Technische Universität Berlin
Fakultät IV: Elektrotechnik und Informatik
Institut für Softwaretechnik und Theoretische Informatik
Algorithmik und Komplexitätstheorie (AKT)

Proportionally Fair Clustering auf Graphen

Bachelorarbeit

von Phuoc Lucky Trinh

zur Erlangung des Grades „Bachelor of Science“ (B. Sc.)
im Studiengang Informatik

Erstgutachter:	Prof. Dr. Mathias Weller
Zweitgutachter:	Prof. Dr. Stefan Schmid
Betreuer:	Jannik Peters

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Die selbstständige und eigenhändige Ausfertigung versichert an Eides statt

Berlin, den _____ _____

Datum Unterschrift

Zusammenfassung

In dieser Bachelorarbeit befassen wir uns mit Fairness im Centroid Clustering. Die Aufgabe ist es, für n gegebene Punkte, k Center auszuwählen, wobei die Kosten eines Punktes die Distanz zum nächsten Center ist. Chen et al. [Che+19] nennen ein Clustering fair, wenn es keine Gruppe von $\lceil n/k \rceil$ Punkten gibt, die einen Center wählen könnten, der näher für jeden dieser Punkte ist. In dieser Bachelorarbeit befassen wir uns speziell mit fairen Clusterings auf Metriken, die durch Graphen definiert werden, im Spezialfall von $\mathcal{N} = \mathcal{M}$.

Für Bäume haben Micha und Shah [MS20] gezeigt, dass ein Exact Proportional Clustering immer existiert. Daher betrachten wir „baumähnliche“ Graphen. Wir zeigen, dass für Graphen mit Baumweite 1 und Graphen mit Baumtiefe 3 immer ein Exact Proportional Clustering existiert und geben Algorithmen an, die solche Clusterings finden. Wir zeigen auch, dass bereits für Graphen mit Baumweite 2 ein Exact Proportional Clustering nicht immer existiert und dass es hier keine $(2-\epsilon)$-Approximation zu Proportional Fairness gibt. Dies beantwortet die offene Frage von Micha und Shah [MS20], ob es für Graphen im Spezialfall $\mathcal{N} = \mathcal{M}$ immer ein Exact Proportional Clustering gibt. Dieser Lowerbound impliziert direkt, dass für Clusterings auf beliebigen Metriken für den Spezialfall $\mathcal{N} = \mathcal{M}$ nicht immer ein $(2-\epsilon)$-Proportional Clustering existiert.

Wir zeigen, dass der Greedy Capture Algorithmus von Chen et al. [Che+19] für Graphen keinen besseren Upperbound als $1 + \sqrt{2}$ liefert. Zuletzt zeigen wir, dass es NP-schwer ist, zu entscheiden, ob für einen gegebenen Graphen ein ρ-Proportional Clustering, für ein festes $\rho < 2$ existiert und dass es keinen Algorithmus gibt, der gleichzeitig das k-Center Objective und die Proportional Fairness approximiert.

Inhaltsverzeichnis

Kapitel 1

Einführung

Machine Learning Algorithmen arbeiten heutzutage mit Datenpunkten, die Menschen repräsentieren können. Folglich stellt sich die Frage, ob diese Algorithmen „fair“ sind. Fairness in Machine Learning wurde schon seit längerer Zeit erforscht. Für eine Übersicht an wissenschaftlichen Artikeln verweisen wir auf Mehrabi et al. [Meh+21].

In dieser Bachelorarbeit befassen wir uns mit Centroid Clustering. Beim Centroid Clustering sind Datenpunkte $\mathcal{N}$ gegeben und die Aufgabe ist es, Zentren der Cluster aus einer Menge von möglichen Zentren $\mathcal{M}$ auszuwählen, die diese Datenpunkte gruppieren sollen. Clustering findet zum Beispiel in der Bildverarbeitung (Segmentierung), Bioinformatik, Datenreduktion oder Data-Mining Anwendung [Sax+17].

1.1 Motivation und Stand der Forschung

Angelehnt an das Beispiel von Chen et al. [Che+19]: Seien mehrere dichte Gruppen an Individuen nah beieinander und eine Gruppe, die eine breite Fläche einnimmt, weiter weg gegeben. Ein Standard k-Median Clustering würde wenige Center bei den dicht besiedelten Gruppen platzieren, weil diese nah beieinander liegen und mehr Center bei der weniger dicht besiedelten Gruppe platzieren. Dies wäre aber verhältnismäßig unfair, weil eine größere Population auch eine höhere Anzahl an Centern verdienen sollte.

Chen et al. [Che+19] haben daher den Begriff von *Proportional Fairness* eingeführt: Ein Clustering heißt fair, wenn es keine verhältnismäßig große Gruppe von Datenpunkten gibt, die „nicht mit dem Clustering einverstanden sind“. Eine Gruppe von Punkten ist nicht einverstanden, wenn für diese eine bessere Wahl eines Centers existiert. Chen et al. [Che+19] betrachten das Clustering über einen beliebigen metrischen Raum. Micha und Shah [MS20] hingegen haben Proportional Clusterings über der L^1, L^2 und L^∞-Norm betrachtet und entsprechende untere und obere Schranken für die Approximation zu Proportional Fairness angegeben. Sie haben auch Proportional Clusterings auf Graphen betrachtet und gezeigt, dass für Bäume immer ein faires Clustering existiert. Eine offene Frage in deren Paper war, ob es auch in allgemeinen Graphen immer ein Exact Proportional Clustering im Spezialfall $\mathcal{N} = \mathcal{M}$ gibt.

In dieser Bachelorarbeit konzentrieren wir uns daher auf durch Graphen induzierte Metriken im Spezialfall $\mathcal{N} = \mathcal{M}$. Graphen sind besonders mächtige Werkzeuge, mit denen viele Probleme in der Wissenschaft und im Alltag gelöst werden können [BM+76].

1.2 Methodik

Es gibt bereits einige Ergebnisse für Proportionally Fair Clustering in anderen metrischen Räumen [Che+19; MS20]. Einige dieser Resultate lassen sich auf Graphen übertragen, indem Datenpunkte durch Sterne ersetzt und deren Distanzen durch Pfade dargestellt werden. Beliebige metrische Räume haben reellwertige Distanzen zur Verfügung, während Distanzen in Graphen immer ganzzahlig sind. Das haben wir modelliert, indem wir reelle Distanzen beliebig gut mit rationalen Distanzen approximiert haben und dann alle Distanzen zu Ganzzahlen skaliert[1] haben. Das funktioniert, weil Proportional Fairness *scale invariant* ist [Che+19]. Die Sterne sorgen dafür, dass das Gewicht der Datenpunkte bei den originalen Punkten bleibt und sich nicht zu sehr auf die Pfade verlagert.

1.3 Verwandte Arbeiten

Die Definition von Proportional Fairness von Chen et al. [Che+19] ist vom Begriff „Core“ aus der Volkswirtschaftslehre inspiriert. Eine Lösung ist im Core, wenn keine verhältnismäßig große Gruppe existiert, für die es eine bessere Lösung gibt. Fain et al. [FMS18] betrachten Fairness mittels des Cores in der Ressourcenzuweisung von unteilbaren Gütern. Für verschiedene Bedingungen an den Gütern geben sie Algorithmen mit additiven Approximationen zum Core an. Jiang et al. [JMW20] und Cheng et al. [Che+20] hingegen befassen sich mit dem Core im Rahmen von Committee Selection. Caragiannis et al. [CMS24] erweitern (in einem Vordruck) Proportionally Fair Clusterings auf nicht-Centroid Clusterings, in der sie sich auch mit dem Core und ihrer Abschwächung, die „Fully Justified Representation“ befassen.

Basierend auf dem Paper von Chen et al. [Che+19] haben Li et al. [Li+21] eine neue Proportional Fairness Definition entwickelt, bei der eine unzufriedene Gruppe die Summe ihrer Distanz verbessern muss. Dort geben sie Ergebnisse zur Existenz, Härte und Approximation ihres Fairnessbegriffs an.

Der Fairnessbegriff *Individual Fairness* von Jung et al. [JKL20] ist, anders als die anderen erwähnten Fairnessbegriffe, kein *Group Fairness*-Begriff. Individual Fairness ist die Idee, dass das Clustering fair für jeden einzelnen Datenpunkt sein soll. Für ein Paper, das beide Individual Fairness und Group Fairness betrachtet, verweisen wir auf das Paper von Kellerhals und Peters [KP23], in dem sie Proportional Fairness und Individual Fairness im Social Choice Rahmen untersuchen. In der zeigen sie unter anderem, dass diese Fairnessbegriffe verwandt sind.

1.4 Ergebnisse und Gliederung

In Kapitel 3 erweitern wir das Ergebnis für Bäume von Micha und Shah [MS20] auf Wälder. Wir zeigen außerdem für triviale Graphen mit Baumweite 2, nämlich Kreise, dass immer ein Exact Proportional Clustering existiert.

In Kapitel 4 zeigen wir, dass für Graphen mit kleiner Baumtiefe — Baumtiefe 3 und zusammenhängende Graphen mit Baumtiefe 4 — ein Exact Proportional Clustering immer existiert und geben einen Algorithmus an, der ein solches Clustering findet. Der

[1] Das geht zum Beispiel mit dem kleinsten gemeinsamen Vielfachen der Nenner.

Algorithmus funktioniert, indem zuerst der Algorithmus von Micha und Shah [MS20] für Bäume auf dem Trémaux Baum läuft und dann mögliche Blocking Coalitions aufgelöst werden.

In Kapitel 5 zeigen wir, dass es für Graphen nicht immer ein $(2-\epsilon)$-Proportional Clustering gibt, indem wir zuerst eine Instanz für $\mathcal{N} \neq \mathcal{M}$ in einer nicht-Graph Metrik angeben und dann die in der Einführung erwähnte Technik anwenden, um diesen in einen Graphen zu transformieren. Der Graph hat eine Baumweite von 2 und hat bereits für Baumtiefe 4 einen Lowerbound größer 1. Das löst die Frage von Micha und Shah [MS20], ob es für Graphen im Spezialfall $\mathcal{N} = \mathcal{M}$ immer ein Exact Proportional Clustering gibt. Dieser Lowerbound impliziert direkt einen Lowerbound von 2 für Clusterings auf beliebigen Metriken im Spezialfall $\mathcal{N} = \mathcal{M}$, was den zuvor bekannten Lowerbound von $1,5$ von Chen et al. [Che+19] verbessert. Der neue Lowerbound gleicht dem für Clusterings im allgemeinen Fall $\mathcal{N} \neq \mathcal{M}$.

In Kapitel 6 zeigen wir, dass der Greedy Capture Algorithmus von Chen et al. [Che+19] für Graphen und bereits für Bäume keinen besseren Upperbound als $1+\sqrt{2}$ liefert.

In Kapitel 7 zeigen wir, dass es NP-schwer ist zu entscheiden, ob in einem gegebenen Graphen ein ρ-Proportional Clustering, für ein festes $\rho < 2$ existiert.

Zuletzt zeigen wir in Kapitel 8, dass es keinen Algorithmus gibt, der für zusammenhängende Graphen gleichzeitig das k-Center Objective und die Proportionalität konstant approximiert.

Kapitel 2

Notation und Grundlegendes

Proportional Fairness. Gegeben sei eine Menge $\mathcal{N}$ von n Datenpunkten und eine Menge $\mathcal{M}$ von möglichen Clusterzentren. Wir betrachten hauptsächlich den Spezialfall $\mathcal{N} = \mathcal{M}$. Sei $d : (\mathcal{N} \cup \mathcal{M}) \times (\mathcal{N} \cup \mathcal{M}) \to \mathbb{R}_{\geq 0}$ eine Distanzfunktion, die die Dreiecksungleichung erfüllt, d. h. es gilt $d(a, c) \leq d(a, b) + d(b, c)$, für alle $a, b, c \in \mathcal{N} \cup \mathcal{M}$. Gesucht ist eine Menge von Cluster Centern $X \subseteq \mathcal{M}$, wobei $|X| = k$.[1] Wir benutzen die Kurznotation $d(i, X) := \min_{x \in X} d(i, x)$. Die Eigenschaft, die das Clustering X aufweisen soll, ist die folgende, wie sie von Chen et al. [Che+19] definiert wurde:

Definition 2.1. Sei $X \subseteq \mathcal{M}$ ein Clustering, mit $|X| = k$ und sei $\rho \geq 1$. Eine Menge $S \subseteq \mathcal{N}$ von Datenpunkten heißt *Blocking Coalition*, falls $|S| \geq \left\lceil \frac{n}{k} \right\rceil$ und es ein $y \in \mathcal{M}$ gibt, sodass für alle $i \in S$ gilt: $\rho \cdot d(i, y) < d(i, X)$. Das Clustering X heißt ρ-*Proportional*, wenn es keine solche Blocking Coalition gibt und *Exact Proportional*, falls zusätzlich $\rho = 1$ gilt.

Graphentheorie. Ein Graph ist ein Paar (V, E), wobei V eine nicht-leere, endliche Menge von Knoten und $E \subseteq V \times V$ eine Menge von Kanten ist. Wir nutzen Standardnotationen von Bondy und Murty [BM+76]. Wir betrachten hier nur ungerichtete, schleifenlose Graphen.

Ein Pfad ist eine Sequenz von Knoten, in der aufeinander folgende Knoten benachbart sind, d. h. eine Kante haben und kein Knoten mehrfach vorkommt. Ein Kreis ist ein Pfad, in der der erste Knoten mit dem letzten übereinstimmt. Ein Graph ist zusammenhängend, wenn es zwischen je zwei Knoten einen Pfad gibt. Eine Zusammenhangskomponente ist ein maximaler, zusammenhängender Teilgraph. Die Distanz $d(u, v)$ zwischen zwei Knoten u, v aus derselben Zusammenhangskomponente ist die Länge eines kürzesten Pfades von u nach v und ist unendlich, falls sie in verschiedenen Zusammenhangskomponenten liegen.

Ein Graph G heißt Wald, wenn G keine Kreise enthält und Baum, wenn G zusätzlich zusammenhängend ist. Sterne sind zusammenhängende Graphen, in denen alle Kanten einen zentralen Knoten als Endpunkt haben.

[1] Die vorgestellten Algorithmen können weniger als k Cluster Center auswählen. Für die Analyse von Blocking Coalitions spielt dies jedoch keine Rolle, da beliebige Punkte hinzugenommen werden können, ohne dass eine Blocking Coalition entsteht.

Die Baumweite eines Graphen wird über die Robertson-Seymour Baumzerlegung definiert [RS86].

