Analyse und Evaluierung von Vergleichsmaßen für Graphclusterungen

Diplomarbeit
am
Institut für Theoretische Informatik
Universität Karlsruhe (TH)

von

Daniel Delling

Betreut durch:

Prof. Dorothea Wagner
Dipl.-Math. Marco Gaertler

Tag der Anmeldung: 1.September 2005
Tag der Abgabe: 22.Februar 2006

Danksagung

Meiner Betreuerin, Prof. Dorothea Wagner, danke ich für die Vergabe des interessanten Themas und die gute Betreuung an ihrem Institut. Ebenso danke ich meinem Betreuer Marco Gaertler sehr herzlich, da seine Tür immer für mich offen stand.

Für die Unterstützung während meines Studiums in jeglicher Hinsicht bin ich meinen Eltern, meinem Bruder und meiner Großmutter sehr dankbar.

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 22. Februar 2006

Inhaltsverzeichnis

1 Einleitung

Der Vergleich von Datenstrukturen hat in vielen Bereichen Anwendungsgebiete. Häufig sind diese Strukturen in gewisser Art gruppiert. Diese Gruppierung, sogenannte Clusterung einer Struktur ist vielen Fällen nicht eindeutig, sodass sich zwei Fragen aufdrängen:

1. Wie gut ist eine Clusterung?
2. Wie verschieden sind zwei Clusterungen?

Mit der zweiten Fragestellung beschäftigt sich diese Arbeit. Anwendungsfälle für diese Fragestellung gibt es in sehr verschiedenen Bereichen. In der Biochemie zum Beispiel gibt es sogenannte Proteininteraktionsdiagramme [LAB04], die ausdrücken, welche Proteine untereinander agieren. Es zeigt sich, dass die Interaktionen nicht gleichmäßig verteilt sind, sondern dass es Gruppen von Proteinen gibt, die viel miteinander und kaum mit anderen Proteinen agieren. Durch Anregung des Organismus´ – z.B. Erwärmung oder auch Erkrankung – kann sich nun das Interaktionsdiagramm und somit auch die Gruppierung der Proteine ändern. Folgende Frage stellt sich: Wie stark hat sich die Gruppierung verändert?

Da solche Proteininteraktionsdiagramme insbesondere für den interessantesten Fall, den Menschen, sehr umfangreich sind – das menschliche Genom codiert nach neuesten Erkenntnissen circa 20.000 - 25.000 Proteine [Int04] – ist es notwendig dieses Problem computergestützt zu lösen.

Ein mögliche Modellierung kommt aus der Graphentheorie. Ein Interaktionsdiagramm wird dabei als Graphen modelliert, bei dem die Knoten die Proteine darstellen. Eine Kante existiert zwischen zwei Proteinen, wenn diese miteinander interagieren. Es sei darauf hingewiesen, dass bereits hier das Problem vereinfacht wird, da nicht jede Proteininteraktion gleich stark ist. Die Gruppierung der Proteine wird durch eine Clusterung des Graphen dargestellt. Eine Clusterung ist hierbei eine disjunkte Aufteilung der Knoten, jede Gruppe von Knoten wird als Cluster bezeichnet. Da die Interaktion vor und nach der Anregung verglichen werden soll, kann man diese Modellierung vor und nach der Anregung durchführen. Somit vergleicht man zwei geclusterte Graphen.

Nicht nur in der Biochemie ist diese Problemstellung von Relevanz. Beispielsweise ist es interessant, die Veränderung der Clusterung des Internets zu untersuchen. Die

Computer stellen hierbei die Knoten und direkte Datenleitungen zwischen Computern die Kanten des Graphen dar. Cluster sind demnach Gruppen von Computern, die in gewisser Weise viele Datenleitungen untereinander haben und wenige zu dem restlichen Internet. In diesem Anwendungsfall wird besonders deutlich, dass eine computergestützte Lösung zwingend notwendig ist. Das Internet besteht heutzutage aus über 350.000.000 Computern (Stand Juli 2005, [ICS]), sodass eine Lösung von Hand nicht einmal im Ansatz Sinn macht.

Auch in der Datenanalyse [JMF99] kann interessant sein, inwieweit sich die Clusterung eines sich verändernden Datensatzes über viele Jahre verändert hat. Hierbei stellen die Datensätze die Knoten des Graphen dar und eine Kante kann beispielsweise eine Referenz zwischen den Datensätze bedeuten. In diesem Fall könnte eine Bestimmung der Veränderung der letzten Jahre sogar für Prognosen über die weitere Entwicklung des gesamten Datensatzes genutzt werden.

Allen drei Beispielen ist gemein, dass zwei in irgendeiner Form geclusterte Graphen vorliegen und man eine Aussage darüber haben möchte, wie verschieden diese beiden Clusterungen sind.

In dieser Arbeit wird intensiv der Lösungsansatz zur Benutzung von Abstandsmaßen untersucht. Eine solches Abstandsmaß hat als Eingabe zwei geclusterte Graphen und gibt einen Wert zwischen Null und Eins aus. Hierbei soll Null für Gleichheit und Eins für völlige Unabhängigkeit der beiden Clustergraphen stehen.

Die Hauptproblematik bei diesem Ansatz liegt darin, dass die menschliche Intuition für das Messen von Abständen von Anwendungsfall zu Anwendungsfall verschieden sein kann. Bereits die Gleichheit von Clusterungen ist nicht eindeutig zu definieren. Beispielsweise kann es Fälle geben, bei denen zwei Clusterungen, die gleich gut sind, strukturell aber völlig verschieden, als gleich bezeichnet werden sollen. In anderen Fällen steht Gleichheit für strukturelle Gleichheit. Noch problematischer wird die Definition von Abstand. Was genau bedeutet ein kleiner Abstand? Und was soll unter völliger Unabhängigkeit zweier Clusterungen verstanden werden?

Aus dieser Motivation heraus wird in dieser Arbeit die Abstandsmessung zwischen zwei Clusterungen mittels eines Abstandsmaßes systematisch analysiert. Dafür werden zunächst bestehende Lösungsansätze analysiert und gefolgert, dass verschiedene Abstandsarten für Clusterungen existieren. Diese Abstandsarten werden gegliedert und daraufhin untersucht, wie bisherige Lösungsansätze mit dieser Gliederung im Zusammenhang stehen. Entsprechend dieser Gliederung werden dann axiomatische Untersuchungen durchgeführt, um bestehende Abstandsmaße bzw. Lösungsansätze entsprechend der erstellten Gliederung einzuordnen. Für Abstandsarten, für die keine Maße in der Literatur zu finden sind, werden neue Abstandsmaße vorgestellt. Eine experimentelle Untersuchung aller Maße klärt, welche Maße sich den Vorgaben entsprechend verhalten. Zum Abschluß werden die Ergebnisse diskutiert und ein Ausblick auf sich anschließende Fragestellungen gegeben.

Gliederung

Um die Übersichtlichkeit zu verbessern, werden die Kapitel dieser Arbeit kurz zusammengefasst.

Kapitel 2 In diesem Kapitel werden mathematische Grundlagen geschaffen, um sich mit der gesamten Thematik Graphclusterung zu beschäftigen. Dabei soll nur ein grober Überblick geschaffen werden, für genauere Analysen einiger Algorithmen wird an entsprechender Stelle auf die Literatur verwiesen.

Kapitel 3 Thema dieses Kapitels ist die genaue Problemstellung für diese Arbeit und Vereinfachungen, die vorgenommen werden, um sich einer Lösung zu nähern. Ferner werden hier bisherige Lösungsansätze und deren Nachteile diskutiert.

Kapitel 4 Hier wird eine Gliederung der möglichen Arten von Abstände zwischen Clusterung vorgenommen. Dabei wird die Möglichkeit gezeigt, wie dies auf Abstandsmaße übertragen werden kann und welche Konsequenzen diese Gliederung für Gleichheit und Abstände hat.

Kapitel 5 Dieses Kapitel betrachtet mögliche Axiome für Abstandsmaße. Dabei werden diese entsprechend der Gliederung der Abstandsarten aus Kapitel 4 aufgeteilt.

Kapitel 6 Die Diskussion der Axiome aus Kapitel 5 findet sich in diesem Kapitel. Dabei ist das Ziel, die Nachteile von – auf den ersten Blick sinnvollen – Axiomen zu zeigen.

Kapitel 7 Dieses umfangreiche Kapitel stellt eine Vielzahl von Abstandsmaßen vor. Dabei wird die Gliederung aus Kapitel 4 beibehalten. Viele Maße sind der Literatur entnommen, einige sind Erweiterungen bekannter Maße und weitere sind Maße, die aus Überlegungen der Kapitel 4 und 5 motiviert sind. Für einige Maße wird zusätzlich angegeben, welche der in Kapitel 5 definierten Axiome das jeweilige Maß erfüllt.

Kapitel 8 Dieses Kapitel umfasst 4 Testszenarien, in denen alle in Kapitel 7 vorgestellt Abstandsmaße untersucht werden. Für jedes Szenario werden hierbei erste Ergebnisse vorgestellt.

Kapitel 9 Die Ergebnisse der verschiedenen Testreihen sind Thema dieses Kapitels. Dabei soll darauf eingegangen werden, welche Maße sich in allen Testreihen sehr intuitiv verhalten haben.

Kapitel 10 Eine Zusammenfassung aller Erkenntnisse dieser Arbeit und ein Ausblick auf sich anschließende Probleme findet sich in diesem Kapitel.

2 Grundlagen

Dieses Kapitel erläutert Grundlagen über Clusterungen und den Vergleich von Clusterungen, allerdings werden Grundlagen der Graphen– [Jun05] und Komplexitätstheorie [Weg03] als bekannt vorausgesetzt.

2.1 Graphen

Ein Graph wird mit $G = (V, E)$ bezeichnet, wobei V die Knotenmenge und $E \subseteq \binom{V}{2}$ die Kantenmenge des Graphen ist. Die Variable n soll die Anzahl der Knoten, die Variable m die Anzahl der Kanten angeben. Mit $A(G)$ wird die Adjazenzmatrix von G bezeichnet, die normalisierte Adjazenzmatrix mit $M(G)$. Sie berechnet sich mit $M(G) = D^{-1}(G)A(G)$, wobei $D(G)$ die Diagonalmatrix der Knotengrade ist.

Alle untersuchten Graphen sind ungerichtet, ungewichtet und einfach. Es werden nun einige Eigenschaften und Zusammenhänge von Graphen wiederholt.

Definition 1 *Ein Graph G heißt d-regulär, falls* $\deg(v) = d$ *für alle* $v \in V$ *gilt.*

Für d-reguläre Graphen ist die Anzahl an Kanten aus der Anzahl der Knoten berechenbar.

Korollar 1 *Sei der Graph G d-regulär. Dann gilt:*

$$m = \frac{dn}{2}$$

Insbesondere gilt für vollständige Graphen somit $m = n(n-1)/2$.

2.2 Clusterungen

Um sich mit Clusterungen beschäftigen zu können, benötigt man zunächst eine Definition für einen Clustergraphen.

Definition 2 *Ein Graph G heißt* Clustergraph*, falls G eine knotendisjunkte Vereinigung von Cliquen ist.*

Es werden aber nicht auschließlich Graphen, die aus knostendisjunkten Cliquen bestehen, betrachtet.

Definition 3 *Eine Partition der Knotenmenge V eines Graphen G in $\{C_1, \ldots, C_k\}$ mit C_i, C_j paarweise disjunkt nennt man* Clusterung $\mathcal{C}$ *des Graphen G. Die C_i einer Clusterung $\mathcal{C}$ werden als Cluster bezeichnet, k ist die Anzahl der Cluster. Mit $\mathbb{P}(V)$ bezeichnet man die Menge aller möglichen Clusterungen.*

Man kann diese Partition als Funktion auffassen, die jedem Knoten einen Cluster der Clusterung $\mathcal{C}$ zuordnet. Diese *Clusterungsfunktion* soll mit $cl_{\mathcal{C}} : V \mapsto \mathcal{C}$ bezeichnet werden. Die Umkehrfunktion $cl_{\mathcal{C}}^{-1} : \mathcal{C} \mapsto V$ gibt dementsprechend für einen Cluster alle in ihm enthaltenen Knoten aus.

Mit dieser Definition einer Clusterung kann man den Schnitt eines Graphen G als Clusterung mit 2 Clustern dieses Graphen G auffassen.

Lemma 1 *Der Schnitt C eines Graphen $G = (V, E)$ ist eine Clusterung des Graphen G mit $\mathcal{C} = \{C, V \setminus C\}$.*

Liegt ein Graph G und eine Clusterung $\mathcal{C}$ dieses Graphen vor, soll von einem *geclusterten* Graphen gesprochen werden.

Nicht jeder gegeclusterte Graph ist ein Clustergraph. Man kann aber jeden geclusterten Graphen durch Hinzunahme und Entfernung von Kanten in einen Clustergraphen überführen.

Definition 4 *Sei $G = (V, E)$ ein Graph mit einer Clusterung $\mathcal{C}$. Dann heißt die Kantenmenge $F_{\mathcal{C}}$, für die $G' = (V, E \Delta F)$ mit $E \Delta F := (F \setminus E) \cup (E \setminus F)$ ein Clustergraph ist,* Clustereditiermenge $F_{\mathcal{C}}$ *der Clusterung $\mathcal{C}$.*

Zu jedem Graphen gibt es mindestens eine Clusterung, für die $|F_{\mathcal{C}}|$ minimal ist. Diese wird mit $\mathcal{C}_{\text{MIN}}$ und die entsprechende Clustereditiermenge mit $F_{\mathcal{C}_{\text{MIN}}}$ bezeichnet.

Problem 1 (Cluster Editing Problem) *Sei $G = (V, E)$ ein Graph und $k \in \mathbb{N}$ ein natürliche Zahl. Das Entscheidungsproblem, ob es eine Clustereditiermenge F mit $|F| \leq k$ gibt, ist NP–vollständig [SST02].*

Somit ist auch die Berechnung von $\mathcal{C}_{\text{MIN}}$ NP-schwer. Im Folgenden zeigt sich, dass viele Probleme, die es im Gebiet der Graphclusterung gibt, NP-schwer sind.

Man kann die Kanten eines geclusterten Graphen in zwei Typen unterscheiden. Solche Kanten, die zwei Knoten des gleichen Clusters verbinden und solche Kanten, die zwei Knoten verschiedener Cluster verbinden.

Definition 5 *Gegeben seien ein Graph G und eine Clusterung $\mathcal{C} = \{C_1, \ldots, C_p\}$ des Graphen G. Man bezeichnet*

$$E_{inter}(\mathcal{C}) := \big\{\{u, v\} \in E \mid u \in C_i, v \in C_j, i \neq j\big\}$$

als Intercluster Kanten und analog

$$E_{intra}(\mathcal{C}) := \big\{\{u, v\} \in E \mid u, v \in C_i\big\}$$

als Intracluster Kanten der Clusterung $\mathcal{C}$.

Im Folgenden soll mit $m(\mathcal{C})$ die Anzahl der Intracluster Kanten und mit $\overline{m}(\mathcal{C})$ die Anzahl der Intercluster Kanten der Clusterung $\mathcal{C}$ bezeichnet werden.

Spezielle Clusterungen

Es gibt zwei triviale Clusterungen eines Graphen. Die Clusterung $\mathcal{C}^1 := \{V\}$ heißt 1-*Clusterung*, die Clusterung $\mathcal{C}^s := \{v_1, \ldots, v_n\}$ *Singleton-Clusterung*.

Desweiteren sollen noch zwei weitere Clusterungen festgelegt werden, die allerdings nur für quadratische Knotenzahlen definiert sind. Dafür sei $n = a^2$. Die Clusterung

$$\mathcal{C}^{\times} := \{C_1, \ldots, C_a\} \text{ mit } C_i = \{v_{(i-1)a+1}, v_{(i-1)a+2}, \ldots, v_{(i-1)a+a}\}$$

heißt *gleichmäßige Clusterung* und die Clusterung

$$\mathcal{C}^{\perp} := \{C_1, \ldots, C_a\} \text{ mit } C_i = \{v_i, v_{i+a} \ldots, v_{i+(a-1)a}\}$$

die *komplementäre gleichmäßige Clusterung*. Die beiden Clusterungen sind *komplementär zueinander*, d.h. es gibt kein Knotenpaar, dass bezüglich beider Clusterungen in einem Cluster ist.

2.3 Vergleich von Clusterungen

Im Folgenden seien $\mathcal{C} = \{C_1, \ldots, C_k\} \in \mathbb{P}(V)$ und $\mathcal{C}' = \{C'_1, \ldots, C'_l\} \in \mathbb{P}(V)$ zwei Clusterungen bezüglich des gleichen Graphen G.

2.3.1 Verknüpfungen auf Clusterungen

In diesem Abschnitt wird die Verfeinerung einer Clusterung $\mathcal{C}'$ und das Produkt bzw. die Vereinigung zweier Clusterungen definiert.

Definition 6 *Clusterung $\mathcal{C}'$ ist eine* Verfeinerung *von Clusterung $\mathcal{C}$, wenn gilt:*

$$\forall C'_j \in \mathcal{C}' \exists C_i \in \mathcal{C} : C'_j \subseteq C_i$$

Clusterung $\mathcal{C}'$ ist also eine *Verfeinerung* von Clusterung $\mathcal{C}$, wenn jeder Cluster von $\mathcal{C}'$ in einem Cluster von $\mathcal{C}$ enthalten ist.

Definition 7 *Die kleinste gemeinsame Verfeinerung von $\mathcal{C}$ und $\mathcal{C}'$, formal*

$$\mathcal{C} \times \mathcal{C}' := \{C_i \cap C'_j \mid C_i \in \mathcal{C}, C'_j \in \mathcal{C}', C_i \cap C'_j \neq \emptyset\}$$

heißt Produkt *von $\mathcal{C}$ und $\mathcal{C}'$.*

Analog kann man die Vereinigung zweier Clusterungen definieren.

Definition 8 *Die Clusterung, von der sowohl $\mathcal{C}$, als auch $\mathcal{C}'$ Verfeinerungen sind, formal*

$$\mathcal{C} \oplus \mathcal{C}' := \{C_i \cup C'_j \mid C_i \in \mathcal{C}, C'_j \in \mathcal{C}'\}$$

heißt Vereinigung *von $\mathcal{C}$ und $\mathcal{C}'$.*

2.3.2 Schnittmengen

Zunächst wird eine $k \times l$-Matrix, deren ij-ter Eintrag die Anzahl der Elemente in dem Schnitt der beiden Cluster $C_i \in \mathcal{C}$ und $C'_j \in \mathcal{C}'$ beinhaltet, definiert.

Definition 9 *Die Matrix*

$$\begin{aligned} \mathcal{CM}(\mathcal{C}, \mathcal{C}') &\in \mathbb{N}^{k \times l} \\ \mathcal{CM}(\mathcal{C}, \mathcal{C}') &:= (m_{ij}) \;\; mit \;\; m_{ij} := |C_i \cap C'_j|, 1 \leq i \leq k, 1 \leq j \leq l \end{aligned}$$

heißt Verschmelzungsmatrix *von $\mathcal{C}$ und $\mathcal{C}'$.*

Korollar 2 *Es gilt $\sum_i \sum_j m_{ij} = n$.*

Lemma 2 *Für $G = K_n$ ist $\sum_i \sum_j \binom{m_{ij}}{2}$ die Anzahl der Intraclusterkanten von $\mathcal{C} \times \mathcal{C}'$.*

Beweis. Es gilt $m_{ij} = |C_i \cap C'_j|$. Somit ist $\binom{m_{ij}}{2}$ die Anzahl der Kanten im Cluster $(C_i \cap C'_j)$. □

Man kann die Verschmelzungsmatrix dahingehend erweitern, dass man für jeden Schnitt nicht die Kardinalität des Schnittes sondern die Knotengradsumme des Schnittes nutzt.

Definition 10 *Die Matrix*

$$\begin{aligned} \mathcal{CM}^d(\mathcal{C},\mathcal{C}') &\in \mathbb{N}^{k\times l} \\ \mathcal{CM}^d(\mathcal{C},\mathcal{C}') &:= (m_{ij}^d) \;\; mit \;\; m_{ij}^d := \sum_{v\in C_i\cap C'_j} deg(V), 1\le i\le k, 1\le j\le l \end{aligned}$$

heißt gewichtete Verschmelzungsmatrix *von* $\mathcal{C}$ *und* $\mathcal{C}'$.

Der Zusammenhang zwischen gewichteter und normaler Verschmelzungsmatrix ist offensichtlich:

Korollar 3 *Wenn G d-regulär ist, gilt:*

$$\mathcal{CM} = \frac{1}{d}\mathcal{CM}^d$$

Insbesondere gilt dies auch für den vollständigen Graphen.

2.3.3 Paarzählung

Bei dem Vergleich von zwei Clusterungen kann man alle ungeordneten Paare der Knotenmenge danach klassifizieren, ob die beiden Knoten bezüglich der beiden Clusterungen gleich oder verschieden geclustert sind. Dies ergibt 4 Paarzählungsmengen:

$$\begin{aligned} S_{11} &:= \big\{\{u,v\} \mid u,v\in V, cl_{\mathcal{C}}(u)=cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u)=cl_{\mathcal{C}'}(v)\big\} \\ S_{00} &:= \big\{\{u,v\} \mid u,v\in V, cl_{\mathcal{C}}(u)\neq cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u)\neq cl_{\mathcal{C}'}(v)\big\} \\ S_{01} &:= \big\{\{u,v\} \mid u,v\in V, cl_{\mathcal{C}}(u)\neq cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u)=cl_{\mathcal{C}'}(v)\big\} \\ S_{10} &:= \big\{\{u,v\} \mid u,v\in V, cl_{\mathcal{C}}(u)=cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u)\neq cl_{\mathcal{C}'}(v)\big\} \end{aligned}$$

Mit $n_{ab} := |S_{ab}|, a,b\in\{0,1\}$ werden die Kardinalitäten der Mengen S_{ab} bezeichnet. Ferner gilt:

$$n_{11}+n_{00}+n_{10}+n_{01} = \binom{n}{2}$$

Da diese Mengen alle Knotenpaare des Graphen betrachten, sollen sie als *globale* Paarzählungsmengen bezeichnet werden.

Lokale Paarzählung

Bei der lokalen Paarzählung betrachtet man nicht alle Paare, sondern nur die Paare, die im Graphen über eine Kante verbunden sind. Entsprechend den globalen

Paarzählungsmengen kann man nun 4 lokale Paarzählungsmengen definieren:

$$\begin{aligned}
E_{11} &:= \left\{\{u,v\} \mid \{u,v\} \in E, cl_{\mathcal{C}}(u) = cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u) = cl_{\mathcal{C}'}(v)\right\} \\
E_{00} &:= \left\{\{u,v\} \mid \{u,v\} \in E, cl_{\mathcal{C}}(u) \neq cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u) \neq cl_{\mathcal{C}'}(v)\right\} \\
E_{01} &:= \left\{\{u,v\} \mid \{u,v\} \in E, cl_{\mathcal{C}}(u) \neq cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u) = cl_{\mathcal{C}'}(v)\right\} \\
E_{10} &:= \left\{\{u,v\} \mid \{u,v\} \in E, cl_{\mathcal{C}}(u) = cl_{\mathcal{C}}(v), cl_{\mathcal{C}'}(u) \neq cl_{\mathcal{C}'}(v)\right\}
\end{aligned}$$

Analog zu den globalen Mengen werden mit $e_{ab} := |E_{ab}|, a, b \in \{0,1\}$ die Kardinalitäten der MengenE_{ab} bezeichnet. Bei genauerer Betrachtung fällt auf, dass die lokalen Paarzählungsmengen Kantenmengen sind. $E_{11}(E_{00})$ enthält dabei genau die Kanten, die sowohl in $\mathcal{C}$, als auch in $\mathcal{C}'$ Intra(Inter)-clusterkanten sind. Dementsprechend enthält $E_{10}(E_{01})$ die Kanten, die in $\mathcal{C}$ Intra(Inter)- und in $\mathcal{C}'$ Inter(Intra)-clusterkanten sind. Da eine Kante entweder Intra- oder Interclusterkante ist, gilt:

$$E_{11} \uplus E_{00} \uplus E_{10} \uplus E_{01} = E$$

Somit gilt auch:

$$e_{11} + e_{00} + e_{10} + e_{01} = |E|$$

Korollar 4 *In dem vollständigen Graphen K_n entsprechen die globalen Paarzählungsmengen den lokalen. Es gilt also $S_{ab} = E_{ab}, a, b \in \{0,1\}$.*

Zusammenhang Zählungsmengen zu Verschmelzungsmatrix Die Kardinalität der globalen Paarzählungsmengen lassen sich über die Verschmelzungsmatrix berechnen.

$$\begin{aligned}
n_{11} &= \sum_i \sum_j \binom{|C_i \cap C'_j|}{2} \\
&= \sum_i \sum_j \binom{m_{ij}}{2} \\
&= \frac{1}{2}\left(\sum_i \sum_j m_{ij}^2 - \sum_i \sum_j m_{ij}\right) \\
&\overset{\text{Kor. 2}}{=} \frac{1}{2}\left(\sum_i \sum_j m_{ij}^2 - n\right)
\end{aligned}$$

Die Menge S_{10} enthält alle Paare, die bezüglich $\mathcal{C}$ in einem Cluster sind abzüglich der Paare, die auch bezüglich $\mathcal{C}'$ in einem Cluster sind. Somit gilt für n_{10}:

$$\begin{aligned} n_{10} &= \sum_i \binom{|C_i|}{2} - n_{11} \\ &= \frac{1}{2}\Big(\sum_i |C_i|^2 - \underbrace{\sum_i |C_i|}_{=n} - \sum_i \sum_j m_{ij}^2 + n\Big) \\ &= \frac{1}{2}\Big(\sum_i |C_i|^2 - \sum_i \sum_j m_{ij}^2\Big) \end{aligned}$$

Analog gilt für n_{01}:

$$n_{01} = \frac{1}{2}\Big(\sum_j |C'_j|^2 - \sum_i \sum_j m_{ij}^2\Big)$$

2.3.4 Informationstheorie

Bei dem Vergleich von Clusterungen kann man auch Ergebnisse aus der Informationstheorie [Sha48] heranziehen. Für ein Verständnis werden kurz informationstheoretische Begriffe für Clusterungen eingeführt. Dabei wird der Entropiebegriff auf kantenbasierte Entropie erweitert.

Entropie

Die *Entropie* S einer beliebigen Information, z.B. eines Textes, mit einem Alphabeth Σ ist definiert durch

$$S(T) = -\sum_{i \in \Sigma} p_i \log_2(p_i)$$

wobei p_i die Wahrscheinlichkeit ist, den Buchstaben i im Text T zu finden. Entropie wird in Bits gemessen und es werden $S(T) \cdot |T|$ Bits für die Darstellung von T benötigt.

Man kann nun einer Clustering $\mathcal{C}$ eine Entropie $\mathcal{H}(\mathcal{C})$ zuordnen: Unter der Annahme, dass jeder Knoten mit gleicher Wahrscheinlichkeit betrachtet wird, ist die Wahrscheinlichkeit, dass ein Knoten v in Cluster $C_i \in \mathcal{C}$ ist, gerade $P(i) := |C_i|/n$. Somit lässt sich die Entropie einer Clusterung mit

$$\mathcal{H}(\mathcal{C}) := -\sum_{i=1}^{k} P(i) \log_2 P(i)$$

definieren. Umgangssprachlich kann man sagen, dass die Entropie einer Clusterung ein Maß für die Unsicherheit von $cl(v)$ mit zufällig gewähltem $v \in V$ ist.

Lemma 3 *Für eine beliebige Clusterung $\mathcal{C}$ gilt $0 \leq \mathcal{H}(\mathcal{C}) \leq \log_2(n)$.*

Beweis. Mit $0 \leq P(i) = |C_i|/n \leq 1$ folgt $\log_2(P(i)) \geq 0$. Somit ist die Entropie immer positiv. Die obere Schranke wird gezeigt mit:

$$\begin{aligned}
\mathcal{H}(\mathcal{C}) &= -\sum_{i=1}^{k} \frac{|C_i|}{n} \log_2(\frac{|C_i|}{n}) \\
&= \underbrace{\sum_{i=1}^{k} \frac{|C_i|}{n}}_{=1} \log_2(n) - \underbrace{\frac{1}{n} \sum_{i=1}^{k} |C_i| \log_2(|C_i|)}_{\geq 0} \\
&\leq \log_2(n)
\end{aligned}$$

□

Diese Schranken werden auch erreicht. Den größten und kleinsten Wert nimmt die Entropie einer Clusterung für die 1–Clusterung $\mathcal{C}^1$ und für die Singleton–Clusterung $\mathcal{C}^s$ an:

$$\begin{aligned}
\mathcal{H}(\mathcal{C}^1) &= -\sum_{i=1}^{1} \frac{n}{n} \underbrace{\log_2(\frac{n}{n})}_{=0} \\
&= 0 \\
\mathcal{H}(\mathcal{C}^s) &= -\sum_{i=1}^{n} \frac{1}{n} \log_2(\frac{1}{n}) \\
&= \log_2(n)
\end{aligned}$$

Korrelationsinformation

Man kann diese Notation der Entropie zu der *Korrelationsinformation* zweier Clusterungen $\mathcal{C}$ und $\mathcal{C}'$ erweitern. Sie beschreibt, inwieweit sich die Unsicherheit für $\mathcal{C}'$ ändert, wenn $\mathcal{C}$ bekannt ist. Voraussetzung hierbei ist, dass $\mathcal{C}$ und $\mathcal{C}'$ auf der gleichen Knotenmenge definiert sind. Mit $P(i,j)$, der Wahrscheinlichkeit, dass ein Knoten in Cluster $C_i \in \mathcal{C}$ und $C'_j \in \mathcal{C}'$ liegt, formal

$$P(i,j) := \frac{|C_i \cap C'_j|}{n}$$

definiert man die Korrelationsinformation zweier Clustering $\mathcal{C}, \mathcal{C}'$ mit:

$$\mathcal{I}(\mathcal{C}, \mathcal{C}') := \sum_{i=1}^{k} \sum_{j=1}^{l} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$

Die Korrelationsinformation ist eine Metrik auf allen Clusterung und ist nach oben durch das Minimum der beiden Clusterungen beschränkt [Mei03].

Korollar 5 *Es gilt* $0 \leq \mathcal{I}(\mathcal{C}, \mathcal{C}') \leq min\{\mathcal{H}(\mathcal{C}), \mathcal{H}(\mathcal{C}')\} \leq \log_2(n)$.

Kantenentropie

In einer Erweiterung der Entropie wählt man nicht jeden Knoten, sondern jeden Endpunkt einer Kante gleich wahrscheinlich. Dementsprechend ist die Wahrscheinlichkeit, dass der Endpunkt einer Kante in Cluster $C_i \in \mathcal{C}$ liegt, gerade

$$P_E(i) := \frac{\sum_{v \in C_i} \deg(v)}{2m}$$

und somit ist die Kantenentropie einer Clusterung mit

$$\mathcal{H}_E(\mathcal{C}) := -\sum_{i=1}^{k} P_E(i) \log_2 P_E(i)$$

definiert. Bei der Kantenentropie muss gefordert werden, dass für alle Knoten aus v der Knotengrad $\deg(v)$ größer Null ist. Ansonsten kann bei bestimmten Clusterungen der Ausdruck im Logarithmus Null werden.

Lemma 4 *Für eine beliebige Clusterung* $\mathcal{C}$ *gilt*

$$0 \leq \mathcal{H}_E(\mathcal{C}) \leq \log_2(2m)$$

Beweis. Da $0 \leq P_E(i) \leq 1$ gilt die Positivität. Die obere Schranke wird analog zum Beweis von Lemma 3 gezeigt:

$$\begin{aligned}
\mathcal{H}_E(\mathcal{C}) &= -\sum_{i=1}^{k} \frac{|\sum_{v \in C_i} \deg(v)|}{2m} \log_2 \left(\frac{|\sum_{v \in C_i} \deg(v)|}{2m}\right) \\
&= \log_2(2m) \underbrace{\sum_{i=1}^{k} \frac{\left|\sum_{v \in C_i} \deg(v)\right|}{2m}}_{=1} - \underbrace{\frac{1}{2m} \sum_{i=1}^{k} \log_2 \left(\sum_{v \in C_i} \deg(v)\right)}_{\geq 0} \\
&\leq \log_2(2m)
\end{aligned}$$

□

Zusammenhang Entropie und Kantenentropie Die Kantenentropie entspricht der Entropie, wenn G d-regulär ist.

Lemma 5 *Sei G d-regulär. Dann gilt für beliebiges $\mathcal{C}$:*

$$\mathcal{H}_E(\mathcal{C}) = \mathcal{H}(\mathcal{C})$$

Beweis. Wegen d-Regularität von G gilt nach Korollar 1 für die Kantenzahl $m = dn/2$. Daher gilt:

$$P_E(i) = \frac{\sum_{v \in C_i} \deg(v)}{2m} = \frac{d|C_i|}{2\frac{d}{2}n} = P(i)$$

Daraus folgt $\mathcal{H}_E(\mathcal{C}) = \mathcal{H}(\mathcal{C})$. □

Somit gilt auch für $G = K_n$ die Gleichheit von Entropie und Kantenentropie.

Kantenkorrelationsinformation

Analog zur Korrelationsinformation kann man auch die *Kantenkorrelationsinformation* für zwei Clusterungen $\mathcal{C}, \mathcal{C}'$ bestimmen. Neben der Knotenmenge muss hier auch die Kantenmenge der zugehörigen Graphen gleich sein. Analog entspricht

$$P_E(i,j) := \frac{\sum_{v \in C_i \cap C'_j} \deg(v)}{2m}$$

der Wahrscheinlichkeit, dass ein Endpunkt einer Kante in Cluster $C_i \in \mathcal{C}$ und $C'_j \in \mathcal{C}'$ liegt. Somit ist

$$\mathcal{I}_E(\mathcal{C}, \mathcal{C}') := \sum_{i=1}^{k} \sum_{j=1}^{l} P_E(i,j) \log_2 \frac{P_E(i,j)}{P_E(i)P_E(j)}$$

die Kantenkorrelationsinformation der beiden Clusterungen $\mathcal{C}, \mathcal{C}'$.

Analog zu Korrolar 5:

Korollar 6 *Es gilt $0 \leq \mathcal{I}_E(\mathcal{C}, \mathcal{C}') \leq min\{\mathcal{H}_E(\mathcal{C}), \mathcal{H}_E(\mathcal{C}')\} \leq \log_2(2m)$.*

Aufgrund von Lemma 5 gilt für reguläre Graphen die Gleichheit von $\mathcal{I}(\mathcal{C}, \mathcal{C}')$ und $\mathcal{I}_E(\mathcal{C}, \mathcal{C}')$:

Lemma 6 *Wenn G d-regulär ist, gilt für beliebige $\mathcal{C}, \mathcal{C}'$*

$$\mathcal{I}_E(\mathcal{C}, \mathcal{C}') = \mathcal{I}(\mathcal{C}, \mathcal{C}')$$

Beweis. Analog zu Beweis von Lemma 5. □

2.4 Indizes

Indizes sind Bewertungsfunktionen für Clusterungen. Eine solche Funktion misst die Signifikanz einer Clusterung mit einem Wert zwischen Null und Eins. Ein Wert nahe Eins soll dabei eine sehr signifikante Clusterungen ausweisen.

Definition 11 *Sei $G = (V, E)$ ein geclusterter Graph mit der zugehörigen Clusterung $\mathcal{C}$. Eine Bewertungsfunktion $i(\mathcal{C})$ der Clusterung $\mathcal{C}$ mit folgenden Eigenschaften*

- $i(\mathcal{C}, G) \in [0, 1]$
- *Ein Wert nahe* 0 *indiziert eine schlechte Clusterung.*
- *Ein Wert von* 1 *indiziert eine optimale Clusterung.*

wird als Index *bezeichnet.*

Durch die NP–Vollständigkeit des Cluster Editing Problems, ist die Berechnung eines Indexes, der $|F|$ minimiert, im Allgemeinen NP–schwer. Außerdem muss die Clusterung $\mathcal{C}_{\text{MIN}}$ eines Graphen nicht zwangsläufig die intuitiv beste sein. Vielmehr kann diese intuitive Clusterung von Anwendungsfall zu Anwendungsfall stark variieren. Somit gibt es wohl keinen ultimativen Index, sondern mehrere, die für verschiedene Anwendungen mehr oder minder sinnvoll sind.

Es werden die in dieser Arbeit genutzten Indizes vorgestellt. Von jedem der hier vorgestellten Indizes exitistiert eine gewichtete Version, d.h. die die Kantengewichte betrachtet. Da nur ungewichtete Graphen betrachtet werden, werden jeweils nur die ungewichteten Versionen vorgestellt. In [BE05] findet sich eine umfangreiche Analyse der Indizes.

2.4.1 Coverage

Der *Coverage* Index misst die Anzahl der Intraclusterkanten im Verhältnis zu der Gesamtkantenzahl.

$$cov(\mathcal{C}) := \frac{m(\mathcal{C})}{m} = \frac{m(\mathcal{C})}{m(\mathcal{C}) + \overline{m}(\mathcal{C})}$$

Dieser Index hat einige Nachteile. Zum Beispiel wird die 1-Clusterung immer als optimal bewertet. Somit wird die Maximierung von Coverage immer in der 1-Clusterung enden.

2.4.2 Peformance

Der *Performance* Index von einer Clusterung zählt die Anzahl der „korrekt interpretierten Knotenpaaren“. Man addiert die Anzahl der Intraclusterkanten zu der Anzahl der Knotenpaare, die richtigerweise nicht verbunden sind, da die Knoten des jeweiligen Paares verschiedenen Clustern zugeordnet sind. Dieser Wert wird in das Verhältnis mit der Anzahl der Kanten in einem vollständigen Graphen mit n Knoten gesetzt.

$$\begin{aligned} per(\mathcal{C}) &:= \frac{m(\mathcal{C}) + \sum_{\{v,w\}\notin E, v\in C_i, w\in C_j, i\neq j} 1}{\frac{1}{2}n(n-1)} \\ &= 1 - \frac{2\overline{m}(\mathcal{C}) - 2m(\mathcal{C}) + \sum_{i=1}^{k} |C_i|(|C_i|-1)}{n(n-1)} \end{aligned}$$

Zusammenhang zu Clustereditiermenge $F_\mathcal{C}$ Die Kardinalität der Clustereditiermenge $F_\mathcal{C}$ ist die Anzahl der Interclusterkanten addiert zu den fehlenden Intraclusterkanten. Daher ergibt sich:

$$|F_\mathcal{C}| = \overline{m}(\mathcal{C}) + \underbrace{\frac{1}{2}\sum_{i=1}^{k} |C_i|(|C_i|-1) - m(\mathcal{C})}_{\text{fehlende Intraclusterkanten}}$$

Somit kann man Performance auch folgendermaßen berechnen:

$$per(\mathcal{C}) = 1 - \frac{2|F_\mathcal{C}|}{n(n-1)}$$

Da die Minimierung von $|F|$ NP-schwer ist, gilt dasselbe auch für die Maximierung von Performance.

2.4.3 Interclusterconductance

Die *Leitfähigkeit* (Conductance) eines Schnittes $\mathcal{C} = \{C, V \setminus C\}$ vergleicht die Anzahl der Kanten im Schnitt mit der Anzahl der Kanten in einem der beiden durch den Schnitt induzierten Subgraphen.

$$\phi(C) := \begin{cases} 1 & \text{für } C \in \{\emptyset, V\} \\ 0 & \text{für } C \notin \{\emptyset, V\} \text{ und } \overline{m}(\mathcal{C}) = 0 \\ \dfrac{\overline{m}(\mathcal{C})}{\min\left(\sum_{v\in C} \deg v, \sum_{v\in V\setminus C} \deg v\right)} & \text{sonst} \end{cases}$$

Ein Schnitt besitzt also eine geringe Leitfähigkeit, wenn die Größe des Schnittes im Verhältnis zur Dichte der beiden durch den Schnitt induzierten Subgraphen klein ist.

Mit Hilfe der Leitfähigkeit lassen sich nun zwei Varianten des Interclusterconductance Index definieren.

Maximale Interclusterconductance Die *maximale Interclusterconductance* $\delta_m(\mathcal{C})$ bewertet eine Clusterung nur nach dem Cluster mit der höchsten Leitfähigkeit.

$$\delta_m(\mathcal{C}) \quad := \quad 1 - \max_{i \in \{1,\dots,k\}} \phi(C_i)$$

Durch die Verwendung des schlechtesten Clusters kann vor allem bei inhomogenen Clusterungen eine intuitiv gut erscheinende Clusterung einen niedrigen Wert haben.

Ferner ist, wenn bereits ein Cluster eine hohe Leitfähigkeit besitzt, die restliche Struktur der Clusterung für diesen Index irrelevant.

Durchschnittliche Interclusterconductance Diese Nachteile lassen sich vermeiden, indem man die *durchschnittliche Interclusterconductance* $\delta_d(\mathcal{C})$ nutzt.

$$\delta_d(\mathcal{C}) \quad := \quad 1 - \sum_{i=1}^{|\mathcal{C}|} \frac{\phi(C_i)}{|\mathcal{C}|}$$

Die Maximierung von Interclusterconductance ist NP-vollständig [BGW03].

2.5 Algorithmen

Als letztes werden noch die genutzten Algorithmen eingeführt. Diese dienen vor allem dem Zweck, in den Experimenten des Kapitels 8 Clusterungen zu vergleichen. Die Algorithmen werden in Generatoren, Clusterer und Clusteroptimierer eingeteilt.

2.5.1 Generatoren

Clustergraphgeneratoren erzeugen einen Graphen mit einer Clusterung. Je nach Wahl der Parameter kann diese Clusterung eine intuitiv sehr gute sein. Die Generatoren werden an dieser Stelle nur kurz eingeführt, eine umfangreichere Diskussion findet sich in [Del04].

Gaußgenerator

Der *Gaußgenerator* erzeugt zunächst eine Partition von n Knoten und verbindet dann Knoten innerhalb eines Clusters mit einer Wahrscheinlichkeit p_{in} und zwei

Knoten, die zu verschiedenen Clustern gehören, mit einer Wahrscheinlichkeit p_{out}. Die Partition stellt eine Gaußverteilung auf den Knoten dar. Die Anzahl c der Cluster ist dabei zufällig zwischen 2 und $\sqrt{n}$. Der Mittelwert der Gaußverteilung wird mit $M = \lfloor n/c \rfloor$, die Standardabweichung s mit $s = \lfloor M/\sqrt{n} \rfloor$ festgelegt. Algorithmus 1 zeigt den Gaußgenerator in Pseudocode.

Algorithmus 1: GaussGenerator(n, p_{in}, p_{out})

```
c ← gleich verteilter Integer ∈ [2, √n] //Clusteranzahl
M ← ⌊n/c⌋ //durchschnittliche Clustergröße
s ← ⌊M/√n⌋ //Standardabweichung der Clustergröße
P ← new Array[c]
for i = 0; i < P.length; i++ do
    d ← normal verteilter Double mit Mittelwert 0 und Standardabweichung 1
    P[i] = (d · V) + M
V ← ∅
E ← ∅
for i=0; i < P.length(); i++ do
    C_i ← ∅
    for j=0; j<P[i]; j++ do
        v ← neuer Knoten
        V ← V ∪ {v}
        C_i ← C_i ∪ {v}
foreach u, v ∈ V do
    if u, v in einem Cluster then
        Füge Kante {u, v} mit Wahrscheinlichkeit p_in zu E
    if u, v in verschiedenen Cluster then
        Füge Kante {u, v} mit Wahrscheinlichkeit p_out zu E
G ← (V, E)
C ← (C_1, ..., C_c)
```

Signifikanz der Clusterung Die Signifikanz der so erzeugten Clusterung hängt stark von der Wahl der Parameter ab. Da die Anzahl der Knotenpaare, die nicht in einem Cluster sind, größer ist als die Anzahl der Knotenpaare, die in einem Cluster liegen, muß p_{in} deutlich größer als p_{out} sein, damit die Clusterung signifikant ist.

Attraktorengenerator

Die Idee bei *Attraktoren* ist, dass man auf einer diskreten Raumaufteilung – in diesem Fall ein zweidimensionales Gitter – einige Knoten als Anziehungspunkte verteilt.

Jeder dieser Knoten repräsentiert einen Cluster. Danach werden weitere Knoten auf dem Gitter verteilt, die dann dem Cluster ihres nächsten Attraktorknoten zugeordnet werden.

Die Konstruktion eines solchen Attraktoren unterteilt sich in 5 Phasen:

1. Erzeuge ein genügend großes zweidimensionales Gitter.
2. Verteile Attraktorenknoten mit gewissem Mindestabstand im Gitter. Jeder Knoten repräsentiert einen eigenen Cluster.
3. Verbinde diese Knoten cliquenartig.
4. An jeder Gitterkoordinate wird mit Wahrscheinlichkeit $1/d$, wobei d der euklidische Abstand zu dem nächsten Attraktoren ist, ein Knoten eingefügt und dieser mit seinem nächsten Attraktoren verbunden und zu dessen Cluster hinzugefügt.
5. Alle Knoten mit einem vorgegebenen maximalen Abstand werden verbunden.

