Seminar im WS 2006/07
Zerlegen und Clustern von Graphen

Correlation Clustering – Minimizing Disagreements on Arbitrary Weighted Graphs

Myriam Freidinger

18. April 2007

1 Einleitung

Ziel dieser Ausarbeitung ist die Aufbereitung einer Arbeit von Dotan Emanuel und Amos Fiat, „Correlation Clustering – Minimizing Disagreements on Arbitrary Weighted Graphs“ [EmFi03] aus dem Department of Computer Science, School of Mathematical Sciences, Tel Aviv University, Israel.

Beim Correlation Clustering werden Knoten mit dem Ziel in Gruppen eingeteilt, dass Knoten innerhalb einer Gruppe möglichst wenige Unterschieden aufweisen und Knoten verschiedener Gruppen möglichst wenige Übereinstimmungen haben. Der Unterschiede beziehungsweise die Übereinstimmung zweier Knoten kann über eine Gewichtsfunktion näher quantifiziert werden. Eine Anwendung für Correlation Clustering ist das automatische Sortieren von Dokumenten nach Thema, ohne vorher eine bestimmte Themenauswahl fest zu legen.

1.1 Problemstellung

Gegeben ist ein ungerichteter Graphen $G = (V, E)$ mit Kantenbeschriftung $l : E \to \{<+>, <->\}$. Für das Minimierungsproblem wird nach einer Einteilung der Knotenmenge V in Cluster $C_1, \ldots, C_n$ mit möglichst wenigen negativen Kanten innerhalb eines Clusters und möglichst wenigen positiven Kanten zwischen den Clustern gesucht.

Alternativ kann auch das Maximierungsproblem betrachtet werden, in dem die Anzahl der negativen Kanten zwischen Clustern sowie die Anzahl der positiven Kanten innerhalb der Cluster maximiert werden soll. In einer Erweiterung wird der Graph um eine Gewichtsfunktion $w : E \to \mathbb{R}^+$ ergänzt. In diesem

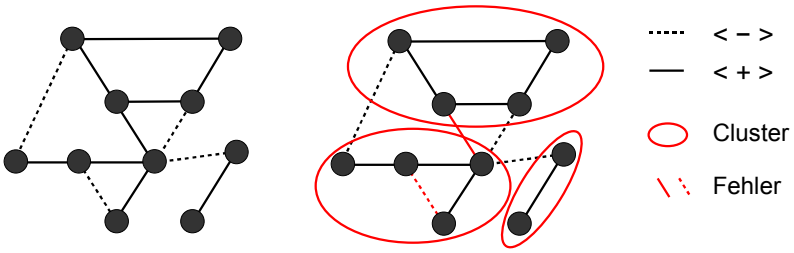

Abbildung 1: Graph mit konstanter Gewichtsfunktion

Fall ist nicht die Anzahl der Kanten sondern die Summe ihrer Gewichte ausschlaggebend. Umgekehrt kann man jede ungewichtete Instanz von Correlation Clustering durch einen konstante Gewichtsfunktion in eine gewichtete Instanz umwandeln.

Die Kanten, die mit $< + >$ beschriftet sind werde ich im Folgenden *positive Kanten*, die mit $< - >$ beschrifteten Kanten *negative Kanten* nennen. Positive Kanten zwischen Clustern sowie negative Kanten innerhalb eines Clusters werde ich als fehlerhafte Kanten oder allgemeiner als Fehler bezeichnen.

Schreibt man $C(u)$ für die Menge aller Knoten, die mit u in einem Cluster liegen, so kann die Zielfunktion für das Minimierungsproblem folgendermaßen beschrieben werden:

$$w(C) \quad = \sum_{\substack{(u,v)\in E,\, v\notin C(u) \\ l(u,v)=<+>}} w(u,v) \quad + \sum_{\substack{(u,v)\in E,\, v\in C(u) \\ l(u,v)=<->}} w(u,v)$$

Abbildung 1 zeigt einen Graphen mit konstanter Gewichtsfunktion sowie eine beispielhafte Einteilung der Knotenmenge in Cluster mit Gewicht $w(C) = 2$ für das Minimierungsproblem.

Das Gewicht einer Einteilung in Cluster für das Maximierungsproblem ist gegeben durch:

$$w(C) \quad = \sum_{\substack{(u,v)\in E,\, v\in C(u) \\ l(u,v)=<+>}} w(u,v) \quad + \sum_{\substack{(u,v)\in E,\, v\notin C(u) \\ l(u,v)=<->}} w(u,v)$$

Für das Maximierungsproblem ist das Gewicht der Einteilung in Cluster C aus Abbildung 1 $w(C) = 11$.