Definition 2.2. Sei $G = (V, E)$ ein Graph. Eine Baumzerlegung von G ist ein Baum T, wobei für die Knoten B von T gilt: $B \subseteq V(G)$. Außerdem muss gelten:

1. $\bigcup_{B \in V(T)} B = V$,
2. für alle $(u, v) \in E$ gibt es ein $B \in V(T)$, mit $u, v \in B$ und
3. für alle $A, B, C \in V(T)$ gilt: Wenn B auf dem Pfad von A nach C liegt, dann gilt: $A \cap C \subseteq B$.

Die *Weite* einer Baumzerlegung T ist $\max_{B \in V(T)} |B| - 1$ und die *Baumweite* $\mathrm{BW}(G)$ eines Graphen G ist die minimale Weite aller Baumzerlegungen des Graphen.

Die Baumtiefe wird wie folgt definiert [NDM06]:

Definition 2.3. Sei $F = (V_F, E_F)$ ein gewurzelter Wald. Der Abschluss $\mathrm{clos}(F)$ von F ist der Graph $(V_F, \{\{u, v\} : u \text{ ist Vorfahre von } v \text{ in } F \text{ und } u \neq v\})$.

Sei $G = (V, E)$ ein Graph. Die *Baumtiefe* $\mathrm{BT}(G)$ von G ist die kleinste Höhe eines gewurzelten Waldes F, sodass G Teilgraph von $\mathrm{clos}(F)$ ist, d. h. $G \subseteq \mathrm{clos}(F)$. F nennen wir Trémaux Baum von G.

Eine alternative Definition ist:

$$\mathrm{BT}(G) = \begin{cases} 1 & \text{, falls } |V(G)| = 1 \\ 1 + \min\limits_{v \in V} \mathrm{BT}(G - v) & \text{, falls } G \text{ zusammenhängend ist} \\ & \text{und } |V(G)| > 1 \\ \max\limits_{i} \mathrm{BT}(G_i) & \text{, sonst} \end{cases}$$

wobei G_i die Zusammenhangskomponenten von G bezeichnen.

Um einige Beispiele für Graphklassen und deren Graphparameter anzugeben: Bäume haben Baumweite 1, während n-Cliquen Baumweite $n - 1$ haben. Sterne haben Baumtiefe 2.

Serien-parallele Graphen werden wie folgt induktiv definiert [TNS82]:

Definition 2.4. Ein *serien-paralleler Graph* ist ein 4-Tupel (V, E, s, t), wobei (V, E) ein Graph ist und $s, t \in V$, und der aus den folgenden Operationen erzeugt werden kann:

- Der Graph, bestehend aus einer einzelnen Kante, ist ein serien-paralleler Graph.
- Wenn (V_1, E_1, s_1, t_1) und (V_2, E_2, s_2, t_2) serien-parallele Graphen sind, dann ist $(V_1 \uplus V_2, E_1 \uplus E_2, s_1, t_2)$, wobei s_2 durch t_1 ersetzt wird, auch ein serien-paralleler Graph.
- Wenn (V_1, E_1, s_1, t_1) und (V_2, E_2, s_2, t_2) serien-parallele Graphen sind, dann ist $(V_1 \uplus V_2, E_1 \uplus E_2, s_1, t_1)$, wobei s_2 durch s_1 und t_2 durch t_1 ersetzt wird, auch ein serien-paralleler Graph.

Die Knoten eines Graphen $G = (V, E)$ können einen Typ / einen Label haben. Wir benutzen die Kurznotation $\{a\}$ für $\{v \in V : v \text{ ist vom Typ } a\}$ und a für einen beliebigen Knoten vom Typ a.

Kapitel 3

Baumweite

Graphen mit Baumweite 1 sind genau die Wälder. Für Bäume haben Micha und Shah [MS20] bereits gezeigt, dass es immer ein Exact Proportional Clustering gibt. In diesem Abschnitt zeigen wir zuerst ein hilfreiches Lemma, das die Analyse von Blocking Coalitions erleichtert. Mithilfe des Lemmas lässt sich die Korrektheit vom Algorithmus 1 von Micha und Shah [MS20] einfacher zeigen und sogar auf Wälder erweitern. Der Algorithmus durchläuft im Wesentlichen den Baum bottom-up und eröffnet einen Cluster Center, sobald $\lceil \frac{n}{k} \rceil$ (neue) Knoten hinzugekommen sind. Anschließend eröffnet der Algorithmus noch einen Cluster Center in der Wurzel, falls diese noch nicht im Clustering war. Diesen letzten Schritt werden wir für Wälder nicht ausführen, da sonst möglicherweise mehr als k Cluster Center eröffnet werden. Wir zeigen, dass selbst ohne einen Center in der Wurzel, das vom Algorithmus ausgegebene Clustering ein Exact Proportional Clustering ist.

Außerdem zeigen wir, dass für triviale Graphen mit Baumweite 2, nämlich für Kreise und generell Graphen mit maximalem Knotengrad 2 ein Exact Proportional Clustering immer existiert. Für allgemeine Graphen mit Baumweite 2 und sogar für serien-parallele Graphen werden wir in Abschnitt 5.2 zeigen, dass dies nicht mehr der Fall ist.

3.1 Baumweite 1

Zuerst beweisen wir ein Lemma, das mögliche Blocking Coalitions auf Teilgraphen „zwischen“ Cluster Centern einschränkt. Wir erinnern, dass eine Blocking Coalition eine Menge von Punkten ist, die einen besseren Cluster Center wählen können. Formell heißt das Lemma:

Lemma 3.1. *Sei X ein Clustering. Sei $G' = G - X$. Eine Blocking Coalition S kann nicht auf mehrere Zusammenhangskomponenten G'_ℓ von G' verteilt sein. Das heißt es gilt $S \subseteq V(G'_\ell)$, für eine Zusammenhangskomponente G'_ℓ von G'.*

Beweis. Es wird ein Widerspruchsbeweis geführt. Seien G'_ℓ, G'_m zwei Zusammenhangskomponenten, mit $V(G'_\ell) \cap S \neq \emptyset$ und $V(G'_m) \cap S \neq \emptyset$.[1] Sei also $i \in V(G'_\ell) \cap S$ und $j \in V(G'_m) \cap S$. Da $V(G'_\ell) \cap V(G'_m) = \emptyset$, kann y nicht in beiden Zusammenhangskomponenten sein. Sei also o. B. d. A. $y \notin V(G'_\ell)$. Da nach Löschung von X der Graph

[1] Falls es nur eine Zusammenhangskomponente gibt, gilt das Lemma trivialerweise.

in Zusammenhangskomponenten zerfällt, muss jeder Pfad, insbesondere der kürzeste, also Distanz-definierende Pfad, zwischen zwei Zusammenhangskomponenten durch ein $x \in X$ verlaufen. Sei also $x \in X$ auf dem kürzesten Pfad von i nach y. Dann gilt aber $d(i, X) \leq d(i, x) < d(i, y)$. Somit würde Knoten i nicht wechseln wollen. □

Das Lemma erlaubt es uns, den Korrektheitsbeweis vom Algorithmus 1 von Micha und Shah [MS20] für Bäume zu vereinfachen, sodass der Algorithmus, etwas abgewandelt, auch für Wälder funktioniert.

Satz 3.2. *Sei G ein Wald. Dann hat G ein Exact Proportional Clustering.*

Beweis. Sei G ein Wald. Der modifizierte Algorithmus 1 von Micha und Shah [MS20] arbeitet wie folgt. Für jeden Baum G_ℓ von G führe folgendes aus: 1) Setzte einen beliebigen Knoten als Wurzel von G_ℓ, um einen gewurzelten Baum zu erhalten. 2) Durchlaufe den Baum G_ℓ bottom-up. Falls an einem Knoten v mindestens $\lceil \frac{n}{k} \rceil$ Knoten im Unterbaum von v sind, dann eröffne einen Cluster Center in v und lösche den Unterbaum.

Zuerst beweisen wir, dass der Algorithmus maximal k Cluster Center öffnet. Für jeden Baum G_ℓ von G werden maximal $\left\lfloor \frac{|G_\ell|}{\lceil n/k \rceil} \right\rfloor$ Cluster Center geöffnet, da der Algorithmus (in Zeile 6) einen Cluster Center erst öffnet, wenn im entsprechenden Unterbaum mindestens $\lceil \frac{n}{k} \rceil$ Knoten sind und weil der Algorithmus keinen extra Center für die Wurzel eröffnet. Also werden in G maximal

$$\sum_\ell \left\lfloor \frac{|G_\ell|}{\lceil \frac{n}{k} \rceil} \right\rfloor \leq \sum_\ell \frac{|G_\ell|}{\frac{n}{k}} = \frac{k}{n} \sum_\ell |G_\ell| = k$$

Cluster Center geöffnet.

Aus Lemma 3.1 folgt, dass eine Blocking Coalition nicht auf mehrere Bäume von G verteilt sein kann. Es genügt also zu zeigen, dass sich in jedem Baum von G keine Blocking Coalition bilden kann.

Sei X das vom modifizierten Algorithmus ausgegebene Clustering. Wir betrachten einen beliebigen Baum H von G. Sei $H' = H - X$. Wegen Lemma 3.1 müssen wir nur die Zusammenhangskomponenten H'_ℓ von H' betrachten. Für jede Zusammenhangskomponente H'_ℓ von H' gilt $|H'_\ell| < \lceil \frac{n}{k} \rceil$, also gibt es nicht genug Knoten, um eine Blocking Coalition zu bilden. Denn angenommen es gibt eine Zusammenhangskomponente H'_ℓ, sodass $|H'_\ell| \geq \lceil \frac{n}{k} \rceil$. Dann hätte der Algorithmus aber spätestens bei der Wurzel r_ℓ von H'_ℓ einen Cluster Center eröffnet. Denn es gilt $|ST(r_\ell)| \geq \lceil \frac{n}{k} \rceil$ (in Zeile 6), weil Knoten aus abgeschnittenen Teilbäumen (Zeile 8) wegen Lemma 3.1 nicht zu H'_ℓ gehören können. Da nun in H'_ℓ ein Center ist, kann H'_ℓ aber keine Zusammenhangskomponente von $H - X$ sein; ein Widerspruch. □

3.2 Baumweite 2

Die einfachsten Graphen mit Baumweite 2 sind Kreise. Wir zeigen, dass Kreise ein Exact Proportional Clustering haben. Als Clustering wählen wir jeden $\lceil \frac{n}{k} \rceil$-ten Knoten. Besteht ein Graph G aus mehreren Kreisen, zeigen wir wie in Abschnitt 3.1, dass in einer Zusammenhangskomponente G_ℓ nur $\left\lfloor \frac{|G_\ell|}{\lceil n/k \rceil} \right\rfloor$ Cluster Center benötigt werden und diese ein Exact Proportional Clustering bilden.

Satz 3.3. *Sei G ein Graph, dessen Zusammenhangskomponenten G_ℓ Kreise sind. Dann hat G ein Exact Proportional Clustering.*

Beweis. Sei G also ein Graph, dessen Zusammenhangskomponenten G_ℓ Kreise sind. Setze $k_\ell = \left\lfloor \frac{|G_\ell|}{\lceil n/k \rceil} \right\rfloor$, für jeden der Kreise G_ℓ. Für jeden Kreis G_ℓ mit Knoten $\{v_1, v_2, \dots\}$ nimm die Knoten $\{v_{i \cdot \lceil \frac{n}{k} \rceil} : i \in \{1, \dots, k_\ell\}\}$ in das Clustering X auf. Analog wie in Satz 3.2 werden maximal k Cluster Center eröffnet.

Nun zeigen wir, dass X ein Exact Proportional Clustering ist. Aus Lemma 3.1 folgt, dass eine Blocking Coalition nicht auf mehrere Kreise von G verteilt sein kann. Es genügt also zu zeigen, dass sich in jedem Kreis von G keine Blocking Coalition bilden kann.

Sei H also ein beliebiger Kreis aus G. Sei $H' = H - X$. Wegen Lemma 3.1 müssen wir nur die Zusammenhangskomponenten H'_ℓ von H' betrachten. Da wir jeden $\lceil \frac{n}{k} \rceil$-ten Knoten als Cluster Center gewählt haben, gilt für alle Zusammenhangskomponenten H'_ℓ, bis auf die Zusammenhangskomponente H'_m zwischen $v_{1 \cdot \lceil \frac{n}{k} \rceil}$ und $v_{k_\ell \cdot \lceil \frac{n}{k} \rceil}$, dass $|H'_\ell| < \lceil \frac{n}{k} \rceil$. Also sind dort nicht genug Knoten, um eine Blocking Coalition zu bilden. Für die letzte größere Zusammenhangskomponente H'_m gilt aber $|H'_m| \leq 2 \lceil \frac{n}{k} \rceil - 2$. Denn wäre $|H'_m| \geq 2 \lceil \frac{n}{k} \rceil - 1$, gäbe es in H'_m einen Center. Sei S nun eine Blocking Coalition, mit $S \subseteq V(H'_m)$ und seien i_1, i_2 die beiden Randknoten von S, also die Knoten, die $d(i_1, i_2)$ maximieren. Sei $y \in V(H'_m)$.[2] Da $|S| \geq \lceil \frac{n}{k} \rceil$ gelten muss, gilt $d(i_1, y) + d(i_2, y) \geq \lceil \frac{n}{k} \rceil - 1$ und da es maximal $\lceil \frac{n}{k} \rceil - 2$ Knoten in $V(H'_m) \setminus S$ gibt, gilt $d(i_1, X) + d(i_2, X) \leq \lceil \frac{n}{k} \rceil$. Werden beide Ungleichungen zusammengesetzt, gilt

$$d(i_1, X) + d(i_2, X) \leq \left\lceil \frac{n}{k} \right\rceil \leq d(i_1, y) + d(i_2, y) + 1.$$

Damit S eine Blocking Coalition ist, muss $d(i_1, X) > d(i_1, y)$ gelten. Also gilt

$$d(i_2, X) + \underbrace{\overbrace{d(i_1, X) - d(i_1, y)}^{>0, \text{ also } \geq 1} - 1}_{\geq 0} \leq d(i_2, y)$$

und somit gilt $d(i_2, X) \leq d(i_2, y)$, also würde Knoten i_2 nicht wechseln wollen. □

Graphen mit maximalem Knotengrad 2 bestehen aus Kreisen, Pfaden und isolierten Knoten. Pfade und isolierte Knoten sind Bäume. Aus Satz 3.2 und Satz 3.3 erhalten wir also

Korollar 3.4. *Sei G ein Graph mit maximalem Knotengrad 2. Dann hat G ein Exact Proportional Clustering.*

[2] Denn sonst wäre S nach ähnlicher Begründung wie in Lemma 3.1 keine Blocking Coalition.

