

## Samenvatting

Dit proefschrift heeft als onderwerp het clusteren van grafen door middel van simulatie van stroming, een probleem dat in zijn algemeenheid behoort tot het gebied der clusteranalyse. In deze tak van wetenschap ontwerpt en onderzoekt men methoden die gegeven bepaalde data een onderverdeling in groepen genereren, waarbij het oogmerk is een onderverdeling in groepen te vinden die *natuurlijk* is. Dat wil zeggen dat verschillende data-elementen in dezelfde groep idealiter veel op elkaar lijken, en dat data-elementen uit verschillende groepen idealiter veel van elkaar verschillen. Soms ontbreken zulke groepjes helemaal; dan is er weinig patroon te herkennen in de data. Het idee is dat de aanwezigheid van natuurlijke groepjes het mogelijk maakt de data te categoriseren. Een voorbeeld is het clusteren van gegevens (over symptomen of lichaamskarakteristieken) van patienten die aan dezelfde ziekte lijden. Als er duidelijke groepjes bestaan in die gegevens, kan dit tot extra inzicht leiden in de ziekte. Clusteranalyse kan aldus gebruikt worden voor *exploratief onderzoek*. Verdere voorbeelden komen uit de scheikunde, taxonomie, psychiatrie, archeologie, marktonderzoek en nog vele andere disciplines. Taxonomie, de studie van de classificatie van organismen, heeft een rijke geschiedenis beginnend bij Aristoteles en culminerend in de werken van Linnaeus. In feite kan de clusteranalyse gezien worden als het resultaat van een steeds meer systematische en abstracte studie van de diverse methoden ontworpen in verschillende toepassingsgebieden, waarbij methode zowel wordt gescheiden van data en toepassingsgebied als van berekeningswijze.

In de cluster analyse kunnen grofweg twee richtingen onderscheiden worden, naargelang het type data dat geclassificeerd moet worden. De data-elementen in het voorbeeld hierboven worden beschreven door vectoren (lijstjes van scores of metingen), en het verschil tussen twee elementen wordt bepaald door het verschil van de vectoren. Deze dissertatie betreft cluster analyse toegepast op data van het type 'graaf'. Voorbeelden komen uit de patroonherkenning, het computer-ondersteund ontwerpen, databases voorzien van hyperlinks en het World Wide Web. In al deze gevallen is er sprake van 'punten' die verbonden zijn of niet. Een stelsel van punten samen met hun verbindingen heet een graaf. Een goede clustering van een graaf deelt de punten op in groepjes zodanig dat er weinig verbindingen lopen tussen (punten uit) verschillende groepjes en er veel verbindingen zijn in elk groepje afzonderlijk. Het eerste deel van de dissertatie, bestaande uit de hoofdstukken 2 en 3, behandelt de positie van clusteranalyse in het algemeen en de positie van graafclusteren binnen de clusteranalyse in het bijzonder, alsmede de relatie van graafclusteren tot het aanverwante probleem van het *partitioneren* van grafen. In het cluster probleem zoekt men een 'natuurlijke' onderverdeling in groepjes en is het aantal en formaat van de groepjes niet voorgeschreven. In het partitie probleem zijn aantal en afmetingen wel voorgeschreven en zoekt men gegeven deze restricties een toewijzing van de elementen aan de groepjes zodanig dat er een minimale hoeveelheid verbindingen tussen de groepjes is.

