get Back in the U.S.S.R.* De collega die met deze suggestie kwam, heeft hier abusievelijk twee songtitels aan elkaar gekoppeld, te weten *Get Back* (een single uit 1969) en *Back in the U.S.S.R.* (het openingsnummer van het in 1968 uitgebrachte album *The Beatles*, beter bekend als *The White Album*). Dit bracht de eerste auteur op het idee om op zoek te gaan naar ketens van songtitels. Dit laat zich eenvoudig vertalen naar een probleem uit de grafentheorie. Uitgangspunt daarbij is een bepaalde verzameling liedjes, bijvoorbeeld alle liedjes die zijn opgenomen door The Beatles. Volgens Wikipedia[1] hebben The Beatles 306 liedjes opgenomen. Elke songtitel die uit minimaal twee woorden bestaat, correspondeert met twee knopen en een verbinding daartussen. De eerste knoop bevat het eerste woord van de songtitel, de tweede knoop het laatste woord van de songtitel. De verbinding tussen de twee knopen wijst van de eerste naar de tweede knoop en bevat de (eventuele) tekst tussen het eerste en het laatste woord in de songtitel. Indien de songtitel uit slechts twee woorden bestaat, bevat de pijl tussen de twee knopen geen tekst. Figuur 1 toont een klein gedeelte van de Beatlessongtitelgraaf die zo wordt verkregen.

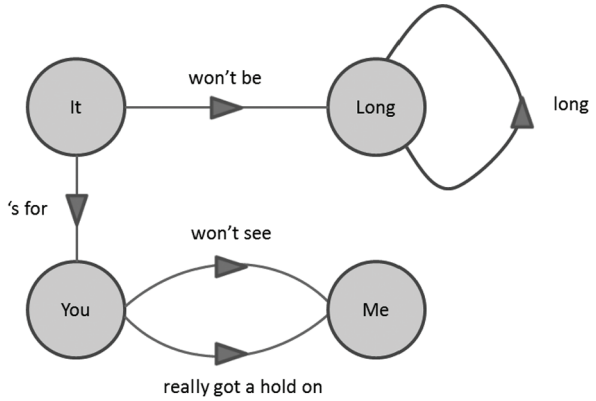

fig. 1 Gedeelte van Beatlessongtitelnetwerk.

Het netwerk dat afgebeeld is in figuur 1 bevat vijf songtitels, namelijk *It won't be long*, *Long, long, long*, *It's for you*, *You won't see me* en *You really got a hold on me*. Uit de figuur blijkt onmiddellijk dat het Beatlessongtitelnetwerk een gerichte multigraaf is met zelflussen. Er zijn namelijk twee verbindingen tussen de nodes 'You' en 'Me', en het liedje *Long, long, long* vormt een zelflus. Merk op dat er voor gekozen is om de titel *It's for you* te splitsen in een knoop 'It' en een knoop 'You' met op de verbinding ''s for'.

Het vinden van de langste keten in een songtitelnetwerk (langste in de zin van het grootste aantal songs), komt dus neer op het vinden van het langste pad in een gerichte multi-graaf met zelflussen, waarbij knopen meerdere malen bezocht kunnen worden, maar verbindingen slechts één keer. Het bepalen van een langste pad binnen een netwerk heeft verschillende toepassingen, zoals de strategische planning van openbare transportnetwerken (Borndörfer, Grötschel & Pfetsch, 2007), de prijsstelling van producten bij verschillende klanten om winst te optimaliseren (Grigoriev, Van Loon, Sitters & Uetz, 2006) en het optimaliseren van de levensduur van verbindingen in ad-hoc draadloze netwerken (Chang & Tassiulas, 2004). Een complicatie bij

dit type problemen is dat het NP-compleet is. Dat wil zeggen dat het vinden van het langste pad niet binnen polynomiale tijd lukt, tenzij er wordt aangetoond dat er een polynomiale tijd-oplossing bestaat voor willekeurig welk NP-compleet probleem dan ook (zie Garey & Johnson, 1979). Dit P versus NP-probleem is echter hardnekkig en het is nu een van de bekendste problemen in de wiskunde. Het probleem is opgenomen in de lijst van zogenaamde millenniumproblemen van het Clay Mathematics Institute, dat één miljoen dollar uitlooft voor de oplossing van een van de problemen[2].