Kapitel 4

Baumtiefe

Es ist einfach zu sehen, dass Graphen mit Baumtiefe 2, Sterne als Zusammenhangskomponenten haben, also Wälder sind. Für Wälder haben wir in Abschnitt 3.1 bewiesen, dass es immer ein Exact Proportional Clustering gibt. In diesem Abschnitt zeigen wir, dass es für Graphen mit Baumtiefe 3 und für zusammenhängende Graphen mit Baumtiefe 4 immer ein Exact Proportional Clustering gibt. Der Algorithmus ist im Wesentlichen der modifizierte Algorithmus 1 von Micha und Shah [MS20] aus Abschnitt 3.1, angewandt auf dem Trémaux Baum des Graphen bzw. der Zusammenhangskomponenten. Dabei können sich Blocking Coalitions bilden. Diese sind aber von der Struktur sehr eingeschränkt und lassen sich durch Hinzufügen oder Verschieben eines entsprechenden Cluster Centers auflösen.

Zuerst führen wir einige Notationen ein. Sei $G = (V, E)$ ein Graph mit Baumtiefe t. Sei T also ein Trémaux Baum von G mit Tiefe t. Bei Knoten v wird mit einem Subskript ℓ der Level [MS20], also die Höhe von v im Baum, kennzeichnet, d. h. $\text{Level}(v_\ell) = \ell$. Die Menge der Knoten im Unterbaum am Knoten v wird mit $\text{ST}(v)$ bezeichnet. Für den Trémaux Baum T sollen Kanten zwischen zwei Knoten aus der tiefsten Ebene auch im Graphen existieren, also soll für v_{t-1}, v_t, mit $v_t \in \text{ST}(v_{t-1})$ gelten: $(v_{t-1}, v_t) \in E$.[1]

Satz 4.1. *Sei G ein Graph mit Baumtiefe $\text{BT}(G) = 3$. Dann hat G ein Exact Proportional Clustering. Algorithmus 1 findet ein solches.*

Beweis. Zuerst beweisen wir, dass Algorithmus 1 maximal k Cluster Center öffnet. In Zeile 2 werden analog zu Abschnitt 3.1 maximal k Cluster Center geöffnet. In Zeilen 12 und 16 verschiebt Algorithmus 1 lediglich einen Cluster Center. In Zeile 8 öffnet Algorithmus 1 nur einen Cluster Center, falls für ein $v_2 \in X \cap V(T_\ell)$ gilt: $|\,\text{ST}(v_2)\,| > \lceil \frac{n}{k} \rceil$, also $|\,\text{ST}(v_2)\,| \geq \lceil \frac{n}{k} \rceil + 1$. Da (in Zeile 5) $|A| = \lceil \frac{n}{k} \rceil - 1$ gilt und A und $\text{ST}(v_2)$ disjunkt sind, gilt

$$|A \cup \text{ST}(v_2)| = |A| + |\,\text{ST}(v_2)\,| \geq \left\lceil \frac{n}{k} \right\rceil - 1 + \left\lceil \frac{n}{k} \right\rceil + 1 = 2 \cdot \left\lceil \frac{n}{k} \right\rceil .$$

Es gibt keinen Cluster Center in A, weil die Teilbäume mit den Cluster Centern abgeschnitten werden und in $\text{ST}(v_2)$ gibt es nur einen Cluster Center, also darf in $A \cup \text{ST}(v_2)$ noch ein weiterer Cluster Center geöffnet werden, ohne k zu überschreiten.

[1] Dies ist immer möglich, denn wenn v_t keine Kante zu v_{t-1} hat, dann hat es nur Kanten zu höheren Vorfahren, kann also im Trémaux Baum eine Ebene hochwandern.

Algorithmus 1 Proportionally Fair Clustering für Baumtiefe 3

Input: Trémaux trees T_ℓ from the connected components of G

1: $X \leftarrow \emptyset$
2: perform the modified algorithm in section 3.1 on a copy of every T_ℓ and store the outputs in X
3: **for** every T_ℓ with $v_1 \notin X$ **do**
4: $\quad A \leftarrow V(T_\ell) \setminus \bigcup_{v_2 \in X} \mathrm{ST}(v_2)$
5: $\quad$ **if** $|A| = \lceil \frac{n}{k} \rceil - 1$ **and** $\nexists_{v_2 \in X}(v_1, v_2) \in E$ **then**
6: $\quad\quad$ **for** every $v_2 \in X \cap V(T_\ell)$ **do**
7: $\quad\quad\quad$ **if** $|\mathrm{ST}(v_2)| > \lceil \frac{n}{k} \rceil$ **then**
8: $\quad\quad\quad\quad X \leftarrow X \cup \{v_1\}$
9: $\quad\quad\quad\quad$ continue with the next Trémaux tree $T_{\ell'}$
10: $\quad\quad$ **if** $\exists_{v_2 \in X} \exists_{v_3, w_3 \in \mathrm{ST}(v_2)} (v_1, v_3), (v_1, w_3) \in E$ **then**
11: $\quad\quad\quad$ **let** $v_2 \in X$ be the vertex for which the condition holds
12: $\quad\quad\quad X \leftarrow (X \setminus \{v_2\}) \cup \{v_1\}$
13: $\quad\quad$ **else**
14: $\quad\quad\quad$ **for** every $v_2 \in X \cap V(T_\ell)$ **do**
15: $\quad\quad\quad\quad$ **let** $v_3 \in \mathrm{ST}(v_2)$ be the (only) vertex with $(v_1, v_3) \in E$
16: $\quad\quad\quad\quad X \leftarrow (X \setminus \{v_2\}) \cup \{v_3\}$
17: **return** X

Nun beweisen wir, dass Algorithmus 1 ein Exact Proportional Clustering ausgibt. Wegen Lemma 3.1 reicht es, einen Trémaux Baum T_ℓ einer beliebigen Zusammenhangskomponente G_ℓ von G zu betrachten. Sei T_ℓ also ein Trémaux Baum einer beliebigen Zusammenhangskomponente von G. Sei Algorithmus 1 bis zur Zeile 2 ausgeführt worden.

Fall 1. Sei $v_1 \in X$. Nach der Definition der Baumtiefe gibt es keine Kanten zwischen Knoten aus verschiedenen Teilbäumen $\mathrm{ST}(v_2), \mathrm{ST}(w_2)$, also gilt für alle v_2, w_2, mit $v_2 \neq w_2$: $\nexists_{v \in \mathrm{ST}(v_2), w \in \mathrm{ST}(w_2)}(v, w) \in E$. Weil $v_1 \in X$ und wegen Lemma 3.1 können wir also die Teilbäume $\mathrm{ST}(v_2)$ einzeln betrachten. Sei $S \subseteq \mathrm{ST}(v_2)$ eine Blocking Coalition. Falls $|\mathrm{ST}(v_2)| < \lceil \frac{n}{k} \rceil$ gilt, hat $\mathrm{ST}(v_2)$ nicht genug Knoten, um eine Blocking Coalition zu bilden. Falls $|\mathrm{ST}(v_2)| \geq \lceil \frac{n}{k} \rceil$ gilt, eröffnet Algorithmus 1 einen Cluster Center im Zentrum v_2 des Sterns $G[\mathrm{ST}(v_2)]$. Dies ist ein Dominating Set[2] für den Stern, also haben alle $i \in S$ Distanz maximal 1 zum nächstgelegenen Cluster Center und müssten Distanz 0 erreichen, um wechseln zu wollen. Also kann sich auch hier keine Blocking Coalition bilden.

Fall 2. Sei $v_1 \notin X$. Sei S eine Blocking Coalition und sei A wie im Algorithmus 1 definiert. (A entspricht der Menge G^0 im Algorithmus 1 von Micha und Shah [MS20].) Wenn $|A| \geq \lceil \frac{n}{k} \rceil$ ist, dann hätte Algorithmus 1 einen Cluster Center in v_1 eröffnet, was ein Widerspruch ist. Sei also $|A| < \lceil \frac{n}{k} \rceil$. Da $v_1 \notin X$ hatte Algorithmus 1 nur Cluster Center in Knoten v_2, mit $\mathrm{Level}(v_2) = 2$ eröffnet. Betrachten wir den Graphen $G_\ell - X$, so sind neben A nur Blattknoten v_3 aus Teilbäumen $\mathrm{ST}(v_2)$, mit $v_2 \in X$ übrig. Da für diese Blattknoten v_3 gilt: $d(v_3, X) = 1$, können nicht mehrere dieser Blattknoten in S sein, da

[2] Ein Dominating Set ist eine Menge X von Knoten, sodass jeder andere Knoten benachbart zu einem Knoten in X ist.

sonst einer von ihnen nicht wechseln wollen würde. D. h. es gilt $\left|S \cap \bigcup_{v_2 \in X} \mathrm{ST}(v_2)\right| \leq 1$. Zusammen mit $|A| < \lceil \frac{n}{k} \rceil$ muss die Blocking Coalition S nun wie folgt aussehen, damit $|S| \geq \lceil \frac{n}{k} \rceil$ gilt: $|A| = \lceil \frac{n}{k} \rceil - 1$ und $\exists_{v_3} S \cap \left(\bigcup_{v_2 \in X} \mathrm{ST}(v_2)\right) = \{v_3\}$ und $S = A \cup \{v_3\}$. Da v_3 schon Distanz 1 zum nächstgelegenen Cluster Center hat, darf v_1 nicht auch Distanz 1 haben. Also darf v_1 keine Kanten zu einem der Cluster Center v_2 haben, d. h. es gilt $\nexists_{v_2 \in X}(v_1, v_2) \in E$. Außerdem gilt, dass alle anderen Knoten $a \in A \setminus \{v_1\}$ wegen der Definition von Baumtiefe nicht zu einem $v_2 \in X$ benachbart sein können.

Falls es ein $v_2 \in X \cap V(T_\ell)$ mit $|\mathrm{ST}(v_2)| > \lceil \frac{n}{k} \rceil$ gibt, wird v_1 zu X hinzugefügt und es kann sich wie in Fall 1 keine Blocking Coalition bilden. Falls nicht, dann gilt für alle $v_2 \in X \cap V(T_\ell)$: $|\mathrm{ST}(v_2)| = \lceil \frac{n}{k} \rceil$.

Falls es ein $v_2 \in X$ mit mehreren Kindern $v_3, w_3 \in \mathrm{ST}(v_2)$ gibt, die zur Wurzel v_1 benachbart sind, also $(v_1, v_3), (v_1, w_3) \in E$, dann wird statt diesem Cluster Center v_2 die Wurzel v_1 zum neuen Cluster Center. Für alle anderen $w_2 \in V(T_\ell)$ kann sich wie in Fall 1 keine Blocking Coalition in $\mathrm{ST}(w_2)$ bilden und für das v_2 gilt für die mögliche Blocking Coalition $S \subseteq \mathrm{ST}(v_2)$: $S = \mathrm{ST}(v_2)$, weil $|\mathrm{ST}(v_2)| = \lceil \frac{n}{k} \rceil$. Aber in S gibt es nun zwei Knoten (nämlich v_3 und w_3) mit Distanz 1 zu einem nächstgelegenen Cluster Center (nämlich v_1), also würde einer von ihnen nicht wechseln wollen.

Falls es kein $v_2 \in X$ mit mehreren Kindern $v_3, w_3 \in \mathrm{ST}(v_2)$ gibt, die zur Wurzel v_1 benachbart sind, gilt für alle $v_2 \in X \cap V(T_\ell)$: Es gibt genau einen Kind $w_3 \in \mathrm{ST}(v_2)$, das zur Wurzel v_1 benachbart ist, also $\exists!_{v_3 \in \mathrm{ST}(v_2)}(v_1, v_3) \in E$. Denn sonst wäre $G[V(T_\ell)]$ nicht zusammenhängend, da es zwischen der Wurzel v_1 und den $v_2 \in X$ wegen Zeile 5 keine Kanten gibt. Dann wird in jedem $\mathrm{ST}(v_2)$, mit $v_2 \in X \cap V(T_\ell)$, dasjenige Kind v_3, mit $(v_1, v_3) \in E$, statt v_2 zum Center gewählt. Nun gilt für den originalen Graphen $G[V(T_\ell)]$, dass jede Zusammenhangskomponente in $G[V(T_\ell)] - X$ weniger als $\lceil \frac{n}{k} \rceil$ Knoten hat, weil 1) für jedes $v_2 \in X \cap V(T_\ell)$ $|\mathrm{ST}(v_2)| = \lceil \frac{n}{k} \rceil$ gilt und die einzige Verbindung zum Rest des Graphen über dem Kind v_3, das zur Wurzel v_1 benachbart ist, geht, welches wegen Algorithmus 1 nun in X ist und 2) $|A| = \lceil \frac{n}{k} \rceil - 1$ gilt. Also kann S wegen Lemma 3.1 und der Bedingung $S \geq \lceil \frac{n}{k} \rceil$ keine Blocking Coalition sein. □

Für zusammenhängende Graphen mit Baumtiefe 4 kann einfach der unmodifizierte Algorithmus 1 von Micha und Shah [MS20] auf dem Trémaux Baum angewendet werden und die möglichen Blocking Coalitions in den $\mathrm{ST}(v_2)$ wie in Algorithmus 1 dann einzeln aufgelöst werden. Das funktioniert, weil die Wurzel immer im Clustering ist [MS20] und wegen Lemma 3.1 und der Definition von Baumtiefe eine Blocking Coalition nicht auf mehrere Zusammenhangskomponenten, also den $\mathrm{ST}(v_2)$, verteilt sein kann.

Korollar 4.2. *Sei G ein zusammenhängender Graph mit Baumtiefe $\mathrm{BT}(G) = 4$. Dann hat G ein Exact Proportional Clustering.*

Kapitel 5

Lowerbounds

In diesem Abschnitt beweisen wir, dass es für allgemeine Graphen nicht immer ein $(2-\epsilon)$-Proportional Clustering gibt. Das impliziert direkt, dass der Lowerbound für beliebige metrische Räume, auch für den Fall $\mathcal{N} = \mathcal{M}$ schon 2 ist. Das verbessert den zuvor bekannten Lowerbound von $1,5$ von Chen et al. [Che+19], sodass dieser mit dem Lowerbound für Clusterings im allgemeinen Fall $\mathcal{N} \neq \mathcal{M}$ übereinstimmt.

Das Beispiel hat einige schöne Grapheigenschaften, wie eine kleine Baumweite und Planarität und liefert bereits für kleine Baumtiefen einen Lowerbound $\rho > 1$. Das Beispiel liefert, etwas abgewandelt, den gleichen Lowerbound für serien-parallele Graphen und für Graphen mit maximalem Knotengrad 3.