Auf die Angabe des Pseudocodes wird an dieser Stelle verzichtet. Die in dieser Arbeit genutzte Implementierung des Generators benötigt lediglich zwei Parameter:

- Die Knotenzahl n. Soviele Knoten hat der Graph im Erwartungswert.
- Die Dichte f ist der Quotient aus minimalem Abstand der initialen Attraktorenknoten und dem Abstand, die zwei Knoten maximal haben dürfen, um verbunden zu werden.

Die Clusteranzahl wird bei jedem Aufruf zufällig gleichverteilt zwischen 2 und $\sqrt[3]{n}$ gewählt. Die Größe des Gitters wird so gewählt, dass die erwartete Knotenzahl der vorgegebenen entspricht.

Signifikanz der Clusterung Mit zunehmender Dichte nimmt für $f < 1$ die Signifikanz der Clusterung zunächst zu, da für sehr kleine f die Cluster nicht sehr dicht verbunden sind. Für $f > 2$ nimmt bei mit zunehmender Dichte allerdings die Signifikanz der Clusterung ab, da dann die Anzahl der Interclusterkanten stark anwächst.

Hierarchiegraphen

Bei *Hierarchiegraphen* verbindet man rekursiv gleiche Subgraphen. Man gibt dabei zwei Parameter an. Die initiale Cliquengröße c und die Anzahl der Rekursionsschritte $L_{\max}$. Es werden zunächst c Knoten erzeugt und cliquenartig verbunden. Nun wird $L_{\max} - 1$ mal folgendes wiederholt:

- Kopiere den Graph $c-1$ mal.
- Wähle aus den c Zusammenhangskomponenten je einen zufälligen Knoten und verbinde die Knoten cliquenartig.

Somit besitzt ein Hierarchiegraph $n = c^{L_{\max}}$ Knoten. Den Pseudocode des Generators zeigt Algorithmus 2.

Algorithmus 2: HierarchieGraph(c, $L_{\max}$)

```
G ← Clique mit c Knoten
for i = 0; j < L_max; i++ do
    erzeuge c − 1 Kopien von G
    foreach Zusammenhangskomponente k do
        n_k ← zufälliger Knoten aus Zusammenhangskomponente k
    verbinde die Knoten n_1, ..., n_c cliquenartig
```

Eine Besonderheit bei diesem Generator gegenüber dem Attraktoren- und Gaußgenerator ist, dass nicht nur eine Clusterung, sondern $L_{\max}+1$ Clusterungen denkbar sind. Diese Clusterungen werden mit *Clusterlevel* $\mathcal{C}_i$, wobei $0 \leq i \leq L_{\max}$ sein muß, bezeichnet und sind rekursiv definiert:

$$\begin{aligned} \mathcal{C}_0 &:= \mathcal{C}^s \\ \mathcal{C}_i &:= \{C_{i,1}, \dots C_{i,c^{(L_{\max}-1)}}\} \text{ mit } C_{i,j} = C_{i-1,cj} \cup C_{i-1,cj+1} \cup C_{i-1,cj+2} \end{aligned}$$

Dabei sollen die Cluster des Levels $i-1$ so gewählt sein, dass sie cliquenartig verbunden sind. Jeder Clusterlevel $\mathcal{C}_i$ ensteht also aus dem vorherigen Clusterlevel $\mathcal{C}_{i-1}$, indem je c Cluster aus $\mathcal{C}_{i-1}$, die cliquenartig verbunden sind, zu einem Cluster zusammengefasst werden. Somit ist $\mathcal{C}_{L_{\max}}$ die 1-Clusterung.

Durch die Art der Graphgenerierung sind die Clusterlevel Verfeinerungen bzw. Vergröberungen voneinander.

Korollar 7 *Für die verschiedenen Clusterlevel $\mathcal{C}_i$ der Hierarchiegraphen gilt:*

$$\forall i, j, 0 \leq i \leq j \leq L_{\max} : \mathcal{C}_i \subseteq \mathcal{C}_j$$

In den Experimenten (Kapitel 8) werden Hierarchiegraphen mit einer Cliquengröße von $c = 3$ und einem maximalem Clusterlevel $L_{\max} = 8$ genutzt. Dort ist eine Übersicht über die einzelnen Clusterlevel angeben (Tabelle 8.2), anhand derer die Zusammenhänge der Clusterlevel noch einmal verdeutlicht werden sollen.

Signifikanz der Clusterlevel Die Signifikanz der einzelnen Clusterlevel hängt stark von der Anzahl der Clusterlevel $L_{\max}$ und der Cliquengröße c ab. Allerdings sind die Cluster der oberen Clusterlevel nicht sehr dicht, sodass tendenziell eher die kleineren Clusterlevel die intuitiv besten sind.

2.5.2 Clusterer

Clusterer sind solche Algorithmen, die für einen Graphen eine Clusterung berechnen.

Markov Clustering (MCL)

Der *Markov Clustering (MCL)* Algorithmus [Don00a] ist ein Clusterer. Die Idee des Algorithmus´ basiert auf folgender Überlegung: Die Wahrscheinlichkeit, dass ein Random Walk auf einem dichten Cluster diesen verlässt, bevor er viele Knoten dieses Clusters besucht hat, ist gering. Anstatt mehrere Random Walks zu simulieren, werden immer wieder zwei Operationen auf der normalisierten Adjazenzmatrix $M = M(G)$ ausgeführt (Anm.: M entspricht den Random Walks der Länge 1 auf G [Beh00]):

Expansion In der Expansion wird M mit $e \in \mathbb{N}$ potenziert. Dies simuliert e Schritte des Random Walks ausgehend von M.

Inflation In der Inflation wird M renormalisiert, indem jeder Eintrag mir $r \in \mathbb{R}^+$ potenziert wird.

Diese beiden Operationen werden ausgeführt, bis ein Fixpunkt oder ein wiederkehrende Matrix erreicht ist. Mit hoher Wahrscheinlichkeit endet MCL in einem Fixpunkt. Diese Wahrscheinlichkeit kann durch Modifikation von M erhöht werden. In dieser Arbeit wurde für den MCL immer der Expansionsparameter $e = 2$ und Inflationsparameter $r = 2$ gewählt. Algorithmus 3 zeigt MCL mit festen e, r–Parametern in Pseudocode:

Algorithmus 3: MCL$(G = (V, E))$

$M \leftarrow M(G)$
while M *ist kein Fixpunkt* **do**
  $M \leftarrow M^2$ //Expansion
  foreach $u, v \in V$ **do**
    //Inflation
    $M_{uv} \leftarrow M_{uv}^2$
    $M_{uv} \leftarrow \frac{M_{uv}}{\sum_{w \in V} M_{uw}}$
$H \leftarrow$ Graph, der durch Nicht-Null Einträge von M induziert wird
$\mathcal{C} \leftarrow$ Clusterung, die durch die Zusammenhangskomponenten von H induziert wird.

Die Laufzeit wird durch die Potenzierung der Matrix während der Expansion dominiert.

Zufallsclusterer

Der *Zufallsclusterer* berechnet zu einem gegebenen Graphen eine zufällige Clusterung, die dementsprechend nur mit sehr geringer Wahrscheinlichkeit signfikant ist. Als Eingabe erfordert der Algorithmus einen Graphen und eine minimale bzw. maximale Clusteranzahl.

Algorithmus 4: ZufallsClusterer$(G = (V, E), cMin, cMax)$

$c \leftarrow$ gleich verteilter Integer $\in [cMin, cMax]$ //Clusteranzahl
foreach $v \in V$ **do**
  $i \leftarrow$ gleich verteilter Integer $\in [1, c]$
  $C_i \leftarrow C_i \cup \{v\}$ // Knoten v wird Cluster C_i zugeordnet
$\mathcal{C} \leftarrow \{C_1, \ldots, C_c\}$

Man erkennt, dass dieser Algorithmus die Kantenmenge nicht betrachtet, sondern nur auf der Knotenmenge operiert.

2.5.3 Clusteroptimierer

Diese Algorithmen benötigen als Eingabe einen Graphen und eine Clusterung. Die Clusteroptimierer versuchen, die Signifikanz der Cluster zu verbessern bzw. zu verschlechtern. In dieser Arbeit wurden dabei Clusteroptimierer benutzt, die einen Index – oder den Durchschnitt mehrerer – lokal zu maximieren.

Lokale Optimierung

Der *lokale Optimierer* versucht eine Clusterung $\mathcal{C}$ auf einem Graphen G mit Hilfe eines Indexes i zu verbessern. Dabei wird für jeden Knoten überprüft, ob das Verschieben des Knotens in den Cluster eines Nachbarknotens eine Verbesserung des Indexes zur Folge hat. Ferner wird überprüft, ob das Abspalten des Knotens aus dem Cluster den Index verbessert. Die Aktion, die die größte Verbesserung des Indexes zur Folge hat, wird ausgeführt. Der Algorithmus terminiert, sobald es für keinen Knoten mehr eine Aktion gibt, die den Index verbessert. Algorithmus 5 zeigt das Vorgehen in Pseudocode.

Es ist offensichtlich, dass die resultierende Clusterung stark abhängig von der Wahl des Indizes ist. In dieser Arbeit wird als Index meist ein Durchschnitt aus Coverage, Performance und durchschnittlicher Interclusterconductance genutzt.

Man kann den lokalen Optimierer noch erweitern, in dem man eine maximale Anzahl von Knotenverschiebungen angibt. Der Algorithmus terminiert dann nach dieser Anzahl von Verschiebungen oder wenn keine Verschiebung mehr möglich ist.

Algorithmus 5: LokalerOptimierer($G = (V, E), \mathcal{C}, i$)

```
change ← TRUE //Marker, der überprüft, ob noch Knoten bewegt werden
while change == TRUE do
    change ← FALSE //Marker setzen
    foreach v ∈ V in zufälliger Reihenfolge do
        C_z ← ∅ //Zielcluster, bleibt leer, falls v nicht verschoben wird
        maxI ← i(C) // aktueller Index
        foreach u mit {u, v} ∈ E do
            if cl(u) ≠ cl(v) then
                i' ← calcMove(u, v) //Index, falls v in cl(u) verschoben würde
                if i' > maxI then
                    maxI ← i' //erhöhe den Schwellwert
                    C_z ← cl(u) //Merke den Zielcluster
        i' ← calcNewCluster(v) //Index, falls v seperiert würde
        if i' > maxI then
            maxI ← i' //erhöhe den Schwellwert
            C_z ← newCluster() //Zielcluster ist neuer Cluster
        if C_z ≠ ∅ then
            move(C, v, C_z) //verschiebe v in C nach C_z
            change ← TRUE //Knoten wurde bewegt ⇒ Marker setzen
            i(C) ← maxI //update des Indexes
```

Lokale Minimierung

Der *lokale Minimierer* agiert ähnlich wie der lokale Optimierer, unterscheidet sich allerdings in zwei Details:

- Das Abspalten des Knotens ist nicht erlaubt. Dies hat zur Folge, dass die resultierende Clusterung mindestens so viele Cluster wie die Ausgangsclusterung besitzt.
- Es wird die Aktion ausgeführt, die den Index am stärksten verschlechtert.

Das Abspalten von Knoten ist nicht erlaubt, da sich gezeigt hat, dass dies meist die größte Verschlechterung des Indexes zur Folge hat und die so errechnete Clusterung dann der Singleton-Clusterung entspricht.

3 Problemstellung

In diesem Kapitel wird nun die Problemstellung vorgestellt. Desweiteren werden die bisherigen Lösungsansätze und deren Nachteile diskutiert. Zum Abschluß des Kapitels wird der Ansatz, der in dieser Arbeit untersucht wird, für die Lösung des Problems vorgestellt.

3.1 Clusterungsvergleich mittels Abstandsmaßen

Bei dem Problem des Clusterungsvergleichs sind im Allgemeinen zwei Clusterungen $\mathcal{C} = \{C_1, \ldots, C_k\} \in \mathbb{P}(V)$ und $\mathcal{C}' = \{C'_1, \ldots, C'_j\} \in \mathbb{P}(V)$ mit zugehörigen Graphen G und G' gegeben. Es wird nun eine Funktion $d(\mathcal{C}, \mathcal{C}')$ gesucht, die den Abstand zwischen diesen beiden Clusterungen messen soll. Ein Wert von Null soll Gleichheit der beiden Clusterungen bedeuten, ein Wert von Eins maximale Verschiedenheit.

Das Hauptproblem liegt darin, dass ein Hauptkriterium für den Abstand zweier Clusterungen die menschliche Intuition ist. Desweiteren kann von Anwendungsfall zu Anwendungsfall die Intuition variieren. In manchen Fällen kann bereits eine kleine Änderung an der Clusterung einen groß empfundenen Abstand bedeuten.

Auch die Tatsache, dass die geclusterten Graphen sehr unterschiedliche Größen haben können, führt zu Problemen. Was bedeutet in diesem Fall Gleichheit? Gibt es überhaupt Gleichheit bei verschiedenen Größen der Graphen? Darf ein Abstandsmaß bereits den Wert Eins bei gleich großen Graphen annehmen? Wenn dies nicht der Fall ist, wie groß muss allein der Größenunterschied sein, damit das Abstandsmaß einen Wert nahe Eins zurückgibt?

Fernziel Erstrebenswert wäre ein Abstandsmaß, dass den Abstand zweier beliebiger Clusterungen mit beliebigen zugehörigen Graphen der Intuition entsprechend messen kann. Unter anderem für die Problematik der dynamischen Clusterungen wäre solch ein Abstandsmaß sehr hilfreich (Dynamische Clusterungen sind ein sehr umfangreiches Problem aus der Graphentheorie, bei der die den Clusterungen zugrundeliegenden Graphen durch Hinzufügen und Entfernen von Knoten bzw. Kanten verändert werden). Studiert man allerdings die Literatur zum Thema der Abstandsmessung, ist das Ergebnis ernüchternd [WW06]. Bisherige Abstandsmaße nutzen lediglich die Knotenmenge der Graphen und die Kardinalität der Knotenmenge muss bei den zu

vergleichenden Clusterungen gleich sein. Ferner existiert nicht einmal für diese stark vereinfachte Problemstellung eine zufriedenstellende Lösung.

Diese Arbeit hat daher nicht den Anspruch, ein Abstandsmaß für dieses Fernziel zu finden. Vielmehr sollen bisherige Abstandsmaße systematisch analysiert werden, welcher Intuition sie entsprechen und wie sie gegebenenfalls verändert werden können, damit sie eine andere Intuition repräsentieren. Dazu werden einige Vereinfachungen der Problemstellung vorgenommen:

Statischer Clusterungsvergleich Es werden Clusterungen bezüglich des gleichen Graphen betrachtet. Der Vergleich zweier Clusterungen bezüglich des gleichen Graphen soll als *statischer Clusterungsvergleich* bezeichnet werden. Als *dynamischen Clusterungsvergleich* wird dementsprechend der Vergleich von Clusterungen genannt, bei dem die zugehörigen Graphen nicht zwingend gleich sein müssen.

Ungewichtete Graphen Alle betrachteten Graphen sind ungewichtet. Die Beschränkung auf ungewichtete Graphen ist darin begründet, dass bisherige Abstandsmaße lediglich die Knotenmenge betrachten, somit sind die Ergebnisse dieser Maße für gewichtete und ungewichtete Graphen gleich.

Unabhängig davon wird an einigen Stellen dieser Arbeit auf Konsequenzen der hier angestellten Überlegungen für den dynamischen Clusterungsvergleich hingewiesen werden.

3.2 Die ideale Lösung?

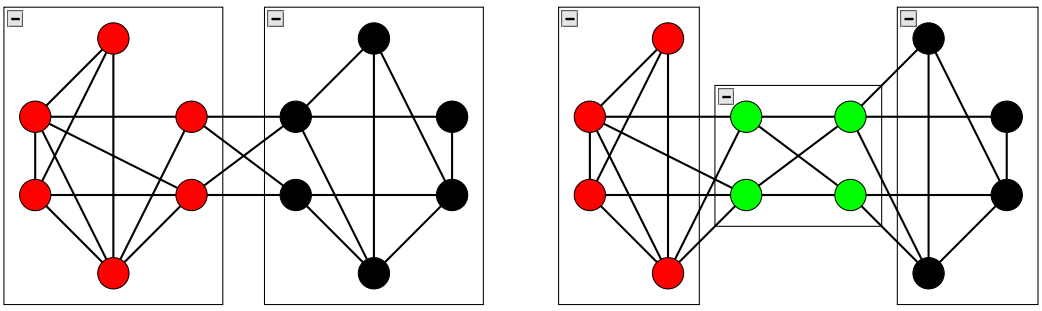
Betrachtet man sich Kapitel 2 noch einmal unter dem Aspekt einer idealen Lösung für den statischen Clusterungsvergleich, könnte man folgenden Ansatz verfolgen:

Die beiden zu vergleichenden Clusterungen besitzen jeweils die entsprechende Clustereditiermenge $F_{\mathcal{C}}$ und $F_{\mathcal{C}'}$. Als Abstand der beiden Clusterungen zueinander könnte man nun das Abstandsmaß

$$d(\mathcal{C}, \mathcal{C}') = \frac{2\big||F_{\mathcal{C}}| - |F_{\mathcal{C}'}|\big|}{n(n-1)}$$

nutzen. Für gleiche Clusterungen liefert dieses Maß den Wert 0, desweiteren gilt $0 \leq d(\mathcal{C}, \mathcal{C}') \leq 1$ für alle denkbaren Clusterungen.

Der Nachteil dieses Ansatzes liegt darin, dass zwei Clusterungen zwar in der Kardinalität der Editiermengen übereinstimmen können, sich strukturell aber stark unterscheiden. Abbildung 3.1 zeigt zwei solche Clusterungen.



Clusterung $\mathcal{C}$ Clusterung $\mathcal{C}'$

Abb. 3.1: Zwei Clusterungen mit gleicher Kardinalität der Editiermengen

Beide Clusterung besitzen eine Clustereditiermenge mit Kardinalität 12. Das Abstandsmaß würde für diese beiden Clusterungen einen Abstand von 0 messen. Dies widerspricht aber der Intuition.

Folgerung Abbildung 3.1 zeigt zwei Clusterungen, die qualitativ gleich, aber strukturell sehr verschieden sind. In denkbaren Anwendungsfällen, in denen der Abstand den Qualitätsunterschied der beiden Clusterungen widerspiegeln soll, möchte man für diese beiden Clusterungen eine Gleichheit ausgewiesen bekommen. Aber es gibt sicherlich Anwendungsfälle, in denen Gleichheit nur bei struktureller Gleichheit der Clusterungen gemessen werden soll.

3.3 Verbandsansatz

Ein anderer Ansatz zur Lösung des Problems ist das Nutzen der Verbandstheorie. Die Grundlagen über Verbände finden sich in [DP02]. Man betrachtet nun die Clusterung als Partition der Knoten und ignoriert die Kantenmenge des Graphen. Bei diesem Ansatz werden zwei Elementaroperationen auf $\mathbb{P}(V)$ definiert:

- Vereinigung zweier Cluster
- Aufteilung eines Clusters

Abbildung 3.2 zeigt alle möglichen Partitionen einer vierelementigen Knotenmenge $V = \{a, b, c, d\}$. Zwei Partitionen sind genau dann verbunden, wenn sie durch eine der beiden Elementaroperationen ineinander überführt werden können.

Solche Diagramme werden in der Literatur *Hassediagramme* genannt. Für die Bestimmung des Abstands zwischen zwei beliebigen Clusterungen sind nun zwei Varianten denkbar:

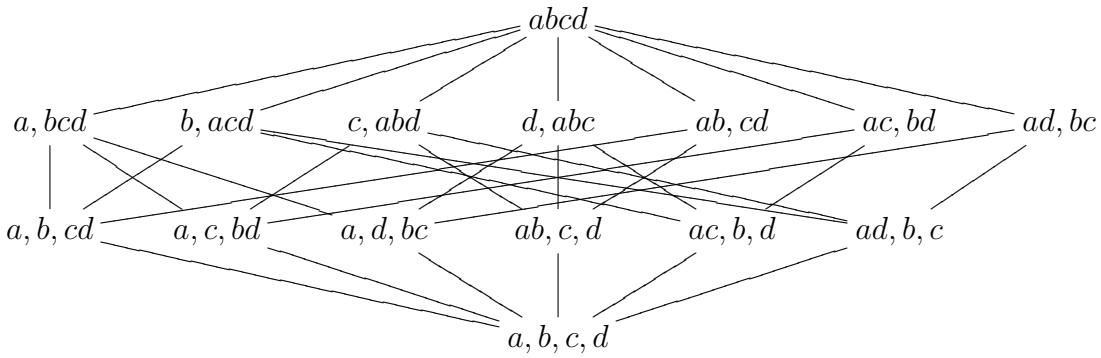


Abb. 3.2: Hassediagramm für eine vierelementige Menge

ungewichtete Elementaroperationen In diesem Falle entspricht jede Kante in Abbildung 3.2 einem Abstand mit Wert α. Nun sei der Abstand D die Anzahl der Kanten des kürzesten Weges zwischen zwei Partitionen $\mathcal{C}$ und $\mathcal{C}'$ in Abbildung 3.2. Es gilt dann $d(\mathcal{C}, \mathcal{C}') = \alpha \cdot D$.

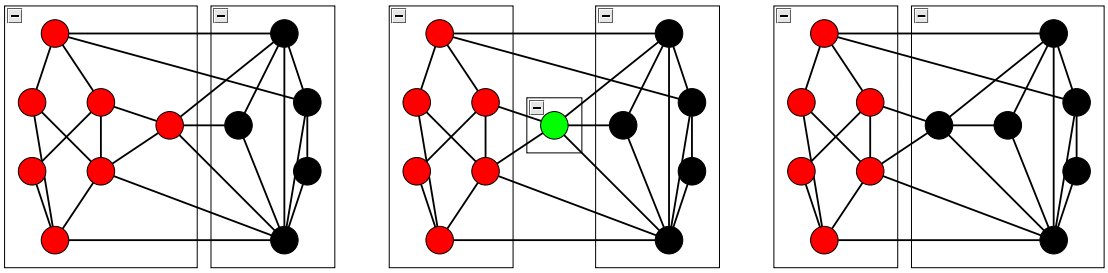
gewichtete Elementaroperationen In diesem Falle wird jeder Kante e in Abbildung 3.2 ein Gewicht $w(e) > 0$ zugeordnet. Dieser Wert soll dem Abstand zwischen den verbundenen Partionen entsprechen. Der Abstand zweier beliebiger Partitionen entspricht dann dem kürzesten Weg in Abbildung 3.2.

Der Vorteil von diesem Ansatz ist, dass Verbände sehr gut untersucht sind und man viele Ergebnisse auf die Clusterung von Graphen übertragen kann. Auch das Definieren zweier Elementaroperationen erscheint sehr schlüssig.

Allerdings erhält man durch das Festlegen auf diese zwei Operationen einen nicht intuitiven Effekt. Das Verschieben eines Knotens v von einem Cluster C_i in einen anderen Cluster C_j führt zu einem gleich großen Abstand wie das Abspalten des Knotens v von C_i und der Vereinigung von v mit C_j. Abbildung 3.3 ist ein Beispiel für ein solches Szenario.

Für ein Abstandsmaß d, dass die Abstände über die Elementaroperationen aufsummiert gilt $d(\mathcal{C}, \mathcal{C}') = d(\mathcal{C}, \mathcal{C} \times \mathcal{C}') + d(\mathcal{C} \times \mathcal{C}', \mathcal{C}')$. Für die Clusterungen in Abbildung 3.3 würde demnach gelten, dass der Abstand von der linken Clusterung zur rechten der Summe der Abstände mit dem Umweg über die mittlere Clusterung entspricht. Dies ist ein recht unintuitives Verhalten.

In Kapitel 6 wird dieser Verbandsansatz erneut aufgegriffen und weitere Nachteile solcher Maße gezeigt.

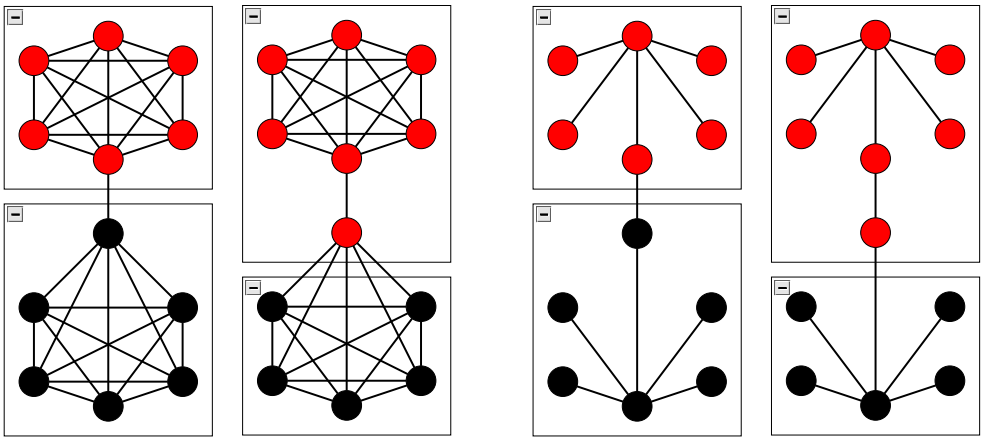


Clusterung $\mathcal{C}$ Clusterung $\mathcal{C} \times \mathcal{C}'$ Clusterung $\mathcal{C}'$

Abb. 3.3: Beispiel für wenig intuitives Verhalten des Verbandsansatzes

3.4 Strukturelle Ansätze

Allen Abstandsmaßen, die in der Literatur zu finden sind [WW06], ist gemein, dass sie lediglich die Partition der Knotenmenge betrachten. Dies scheint auf den ersten Blick unproblematisch, da Clusterungen gerade eine Partition der Knoten sind. Maße, die somit die Struktur der Partition vergleichen, messen dann nur für den Fall $\mathcal{C} = \mathcal{C}'$ einen Abstand von Null. Solche Maße bewerten den Abstand aus Abbildung 3.1 intuitiv richtig. Das Ignorieren der Kantenmenge zur Abstandsbestimmung zweier Clusterungen erscheint bei genauer Betrachtung aber sogar beim statischen Clusterungsvergleich unvorteilhaft. Zur Verdeutlichung zeigt Abbildung 3.4 zwei statische Clusterungsvergleiche.



Clusterung $\mathcal{C}_1$ Clusterung $\mathcal{C}'_1$ Clusterung $\mathcal{C}_2$ Clusterung $\mathcal{C}'_2$

Abb. 3.4: Zwei statische Clusterungsvergleiche auf verschiedenen Graphen

Für ein Abstandsmaß $d(\mathcal{C}, \mathcal{C}')$, das die Kantenmenge des Graphen ignoriert, muss $d(\mathcal{C}_1, \mathcal{C}'_1) = d(\mathcal{C}_2, \mathcal{C}'_2)$ gelten. Doch rein intuitiv ist die Verschiebung des Knoten beim linken Vergleich eine größere Veränderung des Clusterung als beim rechten Vergleich.

Dies liegt daran, dass die Clusterung $\mathcal{C}_1$ qualitativ besser als $\mathcal{C}_2$ und $\mathcal{C}'_1$ qualitativ schlechter als $\mathcal{C}'_2$ ist.

Den gleichen Nachteil haben auch die Maße, die dem Verbandsansatz entsprechen, da diese ebenso die Kantenmenge ignorieren.

3.5 Dreigliedriger Ansatz

Somit scheint eine Abstandsmessung, die sowohl die Knoten- als auch die Kantenmenge der Graphen berücksichtigt, auch für den statischen Clusterungsvergleich sinnvoll. Man kann sagen, dass so eine Abstandsmessung von der Partition der Knoten und der Qualität der Clusterung abhängig ist. Im Rahmen dieser Arbeit soll nun untersucht werden, welche möglichen Lösungsansätze für solch eine Abstandsmessung existieren. Im Hinblick auf den – hier nicht untersuchten – dynamischen Clusterungsvergleich sind diese Lösungsansätze interessant, da bisherige Maße selbst bei dem statischen Clusterungsvergleich gravierende Nachteile besitzen.

Allerdings gibt es sicherlich Anwendungsfälle, in denen der Abstand zweier Clusterungen nur von der Qualität der beiden Clusterungen abhängig sein soll. Ebenso gibt es Fälle, in denen das Ignorieren der Kantenmenge sinnvoll ist.

Vorgehensweise In dieser Arbeit wird ein dreigliedriger Ansatz verfolgt, der zunächst drei Abstandsarten definiert. Daraus können dann verschiedene Aussagen über Gleichheit und minimalen bzw. maximalen Abstand getroffen werden (Kapitel 4). Bezüglich dieser Gliederung der Abstandarten werden daraufhin axiomatische Überlegungen ausgeführt (Kapitel 5 und 6), um anschließend bisherige Maße in die Gliederung einzubinden und zu analysieren bzw. für bestimmte Abstandsarten neue Maße zu definieren (Kapitel 7). Alle Maße werden dann noch experimentell untersucht (Kapitel 8), um ihr Verhalten in bestimmten Testszenarien zu analysieren. Die Ergebnisse dieser Analyse, Schlußfolgerungen für das Messen des Abstands zweier Clusterungen im Allgemeinen (Kapitel 9) sowie ein Ausblick auf sich anschließende Fragestellungen (Kapitel 10) komplettieren diese Analyse.

4 Abstandsarten

In diesem Kapitel wird die Dreigliederung der Abstandsarten vorgstellt. Als Konsequenz ist auch die Gleichheit von zwei Clusterungen nicht mehr eindeutig. Vielmehr existieren je nach Abstandsart drei denkbare Arten der Gleichheit.

Die Dreiteilung hat ebenso Einfluß darauf, wann ein Maß einen minimalen bzw. maximalen Abstand messen soll. Auch diese Problematik ist Thema dieses Kapitels.

4.1 Drei Arten von Abstand

Prinzipiell sind drei Arten von Abstand denkbar. Diese werden zunächst in Worten beschrieben:

Qualitativer Abstand Dieser Abstand ist nur von der Qualität zweier Clusterungen abhängig. Strukturell verschiedene, aber qualitativ gleichwertige Clusterungen besitzen somit keinen qualitativen Abstand.

Knotenstruktureller Abstand Dieser Abstand ist nur abhängig von der Partitionierung der Knoten und ist komplett unabhängig von dem zugrundeliegenden Graphen.

Graphstruktureller Abstand Dieser Abstand ist sowohl von der Partitionierung der Knoten als auch von dem zugrundeliegenden Graphen abhängig.

Die Gegenbeispiele aus Abschnitt 3.3 und 3.4 werfen die Frage auf, warum man zwischen knoten- und graphstrukturellem Abstand unterscheiden sollte. Dies hat mehrere Gründe:

1. Alle in der Literatur zu findenden Abstandsmaße messen den knotenstrukturellen Abstand zweier Clusterungen.
2. Da nur die Partitionierung der Knoten für den knotenstrukturellen Abstand von belang ist, können Überlegungen aus der gut untersuchten Verbandstheorie genutzt werden.
3. Die Graphinstanzen können so groß sein, dass die Berücksichtigung der Kantenmenge die Berechnung des Abstands erheblich vergrößern würde.

4.2 Einteilung der Abstandsmaße

Um die Dreiteilung der Abstandsarten aus Kapitel 4.1 auf Abstandsmaße zu übertragen, sollte zunächst festgestellt werden, welche Informationen einem beim Vergleich von Clusterungen zur Verfügung stehen.

Zur Verfügung stehende Informationen Bei dem Vergleich von Clusterungen kann eine Abstandsfunktion folgende Informationen auswerten:

1. Die Ausgabe der Indizes $i(\mathcal{C})$ und $i(\mathcal{C}')$ der beiden Clusterungen $\mathcal{C}$ und $\mathcal{C}'$.
2. Die beiden Clusterungen $\mathcal{C}$ und $\mathcal{C}'$.
3. Die beiden Graphen G und G'.

In dieser Arbeit soll auf die Problematik der Bewertung von Clusterungen nicht weiter eingegangen werden. Vielmehr wird davon ausgegangen, dass die Bewertung einer Clusterung schon existiert.

Nun kann man die Abstandsmaße entsprechend der Informationen, die sie auswerten, gliedern.

Qualitative Abstandsmaße

Die *qualitativen* Abstandsmaße werten ausschließlich die Ausgaben der Indizes $i(\mathcal{C})$ und $i(\mathcal{C}')$ aus. Sie benötigen weder Informationen über die Graphen noch die Clusterungen. Daher messen diese Abstandsmaße den – entsprechend Kapitel 4.1 – qualitativen Abstand der beiden Clusterungen $\mathcal{C}$ und $\mathcal{C}'$.

Für ein qualitatives Abstandsmaß wird ab nun

$$d_q(\mathcal{C}, \mathcal{C}')$$

geschrieben. Da qualitative Abstandsmaße ausschließlich von den Indizes abhängig sind, ist der Definitionsbereich $\mathbb{D}_q$ hier mit $\mathbb{D}_q := [0; 1]$ festgelegt. Somit ist ein qualitatives Abstandsmaß eine Funktion der Art:

$$\mathbb{D}_q \times \mathbb{D}_q \to [0; 1]$$

Durch die Nichtbetrachtung der Clusterungen $\mathcal{C}$ und $\mathcal{C}'$ und der Graphen G und G' können sämtliche denkbaren qualitativen Abstandsmaße sowohl für den statischen als auch für den dynamischen Clusterungsvergleich genutzt werden. Allerdings stellt sich beim dynamischen Clusterungsvergleich die Frage, welche Aussagekraft ein Vergleich von einem sehr kleinen mit einem sehr großen geclusterten Graphen hat.

Knotenstrukturelle Abstandsmaße

Die *knotenstrukturellen* Abstandsmaße betrachten nur die Clusterungen $\mathcal{C}$ und $\mathcal{C}'$ und ignorieren sowohl die Indizes $i(\mathcal{C})$ und $i(\mathcal{C}')$ als auch die Graphen G und G'. Ein knotenstrukturelles Abstandsmaß soll ab nun mit

$$d_k(\mathcal{C}, \mathcal{C}')$$

bezeichnet werden. Mit dem Definitionsbereich $\mathbb{D}_k := \mathbb{P}(V)$, allen denkbaren Partitionen der Knotenmenge, ist ein knotenstrukturelles Abstandsmaß eine Funktion der Art:

$$\mathbb{D}_k \times \mathbb{D}_k \to [0; 1]$$

Alle in der Literatur befindlichen Abstandsmaße gehören zu dieser Gruppe der Abstandsmaße.

Graphstrukturelle Abstandsmaße

Die *graphstrukturellen* Abstandsmaße werten sowohl die Clusterungen $\mathcal{C}$ und $\mathcal{C}'$ als auch die Graphen G und G' aus. Für diese Maße soll ab jetzt

$$d_g(\mathcal{C}, \mathcal{C}')$$

geschrieben werden. Da bei graphstrukturellen Abstandsmaßen sowohl die Clusterung $\mathcal{C}$, als auch der Graph G ausgewertet wird, ist der Definitionsbereich dieser Maße mit

$$\mathbb{D}_g := \{(\mathcal{C}, G) \mid \mathcal{C} \in \mathbb{P}(V), G \in (V, E)\}$$

definiert. Damit sind graphstrukturelle Abstandsmaße Funktionen der Art:

$$\mathbb{D}_g \times \mathbb{D}_g \to [0; 1]$$

Es sei hierbei angemerkt, dass graphstrukturelle Abstandsmaße auch die Indizes $i(\mathcal{C})$ und $i(\mathcal{C}')$ nutzen können, da sich Indizes aus der Knoten- und Kantenmenge berechnen.

Da graphstrukturelle Abstandsmaße sowohl die Clusterungen, als auch die Graphen auswerten, knotenstrukturelle hingegen nur die Clusterungen, ist auch bei einem statischen Clusterungsvergleich die Berechnungsmächtigkeit der knotenstrukturellen Abstandsmaße eine echt Teilmenge der graphstrukturellen Maße.

Mit dieser Einteilung der Abstandsmaße sind die drei zuvor eingeführten Abstandsarten abgedeckt.

4.3 Konsequenzen der Dreiteilung

Die Dreiteilung der Abstandsarten hat Konsequenzen für die Gleichheit zweier Clusterungen. Ebenso hat die Einteilung einen Einfluß darauf, wann Abstandsmaße einen minimalen bzw. maximalen Abstands messen sollen.

Auf diese drei grundlegenden Eigenschaften wird nun eingegangen, da vor allem minimaler bzw. maximaler Abstand aber stark von der dahinterstehenden Anwendung abhängig ist, werden diese Punkte nur kurz und sehr allgemein dargestellt werden.

4.3.1 Gleichheit

Aus der Aufteilung der Abstandsmaße gemäß Kapitel 4.2 folgen drei Arten der Gleichheit.

Qualitative Gleichheit

Zwei Clusterungen $\mathcal{C}, \mathcal{C}'$ sind genau dann *qualitativ* gleich, wenn ihre Indizes gleich sind. Für qualitative Gleichheit wird ab nun

$$\mathcal{C} =_q \mathcal{C}' \Leftrightarrow i(\mathcal{C}) = i(\mathcal{C}')$$

geschrieben.

Knotenstrukturelle Gleichheit

Zwei Clusterungen $\mathcal{C}, \mathcal{C}'$ sind genau dann *knotenstrukturell* gleich, wenn die Partition der Knoten in beiden Clusterungen gleich sind.

Durch die Art und Weise, wie Clusterung definiert ist, ist knotenstrukturelle Gleichheit die mathematisch sauberste Form der Definition der Gleichheit. Zur besseren Unterscheidung von qualitativer Gleichheit wird ab nun für knotenstrukturelle Gleichheit

$$\mathcal{C} =_k \mathcal{C}' \Leftrightarrow \mathcal{C} = \mathcal{C}'$$

geschrieben werden.

Graphstrukturelle Gleichheit

Zwei Clusterungen $\mathcal{C}, \mathcal{C}'$ sind genau dann *graphstrukturell* gleich, wenn die Partition der Knoten in beiden Clusterungen und die beiden Graphen gleich sind. Formal:

$$\mathcal{C} =_g \mathcal{C}' \Leftrightarrow (\mathcal{C}, G) = (\mathcal{C}', G')$$

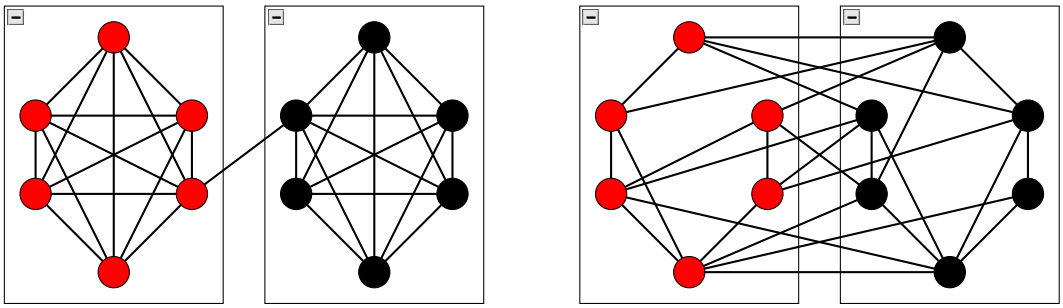
Zusammenhänge

Zwischen diesen drei möglichen Arten, wie Clusterungen zueinander gleich sein können, existieren einige Zusammenhänge. Die nun folgenden Lemmata behandeln diese Zusammenhänge.

Lemma 7 *Da graphstrukturelle Gleichheit zusätzlich zur Gleichheit der Partitionen die Gleichheit der Graphen fordert, gilt:*

$$\mathcal{C} =_g \mathcal{C}' \Rightarrow \mathcal{C} =_k \mathcal{C}'$$

Die Umkehrung gilt bei dynamischen Clustervergleich im Allgemeinen nicht, wie Abbildung 4.1 zeigt. Die beiden Clusterungen sind knotenstrukturell gleich, graphstrukturell allerdings nicht.



Clusterung $\mathcal{C}$ Clusterung $\mathcal{C}'$

Abb. 4.1: Zwei knoten- aber nicht graphstrukturell gleiche Clusterungen

Im statischen Falle gilt hingegen die Umkehrung, da statischer Clusterungsvergleich gerade die Gleichheit der Graphen voraussetzt.

Lemma 8 *Bei statischem Clusterungsvergleich gilt:*

$$\mathcal{C} =_g \mathcal{C}' \Leftrightarrow \mathcal{C} =_k \mathcal{C}'$$

Ebenso folgt aus graphstruktureller Gleichheit auch qualitative, da Indizes ihren Wert aus der Clusterung und dem Graphen berechnen.

Lemma 9 *Es gilt:*

$$\mathcal{C} =_g \mathcal{C}' \Rightarrow \mathcal{C} =_q \mathcal{C}'$$

Die Umkehrung gilt im Allgemeinen nicht, da strukturell verschiedene Clusterungen gleich gut bewertet werden können. Deutlich wird dies bei dem – zugegebenermaßen sehr schlechten – Index, der jeder Clusterung $\mathcal{C}$ den Wert $i(\mathcal{C}) = 1$ zuordnet.

Aus knotenstruktureller Gleichheit folgt im Allgemeinen nicht qualitative Gleichheit, da die zugrundeliegenden Graphen bei knotenstruktureller Gleichheit sehr verschieden sein können und somit die Qualität der Clusterung nicht gleich sein muss.

Lemma 10 *Bei statischem Clusterungsvergleich gilt:*

$$\mathcal{C} =_k \mathcal{C}' \Rightarrow \mathcal{C} =_q \mathcal{C}'$$

Dies folgt aus Lemmata 8 und 9.

Abbildung 4.2 fasst die vorangegangenen Lemmata zusammen, ein „S“ an einem Implikationspfeil bedeutet dabei, dass diese Folgerung nur bei statischem Clusterungsvergleich gilt.

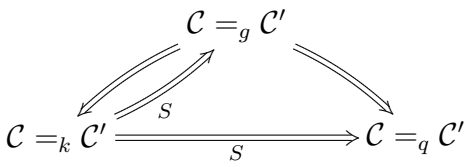


Abb. 4.2: Zusammenhänge der Gleichheiten

4.3.2 Minimaler Abstand

In diesem Abschnitt wird thematisiert, wann zwei Clusterungen einen minimalen Abstand haben sollen. Durch die Dreiteilung der Abstandsarten wird der minimale Abstand für jede Abstandsart getrennt betrachtet.

Minimaler qualitativer Abstand

Da für die qualitative Abstandsmessung ein Index benutzt wird, wird die Problematik an den Index transferiert. Einen minimalen Abstand zweier Clusterungen erhält man, wenn die Bewertungen der Clusterungen sehr ähnlich, aber nicht gleich sind.

Minimaler knotenstruktureller Abstand

Je nach Intuition kann man bei knotenstrukturellem Abstand Verschiedenes fordern, zum Beispiel, dass das Verschieben eines einzelnen Knoten in einer gleichmäßigen Clusterung ein minimaler Abstand sein soll. Auch das Aufteilen eines Clusters könnte als minimal bezeichnet werden.

Solche Elementaroperationen als minimalen Abstand auf der Menge $\mathbb{P}(V)$ verfolgt zum Beispiel der Verbandsansatz. Die Nachteile dieses Ansatzes werden unter anderem Thema von Kapitel 6 sein.

Minimaler graphstruktureller Abstand

Auch hier ist die Frage sicherlich nicht eindeutig zu klären. Aber ein minimaler Abstand ist dann sinnvoll, wenn man eine geringe Änderung an der Struktur der Clusterung vornimmt, die die Qualität der Clusterung kaum ändert.

Auch für den graphstrukturellen Abstand wäre es wünschenswert, dass einige Elementaroperationen auf den Clusterungen existieren, die jede für sich genommen einen minimalen Abstand induziert.

4.3.3 Maximaler Abstand

Zum Abschluss dieses Kapitels soll nun noch diskutiert werden, wann zwei Clusterungen den maximalen Abstand von Eins haben sollen. Auch der maximale Abstand wird entsprechend der Dreiteilung der Abstände getrennt voneinander besprochen.

Maximaler qualitativer Abstand

Beim qualitativen Abstand wird wie beim minimalem Abstand die Problematik des maximalen Abstands an den Index übergeben. Eine qualitative Abstandsfunktion soll genau dann den Wert Eins liefern, wenn eine Clusterung als optimal und die andere als schlecht bewertet wird.

$$d_q(\mathcal{C}, \mathcal{C}') = 1 \Leftrightarrow |i(\mathcal{C}) - i(\mathcal{C}')| = 1$$

Es ist also dem Index überlassen, zwischen welchen Clusterungen ein maximaler Abstand gemessen wird.