De dissertatie beschrijft voorts theorie, implementatie en abstracte toetsing van een krachtig nieuw cluster algoritme voor grafen genaamd Markov Cluster algoritme of *MCL* algoritme. Het algoritme maakt gebruik van (en is in feite niet meer dan een schil om) een algebraïsch proces (genaamd *MCL* proces) gedefinieerd voor Markov grafen, i.e. grafen waarvoor de geassocieerde matrix stochastisch is. In dit proces wordt de aanvangsgraaf successievelijk getransformeerd door alternatie van de twee operatoren *expansie* en *inflatie*. Expansie is het nemen van de macht van een matrix volgens het klassieke matrix product. Stochastisch gezien betekent dit het uitrekenen van de overgangskansen behorend bij een meerstapsrelatie. Inflatie valt samen met het nemen van de macht van een matrix volgens het elementsgewijze Hadamard-Schur product, gevolgd door een kolomsgewijze herschaling zodat het uiteindelijke resultaat weer een (kolom) stochastische matrix is. Dit is een ongebruikelijke operator in de wereld van de stochastiek; zijn introductie is geheel en al gemotiveerd door de beoogde werking op grafen waar clusterstructuur aanwezig is. Het is namelijk te verwachten dat bij meerstapsrelaties die corresponderen met puntparen liggend binnen een natuurlijke cluster grotere overgangskansen zullen horen dan bij puntparen waarvan de punten in verschillende clusters liggen. De inflatie operator bevoordeelt meerstapsrelaties met grote bijbehorende kans en benadeelt meerstapsrelaties met kleine bijbehorende kans. De verwachting is dus dat het *MCL* proces meerstapsrelaties zal creëren en bestendigen die horen bij relaties liggend in één cluster, en dat het alle meerstapsrelaties zal decimeren die behoren bij relaties tussen verschillende clusters. Dit blijkt inderdaad het geval te zijn. Het *MCL* proces convergeert over het algemeen naar een idempotente matrix die zeer ijl is en bestaat uit meerdere componenten. De componenten worden geïnterpreteerd als een clustering van de aanvangsgraaf. Doordat de inflatie operator geparametriseerd is kunnen clusteringen op verschillend niveau van granulariteit ontdekt worden.

Het *MCL* algoritme bestaat ten eerste uit een transformatiestap van een gegeven graaf naar een stochastische aanvangsgraaf, gebruik makend van het standaard concept van een willekeurige wandeling op een graaf. Ten tweede vergt het de specificatie van twee rijen van waarden die de opeenvolgende expansie en inflatie parametrizingen definiëren. Tenslotte berekent het algoritme het bijbehorende proces en interpreteert het de resulterende limiet. Het idee om willekeurige wandelingen te gebruiken om clusterstructuur te ontdekken is niet nieuw, maar de wijze van uitvoering wel. Het idee wordt als 'graafcluster paradigma' geïntroduceerd in hoofdstuk 5, gevolgd door enige combinatorische voorstellen tot het clusteren van grafen. Getoond wordt dat er een verband is tussen de combinatorische en probabilistische clustermethoden, en dat een belangrijk onderscheid de localisatiestap is die probabilistische methoden over het algemeen introduceren. Het hoofdstuk besluit met een voorbeeld van een *MCL* proces en de formele definitie van zowel proces als algoritme. Notaties en definities zijn dan reeds geïntroduceerd in hoofdstuk 4. In hoofdstuk 6 wordt de interpretatiefunctie van idempotente matrices naar clusteringen geformaliseerd, worden simpele eigenschappen van de inflatie operator beschreven, en wordt de stabiliteit van *MCL* limieten en de geassocieerde clusteringen geanalyseerd. Het fenomeen van overlappende clusters is in principe mogelijk<sup>13</sup> en maakt intrinsiek deel uit van de interpretatiefunctie, maar blijkt

---

<sup>13</sup>De tot nu toe waargenomen overlap van clusters correspondeerde altijd met een graafautomorfisme dat het overlappende deel van clusters op zichzelf afbeeldde.

instabiel te zijn. Hoofdstuk 7 introduceert de klassen van *diagonaal symmetrische* en *diagonaal positief semi-definiete* matrices (matrices die diagonaal gelijkvormig zijn met een symmetrische respectievelijk positief semi-definiete matrix). Beide klassen worden in zichzelf overgevoerd door zowel expansie als inflatie<sup>14</sup>. Getoond wordt dat diagonaal positief semi-definiete matrices structuur bevatten die de interpretatiefunctie van idempotente matrices naar clusterings generaliseert. Hieruit volgt een preciezere duiding van het inflatoire effect van de inflatie-operator op het spectrum van de argumentmatrix. Ontkoppelingsaspecten van grafen en matrices zijn altijd nauw verbonden met karakteristieken van de geassocieerde spectra. Hoofdstuk 8 beschrijft een aantal bekende resultaten die ten grondslag liggen aan de meest gebruikte technieken ten behoeve van het partitioneren van grafen. De hoofdstukken 4 tot en met 8 vormen het tweede deel van de dissertatie.