5.1 Existenz von Exact Propotional Clusterings

Die Kernidee des Beweises besteht darin, eine Instanz für $\mathcal{N} \subseteq \mathcal{M}$ (mit nicht-Graph-Metrik), in der es kein $(2-\epsilon)$-Proportional Clustering gibt, in eine Instanz für $\mathcal{N} = \mathcal{M}$ (mit Graph-Metrik) zu transformieren. Die Datenpunkte ersetzen wir mit Sternen und die Distanzen mit langen Pfaden. Wegen der Bedingung $\mathcal{N} = \mathcal{M}$ verlagert sich das Gewicht der Datenpunkte $\mathcal{N}$ bei großen Distanzen auf die Pfade. Um dem entgegenzuwirken, platzieren wir bei den originalen Punkten eine hohe Anzahl an Knoten, d. h. sie werden zu Sternen. Für immer länger werdende Pfade gleicht sich der Lowerbound für Graphen dem der Instanz für $\mathcal{N} \subseteq \mathcal{M}$ an.

Instanz für $\mathcal{N} \subseteq \mathcal{M}$. Sei $k = 5$. Es gibt drei identische Gruppen mit je 3 Punkten, also $n = 9$. Eine Gruppe sieht wie folgt aus. Die drei Punkte a_1, a_2, a_3 liegen je auf einem der Eckpunkte eines gleichseitigen Dreiecks mit Seitenlänge 1. Sei $\mathcal{M}$ die Menge aller Punkte der Dreiecke (Seiten und Eckpunkte). Die Distanz zwischen zwei Punkten innerhalb einer Gruppe ist die Länge eines kürzesten Streckenzugs, der den Seiten des Dreiecks folgt. Die Distanz zwischen Punkten aus verschiedenen Gruppen ist unendlich. Abbildung 5.1 zeigt beispielhaft eine der Gruppen.

Da $k = 5$ und es 3 Gruppen gibt, gibt es in einer der Gruppen maximal einen Center x. Diese Gruppe wird nun betrachtet. Da die Instanz symmetrisch ist, sei o. B. d. A. x auf der Strecke von a_1 nach a_2 und näher oder gleich nah an a_1. Sei $\delta_x = d(a_1, x)$, also $\delta_x \in [0; 0,5]$. Da $\left\lceil \frac{n}{k} \right\rceil = \left\lceil \frac{9}{5} \right\rceil = 2$, ist jede Menge mit 2 Punkten berechtigt, eine Blocking

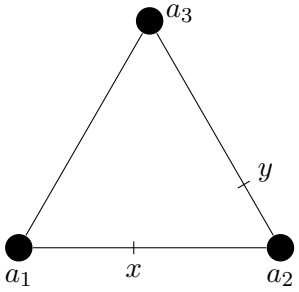

Abbildung 5.1: Eine Gruppe der Instanz für $\mathcal{N} \subseteq \mathcal{M}$ mit Auswahl von y, für ein x.

Coalition zu bilden. Wähle $S = \{a_2, a_3\}$ als Blocking Coalition und y auf der Strecke von a_2 nach a_3 mit Distanz $\delta_y = d(a_2, y)$ zu a_2. Wir wählen δ_y so, dass es die minimale multiplikative Verbesserung von a_2 und a_3 maximiert. Sei also

$$\begin{aligned}\delta_y &= \operatorname*{argmax}_{\delta_y} \left(\min \left(\frac{d(a_2, x)}{d(a_2, y)}, \frac{d(a_3, x)}{d(a_3, y)} \right) \right) \\ &= \operatorname*{argmax}_{\delta_y} \left(\min \left(\frac{1 - \delta_x}{\delta_y}, \frac{1 + \delta_x}{1 - \delta_y} \right) \right).\end{aligned}$$

Da $\frac{1-\delta_x}{\delta_y}$ als Funktion von δ_y monoton fallend ist und $\frac{1+\delta_x}{1-\delta_y}$ als Funktion von δ_y monoton wachsend und keine der Funktionen immer größer als die andere ist, berechnen wir den Schnittpunkt der Funktionen:

$$\begin{aligned}\frac{1 - \delta_x}{\delta_y} &= \frac{1 + \delta_x}{1 - \delta_y} \\ \Leftrightarrow (1 - \delta_x)(1 - \delta_y) &= (1 + \delta_x)\delta_y \\ \Leftrightarrow 1 - \delta_y - \delta_x + \delta_x\delta_y &= d_y + \delta_x\delta_y \\ \Leftrightarrow 1 - \delta_x &= 2\delta_y \\ \Leftrightarrow \delta_y &= \frac{1 - \delta_x}{2}. \end{aligned} \tag{5.1}$$

Setzten wir nun Gleichung (5.1) in einer der Verbesserungen von a_2 oder a_3 ein (hier für a_2), so gilt

$$\rho = \frac{1 - \delta_x}{\left(\frac{1-\delta_x}{2}\right)} = 2.$$

Also gibt es in dieser Instanz kein ρ-Proportional Clustering, für $\rho < 2$.

Nun transformieren wir die Instanz in einen Graphen, indem wir die Seiten des Dreiecks durch lange Pfade darstellen und die Eckpunkte des Dreiecks, also die Elemente aus $\mathcal{N}$, durch Sterne. Verlängern wir die Pfade beliebig groß, so erreichen wir den gleichen Lowerbound.

Satz 5.1. *Für alle $\rho < 2$ gibt es nicht immer ein ρ-Proportional Clustering für Graphen.*

Beweis. Sei $n = 9\alpha + 9\delta$ und $k = 5$. Es gibt drei identische Gruppen (Zusammenhangskomponenten) mit je $3\alpha + 3\delta$ Knoten. Eine Gruppe sieht wie folgt aus. Es gibt zunächst

drei spezielle Typen von Knoten, a_1^*, a_2^*, a_3^*. Ein a_i^* zusammen mit α weiteren Knoten vom Typ a_i bilden einen Stern mit a_i^* als Zentrum, für $i \in \{1,2,3\}$. Die restlichen $3\delta - 3$ Knoten werden benutzt, um je a_i^* mit $a_{i+1 \bmod 3}^*$ durch einen Pfad der Länge δ zu verbinden, für $i \in \{1,2,3\}$. Die neuen Knoten auf dem Pfad von a_1^* nach a_2^* heißen $p_1, \ldots, p_{\delta-1}$, die von a_2^* nach a_3^* heißen $q_1, \ldots, q_{\delta-1}$ und die von a_3^* nach a_1^* heißen $r_1, \ldots, r_{\delta-1}$. Abbildung 5.2 zeigt diese Konstruktion. Durch die zusätzlichen Pfadknoten p, q, r, also ein größeres n, als im Fall $\mathcal{N} \subseteq \mathcal{M}$, soll es folgende Bedingung an α geben: $2\alpha \geq \lceil \frac{n}{k} \rceil = \lceil \frac{9\alpha+9\delta}{5} \rceil$, also

$$2\alpha \overset{!}{\geq} \frac{9\alpha + 9\delta}{5} + 1 \Leftrightarrow \alpha \geq 9\delta + 5 \tag{5.2}$$

Da $k = 5$ und es 3 Gruppen gibt, gibt es in einer der Gruppen maximal einen Center x. Diese Gruppe wird nun betrachtet. Da die Instanz symmetrisch ist, sei o. B. d. A. x auf dem Pfad von a_1^* nach a_2^* und näher oder gleich nah an a_1^* oder sei $x = a_1$ ein Sternknoten. Es soll p_0 den Knoten a_1^* bezeichnen und q_0 den Knoten a_2^*.

Fall 1. Sei $x = p_i$, für $i \in \{0, \ldots, \lfloor \frac{\delta}{2} \rfloor\}$. Da wegen Ungleichung (5.2) $2\alpha \geq \lceil \frac{n}{k} \rceil$ gilt, ist jede Menge mit 2α Knoten berechtigt, eine Blocking Coalition zu bilden. Wähle $S = \{a_2, a_3\}$ als Blocking Coalition und $y = q_j$, mit $j = \lfloor \frac{\delta - i}{2} \rfloor$. Dann gilt

$$\begin{aligned}
\rho &= \min\left(\frac{d(a_2,x)}{d(a_2,y)}, \frac{d(a_3,x)}{d(a_3,y)}\right) = \min\left(\frac{\delta - i + 1}{j+1}, \frac{\delta + i + 1}{\delta - j + 1}\right) \\
&= \min\left(\frac{\delta - i + 1}{\lfloor \frac{\delta - i}{2} \rfloor + 1}, \frac{\delta + i + 1}{\delta - \lfloor \frac{\delta - i}{2} \rfloor + 1}\right).
\end{aligned}$$

Wählen wir δ beliebig groß, so gilt im Grenzwert mithilfe des Sandwich-Theorems und weil $i \in \{0, \ldots, \lfloor \frac{\delta}{2} \rfloor\}$, also $\delta - i \geq \frac{\delta}{2}$ und $\lim_{\delta \to \infty} \frac{c}{\delta - i} = 0$, für alle $c \in \mathbb{N}$:

$$\begin{aligned}
2 &= \min\left(\lim_{\delta\to\infty} \frac{2 + \frac{2}{\delta - i}}{1 + \frac{2}{\delta - i}}, \lim_{\delta\to\infty} \frac{2 + \frac{2}{\delta + i}}{1 + \frac{4}{\delta + i}}\right) \\
&= \min\left(\lim_{\delta\to\infty} \frac{2\delta - 2i + 2}{(\delta - i) + 2}, \lim_{\delta\to\infty} \frac{2\delta + 2i + 2}{2\delta - (\delta - i - 2) + 2}\right) \\
&= \min\left(\lim_{\delta\to\infty} \frac{\delta - i + 1}{\left(\frac{\delta - i}{2}\right) + 1}, \lim_{\delta\to\infty} \frac{\delta + i + 1}{\delta - \left(\frac{\delta - i}{2} - 1\right) + 1}\right) \\
&\leq \min\left(\lim_{\delta\to\infty} \frac{\delta - i + 1}{\lfloor \frac{\delta - i}{2} \rfloor + 1}, \lim_{\delta\to\infty} \frac{\delta + i + 1}{\delta - \lfloor \frac{\delta - i}{2} \rfloor + 1}\right) \\
&\leq \min\left(\lim_{\delta\to\infty} \frac{\delta - i + 1}{\left(\frac{\delta - i}{2} - 1\right) + 1}, \lim_{\delta\to\infty} \frac{\delta + i + 1}{\delta - \left(\frac{\delta - i}{2}\right) + 1}\right) \\
&= \min\left(\lim_{\delta\to\infty} \frac{2\delta - 2i + 2}{(\delta - i - 2) + 2}, \lim_{\delta\to\infty} \frac{2\delta + 2i + 2}{2\delta - (\delta - i) + 2}\right) \\
&= \min\left(\lim_{\delta\to\infty} \frac{2 + \frac{2}{\delta - i}}{1}, \lim_{\delta\to\infty} \frac{2 + \frac{2}{\delta + i}}{1 + \frac{2}{\delta + i}}\right) = 2.
\end{aligned}$$

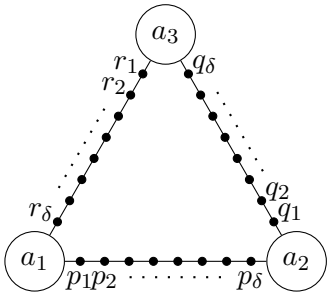

Abbildung 5.2: Eine Gruppe der Instanz für $\mathcal{N} = \mathcal{M}$ mit Graphmetrik. Die Kreise von den Knoten a stellen Sterne dar.

Also gilt für den Verbesserungsfaktor

$$\rho = \min\left(\lim_{\delta\to\infty} \frac{d(a_2,x)}{d(a_2,y)}, \lim_{\delta\to\infty} \frac{d(a_3,x)}{d(a_3,y)}\right) = 2.$$

Fall 2. Sei $x = a_1$. Dann wähle $S = \{a_2, a_3\}$ und $y = q_j$, mit $j = \lfloor \frac{\delta}{2} \rfloor$. Dann ist

$$\begin{aligned}\rho &= \min\left(\lim_{\delta\to\infty} \frac{d(a_2,x)}{d(a_2,y)}, \lim_{\delta\to\infty} \frac{d(a_3,x)}{d(a_3,y)}\right) \\ &= \min\left(\lim_{\delta\to\infty} \frac{\delta+2}{\lfloor \frac{\delta}{2} \rfloor + 1}, \lim_{\delta\to\infty} \frac{\delta+2}{\delta - \lfloor \frac{\delta}{2} \rfloor + 1}\right) = 2\end{aligned}$$

nach ähnlichen Berechnungen wie im Fall 1.

Also gibt es für alle $\rho < 2$ nicht immer ein ρ-Proportional Clustering. □

Das Ergebnis impliziert direkt den gleichen Lowerbound für beliebige Metriken im Fall $\mathcal{N} = \mathcal{M}$. Wir erhalten daraus folgendes Korollar.

Korollar 5.2. *Für alle $\rho < 2$ gibt es nicht immer ein ρ-Proportional Clustering für beliebige metrische Räume im Spezialfall $\mathcal{N} = \mathcal{M}$.*

5.2 Baumparameter des Gegenbeispiels

In Abschnitt 3.1 und Kapitel 4 haben wir gezeigt, dass Graphen mit Baumweite 1 und Graphen mit Baumtiefe 3 immer ein Exact Proportional Clustering haben. Wir zeigen hier, dass dies für die nächstgrößere Baumweite 2 und Baumtiefe 4 nicht mehr der Fall ist, indem wir die Baumweite und Baumtiefe der Instanz aus Satz 5.1 berechnen.

Proposition 5.3. *Der Graph G in Satz 5.1 hat für ein beliebiges δ Baumweite 2 und für alle $\rho < 2$ gibt es nicht immer ein ρ-Proportional Clustering für Graphen mit Baumweite 2.*

Beweis. Wir zeigen, dass $\mathrm{BW}(G) = 2$ gilt. Dann folgt der zweite Teil der Proposition direkt. **Teil 1.** Wir zeigen zuerst $\mathrm{BW}(G) \leq 2$, indem wir eine Baumzerlegung mit

Weite 2, für eine der Zusammenhangskomponenten G' von G angeben.[1] Es soll p_δ und q_0 den Knoten a_2^* bezeichnen und q_δ und r_0 den Knoten a_3^*. Die Sternknoten heißen $a_{i,j}$, mit $j \in \{1, \dots, \alpha\}$. Die Baumzerlegung von G' ist $T = (V, E)$, wobei V und E wie folgt definiert sind:
Sei $V = V' \cup V_1 \cup V_2 \cup V_3$, wobei:

- $V' = \{B'_{i,j} : i \in \{1,2,3\}, j \in \{1, \dots, \alpha\}\}$, wobei $B'_{i,j} = \{a_i^*, a_{i,j}\}$, für $i \in \{1,2,3\}$ und $j \in \{1, \dots, \alpha\}$,
- $V_1 = \{B_{1,j} : j \in \{1, \dots, \delta - 1\}\}$, wobei $B_{1,j} = \{a_1^*, p_j, p_{j+1}\}$, für $j \in \{1, \dots, \delta - 1\}$,
- $V_2 = \{B_{2,j} : j \in \{0, \dots, \delta - 1\}\}$, wobei $B_{2,j} = \{a_1^*, q_j, q_{j+1}\}$, für $j \in \{0, \dots, \delta - 1\}$, und
- $V_3 = \{B_{3,j} : j \in \{0, \dots, \delta - 2\}\}$, wobei $B_{3,j} = \{a_1^*, r_j, r_{j+1}\}$, für $j \in \{0, \dots, \delta - 2\}$.