Die optimale Lösung einer Instanz (G, l, w) des Maximierungsproblems und des Minimierungsproblems unterscheiden sich nicht, die Gütegarantie eines festen Approximationsalgorithmus dagegen schon. Für das Maximierungsproblem kann leicht ein 2-approximativer Algorithmus angegeben werden. Dazu betrachtet man zwei Einteilungen der Knotenmenge in Cluster:

1. Alle Knoten liegen im selben Cluster.

2. Jeder Knoten ist in einem eigenen Cluster.

Ist das Gewicht aller positiven Kanten größer als das Gewicht aller negativen Kanten, so ist die erste Variante besser, andernfalls die zweite Variante. Die besser der beiden Einteilungen enthält in jedem Fall mindestens die Hälfte aller Kantengewichte und damit mindestens halb so viele Kantengewichte wie die optimale Lösung. Für das Minimierungsproblem kann zu diesem Algorithmus keine relative Gütegarantie angegeben werden. Die optimale Einteilung in Cluster könnte fehlerfrei sein, die des Algorithmus im schlechtesten Fall ein Gewicht von $w(C) = 1/2 \sum_{(u,v)\in E} w(u,v)$ haben, was zu einem Approximationsfaktor von unendlich führt. Es zeigt sich also, dass das Minimierungsproblem im Allgemeinen schwieriger zu approximieren ist als das Maximierungsproblem. Die Ergebnisse der hier vorgestellten Arbeit [EmFi03] beziehen sich ausschließlich auf das schwierigere Minimierungsproblem.

1.2 Ziele

Vor dem Erscheinen der Arbeit von Emanuel und Fiat gab es bereits einen polynomialen c-approximativer Algorithmus mit $c \in \mathbb{N}$ für Correlation Clustering in ungewichteten vollständigen Graphen. Des weiteren war bekannt, dass Correlation Clustering in gewichteten allgemeinen Graphen APX-schwer ist.

Die angestrebten neuen Resultate sind ein polynomialer $O(log(n))$-approximativer Algorithmus für Correlation Clustering in beliebigen gewichteten und ungewichteten Graphen sowie der Beweis, dass Correlation Clustering auch in ungewichteten Graphen APX-schwer ist. Das gewählte Hilfsmittel ist die polynomiale Transformation von Correlation Clustering auf Multicut und umgekehrt.

1.3 Multicut

Bei Multicut ist ein Graph $G = (V, E)$ mit Gewichtsfunktion $c : E \to \mathbb{R}^+$ sowie einer Menge von k Paaren $M = \{< s_1, t_1 >, \ldots, < s_k, t_k >\}$ mit $< s_i, t_i > \in V \times V$, $s_i \neq t_i$, $i \in \{1, \ldots, k\}$ gegeben.

Gesucht ist eine Menge von Kanten $S \subseteq E$ mit minimalem Gewicht, so dass nach Entfernen der Kanten von S aus G die Paare $< s_i, t_i >$ nicht mehr durch einen Pfad miteinander verbunden sind.

Zu Multicut ist ein $O(log(k))$-approximativer Algorithmus bekannt [GaVa96] sowie die Tatsache, dass das Problem schon in ungewichteten Graphen APX-schwer ist [DaJo94].

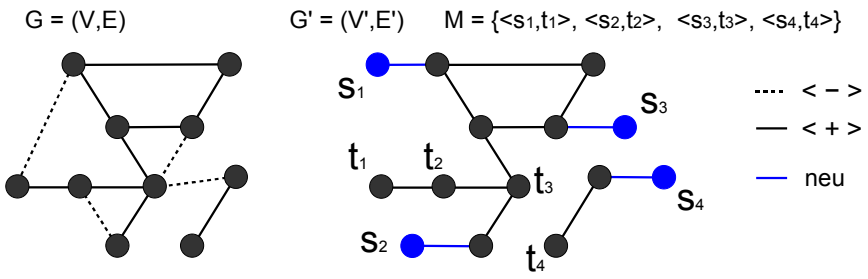

Abbildung 2: Umwandlung: Correlation Clustering in Multicut

2 Correlation Clustering nach Multicut

Um das erste Ziel zu erreichen, einen $O(log(n))$-approximativen Algorithmus für Correlation Clustering, geben Emanuel und Fiat eine polynomiale Transformation einer Instanz von Correlation Clustering $CC = (G, l, w)$ in einem beliebigen Graphen in eine Instanz von Multicut $MC = (G', w', M)$ an. Dabei wird jede ungewichtete Instanz von Correlation Clustering zu einer ungewichteten Instanz von Multicut und jede gewichtete Instanz zu einer gewichteten Instanz. Sie beweisen, das eine Lösung von MC mit Gewicht W einen Lösung von CC mit dem selben Gewicht induziert und umgekehrt.

2.1 Transformation

Um eine Instanz von Correlation Clustering $CC = (G, l, w)$ in eine Instanz $MC = (G', w', M)$ von Multicut umzuwandeln, wird jede negative Kante (u, v) aus dem Graphen G entfernt. Anstelle dessen werden ein neuer Knoten $x_{(u,v)}$ sowie eine neue Kante $(x_{(u,v)}, u)$ mit Gewicht $w(u, v)$ zu G hinzugefügt. Außerdem wird M um ein neues Paar $< x_{(u,v)}, v >$ erweitert.