Aangezien het netwerk met Beatlessongtitels klein is, kunnen we, door simpelweg alle ketens in het netwerk te bepalen, eenvoudig aantonen dat we binnen dit netwerk niet verder komen dan een keten van lengte drie: *All you need is Love me Do you want to know a Secret*.

Om het probleem van het vinden van de langste keten van songtitels uitdagender te maken, gaan we uit van de liedjes van de Top 2000. De Top 2000 is een jaarlijks programma van Radio 2. In dit artikel zullen we ons richten op het vinden van de langste keten van songtitels in de Top 2000, versie 2011. In de volgende sectie beschrijven we daartoe de beschikbare Top 2000-data. Vervolgens construeren we het Top 2000-songtitelnetwerk en geven hiervan enkele karakteristieken. Daarna gaan we op zoek naar het langste pad. Tot slot vinden we het meest waarschijnlijke langste pad in het Top 2000-netwerk, dat alle woorden van songtitels van minimaal twee woorden bevat.

Data

De data die nodig zijn om het Top 2000-songtitelnetwerk te analyseren, is beschikbaar via de website van de Top 2000, zie http://top2011.radio2.nl/lijst/. Behalve voor het zoeken naar het langste pad kunnen de data ook worden gebruikt voor het analyseren van allerlei andere statistieken, zoals beschreven in *De muziek zegt alles – De Top 2000 onder professoren* (Draaisma, Wijfjes, Vingerhoets e.a., 2011).

Voordat we overgaan tot een analyse van het Top 2000-songtitelnetwerk, zullen we eerst toelichten hoe de knopen in het netwerk tot stand zijn gekomen. We hebben een aantal keuzes moeten maken voor woorden die, met name in het Engels, een apostrof bevatten. In het geval van een genitief gebruik van ‘’s’ kiezen we ervoor het woord als één woord te beschouwen. Dus de songtitel *Nobody’s wife* bevatte twee woorden. Echter, indien de ‘’s’ wordt gebruikt als afkorting van ‘is’, beschouwen we de ontstane samenvoeging als twee woorden. Dus *She’s the one* bestaat uit vier woorden. Andere samenvoegingen met werkwoorden worden eveneens als twee woorden beschouwd, zoals ‘You’ve’, ‘I’ll’ and ‘I’d’. Een samenvoeging van een werkwoord met de ontkenning ‘not’ wordt wel als één woord beschouwd. Dus bijvoorbeeld *Ain’t no sunshine* bestaat uit drie woorden. Indien een gedeelte van de songtitel tussen haakjes staat, worden de woorden tussen haakjes weggelaten. Dus, *(Everything I do) I do it for you* wordt *I do it for you* en bestaat uit vijf woorden.