Sei $E = E' \cup E_1 \cup E_2 \cup E_3 \cup \tilde{E}$, wobei:

- $E' = \{\{B_{1,1}, B'_{1,j}\} : j \in \{1, \dots, \alpha\}\} \cup \{\{B_{i,0}, B'_{i,j}\} : i \in \{2,3\}, j \in \{1, \dots, \alpha\}\}$,
- $E_1 = \{\{B_{1,j}, B_{1,j+1}\} : j \in \{1, \dots, \delta - 2\}\}$,
- $E_2 = \{\{B_{2,j}, B_{2,j+1}\} : j \in \{0, \dots, \delta - 2\}\}$,
- $E_3 = \{\{B_{3,j}, B_{3,j+1}\} : j \in \{0, \dots, \delta - 3\}\}$ und
- $\tilde{E} = \{\{B_{1,\delta-1}, B_{2,0}\}, \{B_{2,\delta-1}, B_{3,0}\}\}$.

Es ist leicht zu sehen, dass T eine Baumzerlegung für G' mit Weite 2 ist.

Teil 2. Nun zeigen wir, dass $\mathrm{BW}(G) > 1$, also $\mathrm{BW}(G) \geq 2$. Da G einen Kreis $(a_1^*, p_1, \dots, a_2^*, q_1, \dots, a_3^*, r_1, \dots, r_{\delta-1}, a_1^*)$ enthält, kann G nicht Baumweite 1 haben. □

Für größer werdende δ gibt es in der Instanz den länger werdenden Pfad $(a_1^*, p_1, \dots, a_2^*, q_1, \dots, a_3^*, r_1, \dots, r_{\delta-1})$. Lange Pfade sind verbotene Minoren für die Baumtiefe [NDM06]. Wir zeigen aber, dass die Instanz bereits für $\delta = 2$ einen Lowerbound $\rho > 1$ liefert und Baumtiefe 4 hat.

Proposition 5.4. *Der Graph $G = (V, E)$ in Satz 5.1 hat für $\delta = 2$ Baumtiefe 4 und für Graphen mit Baumtiefe 4 gibt es nicht immer ein Exact Proportional Clustering.*

Beweis. Wir zeigen zuerst, dass G kein Exact Proportional Clustering hat. Wir betrachten wie in Satz 5.1 die Gruppe mit nur einem Center x. Es wird eine Fallunterscheidung gemacht, wo x liegen kann:

- Sei $x = a_1$. Dann gilt für $S = \{a_2, a_3\}$ und $y = q_1$, dass

$$\rho = \min\left(\frac{d(a_2, x)}{d(a_2, y)}, \frac{d(a_3, x)}{d(a_3, y)}\right) = \min\left(\frac{4}{2}, \frac{4}{2}\right) = 2.$$

[1] Die einzelnen Baumzerlegungen der Zusammenhangskomponenten ergeben einen Wald und können einfach mit zwei Kanten zu einem Baum erweitert werden, ohne die Baumzerlegungseigenschaften zu verletzen.

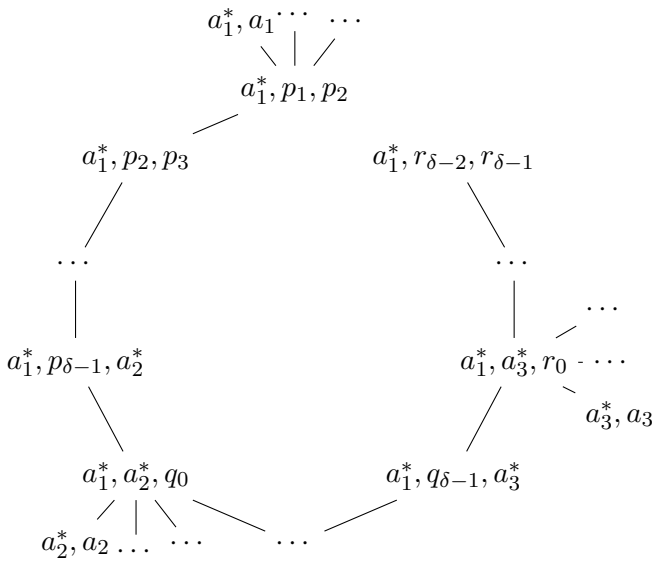

Abbildung 5.3: Baumzerlegung einer Zusammenhangskomponente von G.

- Sei $x = a_1^*$. Dann gilt für $S = \{a_2, a_3\}$ und $y = q_1$, dass
$$\rho = \min\left(\frac{d(a_2,x)}{d(a_2,y)}, \frac{d(a_3,x)}{d(a_3,y)}\right) = \min\left(\frac{3}{2}, \frac{3}{2}\right) = 1,5.$$
- Sei $x = p_1$. Dann gilt für $S = \{a_2, a_3\}$ und $y = a_2^*$, dass
$$\rho = \min\left(\frac{d(a_2,x)}{d(a_2,y)}, \frac{d(a_3,x)}{d(a_3,y)}\right) = \min\left(\frac{2}{1}, \frac{4}{3}\right) = 1,\overline{3}.$$

Die anderen Fälle sind symmetrisch. Also gibt es für G kein Exact Proportional Clustering.

Nun zeigen wir, dass $\mathrm{BT}(G) = 4$ gilt. Dann folgt der zweite Teil der Proposition direkt. Da G nicht zusammenhängend ist und die Zusammenhangskomponenten von G aber isomorph sind, reicht es aus, die Baumtiefe von einer der Zusammenhangskomponenten G' von G zu berechnen. **Teil 1.** Wir zeigen zuerst $\mathrm{BW}(G') \leq 4$, indem wir einen Trémaux Baum mit Tiefe 4 angeben. Der Trémaux Baum von G' ist $T = (V, \tilde{E})$, wobei $\tilde{E}$ wie folgt definiert ist: $\tilde{E} = \{\{a_1^*, a_2^*\}\} \cup \{\{a_3^*, v\} : v \in V \setminus \{a_1^*, a_3^*\}\}$. Es ist leicht zu sehen, dass T ein Trémaux Baum von G' mit Tiefe 4 ist.

Teil 2. Nun zeigen wir, dass $\mathrm{BT}(G') > 3$. Es gilt per Definition der Baumtiefe: $\mathrm{BT}(G') = 1 + \min_{v \in V(G')} \mathrm{BT}(G' - v)$. Da G' einen Kreis C_6 enthält, gilt für alle $v \in V(G')$, dass $G' - v$ einen Pfad P_4 enthält, was ein verbotener Minor für die Baumtiefe 2 ist [NDM06]. Also gilt für alle $v \in V(G')$: $\mathrm{BT}(G' - v) > 2$. Also gilt

$$\mathrm{BT}\left(G'\right) = 1 + \min_{v \in V(G')} \mathrm{BT}\left(G' - v\right) > 1 + 2 = 3.$$

□

Es ist anzumerken, dass dies kein Gegenbeispiel zu Korollar 4.2 ist, da der Graph G nicht zusammenhängend ist.

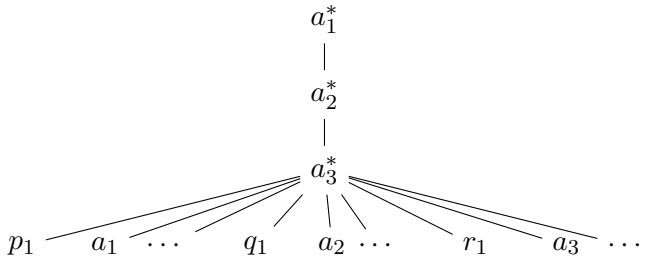

Abbildung 5.4: Trémaux Baum einer Zusammenhangskomponente von G.

5.3 Abwandlung des Beispiels zu anderen Graphparametern

Im vorigen Abschnitt 5.2 haben wir gezeigt, dass es für die Baumweite 2 nicht immer ein Exact Proportional Clustering gibt. Nun stellt sich die Frage, ob es bei serien-parallelen Graphen, die eine einfachere Struktur als allgemeine Graphen mit Baumweite 2 haben, immer möglich ist, ein Exact Proportional Clustering zu finden oder ob zumindest der Lowerbound kleiner ist. Hier zeigen wir, dass auch bei serien-parallelen Graphen der Lowerbound 2 bleibt. Die Idee besteht darin, statt Sterne, Diamant-Graphen zu nutzen und die Gruppen durch lange Pfade zu verbinden. Wir zeigen außerdem, dass bei Graphen mit maximalem Knotengrad 3 der Lowerbound 2 bleibt, mit der gleichen Idee, die Sterne zu ersetzen.

Proposition 5.5. *Für alle $\rho < 2$ gibt es nicht immer ein ρ-Proportional Clustering für serien-parallele Graphen.*

Beweis. Sei $n = 9\alpha + 21\delta + 7$ und $k = 5$. Es gibt drei identische Gruppen mit je $3\alpha + 3\delta + 3$ Knoten. Eine Gruppe sieht wie folgt aus. Es gibt zunächst sechs spezielle Typen von Knoten, $a_1^*, a_2^*, a_3^*, b_1^*, b_2^*, b_3^*$. Die Knoten a_i^* und b_i^* zusammen mit α weiteren Knoten vom Typ a_i bilden einen Diamant-Graphen $D_\alpha = (V_\alpha, E_\alpha)$, wobei $V_\alpha = \{a_i^*, b_i^*\} \cup \{a_{i,j} : j \in \{1, \dots, \alpha\}\}$ und $E_\alpha = \{\{u, a_i\} : u \in \{a_i^*, b_i^*\}\}$, für $i \in \{1, 2, 3\}$. Die restlichen $3\delta - 3$ Knoten werden benutzt, um a_1^* und b_2^*, a_2^* und b_3^*, und b_3^* und a_1^* durch je einen Pfad der Länge δ zu verbinden.

Die restlichen $12\delta - 2$ Knoten werden benutzt, um b_1^* der ersten Gruppe mit a_3^* der zweiten Gruppe und b_1^* der zweiten Gruppe mit a_3^* der dritten Gruppe und durch je einen Pfad P_1 bzw. P_2 der Länge 6δ zu verbinden. Setze $s = a_3^*$ der ersten Gruppe und $t = b_1^*$ der dritten Gruppe. Es ist leicht zu sehen, dass $G = (V, E, s, t)$ ein serien-paralleler Graph ist, da G eine Serien- und Parallelkomposition von Diamant-Graphen und Pfaden ist. Abbildung 5.5 zeigt beispielhaft eine der Gruppen und wie sie erzeugt werden kann.

Wie im Satz 5.1 soll es folgende Bedingung an α geben: $2\alpha \geq \lceil \frac{n}{k} \rceil = \lceil \frac{9\alpha+21\delta+7}{5} \rceil$, also

$$2\alpha \overset{!}{\geq} \frac{9\alpha + 21\delta + 7}{5} + 1 \Leftrightarrow \alpha \geq 21\delta + 12.$$

Da $k = 5$ und es 3 Gruppen gibt, gibt es in einer der Gruppen maximal einen Center x, wobei wir Knoten auf dem Pfad P zu der Gruppe zählen, zu der sie näher dran ist.

Diese Gruppe wird nun betrachtet. Falls in dieser Gruppe (ohne die Pfadknoten P) ein Center eröffnet wurde, dann gilt $d(a_i, X) \leq 3\delta + 1$.[2] Also bezieht keiner der Knoten a_i seine Distanz zu einem Center aus einer anderen Gruppe, da $d(a_i, v) \geq 3\delta+1$ für Knoten v aus anderen Gruppen gilt. Nun gelten im Grenzwert von $\delta \to \infty$ die gleichen Rechnungen wie im Satz 5.1, da sich die Distanzen nur in additiven Konstanten unterscheiden. Also ist der Grenzwert des Verbesserungsfaktors für $\delta \to \infty$ bei beiden Instanzen gleich.

Falls x auf einem der Pfade P liegt, dann gelten die Rechnungen, als wenn $x = a_3^*$ bzw. $x = b_1^*$ gelten würde, da für die Blocking Coalition $d(a_i, a_3^*) < d(a_i, x)$ bzw. $d(a_i, b_1^*) < d(a_i, x)$ gilt und der Verbesserungsfaktor auch hier dann mindestens 2 ist. □

Für Graphen mit maximalem Knotengrad 3 verwenden wir die gleiche Idee, die Sterne zu ersetzen. Wir benutzen hier Bäume, damit die Distanz von den α Knoten a_i zum Verbindungsknoten a_i^* logarithmisch (in δ) ist und folglich im Grenzwert verschwindet. Wir zeigen das Ergebnis für einen allgemeinen maximalen Knotengrad größer gleich 3.

Proposition 5.6. *Für alle $\rho < 2$ gibt es nicht immer ein ρ-Proportional Clustering für Graphen mit beschränktem Knotengrad $\Delta \geq 3$.*

Beweis. Sei $n = 9\alpha + 9\delta$ und $k = 5$. Sei $\Delta \geq 3$ der maximale Knotengrad, den der Graph haben soll. Die Konstruktion des Graphen ist bis auf die Sterne identisch zu dem Graphen aus Satz 5.1. Anstatt die α Knoten a_i zu dem a_i^* zu verbinden, werden wir nun die α Knoten a_i benutzen, um einen Baum zu erstellen, in dem jeder Knoten maximal $\Delta - 1$ Kinder hat, für $i \in \{1, 2, 3\}$. Diese Bäume haben Tiefe $\lfloor \log_{\Delta-1} \alpha \rfloor$. Wir verbinden die Wurzel des Baumes mit a_i^*, für $i \in \{1, 2, 3\}$. Wegen Ungleichung (5.2) wählen wir $\alpha = 9\delta + 5$. Dann ist der maximale Abstand eines a_i zu seinem a_i^*:

$$\begin{aligned} \max_{a_i} d(a_i, a_i^*) &= 1 + \lfloor \log_{\Delta-1} \alpha \rfloor = 1 + \lfloor \log_{\Delta-1}(9\delta + 5) \rfloor \\ &\in \mathcal{O}(\log_{\Delta-1} \delta) = \mathcal{O}(\log \delta). \end{aligned} \tag{5.3}$$

Da $k = 5$ und es 3 Gruppen gibt, gibt es in einer der Gruppen maximal ein Center x. Diese Gruppe wird nun betrachtet. Da die Instanz symmetrisch ist, sei o. B. d. A. x auf dem Pfad von a_1^* nach a_2^* und näher oder gleich nah an a_1^* oder $x = a_1$, für eines der Knoten aus dem Baum. Es soll p_0 den Knoten a_1^* bezeichnen und q_0 den Knoten a_2^*.