Formal lässt sich die polynomiale Transformation wie folgt beschreiben:

CC: $G = (V, E) \qquad l : E \to \{<+>, <->\} \qquad w : E \to \mathbb{R}^+$

MC: $G' = (V', E') \quad$ mit

$$\begin{aligned} V' &= V \cup \{x_{(u,v)} \,|\, (u,v) \in E\,,\ l(u,v) = <->\} \\ E' &= \{(u,v) \in E \,|\, l(u,v) = <+>\} \\ &\quad \cup \{(x_{(u,v)}, u) \,|\, (u,v) \in E\,,\ l(u,v) = <->\} \end{aligned}$$

$w' : E \to \mathbb{R}^+ \quad$ mit

$$\begin{aligned} w'(u,v) &= w(u,v) \quad \forall\,(u,v) \in E \quad \text{mit} \quad l(u,v) = <+> \\ w'(x_{(u,v)}, u) &= w(u,v) \quad \forall\,(u,v) \in E \quad \text{mit} \quad l(u,v) = <-> \end{aligned}$$

$$M = \{< x_{(u,v)}, v > \ |\, (u,v) \in E\,,\ l(u,v) = <->\}$$

Abbildung 2 zeigt die Umwandlung des Graphen aus Abbildung 1.

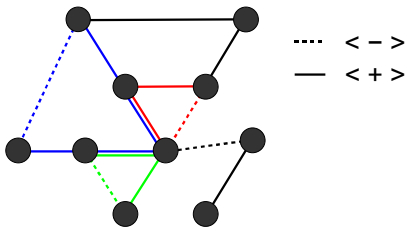

Abbildung 3: fehlerhafte Kreise

2.2 Korrektheit

Um zu zeigen, dass der $O(log(k))$-approximative Algorithmus für Multicut durch die Transformation zu einem $O(log(n))$-approximativen Algorithmus von Correlation Clustering wird, gehe ich wie folgt vor:

Ich werde sogenannte fehlerhafte Kreise einführen und einen entscheidenden Zusammenhang zwischen diesen Kreisen und einer optimalen Einteilung in Cluster herstellen. Damit wird es möglich, das folgende Theorem zu beweisen.

Theorem 1. *(G', M) besitzt einen Multicut mit Gewicht W genau dann, wenn eine eine Einteilung von G in Cluster mit Gewicht W existiert. Die entsprechende Lösung von Multicut kann aus der Lösung für Correlation Clustering konstruiert werden und umgekehrt.*

Definition 2: Ein *fehlerhafter Kreis* ist ein einfacher Kreis, der genau eine negative Kante enthält.

Abbildung 3 zeigt drei der fehlerhaften Kreise in meinem Beispielgraphen.

Lemma 3. *G besitzt eine fehlerfreie Einteilung in Cluster genau dann, wenn G keinen fehlerhaften Kreise enthält.*

Beweis. Falls eine fehlerfreie Einteilung in Cluster existiert, so ist auch die Einteilung fehlerfrei, in der jedes Cluster definiert wird durch

$$C(u) = \{v \in V \mid u \text{ und } v \text{ sind durch einen Pfad aus positiven Kanten verbunden}\}$$

wobei u ein beliebiger Knoten des Clusters $C(u)$ ist. Ich werde diese Einteilung im folgenden Standardeinteilung nennen.

Angenommen, es gibt einen Knoten $v \in V$ der durch einen Pfad aus positiven Kanten mit u verbunden ist aber in einem anderen Cluster als u liegt. Dann verbindet mindestens eine der Kanten auf dem Pfad von u nach v Knoten aus verschiedenen Clustern und die Einteilung ist nicht fehlerfrei. Ist ein Knoten $w \in V$ durch keinen Pfad aus positiven Kanten mit u verbunden, so erzwingen die positiven Kanten auch nicht, dass er in einem Cluster mit u liegt. Die Standardeinteilung ist also eine Einteilung ohne positive Fehler mit minimalen Clustergößen.

Nun besitze G eine fehlerfreie Einteilung in Cluster. Angenommen G enthält auch einen fehlerhaften Kreis, dann sind alle Knoten auf diesem Kreis im selben

Cluster, da sie alle durch einen positiven Pfad miteinander verbunden sind. Die negative Kante des fehlerhaften Kreises verbindet also zwei Knoten des selben Clusters, was einen Wiederspruch zur Fehlerfreiheit der Einteilung ist.