Het derde deel doet verslag van experimenten met het *MCL* algoritme. Hoofdstuk 9 is theoretisch van aard en introduceert functies die gebruikt kunnen worden als maat voor de kwaliteit van een graafclustering. Ondermeer wordt een generieke maat afgeleid die uitdrukt hoe goed een karakteristieke vector de massa van een andere (niet negatieve) vector representeert. Elements- of kolomsgewijze toepassing van de maat geeft een uitdrukking voor de mate waarin een clustering de massa van een gewogen graaf of matrix representeert. Tevens wordt een metriek op de ruimte van clusterings of partities afgeleid, die gebruikt wordt om de continuïteitseigenschappen en het onderscheidend vermogen van het *MCL* algoritme te toetsen in hoofdstuk 12. Hoofdstuk 10 doet verslag van experimenten op kleine symmetrische grafen met welbepaalde dichtheidskarakteristieken zoals rastervormige grafen. Het *MCL* algoritme blijkt — experimenteel — een sterk scheidend vermogen te hebben. Experimenten met buurgrafen<sup>15</sup> wijzen uit dat het algoritme niet geschikt is indien de diameter van de natuurlijke clusters groot is. Dit verschijnsel kan begrepen worden in termen van de (stochastische) stromingseigenschappen van het algoritme. Hoofdstuk 11 gaat in op de schaalbaarheid van het algoritme. Cruciaal is dat de limiet van het *MCL* proces over het algemeen zeer ijl is en dat de iteranden van het proces ijl zijn in een gewogen interpretatie van het begrip ijl. Dat wil zeggen, de inflatie operator zorgt ervoor dat de meeste nieuwe niet-nul elementen (corresponderend met meerstapsrelaties) zeer klein blijven en uiteindelijk weer verdwijnen. Dit is des te meer waar naarmate de diameter van de natuurlijke clusters klein is, en naarmate de connectiviteit van de totale graaf laag is. Dit suggereert dat tijdens elke expansie stap — die ervoor zorgt dat de matrix vol loopt — de kolommen van de nieuw berekende matrix uitgedund kunnen worden door simpelweg de  $k$  grootste elementen van een nieuw berekende (stochastische) kolom te nemen, en deze elementen te herschalen op 1, waar  $k$  afhangt van de aanwezige rekencapaciteit. Omdat het berekenen van de  $k$  grootste waarden van een vector in principe niet in lineaire tijd kan, blijkt het in praktijk noodzakelijk een verfijnder schema te hanteren waarin de vector eerst uitgedund wordt door middel van drempelwaardes die afhangen van homogeniteitseigenschappen van de vector. Dit leidt in principe tot een complexiteit in de orde van grootte  $\mathcal{O}(Nk^2)$ , waar  $N$  de dimensie van de matrix is. Hoofdstuk 12 doet verslag van

---

<sup>14</sup>Voor diagonaal positief semi-definiete matrices geldt dit voor slechts een deel van de parameteriseringsruimte van de inflatie operator.

<sup>15</sup>Rasterachtige grafen gedefinieerd op punten in de Euclidische ruimte.

experimenten op testgrafen met tienduizend punten waarvan de verbindingen op zo'n manier (willekeurig) zijn gegenereerd dat een a priori beste clustering bekend is. Deze grafen hebben natuurlijke clusters met kleine diameter maar hebben als geheel hoge tot zeer hoge connectiviteit. Het geschaalde *MCL* algoritme blijkt zeer goede clusteringen te genereren die dicht bij de a priori bekende clustering liggen. De parameter  $k$  kan laag gekozen worden, maar de prestaties van het algoritme nemen sterker af naarmate  $k$  lager is en de totale connectiviteit van de input graaf hoger. De appendix *A cluster miscellany* beginnend op pagina 149 is geschreven voor een algemeen publiek en bevat korte uiteenzettingen over diverse aspecten van clusteranalyse, zoals de geschiedenis van het vakgebied en de rol van de computer.