Fall 1. Sei $x = p_i$, für $i \in \{0, \ldots, \lfloor \frac{\delta}{2} \rfloor\}$. Da wegen Ungleichung (5.2) $2\alpha \geq \lceil \frac{n}{k} \rceil$ gilt, ist jede Menge mit 2α Knoten berechtigt, eine Blocking Coalition zu bilden. Wähle wie

[2] Das ist eine grobe Abschätzung mit dem längsten Pfad in einer Gruppe (ohne die Pfadknoten P), was eine obere Schranke für die maximale Distanz ist.

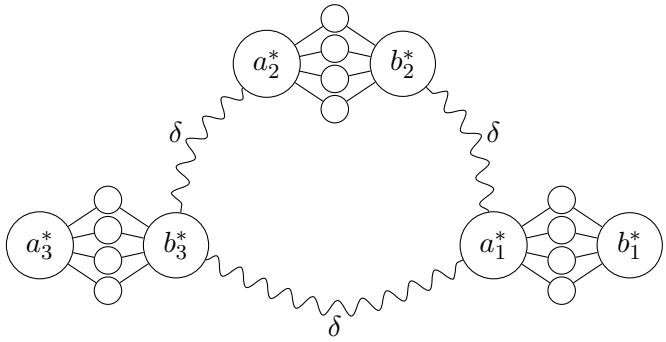

Abbildung 5.5: Eine Gruppe vom serien-parallelen Graph, der den gleichen Lowerbound wie der Graph in Satz 5.1 liefert. Der Graph kann wie folgt erzeugt werden: 1) Erstelle den Diamant-Graphen mit a_2^* und b_2^* als Terminalknoten. 2) Serienkomposition mit den Pfaden $(b_3^*, \dots, a_2^*)$ und $(b_2^*, \dots, a_1^*)$. 3) Parallelkomposition mit dem Pfad $(b_3^*, \dots, a_1^*)$. 4) Serienkomposition mit den restlichen Diamant-Graphen.

in Satz 5.1 als Blocking Coalition $S = \{a_2, a_3\}$ und $y = q_j$, mit $j = \left\lfloor \frac{\delta - i}{2} \right\rfloor$. Dann gilt

$$\begin{aligned}
\rho &= \min\left(\frac{d(a_2,x)}{d(a_2,y)}, \frac{d(a_2',x)}{d(a_2',y)}, \dots, \frac{d(a_3,x)}{d(a_3,y)}, \frac{d(a_3',x)}{d(a_3',y)}, \dots\right) \\
&\geq \min\left(\frac{\delta - i + 1}{\left\lfloor \frac{\delta-i}{2} \right\rfloor + 1 + \left\lfloor \log_{\Delta-1} \alpha \right\rfloor}, \frac{\delta + i + 1}{\delta - \left\lfloor \frac{\delta-i}{2} \right\rfloor + 1 + \left\lfloor \log_{\Delta-1} \alpha \right\rfloor}\right) \\
&\geq \min\left(\frac{\delta - i + 1}{\frac{\delta-i}{2} + 1 + \log_{\Delta-1} \alpha}, \frac{\delta + i + 1}{\delta - \frac{\delta-i}{2} + 2 + \log_{\Delta-1} \alpha}\right) \\
&= \min\left(\frac{2\delta - 2i + 2}{\delta - i + 2 + 2\log_{\Delta-1} \alpha}, \frac{2\delta + 2i + 2}{2\delta - (\delta - i) + 4 + 2\log_{\Delta-1} \alpha}\right) \\
&= \min\left(\frac{2 + \frac{2}{\delta-i}}{1 + \frac{2+2\log_{\Delta-1} \alpha}{\delta-i}}, \frac{2 + \frac{2}{\delta+i}}{1 + \frac{4+2\log_{\Delta-1} \alpha}{\delta+i}}\right).
\end{aligned}$$

Wählen wir δ beliebig groß, so gilt wegen Relation (5.3) und $\log n \in o(n)$, dass der Verbesserungsfaktor ρ für $\delta \to \infty$ gegen 2 konvergiert.[3]

Fall 2. Sei $x = a_1$, für einen Knoten a_1 im Baum. Dann gilt die gleiche Abschätzung, wie wenn $x = a_1^*$ gelten würde, da für die Blocking Coalition $S = \{a_2, a_3\}$ und $y = q_j$, für $j = \left\lfloor \frac{\delta}{2} \right\rfloor$ gilt: $d(a_i, a_1^*) < d(a_i, x)$, für $i \in \{2, 3\}$, also

$$\rho = \min\left(\frac{d(a_2,a_1)}{d(a_2,y)}, \dots, \frac{d(a_3,a_1)}{d(a_3,y)}, \dots\right) \geq \min\left(\frac{d(a_2,a_1^*)}{d(a_2,y)}, \dots, \frac{d(a_3,a_1^*)}{d(a_3,y)}, \dots\right) \overset{\lim}{\rightsquigarrow} 2.$$

Also gibt es für alle $\rho < 2$ nicht immer ein ρ-Proportional Clustering für Graphen mit beschränktem Knotengrad $\Delta \geq 3$. □

[3] Die Abschätzung nach oben ist auch 2, weil für das a_i, mit $(a_i, a_i^*) \in E$ die gleichen Rechnungen wie in Satz 5.1 gelten.

Kapitel 6

Greedy Capture

Für beliebige metrische Räume gibt Greedy Capture immer ein $(1+\sqrt{2})$-Proportional Clustering aus [Che+19]. Der Algorithmus lässt gleichmäßig Kreise um die möglichen Center wachsen und wählt gierig einen Center aus, sobald dieser $\lceil \frac{n}{k} \rceil$ (neue) Punkte enthält. Bereits geöffnete Center wachsen weiter und nehmen (neue) Punkte mit auf. Nun stellt sich die Frage, ob der Algorithmus auf Graphen eine bessere Approximation liefert. In diesem Abschnitt zeigen wir, dass dies nicht der Fall ist. Die Idee des Beweises ist das Beispiel 18 von Micha und Shah [MS20] in einen Graphen umzuwandeln. Dabei werden aus dem Punkt x_5 vom Beispiel zwei Sterne, da Greedy Capture sonst einen Cluster Center auf dem Pfad von x_4 nach x_5 platzieren würde.

Satz 6.1. *Für alle $\rho < 1+\sqrt{2}$ findet Greedy Capture auf Graphen nicht immer ein ρ-Proportional Clustering.*

Beweis. Sei $n = 35\alpha + 5\delta$ und $k = 9$. Es gibt fünf identische Gruppen (Zusammenhangskomponenten) mit je $7\alpha + \delta$ Knoten. Eine Gruppe sieht wie folgt aus. Es gibt zunächst sechs spezielle Typen von Knoten, $a_1^*, a_2^*, a_3^*, a_4^*, a_5^*, a_6^*$. Der Knoten a_1^* zusammen mit 2α weiteren Knoten vom Typ a_1 bilden einen Stern mit a_1^* als Zentrum. Die Knoten a_i^* zusammen mit α weiteren Knoten vom Typ a_i bilden einen Stern mit a_i^* als Zentrum, für $i \in \{2, \dots, 6\}$. Die restlichen δ Knoten werden benutzt, um je a_i^* mit a_{i+1}^* durch einen Pfad der Länge δ_i zu verbinden, für $i \in \{1,2,3,4\}$ und a_4^* mit a_6^* durch einen Pfad der Länge δ_5 zu verbinden. Also ist $\delta = 1 + \sum_{i=1}^{5} \delta_i$. Abbildung 6.1 zeigt beispielhaft eine der Gruppen.

Die Distanzen $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5$ sollen approximiert folgendes Verhältnis haben: $1/1/(\sqrt{2}-1)/1/1$. Außerdem soll $\delta_1 = \delta_2$, $\delta_4 = \delta_5$ und $\delta_4 > \delta_1$ gelten. Ein mögliches Vorgehen, um die δ_i zu bestimmen, ist:

1. Approximiere $\sqrt{2}-1$ durch einen Bruch $\frac{s}{t}$ beliebig genau, mit $s, t \in \mathbb{N}$.
2. Setze $\delta_1 = t, \delta_2 = t, \delta_3 = s, \delta_4 = t+1$ und $\delta_5 = t+1$.

Wegen den Pfadknoten soll es folgende Bedingung an α geben: $4\alpha \geq \lceil \frac{n}{k} \rceil = \lceil \frac{35\alpha+5\delta}{9} \rceil$, also:

$$4\alpha \overset{!}{\geq} \frac{35\alpha + 5\delta}{9} + 1 \Leftrightarrow \alpha \geq 5\delta + 9 \tag{6.1}$$

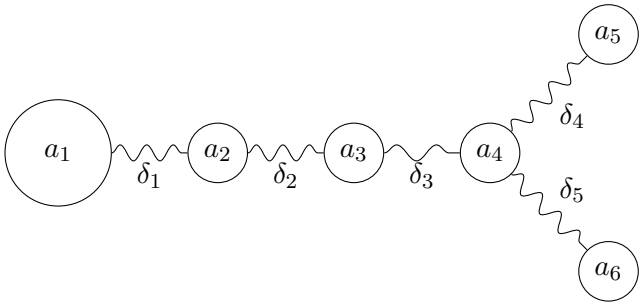

Abbildung 6.1: Graph vom Satz 6.1. Die Kreise stellen die Sterne dar. Der Kreis von a_1 ist größer, da es einen größeren Stern darstellt. Wellenförmige Kanten stellen Pfade dar, dessen Längen deren Beschriftung ist. Pfade verbinden die Sternzentren a_i^* der jeweiligen Sterne.

Greedy Capture öffnet zuerst fünf Cluster Center bei den Knoten a_2^*. Das geschieht beim Kreisradius $\delta = \delta_1 + 1$ im Algorithmus. Dann gibt es in maximal vier Gruppen einen weiteren Center. Da wegen Ungleichung (6.1) $4\alpha \geq \lceil \frac{n}{k} \rceil$ gilt, ist jede Menge mit 4α Knoten berechtigt, eine Blocking Coalition zu bilden. In der Gruppe mit nur einem Center gilt nun mit der Blocking Coalition $S = \{a_3, a_4, a_5, a_6\}$ und $y = a_4^*$, für den Verbesserungsfaktor

$$\begin{aligned}
\rho &= \min\left(\frac{d(a_3,X)}{d(a_3,y)}, \frac{d(a_4,X)}{d(a_4,y)}, \frac{d(a_5,X)}{d(a_5,y)}, \frac{d(a_6,X)}{d(a_6,y)}\right) \\
&= \min\left(\frac{\delta_2+1}{\delta_3+1}, \frac{\delta_2+\delta_3+1}{1}, \frac{\delta_2+\delta_3+\delta_4+1}{\delta_4+1}, \frac{\delta_2+\delta_3+\delta_5+1}{\delta_5+1}\right) \\
&\approx \min\left(\frac{1}{\sqrt{2}-1}, \infty, \frac{1+\sqrt{2}}{1}, \frac{1+\sqrt{2}}{1}\right) = 1+\sqrt{2}.
\end{aligned}$$

Im Grenzwert $\delta \to \infty$ und mit besseren Approximationen der δ_i an den vorgegebenen Verhältnissen konvergiert der Verbesserungsfaktor ρ gegen $1+\sqrt{2}$. Also ist die Ausgabe von Greedy Capture nicht ρ-proportional für $\rho < 1+\sqrt{2}$. □

Da der Lowerbound mit dem Upperbound übereinstimmt, erhalten wir also folgendes Korollar.

Korollar 6.2. *Der Upperbound von $1+\sqrt{2}$ von Greedy Capture bei beliebigen metrischen Räumen ist bei Graphen nicht kleiner.*

Der Graph in Satz 6.1 zeigt, dass Greedy Capture selbst bei Bäumen keinen besseren Upperbound als $1+\sqrt{2}$ liefert. Der Graph lässt sich in einen Baum umwandeln, indem z. B. die Knoten a_1^* durch einen langen Pfad mit einem neuen Knoten a_0^* verbunden werden und dann das α entsprechend erhöht wird.

Kapitel 7

Komplexität

In diesem Abschnitt beweisen wir, dass es NP-vollständig ist zu entscheiden, ob es für einen gegebenen Graphen G und eine Zahl k ein ρ-Proportional Clustering, für ein festes $\rho < 2$ gibt, indem wir von dem NP-schweren Problem Hitting Set reduzieren.[1] Die Idee der Reduktion ist es, pro Hitting-Menge das Dreieck aus Satz 5.1 und pro Element aus dem Universum einen Knoten zu erstellen. Dann verbinden wir einen Knoten mit einem Dreieck, wenn das Element in der Hitting-Menge vorkommt. Wenn es ein Hitting Set gibt, dann können wir die entsprechenden Element-Knoten und ein Knoten pro Dreieck als Clustering wählen, sodass jedes Dreieck zwei Cluster Center in der Nähe hat. Wenn es kein Hitting Set gibt, dann wird es in einem der Dreiecke, mit nur maximal einem Cluster Center in der Nähe, eine Blocking Coalition geben.

Problem 1: Hitting Set (HS)
Eingabe: Eine Grundmenge $\mathcal{U} = \{s_1, \ldots, s_n\}$, eine Teilmengenfamilie $\mathcal{S} = \{S_1, \ldots, S_m\}$, mit $S_j \subseteq \mathcal{U}$, für $j \in \{1, \ldots, m\}$ und eine Zahl $k \in \mathbb{N}$.
Frage: Existiert eine Menge $X \subseteq \mathcal{U}$, mit $|X| \leq k$ und $S_j \cap X \neq \emptyset$, für alle $j \in \{1, \ldots, m\}$?

Problem 2: Proportional Clustering on Graphs (PCG)
Eingabe: Ein Graph $G = (V, E)$ und eine Zahl $k \in \mathbb{N}$.
Frage: Existiert eine Menge $X \subseteq V$ von Cluster Centern, mit $|X| = k$, sodass X ein ρ-Proportional Clustering ist?

Satz 7.1. *Das Problem* Proportional Clustering on Graphs *ist* NP*-schwer.*

Beweis. Um zu zeigen, dass Proportional Clustering on Graphs NP-schwer ist, reduzieren wir von Hitting Set auf PCG, zeigen also HS $\leq_m^p$ PCG.

Sei $\rho < 2$ beliebig aber fest. Die Reduktion arbeitet wie folgt. Sei $\mathcal{U} = \{s_1, \ldots, s_n\}$, $\mathcal{S} = \{S_1, \ldots, S_m\}$, mit $S_j \subseteq \mathcal{U}$ und $k \in \mathbb{N}$ eine Eingabe für HS.