Maximaler knotenstruktureller Abstand

Beim knotenstrukturellem Abstand ist die Frage des maximalen Abstands nicht eindeutig zu klären, da dies stark von der Intuition und dem jeweiligen Anwendungsfall abhängig ist. Im statischen Fall sind drei Betrachtungsweisen sinnvoll:

- Zwei Clusterungen sollen dann den maximalen knotenstrukturellen Abstand haben, wenn eine Clusterung die Singleton- und die andere die 1-Clusterung ist. Das bedeutet:
$$\mathcal{C} = \mathcal{C}^s \wedge \mathcal{C}' = \mathcal{C}^1 \Rightarrow d_k(\mathcal{C}, \mathcal{C}') = 1$$

- Die beiden komplementären Clusterungen $\mathcal{C}^\times$ und $\mathcal{C}^\perp$ sollen maximalen knotenstrukturellen Abstand voneinander haben:

$$\mathcal{C} = \mathcal{C}^\times \wedge \mathcal{C}' = \mathcal{C}^\perp \Rightarrow d_k(\mathcal{C}, \mathcal{C}') = 1$$

- Zwei zufällige unabhängige Clusterungen sollen im Erwartungswert einen maximalen Abstand voneinander haben.

Natürlich sind diese drei Betrachtungsweisen miteinander kombinierbar, sodass ein mögliches Abstandsmaß auch in allen drei Fällen einen maximalen Abstand messen könnte.

Bei dynamischem Clusterungsvergleich ist die Definition eines maximalen Abstands noch problematischer, da mit zunehmender Abweichung in der Kardinalität der beiden Knotenmengen ein immer größerer Abstand gemessen werden sollte. Die Größe von Graphen ist nach oben theoretisch allerdings nicht begrenzt.

Maximaler graphstruktureller Abstand

Bei graphstrukturellem Abstand ist es vom Anwendungsfall abhängig, wann zwei Clusterungen einen maximalen Abstand haben sollen. Aber für den Fall, dass die Knotenmenge in beiden Graphen gleich ist, erscheint folgende Überlegung für einen maximalen Abstand sinnvoll:

Sei G ein Clustergraph mit der Clusterung $\mathcal{C}$, der Graph $G = (V, E)$ ist also sehr signifikant geclustert. Dann ist der Komplementgraph $\overline{G} = (V, \binom{V}{2} \setminus E)$ mit der gleichen Clusterung $\mathcal{C}$ intuitiv sehr schlecht geclustert. Wenn man nun noch fordert, dass es möglichst viele Intraclusterkanten in G und Interclusterkanten in $\overline{G}$ geben soll, ergibt das folgendes Resultat:

$$\mathcal{C} = \mathcal{C}^1, G = K_n \wedge \mathcal{C}' = \mathcal{C}^1, G' = (V, \emptyset) \Rightarrow d_g(\mathcal{C}, \mathcal{C}') = 1$$

oder

$$\mathcal{C} = \mathcal{C}^s, G = (V, \emptyset) \wedge \mathcal{C}' = \mathcal{C}^s, G' = K_n \Rightarrow d_g(\mathcal{C}, \mathcal{C}') = 1$$

Viel komplizierter gestaltet es sich, wenn sich die zugrundeliegenden Graphen in der Knotenanzahl stark unterscheiden. Dann gibt es sicherlich Fälle, in denen ein maximaler Abstand den Abstand zwischen einem wenig signifikant geclusterten kleinen Graphen und einem signifikant geclusterten großen Graphen bedeuten soll.

5 Axiome für Abstandsmaße

In diesem Kapitel werden Axiome für Abstandsmaße aufgestellt. Dabei werden zunächst solche Axiome festgelegt, die unabhängig von der betrachteten Abstandsart sind, im Anschluss daran werden abstandsartabhängige Axiome definiert. In diesem Kapitel werden die Axiome lediglich vorgestellt, eine Diskussion über – nicht sofort ersichtliche – Nachteile einiger Axiome folgt in Kapitel 6.

5.1 Abstandsartinvariante Axiome

Zunächst kann man sich überlegen, dass die verwendeten Maße eine *Metrik* sein sollten. Da aber beispielsweise die Gleichheit von Clusterungen von der Art des zu messenden Abstands abhängig ist, bedeutet eine Metrik für jede Abstandsart etwas Verschiedenes. Die vier Eigenschaften einer Metrik lauten:

1. Eine wichtige Forderung an jedes Abstandsmaß ist die Symmetrieeigenschaft.

   Axiom 1 (Symmetrie) *Ein beliebiges Abstandsmaß d ist* symmetrisch*, wenn gilt:*

   $$\forall \mathcal{C}, \mathcal{C}' : d(\mathcal{C}, \mathcal{C}') = d(\mathcal{C}', \mathcal{C})$$

   Die Verwendung eines asymmetrischen Abstandsmaßes ist nur dann sinnvoll, wenn man eine gegebene optimale Clusterung $\mathcal{C}_{\text{OPT}}$ und eine anders berechnete Clusterung $\mathcal{C}_{\text{CALC}}$ hat, und man die Abweichung der berechneten von der optimalen bestimmen möchte. Bei Graphclusterungen ist allerdings die optimale Clusterung des Graphen häufig nicht bekannt.

2. Jedes Abstandsmaß sollte nur dann einen Abstand von Null messen, wenn die beiden Clusterungen gleich sind. Wie in Kapitel 4 gezeigt wurde, folgt aus der Dreiteilung der Abstandsarten auch drei Arten des Gleichheit.

   Axiom 2 (Identität) *Ein beliebiges Abstandsmaß d wahrt die* Identität*, wenn*

   $$d_a(\mathcal{C}, \mathcal{C}') = 0 \Leftrightarrow \mathcal{C} =_a \mathcal{C}'$$

   mit $a \in \{q, k, g\}$ gilt.

3. Der Abstand zweier Clusterungen sollte unabhängig von der betrachteten Abstandsart immer positiv sein.

   Axiom 3 (Positivität) *Ein beliebiges Abstandsmaß d ist* positiv*, wenn gilt:*

   $$\forall \mathcal{C}, \mathcal{C}' : d(\mathcal{C}, \mathcal{C}') \geq 0$$

4. **Axiom 4 (Dreiecksungleichung)** *Ein beliebiges Abstandsmaß d erfüllt die* Dreiecksungleichung*, wenn gilt:*

   $$\forall \mathcal{C}, \mathcal{C}', \mathcal{C}'' : d(\mathcal{C}, \mathcal{C}'') \leq d(\mathcal{C}, \mathcal{C}') + d(\mathcal{C}', \mathcal{C}'')$$

   An dieser Stelle sei noch einmal darauf hingewiesen, dass die Dreiecksungleichung für alle drei Abstandsarten verschiedene Konsequenzen hat.

Die Metrikeigenschaften beinhalten durch die Positivität nur eine Beschränkung nach unten. Da ein maximaler Abstand mit dem Wert Eins gemessen werden soll, sollten Abstandsmaße nach oben mit Eins beschränkt sein.

Axiom 5 (1–Beschränktheit) *Ein beliebiges Abstandsmaß d heißt* 1–beschränkt*, wenn gilt:*

$$\forall \mathcal{C}, \mathcal{C}' : d(\mathcal{C}, \mathcal{C}') \leq 1$$

Abstandsmaße, die nicht 1-beschränkt sind, können durch Normierung des Maßes in 1–beschränkte Abstandsmaße transferiert werden können. Durch die Normierung kann es allerdings vorkommen, dass das Maß Werte von Eins nicht mehr annimmt.

Axiom 6 (1–Maximalität) *Ein 1-beschränktes Abstandsmaß d heißt* 1–maximal*, wenn gilt:*

$$\exists \mathcal{C}, \mathcal{C}' : d(\mathcal{C}, \mathcal{C}') = 1$$

Eine Abschwächung dieses Axioms ist die grenzwertige 1-Maximalität. Hierbei wird gefordert, dass mit steigender Knotenzahl der gemessene Abstand zweier Clusterungen der Eins annähert.

Axiom 7 (Grenzwertige 1–Maximalität) *Ein 1-beschränktes Abstandsmaß d ist* grenzwertig 1–maximal*, wenn gilt:*

$$\exists \mathcal{C}, \mathcal{C}' : \lim_{n \to \infty} d(\mathcal{C}, \mathcal{C}') = 1$$

Manche Maße messen einen kleinen Abstand von zwei Zufallsclusterungen, wenn die Clusteranzahl in beiden Clusterungen sehr hoch ist, obwohl die Clusterungen unabhängig voneinander sind.

Axiom 8 (Clusteranzahlunabhängigkeit) *Ein beliebiges Abstandsmaß d heißt* clusteranzahlunabhängig*, wenn für zwei unabhängige Zufallsclusterungen die Anzahl der Cluster keine Auswirkung auf den Erwartungswert des gemessenen Abstands hat.*

Prinzipiell sollten Abstandsmaße alle Informationen nutzen, die ihnen zur Verfügung stehen. Unterschiede, die sich in den nicht genutzten Informationen befinden, führen sonst dazu, dass das Maß keinen Abstand feststellen kann.

Axiom 9 (Informationsvollständigkeit) *Ein beliebiges Abstandsmaß d ist* informationsvollständig*, wenn es alle ihm zur Verfügung stehenden Informationen vollständig nutzt.*

Zum Abschluss der abstandsartunabhängigen Axiome ist es natürlich wichtig, dass die Maße auch für große Instanzen von Graphen und Clusterungen berechnet werden können.

Axiom 10 (polynomielle Berechenbarkeit) *Ein Abstandsmaß d ist* polynomiell berechenbar*, wenn die Berechnung von d in $O(p(n))$, wobei p ein beliebiges Polynom sei, möglich ist.*

Für theoretische Überlegungen mag diese Eigenschaft von geringerer Bedeutung sein. Da in der Praxis, wie in der Einleitung beschrieben, die zu betrachtenden geclusterten Graphen sehr groß sein können, ist eine polynomielle Berechenbarkeit des Abstandsmaßes essentiell.

5.2 Abstandsartabhängige Axiome

Die Axiome, die abhängig von der gemessenen Abstandsart sind, werden nun vorgestellt. Dabei wird die eingeführte Dreigliederung der Abstandsarten beibehalten.

5.2.1 Qualitative Axiome

In diesem Abschnitt werden Axiome vorgestellt, die nur für qualitative Abstandsmessungen sinnvoll sind. Dabei werden hauptsächlich Zusammenhänge zwischen Abweichung der beiden Indizes $i(\mathcal{C})$ und $i(\mathcal{C}')$ und dem qualitativen Abstand $d_q(\mathcal{C}, \mathcal{C}')$ hergestellt. In den folgenden Axiomen ist i jeweils ein Index und d_q ein qualitatives Abstandsmaß.

Axiom 11 (Wachstumsfaktor) *Seien $\mathcal{C}, \mathcal{C}', \mathcal{C}''$ drei Clusterungen. Wenn*

$$i(\mathcal{C}') = x \cdot i(\mathcal{C}'') \Leftrightarrow d_q(\mathcal{C}, \mathcal{C}') = f(x) \cdot d_q(\mathcal{C}, \mathcal{C}'')$$

mit $f : \mathbb{R}^+ \mapsto \mathbb{R}^+$ *gilt, ist* d_q f-wachsend *und* f *der* Wachstumsfaktor *bezüglich des Indexes* i*. Gilt* $f(x) = x$*, ist* d_q linear *wachsend. Wenn* $f(x) = x^c, c \in \mathbb{R}^+, c > 1$ *gilt, ist* d_q exponentiell *wachsend.*

Damit kleine Abweichungen beispielsweise nur in den signifikanten Bereichen der Clusterungen zu größeren qualitativen Abständen führen, kann man den Begriff der *Lastigkeit* eines qualitativen Abstandsmaßes einführen.

Axiom 12 (Einseitige Lastigkeit) *Seien* $\mathcal{C}_1, \mathcal{C}'_1, \mathcal{C}_2, \mathcal{C}'_2$ *vier Clusterungen. Ferner sei* $i(\mathcal{C}_1) > i(\mathcal{C}_2)$*,* $i(\mathcal{C}'_1) > i(\mathcal{C}'_2)$ *und* $|i(\mathcal{C}_1) - i(\mathcal{C}'_1)| = |i(\mathcal{C}_2) - i(\mathcal{C}'_2)|$*. Folgt daraus für* d_q*, dass*

$$d_q(\mathcal{C}_1, \mathcal{C}'_1) > (<)\ d_q(\mathcal{C}_2, \mathcal{C}'_2)$$

gilt, wird das Abstandsmaß als gutlastig (schlechtlastig) *bezüglich des Indexes* i *bezeichnet.*

Gut- und Schlechtlastigkeit lassen sich auch kombinieren. Ein solches Maß reagiert in hohen und niedrigen Indexbereichen sensibel, im mittleren Bereich weniger sensibel.

Axiom 13 (Doppelte Lastigkeit) *Seien* $\mathcal{C}_1, \mathcal{C}'_1, \mathcal{C}_2, \mathcal{C}'_2$ *vier Clusterungen. Ferner sei* $i(\mathcal{C}_1) > i(\mathcal{C}_2)$*,* $i(\mathcal{C}'_1) > i(\mathcal{C}'_2)$ *und* $|i(\mathcal{C}_1) - i(\mathcal{C}'_1)| = |i(\mathcal{C}_2) - i(\mathcal{C}'_2)|$*. Gibt es ein* $a \in [0; 2]$*, sodass*

$$\begin{aligned} i(\mathcal{C}_2) + i(\mathcal{C}'_2) > a \quad &\Rightarrow \quad d_q(\mathcal{C}_1, \mathcal{C}'_1) > d_q(\mathcal{C}_2, \mathcal{C}'_2) \\ i(\mathcal{C}_1) + i(\mathcal{C}'_1) < a \quad &\Rightarrow \quad d_q(\mathcal{C}_1, \mathcal{C}'_1) < d_q(\mathcal{C}_2, \mathcal{C}'_2) \end{aligned}$$

gilt, wird das Abstandsmaß als beidlastig *bezüglich des Indexes* i *und* a *als* Umschlagspunkt *bezeichnet.*

Der Umschlagspunkt a gibt dabei an, für welchen Bereich das Maß schlecht- bzw. gutlastig ist. Beispiele für lastige Abstandsmaße finden sich in Kapitel 7.1.

5.2.2 Knotenstrukturelle Axiome

Die knotenstrukturellen Axiome sind teilweise der Literatur entnommen und sind zum Beispiel in [Mei05] zu finden.

Knotenstrukturelle Abstandsmaße können den Abstand zwischen zwei Partitionen messen, indem sie die Anzahl an Elementaroperationen zählen, die die beiden zu vergleichenden Clusterungen ineinander überführen.

Axiom 14 (Elementare Äquidistanz) *Sei δ der Abstand für eine oder mehrere Elementaroperationen. Ferner sei $\alpha(\mathcal{C}, \mathcal{C}')$ die minimale Anzahl, um die Clusterung $\mathcal{C}$ in die Clusterung $\mathcal{C}'$ zu überführen. Ein knotenstrukturelles Abstandsmaß d_k ist* elementar äquidistant*, wenn gilt:*

$$\forall \mathcal{C}, \mathcal{C}' : d_k(\mathcal{C}, \mathcal{C}') = \delta \cdot \alpha(\mathcal{C}, \mathcal{C}')$$

Handelt es sich hierbei um die beiden Operationen Aufteilen und Vereinigen von Clustern, entspricht ein solches elementar äquidistantes Maß dem ungewichteten Verbandsansatz aus Kapitel 3.3.

Aus der Verbandstheorie sind auch die folgenden vier Axiome motiviert. Bei Betrachtung des Hassediagramms (Abbildung 3.2 aus Kapitel 3.3) kann man sich zunächst überlegen, dass eine Additivität für Verfeinerungen von Clusterungen sinnvoll sein kann. Im Hassedigramm bedeutet dies informell, dass wenn man $\mathcal{C}'$ von $\mathcal{C}$ aus nur durch Absteigen im Hassediagramm erreichen kann, die Wahl des Weges dabei irrelevant ist.

Axiom 15 (Additivität bzgl. der Verfeinerung) *Ein knotenstrukturelles Abstandsmaß d_k ist* additiv bzgl. der Verfeinerung*, wenn*

$$\forall \mathcal{C}, \mathcal{C}', \mathcal{C}'' : d_k(\mathcal{C}, \mathcal{C}'') = d_k(\mathcal{C}, \mathcal{C}') + d_k(\mathcal{C}', \mathcal{C}'')$$

mit $\mathcal{C}'' \subseteq \mathcal{C}' \subseteq \mathcal{C}$ gilt.

Da bei dem Verbandsansatz die Abstände zwischen zwei Clusterungen als Summe der einzelnen Abstände entsprechend dem Hassediagramm berechnet werden, kann es Anwendungsfälle geben, in denen dieser Abstand additiv für jede kleinste gemeinsame Vereinigung der beiden Clusterungen ist.

Axiom 16 (Additivität bzgl. der Vereinigung) *Ein knotenstrukturelles Abstandsmaß d_k ist* additiv bzgl. der Vereinigung*, wenn gilt:*

$$\forall \mathcal{C}, \mathcal{C}' : d_k(\mathcal{C}, \mathcal{C}') = d_k(\mathcal{C}, \mathcal{C} \oplus \mathcal{C}') + d_k(\mathcal{C} \oplus \mathcal{C}', \mathcal{C}')$$

Informell steigt man im Hassedigramm von $\mathcal{C}$ soweit aufwärts wie nötig und steigt dann zu $\mathcal{C}'$ ab. Dabei summiert man die Abstände der Einzelabstände.

Analog dazu kann man auch zuerst absteigen und dann aufsteigen. Dies ergibt die Additivität des Produktes.

Axiom 17 (Additivität bzgl. des Produktes) *Ein knotenstrukturelles Abstandsmaß d_k ist* additiv bzgl. des Produktes*, wenn gilt:*

$$\forall \mathcal{C}, \mathcal{C}' : d_k(\mathcal{C}, \mathcal{C}') = d_k(\mathcal{C}, \mathcal{C} \times \mathcal{C}') + d_k(\mathcal{C} \times \mathcal{C}', \mathcal{C}')$$

Eine weitere Additivität ist bei Betrachtung des Hassedigramms definierbar.

Axiom 18 (Konvexe Additivität) *Sei $\mathcal{C} = \{C_1, \ldots, C_p\}$ und $\mathcal{C}', \mathcal{C}''$ zwei Verfeinerungen von $\mathcal{C}$. $\mathcal{C}'_k, \mathcal{C}''_k$ seien die Partionen, die durch $\mathcal{C}', \mathcal{C}''$ von C_k induziert werden. Ein knotenstrukturelles Abstandsmaß d_k ist* konvex additiv*, wenn gilt:*

$$d_k(\mathcal{C}', \mathcal{C}'') = \sum_{k=1}^{p} \frac{|C_k|}{n} d_k(\mathcal{C}'_k, \mathcal{C}''_k)$$

In Worten bedeutet diese Additivität, dass bei Veränderungen von einzelnen Clustern der Abstand unabhängig von der restlichen Clusterung sein soll.

Ein Maß, das alle Axiome des Verbandsansatzes erfüllt, soll *Verbandsmaß* genannt werden. Es sind allerdings auch Maße konstruierbar, die nur einzelne Axiome der Verbandstheorie erfüllen.

5.2.3 Graphstrukturelle Axiome

Diese hier aufgeführten graphstrukturellen Axiome sollen lediglich als Anhaltspunkt für künftige Untersuchungen dienen. Da das Thema dieser Arbeit die Analyse bisheriger und nicht das Design neuer Maße ist, ist der Abschnitt recht kurz gehalten.

Graphstrukturelle Abstandsmaße können die Eigenschaft haben, dass die gleiche Operation auf der Knotenmenge verschiedene Auswirkungen auf den Abstand hat. Dies hängt von der Signifikanz der Clusterungen bezüglich ihrer Graphen sind.

Axiom 19 (Qualitative Sensivität) *Seien $\mathcal{C}_1$ und $\mathcal{C}'_1$ zwei Clusterungen auf einem Graphen G_1 und $\mathcal{C}_2$ und $\mathcal{C}'_2$ zwei Clusterungen auf einem Graphen G_2. Ferner sei d_g ein graphstrukturelles Abstandsmaß und i ein Index. Desweiteren gelte $\mathcal{C}_1 =_k \mathcal{C}_2$ und $\mathcal{C}'_1 =_k \mathcal{C}'_2$. Wenn*

$$|i(\mathcal{C}_1) - i(\mathcal{C}'_1)| > |i(\mathcal{C}_2) - i(\mathcal{C}'_2)| \Rightarrow d_g(\mathcal{C}_1, \mathcal{C}'_1) > d_g(\mathcal{C}_2, \mathcal{C}'_2)$$

gilt, wird d_g als qualitativ sensitiv *bezüglich des Indexes i bezeichnet.*

Nun kann man sich überlegen, dass in manchen Fällen ein Maß das Verschieben von Knoten mit hohem Knotengrad stärker bewertet als solche mit niedrigem Knotengrad.

Axiom 20 (Knotengradabhängigkeit) *Seien $\mathcal{C}, \mathcal{C}', \mathcal{C}''$ drei Clusterungen und d_g ein graphstrukturelles Abstandsmaß. Clusterung $\mathcal{C}'$ ($\mathcal{C}''$) entsteht aus $\mathcal{C}$, indem ein Knoten v (u) von einem Cluster in einen anderen verschoben wird. Wenn*

$$\deg(v) > \deg(u) \Rightarrow d_g(\mathcal{C}, \mathcal{C}') > d_g(\mathcal{C}, \mathcal{C}'')$$

gilt, wird das Maß als knotengradabhängig *bezeichnet.*

Dies ist dahingehend veränderbar, dass nicht der Gesamtknotengrad des Knotens v relevant ist, sondern lediglich der auf den Cluster $cl(v)$ eingeschränkte Knotengrad

$$\deg_{cl(v)}(c) = \left|\left\{\{u, v\} \in E \mid cl(u) = cl(v)\right\}\right|$$

des Knotens v.

Axiom 21 (Intraclusterknotengradabhängigkeit) *Seien $\mathcal{C}, \mathcal{C}', \mathcal{C}''$ drei Clusterungen und d_g ein graphstrukturelles Abstandsmaß. Clusterung $\mathcal{C}'$ ($\mathcal{C}''$) entsteht aus $\mathcal{C}$, indem ein Knoten v (u) von einem Cluster in einen anderen verschoben wird. Wenn*

$$\deg_{cl(v)}(v) > \deg_{cl(u)}(u) \Rightarrow d_g(\mathcal{C}, \mathcal{C}') > d_g(\mathcal{C}, \mathcal{C}'')$$

gilt, wird das Maß als intraclusterknotengradabhängig *bezeichnet.*

Im Hinblick auf den dynamischen Clusterungsvergleich kann auch das Übertragen der Idee der Elementaroperationen des Axioms 14 ein interessanter Ansatz sein. Dazu legt man nicht nur Elementaroperationen auf der Menge aller Clusterungen, sondern auch auf den Graphen fest. Elementaroperationen auf einem Graphen könnte dabei das Entfernen bzw. Hinzufügen von Kanten oder Knoten sein.

Axiom 22 (Graphstrukturelle Elementaroperationen) *Seien $\mathcal{C}$ und $\mathcal{C}'$ zwei Clusterungen auf den Graphen G bzw. G'. Ferner seien einige Elementaroperationen auf den Clusterungen und Graphen definiert. Gilt*

$$d_g(\mathcal{C}, \mathcal{C}') = \sum_e d(e)$$

wobei e Elementaroperation und $d(e)$ Abstand dieser Elementaroperation sei, heißt das Maß graphstukturell elementar.

Es sei angemerkt, dass man Abstandsmaße, die Axiom 14 oder 22 erfüllen, als Editierprobleme auf Clusterungen ansehen kann.

6 Diskussion der Axiome

In diesem Kapitel sollen nun einzelne Axiome diskutiert werden. Die Motivation für diese Diskussion liegt darin, dass jedes in Kapitel 5 vorgestellte Axiom auf den ersten Blick sinnvoll erscheint, bei genauerer Betrachtung besitzen manche Axiome allerdings weniger intuitive Effekte. Daher zeigt dieses Kapitel einige Beispiel, bei denen Abstandsmaße, die bestimmte Axiome erfüllen, ein wenig intuitives Verhalten aufweisen.

6.1 Elementare Äquidistanz

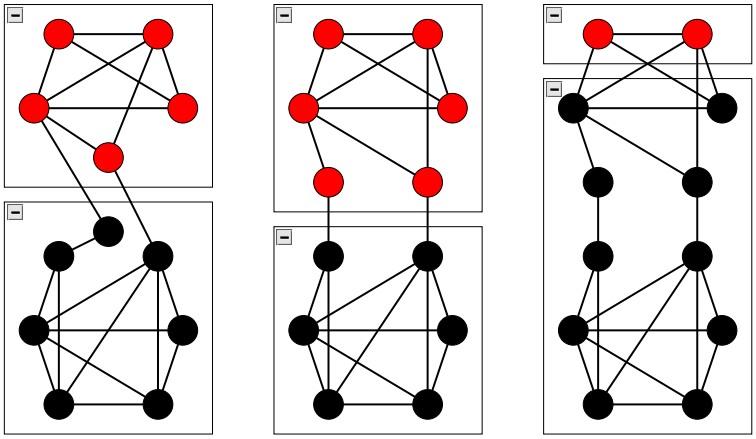
Axiom 14 sagt aus, dass man eine oder mehrere Elementaroperationen mit festem Abstand δ auf Clusterungen definieren kann. Dabei sind zwei Szenarien denkbar:

Aufteilen In diesem Fall ist der Abstand für das Aufteilen eines Clusters mit δ festgelegt. Es ist offensichtlich, dass dann auch das Verschmelzen zweier Cluster einem Abstand von δ entspricht. Da der Abstand für alle Arten der Aufteilung gleich sein soll, ist die Anzahl der Knoten, die von einem Cluster abgetrennt werden für die Größe des Abstands irrelevant.

Verschieben Hier wird das Verschieben eines einzelnen Knotens mit dem Abstand δ festgelegt. Bei Verwendung der Verschiebe-Operation können das Aufteilen und Verschmelzen keine Elementaroperationen sein, da das Verschmelzen zweier Cluster C_i und C_j dem Verschieben von $\min(|C_i|, |C_j|)$ Knoten entspricht.

In beiden Fällen kann jede Clustering $\mathcal{C} \in \mathbb{P}(V)$ Knoten mit maximal n Elementaroperationen in jede andere möglich Clusterung $\mathcal{C}' \in \mathbb{P}(V)$ überführt werden. Da ein maximaler Abstand mit 1 bewertet werden soll, bedeutet dies, dass $\delta = 1/n$ gilt.

Nachteile Da jegliche Aufteilung eines Clusters bei dem ersten Szenario gleich bewertet wird, ist die Anzahl der Knoten, die von einem Cluster in einen anderen verschoben werden, für die Größe des Abstands irrelevant. Diese Bewertung ist wenig intuitiv, was Abbildung 6.1 verdeutlichen soll. Clusterung $\mathcal{C}'$ ($\mathcal{C}''$) entsteht dabei aus Clusterung $\mathcal{C}$, indem ein (vier) Knoten aus dem oberen in den unteren Cluster verschoben wird.



Clusterung $\mathcal{C}'$ Clusterung $\mathcal{C}$ Clusterung $\mathcal{C}''$

Abb. 6.1: Beispiel für elementare Äquidistanz

Ein Abstandsmaß d, das elementar äquidistant ist, würde den Abstand zwischen $\mathcal{C}$ und $\mathcal{C}'$ gleich dem Abstand zwischen $\mathcal{C}$ und $\mathcal{C}''$ messen. Dies widerspricht massiv der Intuition.

Für das zweite Szenario lassen sich ähnliche Clusterungen finden, die zeigen, dass Maße, die elementar äquidistant sind, ein wenig intuitives Verhalten besitzen.

6.2 Axiome der Verbandstheorie

Thema dieses Abschnitts sollen Nachteile bzw. nicht intuitives Verhalten von Maßen sein, die einzelne Axiome der Verbandstheorie erfüllen.

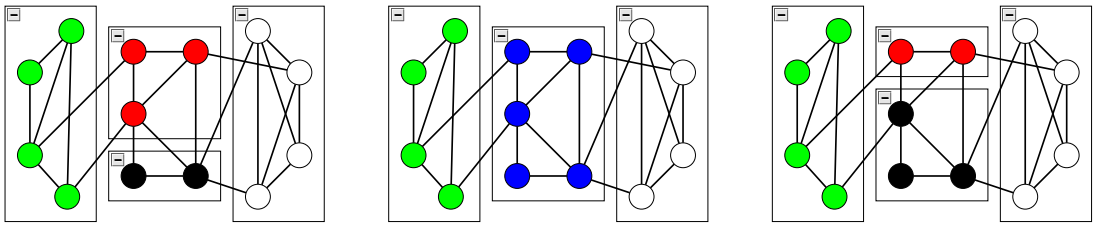
Additivität bzgl. des Produktes

Ein Abstandsmaß, dass dieses Axiom erfüllt, bewertet das Verschieben eines Knotens v von einem Cluster C_i in einen anderen Cluster C_j als die Addition von Separierung des Knotens von C_i und anschließender Vereinigung von v und C_j. Dies entspricht genau dem Szenario aus Kapitel 3.3. Dort zeigt Abbildung 3.3, dass diese Art der Berechnung von Abständen für Graphclusterungen wenig intuitiv ist.

Additivität bzgl. der Vereinigung

Analog zur Additivität bzgl. des Produktes ist auch dieses Axiom wenig intuitiv. Das Verschieben eines Knotens v von Cluster C_i in Cluster C_j entspricht in diesem Fall der Vereinigung von C_i mit C_j und der anschließenden Trennung des Clusters. Abbildung

6.2 zeigt hierfür ein Beispiel. Dabei entsteht Clusterung $\mathcal{C}'$ aus $\mathcal{C}$, indem ein Knoten aus dem mittleren oberen Cluster in den unteren verschoben wird. Clusterung $\mathcal{C}''$ ist die Vereinigung der beiden Clusterungen.



Clusterung $\mathcal{C}$ | Clusterung $\mathcal{C}'' = \mathcal{C} \oplus \mathcal{C}'$ | Clusterung $\mathcal{C}$

Abb. 6.2: Beispiel für die Additivität bzgl. der Vereinigung

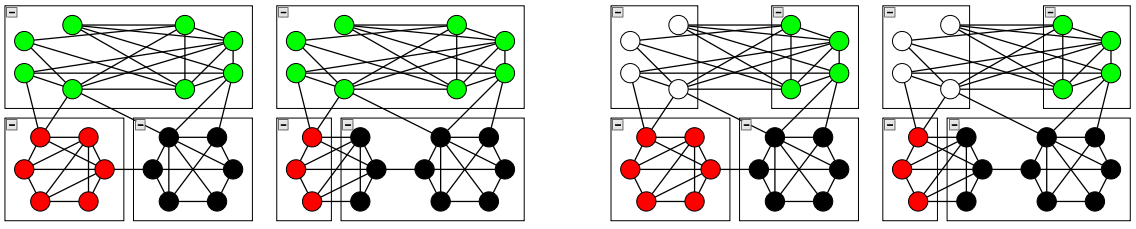
Da für ein Abstandsmaß d, welches additiv bzgl. der Vereinigung ist

$$d(\mathcal{C}, \mathcal{C}') = d(\mathcal{C}, \mathcal{C} \oplus \mathcal{C}') + d(\mathcal{C} \oplus \mathcal{C}', \mathcal{C}')$$

gilt, bedeutet das in diesem Fall, dass ein solches Maß den Abstand von der linken zur rechten Clusterung gleich bewerten würde wie den Umweg über die mittlere Clusterung. Dies widerspricht der Intuition.

Konvexe Additivität

Informell sagt konvexe Additivität aus, dass bei Veränderungen innerhalb eines Clusters der Abstand unabhängig von der restlichen Clusterung sein soll. Dies bedeutet aber bei Graphclusterung, dass beispielsweise der gemessen Abstand für das Teilen eines Clusters unabhängig von der Signifikanz der restlichen Clusterung ist. Abbildung 6.3 zeigt hierfür ein Beispiel. Clusterung $\mathcal{C}_1'$ ($\mathcal{C}_2'$) entsteht dabei jeweils durch Verschieben von 3 Knoten aus dem unteren linken in den unteren rechten Cluster.



Clusterung $\mathcal{C}_1$ | Clusterung $\mathcal{C}_1'$ | Clusterung $\mathcal{C}_2$ | Clusterung $\mathcal{C}_2'$

Abb. 6.3: Beispiel für konvexe Additivität

Die Abbildung zeigt vier Clusterungen bezüglich eines Graphen G. Für ein Abstandsmaß d, dass konvex additiv ist, würde

$$d(\mathcal{C}_1, \mathcal{C}_1') = d(\mathcal{C}_2, \mathcal{C}_2')$$

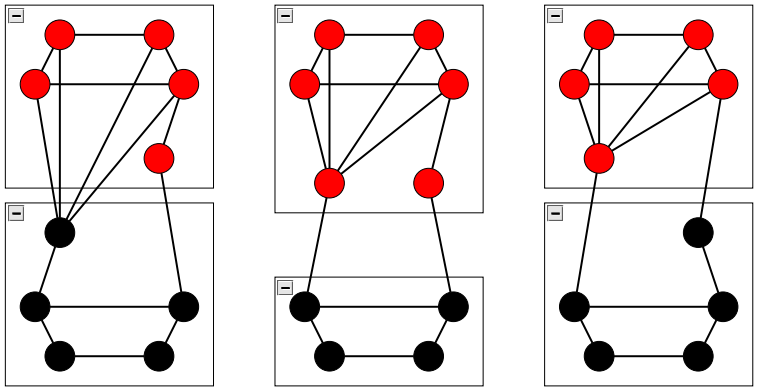
gelten, da bei beiden Vergleichen nur die unteren beiden Cluster verändert wurden. Intuitiv ist der Abstand im linken Vergleich höher, da dort das Verschieben der Knoten eine signifikante Clusterung merklich verschlechtert hat. Beim rechten Vergleich sind beide Clusterungen nicht sehr signifikant, da die oberen beiden Cluster sehr viele Interclusterkanten besitzen. Daher empfindet man dort das Verschieben der Knoten nicht als große Veränderung der Clusterung.

6.3 Knotengradabhängigkeit

Axiom 20 sagt in Worten aus, dass das Verschieben eines Knotens mit hohem Knotengrad einen größeren Einfluß auf den gemessen Abstand hat als das Verschieben eines Knotens mit kleinem Knotengrad.

Nun sollen zwei Beispiele angegeben werden, in denen gezeigt wird, dass Knotengradabhängigkeit als einzige graphstrukturelle Eigenschaft eines Maßes nicht unbedingt zu einem Vorteil gegenüber knotenstrukturellen Maßen führen muss.

Beispiel 1 Abbildung 6.4 zeigt drei Clusterungen auf einem Graphen. Die Clusterung $\mathcal{C}'$ ($\mathcal{C}''$) entsteht aus $\mathcal{C}$ durch Verschieben des linken (rechten) zentralen Knotens von dem oberen in den unteren Cluster.

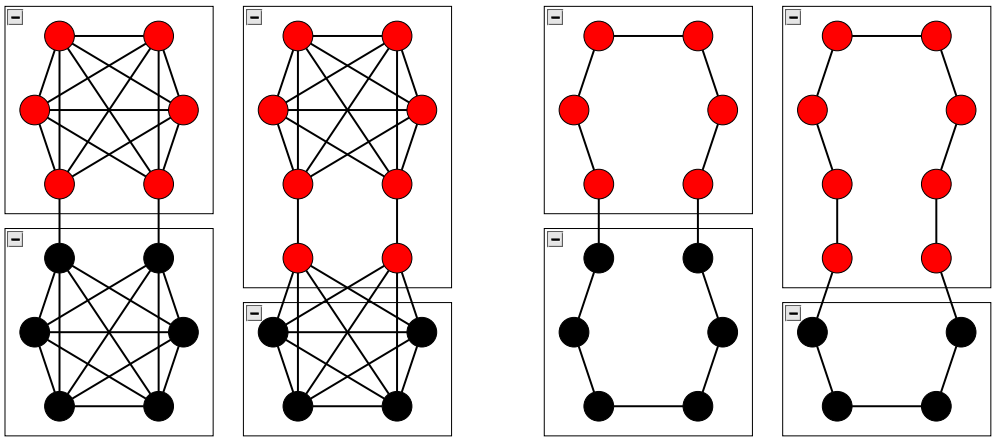


Clusterung $\mathcal{C}'$ Clusterung $\mathcal{C}$ Clusterung $\mathcal{C}''$

Abb. 6.4: Beispiel für den Vorteil von Knotengradabhängigkeit

Das Verschieben des linken Knotens von dem oberen Cluster in den unteren ist intuitiv gesehen eine größere Veränderung an der Clusterung als das Verschieben des rechten. Ein Abstandsmaß, das knotengradabhängig ist, kann diesen Unterschied messen. Ein knotenstrukturelles Maß d_k kann zwischen den beiden Fällen nicht unterscheiden.

Beispiel 2 Abbildung 6.5 zeigt nun zwei statische Clusterungsvergleiche auf unterschiedlichen Graphen.



Clusterung $\mathcal{C}_1$ Clusterung $\mathcal{C}'_1$ Clusterung $\mathcal{C}_2$ Clusterung $\mathcal{C}'_2$

Abb. 6.5: Beispiel für die Wirkungslosigkeit von Knotengradabhängigkeit

Für die Clusterungen gilt $\mathcal{C}_1 =_k \mathcal{C}_2$ und $\mathcal{C}'_1 =_k \mathcal{C}'_2$. Intuitiv ist der Abstand in der linken Grafik höher als in der rechten, da links die Clusterung vor dem Verschieben intuitiv besser als die Clusterung rechts ist. Beide Graphen sind allerdings regulär, sodass hier die Knotengrababhängigkeit wirkungslos bleibt.

Folgerung Die beiden Beispiele zeigen, dass die Knotengradabhängigkeit als einzige graphstrukturelle Eigenschaft eines Maßes nur bei Graphen mit stark schwankendem Knotengraden einen Vorteil gegenüber den knotenstrukturellen Maßen mit sich bringt. Für reguläre Graphen bleibt diese Eigenschaft wirklungslos.

6.4 Qualitative Sensivität

Die Qualitative Sensivität (Axiom 19) sagt aus, dass die gleiche knotenstrukturelle Veränderung einer Clusterung bei einer signfikanten Clusterung zu einem größeren Abstand führt als bei einer weniger signifikanten. Ein Maß mit solcher Eigenschaft würde sich sowohl für das Beispiel in Abbildung 6.4 als auch Abbildung 6.5 intuitiv verhalten.

Allerdings ist die Konstruktion eines solchen Maßes nicht trivial, da die Bewertung von Clusterungen sehr stark von der Intuition abhängt. Außerdem gibt es im Bereich der algorithmischen Clusterung einige axiomatische Negativergebnisse [Kle02].

7 Abstandsmaße

In diesem Kapitel werden einige Abstandsmaße vorgestellt. Dabei wird die in Kapitel 4 eingeführte Dreigliederung beibehalten. Die Analyse teilt sich demnach in qualitative, knotenstrukturelle und graphenstrukturelle Maße. Die knotenstrukturellen Maße sind hierbei der Literatur entnommen, die qualitativen und graphstrukturellen Maße basieren auf eigenen Überlegungen.

Bei den qualitativen und knotenstrukuturellen Abstandsmaßen wird angegeben, welche der in Kapitel 5 vorgestellten Axiome das jeweilige Maß erfüllt. Dabei werden Tabellen verwendet, in denen ein „✔“ das Erfüllen des Axioms, ein „–“ das Nichterfüllen bedeutet. Einen Überblick über die Axiome findet sich in Anhang A.

Ein Teil der in der Literatur angegebenen Maße sind Vergleichsmaße, d.h. bei identischen Clusterungen geben diese Maße den Wert Eins zurück. Darauf wird an entsprechender Stelle hingewiesen.

7.1 Qualitative Abstandsmaße

Zunächst werden einige qualitative Abstandsmaße vorgestellt. Generell gilt bei solchen Maßen, dass das Ergebnis stark vom verwendeten Index abhängt. An dieser Stelle wird auf diese Problematik nicht weiter eingegangen, sondern das Verhalten des Maßes in $[0;1] \times [0;1]$ untersucht.

Alle hier vorgestellten Maße erfüllen die Metrikeigenschaften. Desweiteren sind sie 1-maximal, 1-beschränkt, informationsvollständig (qualitativen Maßen stehen nur die Indizes zur Verfügung) und polynomiell berechenbar. Ob diese Maße Clusteranzahlunabhängig sind, hängt vom verwendeten Index ab. Die Lastigkeit und der Wachstumsfaktor wird – sofern vorhanden – angegeben.

Gerade bei qualitativen Abstandsmaßen ist eine Vielzahl an Variationen denkbar. Man kann sich sicherlich noch viele weitere Maße überlegen, die ein bestimmtes erwünschtes Verhalten besitzen. Bei den hier vorgestellten Maßen soll ein erste Intuition für die Auswirkung von Lastigkeit und Wachstumsfaktoren bei qualitativen Abstandsmaßen vermittelt werden.

7.1.1 Indexquotient

Bei dem *Indexquotienten* wird das Minimum der beiden Indizes durch das Maximum geteilt. Falls das Maximum Null ist, sind die beiden Clusterungen qualitativ gleich, somit wird für diesen Fall ein qualitativer Abstand von Null gemessen.

$$\mathcal{IQ}(\mathcal{C},\mathcal{C}') := \begin{cases} 1 - \frac{\min\{i(\mathcal{C}),i(\mathcal{C}')\}}{\max\{i(\mathcal{C}),i(\mathcal{C}')\}} & \text{für } \max\{i(\mathcal{C}),i(\mathcal{C}')\} \neq 0 \\ 0 & \text{sonst} \end{cases}$$

Das Verhalten des Maßes zeigt Abbildung 7.1. Wie man erkennen kann, ist dieses Maß sehr schlechtlastig. Kleine Abweichungen im unteren Bereich führen zu sehr großen Abständen.

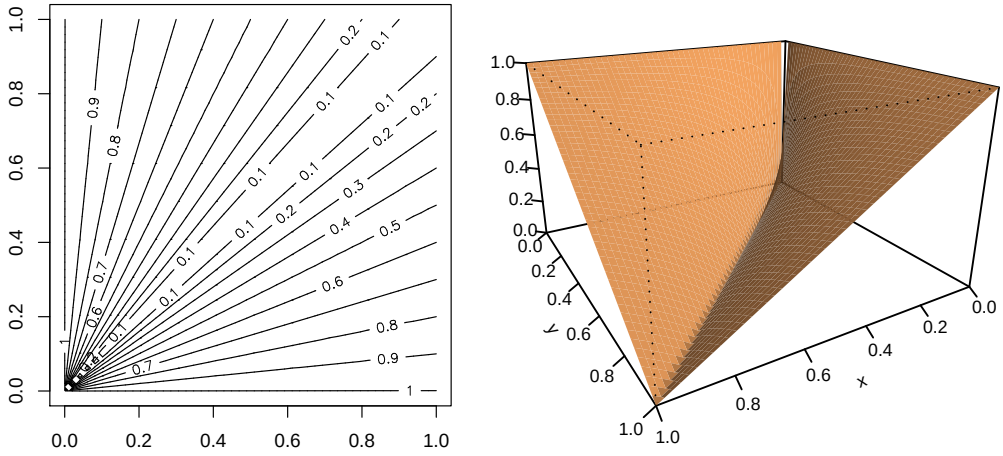


Abb. 7.1: Indexquotient in $[0;1] \times [0;1]$

Der Nachteil des Maßes ist der Fall, in dem eine Clusterung den Index Null besitzt. Dann ist der gemessene Abstand zu jeder anderen Clusterung mit einem Index ungleich Null bereits maximal.

Tabelle 7.1 zeigt, welche der Axiome aus Kapitel 5 dieses Maß erfüllt. Für jedes der nun folgenden Maße wird ein solche Tabelle angeben.

Maß	1-4	5	6	7	8	9	10	11	12	13
$\mathcal{IQ}(\mathcal{C},\mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	–	✓	–

Tabelle 7.1: Axiome des Indexquotient

Der Indexquotient besitzt keinen Wachstumsfaktor. Es ist aber ein schlechtlastiges Maß, was in Abbildung 7.1 sehr gut zu sehen ist.

7.1.2 Indexdifferenz

Ein sehr einfaches qualitatives Abstandsmaß ist die *Indexdifferenz*. Das Maß ist der Betrag der Differenz der beiden Indizes.

$$\mathcal{ID}(\mathcal{C}, \mathcal{C}') := |i(\mathcal{C}) - i(\mathcal{C}')|$$

Das Maß besitzt einen linearen Wachstumsfaktor. Abbildung 7.2 zeigt sein Verhalten.