Wenn G umgekehrt keinen fehlerhaften Kreis besitzt, so ist die Standardeinteilung fehlerfrei. Die durch eine positive Kante verbundenen Knoten sind durch einen positiven Pfad verbunden also im selben Cluster. Angenommen es gäbe eine negative Kante innerhalb eines Clusters, so wären ihre inzidenten Knoten durch einen Pfad aus positiven Kanten verbunden und damit ein fehlerhafter Kreis gefunden. q.e.d.

Lemma 4. *Das Gewicht einer optimalen Einteilung in Cluster ist gleich dem minimalen Gewicht einer Kantenmenge, deren Entfernen alle fehlerhaften Kreise entfernt.*

Beweis. Sei S eine Menge minimalen Gewichts mit der geforderten Eigenschaft. In $\tilde{G} = (V, E \backslash S)$ gibt es nach Lemma 3 eine fehlerfreie Einteilung. Diese Einteilung hat in G ein Gewicht von höchstens $W = \sum_{(u,v)\in S} w(u, v)$. Angenommen das Gewicht der Einteilung wäre echt kleiner als W, dann ist eine der Kanten aus S kein Fehler der Einteilung in G. Sie kann aus S entfernt werden, ohne dass S die geforderte Eigenschaft verliert. Das ist ein Widerspruch zur Minimalität von S. q.e.d.

Meiner Meinung nach sind diese Aussagen zu fehlerhaften Kreisen eine der zentralen Errungenschaften der Arbeit von Emanuel und Fiat. Sie liefern ein sehr gutes Kriterium dafür, wie viele Fehler beim Correlation Clustering mindestens gemacht werden müssen und aus welcher Kantenmenge diese fehlerhaften Kanten stammen. Mit Hilfe von Lemma 3 und Lemma 4 ist es nun auch möglich, Theorem 1 zu beweisen.

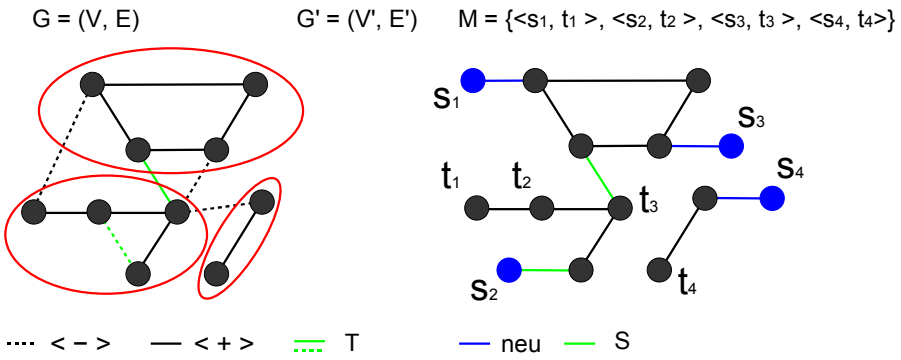

Abbildung 4: Clustering in G wird zum Multicut in G'

Beweis zu Theorem 1. Sei $C = (C_1, \ldots, C_n)$ eine Einteilung von G in Cluster, T die Menge aller fehlerhaften Kanten in C mit Gewicht $w(T) = \sum_{(u,v)\in T} w(u, v)$. Die Einteilung in Cluster C ist also in $\tilde{G} = (V, E \backslash T)$ fehlerfrei. Die Menge S sei

definiert durch

$$S := \{(u,v) \in T \,|\, l(u,v) =< + >\} \cup \{(x_{(u,v)}, u) \,|\, (u,v) \in T,\, l(u,v) =< - >\}$$

Die Menge S hat das selbe Gewicht wie T, da jede Kanten $(x_{(u,v)}, u) \in E'$ mit $(u,v) \in T$ und $l(u,v) =< - >$ das gleich Gewicht hat wie (u,v). Ein Beispiel zu den Mengen T und S ist in Abbildung 4 gegeben.

Angenommen S ist kein Multicut in (G', w', M), dann gibt es ein Paar $< s_i, t_i > \in M$ und einen Pfad p in $E' \backslash S$, der s_i und t_i miteinander verbindet. Dabei sei s_i o.B.d.A der Knoten aus $V' \backslash V$, also ein Knoten $x_{(u,v)}$ mit $(u,v) \in E$ und $l(u,v) =< - >$. Damit entspricht t_i dem Knoten v. Abbildung 5 zeigt die Knoten s_i, t_i sowie den Pfad p in $E' \backslash S$.