Het Top 2000-songtitelnetwerk

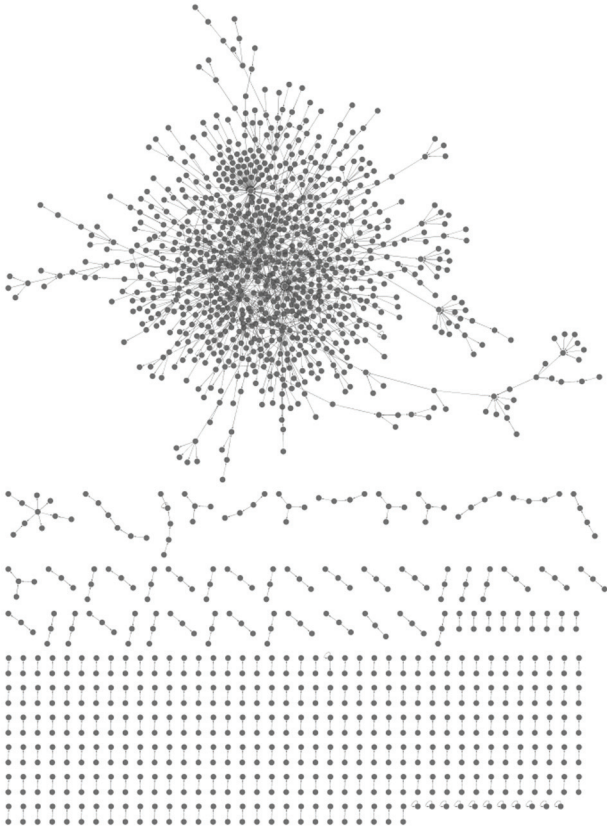

fig. 2 Het Top 2000-songtitelnetwerk.

Figuur 2 toont het netwerk dat wordt verkregen op de manier die we in de introductie beschreven. Het netwerk is gevisualiseerd met behulp van de freeware tool CytoScape[3]. Meteen valt op dat het Top 2000-songtitelnetwerk één gigantische component bevat, alsmede een groot aantal veel kleinere componenten. Met behulp van de netwerkanalysemodule binnen Cytoscape kunnen we aantonen dat het netwerk in totaal uit 295 verschillende componenten bestaat, waarvan er 237 een grootte van twee hebben. De grootste component bestaat uit 905 nodes. De één-na-grootste component, met tien nodes, bestaat uit alleen Nederlandstalige titels (figuur 3). Er zijn elf componenten met omvang één. Dit zijn ‘self loops’: titels die beginnen en eindigen met hetzelfde woord. Voorbeelden zijn: *Honey, honey*, *Na, na, na*, *Woorden*

zonder woorden en *Ashes to ashes*. Merk op dat we titels die uit slechts één woord bestaan buiten beschouwing hebben gelaten; deze titels worden dus niet meegeteld als componenten met omvang één.

De verbinding tussen twee knopen die het vaakst voorkomt, loopt van 'I' naar 'you': er zijn vijftien songtitels die beginnen met 'I' en eindigen op 'you' (bijvoorbeeld *I belong to you*, *I put a spell on you*). Er loopt geen verbinding terug van 'you' naar 'I', dus er zijn geen songtitels in de Top 2000 die beginnen met 'you' en eindigen op 'I'; wel zijn er vier titels die beginnen met 'you' en eindigen op 'me'.

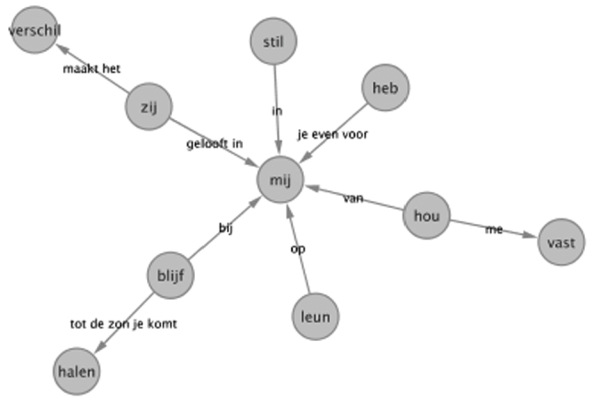

fig. 3 De op één-na-grootste component in het Top 2000-netwerk.

De eerstvolgende meest voorkomende verbinding loopt van 'the' naar 'love': deze komt vijf keer voor. De woorden die het meest als eerste woord voorkomen, zijn 'the' (82 keer), 'I' (66) en 'you' (27). De woorden die het vaakst als laatste woord voorkomen zijn 'love' (38), 'you' (30) en 'me'(26).