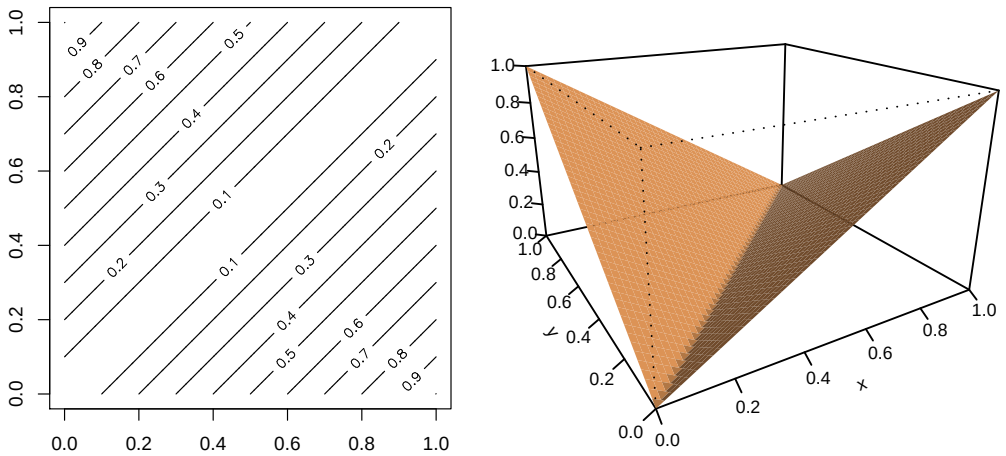


Abb. 7.2: Indexdifferenz in $[0;1] \times [0;1]$

An den parallelen Linien der linken Grafik ist der lineare Wachstumsfaktor zu erkennen. Außerdem ist schnell ersichtlich, dass dieses Maß keine Lastigkeit besitzt. Tabelle 7.2 gibt einen Überblick über die erfüllten Axiome.

Maß	1-4	5	6	7	8	9	10	11	12	13
$\mathcal{ID}(\mathcal{C}, \mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	linear	–	–

Tabelle 7.2: Axiome der Indexdifferenz

Die Indexdifferenz kann man als Ausgangspunkt für viele Varianten nutzen.

7.1.3 Varianten der Indexdifferenz

Nun werden einige Varianten der Indexdifferenz vorgestellt. Dabei liegt der Hauptaugenmerk auf verschiedenen Lastigkeiten und Wachstumsfaktoren.

Wachstumsfaktoren

Man kann ndie Indexdifferenz dahingehend verändern, dass man die Differenz der beiden Indizes noch mit einen $w \in \mathbb{R}_0^+$ potenziert. Somit ergibt

$$\mathcal{PID}(\mathcal{C}, \mathcal{C}') := |i(\mathcal{C}) - i(\mathcal{C}')|^w \text{ mit } w \in \mathbb{R}_0^+$$

die *potenzierte Indexdifferenz*. Für $w < 1$ führen kleine Abweichungen in der Differenz zu kleineren Abständen als bei der Indexdifferenz, für $w > 1$ zu größeren.

Quadratische Indexdifferenz Bei der *quadratischen Indexdifferenz* handelt es sich um eine potenzierte Indexdifferenz mit $w = 2$. Das bedeutet, dass dieses Maße 2–exponentiell wachsend ist.

$$\mathcal{QID}(\mathcal{C}, \mathcal{C}') := |i(\mathcal{C}) - i(\mathcal{C}')|^2$$

Abbildung 7.3 zeigt das Verhalten des Maßes im Bereich $[0; 1] \times [0; 1]$.

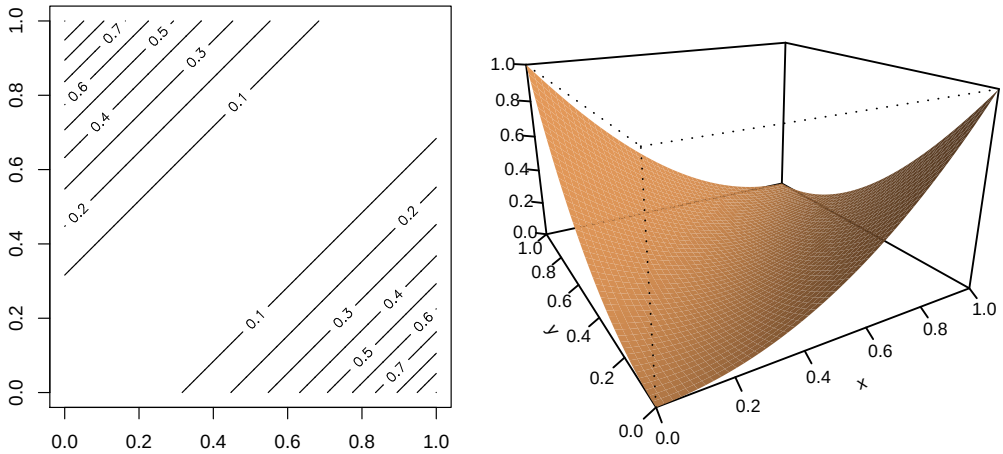


Abb. 7.3: Quadratische Indexdifferenz in $[0; 1] \times [0; 1]$

Man erkennt, dass eine Abweichung von ca. 0.25 in der Differenz der Maße lediglich zu einem gemessenen Abstand von 0.0625 führt. Das muss bei der Wahl dieses Maßes bedacht werden. Um den Effekt gegebenenfalls abzuschwächen, kann man auch einen kleineres $w > 1$ wählen.

Gewurzelte Indexdifferenz Die gewurzelte Indexdifferenz ist 0.5-exponentiell wachsend und somit mit

$$\mathcal{WID}(\mathcal{C}, \mathcal{C}') := \sqrt{|i(\mathcal{C}) - i(\mathcal{C}')|}$$

definiert.

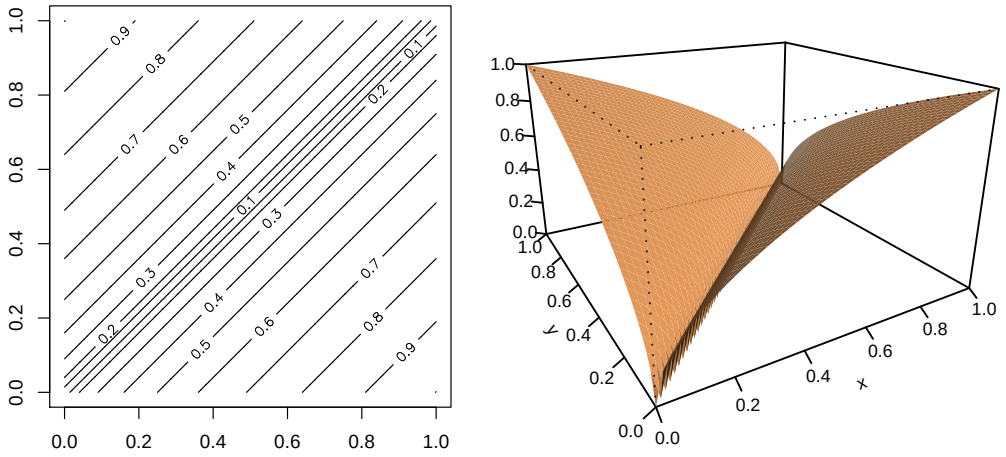


Abb. 7.4: Gewurzelte Indexdifferenz in $[0;1] \times [0;1]$

Man erkennt, dass bereits eine sehr kleine Abweichung in den Indizes zu einem großen gemessenen Abstand führt. Analog zu der quadratischen Indexdifferenz kann man diesen Effekt mit einem anderem $w < 1$ abschwächen bzw. verstärken.

Lastigkeit

Als Nächstes wird gezeigt, wie man die Indexdifferenz in ein lastiges Maß transformieren kann. Die allgemeine Formel für die *lastige Indexdifferenz* lautet:

$$\mathcal{LID}(\mathcal{C},\mathcal{C}') := \underbrace{g_1 \left| \sqrt[l_1]{i(\mathcal{C})} - \sqrt[l_1]{i(\mathcal{C}')} \right|}_{\text{Schlechtlastigkeit}} + \underbrace{g_2 \left| \sqrt[l_2]{1-i(\mathcal{C})} - \sqrt[l_2]{1-i(\mathcal{C}')} \right|}_{\text{Gutlastigkeit}}$$

$$\text{mit } g_1 + g_2 = 1 \text{ und } l_1, l_2 \in \mathbb{R}^+, l_1, l_2 \geq 1$$

Der vordere Teil der Formel ist hierbei für die Schlechtlastigkeit des Maßes verantwortlich. Mit wachsendem l_1 nimmt auch die Ausprägung des Schlechtlastigkeit zu. Der hintere Teil dient analog der Gutlastigkeit mit gleichem Einfluss von l_2. Mit g_1 und g_2 kann man festlegen, inwieweit sich die Lastigkeiten zueinander verhalten. Damit das Maß in jedem Falle 1-beschränkt bleibt, muss $g_1 + g_2 = 1$ gelten.

Man kann die Lastigkeit natürlich mit den Wachstumsfaktoren kombinieren, um einzelne Effekte weiter zu verstärken, bzw. abzuschwächen.

Schlechtlastigkeit Die *schlechtlastige Indexdifferenz* bewertet Abweichungen im qualitativ schlechteren Bereich von Clusterungen stärker als im besseren Bereich. Man erhält dieses Maß aus der lastigen Indexdifferenz mit $g_1 = 1, g_2 = 0$ und $l_1 = 2$. Dies ergibt:

$$\mathcal{SID}(\mathcal{C},\mathcal{C}') := \left| \sqrt{i(\mathcal{C})} - \sqrt{i(\mathcal{C}')} \right|$$

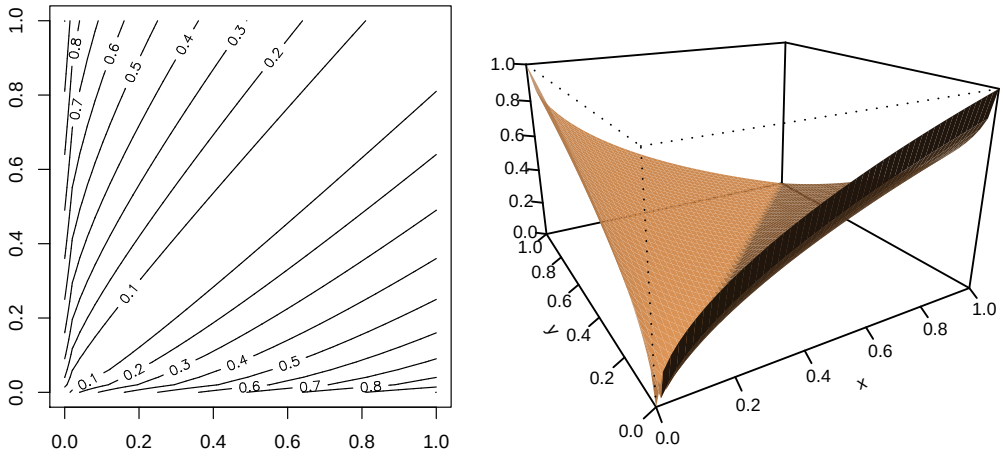


Abb. 7.5: Schlechtlastige Indexdifferenz in $[0;1] \times [0;1]$

Man erkennt, dass Abweichungen im besseren Bereich weniger relevant sind als im schlechteren. Bei Verwendung dieses Maßes sollte das beachtet werden.

Gutlastigkeit Die *gutlastige Indexdifferenz* entspricht der lastigen Indexdifferenz mit $g_1 =, g_2 = 1$ und $l_2 = 1$:

$$\mathcal{GID}(\mathcal{C}, \mathcal{C}') := |\sqrt{1 - i(\mathcal{C})} - \sqrt{1 - i(\mathcal{C}')}|$$

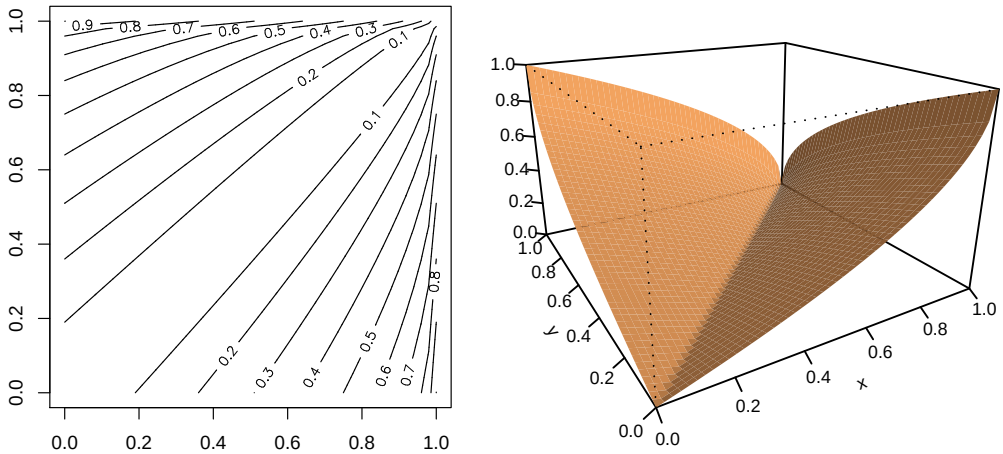


Abb. 7.6: Gutlastige Indexdifferenz in $[0;1] \times [0;1]$

Man erkennt, dass das Maß den Sachverhalt der schlechtlastigen Indexdifferenz umdreht. Abweichungen im signifikanten Bereich der Clusterung führen zu einem größeren gemessenen qualitativen Abstand.

Doppelte Lastigkeit Die *doppeltlastige Indexdifferenz* kombiniert Gut- und Schlechtlastigkeit. Mit Wahl von $g_1 = g_2 = 1/2$ und $l_1 = l_2 = 2$ erhält man:

$$\mathcal{BID}(\mathcal{C}, \mathcal{C}') := \frac{1}{2}\left(\left|\sqrt{1 - i(\mathcal{C})} - \sqrt{1 - i(\mathcal{C}')}\right| + \left|\sqrt{i(\mathcal{C})} - \sqrt{i(\mathcal{C}')}\right|\right)$$

Bei diesem Maß ist $a = 1$ der Umschlagspunkt der Lastigkeit, d.h. für $i(\mathcal{C}) + i(\mathcal{C}') < 1$ ist das Maß schlechtlastig und für $i(\mathcal{C}) + i(\mathcal{C}') > 1$ gutlastig.

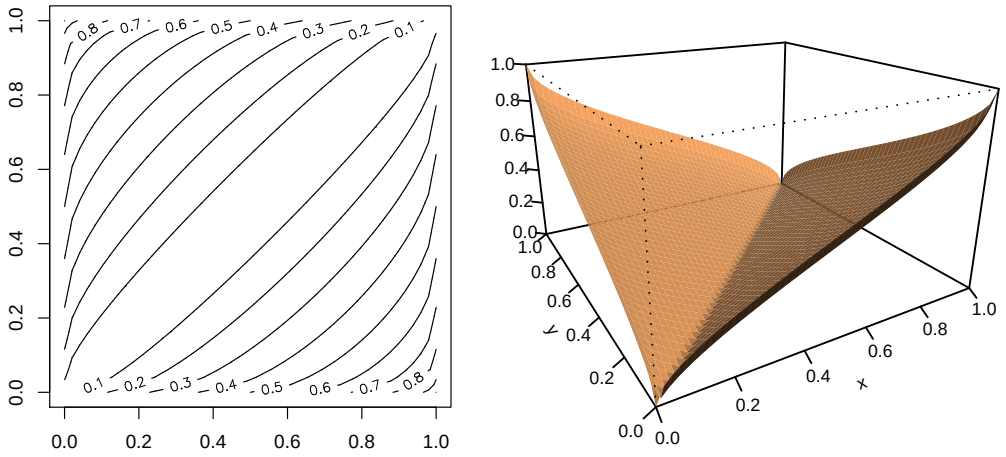


Abb. 7.7: Beidlastige Indexdifferenz in $[0; 1] \times [0; 1]$

Wie man erkennt, ist die Ausprägung der Lastigkeit hier nicht sehr hoch, dieser Effekt lässt sich durch Exponentiation des Maßes verstärken.

Axiome der Varianten

Tabelle 7.3 gibt nun noch eine Übersicht über die verschiedenen Varianten der Indexdifferenz, wobei sich die Maße ausschließlich in der Art der Lastigkeit bzw. dem Wachstumsfaktor unterscheiden.

Maß	1-4	5	6	7	8	9	10	11	12	13
$\mathcal{QID}(\mathcal{C}, \mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	2-exp.	–	–
$\mathcal{WID}(\mathcal{C}, \mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	0.5-exp.	–	–
$\mathcal{SID}(\mathcal{C}, \mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	–	✓	–
$\mathcal{GID}(\mathcal{C}, \mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	–	✓	–
$\mathcal{BID}(\mathcal{C}, \mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	–	–	✓

Tabelle 7.3: Axiome der Varianten der Indexdifferenz

7.2 Knotenstrukturelle Abstandsmaße

In diesem Abschnitt werden die untersuchten knotenstrukturellen Maße vorgestellt, sie sind alle der jeweils angegebenen Literatur entnommen. Generell kann man die Maße nach ihrem Ansatz unterscheiden:

Paarmaße Bei diesen Maßen werden die in Abschnitt 2.3.3 vorgestellten globalen Paarzählungsmengen verwendet. Da diese Mengen nicht die Kantenmengen der Graphen betrachten, sind Paarmaße knotenstrukturelle Maße.

Schnittmaße Die Schnittmaße basieren auf der in Abschnitt 2.3.2 vorgestellten Verschmelzungsmatrix. Sie benötigt ebenfalls nur die Partitionen der Knoten. Somit sind diese Maße auch knotenstrukturell. Die Maße sollen als Schnittmaße bezeichnet werden, da die Verschmelzungsmatrix auf den Schnittmengen der Cluster basiert.

Entropiemaße Diese Maße nutzen die in Abschnitt 2.3.4 vorgestellte Entropie der beiden Clusterungen sowie die Korrelationsinformation der beiden Clusterungen. Entropie und Korrelationsinformation nutzen nur die Partitionen und betrachten nicht die Graphen. Daher sind auch Entropiemaße knotenstrukturelle Maße.

Allen Ansätzen ist gemein, dass die beiden Clusterungen die gleiche Knotenmenge V partitionieren müssen. Da in dieser Arbeit nur der statische Clusterungsvergleich untersucht wird, ist das kein Nachteil. Dennoch muss es bei Untersuchungen über den dynamischen Clusterungsvergleich bedacht werden.

Außerdem werden bei den knotenstrukturellen Maßen die Abstände zwischen den Clusterungen $\mathcal{C}^s, \mathcal{C}^1, \mathcal{C}^\times$ und $\mathcal{C}^\perp$ auf einem Graphen mit 1024 Knoten angegeben. Da die Maße die Kantenmenge nicht betrachten, ist die Anzahl der Kanten für das Ergebnis irrelevant. Die Abstände werden angegeben, um einzelne Eigenschaften zu zeigen und Axiome zu widerlegen.

7.2.1 Paarmaße

Rand–Maß

Das Maß von Rand [Ran71] ist durch ein Klassifikationsproblem motiviert. Dabei wird eine Klassifizierung mit einer bekannten richtigen Klassifizierung der Daten verglichen, indem berechnet wird, wieviele Elemente richtig klassifiziert wurden.

$$\mathcal{R}(\mathcal{C}, \mathcal{C}') := \frac{2(n_{11} + n_{00})}{n(n-1)}$$

Das Maß ist ein Vergleichsmaß, da $\mathcal{R}(\mathcal{C},\mathcal{C}) = 1$ gilt. Ein Hauptnachteil des Maßes ist, dass es von der Anzahl der Knoten und Cluster abhängig ist [MA84]. Außerdem konvergiert es Maß für Zufallsclusterungen mit steigender Clusteranzahl gegen Eins und genau das macht dieses Maß für den Vergleich von Clusterungen uninteressant.

Das Rand Maß ist ein Vergleichmaß. Daher wird die Abstandsversion $\mathcal{R}' := 1-\mathcal{R}$ genutzt. Tabelle 7.4 zeigt die Abstände zwischen den vier vorgegebenen Clusterungen.

$\mathcal{R}'(\mathcal{C}^s,\mathcal{C}^1)$	$\mathcal{R}'(\mathcal{C}^s,\mathcal{C}^\times)$	$\mathcal{R}'(\mathcal{C}^\times,\mathcal{C}^1)$	$\mathcal{R}'(\mathcal{C}^\times,\mathcal{C}^\perp)$
1.0	0.03	0.97	0.06

Tabelle 7.4: Abstände des Rand–Maßes für $n = 1024$

Der geringe Abstand zwischen den beiden komplementären Clusterungen zeigt das Phänomen, dass für eine große Clusteranzahl der gemessene Abstand sehr gering ist, obwohl die Clusterungen stark verschieden sind. Die Tabelle führt zu der Vermutung, dass das Maß additiv bzgl. der Verfeinerung und des Produktes ist. Dies wird von Meila in [Mei05] gezeigt. Tabelle 7.5 gibt einen Überblick über die Axiome.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{R}'(\mathcal{C},\mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	–	✓	–	✓	–

Tabelle 7.5: Axiome des Rand–Maßes

Meila zeigt ebenfalls, dass dieses Maß nicht konvex additiv ist. Obwohl es recht viele Axiome erfüllt, scheint das Rand–Maß wegen des geringen Abstands zwischen den komplementären Clusterungen wenig geeignet zu sein.

Angepasstes Rand–Maß

Um die Problematik mit der steigenden Clusteranzahl zu umgehen, gibt es mehrere Anpassungen des Rand–Maßes. Davon wird exemplarisch eine vorgestellt: Bei dieser Anpassung wird die Differenz zwischen dem Rand–Maß und dem erwarteten Rand–Maß für eine hypergeometrische Verteilung gebildet [HA85]. Formal ergibt dies:

$$\mathcal{AR}(\mathcal{C},\mathcal{C}') := \frac{n_{11} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}$$

$$\text{mit } t_1 := \sum_{i=1}^{k} \binom{|C_i|}{2}, t_2 := \sum_{j=1}^{l} \binom{|C'_j|}{2}, t_3 := \frac{2t_1t_2}{n(n-1)}$$

Das Maß hat einen erwarteten Wert von Null für unabhängige Clusterungen und ist Eins für identische Clusterungen. Somit ist auch das angepasste Rand–Maß ein

Vergleichsmaß. Daher soll die Abstandsversion

$$\mathcal{AR}'(\mathcal{C},\mathcal{C}') := 1 - \mathcal{AR}(\mathcal{C},\mathcal{C}')$$

untersucht werden. Tabelle 7.6 zeigt die Abstände zwischen den vier Clusterungen.

$\mathcal{AR}'(\mathcal{C}^s,\mathcal{C}^1)$	$\mathcal{AR}'(\mathcal{C}^s,\mathcal{C}^\times)$	$\mathcal{AR}'(\mathcal{C}^\times,\mathcal{C}^1)$	$\mathcal{AR}'(\mathcal{C}^\times,\mathcal{C}^\perp)$
1.0	1.0	1.0	1.03

Tabelle 7.6: Abstände des angepassten Rand–Maßes für $n = 1024$

Man erkennt, dass alle Abstände nun maximal sind, der Abstand zwischen den komplementären Clusterungen erreicht sogar einen Wert größer Eins. Somit ist dieses Maß nicht mehr 1-maximal. Ebenso erfüllt es keine Axiome der Verbandstheorie. Einen Überblick gibt Tabelle 7.7.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{AR}'(\mathcal{C},\mathcal{C}')$	✓	–	–	–	–	✓	✓	–	–	–	–	–

Tabelle 7.7: Axiome des angepassten Rand–Maßes

Dieses Maß scheint wenig geeignet für den Vergleich von Clusterungen, da die Annahme der hypergeometrischen Verteilung eine sehr starke ist.

Fowlkes–Mallows

Das von Fowlkes und Mallows in [FM83] eingeführte Maß ist definiert durch:

$$\begin{aligned}\mathcal{FM}(\mathcal{C},\mathcal{C}') &:= \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij}^2 - n}{\sqrt{(\sum_i |C_i|^2 - n)(\sum_j |C'_j|^2 - n)}} \\ &\overset{\text{Kap. 2.3.3}}{=} \frac{n_{11}}{\sqrt{(n_{11}+n_{10})(n_{11}+n_{01})}}\end{aligned}$$

Dieses Maß ist ein Vergleichmaß. In der Literatur finden sich keine Informationen über den Spezialfall, falls eine Clusterung die Singleton-Clusterung ist. In diesem Fall ist der Nenner gleich Null, weshalb eine Variante genutzt wird, die zugleich ein Abstandsmaß ist.

$$\mathcal{FM}'(\mathcal{C},\mathcal{C}') := \begin{cases} 1 - \mathcal{FM}(\mathcal{C},\mathcal{C}') & \text{für } n_{01}, n_{10} \neq 0 \vee n_{11} \neq 0 \\ 1 & \text{für } n_{11}, n_{01} = 0 \vee n_{11}, n_{10} = 0 \\ 0 & \text{sonst} \end{cases}$$

Das hat den Nachteil, dass für den Fall, in dem nur eine Clusterung die Singleton–Clusterung ist, der Abstand maximal gemessen wird, unabhängig davon, wie die zweite Clusterung strukturiert ist. Es sei denn, beide Clusterungen sind die Singleton–Clusterung, dann wird der Abstand auf Null gesetzt.

Wie das Rand–Maß besitzt auch dieses Maß den Nachteil, dass es von der Clusteranzahl stark abhängig ist. Außerdem hat es die Eigenschaft, kein Axiom aus der Verbandstheorie zu erfüllen, was man sehr gut an den Abständen in Tabelle 7.8 ablesen kann.

$\mathcal{FM}'(\mathcal{C}^s, \mathcal{C}^1)$	$\mathcal{FM}'(\mathcal{C}^s, \mathcal{C}^\times)$	$\mathcal{FM}'(\mathcal{C}^\times, \mathcal{C}^1)$	$\mathcal{FM}'(\mathcal{C}^\times, \mathcal{C}^\perp)$
1.0	1.0	0.83	1.0

Tabelle 7.8: Abstände des Fowlkes–Mallows–Maßes für $n = 1024$

Das Maß ist allerdings 1-maximal, 1-beschränkt und ist auch eine Metrik. Tabelle 7.11 gibt einen Überblick.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{FM}'(\mathcal{C}, \mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	–	–	–	–	–

Tabelle 7.9: Axiome des Fowlkes–Mallows–Maßes

Mirkin–Metrik

Die Mirkin–Metrik wird in [Don00b] eingeführt und ist als

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') := \sum_{i=1}^{k} |C_i|^2 + \sum_{j=1}^{l} |C'_j|^2 - \sum_{i=1}^{k} \sum_{j=1}^{l} m_{ij}^2$$

definiert. Durch Umformung zeigt man, dass die Mirkin–Metrik dem nicht–normierten Rand–Maß als Abstandsmaß entspricht:

$$\begin{aligned} \mathcal{M}(\mathcal{C}, \mathcal{C}') &= \sum_{i=1}^{k} |C_i|^2 + \sum_{j=1}^{l} |C'_j|^2 - \sum_{i=1}^{k} \sum_{j=1}^{l} m_{ij}^2 \\ &\overset{\text{Kap. 2.3.3}}{=} 2(n_{01} + n_{10}) \\ &= n(n-1)\mathcal{R}'(\mathcal{C}, \mathcal{C}') \end{aligned}$$

Somit ist eine gesonderte Untersuchung der Mirkin–Metrik hinfällig, da die Axiome, die durch das Rand–Maß erfüllt werden, auch durch die normierte Mirkin–Metrik erfüllt werden.

Jaccard

Das Jaccard- oder Sörensen–Maß [Sör48] findet häufig Verwendung in der Geologie und Ökologie. Es ist definiert durch:

$$\mathcal{J}(\mathcal{C}, \mathcal{C}') := \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

Auch das Jaccard–Maß ist dem Rand–Maß sehr ähnlich, es beachtet lediglich die Elemente der Menge S_{00} nicht. Desweiteren ist – wie bei dem Maß von Fowlkes–Mallows – keine Aussage über den Spezialfall, wenn nämlich der Nenner Null ist, zu finden. Anders als beim Maß von Fowlkes–Mallows kann dies allerdings nur in dem Fall geschehen, wenn beide Clusterungen die Singleton-Clusterung sind. Daher wird folgende Abstandsversion des Maßes genutzt:

$$\mathcal{J}'(\mathcal{C},\mathcal{C}') := \begin{cases} 1 - \mathcal{J}(\mathcal{C},\mathcal{C}') & \text{für } n_{11}+n_{10}+n_{01} \neq 0 \\ 0 & \text{sonst} \end{cases}$$

Für diese Version ergeben sich die in Tabelle 7.10 angegebenen Abstände. Das Maß

$\mathcal{J}'(\mathcal{C}^s,\mathcal{C}^1)$	$\mathcal{J}'(\mathcal{C}^s,\mathcal{C}^\times)$	$\mathcal{J}'(\mathcal{C}^\times,\mathcal{C}^1)$	$\mathcal{J}'(\mathcal{C}^\times,\mathcal{C}^\perp)$
1.0	1.0	0.97	1.0

Tabelle 7.10: Abstände des Jaccard–Maßes für $n = 1024$

misst alle Abstände annähernd maximal. Daher erfüllt auch dieses Maß die Axiome der Verbandstheorie nicht.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{J}'(\mathcal{C},\mathcal{C}')$	✓	✓	✓	✓	–	✓	✓	–	–	–	–	–

Tabelle 7.11: Axiome des Jaccard–Maßes

Partitionsdifferenz

Die Partitionsdifferenz zählt lediglich diejenigen Paare, die sich bezüglich beider Clusterungen in verschiedenen Clustern befinden.

$$\mathcal{PD}(\mathcal{C},\mathcal{C}') := n_{00}$$

Bei diesem Maß handelt es sich weder um ein Abstandsmaß, da $n_{00} \neq 0$ für $\mathcal{C} = \mathcal{C}'$, noch handelt es sich um Vergleichsmaß, da für $n_{00} = 0$ nicht unbedingt $\mathcal{C} = \mathcal{C}'$ gelten muss. Obwohl es nach [LOM04] häufig genutzt wird, wird die Partitions Differenz in den Experimenten nicht untersucht.

7.2.2 Schnittmaße

Als Schnittmaße sollen die Maße bezeichnet werden, die auf der Verschmelzungsmatrix basieren. Dabei werden in den meisten Fällen Cluster aus $\mathcal{C}$ einem Cluster aus $\mathcal{C}'$ zugeordnet und überprüft, wie groß die Schnittmengen sind. Allerdings werden hierbei die Elemente außerhalb der betrachteten Schnitte nicht für die Berechnung des Ergebnisses genutzt. Somit sind diese Maße nicht informationsvollständig.

F-Maß

Das F-Maß hat seinen Ursprung in der Clusterung von Dokumenten [FWE03], wobei hier eine optimale Clusterung bekannt ist, und eine andere Clusterung mit dieser verglichen wird. Hierzu wird berechnet, wie gut der Cluster C'_j den Cluster C_i beschreibt. Dies geschieht über das F-Maß:

$$\mathcal{F}(C_i, C'_j) := \frac{2m_{ij}}{|C_i| + |C'_j|}$$

Das F-Maß für die beiden Clusterungen $\mathcal{C}$ und $\mathcal{C}'$ ist dann die gewichtete Summe der maximalen F-Maße der Cluster in $\mathcal{C}'$:

$$\mathcal{F}(\mathcal{C}, \mathcal{C}') := \frac{1}{n} \sum_{i=1}^{k} |C_i| \max_{j=1}^{l} \{\mathcal{F}(C_i, C'_j)\}$$

Das bedeutet, dass für jeden Cluster C'_j in $\mathcal{C}'$ derjenige Cluster in $\mathcal{C}$ gesucht wird, der C'_j am besten beschreibt. Das F-Maß ist nicht symmetrisch, womit dieses Maß keine Metrik ist. Die Tatsache der Asymmetrie macht eine Deutung des Maßes schwierig.

Für die Abstände der vier Sonderfälle in Tabelle 7.12 gilt allerdings die Symmetrie, daher werden nur die Ergebnisse für eine Richtung angegeben. Das F-Maß ist ein Vergleichmaß, sodass erneut die Abstandsversion $\mathcal{F}'$ genutzt wird.

$\mathcal{F}'(\mathcal{C}^s, \mathcal{C}^1)$	$\mathcal{F}'(\mathcal{C}^s, \mathcal{C}^\times)$	$\mathcal{F}'(\mathcal{C}^\times, \mathcal{C}^1)$	$\mathcal{F}'(\mathcal{C}^\times, \mathcal{C}^\perp)$
0.998	0.939	0.939	0.969

Tabelle 7.12: Abstände des F-Maßes für $n = 1024$

Dieses Maß nicht 1-maximal, sondern lediglich grenzwertig 1-maximal. Die Abstände zeigen zudem, dass kein Axiom der Verbandstheorie erfüllt ist. Einen Überblick gibt Tabelle 7.13.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{F}'(\mathcal{C}, \mathcal{C}')$	–	✓	–	✓	–	–	✓	–	–	–	–	–

Tabelle 7.13: Axiome des F-Maßes

Aufgrund der Asymmetrie scheint eine Verwendung dieses Maßes wenig sinnvoll.

Meila–Heckerman

Meila und Heckerman führen in [MH01] ebenfalls ein asymmetrisches Schnittmaß ein, welches dem F-Maß sehr ähnlich ist. Hierbei wird nicht das F-Maß eines Clusters

aus $\mathcal{C}$ maximiert, sondern nur die Größe des Schnittes. Somit ergibt sich:

$$\mathcal{MH}(\mathcal{C}, \mathcal{C}') := \frac{1}{n} \sum_{i=1}^{k} \max_{C'_j \in \mathcal{C}'} m_{ij}$$

Auch hier macht die Asymmetrie eine Deutung schwierig. Somit ist auch dieses Maß keine Metrik. Desweiteren handelt es sich hier erneut um ein Vergleichsmaß, so dass die Abstandsversion $\mathcal{MH}'$ genutzt wird. Für diese ergeben sich die Abstände in Tabelle 7.14, wobei bei den Ergebnissen in der dritten Zeile jeweils die beiden Clusterungen bei der Eingabe vertauscht wurden.

$\mathcal{MH}'(\mathcal{C}^s, \mathcal{C}^1)$	$\mathcal{MH}'(\mathcal{C}^s, \mathcal{C}^\times)$	$\mathcal{MH}'(\mathcal{C}^\times, \mathcal{C}^1)$	$\mathcal{MH}'(\mathcal{C}^\times, \mathcal{C}^\perp)$
0.0	0.0	0.0	0.969
0.999	0.969	0.969	0.969

Tabelle 7.14: Abstände des Maßes von Meila–Heckerman für $n = 1024$

Hier hat die Asymmetrie des Maßes zur Folge, dass manche Abstände nicht gemessen werden können, was eine Verwendung des Maßes sehr schwierig macht. Dieses Maß erfüllt die gleichen Axiome wie das F-Maß, weshalb auf die Angabe einer gesonderten Tabelle verzichtet wird.

Maximum–Match

Das Maximum–Match–Maß [MH01] ist eine symmetrische Generalisierung des Meila–Heckerman–Maßes. Hierbei werden folgende zwei Schritte auf der Verschmelzungsmatrix wiederholt, bis die Matrix keine Enträge mehr enthält:

1. Finde $m_{ab} = \max m_{ij}$ in der Verschmelzunsmatrix.
2. Lösche Zeile a und Spalte b aus der Verschmelzungsmatrix.

Danach werden die Maxima aufsummiert und durch die Anzahl der Elemente dividiert. Dies ergibt:

$$\mathcal{MM}(\mathcal{C}, \mathcal{C}') := \frac{1}{n} \sum_{i=1}^{\min\{k,l\}} m_{ii'}$$

Auch dieses Maß ist ein Vergleichmaß, daher wird die Abstandsversion $\mathcal{MM}'$ genutzt. Tabelle 7.15 enthält die Abstände für die vier Sonderfälle.

Dieses Maß ist eine Metrik, allerdings ignoriert es ganze Cluster, falls $|\mathcal{C}| \neq |\mathcal{C}'|$. Somit ist es auch nicht informationsvollständig.

Vor allem bei Vergleichen zwischen Clusterungen mit stark abweichender Clusteranzahl sollte dieses Maß nicht genutzt werden.

$\mathcal{MM}'(\mathcal{C}^s, \mathcal{C}^1)$	$\mathcal{MM}'(\mathcal{C}^s, \mathcal{C}^\times)$	$\mathcal{MM}'(\mathcal{C}^\times, \mathcal{C}^1)$	$\mathcal{MM}'(\mathcal{C}^\times, \mathcal{C}^\perp)$
0.999	0.969	0.969	0.969

Tabelle 7.15: Abstände des Maximum–Match–Maßes für $n = 1024$

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{MM}'(\mathcal{C}, \mathcal{C}')$	✓	✓	–	✓	–	–	✓	–	–	–	–	–

Tabelle 7.16: Axiome des Maximum–Match–Maßes

Van Dongen

Van Dongen benutzt in [Don00b] ein symmetrisches Schnittmaß für den Vergleich von Clusterungen.

$$\mathcal{VD}(\mathcal{C}, \mathcal{C}') := 2n - \sum_{i=1}^{k} \max_j m_{ij} - \sum_{j=1}^{l} \max_i m_{ij}$$

Dieses Maß ist eine Metrik auf allen Clusterungen, allerdings ist es nicht 1-beschränkt. Wie alle Schnittmaße ist auch dieses Maß nicht informationsvollständig.

Normierung Man kann dieses Maß mit $2n$ normieren:

$$\mathcal{NVD}(\mathcal{C}, \mathcal{C}') := \frac{\mathcal{VD}(\mathcal{C}, \mathcal{C}')}{2n}$$

Durch diese Normierung ist es zwar 1-beschränkt, allerdings nicht 1-maximal. Das liegt daran, dass die beiden Summen niemals Null ergeben.

$\mathcal{NVD}(\mathcal{C}^s, \mathcal{C}^1)$	$\mathcal{NVD}(\mathcal{C}^s, \mathcal{C}^\times)$	$\mathcal{NVD}(\mathcal{C}^\times, \mathcal{C}^1)$	$\mathcal{NVD}(\mathcal{C}^\times, \mathcal{C}^\perp)$
0.5	0.484	0.484	0.969

Tabelle 7.17: Abstände des normalisierten van Dongen–Maßes für $n = 1024$

Auffällig ist, dass der Abstand zwischen Singleton- und 1-Clusterung mit 0.5 gewertet wird. Der Abstand zwischen den Komplementärclusterungen wird hingegen mit 0.969 gewertet. Es kann sogar gezeigt werden, dass folgende Grenzwerte gelten:

$$\begin{aligned} \lim_{n\to\infty} \mathcal{VD}(\mathcal{C}^s, \mathcal{C}^1) &= 0.5 \\ \lim_{n\to\infty} \mathcal{VD}(\mathcal{C}^\times, \mathcal{C}^\perp) &= 1 \end{aligned}$$

Das bedeutet, dass das van Dongen–Maß den Abstand zwischen Komplementärclusterungen doppelt so groß misst wie zwischen Singleton- und 1-Clusterung. Zusätzlich folgt aus den Grenzwerten, dass das Maß grenzwertig 1-maximal ist.

Meila zeigt in [Mei05], dass das van Dongen–Maß konvexadditiv und additiv bzgl. des Produktes ist. Sie zeigt auch, dass das Maß nicht additiv bzgl. der Verfeinerung ist. Einen Überblick über die erfüllten Axiome der normierten und nichtnormierten Version des van Dongen–Maßes gibt Tabelle 7.18.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{VD}(\mathcal{C},\mathcal{C}')$	✓	–	–	–	–	–	✓	–	–	✓	✓	✓
$\mathcal{NVD}(\mathcal{C},\mathcal{C}')$	✓	✓	–	✓	–	–	✓	–	–	✓	✓	✓

Tabelle 7.18: Axiome der van Dongen–Maße

Das van Dongen–Maß erfüllt zwar recht viele Axiome, allerdings muss man bei Verwendung des Maßes den geringen gemessenen Abstand zwischen $\mathcal{C}^s$ und $\mathcal{C}^1$ beachten.

7.2.3 Entropiemaße

Die entropiebasierten Abstandsmaße kombinieren alle die Entropien der beiden Clusterungen $\mathcal{C}$ und $\mathcal{C}'$ und die Korrelationsinformation $\mathcal{I}(\mathcal{C},\mathcal{C}')$. Die Maße von Strehl & Ghosh und Fred & Jain normalisieren auf zwei verschiedene Arten die Korrelationsinformation, wohingegen Meilas Variation der Information einen anderen Ansatz wählt.

Strehl & Ghosh

In [SG03] stellen Strehl & Gosh eine Möglichkeit vor, mehrere Clusterungen zu einer einzigen zu kombinieren. Dabei nutzen sie unter anderem folgendes Vergleichsmaß:

$$\mathcal{SG}(\mathcal{C},\mathcal{C}') := \frac{\mathcal{I}(\mathcal{C},\mathcal{C}')}{\sqrt{\mathcal{H}(\mathcal{C})\mathcal{H}(\mathcal{C}')}}$$

In [SG03] bleibt unerwähnt, wie mit der Möglichkeit, dass der Nenner Null ergeben kann, umgegangen werden soll. Daher wird folgendes Maß als Abstandsmaß genutzt:

$$\mathcal{SG}'(\mathcal{C},\mathcal{C}') := \begin{cases} 1-\mathcal{SG}(\mathcal{C},\mathcal{C}') & \text{für } \mathcal{H}(\mathcal{C}),\mathcal{H}(\mathcal{C}') \neq 0 \\ 0 & \text{für } \mathcal{H}(\mathcal{C}) = \mathcal{H}(\mathcal{C}') = 0 \\ 1 & \text{sonst} \end{cases}$$

Die dadurch bedingten Abstände sind in Tabelle 7.19 abzulesen.

$\mathcal{SG}'(\mathcal{C}^s,\mathcal{C}^1)$	$\mathcal{SG}'(\mathcal{C}^s,\mathcal{C}^\times)$	$\mathcal{SG}'(\mathcal{C}^\times,\mathcal{C}^1)$	$\mathcal{SG}'(\mathcal{C}^\times,\mathcal{C}^\perp)$
1.0	0.293	1.0	1.0

Tabelle 7.19: Abstände des Maßes von Strehl & Gosh für $n = 1024$

Dieses Maß ist eine Metrik, da die Korrelationsinformation eine Metrik ist. Eebenso gelten 1-Beschränktheit und 1-Maximalität. Mit den Abständen zwischen den gegebenen Clusterungen ergibt sich ebenso, dass die Axiome der Verbandstheorie nicht gelten.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{SG}(\mathcal{C},\mathcal{C}')$	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–

Tabelle 7.20: Axiome des Maßes von Strehl & Ghosh

Das Maß von Strehl & Gosh erscheint vielversprechend, allerdings scheint die Tatsache, dass der Abstand zur 1-Clusterung immer maximal gewertet wird, ein gravierender Nachteil des Maßes zu sein.

Fred & Jain

In [FJ03] benutzen Fred und Jain ebenfalls eine Normierung der Korrelationsinformation. Diese unterscheidet sich nur geringfügig von der von Strehl & Gosh:

$$\mathcal{FJ}(\mathcal{C},\mathcal{C}') := \frac{2\mathcal{I}(\mathcal{C},\mathcal{C}')}{\mathcal{H}(\mathcal{C})+\mathcal{H}(\mathcal{C}')}$$

Wie bei Strehl & Gosh bleibt unerwähnt, wie mit der Möglichkeit, dass der Nenner Null ergeben kann, umgegangen werden soll. Allerdings passiert das bei diesem Maß nur, wenn $\mathcal{C}$ und $\mathcal{C}'$ die 1-Clusterung sind. Diese Clusterungen sind offensichtlich gleich. Analog zum Maß von Strehl & Gosh wird folgendes Maß als Abstandsmaß genutzt:

$$\mathcal{FJ}'(\mathcal{C},\mathcal{C}') := \begin{cases} 1-\mathcal{FJ}(\mathcal{C},\mathcal{C}') & \text{für } \mathcal{H}(\mathcal{C})+\mathcal{H}(\mathcal{C}')\neq 0 \\ 0 & \text{sonst} \end{cases}$$

Von diesem Maß sind die Abstände zwischen den vier Clusterungen in Tabelle 7.21 angegeben.

$\mathcal{FJ}'(\mathcal{C}^s,\mathcal{C}^1)$	$\mathcal{FJ}'(\mathcal{C}^s,\mathcal{C}^\times)$	$\mathcal{FJ}'(\mathcal{C}^\times,\mathcal{C}^1)$	$\mathcal{FJ}'(\mathcal{C}^\times,\mathcal{C}^\perp)$
1.0	0.333	1.0	1.0

Tabelle 7.21: Abstände des Maßes von Fred & Jain für $n = 1024$

Durch die andere Normalisierung gibt es im Vergleich zum Maß von Strehl & Gosh keine Veränderung der erfüllten Axiomen. Tabelle 7.22 verdeutlicht dies.