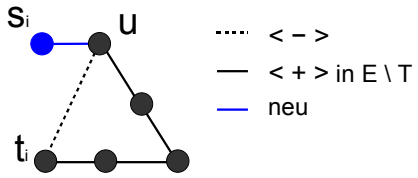

Abbildung 5: fehlerhafter Kreis

Der Knoten s_i hat Grad 1, seine einzige Kante ist mit dem Knoten u verbunden. Also ist die Kante (s_i, u) nicht in S enthalten und damit die negative Kante $(u,v) \in E$ nicht in T. Auch alle anderen neu entstandenen Knoten $x_{(u',v')}$ haben Grad 1 und damit liegt keiner von ihnen auf dem Pfad p. Die Kanten des Pfades sind folglich positive Kanten aus $E \backslash T$. Zusammen mit der negativen Kante (u,v) ergibt das einen fehlerhaften Kreis in $\tilde{G} = (V, E \backslash T)$, was mit Lemma 3 und dem Beweis zu Lemma 4 ein Widerspruch dazu ist, dass T die Menge der fehlerhaften Kanten des Clusterings in G ist. Damit ist bewiesen, dass aus einem Clustering in (G, w, l) ein Multicut in (G', w', M) mit dem selben Gewicht konstruiert werden kann.

Sei nun umgekehrt S ein Multicut in (G', w', M). Die Menge T sei definiert durch

$$T := \{(u,v) \in S \,|\, (u,v) \in E\} \cup \{(u,v) \,|\, (x_{(u,v)}, u) \in S\}$$

Die Standardeinteilung in Cluster auf $G \backslash T$, gegeben durch

$$C(u) = \{v \in V \,|\, \begin{array}{l} u \text{ und } v \text{ sind durch einen Pfad aus} \\ \text{positiven Kanten in } G \backslash T \text{ verbunden}\} \end{array}$$

ist auf $G \backslash T$ fehlerfrei, hat also ein Gewicht von maximal $w(T) = w(S)$. Angenommen die Standardeinteilung auf $G \backslash T$ hätte einen Fehler in $G \backslash T$, so enthielte sie nach Lemma 3 einen fehlerhaften Kreis. Dieser fehlerhafte Kreis induziert analog zu oben einen Pfad in $E' \backslash S$, der ein Paar $< s_i, t_i > \in M$ verbindet. Das ist ein Widerspruch dazu, dass S in (G', w', M) ein Multicut ist.

Damit ist folgende Ungleichungskette bewiesen:

w(optimales Correlation Clustering)

$$
\begin{aligned}
&= w(\text{Multicut induziert durch optimales Correlation Clustering}) \\
&\geq w(\text{minimaler Multicut}) \\
&\geq w(\text{Correlation Clustering induziert durch minimalen Multicut}) \\
&\geq w(\text{optimales Correlation Clustering})
\end{aligned}
$$

Hieraus wiederum folgt die Gleichheit und Theorem 1 ist bewiesen. q.e.d.

Als Folgerung aus Theorem 1 kann nun jeder Approximationsalgorithmus von Multicut für Correlation Clustering verwendet werden. Der $O(log(k))$–Approximationsalgorithmus für Multicut aus [GaVa96] wird dabei zu einem $O(log(n))$–Approximationsalgorithmus für Correlation Clustering.

$$
\begin{aligned}
O(log(k)) &= O(log(|\text{negative Kanten in } G|)) \\
&\subseteq O(log(|\text{Kanten in } G)) \\
&\subseteq O(log(n^2)) = O(log(n))
\end{aligned}
$$

3 Multicut nach Correlation Clustering

Um zu beweisen, dass Correlation Clustering auch in ungewichteten Graphen APX–schwer ist, geben Fiat und Emanuel eine polynomiale Transformation einer ungewichteten Instanz von Multicut $MC = (G', M)$ in eine ungewichtete Instanz von Correlation Clustering $CC = (G, l)$ an. Das Prinzip dieser Transformation ist im gewichteten Fall einfacher zu verstehen, daher zeigen sie zunächst noch einmal, dass Correlation Clustering in gewichteten Graphen APX–schwer ist.

3.1 Der gewichtete Fall

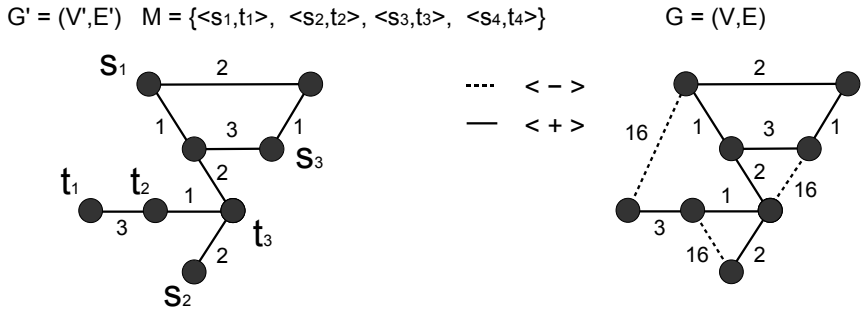

Abbildung 6: Umwandlung: Multicut in Correlation Clustering

Um eine Instanz von Multicut $MC = (G', w', M)$ in eine Instanz von Correlation Clustering $CC = (G, l, w)$ umzuwandeln, wird jede Kante (u, v) als positiven Kante beibehalten, außer $< u, v >$ ist ein Element von M. Eine Kante (u, v)