[1] In Korollar 7.3 zeigen wir, wie in Polynomzeit überprüft werden kann, ob ein gegebenes Clustering ρ-Proportional ist. Dann folgt, dass das Problem in NP ist, mit dem Guess & Check Ansatz.

- Sei $\delta > 14$ beliebig groß, sodass das Dreieck in Satz 5.1 kein ρ-Proportional Clustering hat, mit geradem δ.[2] Setze $\tau = \sum_{i=1}^{m} |S_i| \cdot \left(\left\lfloor \left(\rho - \frac{1}{2}\right) \cdot \delta \right\rfloor - 1\right)$ und

$$\alpha = \left\lceil \max\left(\frac{n + \tau + 6\delta m + 3\delta + k + 3m + 2}{2k + 1}, \right.\right. \\ \left.\left. \frac{(n + \tau + 3\delta m) \cdot (k + 3m) + n + \tau - 3\delta}{2m + 1} + 1 \right) \right\rceil .$$

- Für jedes Element $s_i \in \mathcal{U}$ erstelle einen Knoten s_i.
- Für jede Hitting-Menge $S_j \in \mathcal{S}$ erstelle ein Dreieck S_j mit Seitenlänge δ und Sternen an den Eckpunkten mit je α Armen (analog zu Satz 5.1). Knoten im Dreieck S_j werden mit einem Superskript j versehen und haben sonst die gleiche Bezeichnung wie in Satz 5.1.
- Für jede Hitting-Menge $S_j \in \mathcal{S}$: Verbinde s_i mit a_3^{*j} durch einen Pfad der Länge $\left\lfloor \left(\rho - \frac{1}{2}\right) \cdot \delta \right\rfloor$ gdw. $s_i \in S_j$. Knoten aus diesem Pfad sind vom Typ t_i^j.
- Erstelle $m+1$ isolierte Dreiecke mit Seitenlänge δ und Sternen an den Eckpunkten mit je α Armen (analog zu Satz 5.1).
- Setze $k' = k + 3m + 2$.

Totalität und Polynomzeit. Die Reduktion ist total, da zu jeder Hitting Set-Eingabe ein Graph $G = (V, E)$ und eine Zahl k' erzeugt wird. Für die Größe der Ausgabe gilt: Das δ ist konstant und τ und α sind nur polynomiell groß. Es werden $n' = n + \tau + (2m+1) \cdot (3\alpha + 3\delta)$ Knoten erzeugt, also gibt es auch nur polynomiell viele Kanten und $k' = k + 3m + 2$ ist polynomiell groß. Die Erstellung dieser Knoten und Kanten erfolgt in Polynomzeit.

Korrektheit. Sei die Eingabe für Hitting Set eine Ja-Instanz. Sei also X, mit $|X| \leq k$ ein Hitting Set. Dann konstruiere ein ρ-Proportional Clustering X' wie folgt:

- Für jedes der $m+1$ isolierten Dreiecke füge a_1^* und a_2^* zu X' hinzu.
- Für jedes Dreieck S_j, mit $j \in \{1, \ldots, m\}$ füge $p_{\frac{\delta}{2}}^j$ zu X' hinzu.
- Für jedes $x \in X$ füge x zu X' hinzu.

Es wurden maximal $2 \cdot (m+1) + m + k = k'$ Cluster Center benutzt.

Nun zeigen wir, dass X' ein ρ-Proportional Clustering ist. Die Wahl von α garantiert, dass $2\alpha \geq \left\lceil \frac{n'}{k'} \right\rceil$ und $\left\lceil \frac{n'}{k'} \right\rceil > \alpha + n + \tau + 3\delta m$, da:

$$\begin{aligned} \alpha &\geq \frac{n + \tau + 6\delta m + 3\delta + k + 3m + 2}{2k + 1} \\ \Leftrightarrow 2\alpha k + 6\alpha m + 4\alpha &\geq n + \tau + 6\alpha m + 6\delta m + 3\alpha + 3\delta + k + 3m + 2 \\ \Leftrightarrow 2\alpha &\geq \frac{n + \tau + (2m+1) \cdot (3\alpha + 3\delta)}{k + 3m + 2} + 1 \\ \Rightarrow 2\alpha &\geq \left\lceil \frac{n'}{k'} \right\rceil \end{aligned} \tag{7.1}$$

[2] Diese Bedingung fordern wir nur, um uns unnötige Rundungsklammern zu sparen.

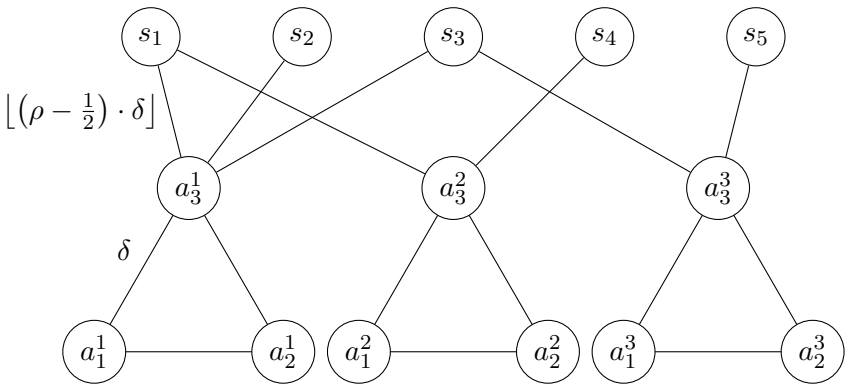

Abbildung 7.1: Ausgabe der Reduktion (ohne die isolierten Dreiecke). Die Pfade zwischen Knoten s und a_3^* haben Länge $\lfloor(\rho - \frac{1}{2}) \cdot \delta\rfloor$ und die Pfade in den Dreiecken Länge δ. Die Pfadlänge $\lfloor(\rho - \frac{1}{2}) \cdot \delta\rfloor$ ist so gewählt, sodass bei einer Ja-Instanz für Hitting Set der Ausgabegraph gerade noch ein ρ-Proportional Clustering hat. Werden diese Pfade zu kurz gewählt, kann es passieren, dass Knoten a_3^j ihre Distanz aus einem Center eines anderen Dreiecks $S_{j'}$ beziehen. Die Eckpunkte a der Dreiecke stellen wie in Satz 5.1 Sterne dar.

und

$$\begin{aligned} \alpha &> \frac{(n+\tau+3\delta m)\cdot(k+3m)+n+\tau-3\delta}{2m+1} \\ \Rightarrow 3\alpha m+\alpha &> \alpha k+(n+\tau+3\delta m)\cdot(k+3m)+n+\tau-3\delta \\ \Leftrightarrow \frac{n+\tau+(2m+1)\cdot(3\alpha+3\delta)}{k+3m+2} &> \alpha+n+\tau+3\delta m \\ \Rightarrow \left\lceil\frac{n'}{k'}\right\rceil &> \alpha+n+\tau+3\delta m, \end{aligned} \tag{7.2}$$

wobei wir im ersten Schritt $\alpha m > \alpha k$ verwendet haben.[3] In den isolierten Dreiecken kann sich keine Blocking Coalition bilden, da $d(a_1, X') = 1$ und $d(a_2, X') = 1$ bereits gilt und die $\alpha + 3\delta$ restlichen Knoten wegen Ungleichung (7.2) zu wenige sind. Wir betrachten also in der restlichen Instanz mögliche Blocking Coalitions. Wegen Ungleichung (7.1) betrachten wir zwei Sterne. Dabei reicht es nicht aus, nur α Knoten von einem Stern zusammen mit den restlichen $n + \tau + 3m\delta$ Knoten zu betrachten, da diese wegen Ungleichung (7.2) zu wenige Knoten sind. Es wird eine Fallunterscheidung gemacht, wo die Sternknoten liegen können:

Fall 1. Die Knoten sind aus verschiedenen Dreiecken. Sei also die Blocking Coalition $S = \{a_i^j, a_{i'}^{j'}\}$. Für jedes Dreieck S_j ist ein $s_i \in S_j$ im Clustering, also gilt $d(a_i^j, X') \leq \lfloor(\rho - \frac{1}{2}) \cdot \delta\rfloor + 1$ und $d(a_{i'}^{j'}, X') \leq \lfloor(\rho - \frac{1}{2}) \cdot \delta\rfloor + 1$. Also müssen $\rho \cdot d(a_i^j, y) < \lfloor(\rho - \frac{1}{2}) \cdot \delta\rfloor + 1$ und $\rho \cdot d(a_{i'}^{j'}, y) < \lfloor(\rho - \frac{1}{2}) \cdot \delta\rfloor + 1$ gelten. Dann gilt aber

$$d(a_i^j, a_{i'}^{j'}) \leq \rho \cdot d(a_i^j, a_{i'}^{j'}) \leq \rho \cdot d(a_i^j, y) + \rho \cdot d(y, a_{i'}^{j'}) < 2 \cdot \left\lfloor\left(\rho - \frac{1}{2}\right) \cdot \delta\right\rfloor + 2.$$

[3] Wenn $m \leq k$, dann können wir einfach eine triviale Ja-Instanz ausgeben.

Da die Knoten aber aus verschiedenen Dreiecken sind, gilt $d(a_i^j, a_{i'}^{j'}) \geq 2 \cdot \lfloor (\rho - \frac{1}{2}) \cdot \delta \rfloor + 2$; ein Widerspruch.

Fall 2. Die Knoten sind aus demselben Dreieck. Sei also die Blocking Coalition $S = \{a_i^j, a_{i'}^j\}$. **Fall (a).** Falls $i = 1, i' = 2$, dann gilt $d(a_1^j, X') = d(a_2^j, X') = \frac{\delta}{2} + 1$, also müssen $\rho \cdot d(a_1^j, y) < \frac{\delta}{2} + 1$ und $\rho \cdot d(a_2^j, y) < \frac{\delta}{2} + 1$ gelten. Dann gilt aber

$$d(a_1^j, a_2^j) \leq \rho \cdot d(a_1^j, a_2^j) \leq \rho \cdot d(a_1^j, y) + \rho \cdot d(y, a_2^j) < \delta + 2.$$

Aber es gilt $d(a_1^j, a_2^j) = \delta + 2$; ein Widerspruch. **Fall (b).** Falls $i = 1, i' = 3$, dann gilt $d(a_1^j, X') = \frac{\delta}{2} + 1$ und $d(a_3^j, X') = \lfloor (\rho - \frac{1}{2}) \cdot \delta \rfloor + 1$, also müssen $\rho \cdot d(a_1^j, y) < \frac{\delta}{2} + 1$ und $\rho \cdot d(a_3^j, y) < \lfloor (\rho - \frac{1}{2}) \cdot \delta \rfloor + 1$ gelten. Dann gilt aber

$$\begin{aligned} \rho(\delta + 2) = \rho \cdot d(a_1^j, a_3^j) &\leq \rho \cdot d(a_1^j, y) + \rho \cdot d(y, a_3^j) \\ &< \frac{\delta}{2} + 1 + \left\lfloor \left(\rho - \frac{1}{2}\right) \cdot \delta \right\rfloor + 1 \\ &\leq \frac{\delta}{2} + 1 + \left(\rho - \frac{1}{2}\right) \cdot \delta + 1 \\ &= \rho\delta + 2, \end{aligned}$$

also $\rho < 1$; ein Widerspruch. Die anderen Fälle sind symmetrisch. Also ist das Clustering X' ein ρ-Proportional Clustering.

Sei die Eingabe für Hitting Set eine Nein-Instanz. Dann gibt es kein Hitting-Set X, mit $|X| \leq k$. Sei X', mit $|X'| \leq k'$ ein beliebiges Clustering. Falls in den isolierten Dreiecken nicht je zwei Cluster Center benutzt werden, gibt es in mindestens einem dieser Dreiecke analog zu Satz 5.1 eine Blocking Coalition und wir sind fertig. Seien also in den $m + 1$ isolierten Dreiecken je zwei Cluster Center benutzt worden. Dann bleiben für X' noch $m + k$ Cluster Center übrig. Falls nicht in jedem der m Dreiecke S_j ein Cluster Center ist, dann gibt es in dem Dreieck S_j ohne einen Cluster Center die Blocking Coalition $S = \{a_1^j, a_2^j\}$, mit $y = p_{\frac{\delta}{2}}^j$, da $d(a_1^j, X') \geq \delta + 2$ und $d(a_2^j, X') \geq \delta + 2$ und $d(a_1^j, y) = d(a_2^j, y) = \frac{\delta}{2} + 1$, mit Verbesserungsfaktor 2 und wir sind fertig. Seien also in jedem der m Dreiecke S_j je ein Cluster Center benutzt worden. Dann bleiben für X' noch k Cluster Center übrig.

Wir argumentieren nun, dass es ausreicht, nur die Knoten s_i zu betrachten oder dass sonst X' kein ρ-Proportional Clustering ist.

Lemma 7.2. *Sei X' ein ρ-Proportional Clustering, mit $x^j, t_i^j \in X'$, für den einen Center x^j im Dreieck S_j (bzw. $x^j, x'^j \in X'$ für zwei Center im Dreieck S_j). Dann ist $X' \setminus \{x^j, t_i^j\} \cup \{p_{\frac{\delta}{2}}^j, s_i\}$ (bzw. $X' \setminus \{x^j, x'^j\} \cup \{p_{\frac{\delta}{2}}^j, s\}$, für ein beliebiges $s \in S_j$) auch ein ρ-Proportional Clustering.*

Beweis. Für die Sternknoten a^j im Dreieck S_j gilt wie beim Beweis bei der Ja-Instanz, dass die Knoten a^j keine Blocking Coalition bilden können. Für die Sternknoten aus den Dreiecken $S_{j'}$, mit $s_i \in S_{j'}$, gilt, dass sich die Distanz zu den Cluster Centern nur verbessert hat. Letztlich gilt für Dreiecke $S_{j''}$, mit $s_i \notin S_{j''}$, dass die Sternknoten $a^{j''}$ keine Distanz aus t_i^j bezogen haben, mit Distanz $d(a^{j''}, t_i^j) \geq 2 \lfloor (\rho - \frac{1}{2}) \cdot \delta \rfloor + 1$,

also sich die Distanz zum nächstgelegenen Cluster Center nicht verändert hat. Denn wenn dies nicht der Fall wäre, dann würde im Dreieck $S_{j''}$ gelten, dass kein $s \in S_{j''}$ ein Cluster Center ist, da die Distanz zu diesen Centern kleiner ist: Distanz $d(a_3^{j''}, s) = \lfloor(\rho - \frac{1}{2}) \cdot \delta\rfloor + 1$. Dann gäbe es in diesem Dreieck $S_{j''}$ aber eine Blocking Coalition, was wir nun noch im Beweis vom Satz 7.1 beweisen. □

Da die Hitting Set-Instanz eine Nein-Instanz ist, muss es ein Dreieck S_j geben, für das in keinem der s_i, mit $s_i \in S_j$ ein Cluster Center eröffnet wurde (und in dem Dreieck S_j nur ein Cluster Center eröffnet wurde). Ansonsten wäre $X = X' \cap \mathcal{U}$ ein Hitting Set nach Verschiebung der Cluster Center. Wir betrachten nun dieses Dreieck S_j. Sei x^j der eine Center im Dreieck S_j. Nun wird eine Fallunterscheidung über x^j gemacht:

Fall 1. Sei $x^j \in \{r_1^j, \dots, r_{\frac{\delta}{2}}^j\} \cup \{a_3^{*j}, a_3^j\}$. Dann ist analog zu Satz 5.1, $S = \{a_1^j, a_2^j\}$ mit dem entsprechenden y eine Blocking Coalition.