Dieses Maß ist dem Maß von Strehl & Gosh vorzuziehen, da es gegenüber Strehl & Gosh keine Nachteile besitzt, dabei aber nicht den Abstand zur 1-Clusterung immer mit Eins bewertet.

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{FJ}(\mathcal{C},\mathcal{C}')$	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–

Tabelle 7.22: Axiome des Maßes von Fred & Jain

Variation der Information

Meila führt in [Mei03] ein weiteres Maß ein, welches auf der Korrelationsinformation basiert. Dieses Maß ist allerdings nicht normiert:

$$\begin{aligned}\mathcal{VI}(\mathcal{C},\mathcal{C}') &:= \mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}') - 2\mathcal{I}(\mathcal{C},\mathcal{C}') \\ &= (\mathcal{H}(\mathcal{C}) - \mathcal{I}(\mathcal{C},\mathcal{C}')) + (\mathcal{H}(\mathcal{C}') - \mathcal{I}(\mathcal{C},\mathcal{C}'))\end{aligned}$$

Meila stellt in [Mei05] eine sehr detaillierte Analyse ihres Maßes vor, in der sie unter anderem zeigt, dass ihr Maß die Axiome der Verbandstheorie erfüllt.

Normierung Meila zeigt auch, das $\mathcal{VI}(\mathcal{C},\mathcal{C}') \leq \log_2(n)$ gilt. Somit kann man die Variation der Information mit $\log_2(n)$ normieren. Dies ergibt:

$$\mathcal{NVI}(\mathcal{C},\mathcal{C}') := \frac{\mathcal{VI}(\mathcal{C},\mathcal{C}')}{log_2(n)}$$

Mit der Normierung ergibt es die gemessenen Abstände der vier vorgegebenen Clusterungen entsprechend Tabelle 7.23.

$\mathcal{NVI}(\mathcal{C}^s,\mathcal{C}^1)$	$\mathcal{NVI}(\mathcal{C}^s,\mathcal{C}^\times)$	$\mathcal{NVI}(\mathcal{C}^\times,\mathcal{C}^1)$	$\mathcal{NVI}(\mathcal{C}^\times,\mathcal{C}^\perp)$
1.0	0.5	0.5	1.0

Tabelle 7.23: Abstände der normierten Variation der Information für $n = 1024$

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
$\mathcal{VI}(\mathcal{C},\mathcal{C}')$	✓	–	–	–	✓	✓	✓	–	✓	✓	✓	✓
$\mathcal{NVI}(\mathcal{C},\mathcal{C}')$	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓

Tabelle 7.24: Axiome der normierten und nichtnormierten Variation der Information

Für einen knotenstrukturellen Vergleich scheint dieses Maß vom axiomatischen Standpunkt aus allen anderen Maßen überlegen zu sein. Allerdings soll an dieser Stelle auf Kapitel 6 verwiesen werden, in welchem die Nachteile des Verbandsansatzes gezeigt wurden.

7.2.4 Anmerkungen zu den knotenstrukturellen Maßen

Die Entropiemaße scheinen ein vielversprechender Ansatz für den knotenstrukturellen Vergleich zu sein. Allerdings ist allen kontenstrukturellen Maßen gemein, dass sie nur für den statischen Clusterungsvergleich definiert sind. Eine Verwendung von knotenstrukturellen Maßen für einen dynamischen Clusterungsvergleich erscheint aber auch wenig sinnvoll, da im dynamischen Fall das Ignorieren der Kantenmenge der verschiedenen Graphen ein weitaus größeren Nachteil darstellt als im statischen Fall.

7.3 Graphstrukturelle Abstandsmaße

Dieser Abschnitt beinhaltet zum größten Teil die graphstrukturellen Erweiterungen der in Kapitel 7.2 vorgestellten Maße. Eine Vorgabe für die Erweiterung ist, dass für $G = K_n$ die graphstrukturellen Maße den gleichen Wert wie die entsprechenden knotenstrukturellen Versionen liefern sollen. Im Wesentlich wird für die graphstrukturelle Erweiterung der Maße die in Kapitel 2 vorgestellten kantenbasierten Erweiterungen der Ansätze Paarzählung, Schnittmengen und Entropie genutzt. Desweiteren werden noch zwei neue, auf den ersten Blick vielversprechende, Maße vorgestellt.

Auf eine axiomatische Analyse der graphstrukturellen Maße wird verzichtet, da sich darüber bisher – wie über die Maße an sich – in der Literatur keine Informationen finden lassen und die exakte Überprüfung der einzelnen Axiome den Rahmen dieser Arbeit sprengen würde. Außerdem ist beispielsweise die qualitative Sensivität stark abhängig von dem verwendeten Index.

In diesem Kapitel wird davon ausgegangen, dass jeder Knoten mindestens den Knotengrad Eins hat und somit die Kantenmenge des Graphen nicht leer ist.

7.3.1 Graphstrukturelle Paarmaße

Bei der lokalen Paarzählung werden nun für die Abstandsmaße nicht die globalen, sondern die lokalen Paarzählungsmengen als Grundlage genutzt. Da die lokalen Paarzählungsmengen die Kanten des Graphen nutzen, handelt es sich somit um graphstrukturelle Abstandsmaße. Ferner gilt aufgrund von Lemma 4, dass für $G = K_n$ die graphstrukturellen den knotenstrukturellen Abstandsmaßen entsprechen.

Durch die Verwendung der lokalen Paarzählungsmengen sind diese Maße nur für den statischen Clusterungsvergleich geeignet.

Rand–Maß (g)

Die graphstrukturelle Version des Rand–Maßes ist definiert durch:

$$\mathcal{R}_g(\mathcal{C},\mathcal{C}') := \frac{e_{11}+e_{00}}{m}$$

Die Abstandsversion des Maßes ist analog mit $\mathcal{R}'_g := 1-\mathcal{R}_g$ festgelegt. Da auch dieses Maß zumindestenst für vollständige Graphen die gleichen Nachteile wie die knotenstrukturelle Version besitzt, ist hier eine Anpassung des Maßes ebenfalls sinnvoll.

Angepasstes Rand–Maß (g)

Das angepasste Rand–Maß kann man auch in ein graphstrukturelles Maß tranferieren. Die Variablen t_1 und t_2 sind im vollständigen Graphen alle Kanten innerhalb der beiden Clusterungen. Die graphstrukturelle Version nutzt deshalb $t_1 := m(\mathcal{C})$ und $t_2 := m(\mathcal{C}')$. Analog zur knotenstrukturellen Variante wird t_3 definiert.

$$\mathcal{AR}_g(\mathcal{C},\mathcal{C}') := \frac{e_{11}-t}{\frac{1}{2}(m(\mathcal{C})+m(\mathcal{C}'))-t_3} \quad \text{mit} \quad t_3 := \frac{m(\mathcal{C})m(\mathcal{C}')}{m}$$

Für die Experimente wird erneut die Abstandsversion $\mathcal{AR}'_g$ genutzt.

Fowlkes–Mallows (g)

Durch Ersetzen von n_{ab} durch e_{ab} erhält man die graphstrukturelle Version des Fowlkes–Mallows–Maßes. Allerdings wird auch hier folgende Abstandsversion des Maßes genutzt:

$$\mathcal{FM}'_g(\mathcal{C},\mathcal{C}') := \begin{cases} 1-\frac{e_{11}}{\sqrt{(e_{11}+e_{10})(e_{11}+e_{01})}} & \text{für } e_{01},e_{10}\neq 0 \vee e_{11}\neq 0 \\ 1 & \text{für } e_{11},e_{01}=0 \vee e_{11},e_{10}=0 \\ 0 & \text{sonst} \end{cases}$$

Es ist zu beachten, dass die Sonderfälle im graphstrukturellen Fall nicht nur für Singleton-Clusterungen auftreten. Wenn eine Clusterung keine Interclusterkanten besitzt, gilt entweder $e_{01}=0$ oder $e_{10}=0$.

Mirkin–Metrik (g)

Auch die graphstrukturelle Variante entspricht der nicht normierten Abstandsversion des graphstrukturellen Rand–Maßes. Somit gilt:

$$\begin{aligned} \mathcal{M}_g(\mathcal{C},\mathcal{C}') &:= 2(e_{01}+e_{10}) \\ &= 2m\mathcal{R}'_g(\mathcal{C},\mathcal{C}') \end{aligned}$$

Aus den gleichen Gründen wie bei der knotenstrukturellen Version wird auf eine tiefere Analyse dieses Maßes verzichtet.

Jaccard (g)

Bei der graphstrukturellen Variante des Jaccard–Maßes werden lediglich die Variablen n_{ab} durch e_{ab} ersetzt. Dies ergibt:

$$\mathcal{J}'_g(\mathcal{C}, \mathcal{C}') := \begin{cases} 1 - \frac{e_{11}}{e_{11}+e_{10}+e_{01}} & \text{für } e_{11} + e_{10} + e_{01} \neq 0 \\ 0 & \text{sonst} \end{cases}$$

Ebenso wie bei der graphstrukturellen Version des Fowlkes–Mallows–Maßes kann der Sonderfall nicht nur für 1-Clusterungen auftreten. Falls weder $\mathcal{C}$ noch $\mathcal{C}'$ Interclusterkanten besitzen, gilt $e_{00} = |E|$ und somit $e_{11} + e_{10} + e_{01} = 0$.

Partitionsdifferenz (g)

Der Vollständigkeit wegen wird hier noch die graphstrukturelle Version der unbrauchbaren Partitionsdifferenz aufgeführt.

$$\mathcal{PD}_g(\mathcal{C}, \mathcal{C}') := e_{00}$$

Die Nachteile der knotenstrukturelle Variante gelten auch hier. Daher wird auch auf eine Untersuchung dieser Version des Maßes verzichtet.

7.3.2 Graphstrukturelle Schnittmaße

Bei diesen Maßen wird anstatt der normalen Verschmelzungsmatrix die gewichtete Version der Matrix verwendet. Das hat zur Folge, dass diese Maße für reguläre Graphen keine Vorteile gegenüber den jeweiligen knotenstrukturellen Versionen besitzen. Da nicht für alle Maße sofort ersichtlich ist, dass für $G = K_n$ die graphstrukturellen Versionen den knotenstrukturellen entsprechen, wird dies gegebenenfalls gezeigt.

F–Maß (g)

Analog zum knotenstrukturellen F–Maß wird auch hier die gewichtete Summe der maximalen F–Maße der Cluster in $\mathcal{C}$ berechnet. Dabei unterscheidet sich das F–Maß zweier Cluster und die Normierung von der knotenstrukturellen Version.

Das graphstrukturelle F–Maß zweier Cluster berechnet sich mit:

$$\mathcal{F}_g(C_i, C'_j) := \frac{2m^d_{ij}}{\sum_{v \in C_i} \deg(v) + \sum_{v \in C'_j} \deg(v)}$$

Bei der graphstrukturellen Version wird nun anstatt mit n mit $2m$ normiert und $|C_i|$ durch $\sum_{v \in C_i} \deg(v)$ ersetzt.

$$\mathcal{F}_g(\mathcal{C}, \mathcal{C}') := \frac{1}{2m} \sum_{i=1}^{k} (\sum_{v \in C_i} \deg(v)) \max_{j=1}^{l} \{\mathcal{F}_g(C_i, C'_j)\}$$

Auch dieses Maß ist offensichtlich nicht symmetrisch. Nun soll noch gezeigt werden, dass für einen d-regulären Graphen G die Gleichheit von graph- und knotenstruktureller Version gilt.

$$\begin{aligned} \mathcal{F}_g(C_i, C'_j) &= \frac{2m_{ij}^d}{\sum_{v \in C_i} \deg(v) + \sum_{v \in C'_j} \deg(v)} \\ &= \frac{2dm_{ij}}{d|C_i| + d|C'_j|} \\ &= \mathcal{F}(C_i, C'_j) \end{aligned}$$

Daraus folgt für das F–Maß:

$$\begin{aligned} \mathcal{F}_g(\mathcal{C}, \mathcal{C}') &= \frac{1}{dn} \sum_{i=1}^{k} d|C_i| \max_{j=1}^{l} \{\mathcal{F}(C_i, C'_j)\} \\ &= \mathcal{F}(\mathcal{C}, \mathcal{C}') \end{aligned}$$

Da der vollständige Graph ebenfalls regulär ist, erfüllt diese Erweiterung die Vorgabe.

Meila–Heckerman (g)

Die Erweiterung des Meila–Heckerman–Maßes ist analog zu der Erweiterung des F–Maßes.

$$\mathcal{MH}_g(\mathcal{C}, \mathcal{C}') := \frac{1}{2m} \sum_{i=1}^{k} \max_{C'_j \in \mathcal{C}'} m_{ij}^d$$

Für d-reguläre Graphen gilt hier ebenso die Gleichheit der graph- und knotenstrukturellen Variante des Maßes.

Maximum Match (g)

Wie bei den vorangegangen Maßes ist die Erweiterung des Maximum–Match–Maßes einfach:

$$\mathcal{MM}_g(\mathcal{C}, \mathcal{C}') := \frac{1}{2m} \sum_{i=1}^{\min\{k,l\}} m_{ii'}^d$$

Analog gilt für reguläre Graphen $\mathcal{MM} = \mathcal{MM}_g$.

Van Dongen (g)

Als graphstrukturelles van Dongen–Maß wird definiert:

$$\mathcal{VD}_g(\mathcal{C},\mathcal{C}') := 2n - \frac{n}{2m}\sum_{i=1}^{k}\max_j m_{ij}^d - \frac{n}{2m}\sum_{j=1}^{l}\max_i m_{ij}^d$$

Mit $m_{ij}^d = dm_{ij}$ und $2m = dn$ für d-reguläre Graphen folgt die Gleichheit von knoten- und graphstruktureller Version des Maßes.

Normierung Analog zur knotenstrukturellen Version kann das graphstrukturelle van Dongen–Maß ebenso mit $2n$ normiert werden.

$$\mathcal{NVD}_g(\mathcal{C},\mathcal{C}') := \frac{\mathcal{VD}_g(\mathcal{C},\mathcal{C}')}{2n}$$

Somit folgt für reguläre Graphen erneut $\mathcal{NVD}_g = \mathcal{NVD}$.

7.3.3 Kantenentropiemaße

Bei den Kantenentropieversionen der Entropiemaße wird anstatt der Entropie die Kantenentropie als Grundlage der Berechnung verwendet. Nach Lemmata 5 und 6 gilt für reguläre Graphen sowohl $\mathcal{H}(\mathcal{C}) = \mathcal{H}_E(\mathcal{C})$ als auch $\mathcal{I}(\mathcal{C}) = \mathcal{I}_E(\mathcal{C})$.

Strehl & Ghosh (g)

Durch Ersetzen von Entropie und Korrelationsinformation ist die graphstrukturelle Version des Maßes von Strehl & Gosh formal definiert durch:

$$\mathcal{SG}'_g(\mathcal{C},\mathcal{C}') := \begin{cases} 1 - \frac{\mathcal{I}_E(\mathcal{C},\mathcal{C}')}{\sqrt{\mathcal{H}_E(\mathcal{C})\mathcal{H}_E(\mathcal{C}')}} & \text{für } \mathcal{H}_E(\mathcal{C}), \mathcal{H}_E(\mathcal{C}') \neq 0 \\ 0 & \text{für } \mathcal{H}_E(\mathcal{C}) = \mathcal{H}_E(\mathcal{C}') = 0 \\ 1 & \text{sonst} \end{cases}$$

Nach Lemmata 5 und 6 gilt für reguläre Graphen die Gleichheit der graph- und knotenstrukturellen Version des Maßes.

Fred & Jain (g)

Die graphstrukturelle Version des Fred & Jain Maßes ist analog festgelegt:

$$\mathcal{FJ}'_g(\mathcal{C},\mathcal{C}') := \begin{cases} 1 - \frac{2\mathcal{I}_E(\mathcal{C},\mathcal{C}')}{\mathcal{H}_E(\mathcal{C})+\mathcal{H}_E(\mathcal{C}')} & \text{für } \mathcal{H}_E(\mathcal{C}) + \mathcal{H}_E(\mathcal{C}') \neq 0 \\ 0 & \text{sonst} \end{cases}$$

Auch hier gilt für reguläre Graphen $\mathcal{FJ}_g = \mathcal{FJ}$.

Variation der Information (g)

Die graphstrukturelle Variante dieses Maßes nutzt ebenso lediglich die Kantenentropie bzw. Kantenkorrelationsinformation.

$$\begin{aligned} \mathcal{VI}_g(\mathcal{C},\mathcal{C}') &:= \mathcal{H}_E(\mathcal{C}) + \mathcal{H}_E(\mathcal{C}') - 2\mathcal{I}_E(\mathcal{C},\mathcal{C}') \\ &= (\mathcal{H}_E(\mathcal{C}) - \mathcal{I}_E(\mathcal{C},\mathcal{C}')) + (\mathcal{H}_E(\mathcal{C}') - \mathcal{I}({}_E\mathcal{C},\mathcal{C}')) \end{aligned}$$

Somit gilt für reguläre Graphen die Gleichheit zwischen knoten- und graphstruktureller Version.

Normierung Nach Lemma 6 gilt die Beziehung $\mathcal{I}_E(\mathcal{C},\mathcal{C}') \leq \log_2(2m)$. Meila zeigt allerdings in [Mei03], dass für gewichtete Partition – jedem Element ist ein Gewicht w mit $0 \leq w \leq 1$ zugeordnet – ebenso $\mathcal{VI} \leq \log_2(n)$ gilt. Da für die $P_E(i)$ der Kantenentropie auch $0 \leq P_E(i) \leq 1$ gilt, kann man die graphstrukturelle Variation der Information ebenfalls mit $\log_2(n)$ normieren.

$$\mathcal{NVI}_g(\mathcal{C},\mathcal{C}') := \frac{\mathcal{VI}_g(\mathcal{C},\mathcal{C}')}{log_2(n)}$$

Auch mit dieser Normierung des Maßes gilt für reguläre Graphen weiterhin die Gleichheit zwischen knoten- und graphstruktureller Version.

7.3.4 Anmerkungen zu den graphstrukturellen Erweiterungen

Der Vorteil dieser Erweiterungen ist die leichte Transferierung der Maße in graphstrukturelle Versionen. Allerdings sind die Schnitt- und Entropiemaße bereits für reguläre Graphen gleich. Das bedeutet, dass in diesem Fall die graphstrukturellen Erweiterungen ohne Wirkung bleiben. Dass solche Fälle auftreten können, wurde in Kapitel 6.3 gezeigt.

Außerdem sind sämtliche Erweiterungen nur für den statischen Clusterungsvergleich geeignet, da die lokalen Paarzählungsmengen, die gewichtete Verschmelzungsmatrix und auch die Kantenkorrelationsinformation die Gleichheit der beiden zugrundeliegenden Graphen voraussetzen.

7.3.5 Editiermengendifferenz

Die Idee der Editiermengendifferenz ist, dass für gleiche Clusterungen und gleiche Graphen die beide Clustereditiermengen $F_{\mathcal{C}}, F_{\mathcal{C}'}$ übereinstimmen. Desweiteren gilt für die graphstrukturelle Gleichheit von $\mathcal{C}$ und $\mathcal{C}'$:

$$F_{\mathcal{C}} = F_{\mathcal{C}'} = F_{\mathcal{C}} \cup F_{\mathcal{C}'} = F_{\mathcal{C}} \cap F_{\mathcal{C}'}$$

Die Editiermengendifferenz berechnet sich nun aus der Differenz der Kardinalitäten der Vereinigung der Clustereditiermengen und dem Schnitt der beiden.

$$\mathcal{EMD}(\mathcal{C}, \mathcal{C}') := |F_{\mathcal{C}} \cup F_{\mathcal{C}'}| - |F_{\mathcal{C}} \cap F_{\mathcal{C}'}|$$

Da man für die Berechnung des Schnittes bzw. der Vereinigung der Mengen sowohl die Clusterung als auch die zugrundeliegenden Graphen benötigt, handelt es sich hierbei um ein graphstrukturelles Maß.

Normierung Die Editiermengendifferenz ist offensichtlich nicht 1-beschränkt. Daher erscheint zur besseren Vergleichbarkeit mit den anderen Maßen eine Normierung sinnvoll zu sein. Da $\mathcal{EMD}(\mathcal{C}, \mathcal{C}') \leq |F_{\mathcal{C}} \cup F_{\mathcal{C}'}|$ gilt, ist folgende normierte Version 1-beschränkt.

$$\begin{aligned} \mathcal{NEMD}_{\cup}(\mathcal{C}, \mathcal{C}') &:= \frac{|F_{\mathcal{C}} \cup F_{\mathcal{C}'}| - |F_{\mathcal{C}} \cap F_{\mathcal{C}'}|}{|F_{\mathcal{C}} \cup F_{\mathcal{C}'}|} \\ &= 1 - \frac{|F_{\mathcal{C}} \cap F_{\mathcal{C}'}|}{|F_{\mathcal{C}} \cup F_{\mathcal{C}'}|} \end{aligned}$$

Der Nachteil dieser *vereinigungsnormierten* Editiermengendifferenz ist, dass für den Fall $|F_{\mathcal{C}} \cup F_{\mathcal{C}'}| = 2$ und $|F_{\mathcal{C}} \cap F_{\mathcal{C}'}| = 1$ ein Abstand von 0.5 ausgegeben wird, und zwar unabhängig von der Anzahl der Kanten des Graphen. Dies kann man umgehen, indem man mit der Kantenzahl des Graphen nomiert.

$$\mathcal{NEMD}_{E}(\mathcal{C}, \mathcal{C}') := \frac{|F_{\mathcal{C}} \cup F_{\mathcal{C}'}| - |F_{\mathcal{C}} \cap F_{\mathcal{C}'}|}{m}$$

Diese *Kantennormierung* hat allerdings den Nachteil, dass unter Umständen die so normierte Editiermengendifferenz einen Wert größer als Eins annehmen kann. Es zeigt sich sogar, dass dieser Fall sehr häufig eintritt.

Um diesen Nachteil auszugleichen, kann man die Editiermengendifferenz mit der möglichen Maximalzahl an Kanten im Graphen normieren. Das ergibt folgende Normierung:

$$\mathcal{NEMD}_{V}(\mathcal{C}, \mathcal{C}') := \frac{2(|F_{\mathcal{C}} \cup F_{\mathcal{C}'}| - |F_{\mathcal{C}} \cap F_{\mathcal{C}'}|)}{n(n-1)}$$

Bei dieser *Knotennormierung* wird nur dann ein maximaler Abstand gemessen, wenn die Editiermengen komplementär zueinander sind und die Vereinigung der beiden Mengen der Kantenmenge des vollständigen Graphen entsprechen. Für Verfeinerungen $\mathcal{C} \subseteq \mathcal{C}'$ gilt außerdem:

$$\begin{aligned} \mathcal{NEMD}_{V}(\mathcal{C}, \mathcal{C}') &= \frac{2(|F'_{\mathcal{C}}| - |F_{\mathcal{C}}|)}{n(n-1)} \\ &= per(\mathcal{C}) - per(\mathcal{C}') \end{aligned}$$

Das bedeutet, dass dieses Maß für Verfeinerungen bzw. Vergröberungen der Indexdifferenz bzgl. Performance entspricht.

Anmerkung Die Editiermengendifferenz fordert keine Gleichheit der Graphen. Allerdings fordert die Kantennormierung die Gleichheit der Kantenmenge, die Knotennormierung die Gleichheit der Knotenmenge. Da sich auf den statischen Clusterungsvergleich konzentriert wird, ist dies zunächst kein Nachteil. Allerdings sollte dies bei künftigen Betrachtungen des dynamischen Clusterungsvergleiches bedacht werden.

7.3.6 Strukturelles Indexmaß

Dieses Maß greift die Idee von Axiom 19 auf, welches aussagt, dass die gleiche knotenstrukturelle Änderung einer Clusterung auf signifikanten Clusterungen zu einem größeren Abstand als auf weniger signifikanten Clusterungen führt. Nun könnte man den graphstrukturellen Abstand als Mittelwert zwischen dem qualitativen und knotenstrukturellem Abstand interpretieren:

$$d_g(\mathcal{C}, \mathcal{C}') = \frac{d_q(\mathcal{C}, \mathcal{C}') + d_k(\mathcal{C}, \mathcal{C}'')}{2}$$

Das Strukturelle Indexmaß bildet nun lediglich den Mittelwert aus zwei gemessen Abständen, wobei $\mathcal{QM}$ ein beliebiges qualitatives und $\mathcal{KM}$ ein beliebiges knotenstrukturelles Abstandsmaß sein soll.

$$\mathcal{SIM}(\mathcal{C}, \mathcal{C}', \mathcal{QM}, \mathcal{KM}) := \frac{\mathcal{QM}(\mathcal{C}, \mathcal{C}') + \mathcal{KM}(\mathcal{C}, \mathcal{C}')}{2}$$

Es ist offensichtlich, dass der so gemessene Abstand von der Wahl der jeweiligen Maße und des Indexes für die Berechnung des qualitativen Abstands stark abhängig ist.

Da die Entropiemaße der vielversprechenste Ansatz für den knotenstrukturellen Vergleich von Clusterungen zu sein scheint, soll ein Hauptaugenmerk bei den Experimenten in Kapitel 8 auf den bewerteten Entropiemaßen liegen. Dabei sollen

$$\mathcal{FJ}_{\mathcal{ID}}(\mathcal{C}, \mathcal{C}') := \mathcal{SIM}(\mathcal{C}, \mathcal{C}', \mathcal{ID}, \mathcal{FJ})$$

das *bewertete Fred & Jain* Maß und analog

$$\mathcal{NVI}_{\mathcal{ID}}(\mathcal{C}, \mathcal{C}') := \mathcal{SIM}(\mathcal{C}, \mathcal{C}', \mathcal{ID}, \mathcal{NVI})$$

die *bewertete normalisierte Variation der Information* sein.

Natürlich sind noch weitere Kombinationen denkbar, die je nach Anwendungsfall mehr oder minder sinnvoll sind.

8 Experimente

In diesem Kapitel werden nun die Experimente des Verhaltens der in Kapitel 7 diskutierten Abstandsmaße vorgestellt. Dabei wurden vier Szenarios untersucht:

Initial- und Zufallsclusterungen (Kapitel 8.1) Bei diesen Experimenten werden zwei Vergleiche gegenübergestellt werden. Zum einen wird der Abstand zwischen der Initialclusterung eines Attraktors und einer Zufallsclusterung mit gleicher erwarteter Clusteranzahl gemessen. Zum anderen der Abstand zweier Zufallsclusterungen zueinander. Ziel ist hierbei die Verdeutlichung, dass knotenstrukturelle Maße keinen Unterschied zwischen diesen Abständen messen, wohingegen qualitative und graphstrukturelle Maße unterschiedliche Abstände messen.

Initial- und Algorithmenclusterungen (Kapitel 8.2) Bei diesen Testreihen wird die Initialclusterung eines Gaußgenerators mit der Clusterung, die der MCL–Algorithmus auf dem gleichen Graphen berechnet, verglichen. Dabei sollen drei Bereiche untersucht werden. In einem Berech stimmen die beiden Clusterungen überein, in den beiden andern besitzt die Algorithmusclusterung mehr bzw. weniger Cluster. Welche Auswirkungen das auf den gemessenen Abstand hat, soll Ziel dieser Testreihen sein.

Lokale Minimierung (Kapitel 8.3) Ähnlich zum ersten Szenario werden hier zwei Vergleiche gegenübergestellt, die sich knotenstrukturelle wenig unterscheiden, qualitativ hingegen stark. Hierzu werden die Initialclusterungen von Attraktorengraphen mit Clusterungen verglichen, die aus der Initialclusterung durch Verschieben einer steigenden Anzahl von Knoten entstehen.

Verfeinerung und Vergröberung (Kapitel 8.4) Diese Experimente dienen dem Ziel, wie die verschiedenen Maße Verfeinerungen bzw. Vergröberungen von Clusterungen bewerten. Hierzu werden die verschiedenen Clusterlevel der Hierarchiegraphen genutzt.

Für jeden dieser Tests wird der qualitative, knotenstrukturelle und graphstrukturelle Abstand getrennt betrachtet. Ziel dieser Experimente ist die Verdeutlichung, dass bisherige Lösungsansätze unzureichend sind. Ferner soll herausgearbeitet werden, in welchen Fällen die graphstrukturellen Erweiterungen der knotenstrukturellen Maße gute bzw. schlechte Ergenisse liefern.

Testumgebung Die in Kapitel 7 vorgstellten Maße und Algorithmen bzw. Datenstrukturen aus Kapitel 2 wurden vollständig in Java implementiert. Dabei wurde auf der kommerziellen yFiles–Bibliothek [Y], einer mächtigen Bibliothek für Graphenalgorithmik, aufgebaut. Die Ergebnisse folgender Experimente basieren auf mindestens 50 Messungen und besitzen ein Vertrauensintervall von 0.1 bei einer statistischen Signifikanz von 95% [Düm03].

Anmerkung Da eine Betrachtung aller Varianten der Maße den Rahmen dieser Arbeit spränge, werden nur die jeweiligen normierten Maße und Ergebnisse vorgestellt. Desweiteren wurde zur besseren Übersichtlichkeit immer die Abstandsversion des Maßes genutzt.

8.1 Initial- und Zufallsclusterung

In dieser Testreihe sollen zwei statische Clusterungsvergleiche gegenübergestellt werden. Hierzu werden die Initialclusterung eines Attraktorengenerators und zwei Zufallsclusterungen auf dem gleichen Graphen genutzt. Die beiden Vergleiche lauten dann:

Initial- gegen Zufallsclusterung (IgZ) Gemessen wird der Abstand der Initialclusterung des Attraktors zu einer Zufallsclusterung. Dabei soll die erwartete Clusteranzahl der Zufallsclusterung mit der der Initialclusterung übereinstimmen.

Zufall- gegen Zufallsclusterung (ZgZ) Bei diesen Tests wird der Abstand zweier unabhängiger Zufallsclusterungen zueinander gemessen. Auch hier soll die jeweilige Clusteranzahl der beiden Zufallsclusterungen im Erwartungswert übereinstimmen.

Die Ziele der Testreihe sind:

- Bei dem ZgZ–Vergleich soll für die knotenstrukturellen Abstandsmaße überprüft werden, welche Maße den Abstand zweier unabhängiger Zufallsclusterungen mit Eins messen. Das ist eine mögliche Art, wie der maximale Abstand in Kapitel 4.3.3 festgelegt wurde.
- Da die Erwartungswerte der Clusteranzahl der Initialclusterung mit der der Zufallsclusterung übereinstimmen, kann man argumentieren, dass sich diese beiden Clusterungen knotenstrukturell nicht unterscheiden, qualitativ hingegen sehr. Dies bedeutet, dass der knotenstrukturelle Abstand des IgZ–Vergleiches mit dem des ZgZ–Vergleiches übereinstimmt, der qualitative bzw. graphstrukturelle hingegen nicht. Inwieweit sich diese Argumentation in den Abstandsmaßen wiederfindet, ist ebenfalls Ziel dieser Auswertung.

8.1.1 Setup

Für die Generierung der Inititalclusterungen wurden Attraktorengeneratoren mit $n = 1000$ Knoten verwendet. Damit ist die Anzahl der Cluster der Initialclusterung zufällig zwischen 2 und $\sqrt[3]{1000} = 10$ (siehe Abschnitt 2.5.1). Die Dichte f des Graphen wurde nun sukzessiv von 0.05 beginnend mit einer Schrittweite von 0.05 bis auf $f = 5.0$ erhöht.

Die Zufallsclusterungen wurden mit Hilfe des Zufallclusterers (Algorithmus 4 aus Abschnitt 2.5.2) auf dem gleichen Graphen erzeugt. Dabei wurde die Clusterzahl analog zur Initialclusterung zufällig zwischen 2 und 10 gewählt.

Abbildung 8.1 zeigt die Indizes für eine Initial- und Zufallsclusterung mit zunehmender Dichte f des Graphen.

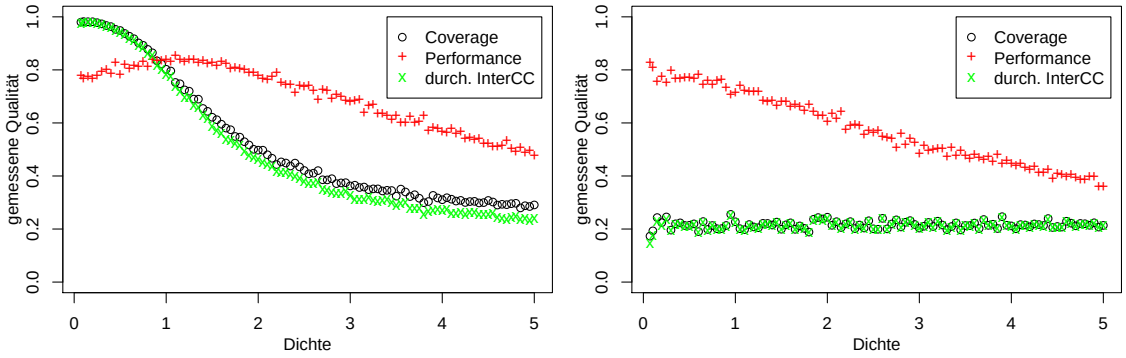


Abb. 8.1: Indizes von Initial-(links) und Zufallsclusterung(rechts)

Hierbei ist auffällig, dass Performance die Intuition für die Bewertung der Initialclusterung am besten wiederspiegelt. Mit zunehmender Dichte nimmt die Qualität der Clusterung zuerst zu, da die Cluster bei sehr kleinen Dichten nur wenige Intraclusterkanten besitzen. Bei weiter ansteigender Dichte nimmt die Zahl der Interclusterkanten zu, sodass die Signifikanz der Clusterung abnimmt. Bei den Zufallsclusterungen hingegen zeigt Performance ein wenig intuitives Verhalten: Eine solche Clusterung sollte nicht ähnlich hoch wie die Initialclusterung bewertet werden.

8.1.2 Qualitativer Abstand

Für die Messung des qualitativen Abstands wurden die qualitativen Maße Indexquotient, Indexdifferenz, gut–, schlecht– und beidlastige Indexdifferenz genutzt. Als Index wurde der Durchschnitt von Coverage, Performance und durchschnittlicher Interclusterconductance verwendet. Abbildung 8.2 zeigt die gemessenen Abstände der Maße, wobei die linke Grafik den Abstand beim IgZ– und die rechte Grafik beim ZgZ–Vergleich zeigt.

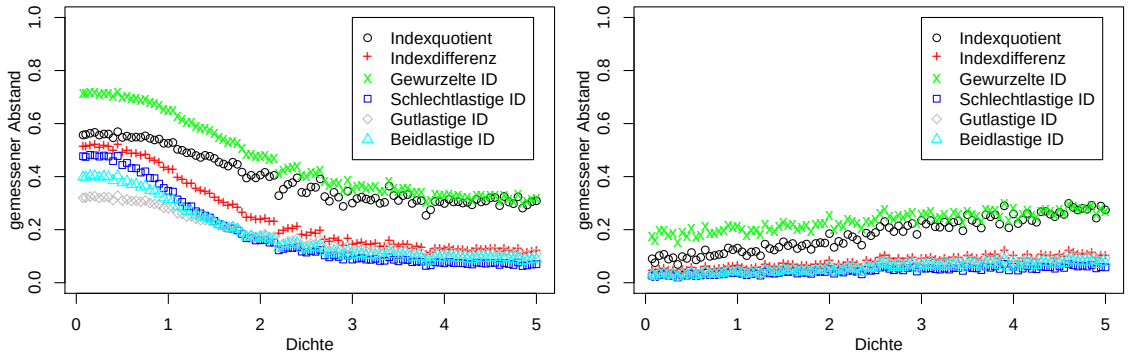


Abb. 8.2: Qualitative Maße beim IgZ–(links) und ZgZ–Vergleichen(rechts)

Die qualitativen Maße verhalten sich wie erwartet. Der qualitative Abstand zweier Zufallsclusterungen zueinander ist Nahe Null, wohingegen der IgZ–Abstand bei kleinen Dichten durch die hohe Signifikanz der Initialclusterung größer ist.

8.1.3 Knotenstruktureller Abstand

Als Nächstes soll der knotenstrukturellen Abstand betrachtet werden. Zur besseren Übersicht sind Paar-, Schnitt- und Entropiemaße getrennt voneinander aufgeführt.

Paarmaße

Abbildung 8.3 zeigt die gemessenen Abstände des Rand–, angepassten Rand–, Fowlkes–Mallows– und Jaccard–Maßes. Erneut repräsentieren die linken Grafiken den IgZ-Vergleich und die rechten den ZgZ-Vergleich.

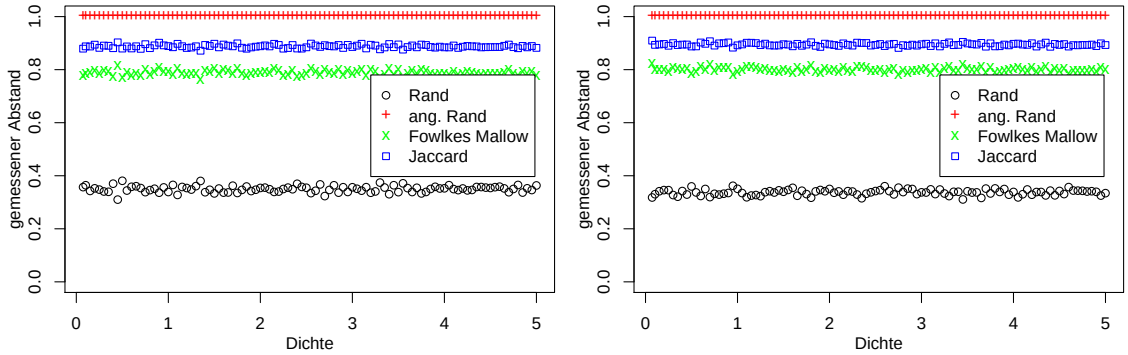


Abb. 8.3: Knotenstrukturelle Paarmaße beim IgZ– und ZgZ–Vergleich

Man erkennt, dass alle drei Maße keinen Unterschied zwischen IgZ– und ZgZ–Vergleich messen können. Außerdem misst nur das angepasste Rand–Maß einen Abstand von Eins. Einen auffällig kleinen Abstands misst vor allem das Rand–Maß.

Schnittmaße

Abbildung 8.4 zeigt die gemessenen Abstände des F–, Meila–Heckermann–, Maximum–Match– und van Dongen–Maßes. Das F– und Meila–Heckermann–Maß sind beide asymmetrisch, sodass in der Grafik die Maße jeweils zweimal enthalten sind.

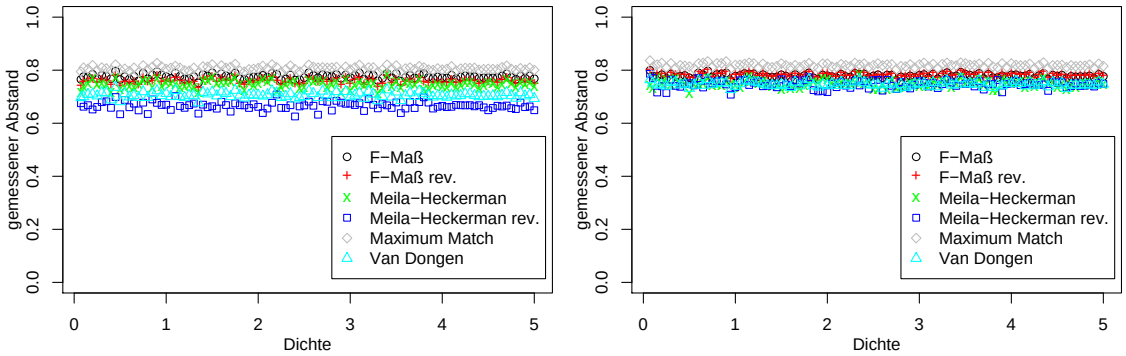


Abb. 8.4: Knotenstrukturelle Schnittmaße beim IgZ– und ZgZ–Vergleich

Auch die knotenstrukturellen Schnittmaße unterscheiden nicht signifikant zwischen IgZ– und ZgZ–Vergleich. Alle Maße weisen außerdem ungefähr den gleichen Abstand auf.

Entropiemaße

Abbildung 8.5 zeigt die gemessenen Abstände der Entropiemaße Fred & Jain, Strehl & Gosh und der normierten Variation der Information.

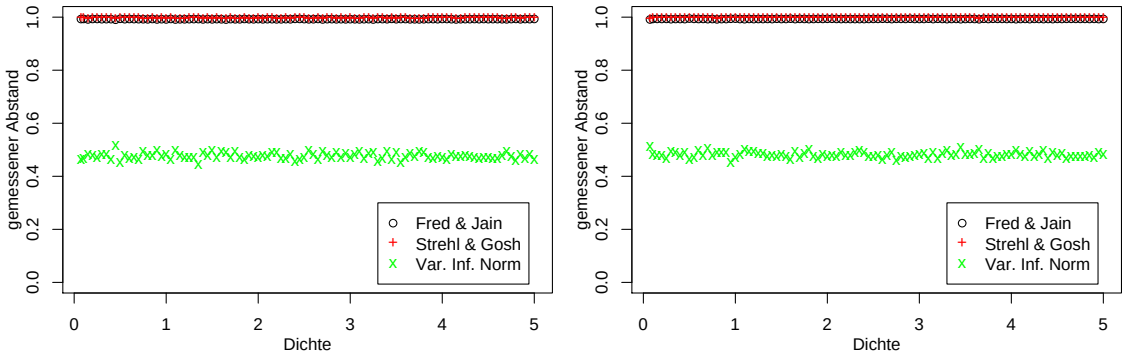


Abb. 8.5: Knotenstrukturelle Entropiemaße beim IgZ– und ZgZ–Vergleich

Wie bei den knotenstrukturellen Paar– und Schnittmaßen ist kein Unterschied zwischen dem IgZ– und ZgZ–Vergleich auszumachen. Neben dem angepassten Rand–Index sind die Maße von Fred & Jain bzw. Strehl & Gosh die einzigen knotenstrukturellen Maße, die den Abstand maximal bewerten. Der Variantion der Information Index erreicht durch die Normierung nur einen Wert von 0.5.

8.1.4 Graphstruktureller Abstand

Graphstrukturelle Maße sollten einen unterschiedlichen Abstand beim IgZ– und ZgZ–Vergleich messen. Bei dem IgZ–Vergleich sollte der Abstand mit zunehmender Dichte abnehmen, der ZgZ–Vergleich sollte ähnlich zu dem gemessenen Abstand der knotenstrukturellen Maße aussehen.

Paarmaße

Abbildung 8.6 zeigt die gemessenen Abstände der jeweiligen graphstrukturellen Version des Rand-, angepassten Rand-, Fowlkes–Mallows- und Jaccard–Maßes.

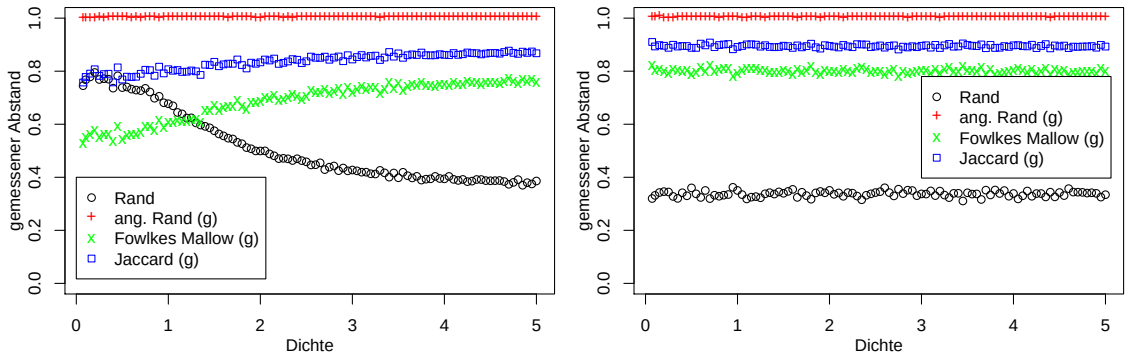


Abb. 8.6: Graphstrukturelle Paarmaße beim IgZ– und ZgZ–Vergleich

Nur das Rand–Maß zeigt das erwartete Verhalten beim IgZ–Vergleich. Das Fowlkes–Mallows- und Jaccard–Maß bewerten den Abstand mit zunehmender Dichte sogar größer, wohingegen das angepasste Rand–Maß immer den maximalen Abstand misst. Beim ZgZ–Vergleich messen die graphstrukturellen Versionen den gleichen Abstand wie die knotenstrukturellen Versionen.

Schnittmaße

Abbildung 8.7 zeigt die gemessenen Abstände der graphstrukturellen Versionen des F–, Meila–Heckermann–, Maximum–Match– und van Dongen–Maßes. Das F– und Meila–Heckermann–Maß sind auch in der graphstrukturellen Version asymmetrisch, sodass in der Grafik die Maße jeweils zweimal enthalten sind.