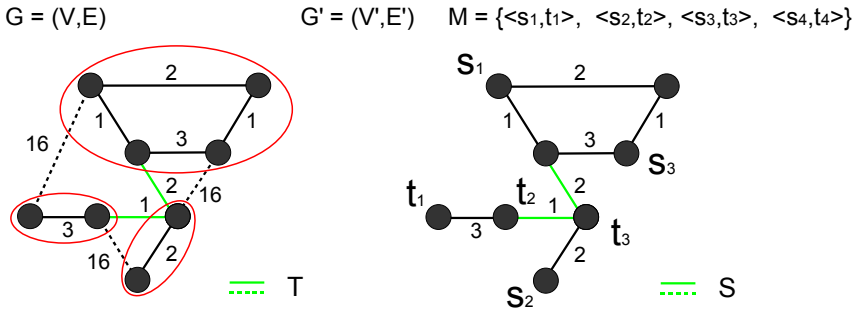

Abbildung 7: T ohne negative Kanten

mit $< u, v > \in M$ muss in einer Lösung von MC auf jeden Fall in S aufgenommen werden und kann daher o.B.d.A. entfallen.

Anschließend wird für jedes Paar $< s_i, t_i > \in M$ eine negative Kante (s_i, t_i) mit Gewicht $w(s_i, t_i) := \sum_{(u,v)\in E'} w(u, v) + 1$ eingefügt. Die Transformation ist polynomial, da sie höchstens $O(n^2)$ viele Kanten hinzufügt und das maximale Kantengewicht um nicht mehr als einen multiplikativen Faktor von n erhöht.

Abbildung 6 zeigt beispielhaft eine solche Transformation. Formal lässt sich die Konstruktion folgendermaßen beschreiben:

$$\text{MC:} \quad G' = (V', E') \qquad w' : E \to \mathbb{R}^+ \qquad M = \{< s_1, t_1 >, \ldots, < s_k, t_k >\}$$

$$\text{CC:} \quad G = (V, E) \qquad w : E \to \mathbb{R}^+$$

$$\begin{aligned} V &:= V' \\ E &:= E' \cup E_{\text{neu}} \\ E_{\text{neu}} &:= \{(s_i, t_i) \mid < s_i, t_i > \in M\} \end{aligned}$$

$$\begin{aligned} l(u,v) &:= < + > && \forall\, (u,v) \in E' \\ l(u,v) &:= < - > && \forall\, (u,v) \in E_{\text{neu}} \\ w(u,v) &:= w'(u,v) && \forall\, (u,v) \in E' \\ w(u,v) &:= \textstyle\sum_{(u,v)\in E'} w(u,v) + 1 && \forall\, (u,v) \in E_{\text{neu}} \end{aligned}$$

Theorem 5. *Eine Einteilung von G in Cluster mit Gewicht W induziert einen Multicut in (G', M) mit einem Gewicht von maximal W. Eine optimale Einteilung von G in Cluster induziert einen optimalen Multicut in (G', M).*

Beweis. Sei T die Menge der fehlerhaften Kanten in einem Correlation Clustering von (G, w, l). Falls T keine negativen Kanten enthält, so ist $S = T$ ein Multicut in (G', w, M). Jeder Pfad in $E'\backslash T$, der ein Paar $< s_i, t_i > \in M$ verbindet erzeugt zusammen mit der negativen Kante $(s_i, t_i) \in E$ einen fehlerhaften Kreis in G. Abbildung 7 zeigt einen solchen Fall.

Enthält T eine negative Kanten (s, t), so bilde o.B.d.A aus s ein neues Cluster. Da $w(s, t) > \sum_{(u,v)\in E'} w(u, v)$ hat die neue Einteilung in Cluster geringeres Gewicht als die vorherige (siehe Abbildung 8). So entsteht nach und nach eine

Einteilung in Cluster ohne negative Kanten die ein geringeres Gewicht hat als die ursprüngliche. Die Fehlermenge dieser Einteilung ist dann analog zu oben ein Multicut S in (G', w', M), in diesem Fall mit kleinerem Gewicht als T.

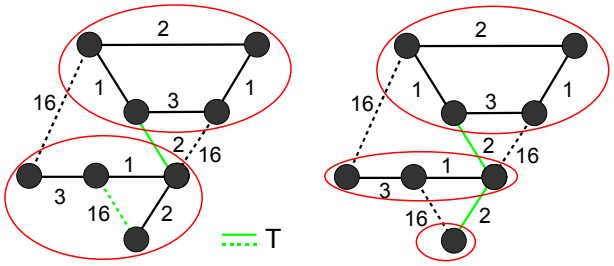

Abbildung 8: Entfernen einer negativen Kante.