Het langste pad in het Top 2000-songtitelnetwerk

Met een brute force algoritme hebben we het langste pad in het netwerk bepaald. Het algoritme itereert over elk van de links in het netwerk en genereert per link alle paden die beginnen met die edge. Dit geeft de langst mogelijke aaneenschakeling van songtitels met twee of meer woorden, waarbij het laatste woord van een songtitel het eerste woord vormt van de volgende songtitel in de keten. Toepassing van dit algoritme leert ons dat het langst voorkomende pad lengte 29 heeft. Op een iMac met 3.4 GHz processor en 8 GB intern geheugen was de rekentijd drie seconden. Er blijken zelfs 191 paden te zijn met deze lengte. Eén van deze paden is:

I am a rock | Rock 'n roll | Roll over lay down | Down down | Down under | Under the bridge | Bridge over troubled water | Water of love | Love is all | All this time | Time waits for no one | One day like this | This is the last time | Time after time | Time stood still | Still loving you | You got it | It must have been love | Love of my life | Life is what you make it | It is my life | Life on Mars | Mars needs woman | Woman in love | Love song | Song sung blue | Blue hotel | Hotel California | California girls

Figuur 4 toont dit langste pad als netwerk. Het bijzondere eraan is dat alle andere paden met lengte 29 minimaal één verbinding (dus songtitel) gemeen hebben met dit langste pad.

Top 2000-songtitelnetwerk met alle woorden

Een andere manier om naar het Top 2000-netwerk te kijken is als een zogenaamde precedence-graaf. In deze precedence-graaf bestaan de knopen uit alle losse woorden die in de Top 2000 voorkomen. Er bestaat een gerichte verbinding tussen twee knopen als het ene woord in een titel voorafgegaan wordt door het andere woord, gelabeld met de kans dat het ene woord aan het andere woord voorafgaat.

Precedence-grafen hebben verschillende praktische toepassingen, bijvoorbeeld Google's query suggesties. Google gebruikt de zoekvragen die in het verleden gesteld zijn om de query die je typt aan te vullen met suggesties. Op basis van de oude queries wordt een precedence-graaf opgebouwd die de kansen bevat dat de gebruiker een bepaalde zoekvraag gaat stellen, gegeven de woorden die al zijn ingetypt (zie figuur 5).

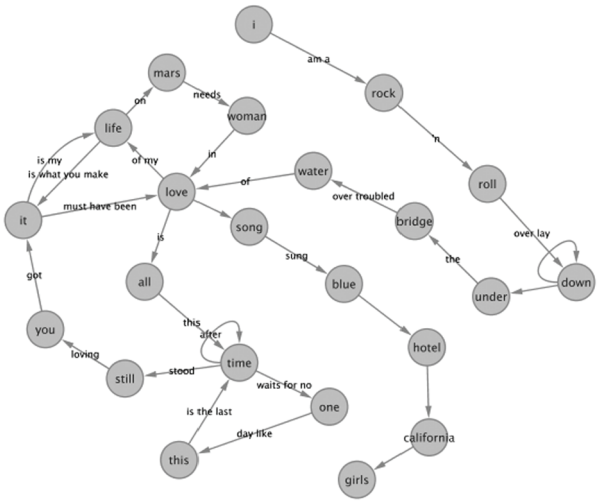

fig. 4 Een langste pad in het Top 2000-songtitelnetwerk met lengte 29.

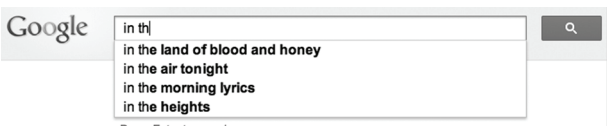

fig. 5 Google's zoekvraagsuggesties.

Net als in de vorige secties nemen we titels die uit slechts één woord bestaan niet mee in de constructie van de graaf omdat ze geen precedence-relatie bevatten. De 1639 titels uit de Top 2000 die uit tenminste twee woorden bestaan, bevatten een totaal van 4051 precedence-relaties, waarvan er 3131 uniek zijn. Verreweg het grootste deel (66% of 2680/4051) van de woordcombinaties komt slechts één keer voor. De meest voorkomende combinatie is die van 'in' en 'the', welke 39 keer in een titel voorkomt. De precedence-graaf die we op deze manier hebben geconstrueerd, bestaat uit 2519 knopen en 3131 gelabelde en gerichte verbindingen.