Die graphstrukturellen Versionen können – wie die knotenstrukturellen – nicht zwischen IgZ– und ZgZ–Vergleich unterscheiden. Außer für sehr kleine Dichten fällt bei einem Vergleich zu Abbildung 8.4 auf, dass die Werte der Maßen den knotenstrukturellen Versionen entsprechen.

Die Abweichung für sehr geringe Dichten lässt sich wohl damit erklären, dass in diesem Bereich der Graph nur Kanten zwischen den initialen Attraktorenknoten und

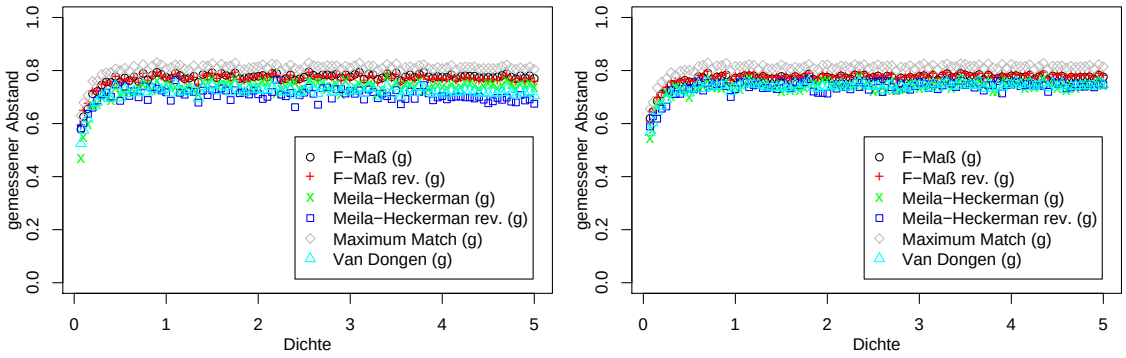


Abb. 8.7: Graphstrukturelle Schnittmaße beim IgZ– und ZgZ–Vergleich

den später hinzugefügten Knoten besitzt. Daher existieren einige wenige Knoten mit hohem Knotengrad (nämlich die Attraktorenknoten), die meisten anderen Knoten haben lediglich einen Knotengrad von Eins.

Entropiemaße

Abbildung 8.8 zeigt die gemessenen Abstände der Kantenentropiemaße Fred & Jain, Strehl & Gosh und der normierten Variation der Information.

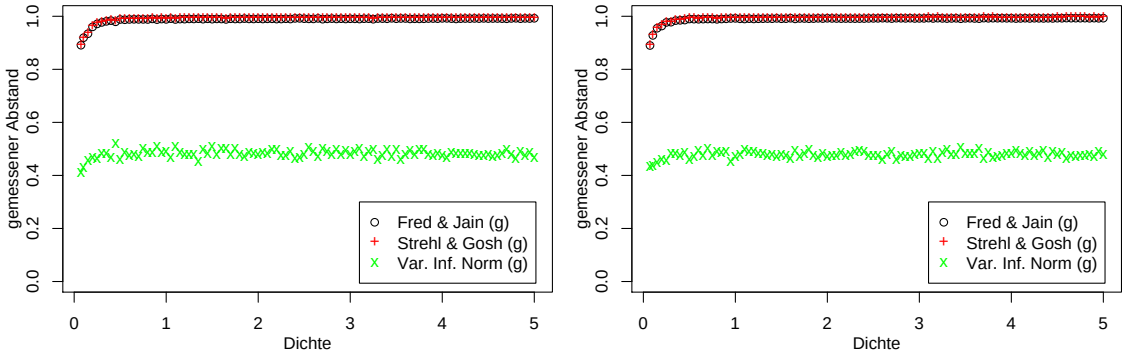


Abb. 8.8: Graphstrukturelle Entropiemaße beim IgZ– und ZgZ–Vergleich

Da nach Lemma 5 die Kantenentropie für reguläre Graphen der Entropie entspricht und die untersuchten Graphen annähernd regulär sind, ist offensichtlich, dass die Kantenentropiemaße den IgZ– und ZgZ–Vergleich gleich bewerten. Die Begründung für das Ansteigen des Abstands bei geringen Dichten ist der gleiche wie bei den Schnittmaßen.

Wie bei den knotenstrukturelllen Entropiemaßen messen die Maße von Strehl & Gosh und Fred & Jain den gleichen Abstand.

Editiermengendifferenz

Den gemessenen IgZ– und ZgZ–Abstand der vereinigungs– und knotennormierten Editiermengendifferenz zeigt Abbildung 8.9.

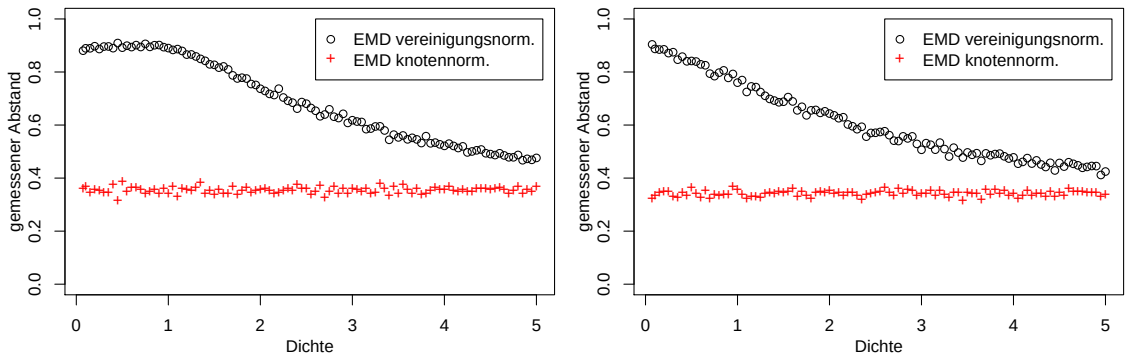


Abb. 8.9: Editiermengendifferenz beim IgZ– und ZgZ–Vergleich

Die vereinigungsnormierte Version zeigt für den IgZ–Vergleich das erwartete Verhalten, nämlich dass der Abstand mit zunehmender Dichte abnimmt. Allerdings ist das Verhalten bei dem ZgZ–Vergleich ähnlich, wobei dieses Verhalten hier wenig intuitiv ist. Die knotennormierte Version hingegen verhält sich wie ein knotenstrukturelles Maß und misst dabei zusätzlich noch einen Abstand von lediglich 0.35.

Strukturelle Indexmaße

Abbildung 8.10 zeigt den gemessenen Abstand des bewerteten Fred & Jain–Maßes und der normierten Variantion der Information. Als Index für die Indexdifferenz wurde der Mittelwert aus Coverage, Performance und durchschnittlicher Interclusterconductance genutzt.

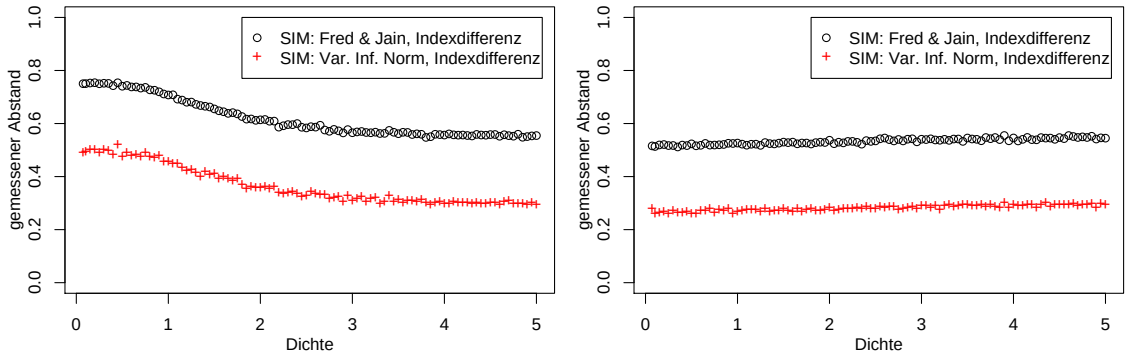


Abb. 8.10: Strukturelle Indexmaße beim IgZ– und ZgZ–Vergleich

Diese Maße sind neben dem graphstrukturellen Rand–Maß die einzigen, die sich so verhalten, wie es der Intuition entspricht. Beim IgZ–Vergleich nimmt der gemessene Abstand mit steigender Dichte ab, beim ZgZ–Vergleich bleibt er gleich.

8.1.5 Ergebnisse dieser Testreihe

Beim ZgZ–Vergleich messen lediglich das angepasste Rand–Maß und die Entropiemaße Fred & Jain und Strehl & Gosh einen maximalen Abstand von Eins. Das bedeutet, wenn man für unabhängige zufällige Clusterungen den maximalen Abstand messen möchte, sind diese Maße die erste Wahl.

Die Ergebnisse der graphstrukturellen Abstände sind ernüchternd. Lediglich die strukturellen Indexmaße und das graphstrukturelle Rand–Maß verhalten sich der Intuition entsprechend.

8.2 Initial- und Algorithmenclusterungen

Im folgenden Abschnitt wird die Initialclusterung von Gaussgeneratoren mit einer Clusterung, die der MCL–Algorithmus (siehe Algorithmus 3 in Kapitel 2.5.2) berechnet, verglichen. Eine vom MCL–Algorithmus berechnete Clusterung soll im Folgenden mit Algorithmusclusterung bezeichnet werden. Es wird sich zeigen, dass es drei Bereiche gibt. In einem der drei stimmen Initial– und Algorithmusclusterung überein, in den beiden anderen besitzt die Algorithmusclusterung mehr bzw. weniger Cluster als die Initialclusterung. Inwieweit die Abstandsmaße diese drei Bereiche bewerten, ist Ziel dieser Testreihe.

8.2.1 Setup

Für diesen Test wurden Gaußgeneratoren mit $n = 100$ Knoten genutzt, somit besitzt die Initialclusterung eine zufällige Clusteranzahl zwischen 2 und $\lfloor\sqrt{n}\rfloor = 10$. Die Wahrscheinlichkeit für Intraclusterkanten p_{in} variierte zwischen 0.05 und 1.0 mit Schrittweiten von 0.05. Die Wahrscheinlichkeit für Interclusterkanten p_{out} variierte für festes p_{in} zwischen 0.05 und p_{in} ebenfalls mit Abstand 0.05. Dies ergibt insgesamt 209 Einzeltests.

Durch die Wahl dieser p_{in}– und p_{out}–Werte ist die Initialclusterung nicht in allen Bereichen signifikant. Vor allem bei $p_{\text{in}} = p_{\text{out}} = 1$ ist der generierte Graph vollständig. Aus diesem Grund werden drei Bereiche unterschieden, die in Abbildung 8.11 dargestellt sind.

signifikanter Bereich Hier besitzt der generierte Graph eine ausgeprägte Initialclusterung. Dies ist der Fall für $p_{\text{in}} \geq 0.5$ und $p_{\text{out}} \leq 0.3$.

Bereich A Für $p_{\text{in}} \geq 0.5$ und $p_{\text{out}} > 0.3$ besitzt der geclusterte Graph bereits viele Interclusterkanten, sodass die Qualität der Initialclusterung nicht mehr sehr hoch ist.

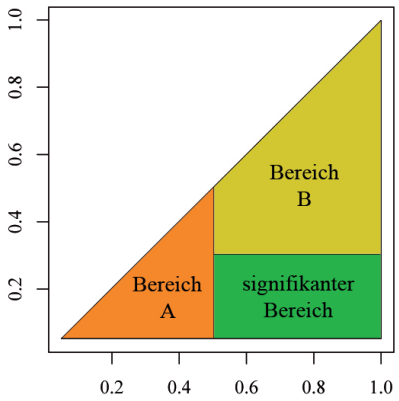


Abb. 8.11: Bereiche des Gaußgenerators. x-Achse: p_{in}, y-Achse: p_{out}

Bereich B Für $p_{\text{in}} < 0.5$ besitzt der Graph wenig Intraclusterkanten, sodass in diesem Bereich die Initialclusterung ebenfalls nicht sehr signifikant ist.

Zum besseren Verständnis von Initial- und Algorithmusclusterung zeigt die Abbildung 8.12 die Qualität und die Abbildung 8.13 die Anzahl der Cluster der jeweiligen Clusterung. Für die Qualitätsmessung wurde der Durchschnitt aus Coverage, Performance und durchschnittlicher Interclusterconductance genutzt. In den nun folgenden Bilder sind auf der x- und y-Achse jeweils p_{in} und p_{out} abgetragen. Die z-Achse gibt den jeweiligen Wert an.

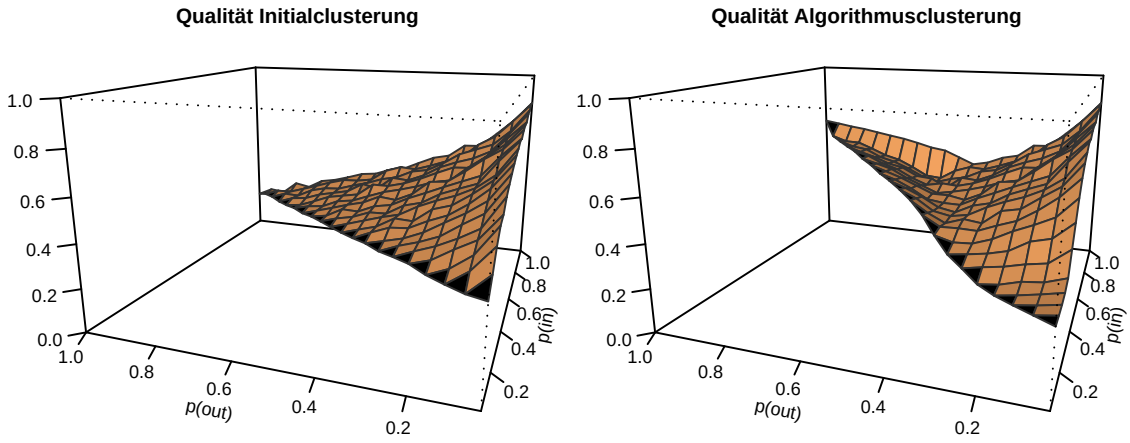


Abb. 8.12: Qualität der Initial– und Algorithmusclusterung

Man erkennt, dass für hohe p_{in} mit niedrigem p_{out} die Initialclusterung einen hohen Index besitzt. Außerdem besitzt die Algorithmusclusterung im signifikanten Bereich den gleichen Index. In Bereich A ist der Index der Algorithmusclusterung niedriger, in Bereich B ist er höher.

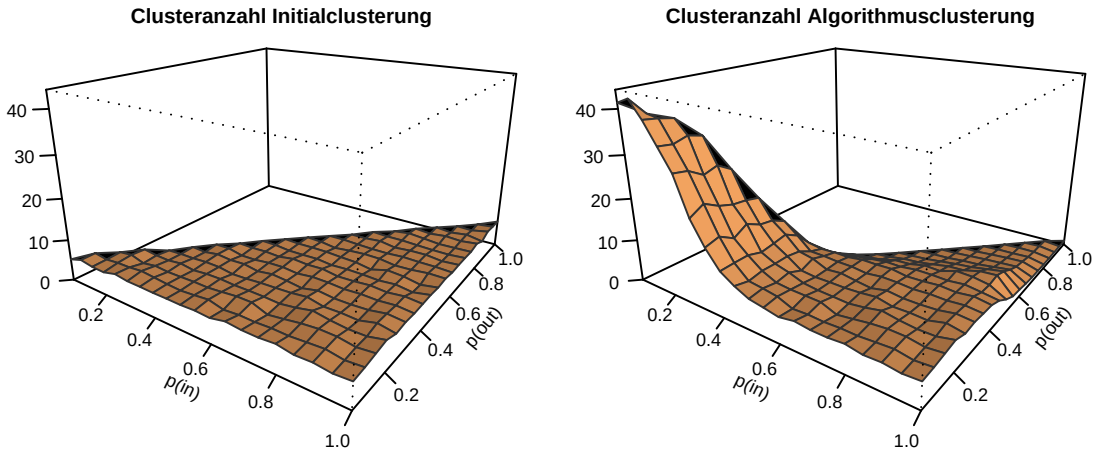


Abb. 8.13: Clusteranzahl der Initial– und Algorithmusclusterung

Da die Test bis zur statistischen Signifikanz wiederholt werden und jedes mal die Clusterzahl zufällig zwischen 2 und 10 liegt, besitzt die Initialclusterung im Schnitt 6 Cluster. Die Algorithmusclusterung hingegen besitzt im signifikanten Bereich ungefähr gleich viele Cluster wie die Initialclusterung. Im Bereich B nähert sich mit steigendem p_{out} die Algorithmusclusterung der 1-Clusterung an. Im Bereich A unterscheidet sich die Algorithmus- von der Initialclusterungen ebenfalls stark.

Folgerung Anscheinend stimmen Initial- und Algorithmusclusterung im signifikanten Bereich überein, wohingegen in den Bereichen A und B die Clusterungen recht stark voneinander abweichen. Inwieweit sich dies in den Abstandsmaßen wiederfindet, soll nun untersucht werden.

8.2.2 Qualitativer Abstand

Bei den qualitativen Maßen wird sich aus Platzgründen auf die Indexdifferenz und den Indexquotienten beschränkt (die Varianten der Indexdifferenz finden sich in Anhang B.1). Abbildung 8.14 zeigt die Auswertungen dieser beiden Maße. Beide nutzen erneut einen Durchschnitt aus Coverage, Performance und durchschnittlicher Interclusterconductance.

Man erkennt, dass die Maße im signifikanten Bereich einen Abstand von Null messen. Der Indexquotient bewertet die Abweichung in Bereich A und B stärker als die Indexdifferenz, da ersterer ein schlechtlastiges Maß ist. Somit haben die geringfügigen Abweichungen der Qualitätsmessung einen größeren Einfluß auf die Abstandsmessung des Indexquotienten.

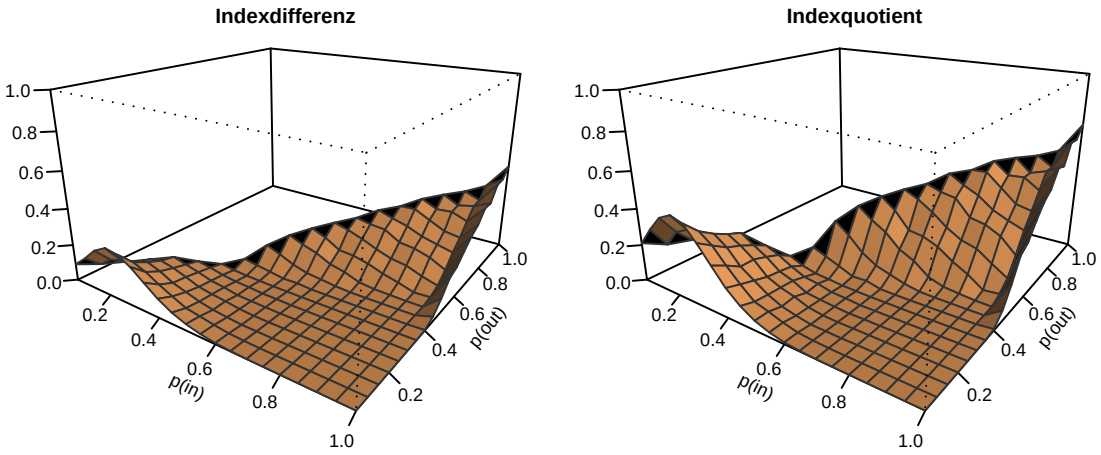


Abb. 8.14: Qualitative Maße: Initial- gegen Algorithmusclusterung

8.2.3 Knotenstruktureller Abstand

Hier werden die gemessenen Abstände der knotenstrukturellen Maße vorgestellt. Dabei wird die Auswertung erneut nach Paar-, Schnitt und Entropiemaßen unterteilt.

Paarmaße

Abbildung 8.15 zeigt den gemessenen knotenstrukturellen Abstand des Rand– und angepassten Rand–Maßes.

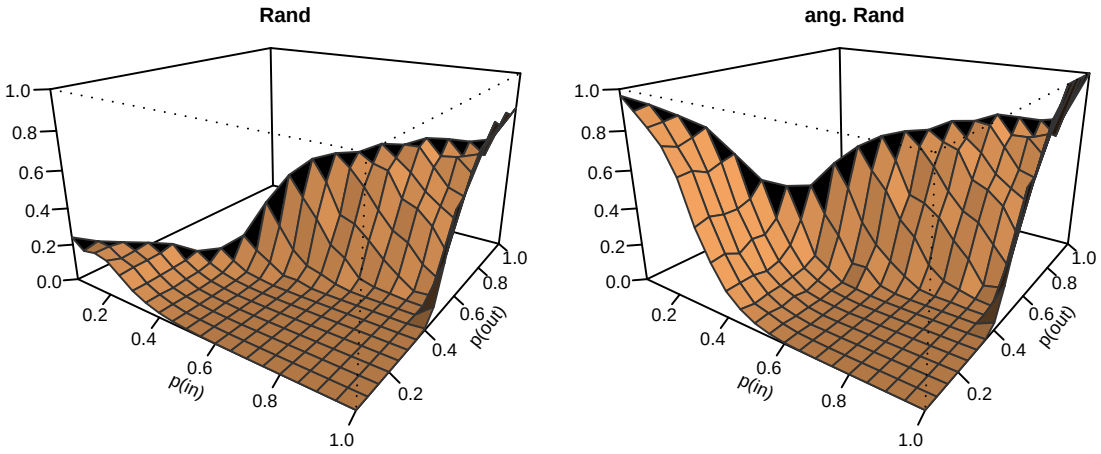


Abb. 8.15: Knotenstrukturelle Paarmaße I: Initial- gegen Algorithmusclusterung

Man erkennt sehr gut die drei Bereiche. Im signifikanten Bereich messen alle Maße einen Abstand von Null, Bereich B wird von beiden Maßen ähnlich bewertet

und im Bereich A misst das Rand–Maß einen signifikant kleineren Abstand als das angepasste Rand–Maß. Dies liegt daran, dass im Bereich A die Clusteranzahl der Algorithmusclusterung stark ansteigt.

Die gemessenen Abstände des Fowlkes–Mallows- und Jaccard–Maßes zeigt Abbildung 8.16. Beide Maße messen die drei Bereiche der Intuition entsprechend.

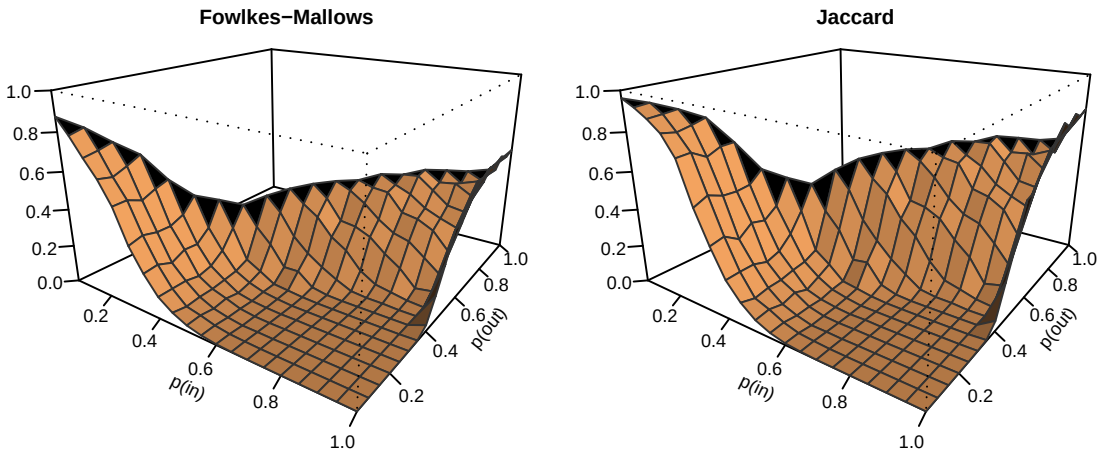


Abb. 8.16: Knotenstrukturelle Paarmaße II: Initial- gegen Algorithmusclusterung

Schnittmaße

Abbildung 8.17 zeigt den gemessenen Abstand des Maximum–Match– und van Dongen–Maßes.

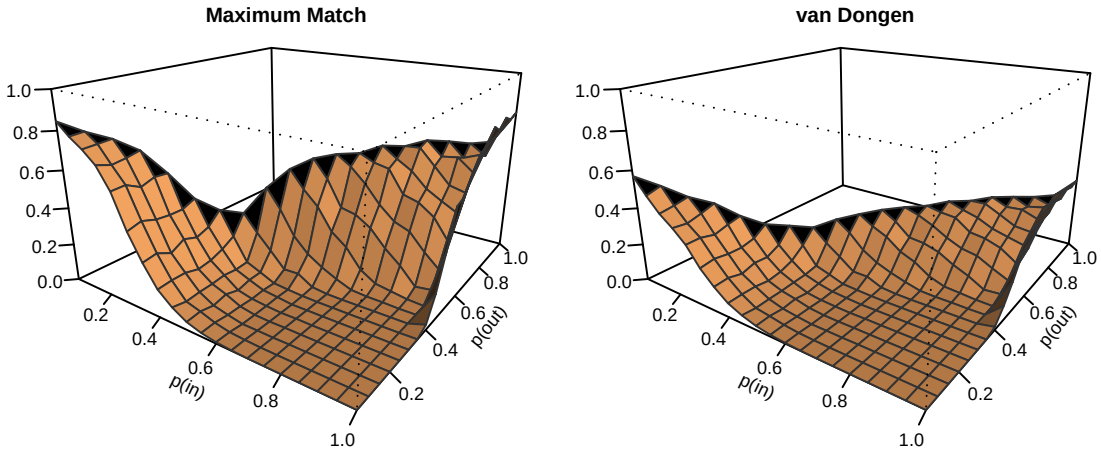


Abb. 8.17: Knotenstrukturelle Schnittmaße I: Initial- gegen Algorithmusclusterung

Beide Maße bewerten die drei Bereiche der Intuition entsprechend, allerdings ist der gemessene Abstand kleiner als beim Maximum–Match–Maß.

Abbildung 8.18 zeigt nun noch den gemessenen Abstand des F– und Meila–Heckerman–Maßes. Beide Maße sind asymmetrisch, aus diesem Grund sind für jedes Maß zwei Auswertungen abgebildet.

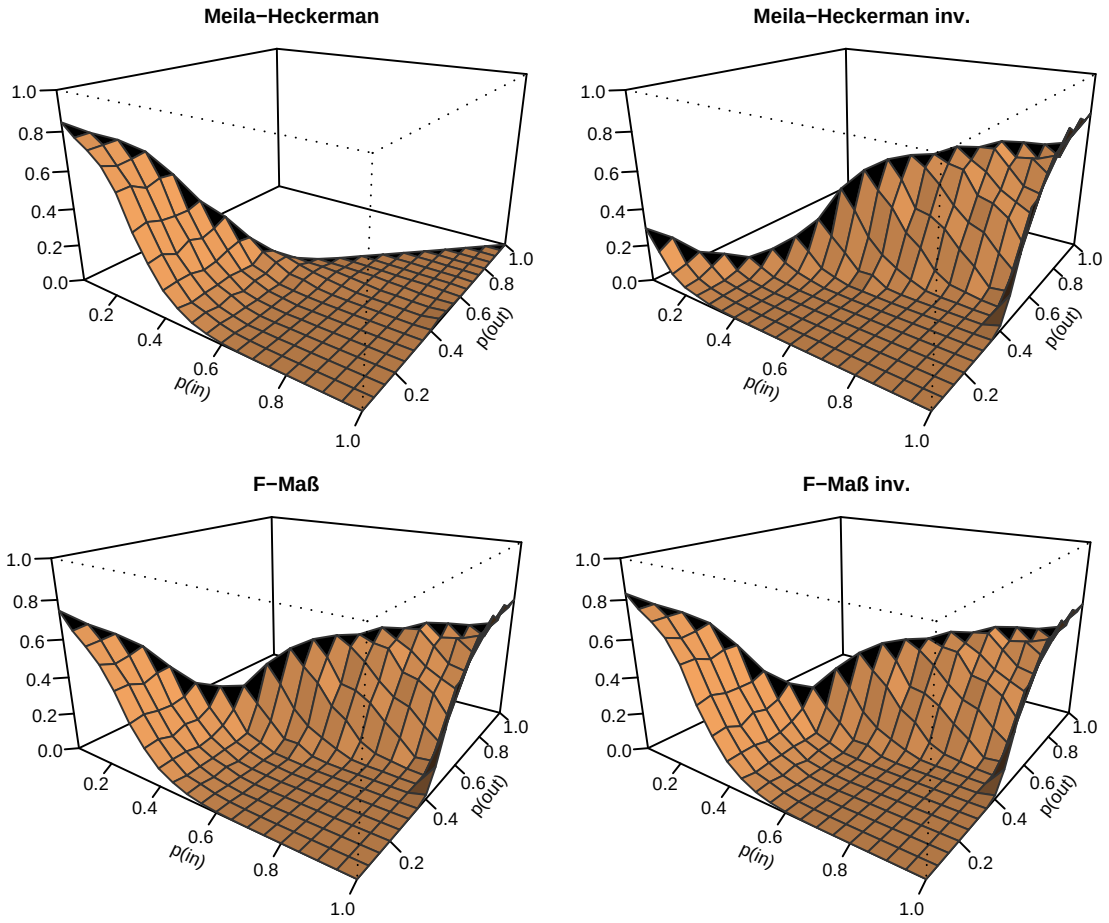


Abb. 8.18: Knotenstrukturelle Schnittmaße II: Initial- gegen Algorithmusclusterung

Man erkennt beim Meila–Heckerman–Maß die Nachteile eines asymmetrischen Maßes. Beide Versionen bewerten den signifikanten Bereich richtig mit einem Abstand von Null, jede Version erkennt aber jeweils nur einen Bereich. Das F–Maß hingegen zeigt kein signifikant asymmetrisches Verhalten und bewertet die drei Bereiche der Intuition entsprechend.

Entropiemaße

Abbildung 8.19 zeigt die Entropiemaße von Fred & Jain, Strehl & Gosh und die normierte Variation der Information.

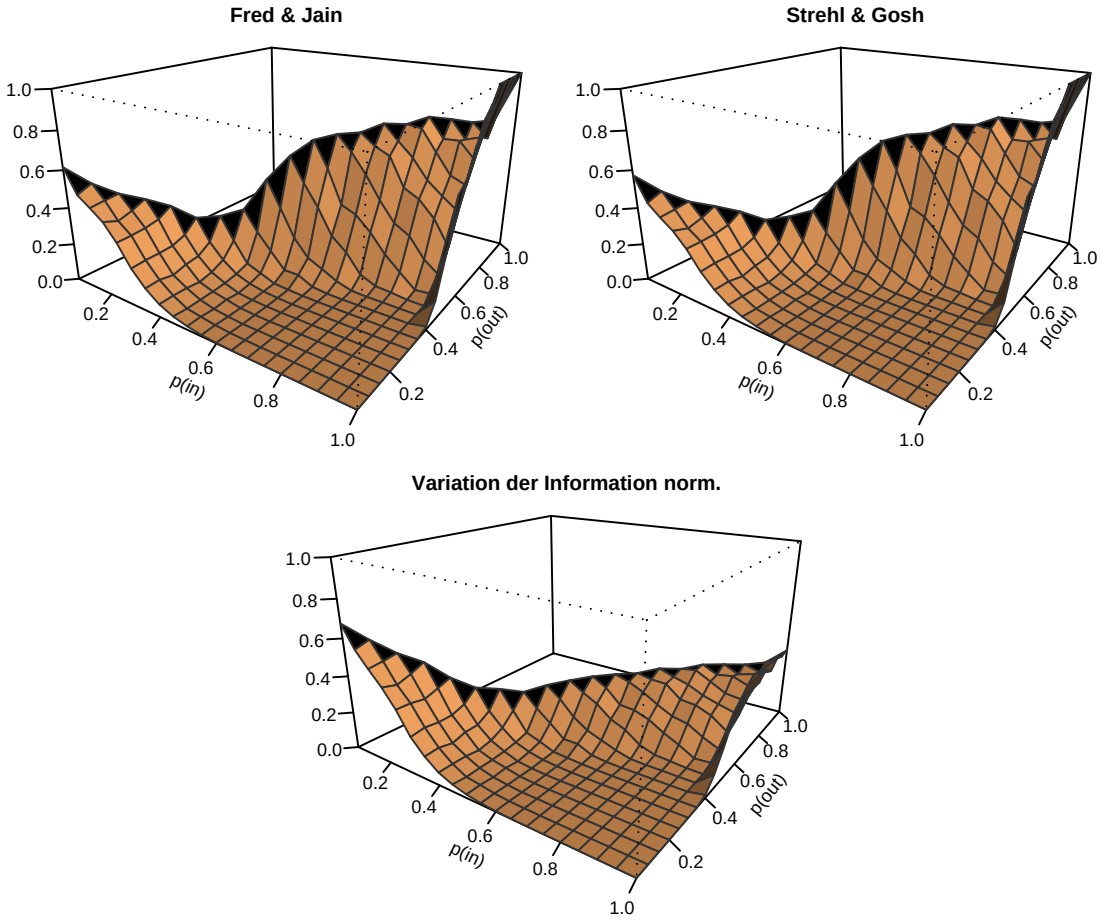


Abb. 8.19: Knotenstrukturelle Entropiemaße: Initial- gegen Algorithmusclusterung

Die Maße von Fred & Jain, Strehl & Gosh und die Variation der Information verhalten sich wie erwartet, wobei die ersteren beiden Maße den Abstand in Bereich B stärker als den Abstand in Bereich A bewerten. Die normierte Variation der Information verhält sich umgekehrt, wobei der maximal gemessene Abstand niedriger als bei den restlichen Entropiemaßen ist.

8.2.4 Graphstruktureller Abstand

Für den Vergleich zwischen Initial- und Algorithmusclusterung soll an dieser Stelle auf eine Aufführung der graphstrukturellen Paar–, Schnitt– und Entropiemaße verzichtet werden. Der Grund hierfür ist, dass in diesem Falle die knotenstrukturellen

Maße sehr gute Ergebnisse liefern und sich die Grafiken der jeweiligen graphstrukturellen Maße von denen in Abschnitt 8.2 nicht unterscheiden. Das liegt unter anderem auch daran, dass die erzeugten Graphen annähernd regulär sind und somit die Erweiterungen bei den Schnitt– und Entropiemaßen wirkungslos bleiben. Die Grafiken dieser Auswertungen finden sich im Anhang B.1.

Es wird nun noch auf die neu eingeführten graphstrukturellen Abstandsmaße eingegangen.

Editiermengendifferenz

Abbildung 8.20 zeigt den gemessenen Abstand der vereinigungs- und knotennormierten Editiermengendifferenz.

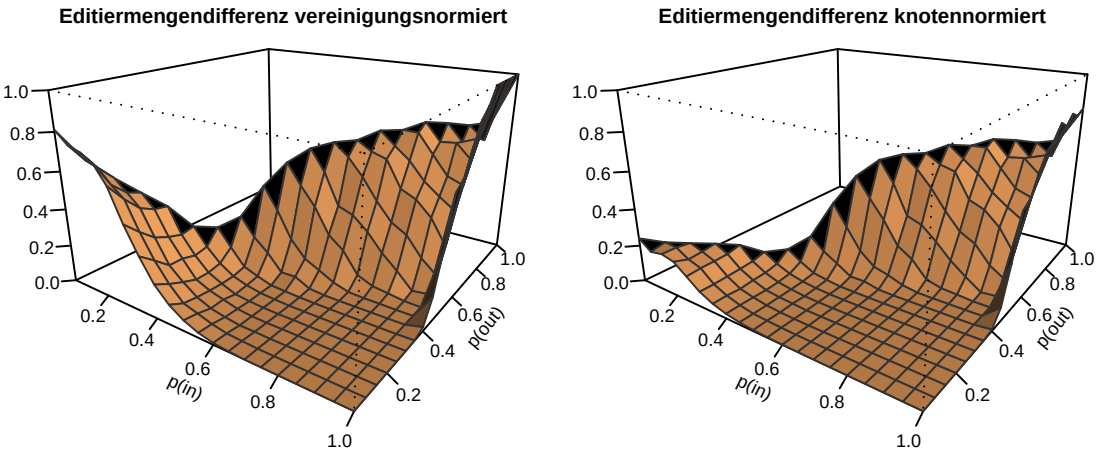


Abb. 8.20: Editiermengendifferenz: Initial- gegen Algorithmusclusterung

Die Maße trennen die drei Bereiche sehr eindeutig, wobei die knotennormierte Version in Bereich A einen deutlich kleineren Abstand misst als die vereinigungsnormierte.

Strukturelle Indexmaße

Abbildung 8.21 zeigt den gemessenen Abstand der strukturellen Indexmaße auf Basis des Fred & Jain Mases bzw. der normierten Variantion der Information kombiniert mit der Indexdifferenz. Als Index für die Indexdifferenz wurde der Mittelwert aus Coverage, Performance und durchschnittlicher Interclusterconductance benutzt.

Die strukturellen Indexmaße Maße zeigen ein intuitives Verhalten und trennen die drei Bereiche eindeutig. Der maximale Abstand ist allerdings deutlich geringer als bei den knotenstrukturellen Maßen.

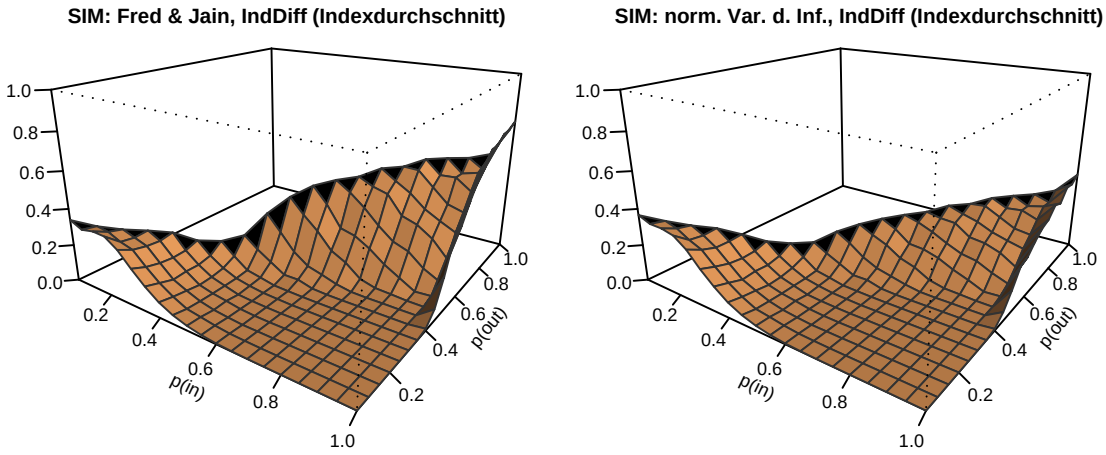


Abb. 8.21: Strukturelle Indexmaße: Initial- gegen Algorithmusclusterung

8.2.5 Ergebnisse dieser Testreihe

Bei dieser Testreihe kann man abschließend festhalten, dass die knotenstrukturellen Maße den Abstand zwischen der Initial– und Algorithmusclusterungen auf Graphen, die mit dem Gaußgenerator erzeugt wurden, intuitiv bewerten. Einzig das asymmetrische Meila–Heckerman–Schnittmaß liefert ein wenig zufriedenstellendes Ergebnis.

Auch die neu eingeführte Editiermengendifferenz und die strukturellen Indexmaße liefern gute Ergebnisse.

8.3 Lokale Minimierung

Die nun folgenden Testreihen sollen zwei Vergleiche gegenübergestellen. In beiden Fällen wird die Initialclusterung von Attraktorengraphen mit einer Clusterung, die der lokale Minimierer (Abschnitt 2.5.3) berechnet, verglichen. Im Gegensatz zu Kapitel 8.1 wurde diese Untersuchung für zwei feste Dichten $f = 1$ und $f = 3$ durchgeführt. Es wurde lokale Minimierung anstatt lokaler Optimierung genutzt, da die Initialclusterungen von Attraktoren verhältnismäßig signifikant sind und somit die lokale Optimierung zu keiner merklichen Verbesserung der Clusterung führt.

8.3.1 Setup

Zunächst wurden 2 Typen von Attraktorengraphen mit je 1000 Knoten und somit einer zufälliger Clusteranzahl zwischen 2 und 10 generiert. Die beiden Typen von Attraktoren unterscheiden sich in ihrem Dichtewert f.

Typ 1 Dieser Typ besitzt einen Dichtewert von $f = 1$. Die Initialclusterungen sind daher signifikant.

Typ 2 Für diesen Typen wurde ein Dichtewert von $f = 3$ gewählt, wodurch die Initialclusterung weniger signifikant als die von Typ 1 ist.

Der Dichtewert hat nur Einfluß auf die Anzahl der Kanten, die Struktur der beiden Initialclusterungen ist somit im Erwartungswert gleich. Die Signifikanz der beiden Clusterungen unterscheidet sich allerdings erheblich. Dies verdeutlicht Tabelle 8.1.

Typ	Dichte (f)	Kanten	$m(\mathcal{C})$	$\overline{m}(\mathcal{C})$	$cov(\mathcal{C})$	$per(\mathcal{C})$	$iccA(\mathcal{C})$
1	1.0	≈ 50000	≈ 40000	≈ 10000	≈ 0.80	≈ 0.83	≈ 0.78
2	3.0	≈ 230000	≈ 90000	≈ 140000	≈ 0.36	≈ 0.67	≈ 0.31

Tabelle 8.1: Werte der Initialclusterung der beiden untersuchten Attraktoren

Die Initialclusterungen der beiden Typen wurde nun als Eingabe für den lokalen Minimierer genutzt. Dabei wurde die Anzahl der maximal erlaubten Knotenverschiebungen sukzessiv von 5 bis 500 in Schritten von je 5 Knoten erhöht. Mit wachsender bewegter Knotenzahl nimmt die Qualität der Clusterung ab. Abbildung 8.22 zeigt die Indizes der so erzeugten Clusterungen. Die Anzahl der bewegten Knoten ist hierbei auf der x-Achse abgetragen. Das linke Bild zeigt Typ 1 und das rechte Bild Typ 2.

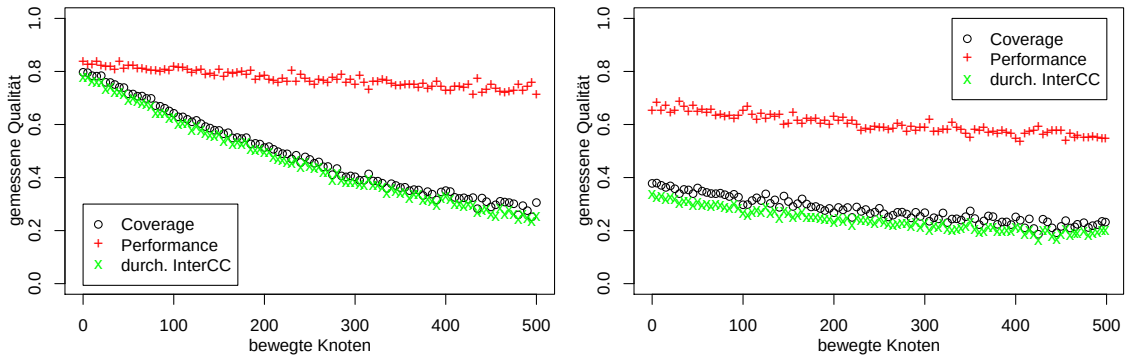


Abb. 8.22: Indizes von Typ 1 (links) und Typ 2 (rechts)

Abbildung 8.22 zeigt, dass mit zunehmender Anzahl der bewegten Knoten die Qualität der Clusterung von Typ 1 stärker abnimmt, als dies bei Typ 2 der Fall ist. Dies erwartet man auch, da die Initialclusterung von Typ 1 signifikanter als die von Typ 2 ist.

Folgerung Die Initial- und Algorithmusclusterungen unterscheiden sich nur qualitativ, sind knotenstrukturell aber im Erwartungswert zumindestens ähnlich. Somit

sollten knotenstrukturelle Abstandsmaße keinen Unterschied zwischen dem Vergleich auf Typ 1 und dem Vergleich auf Typ 2 machen, qualitative und graphstrukturelle Maße schon. Inwieweit dies zutrifft, ist Ziel dieser Testreihen.

Im Folgenden zeigt die linke Grafik immer den Vergleich auf Attraktorentyp 1, die rechte den Vergleich auf Typ 2. Auf der x-Achse ist die Anzahl der bewegten Knoten abgetragen, die y-Achse zeigt den gemessenen Abstand des jeweiligen Maßes.

8.3.2 Qualitativer Abstand

Abbildung 8.23 zeigt den gemessen Abstand verschiedener qualitativer Maße. Als Index wird der Durchschnitt von Coverage, Performance und durchschnittlicher Interclusterconductance genutzt.