Sei nun S ein minimaler Multicut in (G', w', M). Die Standardeinteilung in $\tilde{G} = (V, E \backslash S)$ ist fehlerfrei, denn jeder fehlerhafte Kreis in $\tilde{G}$ induziert einen Pfad in $E' \backslash S$ zwischen einem Paar $< s_i, t_i > \in M$. Die Standardeinteilung auf $\tilde{G}$ enthält also keine negativen Fehler in G und hat damit ein Gewicht von höchstens $w(S)$. Andererseits aber ist S ein minimaler Multicut in (G', w', M), also ist jede Kante $(u, v) \in S$ ein positiver Fehler und damit hat die Standardeinteilung ein Gewicht von $w(S)$. Aus diesen zwei Überlegungen folgt:

w(minimaler Multicut)

$=$ w(Correlation Clustering induziert durch den minimalen Multicut)
$\geq$ w(optimales Correlation Clustering)
$\geq$ w(Multicut induziert durch das optimale Correlation Clustering)
$\geq$ w(minimaler Multicut)

Damit ist bewiesen, dass ein optimales Correlation Clustering (G, w, l) einen optimalen Multicut in (G', w', M) induziert. q.e.d.

Für das gewichtete Correlation Clustering ist hiermit erneut die APX–Schwere bewiesen. Jeder Approximationsalgorithmus für Correlation Clustering kann genutzt werden, um einen Multicut zu finden. Ein konstanter Approximationsfaktor bleibt dabei erhalten.

3.2 Der ungewichtete Fall

Die Grundidee der Transformation im ungewichteten Fall ist identisch zu der Idee für den gewichteten Fall. Man erstellt negative Kanten zwischen den einzelnen Paaren $< s_i, t_i > \in M$ und erzeugen damit für jeden Pfad von s_i nach t_i in E' einen fehlerhaften Kreis in G. Damit das Correlation Clustering nicht die negative Kante als fehlerhaft wählt, konnte sie im gewichteten Fall ein im Vergleich zu den positiven Kanten extrem hohes Gewicht versehen werden. Im ungewichteten Fall ist das nicht möglich, hier müssen einige Hilfsknoten eingeführt werden.

Eine ungewichtete Instanz von Multicut $MC = (G', M)$ mit $G' = (V', E')$ und $M = \{< s_1, t_1 >, \ldots, < s_k, t_k >\}$ wird durch folgende Schritte in eine ungewichtete Instanz von Correlation Clustering $CC = (G, l)$ umgewandelt:

Die Knoten aus V' werden übernommen. Alle Kanten aus E' bekommen eine positive Beschriftung. Anschließend werden $n-1$ neue Knoten zu jeden Knoten $v \in V'$ mit $< u, v > \in M$ oder $< v, u > \in M$ hinzugefügt und zusammen mit v zu einer Clique Q_v aus positiven Kanten verbunden. In jedem Paar $< s_i, t_i > \in M$ wird s_i mit allen Knoten aus Q_{t_i} sowie t_i mit allen Knoten aus Q_{s_i} durch negative Kanten verbunden.

In Abbildung 9 habe ich zur Veranschaulichung eine Transformation am Beispiel durchgeführt.

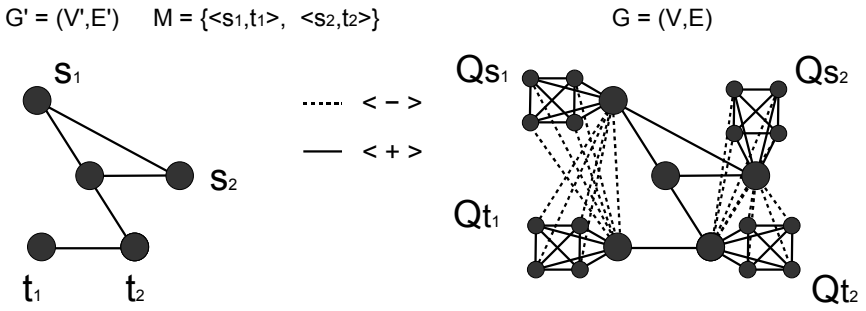

Abbildung 9: Transformation: Multicut nach Correlation Clustering

Analog zum gewichteten Fall muss das folgende Theorem bewiesen werden:

Theorem 6. *Eine Einteilung von G in Cluster mit Gewicht W induziert einen Multicut in (G', M) mit einem Gewicht von maximal W. Eine optimale Einteilung von G in Cluster induziert einen optimalen Multicut in (G', M).*

Zum Beweis sind eine Definition und ein vorbereitendes Lemma notwendig.

Definition 7: Eine Correlation Clustering von G heißt *pur*, wenn keine der Cliquen Q_v getrennt wird und für $< u, v > \in M$ die Cliquen Q_u und Q_v nicht im gleichen Cluster liegen.