Nu we een met kansen gelabelde graaf hebben geconstrueerd, kunnen we ons afvragen wat het meest waarschijnlijke langste pad is in het Top 2000-songtitelnetwerk, waarin alle woorden zijn meegenomen. We beginnen het pad bij 'in', omdat de combinatie 'in' gevolgd door 'the' het meest voorkomt, en we kiezen vanuit een knoop steeds de meest waarschijnlijke bestemming die nog niet bezocht is. Als dit er meer dan één is, kiezen we als volgende knoop degene waarlangs het langste pad voert. Het enorme pad van woorden dat aldus ontstaat, bestaat uit 927 knopen, wisselt regelmatig van taal en bevat delen in het Engels, Nederlands en Frans. Uiteindelijk eindigt de keten met de woorden 'dromen zijn bedrog'.

Conclusie

Om de langste keten van songtitels in de Top 2000 te vinden, hebben we de Top 2000 gerepresenteerd als een gericht netwerk, waarbij de eerste en laatste woorden van alle songtitels die uit meer dan één woord bestaan, de knopen in het netwerk vormen. Hierbij wijst de verbinding van de knoop die het eerste woord van de songtitel vormt naar de knoop die het laatste woord van de songtitel representeert. We hebben aangetoond dat het aldus verkregen Top 2000-songtitelnetwerk één gigantische component bevat dat bestaat uit 905 knopen, en uit 294 veel kleinere componenten. De meest voorkomende verbinding in het netwerk betreft liedjes die beginnen met 'I' en eindigen met 'you'. Deze verbinding komt vijftien keer voor. Het woord waar een liedje het vaakst mee begint, namelijk 82 keer, is 'the'. Songtitels eindigen het vaakst op 'love', te weten 38 keer. Het langste pad in het Top 2000-songtitelnetwerk heeft lengte 29; dit is dus de langste keten van songtitels die we kunnen realiseren in de Top 2000. We hebben 191 verschillende ketens van lengte 29 gevonden. Tot slot hebben we aangetoond dat het meest waarschijnlijke langste pad in het Top 2000-songtitelnetwerk, dat alle woorden van songtitels van minimaal twee woorden bevat, 927 woorden bevat.

We hebben het voornemen om het beschreven onderzoek op een drietal vlakken voort te zetten. Ten eerste zijn we van plan om het Top 2000-netwerk, zoals gevisualiseerd in figuur 2 verder te ontwikkelen tot een interactieve applicatie, waarin, door op de verbindingen te klikken, ook audiofragmenten geïntegreerd zullen worden. Ten tweede zullen we een quiz-game ontwikkelen op basis van het Top 2000-netwerk. Tot slot zullen we het zoeken van langste ketens van songtitels ook toepassen op grotere databases, zoals die van de muziekherkenningsapp Shazam[4] of die van de online muziek catalogus Last.fm[5].

Robert Kooij,
Faculteit EWI, TU Delft,
TNO, Delft, r.e.kooij@tudelft.nl
Erik Boertjes,
TNO, Delft,erik.boertjes@tno.nl
Joost de Wit
TNO, Delft, joost.dewit@tno.nl

Noten

[1] http://en.wikipedia.org/wiki/List_of_The_Beatles_songs
[2] http://www.claymath.org/millennium
[3] http://www.cytoscape.org/
[4] http://www.shazam.com
[5] http://www.last.fm

Literatuur

Borndörfer, R., Grötschel, M., & Pfetsch, M. E. (2007). A Column-Generation Approach to Line Planning in Public Transport. *Transportation Science, 41*(1), 123-132.

Chang, J.-H. & Tassiulas, L. (2004). Maximum lifetime routing in wireless sensor networks. *IEEE/ACM Transactions on Networking, 12*(4), 609-619.

Draaisma, D., Wijfjes, H., Vingerhoets, A. e.a. (2011). *De muziek zegt alles – De Top 2000 onder professoren*. Amsterdam: L.J. Veen.

Garey, M. R., & Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*. New York, NY: W. H. Freeman & Co.

Grigoriev, A., Loon, J. van, Sitters, R., & Uetz, M. (2006). How to Sell a Graph: Guidelines for Graph Retailers. *Graph-Theoretic Concepts in Computer Science, Lecture Notes in Computer Science, 4271*, 125-136.