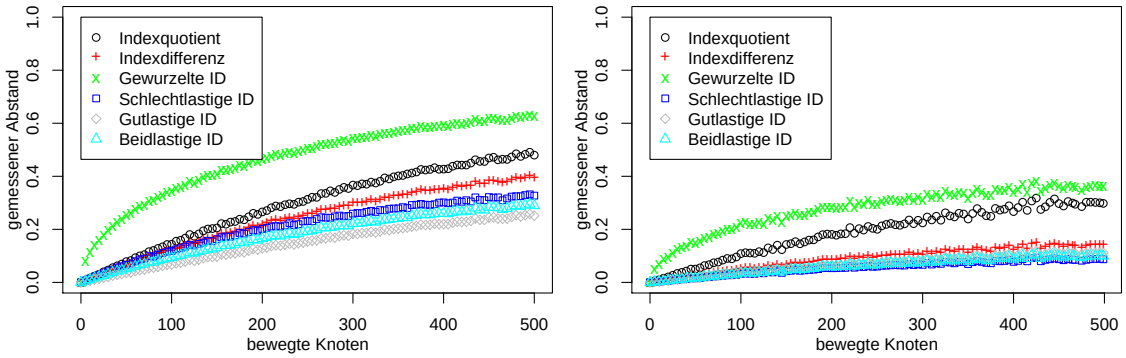


Abb. 8.23: Qualitative Maße auf Typ 1 (links) und Typ 2 (rechts)

Abbildung 8.23 zeigt den direkten Einfluï¿½des Indexes auf die qualitativen Abstandsmaße. Wie erwartet, sind die gemessenen Abstände mit zunehmender bewegter Knotenzahl bei Typ 1 gröï¿½r als bei Typ 2.

8.3.3 Knotenstruktureller Abstand

Da die Struktur der Clusterungen bei Typ 1 und Typ 2 ähnlich sind, erwartet man bei den knotenstrukturellen Maßen, dass sie den Abstand zwischen den Clusterungen auf Typ 1 und Typ 2 gleich groß messen.

Paarmaße

Abbildung 8.24 zeigt die gemessenen Abstände der Paarmaße Rand, angepasster Rand, Fowlkes–Mallows und Jaccard.

Alle vier Maße messen den gleichen Abstand auf Typ 1 und Typ 2. Auffällig ist, dass das Rand–Maß einen merklich kleineren Abstand als die restlichen drei Maße misst.

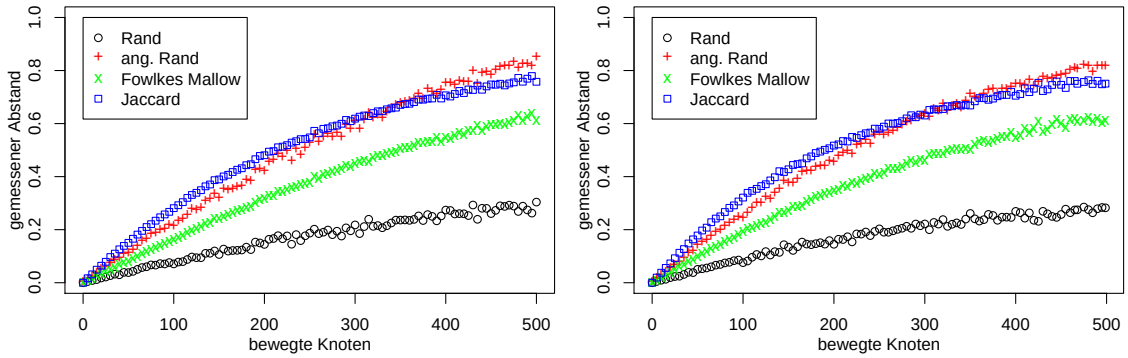


Abb. 8.24: Knotenstrukturelle Paarmaße auf Typ 1 und Typ 2

Schnittmaße

Die gemessenen Abstände der Schnittmaße zeigt Abbildung 8.4. Aufgrund der Asymmetrie des F– und des Meila–Heckerman–Maßes sind diese Maße jeweils zweimal abgetragen.

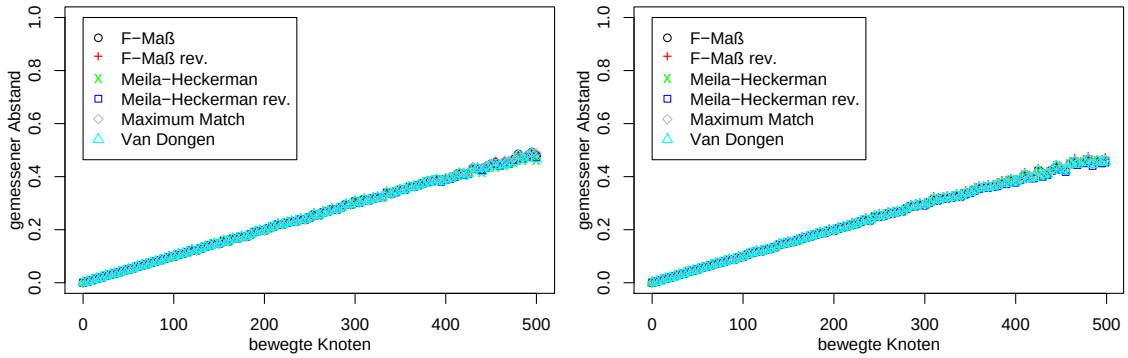


Abb. 8.25: Knotenstrukturelle Schnittmaße auf Typ 1 und Typ 2

Zwei Dinge sind zu erkennen: Zum einen messen alle Maße den gleichen Abstand im Verhältnis zueinander. Zum anderen messen auch die Schnittmaße den gleichen Abstand auf Typ 1 und Typ 2.

Entropiemaße

Abbildung 8.26 zeigt die gemessenen Abstände der Entropiemaße von Fred & Jain, Strehl & Gosh und der normierten Variation der Information.

Die Entropiemaße unterscheiden ebenfalls nicht zwischen Typ 1 und Typ 2. Auch die Gleicheit der Maße Fred & Jain und Strehl & Gosh zeigt sich hier erneut. Der gemessene Abstand der normierten Variation der Information ist wieder nicht so ausgeprägt wie bei den anderen Entropiemaßen.

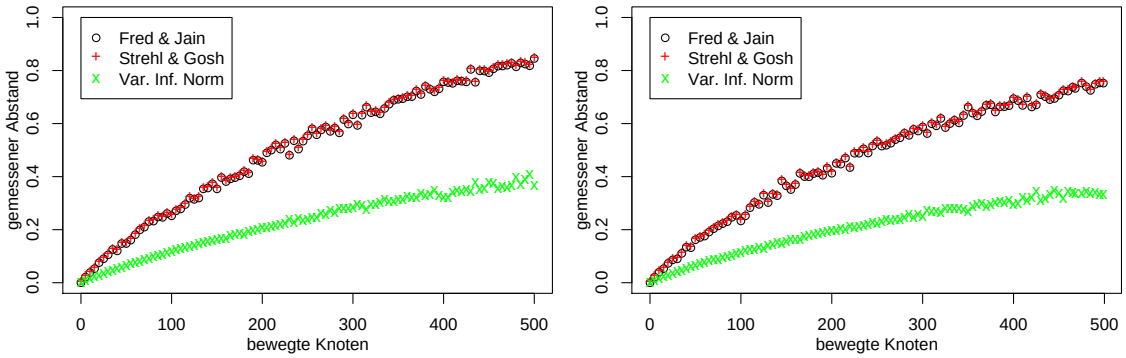


Abb. 8.26: Knotenstrukturelle Entropiemaße auf Typ 1 und Typ 2

8.3.4 Graphstrukturelle Maße

Für die graphstrukturellen Maße erwartet man, dass sie den Abstand auf Typ 1 und Typ 2 nicht gleich bewerten. Sie nutzen nicht nur die Struktur der Clusterung, sondern auch die Struktur der Graphen und somit in gewisser Weise auch die Qualität der Clusterung. Wie Abbildung 8.22 zeigt, unterscheiden sich die Clusterungen hier bezüglich der beiden Typen teilweise erheblich.

Graphstrukturelle Paarmaße

Abbildung 8.27 zeigt die gemessenen Abstände der graphstruturellen Versionen des Rand–, angepassten Rand–, Fowlkes–Mallows und Jaccard–Maßes.

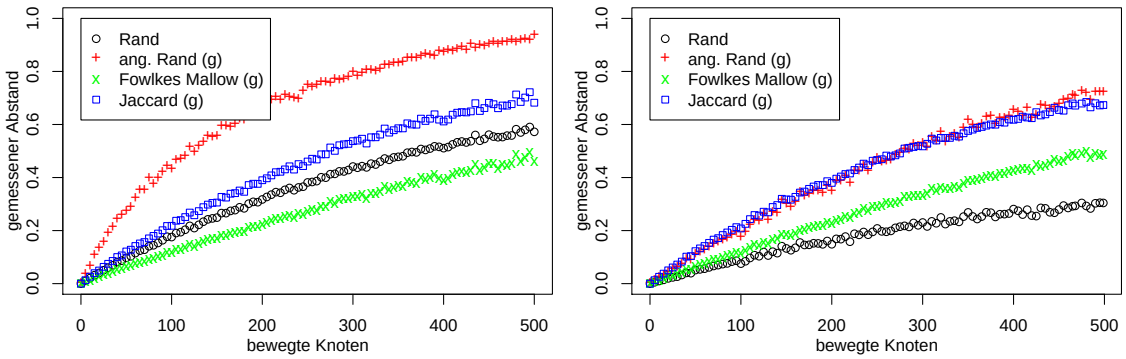


Abb. 8.27: Graphstrukturelle Paarmaße auf Typ 1 und Typ 2

Das Rand– und das angepasste Rand–Maß messen für steigende bewegte Knotenzahl einen signifikant größeren Abstand auf Typ 1 als auf Typ 2. Das Fowlkes–Mallows– und Jaccard–Maß verhalten sich hingegen auf beiden Typen ungefähr gleich. Anscheinend hat die graphstrukturelle Erweiterung in diesem Szenario beim Rand- und angepasstem Rand–Maß einen stärkeren Einfluss als bei den anderen Paarmaßen.

Graphstrukturelle Schnittmaße

Abbildung 8.28 zeigt die gemessenen Abstände der graphstrukturellen Versionen des F– (zweimal), Meila–Heckerman– (ebenfalls zweimal) und Maximum–Match–Maßes.

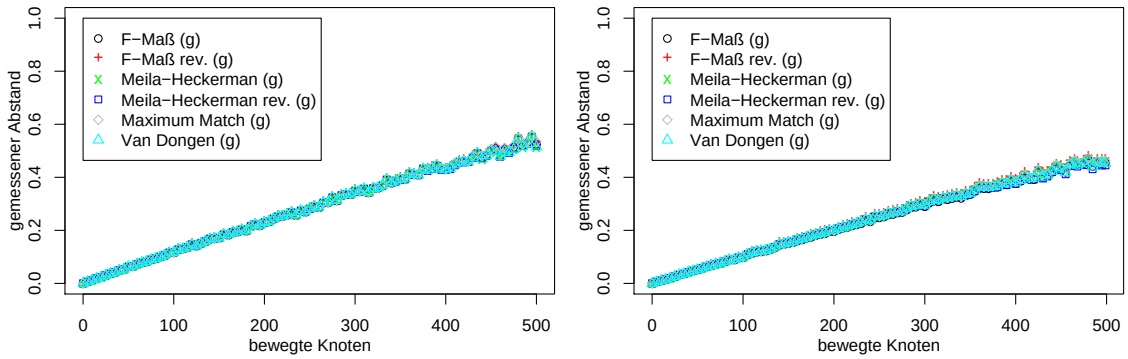


Abb. 8.28: Graphstrukturelle Schnittmaße auf Typ 1 und Typ 2

Da der Graph nahezu regulär ist, ist die graphstrukturelle Erweiterung in diesem Falle nutzlos. Auch diese Maße können keinen Unterschied zwischen Typ 1 und Typ 2 machen. Anscheinend ist die graphstrukturelle Erweiterung der Schnittmaße bei diesem Setup wirkungslos.

Kantenentropiemaße

Die gemessenen Abstände der Kantenentropiemaße zeigt Abbildung 8.29. Gemessen wurde der Abstand mit der jeweiligen graphstrukturellen Version des Fred & Jain–Maßes, des Strehl & Gosh-Maßes und der normierten Variation der Information.

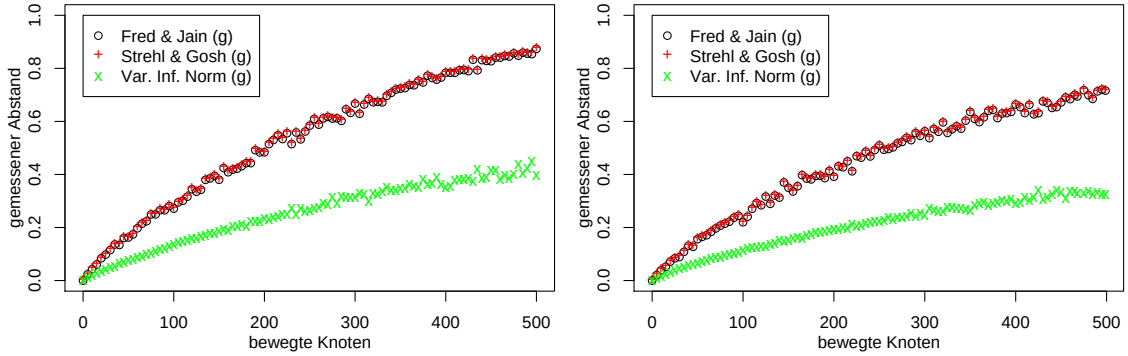


Abb. 8.29: Graphstrukturelle Entropiemaße auf Typ 1 und Typ 2

Alle drei Maße messen einen leicht kleineren Abstand auf Typ 2. Die gemessenen Abstände sind bei den Maßen von Fred & Jain und Strehl & Gosh signifikanter als bei der Variation der Information.

Editiermengen Differenz

Abbildung 8.30 zeigt den gemessen Abstand der vereinigungs- und knotennormierten Editiermengen Differenz.

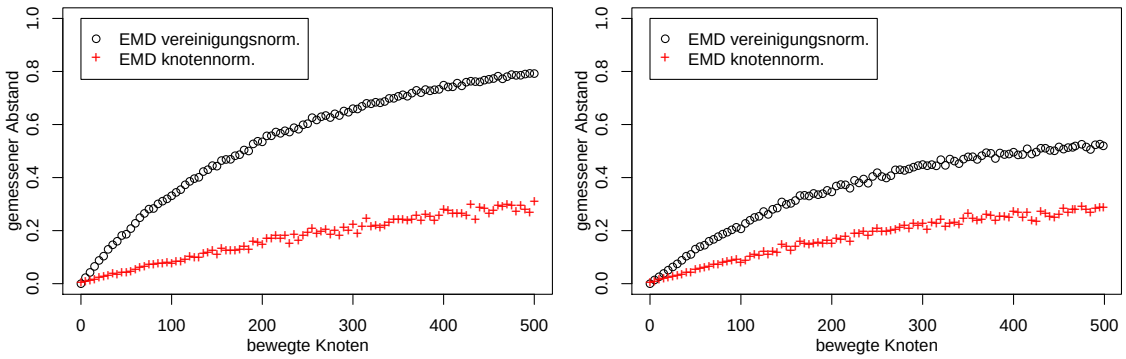


Abb. 8.30: Editiermengendifferenz auf Typ 1 und Typ 2

Die vereinigungsnormierte Version arbeitet den Unterschied zwischen Typ 1 und Typ 2 sehr deutlich heraus, die knotennormierte hingegen misst auf beiden Typen den gleichen Abstand.

Strukturelle Indexmaße

Abbildung 8.31 zeigt den gemessenen Abstand der strukturellen Indexmaße $\mathcal{FJ}_{\mathcal{ID}}$ und $\mathcal{NVI}_{\mathcal{ID}}$.

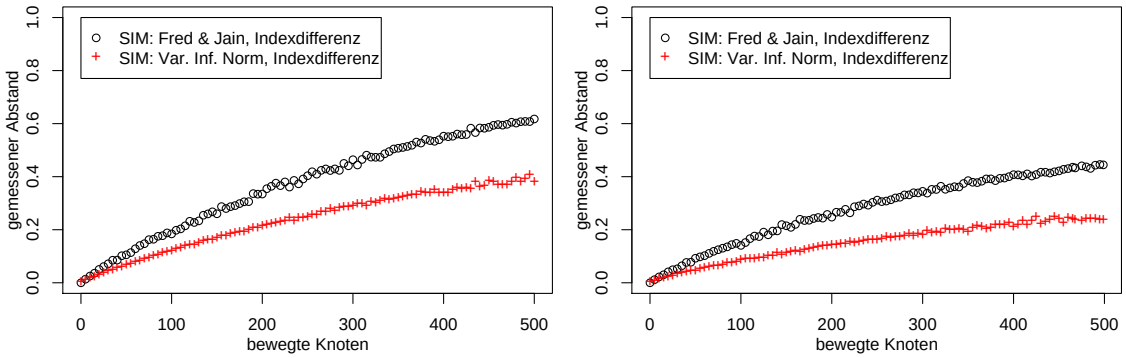


Abb. 8.31: Strukturelle Indexmaße auf Typ 1 und Typ 2

Beide Maße messen auf Typ 1 mit steigendem Abstand einen größeren Abstand als auf Typ 2. Somit zeigen auch diese Maße ein intuitives Verhalten.

8.3.5 Ergebnisse dieser Testreihe

Die gemessenen Abstände in diesem Testszenario spiegeln für die qualitativen und knotenstrukturellen Maße die Erwartung wieder. Die qualitativen Maße messen auf Typ 1 mit steigender Anzahl der bewegten Knoten einen größern Abstand als auf Typ 2. Die knotenstrukturellen Maße hingegen können wie erwartet keinen unterschiedlichen Abstand zwischen den Clusterungen auf Typ 1 und Typ 2 messen. Dieses Verhalten ist in den meisten Anwendungsfällen von großen Nachteil.

Ferner zeigen diese Tests, dass die graphstrukturelle Erweiterung eines knotenstrukturellen Maßes – in diesem Falle für die Schnittmaße – ohne Wirkung bleiben kann. Auch die Editiermengendifferenz kann bei falsch gewählter Normierung die Abstände wenig intuitiv messen.

8.4 Verfeinerung und Vergröberung

Nun soll untersucht werden, wie die Abstandsmaße Verfeinerungen bzw. Vergröberungen bewerten. Hierzu werden Hierarchiegraphen, die in Kapitel 2 vorgestellt worden sind, verwendet. Hierarchiegraphen besitzen mehrere Clusterlevel $\mathcal{C}_i$, die nach Korrolar 7 jeweils Verfeinerung bzw. Vergöberungen voneinander sind.

8.4.1 Setup

Als Cliquengröße des untersuchten Hierarchiegraphen wurde $cli = 3$ und als maximaler Clusterlevel $L_{\max} = 8$ gewählt. Dies ergibt somit einen rekursiven Graphen mit $n = 3^8 = 6561$ Knoten und 9840 Kanten. Die Anzahl der Cluster, Clustergrößen, Intra- bzw. Interclusterkanten und Indizes gibt Tabelle 8.2 an.

$\mathcal{C}_i$	Cluster	$\lvert C\rvert$	$m(\mathcal{C}_i)$	$\overline{m}(\mathcal{C}_i$	$cov(\mathcal{C}_i)$	$per(\mathcal{C}_i)$	$iccA(\mathcal{C}_i)$	$i(\mathcal{C}_i)$
0	6561	1	0	9840	0.0	1.0	0.0	0.3333
1	2187	3	6561	3279	0.6668	0.9999	0.6726	0.7798
2	729	9	8748	1092	0.8890	0.9991	0.8900	0.9261
3	243	27	9477	363	0.9631	0.9965	0.9632	0.9743
4	81	81	9720	120	0.9878	0.9883	0.9878	0.9880
5	27	243	9801	39	0.9960	0.9636	0.9960	0.9852
6	9	729	9828	12	0.9988	0.8895	0.9988	0.9624
7	3	2187	9837	3	0.9997	0.6672	0.9997	0.8889
8	1	6561	9840	0	1.0	0.0001	0.0	0.3333

Tabelle 8.2: Werte der verschiedenen Clusterlevel. Die letzte Spalte gibt hierbei den Durchschnitt aus Coverage, Performance und durchschnittlicher Interclusterconductance an.

Man erkennt, dass der Clusterlevel 0 der Singleton–Clusterung und Level 8 der 1–Clusterung entspricht. Ferner ist das gegenläufige Verhalten der Indizes interessant. Performance bewertet niedrige Clusterlevel besser, Coverage und Interclusterconductance eher höhere. Ein Ausnahme ist die 1-Clusterung, diese wird von Interclusterconductance per Definition mit Null bewertet. Intuitiv ist allerdings bei diesen Graphen schwer zu entscheiden, welcher Clusterlevel der signifikanteste ist. Tendenziell sind wohl die Clusterlevel 3 bis 5 am signifikantesten; in diesem Bereich ist der Durchschnitt der Indizes auch am höchsten.

Im Folgenden werden nun die Abstände zwischen allen möglich Kombination der Clusterlevel gemessen. Wiederum wird dabei nach qualitativen, knotenstrukturellen und graphstrukturellen Maßen unterschieden.

Die folgenden Abbildungen sind erneut dreidimensional. Sie zeigen jeweils den gemessenen Abstand zwischen zwei Clusterleveln (auf x- bzw. y-Achse abgetragen) auf der z-Achse.

8.4.2 Qualitativer Abstand

Abbildung 8.32 zeigt die Indexdifferenz der verschiedenen Clusterlevel. Die linke Grafik nutzt den Durchschnitt von Coverage, Performance und durchschnittlicher Interclusterconductance als Index, die rechte nutzt lediglich Performance. Der Indexquotient und die Varianten der Indexdifferenz finden sich in Anhang B.2.

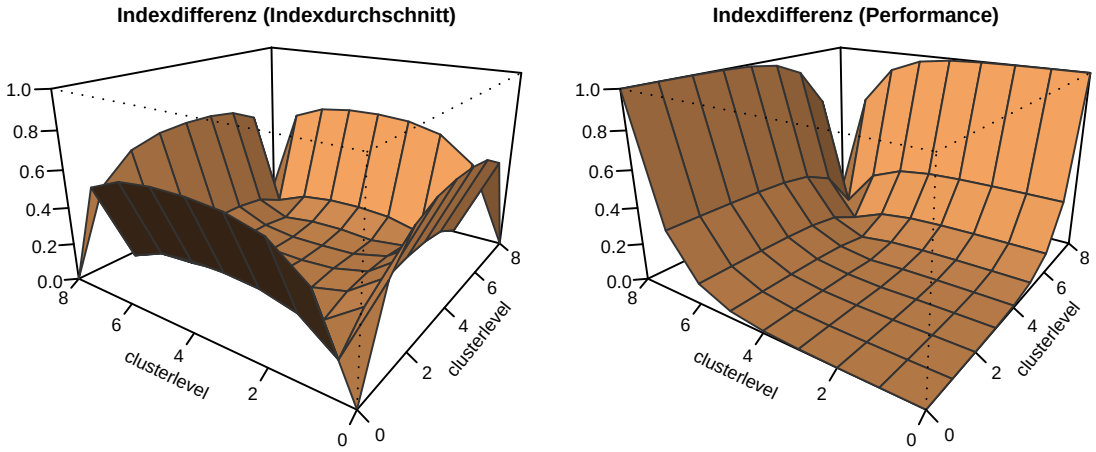


Abb. 8.32: Indexdifferenz auf Hierarchiegraphen

Hier zeigt sich, dass die Wahl des Qualitätsindexes einen groï¿½n Einfluss auf den gemessenen qualitativen Abstand hat. Durch das gegenläufige Verhalten der Indizes ist der Durchschnitt der Indizes für Singleton und 1-Clusterung gleich. Da die Abweichungen der Indizes in den mittleren Clusterlevel nsehr gering sind, sind die

gemessenen Abstände auch sehr gering. Bei Benutzung von Performance als Grundlage, ist dies ebenfalls für die niedrigen Clusterlevel der Fall.

8.4.3 Knotenstruktureller Abstand

Für den knotenstrukturellen Abstand erwartet man, dass zwischen gleichen Clusterlevel kein Abstand gemessen wird. Mit steigender Abweichung der Clusterlevel voneinander sollte auch der gemessene Abstand steigen. Die Untersuchung der knotenstrukturellen Maße unterteilt sich erneut nach Paar–, Schnitt– und Entropiemaßen.

Paarmaße

Abbildung 8.33 zeigt die gemessenen Abstände des Rand– und angepassten Rand–Maßes.

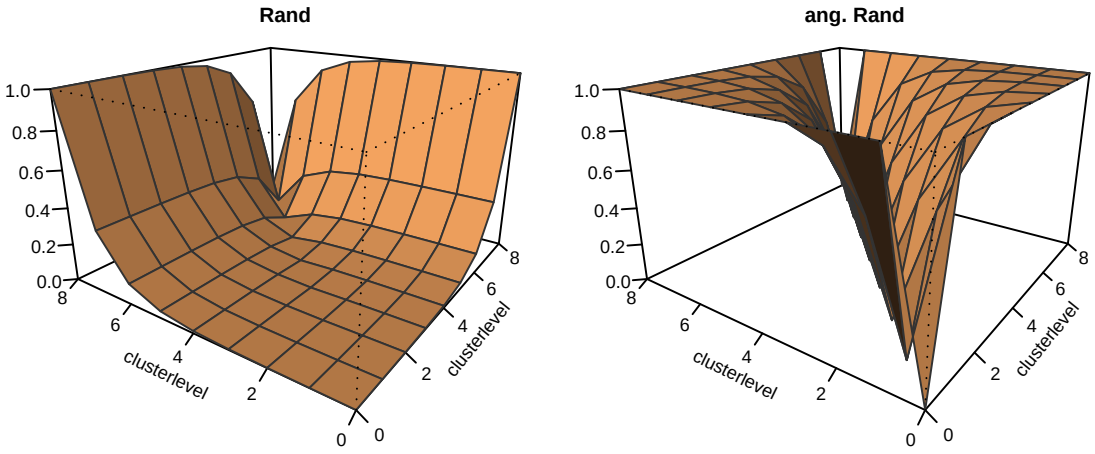


Abb. 8.33: Knotenstrukturelle Paarmaße I auf Hierarchiegraphen

Das Rand–Maß zeigt in seiner nicht angepassten Version das in 7 erwähnte Verhalten. Bei einer hohen Clusteranzahl wird immer ein kleiner Abstand gemessen. Für niedrige Clusterlevel ist die Clusteranzahl hoch, weshalb das Rand Maß hier einen kleinen Abstand misst. Das angepasste Randmaß hingegen misst einen Abstand von 0 für gleiche Clusterlevel, mit steigender Abweichung der Clusterlevel steigt auch der gemessene Abstand.

Abbildung 8.34 zeigt den gemessenen Abstand des Fowlkes-Mallows– und Jaccard–Maßes. Es ist dabei zu beachten, dass im Gegensatz zu Abbildung 8.33 die Perspektive um 180 gedreht wurde.

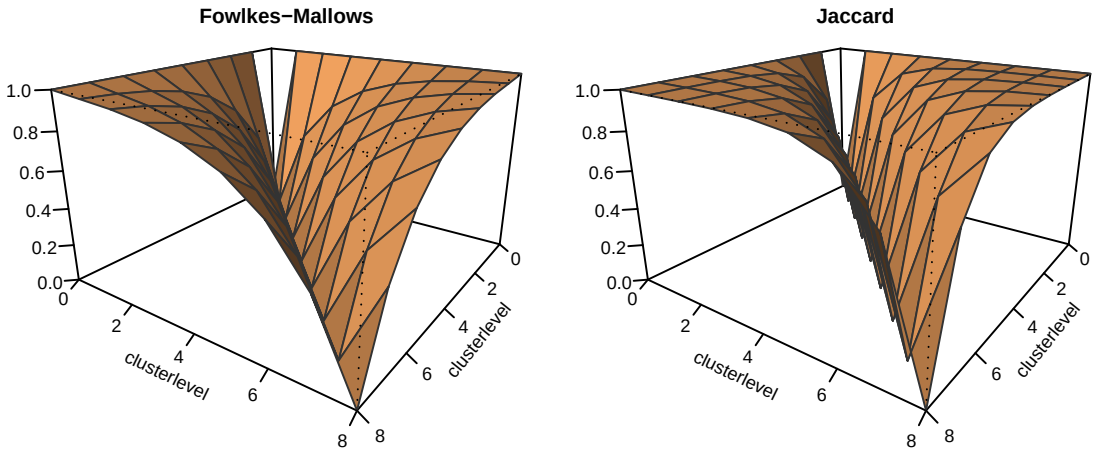


Abb. 8.34: Knotenstrukturelle Paarmaße II auf Hierarchiegraphen

Man erkennt das Greifen des Sonderfalls, in dem der Abstand zur Singletoncluster-ing immer mit 1 gemessen wird. Ansonsten zeigen diese Maße kein ungewöhnliches Verhalten.

Schnittmaße

Abbildung 8.35 zeigt das asymmetrische F– und Meila–Heckerman–Maß. Da sowohl Clusterlevel i mit Clusterlevel j und j mit i verglichen wird, wird für die asymmetrischen Maße jeweils nur eine Grafik benötigen.

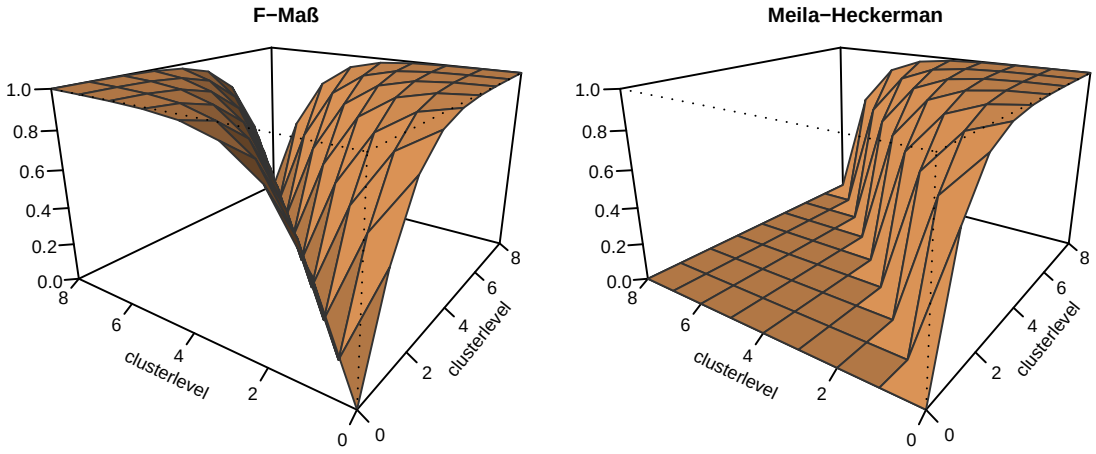


Abb. 8.35: Asymmetrische knotenstrukturelle Schnittmaße auf Hierarchiegraphen

Das F-Maß zeigt kein asymmetrisches Verhalten wohingegen das Meila–Heckerman Vergröberungen immer mit dem Abstand Null misst. Dieses Verhalten ist ein sehr großer Nachteil des Maßes.

Die gemessenen Abstände des Maximum–Match– und van Dongen–Maßes zeigt Abbildung 8.36.

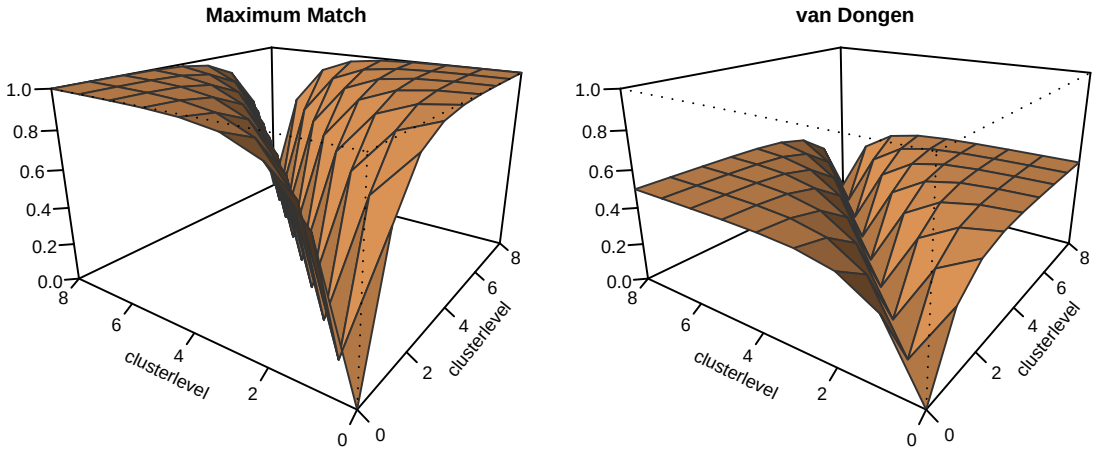


Abb. 8.36: Symmetrische knotenstrukturelle Schnittmaße auf Hierarchiegraphen

Auffällig ist, dass das van Dongen–Maß nur einen maximalen Abstand von 0.5 misst. Dieser maximale Abstand wird außerdem schon für hohe Clusterlevel gemessen. Das Maximum–Match–Maß zeigt hingegen ein intuitives Verhalten.

Entropiemaße

Abbildungen 8.37 und 8.38 zeigen die gemessenen Abstände der Entropiemaße, wobei die Variation der Information mit $\log_2(n)$ normiert ist.

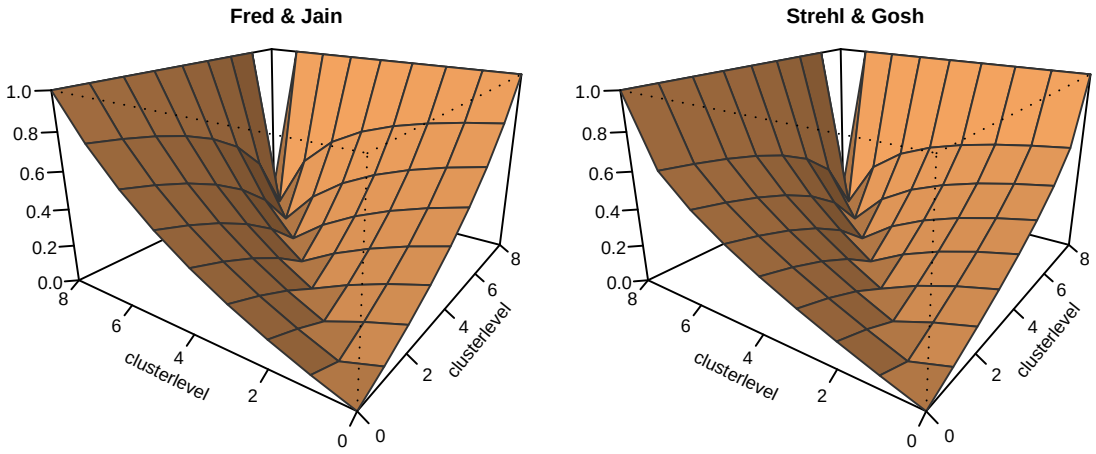


Abb. 8.37: Knotenstrukturelle Entropiemaße I auf Hierarchiegraphen

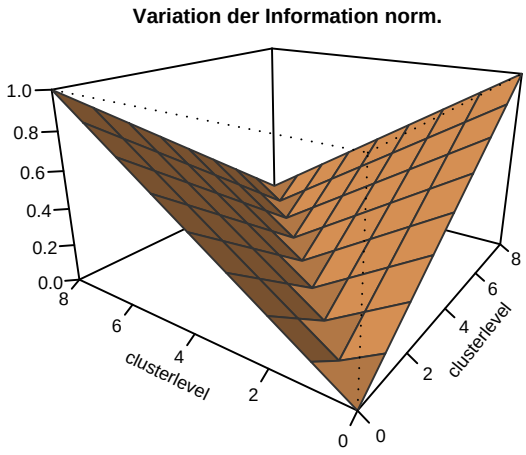


Abb. 8.38: Knotenstrukturelle Entropiemaße II auf Hierarchiegraphen

Die Variation der Information zeigt hier ein lineares Verhalten. Mit steigendem Abstand der Clusterlevel wächst auch der Abstand. Die Maße Fred & Jain und Strehl & Gosh zeigen wieder ein ähnliches Verhalten. Auffällig ist, dass mit sinkendem Clusterlevel Abweichungen des Clusterlevels zu kleineren gemessenen Abständen als bei hohen Clusterleveln führen.

8.4.4 Graphstruktureller Abstand

Bei diesen Graphen zeigen die knotenstrukturellen Maße größtenteils ein zufriedenstellendes Verhalten. Nun wird überprüft, welche Auswirkungen die graphstrukturelle Erweiterungen der Maße bezüglich Verfeinerung bzw. Vergröberung haben. Wie im Setup beschrieben, ist ein mittlerer Clusterlevel intuitiv signifikanter als ein sehr hoher oder sehr niedriger. Daher sollte graphstrukturelle Maße eine Abweichung von einem mittleren Clusterlevel mit einem höheren Abstand messen als die Abweichung von einem niedrigen bzw. hohen Clusterlevel.

Graphstrukturelle Paarmaße

Abbildung 8.39 zeigt die graphstrukturellen Paarmaße Rand als angepasste und nicht angepasste Versionen. Die Perspektive ist bei diesen Grafiken zur besseren Übersichtlichkeit um 180 gedreht.

Das graphstrukturelle Rand–Maß invertiert den gemessenen Abstand im Gegensatz zu seiner knotenstrukturellen Version. Diese Version misst hohe Abstände bei niederigen Clusterleveln, die knotenstrukturelle bei hohen Clusterleveln. Dies lässt sich dadurch erklären, dass bei niedrigem Clusterleveln die überwiegende Zahl der Kanten bereits Intraclusterkanten sind. Da die lokalen Paarzählungsmengen nur verbun-

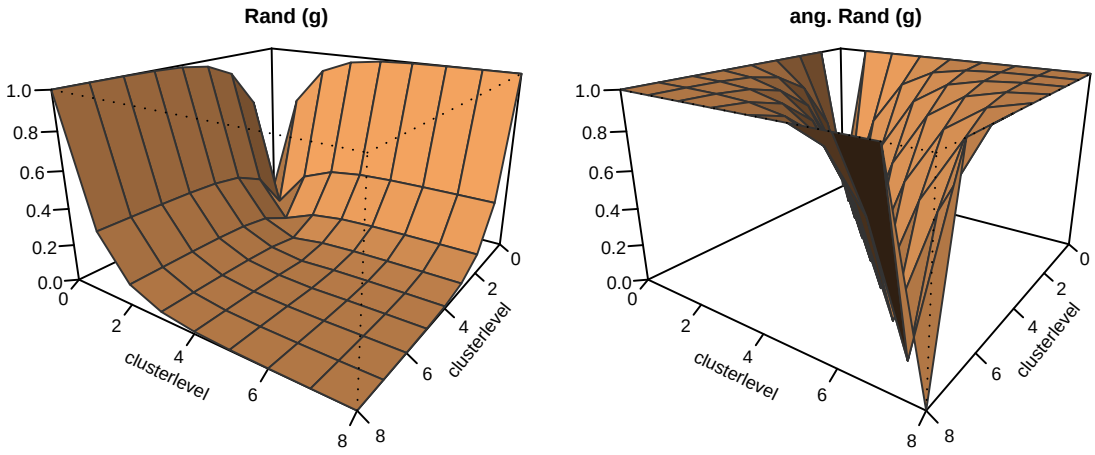


Abb. 8.39: Graphstrukturelle Paarmaße I auf Hierarchiegraphen

dene Knotenpaar betrachtet, bewertet das graphstrukturelle Rand–Maß Abstände zwischen hohen Clusterleveln sehr niedrig. Die graphstrukturelle Variante des angepassten Rand Maßes zeigt hingegen keinen Unterschied zu der knotenstrukturellen Version.

Die gemessenen Abstände des graphstrukturellen Fowlkes–Mallows– und Jaccard–Maßes zeigt Abbildung 8.40. Auch hier ist die Ansicht im Gegensatz zu den restlichen Grafiken um 180 gedreht.

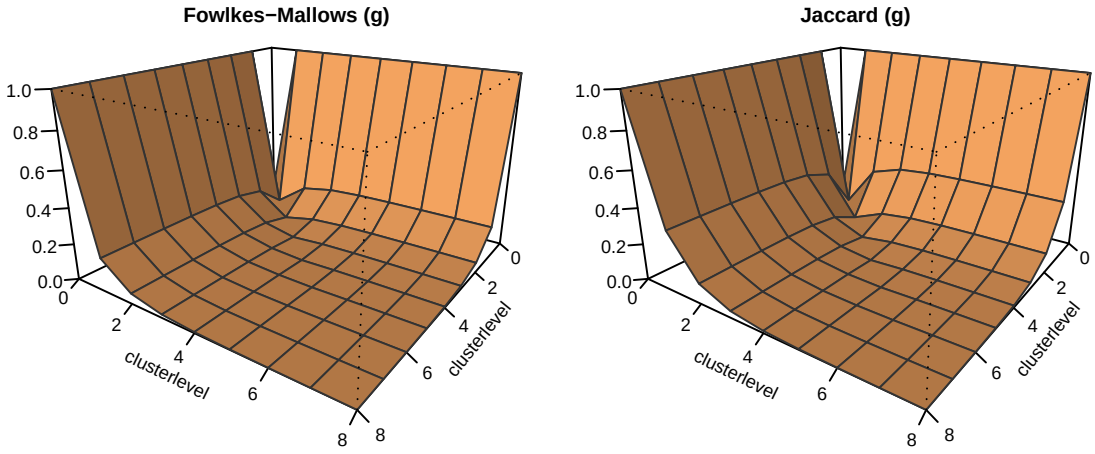


Abb. 8.40: Graphstrukturelle Paarmaße II auf Hierarchiegraphen

Auch diese Maße messen den Abstand ähnlich dem Rand–Maß. Sie unterscheiden sich von den knotenstrukturellen Versionen im Ergebnis erheblich. Der Grund hierfür ist der gleiche wie beim Rand-Maß.

Graphstrukturelle Schnittmaße

Abbildung 8.41 zeigt die graphstrukturellen Schnittmaße. Das F– und das Meila–Heckerman–Maß sind trotz Asymmetrie wegen der Symmetrie der Clusterlevel nur einmal vorhanden.

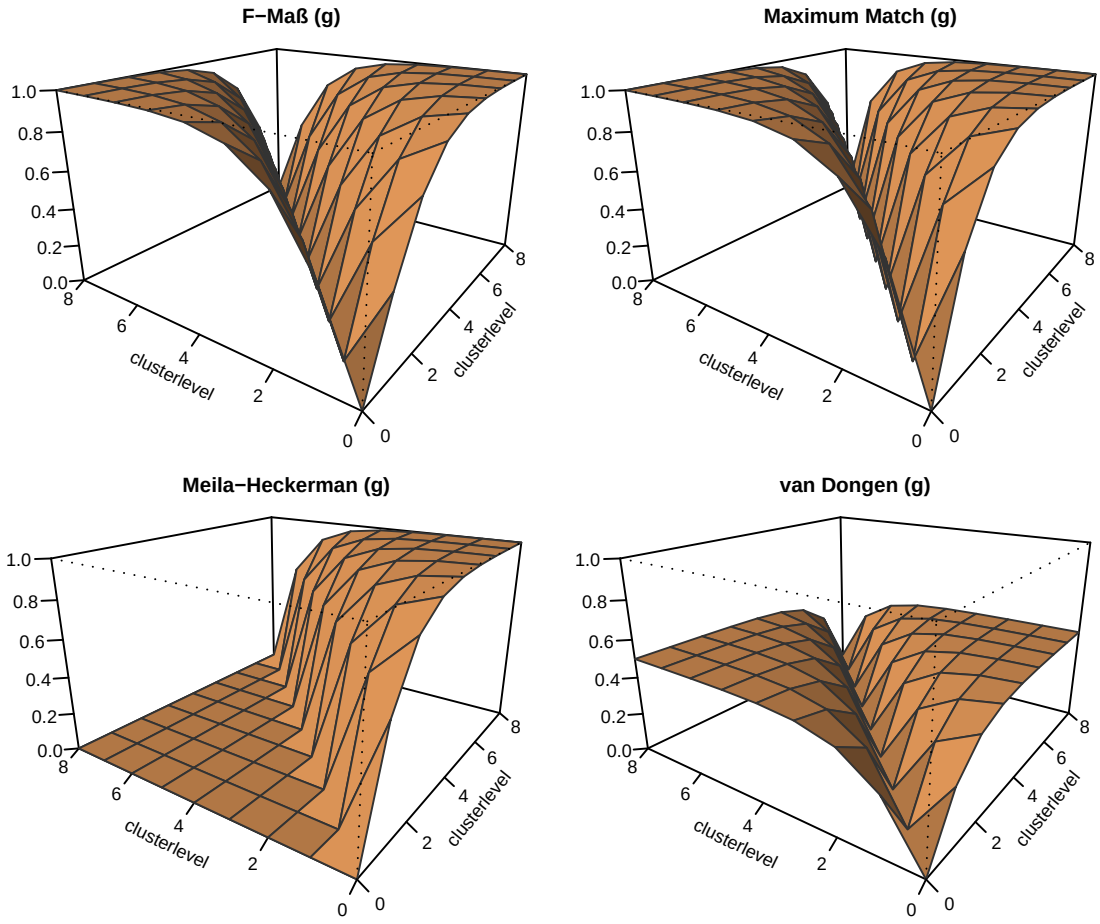


Abb. 8.41: Graphstrukturelle Schnittmaße auf Hierarchiegraphen

Kein Maße zeigt groï¿½ Unterschiede zu den entsprechenden knotenstrukturellen Versionen. Dies liegt an der Tatsache, dass rekursive Graphen annähernd regulär sind.