Fall 2. Sei $x^j \in \{r_{\frac{\delta}{2}+1}^j, \dots, r_{\delta-1}^j\} \cup \{a_1^{*j}, a_1^j\}$. Dann ist analog zu Satz 5.1, $S = \{a_2^j, a_3^j\}$ mit dem entsprechenden y eine Blocking Coalition.

Fall 3. Sei $x^j \in \{p_1^j, \dots, p_{\frac{\delta}{2}}^j\}$. Dann ist $S = \{a_2^j, a_3^j\}$, mit $y = q_{\lfloor \frac{\delta}{2\rho} \rfloor - 1}$ eine Blocking Coalition, da, wenn a_3^j seine Distanz aus x^j bezieht, die gleichen Rechnungen wie in Satz 5.1 gelten oder wenn a_3^j seine Distanz aus einem Center aus einem anderen Dreieck bzw. einem $s \notin S_j$ oder $t_i^{j'}$, mit $s_i \notin S_j$ und $j' \neq j$ bezieht, dann

$$\frac{d(a_2^j, X')}{d(a_2^j, y)} \geq \frac{\frac{\delta}{2} + 1}{\left\lfloor \frac{\delta}{2\rho} \right\rfloor - 1 + 1} > \frac{\frac{\delta}{2}}{\frac{\delta}{2\rho}} = \rho$$

und

$$\frac{d(a_3^j, X')}{d(a_3^j, y)} \geq \frac{2\left\lfloor(\rho - \frac{1}{2}) \cdot \delta\right\rfloor + 1}{\delta - \left(\left\lfloor \frac{\delta}{2\rho} \right\rfloor - 1\right) + 1} \geq \frac{(2\rho - 1)\delta - 1}{\left(1 - \frac{1}{2\rho}\right)\delta + 3} \overset{!}{>} \rho,$$

also

$$\begin{aligned} (2\rho - 1)\delta - 1 &> \left(\rho - \frac{1}{2}\right)\delta + 3\rho \\ \Leftrightarrow \left(2\rho - 1 - \rho + \frac{1}{2}\right)\delta &> 3\rho + 1 \\ \Leftrightarrow \delta &> \frac{3\rho + 1}{\rho - \frac{1}{2}}. \end{aligned}$$

Da $\rho \in [1; 2)$, ist die rechte Seite der Ungleichung höchstens $\frac{3 \cdot 2 + 1}{1 - 0{,}5} = 14$. Da δ größer 14 gewählt wurde, ist S eine Blocking Coalition.

Die restlichen Fälle sind symmetrisch. Also kann es kein ρ-Proportional Clustering X' für G geben. □

Das Ergebnis zeigt auch, dass es schwierig ist, Proportionalität zu approximieren. Genauer gilt

Korollar 7.3. *Es gibt keinen Algorithmus, der zu einer Eingabe G, k in Polynomzeit ein Clustering X ausgibt, das $(\alpha\rho_{\min})$-Proportional, für das optimale $\rho_{\min}$ und $\alpha < 1,5$ ist, unter der Annahme $P \neq NP$.*

Beweis. Angenommen, es gibt einen solchen Algorithmus A, für ein $\alpha < 1,5$. Sei X das ausgegebene Clustering vom Algorithmus für den Graphen aus Satz 7.1, für $\rho = 1$. Wir überprüfen in Polynomzeit, ob X α-Proportional ist, indem wir zuerst die Distanzmatrix und $d(i, X))$ berechnen und dann für jedes y überprüfen, ob $\{i : \alpha \cdot d(i, y) < d(i, X)\} \geq \lceil \frac{n}{k} \rceil$.

Wenn X α-Proportional ist, dann erhalten wir nach dem Verschieben der Cluster Center wie in Lemma 7.2 ein Hitting Set. Wenn wir die Center nicht verschieben könnten, dann wäre X nicht α-Proportional, mit analoger Begründung wie in Satz 7.1, außer im Fall 3, wo y dann so gewählt wird, dass es die minimale multiplikative Verbesserung maximiert: $\max\left(\min\left(\frac{1}{1-z}, \frac{0,5}{z}\right)\right) = 1,5$.

Wenn X nicht α-Proportional ist, dann gab es kein Exact Proportional Clustering und die Eingabe für Hitting Set ist eine Nein-Instanz. □

Kapitel 8

k-Center mit Proportionalität

In diesem Abschnitt zeigen wir, dass es keinen Algorithmus gibt, der für zusammenhängende Graphen gleichzeitig das k-Center Objective und die Proportionalität konstant approximiert. Für beliebige Graphen lässt sich einfach Beispiel 1 von Chen et al. [Che+19] in einen nicht zusammenhängenden Graphen umwandeln, indem Punkte durch Sterne ersetzt werden und die Distanz von a und b durch einen Pfad dargestellt wird.

Satz 8.1. *Es gibt keinen Algorithmus, der für zusammenhängende Graphen ein Clustering ausgibt, das gleichzeitig eine σ-Approximation für k-Center und ρ-Proportional ist, für feste $\sigma, \rho \in \mathbb{R}$.*

Beweis. Angenommen, es gibt einen solchen Algorithmus für feste $\sigma, \rho \in \mathbb{R}$. Sei $\delta \in \mathbb{N}$ (mit δ als Zweierpotenz, um uns Rundungsklammern zu sparen) so gewählt, sodass:

1. $\frac{\frac{1}{2}\delta}{1+\log \delta} > \sigma$,
2. $\log \delta \geq \rho$ und
3. $\delta \geq 5$.

Sei $n = \delta\alpha + \delta \log \delta + \delta + 1$ und $k = \delta + 1$. Es gibt δ Sterne $S_{\alpha+1}$, deren Zentren a^* durch einen $\log \delta$ langen Pfad mit einem Knoten q_0 verbunden sind. Die Knoten von a_i^* nach q_0 heißen p_i^j, für $i \in \{1, \ldots, \delta\}$ und $j \in \{1, \ldots, \log \delta\}$. Von q_0 aus gibt es einen δ langen Pfad zu einem Knoten q_δ, mit Knoten q_i, für $i \in \{1, \ldots, \delta\}$. Abbildung 8.1 zeigt diese Konstruktion. Damit $\alpha \geq \lceil \frac{n}{k} \rceil$ gilt, muss für α gelten:

$$\begin{aligned} \alpha &\overset{!}{\geq} \frac{\delta\alpha + \delta \log \delta + \delta + 1}{\delta + 1} + 1 \\ \Leftrightarrow \delta\alpha + \alpha &\geq \delta\alpha + \delta \log \delta + 2\delta + 2 \\ \Leftrightarrow \alpha &\geq \delta \log \delta + 2\delta + 2. \end{aligned}$$

Falls der Algorithmus nun für einen der Sterne a_i bzw. a_i^* und den zugehörigen Pfadknoten p_i^j, für $j \in \{1, \ldots, \log \delta - 1\}$ keinen Cluster Center gesetzt hatte, dann wäre $S = \{a_i\}$, mit $y = a_i^*$ eine Blocking Coalition mit Verbesserungsfaktor $\frac{1+\log \delta}{1} > \rho$. Also müssen mindestens δ Cluster Center bei den Pfadknoten p oder den Sternen

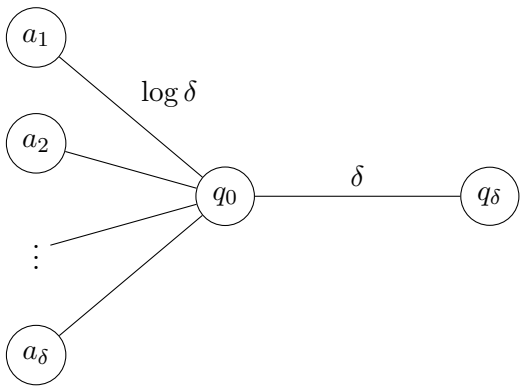

Abbildung 8.1: Zusammenhängender Graph, indem es kein für k-Center gleichzeitig σ-approximatives und ρ-Proportional Clustering gibt. Die Pfadlängen sind so gewählt, sodass der Pfad $(q_0, \ldots, q_\delta)$ asymptotisch schneller wächst als die anderen (hier mit $\log \delta \in o(\delta)$).

sein. Nun ist für den (q_0, q_δ)-Pfad aber nur noch ein Cluster Center übrig. Damit ist der k-Center Wert mindestens $\frac{1}{2}\delta$. Ein optimales k-Center Clustering ist aber $X = \{q_i : i \in \{0, \ldots, \delta - 1\}\}$ mit einem k-Center Wert von $1 + \log \delta$, da wegen Bedingung 3, $\log \delta < \frac{1}{2}\delta$ gilt. Wegen Bedingung 1 ist das Clustering X aber dann keine σ-Approximation für k-Center mehr. □

Das gleiche Ergebnis lässt sich für die Objektiven k-Means und k-Median aufstellen, indem statt des Knotens q_δ mehrere kleine Sterne erstellt werden.

Kapitel 9

Fazit

In dieser Bachelorarbeit haben wir Proportional Fairness auf Graphen untersucht. Wir haben gezeigt, dass für die Baumweite 1 und Baumtiefe 3 Exact Proportional Clusterings noch existieren und dass dies für die nächstgrößeren Baumweiten und Baumtiefen nicht mehr der Fall ist. Wir haben gezeigt, dass in Graphen nicht immer ein $(2 - \epsilon)$-Proportional Clustering existiert. Ob der Lowerbound von 2 für serien-parallele Graphen und allgemein Graphen mit Baumweite 2 erreicht werden kann, bleibt eine offene Frage. Allgemeiner stellt sich die Frage, für welche Graphklassen, die das Dreieck-Gegenbeispiel nicht enthalten, es immer ein Exact Proportional Clustering gibt und für welche Graphklassen, die das Dreieck-Gegenbeispiel enthalten, der Lowerbound von 2 erreicht werden kann.

Wir haben gezeigt, dass der Greedy Capture Algorithmus auf Graphen und bereits auf Bäumen keinen besseren Upperbound als $1+\sqrt{2}$ erreicht. Eine der wichtigsten offenen Fragen ist, ob der Upperbound von Greedy Capture in Graphen verbessert werden kann.

Für Clusterings im Spezialfall $\mathcal{N} = \mathcal{M}$ außerhalb von Graphen kann möglicherweise die Technik, in der Datenpunkte auf jedem Punkt in $\mathcal{M}$ und entsprechend viele Datenpunkte bei den originalen Punkten $\mathcal{N}$ platziert werden, genutzt werden, um Ergebnisse aus dem Fall $\mathcal{N} \subseteq \mathcal{M}$ zu übertragen.

Literatur

[BM+76] John Adrian Bondy, Uppaluri Siva Ramachandra Murty u. a. *Graph theory with applications*. Bd. 290. Macmillan London, 1976 (siehe S. 9, 13).

[Che+19] Xingyu Chen, Brandon Fain, Liang Lyu und Kamesh Munagala. *Proportionally fair clustering*. In: *International Conference on Machine Learning*. PMLR. 2019, S. 1032–1041 (siehe S. 5, 9–11, 13, 23, 33, 41).

[Che+20] Yu Cheng, Zhihao Jiang, Kamesh Munagala und Kangning Wang. *Group fairness in committee selection*. In: *ACM Transactions on Economics and Computation (TEAC)* 8.4 (2020), S. 1–18 (siehe S. 10).

[CMS24] Ioannis Caragiannis, Evi Micha und Nisarg Shah. *Proportional Fairness in Non-Centroid Clustering*. In: (2024) (siehe S. 10).

[FMS18] Brandon Fain, Kamesh Munagala und Nisarg Shah. *Fair allocation of indivisible public goods*. In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. 2018, S. 575–592 (siehe S. 10).

[JKL20] Christopher Jung, Sampath Kannan und Neil Lutz. *Service in your neighborhood: Fairness in center location*. In: *Foundations of Responsible Computing (FORC)* (2020) (siehe S. 10).

[JMW20] Zhihao Jiang, Kamesh Munagala und Kangning Wang. *Approximately stable committee selection*. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, S. 463–472 (siehe S. 10).

[KP23] Leon Kellerhals und Jannik Peters. *Proportional fairness in clustering: A social choice perspective*. In: *arXiv preprint arXiv:2310.18162* (2023) (siehe S. 10).

[Li+21] Bo Li, Lijun Li, Ankang Sun, Chenhao Wang und Yingfan Wang. *Approximate group fairness for clustering*. In: *International conference on machine learning*. PMLR. 2021, S. 6381–6391 (siehe S. 10).

[Meh+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman und Aram Galstyan. *A survey on bias and fairness in machine learning*. In: *ACM computing surveys (CSUR)* 54.6 (2021), S. 1–35 (siehe S. 9).

[MS20] Evi Micha und Nisarg Shah. *Proportionally fair clustering revisited*. In: *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2020 (siehe S. 5, 9–11, 15, 16, 19–21, 33).

[NDM06] Jaroslav Nešetřil und Patrice Ossona De Mendez. *Tree-depth, subgraph coloring and homomorphism bounds*. In: *European Journal of Combinatorics* 27.6 (2006), S. 1022–1041 (siehe S. 14, 27, 28).

[RS86] Neil Robertson und Paul D. Seymour. *Graph minors. II. Algorithmic aspects of tree-width*. In: *Journal of algorithms* 7.3 (1986), S. 309–322 (siehe S. 14).

[Sax+17] Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding und Chin-Teng Lin. *A review of clustering techniques and developments*. In: *Neurocomputing* 267 (2017), S. 664–681 (siehe S. 9).

[TNS82] Kazuhiko Takamizawa, Takao Nishizeki und Nobuji Saito. *Linear-time computability of combinatorial problems on series-parallel graphs*. In: *Journal of the ACM (JACM)* 29.3 (1982), S. 623–641 (siehe S. 14).