Lemma 8. *Sei $C = (C_1, \ldots, C_l)$ ein Correlation Clustering in G. Falls C nicht pur ist, so kann es in ein pures Correlation Clustering verwandelt werden, das kein größeres Gewicht hat.*

Beweis. Existiert eine Clique Q_v, die durch C in zwei Teile, $Q_v \cap C_i$ und $Q_v \cap C_j$, getrennt wird, so erstelle ein neues Cluster aus den Knoten von Q_v. Durch die positiven Kanten der Form (v, x) mit $x \in V'$ können hierbei höchstens $n-1$ $(= |V'|-1)$ neue Fehler entstehen. Innerhalb von Q_v werden dafür mindestens $n-1$ (genau $|Q_v \cap C_i| * |Q_v \cap C_j|$) Fehler behoben.

Falls andererseits ein Paar $< u, v > \in M$ existiert, für das Q_u und Q_v im selben Cluster liegen, so separiere o.B.d.A. Q_v in einem eigenen Cluster. Auch hier

entstehen maximal $n-1$ neue Fehler durch positive Kanten der Form (v, x) mit $x \in V'$. Dafür werden $2n-1$ Fehler durch negative Kanten zwischen Q_u und Q_v behoben. q.e.d.

Mit Hilfe dieses Lemma lässt sich nun auch Theorem 6 beweisen:

Beweis zu Theorem 6. Sei $C = (C_1, \ldots, C_l)$ ein Correlation Clustering in G. Verwandle C nach Lemma 8 in eine pure Einteilung in Cluster C' die kein größeres Gewicht hat. Da nun keine zwei Cliquen Q_u und Q_v im selben Cluster liegen, sind alle fehlerhaften Kanten T in C positive Kanten. Damit ist T analog zum Beweis von Theorem 5 ein Multicut in (G', M). Das Gewicht des induzierten Multicuts ist also nicht größer als das das Correlation Clusterings C.

Ist S ein optimaler Multicut in (G', M), so induziert S auf G eine Einteilung in Cluster. Die Zusammenhangskomponenten von $G' \backslash S$ sind die einzelnen Cluster, Q_v wird zum Cluster von v hinzugefügt. Das Gewicht eines minimalen Multicuts ist damit eben so groß wie das Gewicht des induzierten Correlation Clusterings.

Die folgende Ungleichungskette schließt abermals den Beweis ab.

w(minimaler Multicut)

$$
\begin{aligned}
&= w(\text{Correlation Clustering induziert durch den minimalen Multicut}) \\
&\geq w(\text{optimales Correlation Clustering}) \\
&\geq w(\text{Multicut induziert durch das optimale Correlation Clustering}) \\
&\geq w(\text{minimaler Multicut})
\end{aligned}
$$

q.e.d.

Da das Problem Multicut in ungewichteten Graphen APX-schwer ist, ist nun bewiesen, dass auch Correlation Clustering in ungewichteten Graphen ist APX-schwer ist.

4 Correlation Clustering und Multicut

Durch die polynomiale Transformation von Correlation Clustering auf Multicut und umgekehrt können nun alle Approximationsalgorithmen und Komplexitätsfragen der zwei Probleme aufeinander übertragen werden. Zwei weitere solcher Ergebnisse sind zum Beispiel:

Es gibt einen $O(1)$-Approximation für Correlation Clustering in ungewichteten vollständigen Graphen (siehe [BaBlCh02]). Dieser Algorithmus wird durch die Transformation zu einem $O(1)$-Approximationsalgorithmus für Multicut in ungewichteten Graphen in denen jede Kante (u, v) vorhanden ist, für die $< u, v >$ nicht in M liegt. Außerdem gibt es für Multicut in planaren Graphen einen $O(1)$-Approximationsalgorithmus. Dieser wird zu einem $O(1)$-Approximations algorithmus für alle Instanzen von Correlation Clustering in denen die positiven Kanten einen planaren Graphen induzieren

Die Frage, ob es für Correlation Clustering einen Approximationsalgorithmus mit konstantem Faktor gibt bleibt weiterhin offen. Sie wurde für Multicut schon lange untersucht, bisher leider ohne Erfolg.

Literatur

[EmFi03] Dotan Emanuel und Amos Fiat:
Correlation Clustering
– Minimizing Disagreements on Arbitrary Weighted Graphs
Proc. 11th Annual European Symposium on Algorithms (ESA), 208–220, 2003.

[BaBlCh02] Nikhil Bansal, Avrim Blum und Shuchi Chawla:
Correlation Clustering
Foundation of Computer Science (FOCS), 238–147, 2002.

[GaVa96] Naveen Garg, Vijay V. Vazirani und Mihalis Yannakakis:
Approximate max-flow min-(multi)cut theorems and their applications.
SIAM Journal on Computing, 25:235–251, 1996.

[DaJo94] E. Dahlhaus, D.S. Johnson, C.H. Papadimitriou, P.D. Seymour, M. Yannakakis:
The Complexity of Multiterminal Cuts
SIAM Journal on Computing, 23:864–894, 1994.