Kantenentropiemaße

Die graphstrukturellen Entropiemaße sind in Abbildung 8.42 dargestellt. Die Variation der Information ist mit $\log_2(n)$ normiert.

Wie bei den graphstrukturellen Schnittmaßen zeigt sich bei den Kantenentropiemaßen kein Unterschied zu den knotenstrukturellen Entropiemaßen. Auch dies lässt sich

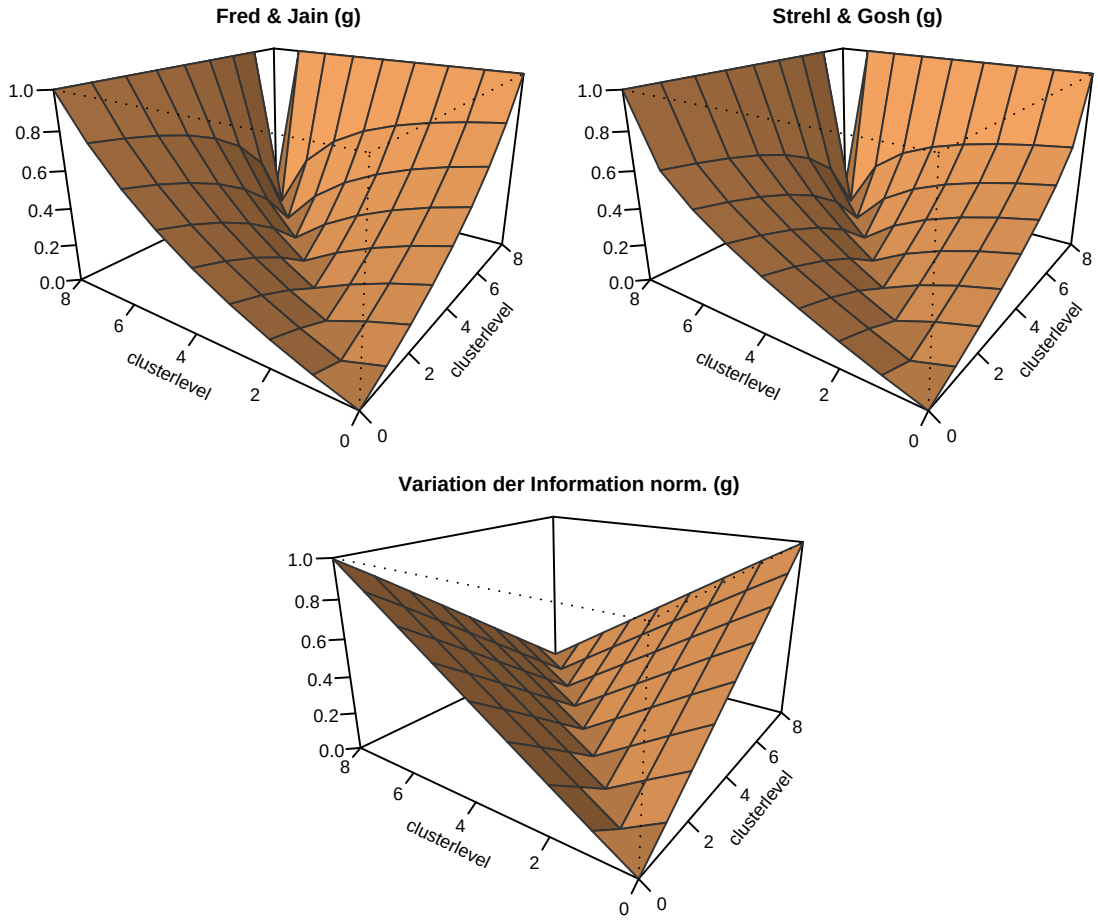


Abb. 8.42: Graphstrukturelle Entropiemaße auf Hierarchiegraphen

mit der Regularität der Hierarchiegraphen erklären.

Editiermengendifferenz

Nun soll noch der gemessene Abstand der vereinigungs- und knotennormierten Editiermengendifferenz betrachtet werden. Abbildung 8.43 zeigt diesen Abstand.

Die knotennormierte Editiermengendifferenz misst nur große Abstände, wenn eine Clusterung die 1-Clusterung ist. Damit ähnelt es der gemessenen Indexdifferenz auf Basis der Performance (Abbildung 8.32). Bei genauerer Überlegung erscheint dieses Verhalten auch logisch, da in diesem Falle die knotennormierte Editiermengendifferenz sich wie diese Indexdifferenz berechnet.

Die vereinigungsnormierte Editiermengendifferenz hingegen verhält sich wie ein knotenstrukturelles Abstandsmaß.

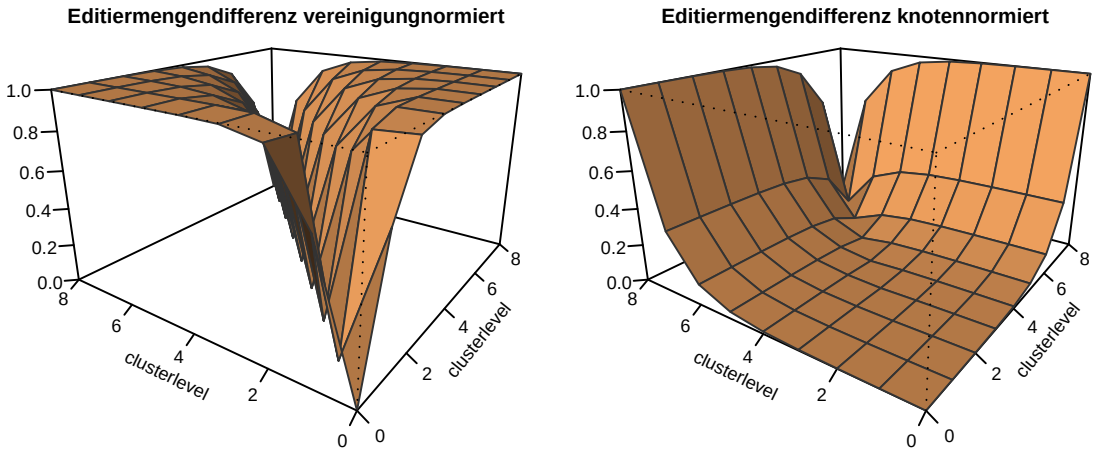


Abb. 8.43: Editiermengendifferenz auf Hierarchiegraphen

Strukturelle Indexmaße

Zum Abschluß zeigt Abbildung 8.44 die gemessenen Abstände des strukturellen Indexmaße $\mathcal{FJ}_{\mathcal{ID}}$. In der Abbildung ist in der rechten Grafik Performance alleinige Grundlage der Indexdifferenz, in der linken der Durchschnitt aus Coverage, Performance und durchschnittlicher Interclusterconductance.

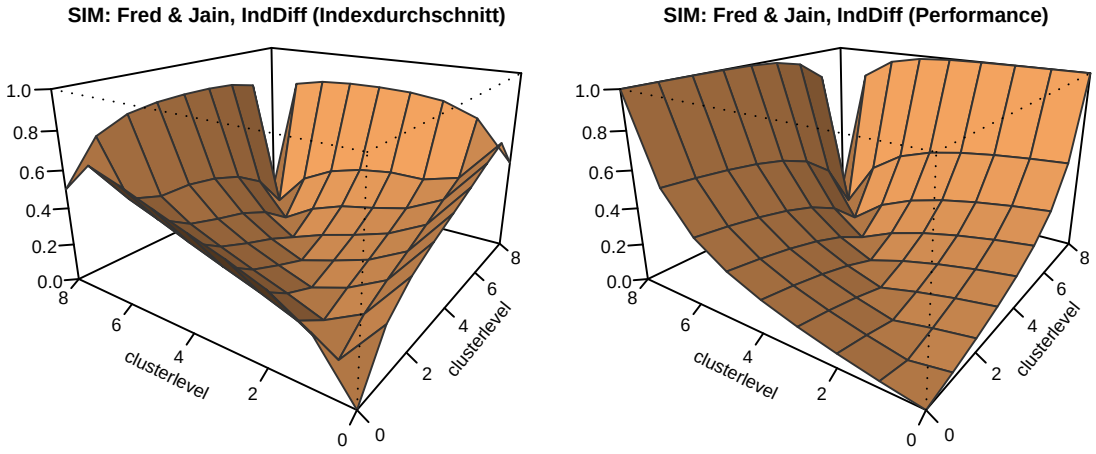


Abb. 8.44: Strukturelle Indexmaße $\mathcal{FJ}_{\mathcal{ID}}$ auf Hierarchiegraphen

An diesen Grafiken sieht man den Einfluß der Indexmaß auf den gemessenen Abstand. Da auf Hierarchiegraphen die Indizes die intuitiv signifikantesten Clusterlevel nicht merklich höher als die anderen bewerten, führt die Addition der Indexdifferenz nicht zu erheblich besseren Ergebnissen.

Der gemessene Abstand des strukturellen Indexmaßes $\mathcal{NVI}_{\mathcal{ID}}$ zeigt Abbildung 8.44

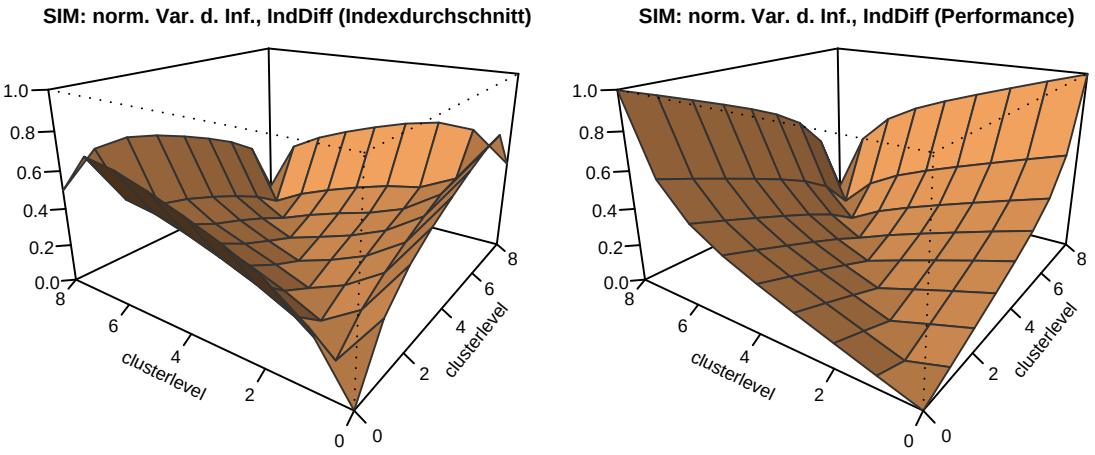


Abb. 8.45: Strukturelle Indexmaße $\mathcal{NVI}_{\mathcal{ID}}$ auf Hierarchiegraphen

Auch hier zeigt sich, dass die Addition eines qualitativen Maßes zu einem knotenstrukturellen nicht zwangsläufig zu besseren Ergebnissen führt.

8.4.5 Ergebnisse dieser Testreihe

Die qualitativen Maße zeigen für diese Tests ein den Indizes entsprechendes Bild. Da die Indizes keinen Clusterunglevel eindeutig als den Besten bewerten, ist der gemessene qualitative Abstand für mittlere Clusterlevel sehr klein. Nur zu der Singleton– bzw. 1-Clusterung wird ein signifikanter Abstand gemessen.

Diese Testreihe hat gezeigt, dass die knotenstrukturellen Abstandsmaße Verfeinerungen gut darstellen. Vor allem das lineare Anwachsen des normierten Variation der Information Maß ist vom knotenstrukturellen Standpunkt her interessant.

Bei den graphstrukturellen Maßen zeigt sich ein zwiespältiges Bild: Die graphstrukturellen Schnitt– und Entropiemaße zeigen das gleiche Verhalten wie die knotenstrukturellen Versionen, da Hierarchiegraphen annähernd regulär sind. Die Maße auf Basis der lokalen Paarzählung hingegen messen Abstände nur für untere Clusterlevel, da für höhere Clusterlevel nahezu alle Kanten Intraclusterkanten sind. Die Editiermengendifferenz zeigt vereinigungsnormiert ein Verhalten wie ein knotenstrukturelles Maß und die knotennormierte Version ist hier ein qualitatives Abstandsmaß und zeigt somit das selbe Verhalten wie ein solches Maß. Durch die wenig eindeutigen Indizes ist die Aussage der strukturellen Indexmaße begrenzt.

9 Ergebnisse

Dieses Kapitel verknüpft Erkenntnisse der axiomatischen Überlegungen in Kapitel 6 und 7 mit denen der Experimente aus Kapitel 8. Diese Analyse ist erneut nach qualitativen, knoten- und graphstrukturellen Maßen unterteilt.

Qualitative Abstandsmaße

Da bei den qualitativen Abstandsmaßen die Problematik der mathematischen Modellierung der menschlichen Intuition an die Indizes übergeben wird, können die Ergebnisse für diese Maße nicht besser als die Indizes sein. Bei den Experimenten in Kapitel 8 hat sich gezeigt, dass der Unterschied in der Qualität der Clusterungen häufig nicht sehr groß ist, sodass man geneigt sein könnte, lastige Maße oder Maße mit Wachstumsfaktoren $f < 1$ zu nutzen.

Knotenstrukturelle Abstandsmaße

Bei den knotenstrukturellen Abstandsmaßen bestätigt sich der Verdacht, dass die Entropiemaße ein sehr vielversprechender Ansatz für das Messen von (knotenstrukturellen) Abständen sind. Sowohl Paar- als auch Schnittmaße haben jeweils signfikante Nachteile wie Asymmetrie, mangelnde Informationsvollständigkeit oder eine starke Clusterzahlabhängigkeit.

Vom axiomatischen Standpunkt scheint die Variation der Information das beste Maß für den knotenstrukturellen Clusterungsvergleich zu sein, doch Abbildung 8.5 zeigt, dass dieses Maß für zwei Zufallsclusterungen im Erwartungswert nicht den maximalen Abstand misst. Gründe hierfür sind zum einen die Normierung mit $\log_2 n$, andererseits wurde in den Kapitel 3 und 6 gezeigt, dass der Verbandsansatz für Graphclusterungen wenig intuitiv ist.

Die Maße Fred & Jain bzw. Strehl & Gosh zeigen hingegen trotz anscheinender axiomatischer Unterlegenheit ein intuitiveres Verhalten, sie messen den Abstand zwischen Singleton- und 1-Clusterung, den beiden komplementären Clusterungen $\mathcal{C}^\times$ und $\mathcal{C}^\perp$ und zwei Zufallsclusterungen (Abbildung 8.5) maximal. In den Testreihen zu Algorithmenclusterungen (Kapitel 8.2), lokaler Minimierung (Kapitel 8.3) und Verfeinerungen (Kapitel 8.4) zeigen diese Maße im Rahmen eines knotenstrukturellen Maßes sehr intuitive Ergebnisse.

Weiterhin fällt in den weiteren Testreihen auf, dass sich die Ergebnisse der beiden Maße sehr ähneln. Allerdings misst das Strehl & Gosh Maß den Abstand zur Singleton-Clusterung konstruktionsbedingt immer maximal. Aus diesem Grund erscheint für die knotenstrukturelle Abstandsmessung bei einem statischen Clusterungsvergleich das Fred & Jain–Maß die erste Wahl zu sein.

Allerdings ist eindeutig gezeigt worden, dass die Nachteile bei ausschließlicher Betrachtung der Knotenmenge nicht nur theoretischer Natur sind. In den Testreihen zu Zufallsclusterungen (Kapitel 8.1) und lokaler Minimierung (Kapitel 8.3) zeigt sich dies deutlich. Kein knotenstrukturelles Maß kann die intuitiven Unterschiede der jeweiligen Vergleiche messen. Der Grund dafür ist, dass vom knotenstrukturellen Standpunkt kein Unterschied zwischen den jeweiligen Vergleichen ist.

Graphstrukturelle Abstandsmaße

Das Fazit für die graphstrukturellen Erweiterungen ist ernüchternd. Bis auf wenige Ausnahmen bringen die Erweiterungen in den Testreihen keine deutlichen Verbesserungen mit sich. Bei den Schnitt- und Entropiemaßen liegt das daran, dass für reguläre Graphen die graphstruktuelle Version eines Maßes konstruktionsbedingt der knotenstrukturellen entspricht.

Die Erweiterung der Paarzählungsmaße durch Verwendung von lokalen Paarzählungsmengen erzeugt sogar schlechtere Ergebnisse. Die graphstrukturellen Versionen des Fowkles–Mallows– und Jaccard–Maßes verhalten sich beispielsweise in Abbildung 8.6 weniger intuitiv als die jeweiligen knotenstrukturellen Pendants. Bei dem Vergleich von Initial- zu Zufallsclusterung nimmt dort der gemessene Abstand zu, obwohl der Abstand intuitiv kleiner wird.

Bei der Editiermengendifferenz sieht man an der knotennormierten Version, dass ein graphstrukturelles Maß in Spezialfällen – hier Verfeinerungen von Clusterungen – einem qualitativen Maß entsprechen kann. Die Experimente haben ferner gezeigt, dass sich die knotennormierte Version ähnlich einem knotenstruturellem Maß verhält. Das Maß erkennt weder einen Unterschied zwischen den beiden Vergleichen bei den Zufallscluster–Testreihen, noch bei den lokalen Minimierungen. Somit scheint dieses Maß keine signifikanten Vorteile für die Abstandsmessung mit sich zu bringen. Die vereinigungsnormierte Version ist vom experimentellen Standpunkt her der knotennormierten Version überlegen, da zumindestens für den Fall der lokalen Minimierung das Maß die intuitiven Unterschiede des Abstands zwischen den beiden Vergleichen misst. Für die Testreihen mit Zufallsclusterungen hingegen entspricht das Maß nicht der Intuition.

Die strukturellen Indexmaße sind kein Garant für gute Ergebnisse, da sie ähnlich den qualitativen Maßen die Problematik der qualitativen Sensivität (Axiom 19) an einen Index übergeben.

10 Abschließende Bemerkungen

10.1 Zusammenfassung

In dieser Arbeit wurde der Vergleich von Clusterungen auf gleichen Graphen systematisch analysiert. Dabei wurden zunächst bisherige Lösungsansätze diskutiert und welche Probleme bzw. Nachteile diese Ansätze mit sich bringen. Es wurde festgestellt, dass bisherigen Abstandsmaße zur Berechnung des Abstands die Kantenmenge der Graphen ignorieren. Dass dies auch beim statischen Clusterungsvergleich ein Nachteil ist, wurde an Beispielen gezeigt.

Da in Anwendungsfällen ein Abstand zweier Clusterungen nur von der Signifikanz der beiden abhängig sein kann, wurde gefolgert, dass es im Bereich des Clusterungsvergleich nicht nur eine Möglichkeit der Abstandsmessung existiert. Aus diesem Grund wurden die drei Abstandsarten qualitativer, knoten- und graphstruktureller Abstand hergeleitet, die jeweils einer anderen Intuition entsprechen. Diese Dreiteilung erforderte, für jede Abstandsart getrennt zu untersuchen, wann ein Abstandsmaß die Gleichheit zweier Clusterungen und einen minimalen bzw. maximalen Abstand zwischen zwei Clusterungen messen soll.

Daraufhin wurden einige axiomatische Überlegungen ausgeführt, wie sich Abstandsmaße verhalten sollten. Hier konnte gezeigt werden, dass viele Axiome auf den ersten Blick sinnvoll erscheinen, bei genauerer Analyse konnte jedoch ein wenig intuitives Verhalten von Maßen, die bestimme Axiome erfüllen, an Beispielen belegt werden.

Ein Literaturrecherche ergab, dass zwar viele Abstandsmaße in der Literatur zu finden sind, diese jedoch immer den knotenstrukturellen Abstand messen. Für diese Abstandsmaße wurde überprüft, welche der eingeführten Axiome die jeweiligen Maße erfüllen. Um auch die anderen Abstandsarten abzudecken, wurden sowohl qualitative Abstandsmaße, graphstrukturelle Erweiterungen der bisherigen Maße und neue graphstrukturelle Maße eingeführt.

Testreihen ergaben, dass die konstruktionsbedingten Nachteile der knotenstrukturellen Maße nicht nur theoretisch sind. Beispielsweise messen knotenstrukturelle Maße den gleichen Abstand zwischen zwei Zufallsclusterungen wie zwischen einer signifikanten und einer zufällig generierten Clusterung. Allerdings ergaben die Tests auch, dass für den Fall, dass die untersuchten Clusterungen bezüglich regulärer Graphen definiert sind, die graphstrukturellen Erweiterung meist wirkungslos bleiben.

Allerdings ergaben die experimentellen Untersuchungen, dass das Maß von Fred & Jain für die knotenstrukturelle Abstandsmessung sehr intuitive Ergebnisse liefert.

10.2 Ausblick

Die hier vorgestellten graphstrukturellen Erweiterungen lieferten in den Testreihen keine signifikant besseren Ergebnisse als die knotenstrukturellen Versionen. Aus diesem Grund ist die Fragestellung interessant, ob andere graphstrukturelle Erweiterungen existieren. Hierbei sollte ein Hauptaugenmerk auf einer Erweiterung des vielversprechenden Fred & Jain–Maßes liegen.

Da kein hier behandeltes Maß den graphstrukturellen Abstand zufriedenstellend misst, kann ebenso das Design neuer graphstruktureller Maße von Interesse sein. In diesem Zusammenhang sei noch einmal auf das Axiom 22 der graphstrukturellen Elementaroperationen verwiesen. Das Design und die Analyse eines Maßes, dass dieses Axiom erfüllt, ist eine herausfordende Aufgabe. Allerdings ist eine intensivere axiomatische Analyse des graphstrukturellen Abstands für das Design neuer Maße hilfreich und nötig.

Man kann ein Abstandsmaß für die Bewertung einer Clusterung $\mathcal{C}$ nutzen. Der gemessene Abstand zu der minimalen Clusterung $\mathcal{C}_{\text{MIN}}$ kann als Bewertung der Clusterung $\mathcal{C}$ aufgefasst werden. Die Analyse des Zusammenhangs zwischen Abstandsmaß und Bewertungsfunktion einer Clusterung ist ebenso eine sehr interessante Aufgabe. Dabei ist die Frage, welche Auswirkungen die hier eingeführte Dreigliederung der Abstandsarten für die Bewertung einer Clusterung hat.

Der gleichen Problematik des Zusammenhangs zwischen Abstand und Bewertung kann sich auch von der anderen Seite genähert werden. Dazu verfolgt man bezüglich Indizes einen ähnlich systematischen Ansatz wie in dieser Arbeit: Eine Analyse der bisherigen Indizes, welche Arten von Clusterungen die Indizes als optimal bewerten und welche Intuition sie somit repräsentieren. Die so gegliederten Maße sollen dann mit der hier vorgestellten Gliederung in Einklag gebracht werden.

In dieser Arbeit wurde der statische Clusterungsvergleich systematisch analysiert. Eine Analyse des dynamischen Clusterungsvergleiches steht somit aus. Dabei könnte man zunächst in einem ersten Schritt nicht wie in dieser Arbeit die Gleichheit von Knoten- und Kantenmenge, sondern lediglich die Gleichheit der Knotenmenge fordern. Hat man dieses Problem zufriedenstellend gelöst, könnte man versuchen, eine Lösung für den dynamischen Clusterungsvergleich zu finden.

Abschließend kann man sagen, dass der Vergleich von Clusterungen auf Graphen eine umfangreiches Themengebiet ist, in dem noch viele ungelöste Probleme existieren.

A Axiomübersicht

Hier findet sich eine Übersicht aller in Kapitel 5 vorgestellten Axiome inklusive einer Kurzbeschreibung, sowie eine Übersicht, welche Axiome von den in Kapitel 7 diskutierten Maßen erfüllt werden. Dabei werden nur die qualitativen und knotenstrukturellen Abstandsmaße berücksichtigt.

Nr	Axiom	Maße	Kurzbeschreibung
1	Symmetrie	alle	$d(\mathcal{C},\mathcal{C}') = d(\mathcal{C}',\mathcal{C})$
2	Identität	alle	$d_a(\mathcal{C},\mathcal{C}') = 0 \Rightarrow \mathcal{C} =_a \mathcal{C}'$ mit $a \in \{q,k,g\}$
3	Positivität	alle	$d(\mathcal{C},\mathcal{C}') > 0$
4	Dreiecksungl.	alle	$d(\mathcal{C},\mathcal{C}'') \leq d(\mathcal{C},\mathcal{C}) + d(\mathcal{C}',\mathcal{C}'')$
5	1–Beschränktheit	alle	$d(\mathcal{C},\mathcal{C}') \leq 1$
6	1–Maximalität	alle	$\exists \mathcal{C},\mathcal{C}' : d(\mathcal{C},\mathcal{C}') = 1$
7	Grenzw. 1–Maxi.	alle	$\exists \mathcal{C},\mathcal{C}' : \lim_{n\to\infty} d(\mathcal{C},\mathcal{C}') = 1$
8	Clusterzahlunabh.	alle	Der Abstand zweier Zufallsclusterung ist unabhängig von der Clusteranzahl.
9	Informationsvoll.	alle	Alle zur Verfügung stehenden Informationen werden genutzt.
10	pol. berechenbar	alle	Der Abstand ist polynomiell berechnbar.
11	Wachstumsfaktor f	d_q	$i(\mathcal{C}') = x \cdot i(\mathcal{C}'') \Leftrightarrow d_q(\mathcal{C},\mathcal{C}') = f(x) \cdot d_q(\mathcal{C},\mathcal{C}'')$
12	Lastigkeit	d_q	Abweichungen des Indexes im guten oder schlechten Bereich führen zu einem größerem Abstand
13	Beidlastigkeit	d_q	Kombination Gut- und Schlechtlastigkeit
14	element. Äquidist.	d_k	Abstände der Elementarop. sind gleich
15	Addi. bzgl. Verfein.	d_k	Abstand ist additiv für Verfeinerungen
17	Addi. bzgl. Produkt	d_k	$d_k(\mathcal{C},\mathcal{C}') = d_k(\mathcal{C},\mathcal{C}\times\mathcal{C}') + d_k(\mathcal{C}\times\mathcal{C}',\mathcal{C}')$
16	Addi. bzgl. Verein.	d_k	$d_k(\mathcal{C},\mathcal{C}') = d_k(\mathcal{C},\mathcal{C}\oplus\mathcal{C}') + d_k(\mathcal{C}\oplus\mathcal{C}',\mathcal{C}')$
18	Konvexe Addi.	d_k	Abstand ist unabh. von der restlicher Clusterung
19	Qual. Sensivität	d_g	Abstand ist für gleiche Veränderung der Partition von der Signifikanz der Clusterung abhängig
20	Knotengradabh.	d_g	Abstand ist vom Knotengrad abhängig
21	Intraknotengradabh.	d_g	Abstand ist vom Intraclustreknotengrad abhängig
22	G-str. Elementarop.	d_g	Übertragung des Verbandsansatzes auf Graphoperationen

Tabelle A.1: Übersicht der Axiome aus Kapitel 5

Die Tabellen A.2 und A.3 geben einen Überblick über die erfüllten Axiome der qualitativen bzw. knotenstrukturellen Maße.

Maß	1-4	5	6	7	8	9	10	11	12	13
Indexquotient	✓	✓	✓	✓	–	✓	✓	–	✓	–
Indexdifferenz (ID)	✓	✓	✓	✓	–	✓	✓	linear	–	–
Gewurzelte ID	✓	✓	✓	✓	–	✓	✓	0.5-exp.	–	–
Schlechtlastige ID	✓	✓	✓	✓	–	✓	✓	–	✓	–
Gutlastige ID	✓	✓	✓	✓	–	✓	✓	–	✓	–
Beidlastige ID	✓	✓	✓	✓	–	✓	✓	–	–	✓

Tabelle A.2: Erfüllte Axiome der qualitativen Abstandsmaße

Maß	1-4	5	6	7	8	9	10	14	15	16	17	18
Rand	✓	✓	✓	✓	–	✓	✓	–	✓	–	✓	–
ang. Rand	✓	–	–	–	–	✓	✓	–	–	–	–	–
Fowlkes-Mallows	✓	✓	✓	✓	–	✓	✓	–	–	–	–	–
Jaccard	✓	✓	✓	✓	–	✓	✓	–	–	–	–	–
F-Maß	–	✓	–	✓	–	–	✓	–	–	–	–	–
Meila–Heckerman	–	✓	–	✓	–	–	✓	–	–	–	–	–
Maximum Match	✓	✓	–	✓	–	–	✓	–	–	–	–	–
Van Dongen	✓	–	–	–	–	–	✓	–	–	✓	✓	✓
Van Dongen norm.	✓	✓	–	✓	–	–	✓	–	–	✓	✓	✓
Streh & Gosh	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–
Fred & Jain	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–
Var.d.Inf.	✓	–	–	–	✓	✓	✓	–	✓	✓	✓	✓
Var.d.Inf. norm.	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	✓

Tabelle A.3: Erfüllte Axiome der knotenstrukturellen Abstandsmaße

B Ergänzung: Abbildungen

In diesem Anhang sind zusätzliche Abbildungen von gemessenen Abständen zu finden, die in Kapitel 8 nicht abgebildet sind. Die Gründe hierfür sind an den entsprechenden Stellen angegeben.

B.1 Initial- und Algorithmenclusterung

Qualitative Maße

Zur Vervollständigung zeigt B.1 noch die Varianten der Indexdifferenz für den Vergleich zwischen Initial– und Algorithmusclusterung auf Gaussgeneratoren aufgeführt. Grundlage ist auch der Durchschnitt aus Coverage, Performance und durchschnittlicher Interclusterconductance.

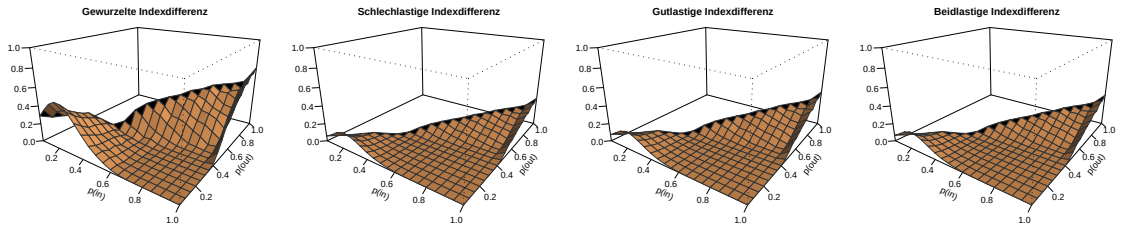


Abb. B.1: Varianten der Indexdifferenz: Initial- gegen Algorithmusclusterung

Die Abbildungen zeigen die Ausmaße der jeweiligen Lastigkeit auf die gemessenen qualitativen Abstände.

Graphstrukturelle Maße

Bei den Testreihen zum Vergleich von Initial– und Algorithmenclusterungen wurde auf die Angabe der Ergebnisse für die graphstrukturellen Maße verzichtet. Die Auswertungen finden sich in den Abbildungen B.2, B.3 und B.4.

Bei deren Betrachtung erkennt man, dass Unterschiede zu den jeweiligen knotenstrukturellen Versionen praktisch nicht auszumachen sind.

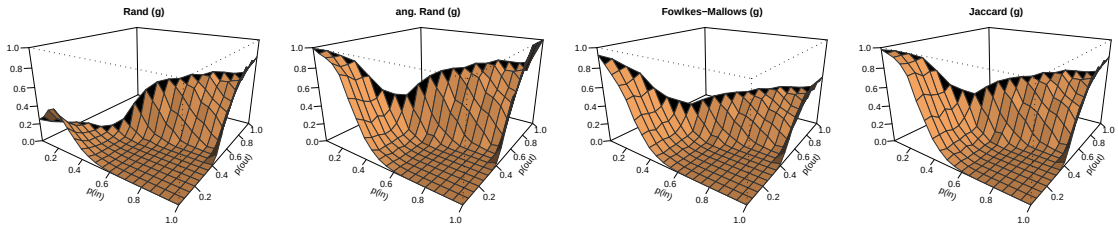


Abb. B.2: Graphstrukturelle Paarmaße: Initial- gegen Algorithmusclusterung

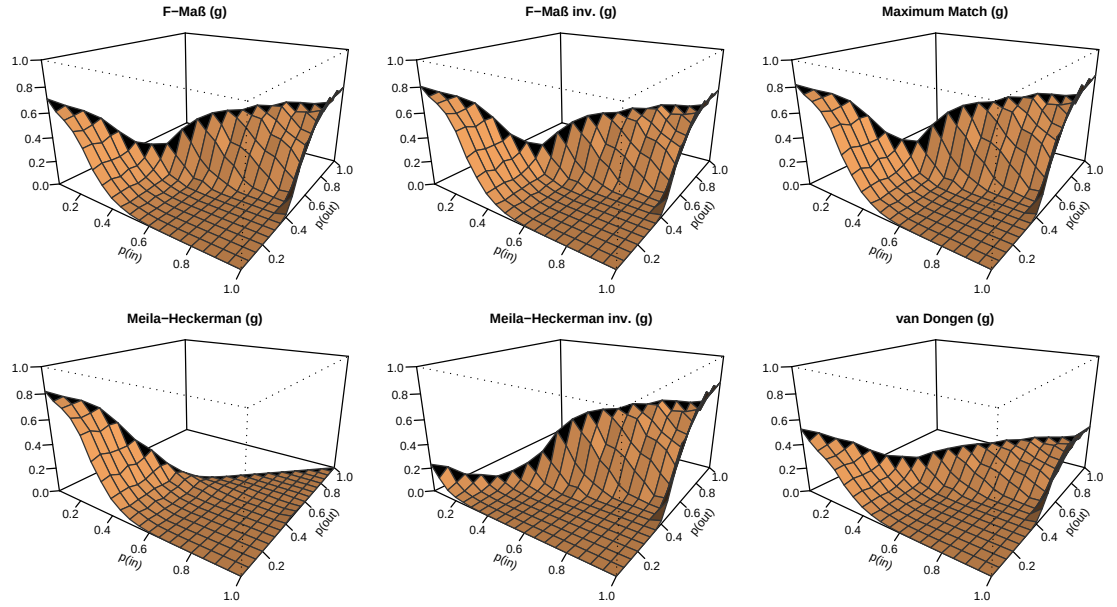


Abb. B.3: Graphstrukturelle Schnittmaße: Initial- gegen Algorithmusclusterung

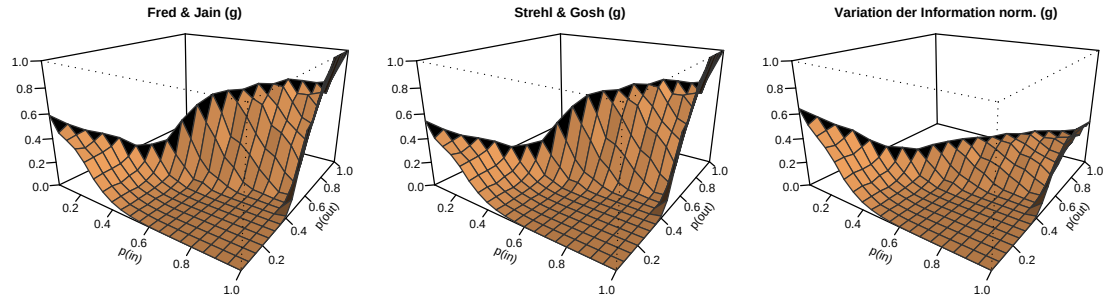


Abb. B.4: Graphstrukturelle Entropiemaße: Initial- gegen Algorithmusclusterung

B.2 Hierarchiegraphen

Zur Vervollständigung zeigt Abbildung B.5 die Varianten der Indexdifferenz zwischen den einzelnen Clusterleveln. Die obere Reihe an Grafiken nimmt als Grundlage den Durchschnitt aus Coverage, Performance und durchschnittlicher Interclusterconductance, die untere lediglich Perfomance.

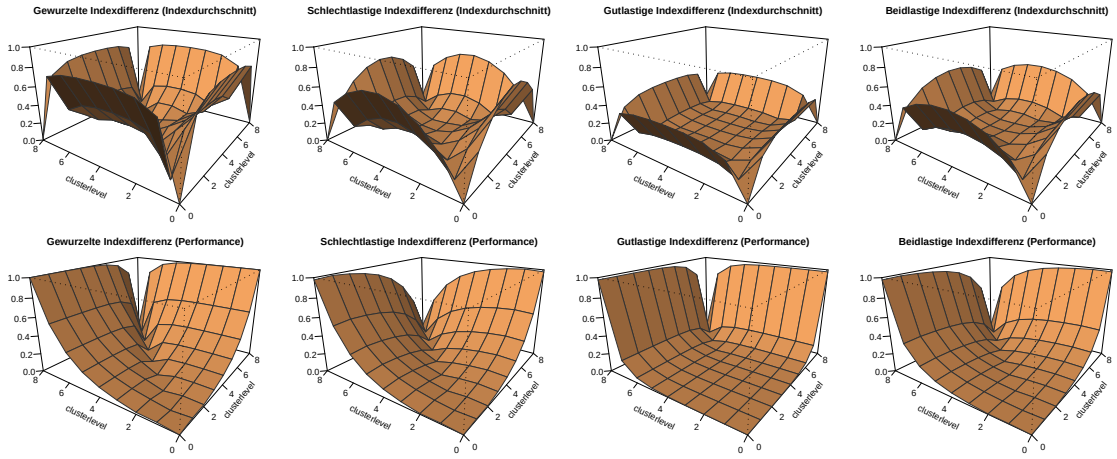


Abb. B.5: Varianten der Indexdifferenz auf Hierarchiegraphen

Zum Abschluß zeigt Abbildung B.6 den gemessenen Abstand des Indexquotienten. Links mit Grundlage des Durchschnitts aus Coverage, Performance und durchschnittlicher Interclusterconductance und rechts mit Performance als alleiniger Grundlage.

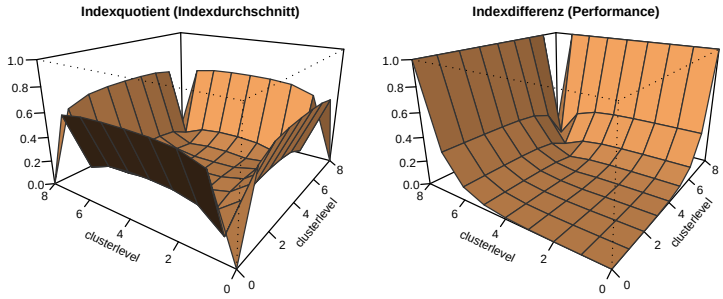


Abb. B.6: Indexquotient auf Hierarchiegraphen

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[BE05] Brandes, U. (Hrsg.) ; Erlebach, T. (Hrsg.): *Lecture Notes in Computer Science*. Bd. 3418: *Network Analysis: Methodological Foundations*. Springer-Verlag, 2005 http://springerlink.metapress.com/openurl.asp?genre=issue&issn=0302-9743&volume=3418

[Beh00] Behrends, E.: *Introduction to Markov Chains*. Vieweg, Braunschweig, 2000

[BGW03] Brandes, U. ; Gaertler, M. ; Wagner, D.: Experiments on Graph Clustering Algorithms. In: *Proceedings of the 11th Annual European Symposium on Algorithms (ESA'03)* Bd. 2832, 2003 (Lecture Notes in Computer Science). – ISSN 0302–9743, 568 - 579

[CSRL01] Cormen, T.H. ; Stein, C. ; Rivest, R.L. ; Leiserson, C.E.: *Introduction to Algorithms*. 2nd Edition. The MIT Press, 2001. – ISBN 0070131511

[Del04] Delling, D.: *Experimentelle Untersuchung von Clusterungsgeneratoren mittels Qualitätsindizes*, Fakultät für Informatik, Universität Karlsruhe, Studienarbeit, 2004

[Don00a] Dongen, S. van: A cluster algorithm for graphs. In: *Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May* (2000)

[Don00b] Dongen, S. van: Performance criteria for graph clustering and Markov cluster experiments. In: *Technical Report INS-R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May* (2000)

[DP02] Davey, B.A. ; Priestley, H.A.: *Introduction to Lattices and Order*. Cambridge University Press, 2002

[Düm03] Dümbgen, L.: *Stochastik für Informatiker*. Springer, 2003. – ISBN 3–540–00061–5

[FJ03] Fred, A.L.N. ; Jain, A.K.: Robust Data Clustering. In: *In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition,CVPR (3)*, 2003, S. 128–136

[FM83] Fowlkes, E.B. ; Mallows, C.L.: A method for comparing two hierarchical clusterings. In: *A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78, 553-569* (1983)

[FWE03] Fung, B. ; Wang, K. ; Ester, M.: Hierarchical Document Clustering Using Frequent Itemsets. In: *Proceedings of the SIAM International Conference on Data Mining.* (2003)

[GJ90] Garey, M.R. ; Johnson, D.S.: *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New York, NY, USA : W. H. Freeman & Co., 1990. – ISBN 0716710455

[HA85] Hubert, L. ; Arabie, P.: Comparing Partitions. In: *Journal of Classification, Vol. 2, pp 193-218* (1985)

[ICS] *Internet Systems Consortium*. http://www.isc.org/ops/ds/,

[Int04] International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. In: *Nature 431, 931-945* (2004)

[JMF99] Jain, A.K. ; Murty, M.N. ; Flynn, P.J.: Data clustering: a review. In: *ACM Computing Surveys 31 264-323* (1999)

[Jun05] Jungnickel, D.: *Graphs, networks and algorithms*. 2. Aufl. Springer, 2005

[Kle02] Kleinberg, J.: An impossibility theorem for clustering. In: *In Proc. of the 16th conference on Neural Information Processing Systems* (2002)

[LAB04] Li, S. ; Armstrong, C.M. ; Bertin, N.: A map of the interactome network of the metazoan C. elegans. In: *Science Jan 23;303(5657):540-3.* (2004)

[LOM04] Li, T. ; Ogihara, M. ; Ma, S.: On combining multiple clusterings. In: *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*. New York, NY, USA : ACM Press, 2004. – ISBN 1–58113–874–1, S. 294–303

[MA84] Morey, R. ; Agresti, A.: The Measurement of Classification Agreement: An Adjustment to the RAND Statistic for Chance Agreement. In: *Educational and Psychological Measurement 44:33-37* (1984)

[Mei03] Meila, M.: Comparing Clusterings by the Variation of Information. In: Schölkopf, B. (Hrsg.) ; Warmuth, M.K. (Hrsg.): *COLT* Bd. 2777, Springer, 2003 (Lecture Notes in Computer Science). – ISBN 3–540–40720–0, S. 173–187

[Mei05] Meila, M.: Comparing Clusterings - An Axiomatic View. In: *In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany* (2005)

[MH01] Meila, M. ; Heckerman, D.: An Experimental Comparison of Model-Based Clustering Methods. In: *Machine Learning* 42 (2001), Nr. 1/2, S. 9–29

[Ran71] Rand, W.M.: Objective criteria for the evaluation of clustering methods. In: *Journal of the American Statistical Association, 66:846-850* (1971)

[SG03] Strehl, A. ; Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. In: *J. Mach. Learn. Res.* 3 (2003), S. 583–617. – ISSN 1533–7928

[Sha48] Shannon, C.E.: A mathematical theory of communication. In: *Bell System Tech. J., 27:379-423, 623-656* (1948)

[Sör48] Sörensen, T.: A method for establishing groups of equal amplitude in plant sociology based similarity of species content. In: *Kong. Danske Vidensk. Selsk Biologische Skrifter 5: 1-34* (1948)

[SST02] Shamir, R. ; Sharan, R. ; Tsur, D.: Cluster Graph Modification Problems. In: *Proceedings of the 28th International Workshop on Graph-Theoretical Conecpts in Computer Science (WG* Bd. 2573, Springer-Verlag, 2002 (Lecture Notes in Computer Science), 379-390

[Weg03] Wegener, I.: *Komplexitätstheorie - Grenzen der Effizienz von Algorithmen.* 3. Aufl. Springer, 2003

[WW06] Wagner, S. ; Wagner, D.: Comparing Clusterings – An Overview / ITI Wagner, Faculty of Informatics, Universität Karlsruhe (TH). 2006 (2006-4). – Forschungsbericht

[Y] *The yFiles Java Library 2.4.* http://www.yworks.com/en/products_yfiles_about.htm,